Methods of Statistical Inference

Sinan Yıldırım

June 6, 2022

Contents

Chapter 1

Some basics in classical statistics

A Samples from an infinite population

Let $X_1, \ldots, X_n$ be a sample of size $n \geq 1$ from an infinite population with mean μ and variance σ^2. Equivalently, we say $X_1, \ldots, X_n$ are *independent and identically distributed (i.i.d.)* with a distribution F with mean μ and variance σ^2. We write

$$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} F, \quad E[X_i] = \mu, \quad V(X_i) = \sigma^2, \quad i = 1, \ldots, n. \tag{1.1}$$

Define the sample mean and the sample variance

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Several useful properties of $\bar{X}$ and S^2 follow.

Exercise 1.1. Assume (1.1) with $\sigma^2 < \infty$. Show that

(a) $E(\bar{X}) = \mu$ and $V(\bar{X}) = \frac{\sigma^2}{n}$.

(b) $\sum_{i=1}^{n}(X_i - \bar{X}) = 0$.

(c) $\bar{X}$ and $X_i - \bar{X}$ are uncorrelated.

(d) For each $a > 0$, we have $P(|\bar{X} - \mu| > a) \leq \sigma^2/(na^2)$ [Hint: Use the Chebyshev inequality, i.e. Markov inequality applied to the second central moment.]

(e) $E(S^2) = \sigma^2$.

Identifying relations between random variables plays a crucial role in statistics, such as in understanding the statistical behaviour of samples, devising practical methods, and assessing the quality of statistical procedures. One useful specification of the distribution of a random variable, which can be quite useful for that aim, is its *moment generating function*.

Definition 1.1 (Moment generating function)**.** The moment generating function of a random variable X is defined as

$$M_X(t) = E(e^{tX}), \quad t \in \mathbb{R}.$$

whenever the expectation exists. More generally, when $X = (X_1, \ldots, X_p)$ is a vector[1], then

$$M_X(t) = E(e^{t^T X}), \quad t = (t_1, \ldots, t_p) \in \mathbb{R}^p.$$

whenever the expectation exists.

The moment generating function of a random variable may or may not exist. Also, as hinted by its definition, the support of the function varies across distributions. Moment generating functions are useful for deriving the moments of a distribution: Consider for simplicity that X is scalar with moment generation function $M_X(t)$. The k'th partial derivative of $M_X(t)$ with respect to t, evaluated at $t = 0$ is given by

$$\left.\frac{\partial^k M_X(t)}{\partial t}\right|_{t=0} = \left.\frac{\partial^k E(e^{tX})}{\partial t}\right|_{t=0} = E(X^k e^{tX})\big|_{t=0} = E(X^k).$$

Similar derivations can be made for the moment generating function of a random vector.

The practicality of the moment generating function is not limited to calculating moments. When the moment generating function of a random variable exists, it uniquely determines the random variable's distribution. This allows one to identify the distribution of a random variable constructed from other random variables by inspecting its moment generating function.

B Samples from normal populations

Normal populations are central to classical statistics. We will define the normal and multivariate normal distributions first, and then derive several properties, test statistics, and other distributions arising from samples from normal populations.

Definition 1.2 (Normal distribution)**.** We say a random variable X has a normal distribution with mean μ and variance $\sigma^2 > 0$, and show it by $\mathcal{N}(\mu, \sigma^2)$, if it has a probability density function given by

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right), \quad -\infty < x < \infty$$

Exercise 1.2. Show that the moment generating function of $X \sim \mathcal{N}(\mu, \sigma^2)$ is

$$M_X(t) = \exp\left(t\mu + \frac{1}{2}\sigma^2 t^2\right), \quad t \in \mathbb{R}.$$

[1]It is more conventional, and usually more convenient, to define random vectors as column vectors. That is why we will use the notation $v = (v_1, \ldots, v_n)$ to mean that v is a column vector of size n, that is, $(v_1, \ldots, v_n) = \begin{bmatrix} v_1 & \ldots & v_n \end{bmatrix}^T$. Also, the space of column vectors of size n will be shown as $\mathbb{R}^n$ or $\mathbb{R}^{n\times 1}$

A multivariate version of the normal distribution exists with very close connections to the univariate one.

Definition 1.3 (Multivariate normal distribution)**.** A random vector $X = (X_1, \ldots, X_n)$ with mean vector $m = (m_1, \ldots, m_n)$ and a positive semidefinite covariance $\Sigma = (\Sigma_{ij})$ has a multivariate normal distribution, shown by $\mathcal{N}(m, \Sigma)$, if its moment generating function is written as

$$M_X(t) = \exp\left(t^T m + \frac{1}{2} t^T \Sigma t\right), \quad t = (t_1, \ldots, t_n) \in \mathbb{R}^n.$$

Note that we do not define the multivariate normal distribution by its probability density function, since the pdf does not always exist. Specifically, if Σ is positive definite (non-singular) with inverse Σ^{-1}, then X has a probability density function given by

$$f(x) = \frac{1}{\sqrt{2\pi \det \Sigma}} \exp\left\{-\frac{1}{2}(x-m)^T \Sigma^{-1}(x-m)\right\}, \quad x \in \mathbb{R}^n.$$

A linear transformation of X is any random variable written in the form of $a^T X + b$, where a is a vector of the same size as X and b is a scalar. Linear transformations help lay out an important relation between multivariate and univariate normal distributions.

Theorem 1.1. *Let $X = (X_1, \ldots, X_n)$ be a random vector with mean m and covariance Σ. Then, X has a multivariate normal distribution if and only if every linear transformation of X has a univariate normal distribution.*

Application of the result above with $n = 1$ reduces to stating that a linear transformation of a random variable with a univariate normal distribution also has a univariate normal distribution.

One can show that a matrix transformation of a random vector with a multivariate normal distribution has also a multivariate normal distribution.

Exercise 1.3. Suppose that X has a multivariate normal distribution with mean vector m and covariance Σ. Show that, for any matrix $A \in \mathbb{R}^{r \times n}$ and a column vector $b \in \mathbb{R}^r$, the vector $Y = AX + b$ has a multivariate normal distribution, $Y \sim \mathcal{N}(Am + b, A\Sigma A^T)$.

We call $\mathcal{N}(0, 1)$ the standard normal distribution. Accordingly, given $X \sim \mathcal{N}(\mu, \sigma^2)$, the random variable is we call

$$Z = (X - \mu)/\sigma \tag{1.2}$$

the standardised normal random variable, or X is said to be standardised when it is transformed as in (1.2).

Exercise 1.4. Show that $Z \sim \mathcal{N}(0, 1)$.

B.1 Single population

Sample from a normal population: Let $X_1, \ldots, X_n$ be a random sample of size n from a (infinite) normal population with mean μ and variance $\sigma^2 < \infty$. Equivalently, we write

$$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2) \tag{1.3}$$

It should be easy to show that $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$ and, in particular,

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1).$$

Definition 1.4 (Chi-square distribution)**.** We say a random variable X has a chi-square distribution with ν degrees of freedom, and show it by χ^2_ν, if it has a probability density

$$f(x) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{\nu/2-1} e^{-x/2}, \quad x > 0.$$

The relevance of the chi-square distribution to normal populations is due to the following fundamental relation between $\mathcal{N}(0, 1)$ and χ^2_1.

Theorem 1.2. *If $Z \sim \mathcal{N}(0, 1)$, we have $Z^2 \sim \chi^2_1$.*

The next set of results establishes that the family of chi-squared distributions is closed under addition.

Exercise 1.5. Use the moment generating function to show the following.

(a) If $X \sim \chi^2_{\nu_1}$, $Y \sim \chi^2_{\nu_2}$, and X, Y are independent, $X + Y \sim \chi^2_{\nu_1+\nu_2}$.

(b) If X and Y are independent, $X \sim \chi^2_{\nu_1}$, and $X + Y \sim \chi^2_{\nu_1+\nu_2}$, then $Y \sim \chi^2_{\nu_2}$.

(c) Let $X_1, \ldots, X_n$ be a sample of size n from a normal population with mean μ and variance $\sigma^2 < \infty$ and let $Z_i = (X_i - \mu)/\sigma$. Then $\sum_{i=1}^n Z_i^2 \sim \chi^2_n$.

A central result for a random sample from a normal population is the independence between $\bar{X}$ and S^2. Independence between the sample mean and the sample variance is not only a nice result on its own (e.g., for providing independent estimators for the parameters of the distribution), but it also leads to a closed-form distribution for (a scaled version of) the sample variance. The next two exercises show the steps for obtaining the independence relation and the sampling distribution of S^2.

Exercise 1.6. Let $X_1, \ldots, X_n$ be a random sample of size n from a normal population with mean μ and variance $\sigma^2 < \infty$. Show that

(a) $\bar{X}$ and $X_i - \bar{X}$ are independent,

(b) the sample mean $\bar{X}$ and the vector $(X_i - \bar{X}, \ldots, X_n - \bar{X})$ are independent,

(c) the sample mean $\bar{X}$ and the sample variance S^2 are independent.

Next, we establish the sampling distribution of S^2 by the following the steps of the exercise below.

Exercise 1.7. Let $X_1, \ldots, X_n$ be a random sample of size n from a normal population with mean μ and variance $\sigma^2 < \infty$.

(a) Show that

$$\sum_{i=1}^{n}(X_i - \mu)^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2 + n(\bar{X} - \mu)^2.$$

(b) Using the above, show that

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2 = \frac{(n-1)S^2}{\sigma^2} + \frac{(\bar{X} - \mu)^2}{\sigma^2/n}.$$

(c) Finally, conclude that $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$.

Another distribution related to a random sample is the t-distribution.

Definition 1.5 (t-distribution)**.** We say a random variable X has a t-distribution with ν degrees of freedom, and show it by t_ν, if it has a probability density

$$f(x) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{\pi\nu}}\left(1 + \frac{x^2}{v}\right)^{-(\nu+1)/2}, \quad -\infty < x < \infty.$$

The t-distribution characterises the ratio between the sample mean and the sample variance of a random sample from a normal distribution. The result in the exercise below indicates the practical relevance of the t-distribution to normal populations.

Theorem 1.3. *If $X \sim \mathcal{N}(0,1)$, $Y \sim \chi^2_\nu$, and X, Y are independent, then,*

$$\frac{X}{\sqrt{Y/\nu}} \sim t_\nu$$

Exercise 1.8. Let $X_1, \ldots, X_n$ be a random sample of size n from a normal population with mean μ and variance $\sigma^2 < \infty$. Using the result of Theorem 1.3, show that the random variable

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$

Finally, we will introduce the F-distribution, which characterises the ratio between two independent chi-squared random variables.

Definition 1.6 (F-distribution)**.** We say a random variable X has a F-distribution with ν_1 and ν_2 degrees of freedom, and show it by F_{ν_1,ν_2}, if it has a probability density

$$f(x) = \frac{\Gamma((\nu_1+\nu_2)/2)(\nu_1/\nu_2)^{\nu_1/2}}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)} \frac{x^{\nu_1/2-1}}{(1+x\nu_1/\nu_2)^{(\nu_1+\nu_2)/2}}, \quad 0 < x < \infty.$$

Theorem 1.4. *If $X \sim \chi^2_{\nu_1}$, $Y \sim \chi^2_{\nu_2}$, and X, Y are independent, then,*

$$\frac{X/\nu_1}{Y/\nu_2} \sim F_{\nu_1,\nu_2}.$$

Exercise 1.9. Show that, if $X \sim t_\nu$, then,

$$X^2 \sim F_{1,\nu}.$$

B.2 Two populations

Sometimes, we are interested in inference about two random samples from two independent normal populations. The following exercise covers some common test statistics relevant to two normal populations.

Exercise 1.10. Suppose that we have two random samples of sizes n_1 and n_2 from two independent normal populations

$$X_{11}, \ldots, X_{1n_1} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_1, \sigma_1^2), \quad X_{21}, \ldots, X_{2n_2} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_2, \sigma_2^2).$$

Denote the sample means and variances of those samples by $\bar{X}_1$, $\bar{X}_2$ and S_1^2, S_2^2, respectively.

(a) Show that

$$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \sim \mathcal{N}(0,1)$$

(b) Show that

$$\frac{(n_1-1)S_1^2}{\sigma_1^2} + \frac{(n_2-1)S_2^2}{\sigma_2^2} \sim \chi^2_{n_1+n_2-2}.$$

(c) Show that, when $\sigma_1^2 = \sigma_2^2$, we have

$$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{1/n_1 + 1/n_2}} \sim t_{n_1+n_2-2}$$

where S_p^2 is called the pooled sample variance and is given by

$$S_p^2 = \frac{S_1^2(n_1-1) + S_2^2(n_2-1)}{n_1+n_2-2}$$

(d) Show that

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{n_1-1,n_2-1}.$$

Paired samples: Sometimes the populations are not independent, and even not necessarily normal. Here, we are interested in such a special case. In particular, we are interested in *paired* samples from two different populations, with means μ_1 and μ_2, such that

$$X_{11}, \ldots, X_{1n} \overset{\text{i.i.d.}}{\sim} F_1, \quad X_{21}, \ldots, X_{2n} \overset{\text{i.i.d.}}{\sim} F_2, \quad Z_1, \ldots, Z_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_d, \sigma_d^2) \tag{1.4}$$

where $Z_i = X_{1i} - X_{2i}$ is the difference, $\mu_d = E(Z_1)$ is the mean of the difference, which necessarily satisfies $\mu_d = \mu_1 - \mu_2$, and σ_d^2 is the variance of the difference. Under (1.4), we can simply treat $Z_1, \ldots, Z_n$ as a random sample from a single normal population.

C Hypotheses tests

Suppose we have a population with a distribution whose probability density (or probability mass) function is denoted by $f(x; \theta)$, where $\theta \in \Theta$ is a parameter vector of this distribution and Θ is the whole support of θ. A *hypothesis test* regarding the parameter vector θ of this population is in the following form

$$H_0 : \theta \in \Theta_0, \quad \text{vs} \quad H_1 : \theta \in \Theta_1.$$

The result of a hypothesis test is the choice of one of the hypotheses over the other. However, the terminology typically is one-sided: H_0 is called the *null hypothesis*, H_1 is called the *alternative hypothesis* and accepting H_1 over H_0 is usually referred to as "rejecting H_0" or "rejecting the null (hypothesis)".

Depending on the *observed* values of $X_1, \ldots, X_n$, typically denoted by $x_1, \ldots, x_n$ (small-case counterparts of the symbols used for the random variables), we decide whether to accept H_0 or H_1. The set of values at which H_0 is rejected is called the *critical region*, which we will show as C.

Usually, whether H_0 is accepted or rejected depends on the observed value of a *test statistic* $T : \mathcal{X}^n \to \mathbb{R}$. Therefore, the critical region C is defined through this test statistic. Typical forms are

$$\begin{aligned}
C &= \{x_1, \ldots, x_n : T(x_{1:n}) \in [t_c, \infty)\}, \\
C &= \{x_1, \ldots, x_n : T(x_{1:n}) \in (-\infty, t_c]\}, \\
C &= \{x_1, \ldots, x_n : T(x_{1:n}) \in [-t_c, t_c]\}, \\
C &= \{x_1, \ldots, x_n : T(x_{1:n}) \in \mathbb{R} \setminus [-t_c, t_c]\}.
\end{aligned}$$

Values like t_c, appearing in the definition of a critical region, usually define the borders of the critical region and are referred to as *critical values*.

As short-hand notation, critical regions are sometimes expressed in terms of the set R of values of the observed test statistic, $t_{\text{obs}} = T(x_{1:n})$, that is

$$C = \{t_{\text{obs}} \in R\}$$

For example, in the first example for C above, $R = [t_c, \infty]$.

C.1 Error of a hypothesis test

There is almost always a non-zero chance that the result of the hypothesis test will be wrong, i.e., we end up choosing the wrong hypothesis among the two. There are two types of errors one can make in a hypothesis test.

Definition 1.7 (Type-I error and size of a test)**.** The error we commit by rejecting H_0 when H_0 is true is called *type-I error*. The probability of committing a type-I error is called the *significance level*, or *size*, of the test, and is usually shown with α.

$$\alpha = P(\text{type I error}) = P(\text{reject } H_0 | H_0 \text{ is true}) = P(X_{1:n} \in C | \theta \in \Theta_0)$$

Definition 1.8 (Type-II error and power of a test)**.** The error we commit by accepting H_0 when H_1 is true is called *type-II error*, and usually shown by β.

$$\beta = P(\text{type II error}) = P(\text{accept } H_0 | H_1 \text{ is true}) = P(X_{1:n} \notin C | \theta \in \Theta_1)$$

The quantity $1 - \beta$ is referred to as the *power* of a test.

Power function: A hypothesis is called *simple* if it fully specifies the distribution of a test statistic; otherwise it is called *composite*. When one or both hypotheses are composite, probabilities α or β may be unavailable. Instead, one can consider the error probability at a specific value of θ and define the following error functions: Given a hypotheses test with given a decision rule, define $\alpha : \Theta_0 \mapsto [0, 1]$ and $\beta : \Theta_1 \mapsto [0, 1]$ (with an abuse of notation we used the same symbols as functions) such that

$$\begin{aligned} \alpha(\theta) &= P(\text{reject } H_0 \mid \text{the true parameter is } \theta), \quad \text{for } \theta \in \Theta_0 \\ \beta(\theta) &= P(\text{accept } H_0 \mid \text{the true parameter is } \theta), \quad \text{for } \theta \in \Theta_1 \end{aligned}$$

A useful function combining $\alpha(\theta)$ and $\beta(\theta)$ to monitor the error behaviour is the *power function* $\pi : \Theta_0 \cup \Theta_1 \mapsto [0, 1]$ defined as

$$\pi(\theta) = \begin{cases} \alpha(\theta), & \text{for } \theta \in \Theta_0 \\ 1 - \beta(\theta), & \text{for } \theta \in \Theta_1. \end{cases}, \quad \theta \in \Theta_0 \cup \Theta_1$$

Simply put, $\pi(\theta)$ is the probability of rejecting H_0 when the true parameter is θ.

C.2 Deriving decision rules systematically

Recall that the decision rule of a hypothesis test is defined via a critical region C. The critical region C may or may not depend on values in Θ_0 and Θ_1. Two different critical regions for the same pair of hypotheses correspond to two different tests. Therefore, "test" and "critical region" are used interchangeably in this context. A natural question is, then, the following: Among all candidates, which critical region (or test) is better?

The power function of a test allows us to compare it with other tests. Let us restrict our question to comparing the tests of the same size. However, since a single type I error probability may not be available for a test, as discussed at the beginning of this section, we generalise the definition of a size of a test as below.

Definition 1.9 (Size of a test)**.** A hypothesis test with power function $\pi(\theta)$ is called to have size α if

$$\sup_{\theta \in \Theta_0} \pi(\theta) = \alpha.$$

A reasonable comparison is in terms of power functions among all tests with sizes equal to (or smaller than) a given α.

Definition 1.10 (Uniformly most powerful test)**.** A critical region C of size α is called *uniformly most powerful* if, for any other critical region C' with size at most α, we have

$$\pi(\theta) \geq \pi'(\theta), \quad \forall \theta \in \Theta_1.$$

where $\pi(\theta)$ and $\pi'(\theta)$ are the power functions corresponding to C and C', respectively. The test that uses C as its critical region is called a uniformly most powerful test.

C.2.1 Neyman-Pearson Lemma

When both hypotheses are simple, so that $f(x;\theta)$ can be written under both hypotheses, the Neyman-Pearson Lemma establishes the most powerful test among tests of a given size.

Assume $X_1, \ldots, X_n$ are i.i.d. with a probability density (or mass) function given by $f(x;\theta)$. Given $X_{1:n} = x_{1:n}$, define the likelihood function

$$L(\theta; x_{1:n}) = \prod_{i=1}^{n} f(x_i;\theta), \quad \theta \in \Theta.$$

Theorem 1.5 (Neyman-Pearson Lemma)**.** *Suppose both hypotheses are simple, i.e., $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$ for some $\theta_0, \theta_1 \in \Theta$. The critical region of the most powerful test is of the form*

$$C = \left\{ x_1, \ldots, x_n : \frac{L(\theta_0; x_{1:n})}{L(\theta_1; x_{1:n})} \leq c \right\}$$

where c is selected such that $P((X_1, \ldots, X_n) \in C | \theta = \theta_0) = \alpha$ (the type-I error probability is α).

The Neyman-Pearson lemma states that, observing $X_{1:n} = x_{1:n}$, the decision rule of the most powerful test is

$$\text{Decision} = \begin{cases} H_0, & \text{for } \frac{L(\theta_0;x_{1:n})}{L(\theta_1;x_{1:n})} > c, \\ H_1, & \text{for } \frac{L(\theta_0;x_{1:n})}{L(\theta_1;x_{1:n})} \leq c. \end{cases}$$

C.2.2 Likelihood ratio test

When one or both hypotheses are composite, the Neyman-Pearson Lemma is not useful. Instead, we introduce a popular test, *likelihood ratio test*, which generalizes the test suggested by the Neyman-Pearson Lemma for composite hypotheses.

Definition 1.11 (Likelihood ratio statistic)**.** Given two hypotheses $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_1$, where Θ_0 and Θ_1 are disjoint subsets of Θ, the likelihood ratio statistic is defined as

$$T_{LR}(X_{1:n}) = -2 \ln \left[\frac{\sup_{\theta \in \Theta_0} L(\theta; X_{1:n})}{\sup_{\theta \in \Theta_0 \cup \Theta_1} L(\theta; X_{1:n})} \right].$$

Definition 1.12 (Likelihood ratio test)**.** A hypothesis test of size α is called a likelihood ratio test if its critical region is of the form

$$C = \{x_1, \ldots, x_n : T_{LR}(x_{1:n}) > c\}$$

where c is such that $\sup_{\theta \in \Theta_0} \pi(\theta) = \alpha$.

Although it is not always easy to derive the exact decision rule of the likelihood ratio test in closed form, the asymptotical behaviour of $T_{LR}(X_{1:n})$ in n is known. Specifically, when θ is p dimensional and the null hypothesis is in the form

$$H_0 : \theta_1 = \theta_1^*, \ldots, \theta_k = \theta_k^*,$$

for some $k \leq p$ then, under some general conditions, we have

$$T_{LR}(X_{1:n}) \xrightarrow{d} \chi_k^2$$

under the null hypothesis H_0.

D Significance tests

We will review null hypothesis significance tests, or shortly significance tests, regarding parameters of normal populations and proportions regarding what is known as Bernoulli populations. Classical tests regarding normal populations are either likelihood ratio tests, or tractable approximations of the likelihood ratio tests when the test statistic has a non-symmetrical distribution. Due to the central limit theorem, tests for proportions are developed based on certain normality approximations of sample proportions.

In significance tests, the size α, that is, the type I error probability, is usually a design parameter. For significance tests for normal populations, this parameter is also called the significance level. For the tests we will consider here, the critical region C, in which we reject H_0, can be determined to make the type I error exactly equal to a desired α. The only exception is for the tests for proportions, where we rely on the central limit theorem to make normality assumptions.

D.1 Confidence intervals

Before going to the tests, we will discuss confidence intervals for an unknown parameter in question, which are random intervals that contain the true value of the parameter with a given high probability. Confidence intervals themselves deserve separate attention. However, for sake of the compactness of the review, we will mostly confine our discussion to their connections to critical regions of the significance tests.

Definition 1.13 (Confidence interval)**.** Let $X_1, \ldots, X_n$ be a sample from a population and let $\theta \in \Theta$ be a particular parameter of interest of the population distribution. A $100(1-\alpha)\%$ confidence interval of θ is a set $\text{CI} \subseteq \Theta$ such that

$$P(\theta \in \text{CI}) = 1 - \alpha.$$

where the randomness involved in the statement above is due to CI depending on $X_{1:n}$.

The above definition indicates that there is not a unique confidence interval for a parameter. For example, we can have left-sided, right-sided (one-sided, to call them with a common name), or (typically more than one) two-sided confidence intervals for the same parameter with the same significance level, depending on the purpose of the analysis.

The definition also indicates that CI is a *random* interval. When $X_{1:n} = x_{1:n}$ is observed, the computed value of CI is only the observed confidence interval associated to $x_{1:n}$. Therefore, *after* we compute a confidence interval from a specific observed sample, we cannot say that that particular confidence interval contains θ with probability $1-\alpha$. A correct interpretation is: the confidence interval computed from a random sample *will* contain θ with probability $1-\alpha$. A (correct) frequentist interpretation: If I have 1000 independent random samples of size n, and compute the confidence interval for θ for each sample, I expect around $1000(1-\alpha)$ of the resulting 1000 confidence intervals to contain θ.

We will use CI both to indicate the confidence interval as a random interval or its computed value when we observe $X_{1:n} = x_{1:n}$. The distinction should be clear from the context.

As shown in the following exercise, there exists a duality between confidence intervals and critical regions. Specifically, noting that CI depends on $X_{1:n}$, a critical region of size α for $H_0 : \theta = \theta_0$ can be defined as

$$C = \{x_{1:n} : \theta_0 \notin \text{CI}\}.$$

Exercise 1.11. Suppose we have significance test with a null hypothesis $H_0 : \theta = \theta_0$ and let CI is a $100(1-\alpha)\%$ confidence interval for θ. Consider the decision rule given as

$$\text{Decision} = \begin{cases} \text{Reject } H_0, & \text{for } \theta_0 \notin \text{CI} \\ \text{Do not reject } H_0, & \text{for } \theta_0 \in \text{CI} \end{cases}$$

Show that this test's size, or type I error probability, is α.

It turns out that, confidence intervals that are dual to the critical regions of the significance tests covered above can easily be detected.

D.2 Tests concerning the mean

D.2.1 One population

Suppose we take a random sample $X_1, \dots X_n$ from a normal population with mean μ and known or unknown variance σ^2. Recall that

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1), \quad T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

Given $X_{1:n} = x_{1:n}$, we will denote their observed values evaluated at $\mu = \mu_0$ as z_{obs} and t_{obs}, respectively, i.e.,

$$z_{\text{obs}} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}, \quad t_{obs} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

We will consider the simple null hypothesis $H_0 : \mu = \mu_0$ against each of the following: a two-sided alternative $\mu \neq \mu_0$ and two one-sided alternatives $\mu > \mu_0$ and $\mu < \mu_0$.

When the variance is known: When the variance is known, the likelihood ratio test for the unknown mean is *a* z-test[2] and its critical region is stated in terms of critical values of the standard normal distribution. Let z_α be the critical value for the standard normal distribution at α such that, when $Z \sim \mathcal{N}(0,1)$ we have $P(Z > z_\alpha) = \alpha$.

Exercise 1.12. Assume $X_1, \dots, X_n$ form a random sample from a normal population with an unknown mean μ and known variance σ^2. Show that the critical region C of the likelihood ratio test for testing $H_0 : \mu = \mu_0$ can be expressed as[3]

$$C = \begin{cases} |z_{\text{obs}}| > z_{\alpha/2}, & \text{if } H_1 : \mu \neq \mu_0 \\ z_{\text{obs}} > z_\alpha, & \text{if } H_1 : \mu > \mu_0 \\ z_{\text{obs}} < -z_\alpha, & \text{if } H_1 : \mu < \mu_0 \end{cases}$$

or, equivalently, $C = \{x_{1:n} : \mu_0 \notin \text{CI}\}$, where

$$\text{CI} = \begin{cases} \left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right), & \text{if } H_1 : \mu \neq \mu_0 \\ \left(\bar{x} - z_{\alpha}\frac{\sigma}{\sqrt{n}}, \infty\right), & \text{if } H_1 : \mu > \mu_0 \\ \left(-\infty, \bar{x} + z_{\alpha}\frac{\sigma}{\sqrt{n}}\right), & \text{if } H_1 : \mu < \mu_0 \end{cases}$$

When the variance is unknown: When the variance is unknown, the likelihood ratio test for the unknown mean is a t-test and its critical region is stated in terms of critical values of a t distribution. Let $t_{\alpha,\nu}$ be the critical value for the t_ν such that, when $Y \sim t_\nu$ we have $P(Y > t_{\alpha,\nu}) = \alpha$.

[2] Any test whose test statistic has the standard normal distribution can be called a z test. Likewise, a t-test, a chi-square test, or an f-test does not uniquely specify a certain test but rather refers to the distribution of the test statistic under the null hypothesis.

[3] The set formalism is relaxed for ease of notation – the RHS is in fact a set.

Exercise 1.13. Assume $X_1, \ldots, X_n$ form a random sample from a normal population with an unknown mean μ and unknown variance σ^2. Show that the critical region C of the likelihood ratio test for testing $H_0 : \mu = \mu_0$ can be expressed as

$$C = \begin{cases} |t_{\text{obs}}| > t_{\alpha/2,n-1}, & \text{if } H_1 : \mu \neq \mu_0 \\ t_{\text{obs}} > t_{\alpha,n-1}, & \text{if } H_1 : \mu > \mu_0 \\ t_{\text{obs}} < -t_{\alpha,n-1}, & \text{if } H_1 : \mu < \mu_0 \end{cases}$$

or, equivalently, $C = \{x_{1:n} : \mu_0 \notin \text{CI}\}$, where

$$\text{CI} = \begin{cases} \left(\bar{x} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}, \bar{x} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right), & \text{if } H_1 : \mu \neq \mu_0 \\ \left(\bar{x} - t_{\alpha,n-1}\frac{S}{\sqrt{n}}, \infty\right), & \text{if } H_1 : \mu > \mu_0 \\ \left(-\infty, \bar{x} + t_{\alpha,n-1}\frac{S}{\sqrt{n}}\right), & \text{if } H_1 : \mu < \mu_0 \end{cases}$$

D.2.2 Two populations

Suppose that we have two random samples of sizes n_1 and n_2 from two independent normal populations

$$X_{11}, \ldots, X_{1n_1} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_1, \sigma_1^2), \quad X_{21}, \ldots, X_{2n_2} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_2, \sigma_2^2).$$

Denote the sample means and variances of those samples by $\bar{X}_1$, $\bar{X}_2$ and S_1^2, S_2^2, respectively. Let $\delta = \mu_1 - \mu_2$ be the difference between the means. We will consider the simple null hypothesis $H_0 : \delta = \delta_0$ against each of the following: a two-sided alternatives $\delta \neq \delta_0$ and two one-sided alternatives $\delta > \delta_0$ and $\delta < \delta_0$.

Recall that the test statistic

$$Z_d = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \sim \mathcal{N}(0, 1)$$

and, provided that $\sigma_1^2 = \sigma_2^2$,

$$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{1/n_1 + 1/n_2}} \sim t_{n_1+n_2-2}$$

where S_p^2 is the pooled sample variance. Denote the observed values of those statistics evaluated at $\mu_1 - \mu_2 = \delta_0$ by

$$z_{d,obs} = \frac{\bar{x}_1 - \bar{x}_2 - \delta_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}, \quad t_{d,obs} = \frac{\bar{x}_1 - \bar{x}_2 - \delta_0}{s_p\sqrt{1/n_1 + 1/n_2}}$$

When variances are known: When the variances are known, the likelihood ratio test for testing $H_0 : \delta = \delta_0$ is a z test.

Exercise 1.14. Assume that σ_1^2 and σ_2^2 are known. Show that the critical region C of the likelihood ratio test for testing $H_0 : \delta = \delta_0$ can be expressed as

$$C = \begin{cases} |z_{d,\text{obs}}| > z_{\alpha/2} & H_1 : \delta \neq \delta_0 \\ z_{d,\text{obs}} > z_\alpha & H_1 : \delta > \delta_0 \\ z_{d,\text{obs}} < -z_\alpha & H_1 : \delta < \delta_0 \end{cases}$$

or, equivalently, $C = \{x_{1:n} : \delta_0 \notin \text{CI}\}$, where

$$\text{CI} = \begin{cases} \left(\bar{x}_1 - \bar{x}_2 - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \bar{x}_1 - \bar{x}_2 + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right), & \text{if } H_1 : \delta \neq \delta_0 \\ \left(\bar{x}_1 - \bar{x}_2 - z_\alpha\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \infty\right), & \text{if } H_1 : \delta > \delta_0 \\ \left(-\infty, \bar{x}_1 - \bar{x}_2 + z_\alpha\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right), & \text{if } H_1 : \delta < \delta_0 \end{cases}$$

When variances are equal and unknown: When the variances are unknown but equal, the likelihood ratio test for for testing $H_0 : \delta = \delta_0$ is a t test.

Exercise 1.15. Assume that the variances are unknown but equal, $\sigma_1^2 = \sigma_2^2$. Show that the critical region C of the likelihood ratio test for testing $H_0 : \delta = \delta_0$ can be expressed as

$$C = \begin{cases} |t_{d,\text{obs}}| > t_{\alpha/2,n_1+n_2-2} & H_1 : \delta \neq \delta_0 \\ t_{d,\text{obs}} > t_{\alpha,n_1+n_2-2} & H_1 : \delta > \delta_0 \\ t_{d,\text{obs}} < -t_{\alpha,n_1+n_2-2} & H_1 : \delta < \delta_0 \end{cases}$$

or, equivalently, $C = \{x_{1:n} : \delta_0 \notin \text{CI}\}$, where

$$\text{CI} = \begin{cases} \left(\bar{x}_1 - \bar{x}_2 - t_{\alpha/2,n_1+n_2-2}s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \bar{x}_1 - \bar{x}_2 + t_{\alpha/2,n_1+n_2-2}s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right), & \text{if } H_1 : \delta \neq \delta_0 \\ \left(\bar{x}_1 - \bar{x}_2 - t_{\alpha,n_1+n_2-2}s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \infty\right), & \text{if } H_1 : \delta > \delta_0 \\ \left(-\infty, \bar{x}_1 - \bar{x}_2 + t_{\alpha,n_1+n_2-2}s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right), & \text{if } H_1 : \delta < \delta_0 \end{cases}$$

D.3 Tests concerning the variance

D.3.1 One population

Suppose we take a random sample $X_1, \ldots X_n$ from a normal population with known or unknown mean μ and unknown variance σ^2. Recall that

$$Y = \frac{\sum_{i=1}^n (X_i - \mu)^2}{\sigma^2} \sim \chi_n^2, \quad \chi = \frac{S^2(n-1)}{\sigma^2} \sim \chi_{n-1}^2$$

Given $X_{1:n} = x_{1:n}$, we will denote their observed values evaluated at $\sigma = \sigma_0$ as y_{obs} and χ^2_{obs}, respectively, i.e.,

$$y_{obs} = \frac{\sum_{i=1}^n (x_i - \mu)^2}{\sigma_0^2}, \quad \chi^2_{obs} = \frac{s^2(n-1)}{\sigma_0^2}$$

Also, let $\chi^2_{\alpha,\nu}$ be the critical value for χ^2_ν such that, when $Y \sim \chi^2_\nu$ we have $P(Y > \chi^2_{\alpha,\nu}) = \alpha$.

Both when the mean is known or unknown, the resulting likelihood ratio test is a chi-square test.

Exercise 1.16. Suppose we take a random sample $X_1, \dots X_n$ from a normal population with known μ and unknown variance σ^2. Show that the critical region C of the likelihood ratio test for testing $H_0 : \sigma^2 = \sigma_0^2$ can be expressed as[4]

$$C = \begin{cases} y_{obs} < \chi^2_{1-\alpha/2,n} \text{ or } y_{obs} > \chi^2_{\alpha/2,n}, & \text{if } H_1 : \sigma \neq \sigma_0 \\ y_{obs} > \chi^2_{\alpha,n}, & \text{if } H_1 : \sigma > \sigma_0 \\ y_{obs} < \chi^2_{1-\alpha,n}, & \text{if } H_1 : \sigma < \sigma_0 \end{cases}$$

or, equivalently, $C = \{x_{1:n} : \sigma_0^2 \notin \text{CI}\}$, where

$$\text{CI} = \begin{cases} \left(\frac{\sum_{i=1}^n (x_i-\mu)^2}{\chi^2_{\alpha/2,n}}, \frac{\sum_{i=1}^n (x_i-\mu)^2}{\chi^2_{1-\alpha/2,n}} \right) & \text{if } H_1 : \sigma \neq \sigma_0 \\ \left(\frac{\sum_{i=1}^n (x_i-\mu)^2}{\chi^2_{\alpha,n}}, \infty \right), & \text{if } H_1 : \sigma > \sigma_0 \\ \left(0, \frac{\sum_{i=1}^n (x_i-\mu)^2}{\chi^2_{1-\alpha,n}} \right), & \text{if } H_1 : \sigma < \sigma_0 \end{cases}$$

Exercise 1.17. Suppose we take a random sample $X_1, \dots X_n$ from a normal population with unknown μ and unknown variance σ^2. Show that the critical region C of the likelihood ratio test for testing $H_0 : \sigma^2 = \sigma_0^2$ can be expressed as[5]

$$C = \begin{cases} \chi^2_{obs} < \chi^2_{1-\alpha/2,n-1} \text{ or } \chi^2_{obs} > \chi^2_{\alpha/2,n-1}, & \text{if } H_1 : \sigma \neq \sigma_0 \\ \chi^2_{obs} > \chi^2_{\alpha,n-1}, & \text{if } H_1 : \sigma > \sigma_0 \\ \chi^2_{obs} < \chi^2_{1-\alpha,n-1}, & \text{if } H_1 : \sigma < \sigma_0 \end{cases}$$

or, equivalently, $C = \{x_{1:n} : \sigma_0^2 \notin \text{CI}\}$, where

$$\text{CI} = \begin{cases} \left(\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}} \right) & \text{if } H_1 : \sigma \neq \sigma_0 \\ \left(\frac{(n-1)s^2}{\chi^2_{\alpha,n-1}}, \infty \right), & \text{if } H_1 : \sigma > \sigma_0 \\ \left(0, \frac{(n-1)s^2}{\chi^2_{1-\alpha,n-1}} \right), & \text{if } H_1 : \sigma < \sigma_0 \end{cases}$$

[4]For the first case, the given critical region is approximate. The exact form of the critical region of the likelihood ratio test is of the form $\chi^2_{1-\alpha',n} < y_{obs} < \chi^2_{\alpha'',n}$ for specific values of α' and α'' satisfying $\alpha' + \alpha'' = \alpha$.

[5]For the first case, the given critical region is approximate. The exact form of the critical region of the likelihood ratio test is of the form $\chi^2_{1-\alpha',n-1} < \chi^2_{obs} < \chi^2_{\alpha'',n-1}$ for specific values of α' and α'' satisfying $\alpha' + \alpha'' = \alpha$.

D.3.2 Two populations

Suppose that we have two random samples of sizes n_1 and n_2 from two independent normal populations

$$X_{11},\dots,X_{1n_1} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_1,\sigma_1^2), \quad X_{21},\dots,X_{2n_2} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_2,\sigma_2^2).$$

Recall that

$$U = \frac{\frac{1}{n_1}\sum_{i=1}^{n_1}(X_{1i}-\mu_1)^2/\sigma_1^2}{\frac{1}{n_2}\sum_{i=1}^{n_2}(X_{2i}-\mu_2)^2/\sigma_2^2} \sim f_{n_1,n_2}, \quad F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim f_{n_1-1,n_2-1}$$

Given $X_{11}=x_{11},\dots,X_{1n_1}=x_{1n_1}$ and $X_{21}=x_{21},\dots,X_{2n_2}=x_{2n_2}$, we will denote their observed values evaluated at $\sigma_1^2/\sigma_2^2=1$ as u_{obs} and f_{obs}, respectively, i.e.,

$$u_{obs} = \frac{\frac{1}{n_1}\sum_{i=1}^{n_1}(x_{1i}-\mu_1)^2}{\frac{1}{n_2}\sum_{i=1}^{n_2}(x_{2i}-\mu_2)^2}, \quad f_{obs} = \frac{s_1^2}{s_2^2}$$

Both when the means are known and unknown we have an f-test for testing H_0 : $\sigma_1^2/\sigma_2^2=1$.

Exercise 1.18. Assume the means μ_1 and μ_2 are known. Show that the critical region C of the likelihood ratio test for testing $H_0 : \sigma_1^2/\sigma_2^2=1$ can be expressed as[6]

$$C = \begin{cases} u_{obs} < f_{1-\alpha/2,n_1,n_2} \text{ or } u_{obs} > f_{\alpha/2,n_1,n_2}, & \text{if } H_1 : \sigma_1/\sigma_2 \neq 1 \\ u_{obs} > f_{\alpha,n_1,n_2}, & \text{if } H_1 : \sigma_1/\sigma_2 > 1 \\ u_{obs} < f_{1-\alpha,n_1,n_2}, & \text{if } H_1 : \sigma_1/\sigma_2 < 1 \end{cases}$$

or, equivalently, $C=\{x_{11},\dots,x_{1n_1};x_{21},\dots,x_{2n_2} : 1 \notin \text{CI}\}$, where

$$\text{CI} = \begin{cases} \left(u_{obs}f_{1-\alpha/2,n_2,n_1}, u_{obs}f_{\alpha/2,n_2,n_1}\right) & \text{if } H_1 : \sigma_1/\sigma_2 \neq 1 \\ \left(u_{obs}f_{1-\alpha,n_2,n_1}, \infty\right), & \text{if } H_1 : \sigma_1/\sigma_2 > 1 \\ \left(0, u_{obs}f_{\alpha,n_2,n_1}\right), & \text{if } H_1 : \sigma_1/\sigma_2 < 1 \end{cases}$$

Exercise 1.19. Assume the means μ_1 and μ_2 are unknown. Show that the critical region C of the likelihood ratio test for testing $H_0 : \sigma_1^2/\sigma_2^2=1$ can be expressed as[7]

$$C = \begin{cases} f_{obs} < f_{1-\alpha/2,n_1-1,n_2-1} \text{ or } f_{obs} > f_{\alpha/2,n_1-1,n_2-1}, & \text{if } H_1 : \sigma_1/\sigma_2 \neq 1 \\ f_{obs} > f_{\alpha,n_1-1,n_2-1}, & \text{if } H_1 : \sigma_1/\sigma_2 > 1 \\ f_{obs} < f_{1-\alpha,n_1-1,n_2-1}, & \text{if } H_1 : \sigma_1/\sigma_2 < 1 \end{cases}$$

[6]For the first case, the given critical region is approximate. The exact form of the critical region of the likelihood ratio test is of the form $f_{1-\alpha',n_1,n_2} < u_{obs} < f_{\alpha'',n_1,n_2}$ for specific values of α' and α'' satisfying $\alpha'+\alpha''=\alpha$.

[7]For the first case, the given critical region is approximate. The exact form of the critical region of the likelihood ratio test is of the form $f_{1-\alpha',n_1-1,n_2-1} < f_{obs} < f_{\alpha'',n_1-1,n_2-1}$ for specific values of α' and α'' satisfying $\alpha'+\alpha''=\alpha$.

or, equivalently, $C = \{x_{11}, \ldots, x_{1n_1}; x_{21}, \ldots, x_{2n_2} : 1 \notin \text{CI}\}$, where

$$\text{CI} = \begin{cases} \left(f_{obs} f_{1-\alpha/2, n_2-1, n_1-1}, f_{obs} f_{\alpha/2, n_2-1, n_1-1}\right) & \text{if } H_1 : \sigma_1/\sigma_2 \neq 1 \\ \left(0, f_{obs} f_{1-\alpha, n_2-1, n_1-1}\right), & \text{if } H_1 : \sigma_1/\sigma_2 > 1 \\ \left(f_{obs} f_{\alpha, n_2-1, n_1-1}, \infty\right), & \text{if } H_1 : \sigma_1/\sigma_2 < 1 \end{cases}$$

Exercise 1.20. The above tests are designed for $H_0 : \sigma_1/\sigma_2 = 1$. How would you modify them for $H_0 : \sigma_1/\sigma_2 = c$ for a general $c > 0$?

D.4 Tests concerning proportions

Let $X_1, \ldots, X_n$ be a random sample from a Bernoulli population with success probability θ, i.e., $P(X_i = 1) = 1 - P(X_i = 0) = \theta$ independently for each X_i. Let $Y = \sum_{i=1}^n X_i$. Then, $Y \sim \text{Binom}(n, \theta)$, the Binomial distribution with number of independent trials n and success probability θ, with probability mass function

$$f(k) = P(Y = k) = \frac{n!}{k!(n-k)!}\theta^k(1-\theta)^{n-k}, \quad k = 0, \ldots, n.$$

If $X_{1:n} = x_{1:n}$ are given with $y = x_1 + \ldots + x_n$, the maximum likelihood estimator for θ is

$$\hat{\theta} = Y/n = \frac{1}{n}\sum_{i=1}^n X_i.$$

Exercise 1.21. Let $X_1, \ldots, X_n$ be a random sample from a Bernoulli population with success probability θ, $Y = X_1 + \ldots + X_n$, and $\hat{\theta} = Y/n$. Show that

- $E(X_i) = \theta$ and $V(X_i) = \theta(1-\theta)$.
- $E(Y) = n\theta$ and $V(Y) = n\theta(1-\theta)$
- $E(\hat{\theta}) = \theta$ and $V(\hat{\theta}) = \theta(1-\theta)/n$.

It is possible to write down the distribution of $\hat{\theta}$ exactly using the distribution of Y; however, because of the combinatorial terms, this distribution is not easy to handle. Instead, for large enough n we can resort to a normal approximation for $\hat{\theta}$, thanks to the central limit theorem.

Theorem 1.6. *Let $X_1, X_2, \ldots$ are i.i.d. with mean μ and variance $\sigma^2 < \infty$, and let $\bar{X}_n$ is the sample mean of $X_1, \ldots, X_n$. Then,*

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, 1).$$

That is, the sample mean converges in distribution to a random variable with standard normal distribution.

The informal interpretation of the theorem above is that, for large n, we have $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$ approximately. Applying this to $\hat{\theta} = \bar{X}$ for when X_i's are i.i.d. with Bernoulli distribution, we can use the central limit theorem and say that

$$\hat{\theta} \sim \mathcal{N}(\theta, \theta(1-\theta)/n)$$

or

$$\frac{\hat{\theta} - \theta}{\sqrt{\theta(1-\theta)/n}} \sim \mathcal{N}(0,1)$$

approximately. Therefore, given $Y = y$, a critical region C for testing $H_0 : \theta = \theta_0$ can be expressed as[8]

$$C = \begin{cases} |\hat{\theta} - \theta_0| > z_{\alpha/2}\sqrt{\theta_0(1-\theta_0)/n}, & \text{if } H_1 : \theta \neq \theta_0 \\ \hat{\theta} - \theta_0 > z_{\alpha}\sqrt{\theta_0(1-\theta_0)/n}, & \text{if } H_1 : \theta > \theta_0 \\ \hat{\theta} - \theta_0 < -z_{\alpha}\sqrt{\theta_0(1-\theta_0)/n}, & \text{if } H_1 : \theta < \theta_0 \end{cases} \tag{1.5}$$

Note that we did not provide a confidence interval that corresponds to the C given above. The reason is that, with θ appearing in the variance, it is difficult to derive a confidence interval for θ. However; the variance of $\hat{\theta}$ can also be estimated by $\hat{\theta}(1-\hat{\theta})/n$ and we have the resulting approximation

$$\frac{\hat{\theta} - \theta}{\sqrt{\hat{\theta}(1-\hat{\theta})/n}} \sim \mathcal{N}(0,1)$$

from which we can build one-sided and two-sided CI's of the form

$$\left(\hat{\theta} - z_{\alpha/2}\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}, \hat{\theta} + z_{\alpha/2}\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}\right), \quad \left(\hat{\theta} - z_{\alpha}\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}, 1\right], \quad \left[0, \hat{\theta} + z_{\alpha}\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}\right)$$

depending on the purpose. (Why are the borders 0 and 1?) If in C in (1.5), we replaced the θ_0's appearing on the right-hand side by $\hat{\theta}$'s, then the confidence intervals and the critical region would be fully consistent.

When two populations are concerned, we can use similar approximations. For example, given $X_{11}, \ldots, X_{1n_1}$ and $X_{21}, \ldots, X_{2n_2}$ from Bernoulli populations with success probabilities θ_1 and θ_2, and the estimators $\hat{\theta}_1 = \bar{X}_1$ and $\hat{\theta}_2 = \bar{X}_2$, we approximately have

$$\hat{\theta}_1 - \hat{\theta}_2 \sim \mathcal{N}\left(\theta_1 - \theta_2, \frac{\theta_1(1-\theta_1)}{n_1} + \frac{\theta_2(1-\theta_2)}{n_2}\right)$$

The null hypothesis $H_0 : \theta_1 - \theta_2 = \theta_d$ can be tested by comparing

$$\frac{\hat{\theta}_1 - \hat{\theta}_2 - \theta_d}{\sqrt{\frac{\hat{\theta}_1(1-\hat{\theta}_1)}{n_1} + \frac{\hat{\theta}_2(1-\hat{\theta}_2)}{n_2}}}$$

[8] The set formalism is relaxed for ease of notation – the RHS is in fact a set.

against suitable critical values, depending on H_1. Approximate confidence intervals derived from the above statistic will be consistent with those tests.

When $H_0 : \theta_1 - \theta_2 = 0$ is to be tested, an alternative approach is to observe that, when $\theta_1 = \theta_2 = \theta$, we have

$$\hat{\theta}_1 - \hat{\theta}_2 \sim \mathcal{N}\left(0, \theta(1-\theta)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$$

Based on that, we can estimate the common success probability as $\hat{\theta} = \frac{Y_1+Y_2}{n_1+n_2}$ and use the test statistic

$$\frac{\hat{\theta}_1 - \hat{\theta}_2}{\sqrt{\hat{\theta}(1-\hat{\theta})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

whose approximate distribution is $\mathcal{N}(0, 1)$ under the null hypothesis.

Chapter 2

The Analysis of Variance

Analysis of variance (ANOVA) is a collection of statistical models and their associated estimation procedures (such as the "variation" among and between groups) used to analyze the differences among group means in a sample.

A The one-way layout

A.1 Setting

We have I treatments, each having n_i observations. Let the X_{ij} be the j'th observation in the i'th sample. The model for the observations are

$$X_{ij} = \mu + a_i + e_{ij}, \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, I$$

Here, μ is the overall mean level, a_i is the differential effect of the i'th treatment, normalised such that

$$\sum_{i=1}^{I} n_i a_i = 0,$$

and e_{ij}'s are independent random error with a normal distribution $\mathcal{N}(0, \sigma^2)$,

$$e_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2).$$

The differential effects are normalised, i.e., The i.i.d. assumption ensures that all X_{ij} are independent and normally distribution with a common variance,

$$X_{ij} \sim \mathcal{N}(\mu + a_i, \sigma^2), \quad j = 1, \ldots, n_i; \quad i = 1, \ldots, I. \tag{2.1}$$

The inferential problem: We would like to know if there is any variation in the mean across the samples. Therefore, we would like to test the null hypothesis

$$H_0 : a_1 = a_2 = \ldots = a_I.$$

Exercise 2.1. Show that, H_0 above is equivalent to

$$H_0 : a_1 = a_2 = \ldots = a_I = 0.$$

Denote $\mu_i = \mu + a_i$. For $I = 2$ (two populations), H_0 reduces to the hypothesis $H_0 : \delta = \mu_1 - \mu_2 = 0$, which was discussed earlier.

A.2 Testing equality of the means in one-way ANOVA

Let $n = n_1 + \ldots + n_I$ be the total number of observations. Define $\bar{X}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}$ to be the sample mean of the i'th sample, and $\bar{X} = \frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{n_i} X_{ij}$ to be the overall sample mean.

Exercise 2.2. Show the basic identity of the analysis of variance

$$\sum_{i=1}^{I}\sum_{j=1}^{n_i}(X_{ij} - \bar{X})^2 = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2 + \sum_{i=1}^{I} n_i(\bar{X}_i - \bar{X})^2. \tag{2.2}$$

[Hint: For each term of the LHS, apply $X_{ij} - \bar{X} = (X_{ij} - \bar{X}_i) + (\bar{X}_i - \bar{X})$, and show that the third terms of the expansion sum to 0.]

Equation (2.2) suggests that the total sum of errors is equal to the sum of squares within groups plus the sum of squares between groups. We write (2.2) according to this interpretation

$$SS_{total} = SS_w + SS_b.$$

where

$$SS_{total} = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(X_{ij} - \bar{X})^2, \quad SS_w = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2, \quad SS_b = \sum_{i=1}^{I} n_i(\bar{X}_i - \bar{X})^2.$$

Theorem 2.1. *Functions of independent random vectors are also independent. That is, if $X = (X_1, \ldots, X_p)$ and $Y = (Y_1, \ldots, Y_q)$ are independent real-valued random vectors, then, for any $f : \mathbb{R}^p \mapsto \mathbb{R}^{p'}$ and $g : \mathbb{R}^q \mapsto \mathbb{R}^{q'}$, $f(X)$ and $g(Y)$ are also independent.*

Exercise 2.3. Show that SS_w and SS_b are independent.

Exercise 2.4. Show that

$$\begin{aligned} E(SS_{total}) &= (n-1)\sigma^2 + \sum_{i=1}^{I} n_i a_i^2 \\ E(SS_w) &= (n-I)\sigma^2 \\ E(SS_b) &= (I-1)\sigma^2 + \sum_{i=1}^{I} n_i a_i^2 \end{aligned}$$

As stated in the last exercise, the total error will be inflated by non-zero a_i's. In the same manner, we expect that the SS_b/SS_w gets larger when the model deviates from H_0. Indeed, the likelihood ratio test confirms this intuition and yields an F test for H_0.

Exercise 2.5. Show the following

(a) $\frac{SS_w}{\sigma^2} \sim \chi^2_{n-I}$.

(b) Under H_0, $\frac{SS_{total}}{\sigma^2} \sim \chi^2_{n-1}$, $\frac{SS_b}{\sigma^2} \sim \chi^2_{I-1}$, $E(SS_b) = (I-1)\sigma^2$, and

$$\frac{SS_b/(I-1)}{SS_w/(n-I)} \sim f_{I-1,n-I}$$

Exercise 2.6. Assume (2.2). Show that the likelihood ratio test for

$$H_0 : a_1 = a_2 = \ldots = a_I = 0, \quad \text{vs} \quad H_1 : \text{ At least one } a_i \text{ is non-zero.}$$

is an F-test, with the critical region of size α given by

$$C = \left\{ \frac{ss_b/(I-1)}{ss_w/(n-I)} > f_{\alpha,I-1,n-I} \right\}$$

A typical one-way ANOVA table, which summarises the analysis necessary for the basic ANOVA null hypothesis, is given below.

Source	df	SS	MS	F
Between groups	$I-1$	ss_b	$ss_b/(I-1)$	$\frac{ss_b/(I-1)}{ss_w/(n-I)}$
Within groups	$n-I$	ss_w	$ss_w/(n-I)$	
Total	$n-1$	ss_{total}		

A.3 Contrasts

Instead of testing the equality of the means, we can be more specific and test a linear combination of the means. Consider any vector of constants $c_1, \ldots, c_I$, let us say we are interested in $\sum_{i=1}^I c_i\mu_i$. Under the ANOVA setting, we have

$$\frac{\sum_{i=1}^I c_i\bar{X}_i - \sum_{i=1}^I c_i\mu_i}{\sigma\sqrt{\sum_{i=1}^I \frac{c_i^2}{n_i}}} \sim \mathcal{N}(0,1), \quad \frac{\sum_{i=1}^I c_i\bar{X}_i - \sum_{i=1}^I c_i\mu_i}{\sqrt{\frac{SS_w}{n-I}\sum_{i=1}^I \frac{c_i^2}{n_i}}} \sim t_{n-I}.$$

Therefore, confidence intervals and hypothesis tests about $\sum_{i=1}^I c_i\mu_i$ are available. A $100(1-\alpha)\%$-level CI for $\sum_{i=1}^I c_i\mu_i$ is given by

$$\left(\sum_{i=1}^I c_i\bar{X}_i - t_{\alpha/2,n-I}\sqrt{\frac{SS_w}{n-I}\sum_{i=1}^I \frac{c_i^2}{n_i}}, \quad \sum_{i=1}^I c_i\bar{X}_i + t_{\alpha/2,n-I}\sqrt{\frac{SS_w}{n-I}\sum_{i=1}^I \frac{c_i^2}{n_i}} \right) \tag{2.3}$$

One family of hypotheses involving the means of groups is in terms of contrasts. Contrasts are important since they formalize more detailed, or specific, comparisons among the means.

Definition 2.1 (Contrast)**.** A linear combination of the means $\sum_{i=1}^I c_i\mu_i$ is called a contrast if the vector of coefficients $\mathbf{c} = (c_1, \ldots, c_I)$ satisfies $\sum_{i=1}^I c_i = 0$.

Let the vector of coefficients $\mathbf{c} = (c_1, \ldots, c_I)$ leading to a contrast be called a contrast vector. Due to the constraint of summing to 0, the set of all contrast vectors

$$\mathcal{C} = \left\{ (c_1, \ldots, c_I) : \sum_{i=1}^{I} c_i = 0 \right\}$$

has dimension $I - 1$, i.e., it is spanned by $I - 1$ linearly independent vectors whose coefficients sum to 0. In other words, any contrast vector can be written as a linear combination of $I - 1$ linearly independent contrast vectors. A set of contrast vectors that span $\mathcal{C}$ is

$$\mathbf{c}_1 = (1, -1, 0, \ldots, 0), \quad \mathbf{c}_2 = (0, 1, -1, 0, \ldots, 0), \quad \mathbf{c}_{I-1} = (0, \ldots, 0, -1, 1). \tag{2.4}$$

(You can think of them as the 'basis vectors' for the space of contrast vectors, in analogy to the basis vectors $(1, 0, \ldots, 0), \ldots, (0, \ldots, 0, 1)$ that span $\mathbb{R}^I$.)

Contrasts are important due to the null hypotheses they correspond to. Given $\mathbf{c} \in \mathcal{C}$, let us consider the corresponding null hypothesis

$$H_0 : \sum_{i=1}^{I} c_i \mu_i = 0, \quad H_1 : \sum_{i=1}^{I} c_i \mu_i \neq 0$$

Then, for the unit basis contrasts in (2.4), the corresponding null hypotheses in the same manner are

$$\mathbf{c}_1 \Rightarrow \mu_1 = \mu_2, \quad \mathbf{c}_2 \Rightarrow \mu_2 = \mu_3, \quad \ldots \quad \mathbf{c}_{I-1} \Rightarrow \mu_{I-1} = \mu_I.$$

Therefore, the combination of all the null hypotheses above leads to the ANOVA null hypothesis that states that all the means are equal,

$$\mathbf{c}_1, \ldots, \mathbf{c}_{I-1} \Rightarrow \mu_1 = \mu_2 = \ldots = \mu_I$$

One can show that the null ANOVA hypothesis is equivalent to saying that all linear combinations of the means by contrasts are 0, as stated in the following exercise.

Exercise 2.7. Show that, $\mu_1 = \mu_2 = \ldots = \mu_I$ if and only if $\sum_{i=1} c_i \mu_i = 0$ for all $\mathbf{c} = (c_1, \ldots, c_I) \in \mathcal{C}$. [Hint: To show $\Leftarrow$, use the contrasts in (2.4)]

An immediate consequence is that the ANOVA null and alternative hypotheses can be written as

$$H_0 : \sum_{i=1}^{I} c_i \mu_i = 0, \quad \text{for all } \mathbf{c} \in \mathcal{C}, \quad H_1 : \sum_{i=1}^{I} c_i \mu_i \neq 0, \quad \text{for some } \mathbf{c} \in \mathcal{C}.$$

One reason contrasts are useful for analysis is that they are interesting on their own, being natural hypotheses to test: Pairwise comparisons between means (such as μ_1 vs μ_2), or comparison between averages of two sets of means (such as μ_1 vs $\frac{\mu_2+\mu_3}{2}$, or $\frac{\mu_1+\mu_2}{2}$ vs $\frac{\mu_3+\mu_4+\mu_5}{3}$).

Second, we will see later that, it is possible to design simultaneous confidence intervals that hold for *all* contrasts with probability $1-\alpha$. Those confidence intervals are narrower than the confidence intervals for arbitrary linear combinations of the means, see Section B.2.4.

Finally, there is a systematic way of decomposing SS_b into smaller pieces (called contrast sum of squares) using what we will call *orthogonal contrasts*.

A.3.1 Orthogonal contrasts

Assume we want to test a set of multiple contrasts. Things get more interesting when those contrasts are *orthogonal*.

Definition 2.2 (Inner product and norm)**.** Given two vectors $v, u \in \mathbb{R}^I$ and $n_1, \ldots, n_I > 0$, define the inner product

$$\langle u, v \rangle = \sum_{i=1}^{I} \frac{u_i v_i}{n_i}.$$

Moreover, given a vector v, we define $\|v\| = \langle v, v \rangle^{1/2}$ as the norm of v.

Definition 2.3 (Orthogonality and orthonormality)**.** Vectors v, u are said to be orthogonal if $\langle u, v \rangle = 0$. If, in addition, $\|v\| = \|u\| = 1$, then v, u are said to be orthonormal.

Definition 2.4. In particular, two contrasts constructed by $\mathbf{c}_1, \mathbf{c}_2 \in \mathcal{C}$ are said to be orthogonal if $\sum_{i=1}^{I} \frac{c_{1,i} c_{2,i}}{n_i} = 0$.

Exercise 2.8 (Orthogonal contrasts are independent)**.** Orthogonal contrasts are independent, that is, given $\mathbf{c}_1$ and $\mathbf{c}_2$ such that $\sum_{i=1}^{I} \frac{c_{1,i} c_{2,i}}{n_i} = 0$, the contrasts $\sum_{i=1}^{I} c_{1,i} \bar{X}_i$ and $\sum_{i=1}^{I} c_{2,i} \bar{X}_i$ are uncorrelated, hence independent.

Definition 2.5 (Orthonormal basis)**.** A set of vectors $\{v_1, \ldots, v_I\}$ form an orthonormal basis if

$$\langle v_i, v_j \rangle = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{if } i \neq j \end{cases}.$$

Exercise 2.9. Show that the size of a set of orthogonal contrasts can be at most $I - 1$.

Orthogonal contrasts enable a perfect decomposition of SS_b. This enables writing the test statistic of the ANOVA null hypotheses (all means are equal) as the average of the test statistics of those contrasts.

Let $Y = (n_1 \bar{X}_1, \ldots, n_I \bar{X}_I)$. Given $\mathbf{c} = (c_1, \ldots, c_I)$, let

$$SS_{\mathbf{c}} = \frac{\langle Y, \mathbf{c} \rangle^2}{\langle \mathbf{c}, \mathbf{c} \rangle} = \frac{\left(c_1 \bar{X}_1 + \ldots + c_I \bar{X}_I \right)^2}{\frac{c_1^2}{n_1} + \ldots + \frac{c_I^2}{n_I}}$$

Theorem 2.2 (Contrasts split SS_b perfectly)**.** *Suppose* $\{\mathbf{c}_1, \ldots, \mathbf{c}_{I-1}\}$ *is set of* $I-1$ *orthogonal contrast vectors. Then,* $SS_{\mathbf{c}_1}, \ldots, SS_{\mathbf{c}_{I-1}}$ *are independent and split* SS_b *perfectly, i.e.,*

$$\sum_{i=1}^{I-1} SS_{\mathbf{c}_i} = SS_b$$

Independence is a direct implication of Exercise 2.8. The proof of the Theorem 2.2 is designed as Exercise 2.10.

Exercise 2.10. By following the steps below, prove Theorem 2.2

(a) Show that $\mathbf{c}_0 = (n_1, \ldots, n_I)$, $SS_{\mathbf{c}_0} = N\bar{X}^2$.

(b) Show that $\|Y\|^2 = \sum_{i=1}^{I} n_i \bar{X}_i^2$.

(c) Let $\{\mathbf{c}_1, \ldots, \mathbf{c}_{I-1}\}$ be a set of orthogonal contrasts and let $\mathbf{c}_0 = (n_1, \ldots, n_I)$. Then, show that $\{\mathbf{c}_0, \mathbf{c}_1, \ldots, \mathbf{c}_{I-1}\}$ is a set of mutually orthogonal vectors. Furthermore, show that

$$\left\{ \frac{\mathbf{c}_0}{\|\mathbf{c}_0\|}, \frac{\mathbf{c}_1}{\|\mathbf{c}_1\|}, \ldots, \frac{\mathbf{c}_{I-1}}{\|\mathbf{c}_{I-1}\|} \right\} \tag{2.5}$$

form an orthonormal basis.

(d) Let $x \in \mathbb{R}^I$ and $\{v_1, \ldots, v_I\}$ form an orthonormal basis. Then x can be written as $x = \langle v_1, x\rangle v_1 + \ldots + \langle v_I, x\rangle v_I$ and therefore $\|x\|^2 = \langle v_1, x\rangle^2 + \ldots + \langle v_I, x\rangle^2$.

Apply the result above with the orthonormal basis in (2.5) to show that

$$\|Y\|^2 = n\bar{X}^2 + \sum_{i=1}^{I-1} SS_{\mathbf{c}_i}$$

(e) Show that $\sum_{i=1}^{I} n_i \bar{X}_i^2 - n\bar{X}^2 = \sum_{i=1}^{I} n_i (\bar{X}_i - \bar{X})^2 = SS_b$.

(f) Finally, show that $\sum_{i=1}^{I-1} SS_{\mathbf{c}_i} = SS_b$.

Another advantage of orthogonal contrasts is, that due to their independence, they help reduce the overall type I error (more formal definitions to come soon). We will see that in more detail in Section B.2.2, see Exercise 2.16.

B Multiple hypotheses

Let $H_{01}, \ldots, H_{0m}$ be a set, or a *family*, of m null hypotheses and

$$H_0 = H_{01} \cap H_{02} \cap \ldots \cap H_{0m}$$

be the "combined", "overall", or "intersection" null hypothesis, which is true when all H_{0i} are true. During the analysis in this section, we will assume that the hypotheses do not contradict altogether, i.e., their intersection is not empty.

Table 2.1: Table of number of outcomes in multiple hypotheses setting

	Null hypothesis is true	Null hypothesis is false	Total
Hypothesis rejected	V	S	R
Hypothesis not rejected	U	T	$m - R$
Total	m_0	$m - m_0$	m

Assumption 2.1. *H_0 is non-empty.*

A table of numbers of outcomes in multiple hypotheses setting is given in Table 2.1. As can be seen from the table, we assume the general setting where some hypotheses are true and some are false. In practice, the knowledge of which hypotheses are true and which are false is not available, which renders the random variables V, S, U, T unobservable. On the contrary, the random variable R is observable at the end of the experiment. It is this general setting for which several type-I error rates are defined. Also, let

$$J = \{j_1, \ldots, j_{m_0}\} \subseteq \{1, \ldots, m\}$$

be the indices of the null hypotheses that are true. Moreover, since every hypothesis corresponds to a set a certain parameter vector is claimed to reside, we can define Θ_j to be that set for H_{0j}.

Note that, for a true hypothesis H_j with $j \in J$, which corresponds to the set Θ_j, we have that $\theta \in \Theta_j$ holds.

Remark 2.1. In the frequentist setting, the variable θ is a static parameter vector whose value is fixed and unknown. It is also a common practice to use θ as a dummy variable in some definitions. For example, see the definition of the power function in Section C.1 in Chapter 1, where θ is used as the argument of the power function. While this use of θ may help with getting increasing the familiarity of the reader to certain concepts, it may also be a source of possible confusion in the discussion to follow in this chapter. To avoid such confusion, we will sometimes use a different symbol, such as ϑ for the dummy variables for θ.

Remark 2.2. In the discussion, we will introduce some error rates, all of which regard the type I error probabilities. As we saw earlier, it was not always possible to talk about a single type I error under a null hypothesis, and that is why we generalised the definition of the size of the test with a null hypothesis $H_0 : \theta \in \Theta_0$ as

$$\sup_{\vartheta \in \Theta_0} P(\text{reject } H_0 | \theta = \vartheta).$$

where the event $\{\text{Reject } H_0\}$ is defined as $\{X_{1:n} \in C\}$ and the conditioned event $\{\theta = \vartheta\}$ is equivalent to $\{\text{the true value of the parameter } \theta \text{ is } \vartheta\}$. However, for the discussions to come, it is convenient to suppose that the type I error probability is same for all $\theta \in \Theta_0$

$$P(\text{Reject } H_0 | \theta = \vartheta) = \alpha, \quad \forall \vartheta \in \Theta_0.$$

the size of the test is simply α. In fact, only in such cases writing $P(\text{Reject } H_0|H_0)$ makes sense and it means

$$\{P(\text{Reject } H_0|H_0) = \alpha\} \Leftrightarrow \{P(\text{Reject } H_0|\theta = \vartheta) = \alpha, \quad \forall \vartheta \in \Theta_0\}$$

In general we can make the following definition.

Definition 2.6 (Probabilities conditional on a hypothesis). For any event A, a conditional probability $P(A|H)$, where the condition is a hypothesis $H : \theta \in \Theta_H$ is well defined and is equal to $p \in (0, 1)$ if $P(A|\theta = \vartheta)$ is the same for all $\vartheta \in \Theta_H$ and equal to p.

In the following, we will assume that such probabilities, whenever they are mentioned, are well defined in the sense of Definition 2.6.

B.1 Error rates regarding multiple hypotheses

We define several useful error rates regarding multiple hypothesis testing. When a probability or expectation is stated as $P(\cdot)$ or $E(\cdot)$ regarding the test statistics or the outcomes of the hypothesis tests, the implicit conditioning on the true (and unknown) value of θ should be assumed.

Definition 2.7 (Per comparison error rate)**.** The per comparison error rate (PCER), or the comparison-wise error rate, is related to the probability of rejecting a particular H_{0i} when H_{0i} is true. PCER is defined as

$$\text{PCER} = E(V/m).$$

Controlling PCER at α means that the PCER is at most α when all the hypotheses are true.

It is common knowledge that testing each hypothesis with a type I error probability of α guarantees that PCER $\leq \alpha$. However, it is worth proving it rigorously to gain a deep insight into the setting.

Exercise 2.11. Suppose that we test each null hypotheses such that $P(\text{reject } H_{0j}|H_{0j}) = \alpha$. Then, $E(V/m) \leq \alpha$.

Proof. Take any $j \in J$, where J is the set of indices of the true hypotheses. We are given $P(\text{reject } H_{0j}|H_{0j}) = \alpha$, which means that

$$P(\text{reject } H_{0j}|\theta = \vartheta) = \alpha, \quad \forall \vartheta \in \Theta_{0j}$$

by Definition 2.6. But, since H_{0j} is a true hypothesis, we have $\theta \in \Theta_{0j}$, which implies that

$$P(\text{reject } H_{0j}) = \alpha.$$

Since $V = \sum_{j \in J} \mathbb{I}(\text{reject } H_{0j})$, we can write the expectation

$$E(V/m) = \frac{1}{m} \sum_{j \in J} E(\mathbb{I}(\text{reject } H_{0j})) = \frac{1}{m} \sum_{j \in J} P(\text{reject } H_{0j}) = \frac{m_0}{m} \alpha.$$

By taking $m_0 = m$, i.e., all hypotheses are true, we show that PCER is controlled by α error for each test. □

Definition 2.8 (Familywise error rate)**.** The familywise error rate or per experiment error rate or experiment wise error rate is the probability of falsely rejecting at least one true null hypothesis, i.e., it is defined as

$$\text{FWER} = P(V \geq 1).$$

Controlling FWER in a weak sense at α means that the FWER is at most α when all the hypotheses are true, i.e., $P(\text{reject at least one null hypothesis} | H_0 \text{ is true}) = \alpha$.

Controlling FWER is a strong sense (S-FWER) means that FWER is at most α for every combination of true/false hypotheses.

Definition 2.9 (Discovery and false discovery)**.** A statistical discovery is the rejection of an H_{0i}. A false discovery is the rejection of an H_{0i} when H_{0i} is true.

Definition 2.10 (False discovery rate)**.** The false discovery rate is the expected number of falsely rejected hypotheses divided by the total number of rejected hypotheses.

$$\text{FDR} = E(Q), \quad Q = \begin{cases} V/R & R > 0 \\ 0 & R = 0 \end{cases}.$$

Exercise 2.12. Show that, FDR $\leq$ FWER, with equality when $m_0 = m$, i.e., all null hypotheses are true.

Proof. Noting that $Q \leq 1$ always holds, and that $Q > 0$ if and only if $V \geq 1$, we have

$$\begin{aligned} \text{FDR} = E(Q) &\leq \underbrace{1 \times P(Q > 0)}_{Q \leq 1} + 0 \times P(Q = 0) \\ &= P(V \geq 1) = \text{FWER}. \end{aligned}$$

When $m = m_0$, we have $V = R$ and $Q = 1$ if and only if $R = V \geq 1$ and $Q = 0$ if and only if $R = V = 0$. Therefore $E(Q) = P(Q = 1) = P(V > 0) = \text{FWER}$. □

B.2 Methods for controlling FWER

In the following, we will cover some of the methods to control FWER.

B.2.1 Bonferroni correction:

The Bonferroni correction method stems from a general lower bound on the probability of the intersection of sets. Generally, for sets $B_1, \ldots, B_m$,

$$P\left(\bigcup_{j=1}^{m} B_j\right) \leq \sum_{j=1}^{m} P(B_j).$$

Therefore, in order to have at most α probability for $\bigcup_{j=1}^{m} B_j$ each B_j can be set to have probability α/m.

When applied to multiple hypotheses, the Bonferroni method guarantees the desired FWER.

Theorem 2.3. *When each test is applied with α/m type I error, then the FWER is guaranteed in the strong sense to be less than or equal to α.*

Proof. Define $B_j = \{\text{rejection of } H_{0j}\}$. We are interested in the probability

$$P\left(V \geq 1\right) = P\left(\bigcup_{j \in J_0} B_j\right)$$

Using the Bonferroni correction, we can bound this probability as

$$\begin{aligned} P\left(\bigcup_{j \in J_0} B_j\right) &\leq \sum_{j \in J_0} P\left(B_j\right) \\ &= \sum_{j \in J_0} \alpha/m = \alpha \frac{m_0}{m} \leq \alpha, \end{aligned}$$

where the second line follows from the fact that $\theta \in \Theta_{0j}$ for all $j \in J$ and therefore the rejection probabilities are α/m. □

An obvious corollary to the above result is that one can choose α_j for the type I error of the test for H_{0j} differently provided that $\sum_{j=1}^{m} \alpha_j \leq \alpha$.

B.2.2 Šidák correction

We call a set of tests independent if their decisions, when viewed as random variables, are independent random variables. When we have m independent tests, each having α_0 type I error, the probability of making no false rejections, i.e., $P(V = 0)$, can be written as (recalling the number of true null hypotheses is $m_0 \leq m$)

$$P(V = 0) = (1 - \alpha_0)^{m_0},$$

which yields

$$\text{FWER} = P(V \geq 1) = 1 - (1 - \alpha_0)^{m_0} \leq 1 - (1 - \alpha_0)^{m}.$$

To bound the last expression by the desired FWER, α, one can take

$$\alpha_0 = 1 - (1 - \alpha)^{1/m}$$

Setting α_0 for each test as above is known as the Šidák correction.

When the tests are independent, as we assumed above, α_0 provided by Šidák correction is tight. However, we can show that FWER can still be controlled at α with Šidák correction if we have

$$P(D_1 = i, \ldots, D_m = i) \geq \prod_{j=1}^{m} P(D_j = i), \quad \text{for each } i = 0, 1. \tag{2.6}$$

where $D_1, \ldots, D_m \in \{0, 1\}$ be the decisions of the m tests, with $D_j = 1$ resembles rejection of H_{0j} so that $P(D_j = 1|H_{0j}) = \alpha_0$. An interpretation of (2.6) is that, given that a true hypotheses is rejected (accepted), it is more likely to reject (accept) another true hypothesis.

Exercise 2.13. Show that, if (2.6) holds, we have $P(V \geq 1) \leq \alpha$.

The condition (2.6) being satisfied is usually referred to as the existence of positive dependence among the tests. There are some formal relations among random variables which imply (2.6). One such relation is associatedness among the random variables.

Definition 2.11 (Positive association)**.** We say that random variables $X_1, \ldots, X_n$ are positively associated if for any non-decreasing functions f and g, we have

$$\text{Cov}(f(X_1, \ldots, X_n), g(X_1, \ldots, X_n)) \geq 0.$$

provided that the required expectations to define the covariance exist.

It is by the following theorem that associated decisions $D_1, \ldots, D_m$ satisfy (2.6).

Theorem 2.4. *If $X_1, \ldots, X_n$ are positively associated, then, for all $x_1, \ldots, x_n$,*

$$P(X_1 \geq x_1, \ldots, X_n \geq x_n) \geq \prod_{i=1}^{n} P(X_i \geq x_i). \tag{2.7}$$

$$P(X_1 \leq x_1, \ldots, X_n \leq x_n) \geq \prod_{i=1}^{n} P(X_i \leq x_i). \tag{2.8}$$

Exercise 2.14. Apply the theorem above to conclude that if $D_1, \ldots, D_m$ are associated (2.6) is satisfied.

Associatedness is implied by another relation what is known as positive regression dependency, another form of 'positive dependence' among random variables. Positive regression dependency is based on increasing sets, which are defined below.

Definition 2.12 (Increasing set)**.** A set $A \in \mathcal{E} \subseteq \mathbb{R}^m$ is called increasing if, for any $x, y \in \mathbb{R}^m$ with $\forall i; x_i \leq y_i$, we have $x \in A \Rightarrow y \in A$.

For example, for $\mathcal{E} = \{0,1\}^3$ the set $A = \{\{0,0,1\},\{1,0,1\},\{0,1,1\},\{1,1,1\}\}$ is an increasing set.

Definition 2.13 (Positive regression dependency)**.** Random variables $X_1, \ldots, X_m \in \mathcal{X}$ are said to have positive regression dependency if for any increasing set $A \in \mathcal{X}^m$ and for any $(j_1, \ldots, j_i) \subseteq \{1, \ldots, m\}$, the probability

$$P((X_1, \ldots, X_m) \in A | X_{j_1} = x_1, \ldots, X_{j_i} = x_i)$$

is increasing in $(x_1, \ldots, x_i)$, i.e., for any $x_1' \geq x_1, \ldots, x_m' \geq x_m$, we have

$$P((X_1, \ldots, X_m) \in A | X_{j_1} = x_1, \ldots, X_{j_i} = x_i) \leq P((X_1, \ldots, X_m) \in A | X_{j_1} = x_1', \ldots, X_{j_i} = x_i')$$

Theorem 2.5. *Positive regression dependency among $X_1, \ldots, X_n$ implies positive association, hence* (2.7).

We say that Šidák correction is 'conservative' for tests satisfying (2.6). For 'negatively dependent' tests (which can be defined by reversing the sign in (2.6)), controlling FWER at α is not anymore guaranteed, hence we say that Šidák correction is 'liberal'.

Beware, though: Pairwise positive correlation among random variables does not imply positive regression dependency.

Exercise 2.15. Suppose $X_1, X_2 \in \{0,1\}$, with $P(X_1 = 1) = P(X_2 = 1) = \rho > 0$, $\mathrm{Cov}(X_1, X_2) > 0$ and $X_3 = X_1 + X_2$. Show that $\mathrm{Cov}(X_1, X_3) > 0$ and $\mathrm{Cov}(X_2, X_3) > 0$, but

$$P(X_1 = 1 | X_2 = 1, X_3 = 1) < P(X_1 = 1 | X_2 = 0, X_3 = 1).$$

How does this violate positive regression dependency?

The above discussion suggests another advantage of using multiple orthogonal contrasts in the sense of having a reduced type I error.

Exercise 2.16 (Positive dependency among test with orthogonal contrasts)**.** Suppose we are in the ANOVA setting, and we have $I-1$ orthogonal contrasts. By using the independence of $SS_{\mathbf{c}_1}, \ldots, SS_{\mathbf{c}_{I-1}}$ (under orthogonality) and SS_w, show that the decisions of the t-tests for those contrasts are positively dependent, in the sense that they satisfy (2.6).

B.2.3 Simultanenous confidence intervals

Recall, again, the duality between confidence intervals and hypothesis tests. Here, we will dwell on that duality, however with a level of abstraction. The discussion here is provided to remove possible confusion and learn a general principle which can be applied to clarify the connection between confidence intervals and hypothesis tests. In particular, the results

we will have by the end of the discussion here should guide the reader through handling multiple hypotheses.

As we said earlier, a $100(1-\alpha)\%$ confidence interval for a component of the parameter vector is an interval whose probability of containing the true value of that component is $1-\alpha$. What is crucial in our context is that this CI can be expressed as subset $\Theta^{\text{CI}} \subseteq \Theta$ of the space of the whole parameter vector, θ. For example, for a sample from the normal population, a $100(1-\alpha)\%$ confidence interval for the mean parameter μ, when the variance is unknown, is traditionally written as

$$\left(\bar{X} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}, \quad \bar{X} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right)$$

While this confidence interval resides in the space of μ, we can convert it to a set for the whole parameter vector $\theta = (\mu, \sigma^2)$. The equivalent set in $\Theta = \mathbb{R} \times [0, \infty)$ to the confidence interval given above is

$$\Theta_{\mu}^{\text{CI}} = \left\{(\mu, \sigma^2) \in \Theta : \bar{X} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}} \text{ and } \sigma^2 > 0\right\}$$

Exercise 2.17. Show that $P(\theta \in \Theta_{\mu}^{\text{CI}}) = 1 - \alpha$.

What does this achieve, other than complicating the notation seemingly unnecessarily? The point here is that every confidence interval corresponds to a subset of Θ. As another example, the two-sided confidence interval for the variance σ^2, which is well known as

$$\left(\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}, \quad \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}\right)$$

corresponds to

$$\Theta_{\sigma^2}^{\text{CI}} = \left\{(\mu, \sigma^2) \in \Theta : -\infty < \mu < \infty \text{ and } \frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}\right\}.$$

What comes to mind after seeing both confidence intervals is providing them as *simultaneous* confidence intervals for (μ, σ^2). This is equivalent to providing the set

$$\Theta^{\text{CI}} = \Theta_{\mu}^{\text{CI}} \cap \Theta_{\sigma^2}^{\text{CI}}$$

for $\theta = (\mu, \sigma^2)$. It is nice to be able to provide simultaneous confidence intervals, however this comes with a cost. Since $\{\theta \in \Theta^{\text{CI}}\} = \{\theta \in \Theta_{\mu}^{\text{CI}})\} \cap \{\theta \in \Theta_{\sigma^2}^{\text{CI}})\}$, with a rough examination using Bonferroni, we can show that

$$1 - 2\alpha \leq P(\theta \in \Theta^{\text{CI}}) \leq 1 - \alpha. \tag{2.9}$$

that is, both intervals hold simultaneously with a probability between $1-2\alpha$ and $1-\alpha$, which is smaller than the $1-\alpha$ coverage probability for each interval.

How can we use Θ^{CI} for testing? Recall that every hypothesis about the parameter θ corresponds to a subset $\Theta_0 \subseteq \Theta$. In the context of the normal population, examples for possible hypotheses are $H_0 : \mu = 0$, $H_0 : \sigma^2 > 1$, $H_0 : \mu = 2, \sigma^2 = 0$, or $H_0 : g(\mu, \sigma^2) = 1$ for some function of g. However, to stress once again, whatever the null hypothesis is, it corresponds to a subset Θ_0 for $\theta = (\mu, \sigma^2)$. Therefore, let us state our null hypothesis as

$$H_0 : \theta \in \Theta_0$$

Given the null hypothesis above, consider the following rejection rule:

$$\text{Decision} = \begin{cases} \text{Reject } H_0 & \text{if } \Theta_0 \cap \Theta^{\text{CI}} = \emptyset \\ \text{Do not reject } H_0 & \text{if } \Theta_0 \cap \Theta^{\text{CI}} \neq \emptyset \end{cases}$$

Exercise 2.18. Show that the size of this test is at most 2α.

Proof. From the definition of the size of a test, we have

$$\begin{aligned} \alpha &= \sup_{\vartheta \in \Theta_0} P\left(\Theta_0 \cap \Theta^{\text{CI}} = \emptyset \middle| \, \theta = \vartheta\right). \\ &\leq \sup_{\vartheta \in \Theta_0} P\left(\theta \notin \Theta^{\text{CI}} \middle| \, \theta = \vartheta\right). \\ &= 1 - \inf_{\vartheta \in \Theta_0} P\left(\theta \in \Theta^{\text{CI}} \middle| \, \theta = \vartheta\right). \\ &\leq 2\alpha \end{aligned}$$

where the second line is from the fact that one element of a set not belonging to another set is more probable than those two sets not intersecting at all, and the last line is by (2.9). So we conclude. □

Let us generalise the discussion above for the normal population to the general setting where we have several confidence intervals which we want to consider simultaneously and a set of null hypotheses. Assume we observe random variables whose distributions are in question, and let θ denote all the parameters of those distributions. Let $g_1, \ldots, g_m$ be any functions with $g_i : \Theta \mapsto \mathbb{R}$ and let Θ_i^{CI} be the set of parameters that corresponds to the $100(1-\alpha_i)\%$ level confidence interval $\text{CI}_i \subseteq \mathbb{R}$ for $g_i(\theta)$. Those sets can be written as

$$\Theta_i^{\text{CI}} = \{\theta \in \Theta : g_i(\theta) \in \text{CI}_i\}, \quad i = 1, \ldots, m \tag{2.10}$$

where each $g_{i,\min} < g_{i,\max} \in \mathbb{R}$. Next, define the intersection $\Theta^{\text{CI}} = \bigcap_{i=1}^m \Theta_i^{\text{CI}}$. This will be the set we will use to decide whether to reject or not reject a null hypothesis. For example, for a null hypothesis $H_0 : \theta \in \Theta_0$, the decision rule

$$\text{Decision} = \begin{cases} \text{Reject } H_0 & \text{if } \Theta_0 \cap \Theta^{\text{CI}} = \emptyset \\ \text{Do not reject } H_0 & \text{if } \Theta_0 \cap \Theta^{\text{CI}} \neq \emptyset \end{cases}$$

yields a test of size at most $1 - \inf_{\vartheta \in \Theta_0} P\left(\theta \in \Theta^{\text{CI}} \middle| \, \theta = \vartheta\right)$ which can be simplified to $1 - P(\theta \in \Theta^{\text{CI}})$ if the probability in the infimum is the same for all $\vartheta \in \Theta_0$.

The discussion above is more relevant to our multiple hypotheses framework when we have multiple null hypotheses. More formally, given $H_{0i} : \theta \in \Theta_{0i}$, for $i = 1, \ldots, m$, consider the decision rule for the i'th hypothesis is

$$\text{Decision} = \begin{cases} \text{Reject } H_{0i} & \text{if } \Theta_{0i} \cap \Theta_i^{\text{CI}} = \emptyset \\ \text{Do not reject } H_{0i} & \text{if } \Theta_{0i} \cap \Theta_i^{\text{CI}} \neq \emptyset \end{cases} \tag{2.11}$$

Exercise 2.19. Show that, size of the i'th test is α_i.

Theorem 2.6. *The FWER of the procedure given above is controlled in the strong sense at* $\alpha = 1 - P(\theta \in \Theta^{CI})$.

Proof. From the definition of FWER,

$$\begin{aligned} \text{FWER} = P(V \geq 1) &= P(\exists j \in J : \Theta_j \cap \Theta_j^{\text{CI}} = \emptyset) \\ &\leq P(\exists j \in J, \theta \notin \Theta_j^{\text{CI}}) \\ &= 1 - P(\theta \in \Theta_j^{\text{CI}}, \forall j \in J) \\ &= 1 - P\left(\theta \in \bigcap_{j \in J} \Theta_j^{\text{CI}}\right) \\ &\leq 1 - P\left(\theta \in \bigcap_{j=1}^{m} \Theta_j^{\text{CI}}\right) \\ &= 1 - P\left(\theta \in \Theta^{\text{CI}}\right) \end{aligned}$$

where the second line comes from the facts that $\theta \in \Theta_j$ holds for all $j \in J$ and the intersection of two sets is a less probable event than an element in one of the sets not belonging to another. □

Typically, the individual hypotheses regard the same functions of the unknown parameter $g_1(\theta), \ldots, g_m(\theta)$ for which we built the confidence intervals. For example, we may have

$$H_{01} : g_1(\theta) = \phi_1, \ldots, H_{0m} : g_m(\theta) = \phi_m. \tag{2.12}$$

A different number of hypotheses than the number of confidence intervals is also possible. The decision rule in (2.11) corresponds to using the i'th confidence interval in (2.10) to test H_{0i} in (2.12), i.e., rejecting H_{0i} when $\phi_i \notin \text{CI}_i$.

A crude upper bound for α in Theorem 2.6 is, as in the Bonferroni method, $\sum_{i=1}^{m} \alpha_i$ or, when the tests are independent or positively correlated, $1 - \prod_{i=1}^{m}(1 - \alpha_i)$ as in the Sidak method. However, in some cases, we can show that this α is much less than those quantities. A well-known instance of that is Scheffe's method, which we will see below.

Simultenous confidence intervals for linear combinations: Recall the ANOVA setting. We are interested in testing linear combinations of the means $\sum_{i=1}^{I} c_i\mu_i$. Moreover, we would like to test those combinations being equal to certain values. As we will see below, it is possible, in a certain sense, to look at all the linear combinations at once. The following theorem provides simultaneous confidence intervals with the overall level of $100(1-\alpha)\%$.

Theorem 2.7 (Scheffe's theorem for linear combinations)**.** *Under the ANOVA setting, for all $\mu_1, \ldots, \mu_I$, we have*

$$P\left(\left|\sum_{i=1}^{I} c_i\bar{X}_i - \sum_{i=1}^{I} c_i\mu_i\right| \le M_{\alpha,I,n-I}\sqrt{\frac{SS_w}{n-I}\sum_{i=1}^{I}\frac{c_i^2}{n_i}}, \quad \forall c_1, \ldots, c_I\right) = 1-\alpha. \tag{2.13}$$

with $M_{\alpha,I,n-I} = \sqrt{If_{\alpha,I,n-I}}$.

We take σ^2 away from consideration since it is not a parameter of interest in the one-way ANOVA setting and therefore define $\theta = (\mu_1, \ldots, \mu_I)$. The probability in Theorem 2.7 can also be rewritten as $P(\theta \in \Theta^{\text{CI}}) = 1-\alpha$ where

$$\Theta^{\text{CI}} = \left\{(\mu_1, \ldots, \mu_I) : \left|\sum_{i=1}^{I} c_i\bar{X}_i - \sum_{i=1}^{I} c_i\mu_i\right| \le M_{\alpha,I,n-I}\sqrt{\frac{SS_w}{n-I}\sum_{i=1}^{I}\frac{c_i^2}{n_i}}, \quad \forall c_1 \ldots, c_I\right\}.$$

Theorem 2.7 can be used for testing multiple hypotheses. Consider the following set of hypothesis tests

$$H_{0j} : \sum_{i=1}^{I} c_{j,i}\mu_i = \phi_j \quad \text{vs} \quad H_{1j} : \sum_{i=1}^{I} c_{j,i}\mu_i \ne \phi_j, \quad j = 1, \ldots, m. \tag{2.14}$$

so that $\Theta_{0j} = \{(\mu_1, \ldots, \mu_I) : \sum_{i=1}^{I} c_{j,i}\mu_i = \phi_j\}$, provided that $H_0 = \bigcap_{i=1}^{m} H_{0i}$ is non-empty. Define the individual confidence intervals, implied by Θ^{CI}, for those linear combinations

$$\text{CI}_j = \left(\sum_{i=1}^{I} c_{j,i}\bar{X}_i - M_{\alpha,I,n-I}\sqrt{\frac{SS_w}{n-I}\sum_{i=1}^{I}\frac{c_{j,i}^2}{n_i}}, \quad \sum_{i=1}^{I} c_{j,i}\bar{X}_i + M_{\alpha,I,n-I}\sqrt{\frac{SS_w}{n-I}\sum_{i=1}^{I}\frac{c_{j,i}^2}{n_i}}\right)$$

and the corresponding sets of θ as $\Theta_j^{\text{CI}} = \{(\mu_1, \ldots, \mu_I) : \sum_{i=1}^{I} c_{j,i}\mu_i \in \text{CI}_j\}$.

Exercise 2.20. For the hypothesis tests in (2.14) and the confidence intervals that follow (2.14), show the following.

(a) $\Theta_j^{\text{CI}} \cap \Theta_{0j} = \emptyset$ if and only if $\phi_j \notin \text{CI}_j$.

(b) For any $m \ge 1$, we have $\Theta^{\text{CI}} \subseteq \bigcap_{j=1}^{m} \Theta_j^{\text{CI}}$.

(c) if each linear combination is tested according to the testing procedure

$$\text{Decision} = \begin{cases} \text{Reject } H_{0j} & \text{if } \phi_j \notin \text{CI}_j \\ \text{Do not reject } H_{0j} & \text{if } \phi_j \in \text{CI}_j \end{cases} \tag{2.15}$$

in (2.15) controls FWER with α in the strong sense.

Exercise 2.21. Consider the testing procedure given in (2.15) for each null hypothesis in (2.14). Show the following results.

(a) For each individual test for $H_{0j} : \sum_{i=1}^{I} c_{j,i}\mu_i = \phi_j$, the critical region of the t-test of size α for H_{0j} covers the critical region of the decision rule in (2.15), i.e., the latter is a subset of the former. Use the fact that for any α, I, and ν, we have $t_{\alpha/2,\nu} \leq \sqrt{I f_{\alpha,I,\nu}}$.

(b) The type I error for each test in (2.15) is less than α.

(c) Therefore, each test is less powerful than the t-test of size α for the same null hypothesis.

B.2.4 Back to contrasts

If we confine the interest linear combinations to only contrasts, we can find narrower simultaneous confidence intervals. Scheffe's method for contrasts is based on the following theorem.

Theorem 2.8 (Scheffe's theorem for contrasts)**.** *Under the ANOVA setting, for all $\mu_1, \ldots, \mu_I$, we have*

$$P\left(\left|\sum_{i=1}^{I} c_i \bar{X}_i - \sum_{i=1}^{I} c_i \mu_i\right| \leq M_{\alpha,I-1,n-I}\sqrt{\frac{SS_w}{n-I}\sum_{i=1}^{I}\frac{c_i^2}{n_i}}, \quad \forall (c_1, \ldots, c_I) \in \mathcal{C}\right) = 1 - \alpha. \tag{2.16}$$

with $M_{\alpha,I-1,n-I} = \sqrt{(I-1) f_{\alpha,I-1,n-I}}$.

Contrast the expression in (2.16) with that in (2.13), the difference is between the constants $M_{\alpha,I-1,n-I}$ and $M_{\alpha,I,n-I}$, respectively. We can show that the former is always smaller. This is the result of a more general fact about *stochastic order*.

Definition 2.14. A random variable X is said to be less than Y in the stochastic order, shown by $X \preceq Y$, if

$$P(X > a) \leq P(Y > a), \quad \forall a \in \mathbb{R}.$$

If the inequality is strict, then the ordering is also strict and we write $X \prec Y$.

The definition is invariant to the joint distribution of X and Y, it only depends on the marginal probability distributions of X and Y. Therefore, we really talk about the stochastic order between two distributions.

Exercise 2.22. Suppose that, $X \preceq Y$. Let U, and V be any pair of random variables such that U and X have the same distribution, and V and Y have the same distribution. Then, $U \preceq V$ also.

Theorem 2.9. *Given ν, let $F_k \sim f_{k,\nu}$. Then, kF_k is stochastically increasing, that is, $(k-1)F_{k-1} \prec kF_k$.*

Proof. Let $X \sim \chi^2_{k-1}$, $Y \sim \chi^2_1$, and $Z \sim \chi^2_\nu$ and X, Y, Z be independent. Define $F_{k-1} = \frac{X/(k-1)}{Z/\nu}$ and $F_k = \frac{(X+Y)/k}{Z/\nu}$, so that $F_{k-1} \sim f_{k-1,\nu}$ and $F_k \sim f_{k,\nu}$. Then $kF_k = \frac{X+Y}{Z/\nu} > \frac{X}{Z/\nu} = (k-1)F_{k-1}$. This shows that kF_k is stochastically increasing, and the validity of the result is independent from how F_k's are constructed. □

Exercise 2.23. Show that $kf_{\alpha,k,\nu} > (k-1)f_{\alpha,k-1,\nu}$.

Proof. Notice that $P((k-1)F_{k-1} > (k-1)f_{\alpha,k-1,\nu}) = P(kF_k > kf_{\alpha,k,\nu}) = \alpha$. But since $(k-1)F_{k-1} \prec kF_k$, we have

$$P((k-1)F_{k-1} > kf_{\alpha,k,\nu}) < P(kF_k > kf_{\alpha,k,\nu}).$$

Finally, by monotonicity of the cdf, we must have $(k-1)f_{\alpha,k-1,\nu} < kf_{\alpha,k,\nu}$ □

Exercise 2.24. Apply the result of Exercise 2.23 to show that

$$t^2_{\alpha/2,n-I} \leq (I-1)f_{\alpha,I-1,n-I} \leq I f_{\alpha,I,n-I}.$$

(Hint: Show that $t^2_{\alpha/2,n-I} = f_{\alpha,1,n-I}$). Therefore, conclude that, with significance levels being equal, the confidence interval for a single linear combination in (2.3) is narrower than the simultaneous Scheffe's interval for all contrasts, which is narrower than the simultaneous Scheffe's intervals for all linear combinations.

Tukey's method for pairwise contrasts: An even smaller family of linear combinations is the family of pairwise comparisons, of the form $\mu_i - \mu_j$ for $i \neq j$. When we confine ourselves to pairwise comparisons, we can get even narrower simultaneous confidence intervals with the same $1-\alpha$ probability. Those simultaneous confidence intervals are due to Tukey.

Theorem 2.10 (Distribution of maximum difference)**.** *Under the ANOVA setting with equal group sizes $n_1 = n_2 = \ldots = n_I$, the statistic*

$$\max_{i,j} \bar{X}_i - \bar{X}_j$$

follows a Studentized range distribution $q_{I,n-I}$ with I populations and $n-I$ degrees of freedom.

The critical values of a Studentized range distribution $q_{I,\nu}$ at α is shown by $q_{\alpha,I,\nu}$ and is tabulated.

Exercise 2.25. Show that $\max_{i,j} |\bar{X}_i - \bar{X}_j| < c$ if and only if $|\bar{X}_i - \bar{X}_j| < c$ for all $i \neq j$.

Theorem 2.10 and the observation in Exercise 2.25 lead to the following simultaneous confidence intervals for pairwise differences

Theorem 2.11. *Tukey confidence intervals for pairwise differences Under the ANOVA setting with equal group sizes $n_1 = n_2 = \ldots = n_I = n_*$, we have, for all $\mu_1, \ldots, \mu_I$,*

$$P\left(\left|\bar{X}_i - \bar{X}_j - (\mu_i - \mu_j)\right| \leq \frac{q_{\alpha,I,n-I}}{\sqrt{2}}\sqrt{\frac{SS_w}{n-I}\frac{2}{n_*}}, \quad \forall i \neq j\right) = 1 - \alpha. \tag{2.17}$$

where $q_{\alpha,I,n-I}$ is the critical value at α of the Studentized range distribution with I populations and $n - I$ degrees of freedom. Therefore, we have $P(\theta \in \Theta^{\text{CI}}) = 1 - \alpha$, where

$$\Theta^{\text{CI}} = \left\{(\mu_1, \ldots, \mu_I) : \left|\bar{X}_i - \bar{X}_j - (\mu_i - \mu_j)\right| \leq \frac{q_{\alpha,I,n-I}}{\sqrt{2}}\sqrt{\frac{SS_w}{n-I}\frac{2}{n_*}}, \quad \forall i \neq j\right\}$$

For the individual comparison between μ_i and μ_j the simultaneous confidence interval is

$$\text{CI}_{i,j} = \left(\bar{X}_i - \bar{X}_j - \frac{q_{\alpha,I,n-I}}{\sqrt{2}}\sqrt{\frac{SS_w}{n-I}\frac{2}{n_*}}, \quad \bar{X}_i - \bar{X}_j + \frac{q_{\alpha,I,n-I}}{\sqrt{2}}\sqrt{\frac{SS_w}{n-I}\frac{2}{n_*}}\right)$$

The use of Tukey's confidence interval is justified by the following theorem:

Theorem 2.12. *For all α, I and ν, we have $q_{\alpha,I,n-I}/\sqrt{2} < M_{\alpha,I-1,n-I}$.*

Data snooping: So far, all the tests we have discussed are *planned* experiments. A planned experiment is designed *before* looking at the data. In contrast to a planned experiment, there are *post hoc* experiments which are conducted *after* looking at the data. This is called 'data snooping', which introduces errors in the experiments if one is not careful. To understand the potential drawback of data snooping, consider the ANOVA setting where we are particularly interested in pairwise comparisons. For each pairwise comparison, we have a hypothesis test available whose type I error is α. For example, for $H_0 : \mu_i = \mu_j$, a suitable test statistic is

$$\frac{\bar{X}_i - \bar{X}_j}{\sqrt{\frac{SS_w}{n-I}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \sim t_{n-I}$$

Suppose that we are also committed to use the critical value $t_{\alpha/2,n-I}$ for testing, i.e., the critical region $\{|t_{obs}| > t_{\alpha/2,n-I}\}$. Now, consider two different scenarios:

- In the first scenario, we decide to compare μ_2 and μ_3 *before* looking at the data. Then, we collect the data and perform our test above for $H_0 : \mu_2 = \mu_3$. The type I error we commit in this case is simply α.

- In the second scenario, we look at the data, find the pair (i^*, j^*) such that $\bar{X}_i - \bar{X}_j$ is maximised at $i = i^*, j = j^*$, and then decided to test the null hypothesis that $H_0 : \mu_i = \mu_j$ by using the same test statistic above. The error of this test is higher than α, since the actual test statistic of this procedure corresponds to

$$\max_{i,j} \frac{\bar{X}_i - \bar{X}_j}{\sqrt{\frac{SS_w}{n-I}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}},$$

  which has a higher probability (than α) of exceeding $t_{\alpha/2,n-I}$ under the null hypothesis of equal means. Therefore, applying the same test procedure for a hypothesis after looking at the data can be misleading.

Exercise 2.26. Why do not we have absolute values of the test statistic in the discussion above?

Now that the danger of data snooping is (hopefully) clear, let us consider a new test procedure that can handle multiple hypotheses and give reliable results even under data snooping. Continuing with the same example, we wish to find a critical value k_α such that

$$P\left(\max_{i,j} \frac{\bar{X}_i - \bar{X}_j}{\sqrt{\frac{SS_w}{n-I}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} > k_\alpha\right) = \alpha.$$

under the null hypothesis $\mu_1 = \ldots = \mu_I$. Notice that the condition that the maximum difference be larger than a value can also be written as all of the pairwise differences being larger than that value, in particular,

$$\max_{i,j} \frac{\bar{X}_i - \bar{X}_j}{\sqrt{\frac{SS_w}{n-I}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} > k_\alpha \Leftrightarrow \frac{\bar{X}_i - \bar{X}_j}{\sqrt{\frac{SS_w}{n-I}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} > k_\alpha \text{ for all } i, j.$$

This k_α is provided by Scheffe's method and by Tukey's method when group sizes are equal. Using them, we can test all the hypotheses of the form $\mu_i = \mu_j$ at the same time with an overall type I error of α. Therefore, we have seen another advantage of simultaneous confidence intervals: We can use them to data snoop safely. For another example, similarly to pairwise comparisons, one can be interested in testing the linear combination $\sum_{i=1}^I c_i\mu_i$ being equal to 0 or not for which

$$\frac{\sum_{i=1}^I c_i\bar{X}_i}{\sqrt{\frac{SS_w}{n-I}\sum_{i=1}^I \frac{c_i^2}{n_i}}}$$

is maximised, and the theorem for linear combinations suggests that this can be done safely.

Exercise 2.27. Suppose we are in the one way ANOVA setting with $I = 5$, with $n_i = 7$ for all $i = 1, \ldots, 5$ so that $n = 35$. We are supposed to test $m \geq 1$ different null hypotheses, each claiming a different linear combination being equal to 0. We want the FWER controlled at $\alpha = 0.05$, however keeping the power of each test as big as possible.

Consider two methods for controlling FWER: (i) Bonferroni correction along with a t-test, and (ii) one that uses simultaneous confidence intervals based on Scheffe's theorem for linear combinations, i.e., Theorem 2.7.

Which method would you prefer to use and how would your choice depend on m? Make a numerical study, and report your answer for $m = 1, 2, \ldots, 100$. (For example, make a plot.) [Hint: Remember the duality between confidence intervals and hypothesis tests, and that a wider confidence interval implies a less powerful test.]

Exercise 2.28. Do Exercise 2.27 again, but this time instead of linear combinations in general we are only interested in contrasts. (Hence you should consider Scheffe's theorem for contrasts for the second method in (ii).)

Exercise 2.29. Do Exercise 2.27 again, but this time instead of linear combinations in general we are only interested in pairwise comparisons. (Hence you should consider Tukey's theorem for pairwise comparisons for the second method in (ii).) Also, limit the range of m values to $m = 1, \ldots, 10$. (Why?)

Exercise 2.30. To asses the relative effects of three toxins and a control on the liver of a certain species of trout, the amounts of deterioration (in standard units) of the liver in each sacrificed fish are measured and the data are shown in the following the table.

Toxin 1	Toxin 2	Toxin 3	Control
28	33	18	11
23	36	21	14
14	34	20	11
27	29	22	16
	31	24	
	24		

Let μ_i be the mean effect of the toxins ($i = 1, 2, 3$) and the control ($i = 4$).

(a) Test the basic ANOVA null hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ against $H_1 :$ not all means are equal with a significance level of $\alpha = 0.1$.

(b) Provide a simultaneous $100(1 - \alpha)\%$ confidence interval for all pairwise differences, with $\alpha = 0.1$. Do your best to make CI be as narrow as possible. (Why do we desire narrower confidence intervals with the same confidence level?) Indicate which method you use to build the confidence interval.

(c) This time we want to test the following hypotheses altogether:

- $H_{01} : \mu_1 = \mu_4$ vs $H_{11} : \mu_1 \neq \mu_4$.
- $H_{02} : \mu_2 = \mu_4$ vs $H_{12} : \mu_2 \neq \mu_4$.

- $H_{03} : \mu_3 = \mu_4$ vs $H_{13} : \mu_3 \neq \mu_4$.
- $H_{04} : \frac{\mu_1+\mu_2+\mu_3}{3} = \mu_4$ vs $H_{14} : \frac{\mu_1+\mu_2+\mu_3}{3} \neq \mu_4$.

(i) Show that all of the linear combinations in the null hypotheses are contrasts.

(ii) Conduct those hypotheses tests while controlling the FWER at $\alpha = 0.1$ by using the Bonferroni method.

(iii) Conduct those hypotheses tests while controlling the FWER at $\alpha = 0.1$ by making use of simultaneous confidence intervals as discussed in Section B.2.3.

(iv) Compare the tests in the last two parts. Which one do you prefer for this case?

C Two-way layout

This time we have two factors, having $I > 1$ and $J > 1$ layers, respectively. For each combination (i, j) of those factors, we have K_{ij} independent observations. In this section we will simplify the setting by allowing $k_{ij} = K$ for all combinations.

A two-way layout is an experimental design involving two factors, each at two or more levels. The levels of one factor might be various drugs, for example, and the levels of the other factor might be sex ($J = 2$). If there are I levels of one factor and J of the other, there are $I \times J$ combinations. We will assume that K independent observations are taken for each of these combinations.

Again, we have the normality and independence assumption. The variations along the first and second factors will be represented by a_i and b_j, respectively, and the interaction between the i'th level of the first factor and the j'th level of the second factor will be represented by δ_{ij}. The resulting observation model is

$$X_{ijk} = \mu + a_i + b_j + \delta_{ij} + \epsilon_{ijk}, \tag{2.18}$$

for all $i = 1, \ldots, I$, $j = 1, \ldots, J$, and $k = 1, \ldots, K$, where

$$\epsilon_{ijk} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$$

and the differential effects a_i and b_j, are normalised

$$\sum_{i=1}^{I} a_i = 0, \quad \sum_{j=1}^{J} b_j = 0, \tag{2.19}$$

and the interaction effects are also normalised

$$\sum_{j=1}^{J} \delta_{ij} = 0, \quad i = 1, \ldots, I; \quad \sum_{i=1}^{I} \delta_{ij} = 0, \quad j = 1, \ldots, J. \tag{2.20}$$

C.1 A motivation: Randomised block design

Blocking is the arranging of experimental units in groups (blocks) in which the member units are similar to one another. An example of a blocking factor might be the sex of a patient. Typically, a blocking factor is a source of variability that is not of primary interest to the experimenter, For example, the primary interest is the effect of a new drug on patients while the blocking factor is the sex of a patient. By blocking, variability due to the blocking factor (in this example, sex) is controlled for, thus leading to greater accuracy (in tests about the effect of the drug).

A nuisance factor is used as a blocking factor if every level of the primary factor occurs the same number of times with each level of the nuisance factor. The analysis of the experiment will focus on the effect of varying levels of the primary factor within each block of the experiment.

Example 2.1. The table below shows a randomized block design for a hypothetical medical experiment. Subjects are assigned to blocks, based on sex. Then, within each block, subjects are randomly assigned to treatments (no vaccine, a placebo, or a cold vaccine).

Treatment	Male	Female
no vaccine	$X_{111}, \ldots, X_{11K}$	$X_{121}, \ldots, X_{12K}$
placebo	$X_{211}, \ldots, X_{21K}$	$X_{221}, \ldots, X_{22K}$
vaccine	$X_{311}, \ldots, X_{31K}$	$X_{321}, \ldots, X_{32K}$

Example 2.2. Randomized block designs originated in agricultural experiments. To compare the effects of I different fertilizers, J relatively homogeneous plots of land, or blocks, are selected, and each is divided into I plots. Within each block, the assignment of fertilizers to plots is made at random. By comparing fertilizers within blocks, the variability between blocks, which would otherwise contribute "noise" to the results, is controlled. This design is a multisample generalization of a matched-pairs design.

Fertilizer	Plot 1	...	Plot J
1			
$\vdots$			
I			

Example 2.3. Another example of a randomized block design is one used by a nutritionist who wants to compare the effects of three different diets on experimental animals. To control for genetic variation in the animals, the nutritionist might select three animals from each of several litters and randomly determine their assignments to the diets.

The model for a randomized block design may be

$$X_{i,j,k} = \mu + a_i + b_j + \epsilon_{i,j,k}, \tag{2.21}$$

which is a simplification over the general 2-way ANOVA in that $\delta_{ij} = 0$.

C.2 Inference

Given observations $X_{ijk} = x_{ijk}$, the log-likelihood function for the parameters μ, a_i's, β_j's, and δ_{ij}'s is given by

$$-\frac{IJK}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(x_{ijk} - \mu - a_i - b_j - \delta_{ij})^2,$$

Define the following averages

$$\bar{X}_{ij} = \frac{1}{K}\sum_{k=1}^{K} X_{ijk}, \qquad \bar{X}_{i.} = \frac{1}{JK}\sum_{j=1}^{J}\sum_{k=1}^{K} X_{ijk},$$

$$\bar{X}_{.j} = \frac{1}{IK}\sum_{i=1}^{I}\sum_{k=1}^{K} X_{ijk} \qquad \bar{X} = \frac{1}{IJK}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} X_{ijk}$$

Exercise 2.31. Show that the maximum likelihood estimator for the parameters μ, a_i, b_j, δ_{ij}, under the given constraints in (2.20), are given by

$$\hat{\mu} = \bar{X}, \quad \hat{a}_i = \bar{X}_{i.} - \bar{X}, \quad \hat{b}_j = \bar{X}_{.j} - \bar{X}, \quad \hat{\delta}_{ij} = \bar{X}_{ij} - \bar{X}_{.j} - \bar{X}_{i.} + \bar{X}$$

Exercise 2.32. Show that $\bar{X}_{ij} = \hat{\mu} + \hat{a}_i + \hat{b}_j + \hat{\delta}_{ij}$.

Define the sums of squares

$$SS_A = JK\sum_{i=1}^{I}(\bar{X}_{i.} - \bar{X})^2$$

$$SS_B = IK\sum_{j=1}^{J}(\bar{X}_{.j} - \bar{X})^2$$

$$SS_{AB} = K\sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{X}_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X})^2$$

$$SS_E = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(X_{ijk} - \bar{X}_{ij})^2$$

$$SS_{total} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(X_{ijk} - \bar{X})^2$$

Exercise 2.33. Show that

(a) The sums of squares split the total error

$$SS_{total} = SS_A + SS_B + SS_{AB} + SS_E.$$

(b) SS_A, SS_B, SS_{AB}, and SS_E are independent.

(c) The sums of squares have the following expectations

$$E(SS_A) = \sigma^2(I-1) + JK\sum_{i=1}^{I} a_i^2$$

$$E(SS_B) = \sigma^2(J-1) + IK\sum_{j=1}^{J} b_j^2$$

$$E(SS_E) = IJ(K-1)\sigma^2$$

$$E(SS_{AB}) = \sigma^2(I-1)(J-1) + K\sum_{i=1}^{I}\sum_{j=1}^{J} \delta_{ij}^2$$

Next, we state the distributions of some test statistics used to test certain hypotheses about the model.

Exercise 2.34. Show that

(a) $SS_E/\sigma^2 \sim \chi^2_{IJ(K-1)}$.

(b) Under $H_0 : a_1 = \ldots = a_I = 0$, we have $SS_A/\sigma^2 \sim \chi^2_{I-1}$ and therefore

$$\frac{SS_A/(I-1)}{SS_E/(IJ(K-1))} \sim f_{I-1,IJ(K-1)}$$

(c) Under $H_0 : b_1 = \ldots = b_J = 0$, we have $SS_B/\sigma^2 \sim \chi^2_{J-1}$ and therefore

$$\frac{SS_B/(J-1)}{SS_E/(IJ(K-1))} \sim f_{J-1,IJ(K-1)}$$

(d) Under $H_0 : \delta_{ij} = 0$ for all i, j, we have $SS_{AB}/\sigma^2 \sim \chi^2_{(I-1)(J-1)}$ and therefore

$$\frac{SS_{AB}/((I-1)(J-1))}{SS_E/(IJ(K-1))} \sim f_{(I-1)(J-1),IJ(K-1)}$$

[Hint: Look up the addition properties of the non-central chi-square distribution to derive the distribution of SS_{AB}/σ^2.]

Exercise 2.35. Suppose you want to determine whether the brand of laundry detergent used and the temperature affects the amount of dirt removed from your laundry. To this end, you buy two different brands of detergent ("Super" and "Best") and choose three different temperature levels ("cold", "warm", and "hot"). Then you divide your laundry randomly into $6 \times r$ piles of equal size and assign each group of r piles into the combination of ("Super" and "Best") and ("cold", "warm", and "hot").

(a) Determine the factors and their layers in this experiment.

(b) With $r = 4$, the amounts of dirt removed when washing sub pile k ($k = 1, 2, 3, 4$) are recorded as

	Cold	Warm	Hot
Super	4, 5, 6, 5	7, 9, 8, 12	10, 12, 11, 19
Best	6, 6, 4, 4	13, 15, 12, 12	12, 13, 10, 13

Using the data in the table, fill in the following table by replacing expressions with actual numbers. (Note: $MS = SS/$degrees of freedom)

Source	degrees of freedom	SS	MS	F
A	$I - 1$	SS_A	MS_A	MS_A/MS_E
B	$J - 1$	SS_B	MS_B	MS_B/MS_E
$A \times B$	$(I - 1)(J - 1)$	SS_{AB}	MS_{AB}	MS_{AB}/MS_E
within	$IJ(K - 1)$	SS_E	MS_E	
Total	$IJK - 1$	SS_{total}		

(c) Test the null hypothesis at $\alpha = 0.1$ that the amount of dirt removed does not depend on the type of detergent.

(d) Find a $100(1 - \alpha)\%$ confidence interval for the difference between the means of the amount of dirt removed with detergents Super and Best.

(e) Test the null hypothesis at $\alpha = 0.1$ that the amount of dirt removed does not depend on the temperature.

Chapter 3

Linear Regression

In this chapter, we discuss a basic model for analysis of multivariate data, the linear regression model. Regression models are used to model the dependence between a random variable, which can be treated as responses, and some other variables, which can be treated as predictor variables, when that dependence is believed to be in a linear fashion. We start with the simple linear regression model, where there is a single predictor variable, and continue with the multiple linear regression model, which has multiple predictor variables. In both models, simple and multiple, we reserve a detailed discussion of the inferential properties under normality assumptions.

A Simple linear regression

In cases where simple linear regression is of consideration, one has a collection of pairs of numbers

$$(x_1, y_1), \ldots, (x_n, y_n)$$

which are simply points on the $x - y$ plane. Linear regression models are examples where a functional dependence of one variable on another is assumed (or sought, as far as testing is concerned). The analysis required to study this functional dependence depends on the nature of (x_i, y_i). However, as long as simple linear regression is concerned, this functional dependency can generally be written as

$$y_i = a + bx_i + e_i, \tag{3.1}$$

regardless of the assumptions that underlie generation of (x_i, y_i)'s. Here, $y = a + bx$ stands for the functional dependency and e_i is the amount of deviation from the exact relationship stated by $y = a + bx$.

In linear regression models, with a minimal statistical flavour, y_i is treated as the observed value of a random variable Y_i, while x_i is kept non-random as before. Therefore, (3.1) is modified as

$$Y_i = a + bx_i + e_i. \tag{3.2}$$

Note that this time the e_i's are necessarily random, which is what makes Y_i's random variables. It is common to assume that e_i's are independent and have zero mean. The random e_i in (3.2) contrasts with e_i in (3.1), which is merely a deviation due to the imperfect fit of the linear line $y = a + bx$ to the data.

The model is built such that the aim is often to predict, or estimate, Y_0, a new observation to be received, from the knowledge of x_0, which is known to be paired with Y_0. That is why the language in regression analysis is asymmetric in general. It is typically said that Y_i *depends* on x_i. In the literature, there are several ways to refer to the variables Y_i and x_i. One common way is to refer to Y_i as the "dependent" variable and x_i as the "independent" variable. However, this terminology may be confusing since x_i and Y_i are not statistically independent. With reference to the usual aim of predicting Y_0 from x_0, another terminology is to refer to Y_i as the *response* variable, and x_i as the *predictor* variable.

The fact that the aim of regression analysis is usually predicting the response variable Y given the predictor variable x is important to keep in mind. This aim underlies most of the aspects of regression analysis and the choices to build a regression model in certain ways (and not in other ways). Since the prediction problem is formalised as predicting Y given x, we are interested in the distribution of Y conditional on x. Equation (3.2) expresses that conditional distribution in a generative manner.

We have two different possible cases for the nature of the predictor variable.

- In one case, x_i's are design variables, on which the experimenter has full (or partial) control. For example, assume that an experimenter wants to find out a relation between the time needed to cook a pizza to a certain degree of crispness and temperature. Then, the predictor variable x_i is the temperature, which can surely be designed by the experimenter, hence is a design variable. The response variable, Y_i, is the cooking time, and, accounting for the stochasticity involved in the cooking process, it should be treated as a random variable. In such an experiment, $x_1, \ldots, x_n$ can be chosen to maximize the precision of the regression analysis. Whatever the choice for x_i's is, we base our inference on the conditional distribution of Y_i, which is expressed in (3.2).

- In the second case, x_i is the observed value of a random variable X_i, which cannot be controlled by the experimenter. (In the models covered in this chapter, we assume that $X_i = x_i$ is observable; there exist models that account for *latent* predictors.) This case naturally occurs when a data pair x_i, y_i are collected together. For example, in an experiment conducted in search of a possible relation between happiness (response) and wealth (predictor), the collected data will be a sample of randomly selected people, whose wealth we do not have a control on. The regression equation relating the random variables can be written as

$$Y_i = a + bX_i + e_i. \tag{3.3}$$

  Since we are interested in predicting the response from the predictor, the inference will be based on the conditional distribution of Y_i given $X_i = x_i$, which is implied by the very same equation (3.2).

$$Y_i = a + bx_i + e_i.$$

Therefore, our inferential analysis will not be different from the first case where x_i is a design variable.

One use of treating x_i as the observed value of a random variable is to justify the linear regression. For example, as we will see later, when X_i, Y_i are bivariate normal, then the relationship in (3.3) follows with suitable choices for a and b in terms of the moments (means, variances, and the cross-covariance) of the bivariate distribution.

Recall that linear regression is an example of functional dependence of the response variable on the predictor variable. But what do we mean by "functional dependence"? One way to define it is via the conditional expectation of the response variable given the predictor variable, which is known as the *population regression function*, or shortly *regression function*.

Definition 3.1 (Population regression function)**.** Let (x, Y) be a pair of variables, the former being the predictor and the latter being the response variable. The conditional expectation of Y given x, denoted by $E(Y|x)$, is called the *population regression function*.

For the relation given in (3.2), assuming $E(e) = 0$, we have

$$E(Y|x) = a + bx. \tag{3.4}$$

The population regression function in (3.4) is linear in x, but also in a and b. In the above equation, a and b are referred to as the parameters of the regression. The term *linear regression* refers to a specification that is *linear in parameters* (not in the predictor variable!).

Definition 3.2 (Linear regression)**.** A population regression function specifies a linear regression if it is linear in the parameters of the regression.

For example, with a and b being the parameters of the regression, $E(Y|x) = a + bx^2$ specifies a linear regression (a linear relationship between Y and x^2). The regression function $E(Y|x) = a\exp(bx)$ does not specify a linear regression, though.

A.1 Least squares solution

A non-statistical, but quite reasonable, method is to fit a straight line through the points $(x_1, y_1), \ldots, (x_n, y_n)$ that is as "close" to the data points as possible. One way of measuring the goodness of a fitted line is to look at the *residual sum of squares (RSS)*. If the fitted line is $y = a + bx$, then RSS is defined as

$$\text{RSS} = \sum_{i=1}^{n} (y_i - a - bx_i)^2. \tag{3.5}$$

The *least squares* solutions of a and b are defined to be those values $\tilde{a}$ and $\tilde{b}$ that minimise the RSS, that is

$$\min_{a,b} \sum_{i=1}^{n} (y_i - a - bx_i)^2 = \sum_{i=1}^{n} (y_i - \tilde{a} - \tilde{b}x_i)^2$$

Another way to express the above is to use the arg min (minimising argument) notation,

$$(\tilde{a}, \tilde{b}) = \arg\min_{a,b} \sum_{i=1}^{n} (y_i - a - bx_i)^2.$$

which is well defined when the solution is unique.[1] The solution is usually expressed in terms of the following quantities. The sample means are defined as usual,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Next, define the sums of squares

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2, \quad S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2. \tag{3.6}$$

and the sum of cross products

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}). \tag{3.7}$$

The quantities defined above will be extensively used both in the non-statistical and statistical settings. We can now state the least square solution for (a, b).

Exercise 3.1. Show that the least-squares solution for (a, b) is given by

$$\tilde{b} = \frac{S_{xy}}{S_{xx}}, \quad \tilde{a} = \bar{y} - \tilde{b}\bar{x}.$$

The least squares solution minimises the RSS, which is one way of measuring the distance between the fitted line $y = a + bx$ and the data points. There are several other ways to measure the distance. If we scatterplot the data points (x_i, y_i), $i = 1, \ldots, n$ and superimpose the line $y = a + bx$, the residual error for the i'th data point, the residual $\tilde{e}_i = y_i - a - bx_i$ can be seen to be the vertical distance between the line and (x_i, y_i). Therefore, $\tilde{a}$ and $\tilde{b}$ minimize the sum of squares of those vertical distances.

Rather than vertical distances, one may be interested in horizontal distances instead. This corresponds to minimizing

$$\sum_{i=1}^{n} (x_i - a' - b' y_i)^2. \tag{3.8}$$

over a' and b', where a' and b' are such that $y = a + bx$ can be rewritten as $x = a' + b'y$. Hence, we have $a' = -a/b$ and $b' = 1/b$. The minimization of (3.8) follows very similar lines as those in minimising the RSS and the following result should be easy to derive.

[1] In general, $\arg\min_x f(x)$ is defined as the set of x values for which the minimum is attained.

Exercise 3.2. Show that the values of a' and b' that minimises (3.8) is given by

$$\tilde{b'} = \frac{S_{xy}}{S_{yy}}, \quad \tilde{a'} = \bar{x} - \tilde{b'}\bar{y}.$$

Exercise 3.2 reveals the slope of the fitted line according to (3.8) is $1/\tilde{b'} = S_{yy}/S_{xy}$. Compare that to the least squares solution $\tilde{b} = S_{xy}/S_{xx}$ that minimises the RSS: which slope is bigger? In fact the ratio between those slopes reveals the answer:

$$\frac{\tilde{b}}{1/\tilde{b'}} = \tilde{b}\tilde{b'} = \frac{S_{xy}}{S_{xx}}\frac{S_{xy}}{S_{yy}} = \frac{S_{xy}^2}{S_{xx}S_{yy}}$$

One can show that the above ratio is always less than or equal to 1. The key to derive that relation is Holder's inequality.

Theorem 3.1 (Hölder's inequality for real valued numbers)**.** *For all $n \geq 1$, $p, q \in (1, \infty)$ such that $1/p + 1/q = 1$, we have*

$$\sum_{i=1}^{n} |a_i b_i| \leq \left(\sum_{i=1}^{n} |a_i|^p \right)^{\frac{1}{p}} \left(\sum_{i=1}^{n} |b_i|^q \right)^{\frac{1}{q}} \quad \textit{for all } (a_1, \ldots, a_n), (b_1, \ldots, b_n) \in \mathbb{R}^n. \tag{3.9}$$

Exercise 3.3. Using Hölder's inequality with $p = q = 2$, $a_i = x_i - \bar{x}$, and $b_i = y_i - \bar{y}$, show that $S_{xy}^2/(S_{xx}S_{yy}) \leq 1$.

The least-square method (or the one we discussed in comparison to it) is a data-oriented method, which only summarises the sample data by fitting a line to it. This data-fitting operation is not a subject of statistics in the classical sense. The available data are not treated as a sample from a population. Hence there are no distribution assumptions about the data. This means that a and b are not treated as population parameters but merely coefficients of the linear fit to the data. This is why "least-squares *solution*" may be a more appropriate term than "least-squares *estimation*" for this method.

The following section will make some minimal statistical assumptions about the data and present an estimator for a and b that has appealing properties under fairly general conditions. It turns out that that estimator is the same as the least-squares solution!

A.2 Best linear unbiased estimator

A.2.1 A general statistical model for simple linear regression

This section considers a fairly general statistical linear regression model for the observed data. The predictor variables $x_1, \ldots, x_n$ are fixed and known. We can assume either that the predictor variables $x_1, \ldots, x_n$ are chosen by the experimenter or that they are observations of random variables; the subsequent analysis will not change. As for the

response variables, we assume that $y_1, \ldots, y_n$ are observed values of uncorrelated random variables $Y_1, \ldots, Y_n$. The linear relationship between x_i and Y_i is given by

$$Y_i = a + bx_i + e_i \tag{3.10}$$

where e_i's are uncorrelated random variables with zero mean $E(e_i) = 0$ and equal unknown variance $V(e_i) = \sigma^2$. Several properties of Y_i's follow. The expectation and the variance are given by

$$E(Y_i) = a + bx_i, \quad V(Y_i) = \sigma^2, \quad i = 1, \ldots, n$$

where we suppress the notation for the conditioning on the x variable in the expectation, since this conditioning is there for the whole analysis that follows. Also, the uncorrelatedness among e_i's is inherited by Y_i's.

Exercise 3.4. Show that Y_i's are uncorrelated, that is, $\mathrm{Cov}(Y_i, Y_j) = 0$ for any $i \neq j$.

Note the generality of this model: the parameters a, b, and the variance parameter σ^2 determine only the first and the second moments of Y_i's. Thus, any sort of inference we make under those assumptions must be valid for all populations satisfying those assumptions. In the following sections, we will be more specific and let Y_i's be normal random variables; this will be mostly because the normality assumption allows us to go further in our inference and derive other sorts of information about the population parameters a and b, such as confidence intervals, tests, etc.

A.2.2 Estimation of regression parameters

How do we estimate a and b in this general simple linear regression model given the data $(x_1, y_1), \ldots, (x_n, y_n)$? It is important to note that we cannot use distribution based methods such as the method of moments or maximum likelihood, since we do not fully know the conditional distribution of Y_i given x_i. Rather, only the first two moments are available. One convenient way to estimate a, b is to restrict the attention to *linear estimators*.

Definition 3.3 (Linear estimator)**.** An estimator is called linear if it is of the form $\sum_{i=1}^n d_i Y_i$.

Among the class of linear estimators, we further restrict the attention to unbiased estimators. Among the unbiased estimators, we want to find the "best" estimators in terms of variance.

Definition 3.4 (Best linear unbiased estimator)**.** A linear and unbiased estimator is called a *best linear unbiased estimator* (BLUE) if it has the minimum variance among all linear and unbiased estimators.

Below we present the best linear unbiased estimators for the parameters a and b as a theorem.

Theorem 3.2 (BLUE for a, b)**.** *Given $(x_1, Y_1), \ldots, (x_n, Y_n)$, the best linear unbiased estimators for a and b are given as*

$$\hat{A} = \bar{Y} - \bar{x}\frac{S_{xY}}{S_{xx}}, \quad \hat{B} = \frac{S_{xY}}{S_{xx}} \tag{3.11}$$

The theorem should be striking to the careful reader: *The observed values of the BLUE estimators, $\hat{a}$ and $\hat{b}$, and the least-squares solution coincide!*

In the following, we will prove Theorem 3.2. Our key result for the proof is the Gauss-Markov theorem, which concerns a general linear model, of which the multiple linear regression model, with any number of predictors, is a special version. We now take a detour by introducing the general linear model, and the multiple regression model, state the Gauss-Markov theorem in the setting of the general linear model and finally deduce the result for the simple linear regression model.

Multiple linear regression as a linear model: Firstly, we will set up the scene for the Gauss-Markov theorem. Let X be an observable $n \times (k+1)$ matrix and $y = (y_1, \ldots, y_n)^T$ be a $n \times 1$ vector,

$$X = \begin{bmatrix} x_{1,0} & x_{1,1} & \ldots & x_{1,k} \\ x_{2,0} & x_{2,1} & \ldots & x_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,0} & x_{n,1} & \ldots & x_{n,k} \end{bmatrix}, \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}.$$

The pair X, y is an often encountered form of data in problems where one is interested in finding a relation between the rows of X and the respective elements of y. One convenient way to model such a relationship between X and y is define a linear relationship with possible deviation. More concretely, one may consider modeling the relation as

$$y = X\beta + e \tag{3.12}$$

where $\beta = (\beta_0, \beta_1, \ldots, \beta_k)^T$ is a parameter vector defining the linear relationship and $e = (e_1, \ldots, e_n)^T$, with e_i being the deviation from the linear relationship between the i'th row of X and y_i, due to noise in the measurements, for example. Furthermore, we assume that X has rank $k+1$, so that $X^T X$ is invertible.

When linear regression is concerned, it is customary to set the first (or the last) column of X to all 1's, so that $x_{i,0} = 1$ for all $i = 1, \ldots, n$. Such a choice of X yields the formulae

$$y_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_k x_{i,k} + e_i, \quad i = 1, \ldots, n, \tag{3.13}$$

(recall $x_{i,0} = 1$) where e_i is the deviation from the linear regression. This form of relation is referred to as multiple linear regression. The component β_0 is called the intercept parameter, as in the simple linear regression model. The model in (3.13) has k predictor variables for each response variable. This should remind the reader the simple linear regression model, which is easily obtained by taking $k = 1$, $x_{i,0} = 1$, $x_{i,1} = x_i$, $\beta_0 = a$ and $\beta_1 = b$.

Least squares solution: As far as the model in (3.12) is concerned, the residual sum of squares RSS can be defined in a similar fashion to the simple linear regression model as

$$\text{RSS} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{i,j} \right)^2$$

With matrix-vector notation, the above can be rewritten as

$$\begin{aligned} \text{RSS} &= ||y - X\beta||_2^2 \\ &= (y - X\beta)^T (y - X\beta) \\ &= y^T y - 2\beta^T X^T y + \beta^T X^T X \beta. \end{aligned}$$

The least-squares solution for β is the value $\tilde{\beta}$ that minimises the RSS.

Theorem 3.3. *The least squares solution for β is given by*

$$\tilde{\beta} = (X^T X)^{-1} X^T y.$$

The following discussion is devoted to the proof of Theorem 3.3. Minimisation with respect to the vector β can be performed by taking the derivative of RSS with respect to β. Differentiating an expression with respect to a vector returns a vector of element-wise differentiations, which is also called the gradient of f. More concretely,

Definition 3.5. For a function $f : \mathbb{R}^m \to \mathbb{R}$, the vector

$$\frac{\partial}{\partial v} f(v) = \left(\frac{\partial}{\partial v_1} f(v), \ldots, \frac{\partial}{\partial v_m} f(v) \right)^T$$

is called the gradient of f with respect to v.

We need the following facts regarding differentiating a scalar, obtained by matrix-vector products, with respect to a vector.

Lemma 3.1. Let x and y be a $m \times 1$ vector, A be a $m \times m$ matrix, and x, y, and A be independent variables (not in statistical terms, but in the sense that one is not defined through any of the others). Then, we have

$$\frac{\partial}{\partial x} x^T y = \frac{\partial}{\partial x} y^T x = y, \quad \frac{\partial}{\partial x} x^T A x = (A + A^T) x.$$

Using Lemma 3.1, the gradient of RSS with respect to β is

$$\frac{\partial}{\partial \beta} (y^T y - 2\beta^T X^T y + \beta^T X^T X \beta) = -2X^T y + 2X^T X \beta. \tag{3.14}$$

Since X has rank $k+1$, the matrix $X^T X$ is invertible. In that case, a solution exists for

$$-2X^T y + 2X^T X \beta = 0 \Leftrightarrow X^T X \beta = X^T y$$

and it is given by

$$\tilde{\beta} = (X^T X)^{-1} X^T y.$$

To establish the least-squares solution fully, we need to ensure the Hessian of RSS, which corresponds to the second derivative in the scalar case, is positive definite. We will make the definitions of those terms.

Definition 3.6 (Hessian)**.** For a function $f : \mathbb{R}^m \to \mathbb{R}$, the square matrix H of second-order partial derivatives, given by

$$H_{ij}(v) = \frac{\partial^2 f(v)}{\partial v_i \partial v_j}, \quad i = 1, \ldots, m; j = 1, \ldots, m$$

is called the Hessian matrix of f at v.

Definition 3.7 (Positive semi-definiteness)**.** An $m \times m$ square matrix A is

- positive definite if $x^T Ax > 0$ for every vector $x \in \mathbb{R}^m$,
- positive semi-definite if $x^T Ax \geq 0$ for every vector $x \in \mathbb{R}^m$,
- negative definite if $x^T Ax < 0$ for every vector $x \in \mathbb{R}^m$,
- negative semi-definite if $x^T Ax \leq 0$ for every vector $x \in \mathbb{R}^m$.

Exercise 3.5. Finish the proof of Theorem 3.3 by showing that the Hessian of RSS at $\tilde{\beta}$ is positive definite.

We have only derived the least-squares solution for the data X, y. Again, the least-squares solution is only a data-driven 'solution' which is not a statistical matter. This is because no statistical assumptions are made for X or y. However, as we will see soon, the least-squares solution is quite relevant to a statistical formulation of the linear model in (3.12).

Just like in the simple linear regression case, a minimal extension towards statistical modelling is to assume that $y_1, \ldots, y_n$ are observed values of the random variables $Y_1, \ldots, Y_n$. Defining the collection of those variables as a vector $Y = (Y_1, \ldots, Y_n)^T$, we have the model

$$Y = X\beta + e$$

where this time e is a vector of uncorrelated random noise terms with $E(e_i) = 0, V(e_i) = \sigma^2$ for some unknown σ^2 and $\mathrm{Cov}(e_i, e_j) = 0$ for $i \neq j$.

In parallel to the simple linear regression model, an estimator of β is said to be linear if each $\hat{\beta}_j$ is of the form $\sum_{i=1}^n d_{j,i} Y_i$. Hence a linear estimator for β is given by DY, where D is a $(k+1) \times n$ matrix given by

$$D = \begin{bmatrix} d_{01} & d_{02} & \ldots & d_{0n} \\ d_{11} & d_{12} & \ldots & d_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{k1} & d_{k2} & \ldots & d_{kn} \end{bmatrix}$$

Consider now the estimator suggested by the least square solution, that is,

$$\hat{\beta} = (X^T X)^{-1} X^T Y. \tag{3.15}$$

Note the difference from the least square solution $\tilde{\beta}$: Here the $\hat{\beta}$ is an *estimator*, defined as a random variable, with the source of randomness being Y. Given X and $Y = y$, the *estimate* produced by $\hat{\beta}$ is equal to the least squares solution, $\tilde{\beta} = (X^T X)^{-1} X^T y$.[2]

Exercise 3.6. Show that $\hat{\beta} = (X^T X)^{-1} X^T Y$ is a linear estimator. Identify D.

Exercise 3.7. Show that $\hat{\beta} = (X^T X)^{-1} X^T Y$ is an unbiased estimator for β.

It is possible to extend the definition of a best linear unbiased estimator for multi-dimensional parameters. When we have an estimator of a parameter vector, its second moment is characterized not merely as a scalar variance but as a covariance matrix. One way to compare estimators in terms of their covariance matrices is to check whether the difference between the covariance matrices is positive (or negative) (semi)-definite. (In the scalar case, the difference being positive reduces to having a bigger variance.) The definition of the best linear unbiased estimator for a parameter vector is indeed based on that comparison.

Definition 3.8 (BLUE (generalised))**.** An estimator $\hat{\beta}$ with covariance $\mathrm{Cov}(\hat{\beta})$ is said to be a best linear unbiased estimator if for any other linear unbiased estimator $\hat{\beta}'$ with $\mathrm{Cov}(\hat{\beta}')$, the difference $\mathrm{Cov}(\hat{\beta}') - \mathrm{Cov}(\hat{\beta})$ is positive semidefinite, which is also shown by $\mathrm{Cov}(\hat{\beta}') \succeq \mathrm{Cov}(\hat{\beta})$.

We are finally ready to set the *Gauss-Markov theorem*, which states that the least-squares estimator has the lowest variance within the linear unbiased estimators.

Theorem 3.4 (Gauss-Markov theorem)**.** *Suppose we have*

$$Y = X\beta + e$$

where X is an observable full rank matrix having less columns than its rows, and e is a vector of uncorrelated random noise terms with $E(e_i) = 0$, $V(e_i) = \sigma^2$ and $\mathrm{Cov}(e_i, e_j) = 0$ for $i \neq j$. Then, the estimator $\hat{\beta} = (X^T X)^{-1} X^T Y$ that is suggested by the least squares solution, or shortly the least squares estimator, is the best linear unbiased estimator.

Reduction to simple linear regression model: The simple linear regression model is recovered by

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \quad Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \beta = \begin{bmatrix} a \\ b \end{bmatrix}, \quad e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} \tag{3.16}$$

and check that, given $Y = y$, the RSS in (3.5) for the simple model is indeed recovered by $(y - X\beta)^T (y - X\beta)$.

[2] While an estimator is defined a random variable, a realized value of an estimator is called an estimate.

Exercise 3.8. Consider the simple linear regression model.

(a) Show that, we have

$$X^T X = \begin{bmatrix} n & n\bar{x} \\ n\bar{x} & S_{xx} + n\bar{x}^2 \end{bmatrix}, \quad X^T y = \begin{bmatrix} n\bar{y} \\ S_{xy} + n\bar{x}\bar{y} \end{bmatrix}, \tag{3.17}$$

(b) Show that

$$(X^T X)^{-1} X^T y = \begin{bmatrix} \bar{y} - \bar{x} S_{xy}/S_{xx} \\ S_{xy}/S_{xx} \end{bmatrix},$$

and verify that the least square solution for the simple linear regression model is indeed deduced from the least-squares solution shown for the general case.

Hint: The inverse of an invertible 2×2 matrix is given by

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

Corollary 3.1 (of Theorem 3.2)**.** Since the least square estimators $\hat{A}$ and $\hat{B}$ in (3.11) is a special case $\hat{\beta}$ where X and y are given in (3.16), they are the best linear unbiased estimators for a and b.

A.3 Models with distribution assumptions

In the previous section, Section A.2, we made some statistical assumptions about the data. Under those assumptions, we were able to show that the least-squares estimator has an optimality property among linear and unbiased estimators of a and b. However, since our assumptions were limited to the first and the second moments only and not the full probability distribution of the data, we were unable to derive confidence bounds or tests for the parameters a and b.

In this section, we present two statistical models that completely specify the conditional distribution of the response variables $Y_1, \ldots, Y_n$ given $x_1, \ldots, x_n$.

A.3.1 Conditional normal model

The conditional normal model is the most common simple linear regression model. The observed data are, as before, the pairs $(x_1, y_1), \ldots, (x_n, y_n)$. The predictor variables $x_1, \ldots, x_n$ are assumed fixed and known, and no statistical assumptions are made for them. We only model the conditional distribution of Y_i's given x_i's, and, by no surprise, we use a normal distribution for that. More specifically, Y_i's are independent and distributed according to

$$Y_i \sim \mathcal{N}(a + bx_i, \sigma^2), \quad i = 1, \ldots, n, \tag{3.18}$$

or, equivalently,

$$Y_i = a + bx_i + e_i, \quad e_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2), \quad i = 1, \ldots, n, \tag{3.19}$$

Note that e_i's were assumed uncorrelated in the previous section. Here we add to it the assumption that they have the normal distribution with zero mean and some variance, and the uncorrelatedness property has been automatically strengthened to independence. (Recall that if jointly normal variables are uncorrelated, they are independent.) Therefore, the only additional assumption that has led to the conditional normal model is the normality e_i's.

Now that the exact (conditional) distribution of $Y_1, \dots, Y_n$ is specified, we can write down the joint probability distribution function $Y_1, \dots, Y_n$ given the parameters a, b, and σ^2. It is given by

$$\begin{aligned} f(y_{1:n}; a, b, \sigma^2) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \prod_{i=1}^{n} \exp\left\{-\frac{1}{2\sigma^2}(y_i - a - bx_i)^2\right\} \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - a - bx_i)^2\right\}. \end{aligned}$$

We will use this joint distribution to find the estimators for a and b.

A.3.2 Bivariate normal model

In the conditional normal model, x_i's are assumed fixed and known and no statistical assumptions are made for them. There are cases where x_i's are observations of random variables. For example, recall the example where the experimenter wants to find out a relation between the wealth and happiness of a person. By the way of collecting the data for this purpose, we can talk about the wealth of the randomly selected people as random variables.

In the bivariate normal model $(x_1, y_1), \dots, (x_n, y_n)$ are observations of independent bivariate random vectors $(X_1, Y_1), \dots, (X_n, Y_n)$, where the bivariate vector has a bivariate normal distribution

$$(X_i, Y_i) \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\begin{bmatrix}\mu_x \\ \mu_y\end{bmatrix}, \begin{bmatrix}\sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2\end{bmatrix}\right), \quad i = 1, \dots, n. \tag{3.20}$$

where ρ is the correlation of X_i and Y_i with $|\rho| < 1$. (We avoid $\rho = 1$ to ensure the existence of the joint pdf). Observe that this probability distribution can be specified by the parameters μ_x, μ_y, ρ, σ_x^2 and σ_y^2, and that is why sometimes we write

$$(X_i, Y_i) \overset{\text{i.i.d.}}{\sim} \text{bivariate-normal}(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho), \quad i = 1, \dots, n.$$

Exercise 3.9. By working out the joint pdf of a multivariate normal distribution and the inverse of the covariance matrix given in (3.20), show that the joint probability density can be explicitly written as

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right) + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right]}$$

As far as simple linear regression is concerned the roles of X_i's and Y_i's are the same: X_i is the predictor variable and Y_i is the response variable. Therefore, we are interested in predicting Y from a given x value. For this, we need the conditional distribution of Y on x.

Exercise 3.10. Show that, when $(X, Y) \sim$ bivariate-normal$(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$, the conditional distribution is also normal with

$$Y|(X = x) \sim \mathcal{N}\left(\mu_y + \rho\frac{\sigma_y}{\sigma_x}(x - \mu_x),\ \sigma_y^2(1 - \rho^2)\right) \tag{3.21}$$

[Hint: To derive the conditional distribution, write down the joint distribution, treat x as fixed, and show that the whole expression, when considered a function of y is proportional to a probability density of a normal distribution for y. This is a trick often used in Bayesian statistics to derive conditional distributions.]

The conditional distribution in (3.21) yields the conditional expectation

$$E(Y|x) = \mu_y + \rho\frac{\sigma_y}{\sigma_x}(x - \mu_x) = \left(\mu_y - \rho\frac{\sigma_y}{\sigma_x}\mu_x\right) + \left(\rho\frac{\sigma_y}{\sigma_x}\right)x \tag{3.22}$$

The bivariate normal model *implies* that the population regression function of Y on x is linear in x. This is a distinction for the bivariate normal model: unlike the other models we have seen so far, we did not assume a linear regression function in x; the bivariate normal model automatically led to that. Also, the conditional variance is given by

$$V(Y|x) = \sigma_y^2(1 - \rho^2).$$

which is independent of x. Therefore, we are back in the setting of the conditional normal model after a reparametrisation: Letting $a = \mu_y - \rho\frac{\sigma_y}{\sigma_x}\mu_x$, $b = \rho\frac{\sigma_y}{\sigma_x}$, and $\sigma^2 = \sigma_y^2(1 - \rho^2)$ we have independent Y_i's with

$$E(Y|x) = a + bx, \quad V(Y|x) = \sigma^2.$$

so that we have a linear regression model.

For the bivariate model, the linear regression analysis is almost always carried out using the conditional distribution of Y_i's given x_i's. If this is the case, we end up in the same situation as the conditional normal model described in Section A.3.1. If we condition on x_i's, whether they are design variables or observed values of random variables does not matter.

Moreover, even when X is random with a different marginal distribution than normal, any type of linear regression analysis based on the conditional distribution of the responses given predictors can be carried out in the same way as long as the joint distribution of (X_i, Y_i)'s is such that Y_i's are independent and $Y|(X = x) \sim \mathcal{N}(a + bx, \sigma^2)$.

We have thus concluded that inference based on point estimators, intervals, or tests for a, b (and σ^2) is the same for both conditional normal and bivariate normal models.

A.3.3 Estimation and testing with normal errors

In the following, we discuss inference procedures under the conditional normal model defined by (3.18) or (3.19).

We begin with estimation of the model parameters a, b, and σ^2 via maximum likelihood. The log-likelihood function of (a, b, σ^2), given the data $x_{1:n}, y_{1:n}$, is given by

$$\ell((a, b, \sigma^2); x_{1:n}, y_{1:n}) = -\frac{n}{2}[\log(2\pi) + \log \sigma^2] - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - a - bx_i)^2$$

For any value of σ^2, maximising the log-likelihood with respect to a and b is equivalent to minimising the sum of squares

$$\sum_{i=1}^{n}(y_i - a - bx_i)^2$$

which yields the least square solution (once again!)

$$\hat{A} = \bar{Y} - \frac{S_{xY}}{S_{xx}}\bar{x}, \quad \hat{B} = \frac{S_{xY}}{S_{xx}}.$$

Hence, the least square estimators (as random variables) $\hat{A}$ and $\hat{B}$ in (3.11) are also the maximum likelihood estimators for a and b. Substituting those in the log-likelihood and maximizing with respect to σ^2, one gets the maximum likelihood estimator for σ^2.

Exercise 3.11. Doing as suggested in the previous sentence, show that the maximum likelihood estimator for σ^2 is given by

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{A} - \hat{B}x_i)^2.$$

Define the *residual from the regression* $\hat{e}_i = Y_i - \hat{A} - \hat{B}x_i$, so that $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{e}_i^2$.

Exercise 3.12. Show that $E(\hat{e}_i) = 0$.

The estimator $\hat{\sigma}^2$ is a biased estimator of σ^2. Specifically, as we will prove later, we have

$$E(\hat{\sigma}^2) = \frac{n-2}{n}\sigma^2.$$

Alternatively, we propose the unbiased estimator

$$S_e^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{A} - \hat{B}x_i)^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{e}_i^2.$$

In the following, we will show several properties regarding $\hat{A}$, $\hat{B}$, and S_e^2. The most remarkable ones are gathered in the following theorem:

Theorem 3.5. *The estimators $\hat{A}$, $\hat{B}$, and S_e^2 are distributed with*

$$\hat{A} \sim \mathcal{N}\left(a, \frac{\sigma^2}{S_{xx}n}\sum_{i=1}^{n} x_i^2\right), \quad \hat{B} \sim \mathcal{N}\left(b, \frac{\sigma^2}{S_{xx}}\right), \quad \frac{(n-2)S_e^2}{\sigma^2} \sim \chi^2_{n-2}. \tag{3.23}$$

Moreover, we have the following relations between $\bar{A}$, $\bar{B}$, S_e^2

- $\mathrm{Cov}(\hat{A}, \hat{B}) = -\frac{\bar{x}}{S_{xx}}\sigma^2$,
- *$\hat{A}$ and S_e^2 are independent,*
- *$\hat{B}$ and S_e^2 are independent.*

We will prove those claims one by one. While proving, and in many other places, some manipulations of quantities S_{xx}, S_{xy} and S_{yy} as below prove useful.

Exercise 3.13. Show that, we have

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})x_i, \quad S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})y_i, \quad S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})y_i = \sum_{i=1}^{n} x_i(y_i - \bar{y}).$$

Sampling distribution of $\hat{B}$: First, we show that $\hat{B}$ is a linear combination of Y_i's. Indeed, using Exercise 3.13, we have

$$\hat{B} = \frac{S_{xY}}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{S_{xx}} = \sum_{i=1}^{n}\frac{x_i - \bar{x}}{S_{xx}}Y_i.$$

Since Y_i's have a normal distribution, so is $\hat{B}$. The mean and the variance of $\hat{B}$ can be derived as

$$\begin{aligned}
E(\hat{B}) &= \sum_{i=1}^{n}\frac{x_i - \bar{x}}{S_{xx}}E(Y_i) \\
&= \sum_{i=1}^{n}\frac{x_i - \bar{x}}{S_{xx}}[a + bx_i] \\
&= \sum_{i=1}^{n}\frac{a(x_i - \bar{x})}{S_{xx}} + b\sum_{i=1}^{n} x_i\frac{x_i - \bar{x}}{S_{xx}} \\
&= \sum_{i=1}^{n}\frac{a(x_i - \bar{x})}{S_{xx}} + b\frac{1}{S_{xx}}\sum_{i=1}^{n} x_i(x_i - \bar{x}) \\
&= 0 + \frac{1}{S_{xx}}bS_{xx} = b
\end{aligned}$$

where the last line follows from the result for S_{xx} in Exercise 3.13. To derive the variance of $\hat{B}$, we make use of the independence of Y_i's and write the variance of the linear combination as a sum.

$$\begin{aligned}
\text{Var}(\hat{B}) &= \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{S_{xx}} \right)^2 \text{Var}(Y_i) \\
&= \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{S_{xx}^2} \sigma^2 \\
&= \frac{S_{xx}}{S_{xx}^2} \sigma^2 \\
&= \frac{\sigma^2}{S_{xx}}.
\end{aligned}$$

Sampling distribution of $\hat{A}$: Just like $\hat{B}$, the estimator $\hat{A}$ can also be written as a linear combination of Y_i's,

$$\hat{A} = \bar{Y} - \bar{x} \frac{S_{xY}}{S_{xx}} = \sum_{i=1}^{n} \left[\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{S_{xx}} \right] Y_i.$$

Therefore, the expectation of $\hat{A}$ is given by

$$\begin{aligned}
E(\hat{A}) &= \sum_{i=1}^{n} \left[\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{S_{xx}} \right] E(Y_i) \\
&= \frac{1}{n} \sum_{i=1}^{n} E(Y_i) - \sum_{i=1}^{n} \frac{\bar{x}(x_i - \bar{x})}{S_{xx}} E(Y_i) \\
&= \frac{1}{n} \sum_{i=1}^{n} (a + b x_i) - \bar{x} \sum_{i=1}^{n} \frac{x_i - \bar{x}}{S_{xx}} E(Y_i) \\
&= a + b\bar{x} - \bar{x} b = a.
\end{aligned}$$

where the last line follows from the fact that the second sum was shown to be b. Again, to derive the variance of $\hat{A}$, we make use of the independence of Y_i's and write the variance

of the linear combination as a sum.

$$\begin{aligned}
\mathrm{Var}(\hat{A}) &= \sum_{i=1}^{n} \left[\frac{1}{n} - \bar{x}\left(\frac{x_i - \bar{x}}{S_{xx}}\right)\right]^2 \mathrm{Var}(Y_i) \\
&= \sum_{i=1}^{n} \left[\frac{1}{n^2} + \bar{x}^2\left(\frac{x_i - \bar{x}}{S_{xx}}\right)^2 - \frac{2}{n}\bar{x}\left(\frac{x_i - \bar{x}}{S_{xx}}\right)\right]\sigma^2 \\
&= \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}^2}\sum_{i=1}^{n}(x_i - \bar{x})^2 - \frac{2}{n}\frac{\bar{x}}{S_{xx}}\sum_{i=1}^{n}(x_i - \bar{x})\right] \\
&= \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right] \\
&= \sigma^2\left[\frac{S_{xx} + n\bar{x}^2}{nS_{xx}}\right] \\
&= \sigma^2\frac{\sum_{i=1}^{n} x_i^2}{nS_{xx}}.
\end{aligned}$$

Covariance between A and B: Although not entirely necessary here, we define covariance between any two random vectors (of possibly different dimensions) below.

Definition 3.9 (Covariance between two vectors)**.** For random vectors $X \in \mathbb{R}^m$ and $Y \in \mathbb{R}^n$, covariance between X and Y is a $m \times n$ matrix defined as

$$\mathrm{Cov}(X, Y) = E(XY^T) - E(X)E(Y)^T.$$

That is, the (i, j)'th element of $\mathrm{Cov}(X, Y)$ is $\mathrm{Cov}(X_i, Y_j)$.

When we want to find the covariance between two variables (or vectors), the following lemmas are useful if those variables are linear combinations of certain other variables whose covariance can be found more easily.

Lemma 3.2. For random vectors $X, Y \in \mathbb{R}^m$ and $Z \in \mathbb{R}^n$, we have

$$\mathrm{Cov}(X + Y, Z) = \mathrm{Cov}(X, Z) + \mathrm{Cov}(Y, Z)$$

Lemma 3.3. For random vectors $X \in \mathbb{R}^m$ and $Y \in \mathbb{R}^n$, and matrices $P \in \mathbb{R}^{k\times m}$ and $Q \in \mathbb{R}^{l\times n}$, we have

$$\mathrm{Cov}(PX, QY) = P\mathrm{Cov}(X, Y)Q^T.$$

Using Lemma 3.3, we can show that

$$\begin{aligned}
\mathrm{Cov}(\hat{A}, \hat{B}) &= \mathrm{Cov}\left(\sum_{i=1}^{n}\left[\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{xx}}\right]Y_i\,,\, \sum_{j=1}^{n}\frac{(x_j - \bar{x})}{S_{xx}}Y_j\right) \\
&= \sum_{i=1}^{n}\left[\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{xx}}\right]\frac{(x_i - \bar{x})}{S_{xx}}\sigma^2 \\
&= -\frac{\bar{x}}{S_{xx}}\,\sigma^2
\end{aligned}$$

Combining the knowledge on the first and the second moments, the distribution of the vector $\begin{bmatrix}\hat{A} & \hat{B}\end{bmatrix}^T$ is fully identified as

$$\begin{bmatrix}\hat{A}\\ \hat{B}\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}a\\ b\end{bmatrix}, \sigma^2 \begin{bmatrix}\frac{\sum_{i=1}^n x_i^2}{nS_{xx}} & -\frac{\bar{x}}{S_{xx}}\\ -\frac{\bar{x}}{S_{xx}} & \frac{1}{S_{xx}}\end{bmatrix}\right).$$

Independence between $\hat{A}$ and S_e^2: For independence between $\hat{A}$ and S_e^2, we first need to show that $\hat{A}$ and each $\hat{e}_i$ are independent. Since $\hat{e}_i = Y_i - \hat{A} - \hat{B}x_i$ is a linear combination of Y_i's, $\hat{A}$ and $\hat{e}_i$ are jointly normal. Therefore, it is sufficient to show that $\hat{A}$ and $\hat{e}_i$ are uncorrelated. We check that

$$\begin{aligned}
\mathrm{Cov}(Y_i - \hat{A} - \hat{B}x_i\,,\, \hat{A}) &= \mathrm{Cov}(Y_i, \hat{A}) - \mathrm{Var}(\hat{A}) - x_i\mathrm{Cov}(\hat{A}, \hat{B})\\
&= \left[\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{xx}}\right]\sigma^2 - \frac{\sum_j x_j^2}{nS_{xx}}\sigma^2 + x_i\frac{\bar{x}\sigma^2}{S_{xx}}\\
&= \frac{\sigma^2}{nS_{xx}}\left(S_{xx} - n(x_i - \bar{x})\bar{x} - \sum_{j=1}^n x_j^2 + nx_i\bar{x}\right)\\
&= \frac{\sigma^2}{nS_{xx}}\left(S_{xx} - nx_i\bar{x} + n\bar{x}^2 - \sum_{j=1}^n x_j^2 + nx_i\bar{x}\right)\\
&= \frac{\sigma^2}{nS_{xx}}\left(S_{xx} + n\bar{x}^2 - \sum_{j=1}^n x_j^2\right)\\
&= \frac{\sigma^2}{nS_{xx}}\left(S_{xx} - S_{xx}\right) = 0.
\end{aligned}$$

The second line in the derivation above is due to the fact that $\hat{A} = \sum_{i=1}^n d_i Y_i$ is a linear combination of Y_i's, with $d_i = 1/n - (x_i - \bar{x})\bar{x}/S_{xx}$, and as a result $\mathrm{Cov}(Y_i, \hat{A}) = \sigma^2 d_i$. Since $\hat{A}$ is independent from each $\hat{e}_i$, it is also independent from S_e^2, a function of $\hat{e}_i$'s.

Independence between $\hat{B}$ and S_e^2: For independence between $\hat{B}$ and S_e^2, we follow a similar path: First need to show that $\hat{B}$ and each $\hat{e}_i$ are independent. Since $\hat{e}_i = Y_i - \hat{A} - \hat{B}x_i$ is a linear combination of Y_i's, $\hat{B}$ and $\hat{e}_i$ are jointly normal. Therefore, it is sufficient to show that $\hat{B}$ and $\hat{e}_i$ are uncorrelated. We check that

$$\begin{aligned}
\mathrm{Cov}(Y_i - \hat{A} - \hat{B}x_i,\, \hat{B}) &= \mathrm{Cov}(Y_i, \hat{B}) - \mathrm{Cov}(\hat{A}, \hat{B}) - x_i\mathrm{Var}(\hat{B})\\
&= \frac{x_i - \bar{x}}{S_{xx}}\sigma^2 + \frac{\bar{x}\sigma^2}{S_{xx}} - x_i\frac{\sigma^2}{S_{xx}} = 0.
\end{aligned}$$

Again, the second line in the derivation above is due to the fact that $\hat{B} = \sum_{i=1}^n c_i Y_i$ is a linear combination of Y_i's, with $c_i = (x_i - \bar{x})/S_{xx}$, and as a result $\mathrm{Cov}(Y_i, \hat{B}) = \sigma^2 c_i$. Since $\hat{B}$ is independent from each $\hat{e}_i$, it is also independent from S_e^2, a function of $\hat{e}_i$'s.

A.3.4 Inference for the parameters

We will present confidence intervals and tests for a, b, and σ^2.

Inference for b: Between a and b, the slope parameter b is often more important in linear regression. For example, the situation $b = 0$ corresponds to no linear regression at all. Therefore, to test the existence of linear regression, the null hypothesis $H_0 : b = 0$ can be considered, with the alternative $H_1 : b \neq 0$.

Observe that the distribution of $\hat{B}$ involves the unknown parameter σ^2. This can be eliminated by factoring in the estimator of σ^2.

Exercise 3.14. Using the results in Theorem 3.5, show that the test statistic

$$\frac{\hat{B} - b}{S_e/\sqrt{S_{xx}}} \sim t_{n-2}. \tag{3.24}$$

The test statistic given in (3.24) can be used to obtain confidence intervals or test a value for b.

Exercise 3.15. Show that, the interval

$$\left(\hat{B} - \frac{S_e}{\sqrt{S_{xx}}} t_{\alpha/2,n-2}, \quad \hat{B} + \frac{S_e}{\sqrt{S_{xx}}} t_{\alpha/2,n-2}\right)$$

is a $100(1-\alpha)\%$ confidence interval for b.

Exercise 3.16. Show that, the null hypothesis $H_0 : b = b_0$ can be tested with a significance level α if the following critical value is used to reject H_0.

$$C = \left\{ \left| \frac{\hat{b} - b_0}{S_e/\sqrt{S_{xx}}} \right| > t_{\alpha/2,n-2} \right\}$$

Testing the existence of a linear regression: The null hypothesis $H_0 : b = 0$ vs $H_1 : b \neq 0$ is worth more investigation, since testing it is equivalent to testing for the existence of linear regression on the predictor variable. Related to testing this null hypothesis, we introduce some important concepts.

Firstly, a partitioning the sum of squares as in the ANOVA setting is also available for the simple linear regression model. Let $\hat{Y}_i = \hat{A} + \hat{B}x_i$ be the 'fitted' value of Y_i based on the estimators of a and b.

Exercise 3.17. Show that in the simple normal linear regression model, the total sum of squares can be partitioned as

$$\underbrace{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}_{\text{total sum of squares}} = \underbrace{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}_{\text{regression sum of squares}} + \underbrace{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}_{\text{residual sum of squares (RSS)}} \tag{3.25}$$

Furthermore, using that e_i's are normal, $\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$ and $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ are independent.

Next, we will see how the partitioning is particularly relevant to the null hypothesis $H_0 : b = 0$ if

$$\left| \frac{\hat{B}}{S_e/\sqrt{S_{xx}}} \right| > t_{\alpha/2,n-2},$$

or, equivalently,

$$\frac{\hat{B}^2}{S_e^2/S_{xx}} > F_{\alpha,1,n-2}. \tag{3.26}$$

Using $\hat{B} = S_{xY}/S_{xx}$ and $\text{RSS} = S_e^2(n-2)$, we have

$$\frac{\hat{B}^2}{S_e^2/S_{xx}} = \frac{S_{xY}^2/S_{xx}}{\text{RSS}/(n-2)} \tag{3.27}$$

The numerator can also be shown to be equal to the regression sum of squares, which was defined as the first term on the right hand side of the partitioning equation in (3.25). More explicitly, it can be shown that

Exercise 3.18. Show that the numerator in (3.29) is equal to the regression sum of squares,

$$\frac{S_{xY}^2}{S_{xx}} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 \tag{3.28}$$

where the right hand side was defined as the regression sum of squares in (3.25).

Using the result in Exercise 3.18, we can write the test statistic in (3.26)

$$\frac{\hat{B}^2}{S_e^2/S_{xx}} = \frac{\text{regression sum of squares}}{\text{residual sum of squares}/(n-2)} \tag{3.29}$$

Equation (3.29) already suggests that the regression sum of squares has a χ_1^2 under $H_0 : b = 0$. This can be shown easily by considering the partitioning in (3.25).

Exercise 3.19. Show that, when $b = 0$, we have the following distributions of the total sum of squares

$$\frac{1}{\sigma^2} \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \sim \chi_{n-1}^2, \quad \frac{1}{\sigma^2} \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 \sim \chi_1^2.$$

[Hint: Use independence of the terms and the distribution of RSS/σ^2]

An ANOVA table is available in Table 3.1 for the simple normal linear regression model, with a focus on testing $b = 0$.

Table 3.1: ANOVA table for simple normal linear regression model

source of variation	d.o.f	Sum of squares	Mean square	F statistic
Regression	1	Reg. SS $= S_{xy}^2/S_{xx}$	MS(Reg) $=$ Reg. SS	$F = \frac{MS(Reg)}{MS(Resid)}$
Residual	$n-2$	RSS $= \sum_{i=1}^n \hat{e}_i^2$	$MS(Resid) = \frac{RSS}{n-2}$	
Total	$n-1$	TSS $= \sum_{i=1}^n (y_i - \bar{y})^2$		

Coefficient of determination: Another statistic that quantifies how well the fitted line describes the data is called the coefficient of determination, and it is given by

$$r^2 = \frac{\text{regression sum of squares}}{\text{total sum of squares}} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{S_{xy}^2}{S_{xx} S_{yy}}$$

From the partitioning equation, we have $0 \leq r^2 \leq 1$. A better fit, which means that $\hat{y}_i$'s are close to y_i's, yields larger values of r^2. We could also define R^2, the random variable version of r^2, through the random variables Y_i's. Under $H_0 : b = 0$, R^2 is expected to take small values.

Exercise 3.20. Show that the test statistic in (3.29) is equal to $(n-2)R^2/(1-R^2)$.

Inference for a: Although the intercept parameter a is less important than the slope parameter b, confidence intervals and tests can be obtained for a as well, by following similar steps as for b. For example, $a = 0$ can be tested to understand whether the linear line crosses the origin. Observe that the distribution of $\hat{A}$ also involves the unknown parameter σ^2, which can be eliminated by factoring in the estimator of σ^2.

Exercise 3.21. Using the results in Theorem 3.5, show that the test statistic

$$\frac{\hat{A} - a}{S_e \sqrt{\frac{\sum_{i=1}^n x_i^2}{S_{xx} n}}} \sim t_{n-2}. \tag{3.30}$$

The test statistic given in (3.30) can be used to obtain confidence intervals or test a value for b.

Exercise 3.22. Show that, the interval

$$\left(\hat{A} - S_e \sqrt{\frac{\sum_{i=1}^n x_i^2}{S_{xx} n}} t_{\alpha/2, n-2}, \quad \hat{A} + S_e \sqrt{\frac{\sum_{i=1}^n x_i^2}{S_{xx} n}} t_{\alpha/2, n-2} \right)$$

is a $100(1-\alpha)\%$ confidence interval for a.

Exercise 3.23. Show that, the null hypothesis $H_0 : a = a_0$ can be tested with a significance level α if the following critical value is used to reject H_0.

$$C = \left\{ \left| \frac{\hat{a} - a_0}{S_e \sqrt{\frac{\sum_{i=1}^n x_i^2}{S_{xx} n}}} \right| > t_{\alpha/2, n-2} \right\}$$

Simultaneous inference for a and b: A simultaneous confidence interval for a and b can be obtained, for example by using Bonferroni correction. Namely, obtain a $100(1-\alpha/2)\%$ confidence interval for a and b separately, and, denoting them as CI_a, CI_b, consider their intersection

$$\text{CI}_{a,b} = \{(a', b') : a' \in \text{CI}_a \text{ and } b' \in \text{CI}_b\}$$

will give a simultaneous confidence interval for a, b with a confidence level that is at least $1-\alpha$. Note also that the $\text{CI}_{a,b}$ can be used to devise a critical region as well. For example, for the null hypothesis $H_0 : a = a_0, b = b_0$, the critical region

$$C = \{(a_0, b_0) \notin \text{CI}_{a,b}\}$$

gives a hypothesis test of size at most α.

However, we can be smarter than joining two tests (or confidence intervals) using standard methods, which can be loose. Recall that the vector $\begin{bmatrix}\hat{A} & \hat{B}\end{bmatrix}^T$ has a bivariate normal distribution. Also, note that the vector $\begin{bmatrix}\hat{A} & \hat{B}\end{bmatrix}^T$ is independent from S_e^2, since each element of the vector is independent from S_e^2. Those observations lead to a test statistic for a and b jointly, thereby enabling, for any α, confidence intervals with confidence levels exactly equal to $1-\alpha$ or tests with significance exactly equal to α. Like the sampling distribution of S_e^2, we also skip presenting such tests or confidence intervals for the time being. We will see those techniques in the more general setting of the multiple normal linear regression model with $k \geq 1$ predictors; the application to $k = 1$ should be clear from then on.

Inference for σ^2: For inference on σ^2, S_e^2 can be used, as it does not depend on the other unknown parameters. Recall from Theorem 3.5 that

$$\frac{S_e^2(n-2)}{\sigma^2} \sim \chi^2_{n-2}. \tag{3.31}$$

The test statistic given in (3.31) can be used to obtain confidence intervals or tests for a value for σ^2. We will not state them here to avoid repetition; see Exercise 1.17 for such tests (and confidence intervals), with the degrees of freedom changed from $n-1$ to $n-2$.

A.3.5 Estimation and prediction at a new predictor

Recall that one of the main aims of linear regression is the prediction of Y_0 given a new predictor variable x_0. There are two ways of expressing our prediction. The first is in terms of confidence intervals for the population mean at x_0, and the other is the prediction interval for the observation Y_0 itself.

Confidence interval for the population mean at a specified x_0: When a new point x_0 is given, consider $\hat{A} + \hat{B}x_0$ as an estimator of the population mean at x_0, that is $E(Y|x_0) = a + bx_0$.

Exercise 3.24. Show that the estimator $\hat{A} + \hat{B}x_0$ for $E(Y|x_0) = a + bx_0$ is unbiased with a normal distribution

$$\hat{A} + \hat{B}x_0 \sim \mathcal{N}\left(a + bx_0, \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)\right) \tag{3.32}$$

Moreover, show that $\hat{A} + \hat{B}x_0$ and S_e^2 are independent.

Once again, combining (3.32) with the residual sum of squares estimator S_e^2 of the variance gives us a test statistic for $a + bx$. Namely,

$$\frac{\hat{A} + \hat{B}x_0 - a - bx_0}{S_e\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \sim t_{n-2}. \tag{3.33}$$

Exercise 3.25. Derive a two-sided $100(1 - \alpha)\%$-CI for $a + bx_0$.

Prediction interval for the observation at a specified x_0: Instead of estimating the population function at x_0, let us focus, instead, on *prediction* of an, as yet, unobserved random variable. A typical scenario for prediction occurs when x_0 is known but Y_0 has not been observed yet. For example, we have data for high school and university GPAs of a number of the students in a school. A new high school graduate with a high school GPA of x_0 is considered for admission to the school. We want to predict this student's performance in the university, Y_0, which is not observed yet.

Assume that the yet unobservable variable Y_0 is paired with the predictor x_0, and its distribution is expressed as

$$Y_0 = a + bx_0 + e_0, \quad e_0 \sim \mathcal{N}(0, \sigma^2).$$

where e_0 is independent from $e_1, \ldots, e_n$, hence Y_0 is independent from the observations so far, $Y_1, \ldots, Y_n$. Hence Y_0 is independent from $\hat{A}$, $\hat{B}$, and S_e^2. Just like $a + bx_0$, the variable Y_0 can be predicted pointwise by $\hat{A} + \hat{B}x_0$ also. The 'error' variable $Y_0 - \hat{A} - \hat{B}x_0$, being a linear combination of independent normally distributed random variable, has a normal distribution.

$$E(Y_0 - \hat{A} - \hat{B}x_0) = a + bx_0 - a - bx_0 = 0.$$

$$\mathrm{Var}(Y_0 - \hat{A} - \hat{B}x_0) = \mathrm{Var}(Y_0) + \mathrm{Var}(\hat{A} - \hat{B}x_0) = \sigma^2 + \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)$$

Using the independence between $Y_0 - \hat{A} - \hat{B}x_0$ and S_e^2, we have

$$\frac{Y_0 - (\hat{A} + \hat{B}x_0)}{S_e\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \sim t_{n-2}. \tag{3.34}$$

Exercise 3.26. Using (3.34), derive a prediction interval for Y_0.

A.3.6 Simultaneous estimation and confidence intervals

Let us return to estimating the population regression function $E(Y|x)$ at a new point $x = x_0$. So far, we have looked at a single x_0. A natural extension is to predict the population regression function at multiple values for x. One way to build simultaneous confidence intervals is to use Bonferroni correction. With points $x_{01}, \ldots, x_{0m}$, we can state that

$$P\left[a + bx_i \in \left(\hat{A} + \hat{B}x_{0i} \pm t_{\alpha/2m,n-2} S_e \sqrt{\frac{1}{n} + \frac{(x_{0i} - \bar{x})^2}{S_{xx}}}\right), \forall i = 1, \ldots, m\right] \geq 1 - \alpha.$$

When m is too large, the confidence intervals get unreasonably large. Besides, ideally, we would like to have a confidence interval for all x's and therefore draw a 'confidence band' around the fitted line. As a remedy, Scheffe derived confidence intervals for all x points that simultaneously hold with probability $1 - \alpha$.

Theorem 3.6 (Scheffe's theorem for all values of x)**.** *Suppose we observe $(x_1, Y_1), \ldots, (x_n, Y_n)$, and construct $\hat{A}$, $\hat{B}$, and S_e^2 based on those observations. Then, we have*

$$P\left[a + bx \in \left(\hat{A} + \hat{B}x \pm \sqrt{2f_{\alpha,2,n-2}} S_e \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}}\right), \forall x \in \mathbb{R}\right] = 1 - \alpha.$$

Exercise 3.27. Houck (1970) studied the bismuth I–II transition pressure as a function of temperature. The data are listed in the following table.

Pressure (bar)	Temperature
25366	20.8
25356	20.9
25336	21.0
25256	21.9
25267	22.1
25306	22.1
25237	22.4
25267	22.5
25138	24.8
25148	24.8
25143	25.0
24731	34.0
24751	34.0
24771	34.1
24424	42.7
24444	42.7
24419	42.7
24417	49.9
24102	50.1
24092	50.1
25202	22.5
25157	23.1
25157	23.0

A simple normal linear regression model $y = a + bx$ is to be fitted to the data, with y being the pressure and x being the temperature.

(a) Make a scatter plot of the data, where each pair is shown on the $x - y$ plane. Limit the x and y axes appropriately to improve visibility.

(b) Calculate $\bar{x}$, $\bar{y}$, S_{xx}, S_{xy}, and S_{yy}, and R^2, the coefficient of determination, for these data.

(c) Find the MLE estimates for a, b, and σ^2.

(d) Find a $100(1-\alpha)\%$ confidence interval for b, with $\alpha = 0.1$.

(e) Test the null hypothesis $H_0 : b = -40$.

(f) In this part you will perform prediction: Calculate Sheffe's simultaneous confidence intervals for the population mean at all the integer values for x between 20 and 50. (You can increase the resolution if you wish.) Let the confidence interval at a specified x value be $(L(x), U(x))$. Plot $L(x)$ vs x and $U(x)$ vs x on the same plot, preferably together with the data pairs. What is supposed to appear in your figure is the Scheffe band for the population mean between 20 and 50.

B Multiple normal linear regression

We have already started discussing the multiple linear regression model in Section A.2 where we introduced linear models and the best linear unbiased estimators for them. This section revisits the linear model and presents the multiple normal linear regression model as a special case. The simple linear regression model was extensively discussed in the earlier section. We will see that many results for the simple linear regression model correspond to a more general result for the multiple linear regression model. Provided that one is confident in matrix-vector operations, dealing with the multiple linear regression model can be even easier and complementary in terms of clarifying the specific results in the simple linear regression model.

Let's remind ourselves of the linear model

$$Y = X\beta + e \tag{3.35}$$

where $Y = \begin{bmatrix} Y_1 & \dots & Y_n \end{bmatrix}^T$, X is a $n \times (k+1)$ design matrix with rank $k+1$, $\beta = \begin{bmatrix} \beta_0 & \dots & \beta_k \end{bmatrix}^T$ is the parameter vector and $e = \begin{bmatrix} e_1 & \dots & e_n \end{bmatrix}^T$ is the vector uncorrelated noise terms with zero mean and common variance σ^2. In parallel to the normal linear regression model, let us assume further that

$$e \sim \mathcal{N}(0, \sigma^2 I_n), \tag{3.36}$$

that is, each e_i is independent and has $\mathcal{N}(0, \sigma^2)$.

Exercise 3.28. Show that $\frac{Y - X\beta}{\sigma} \sim \mathcal{N}(0, I_n)$.

It is worth repeating that the multiple normal linear regression can be obtained with a general $k \geq 1$ and X being

$$X = \begin{bmatrix} 1 & x_{1,1} & \dots & x_{1,k} \\ 1 & x_{2,1} & \dots & x_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \dots & x_{n,k} \end{bmatrix}. \tag{3.37}$$

B.1 Maximum likelihood estimation and properties

Recall that the best linear unbiased estimator for β is given by $\hat{\beta} = (X^T X)^{-1} X^T Y$. One can show that this is also the maximum likelihood estimator for β. Given X and $Y = y$, the log-likelihood function can be written as

$$\ell(\beta, \sigma^2; X, y) = -\frac{n}{2}[\log 2\pi + \log \sigma^2] - \frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta).$$

Exercise 3.29. Show that the maximum likelihood estimators for β and σ^2 are given as

$$\hat{\beta} = (X^T X)^{-1} X^T Y, \quad \hat{\sigma}^2 = \frac{1}{n}(Y - X\hat{\beta})^T(Y - X\hat{\beta})$$

With this estimator of β, we can consider the predictions of Y_i's (not in the real sense, since Y_i's are already given), which are given by

$$\hat{Y} = X\hat{\beta} = X(X^T X)^{-1} X^T Y.$$

Sometimes, the matrix $X(X^T X)^{-1} X^T$ is referred to as the "hat matrix", since it produces $\hat{Y}$. In the subsequent discussion, we will refer to the hat matrix several times, so let us define it for convenience, as

$$H = X(X^T X)^{-1} X^T.$$

Define the vector of residuals

$$\hat{e} = Y - HY. \tag{3.38}$$

The residual sum of squares of those residuals can be written as

$$\text{RSS} = \sum_{i=1}^{n} \hat{e}_i^2 = \hat{e}^T \hat{e} = \|Y - \hat{Y}\|_2^2 = (Y - \hat{Y})^T(Y - \hat{Y}) \tag{3.39}$$

$$= (Y - HY)^T(Y - HY). \tag{3.40}$$

It turns out that $\hat{\sigma}^2$ is a biased estimator of σ^2, with expectation $E(\hat{\sigma}^2) = \frac{n-k-1}{n}\sigma^2$. Hence, similarly to the simple normal linear regression model, we propose

$$S_e^2 = \frac{1}{n-k-1}(Y - X\hat{\beta})^T(Y - X\hat{\beta}) = \frac{\text{RSS}}{n-k-1}.$$

Theorem 3.7. *The estimators $\hat{\beta}$ and S_e^2 have the following properties:*

- *The estimator $\hat{\beta}$ is an unbiased estimator of β with a multivariate normal distribution*

$$\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(X^TX)^{-1}) \tag{3.41}$$

- *The estimator S_e^2 is an unbiased estimator of σ^2 satisfying*

$$\frac{S_e^2(n-k-1)}{\sigma^2} \sim \chi^2_{n-k-1}.$$

- *$\hat{\beta}$ and S_e^2 are independent.*

We will prove each item of Theorem 3.7 in turn.

B.1.1 Distribution of $\hat{\beta}$

Unbiasedness of $\hat{\beta}$ was stated in Exercise 3.7 under looser assumptions (uncorrelated and zero mean noise, normality is not necessary). Moreover, being a linear combination of Y, $\hat{\beta}$ has a normal distribution and its covariance is given by

$$\begin{aligned}
\mathrm{Cov}(\hat{\beta}) &= \mathrm{Cov}((X^TX)^{-1}X^TY) \\
&= [(X^TX)^{-1}X^T]\sigma^2 I_n [(X^TX)^{-1}X^T]^T \\
&= \sigma^2(X^TX)^{-1}X^TX(X^TX)^{-1} \\
&= \sigma^2(X^TX)^{-1}.
\end{aligned}$$

Therefore,

$$\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(X^TX)^{-1}) \tag{3.42}$$

B.1.2 Distribution of S_e^2

Here, Cochran's theorem comes to help.

Theorem 3.8 (Cochran's theorem)**.** *Let $Z = \begin{bmatrix} Z_1 & \dots & Z_n \end{bmatrix}^T$ be a vector of i.i.d. random variables with a standard normal distribution, $Z_i \sim \mathcal{N}(0,1)$, and $Q_1, Q_2, \dots, Q_m$ be a collection of positive semidefinite matrices satisfying $\sum_{i=1}^m Q_i = I_n$. Further, suppose that $r_1 + \dots + r_m = n$, where r_i is the rank of Q_i. Finally, define the random variables of quadratic form*

$$V_i = Z^TQ_iZ, \quad i = 1, \dots, m$$

Then, all V_i's are independent and have chi-square distributions with

$$V_i \sim \chi^2_{r_i}, \quad i = 1, \dots, m.$$

As it is clear from the theorem, the result requires the existence of positive definite matrices $Q_1, \ldots, Q_m$ that sum to the identity matrix. While it can be difficult to find such matrices for general m, the task gets easier for $m = 2$ thanks to *idempotent* matrices.

Definition 3.10 (Idempotent matrices)**.** A square matrix P is idempotent if $PP = P$.

There are two properties of an idempotent matrix, which is relevant to finding positive definite matrices $m = 2$.

Proposition 3.1. *The following hold for idempotent matrices*

- *If P is idempotent and symmetric, it is positive semidefinite.*
- *If P is idempotent, $I_n - P$ is idempotent, too.*

Exercise 3.30. Prove Proposition 3.1.

Therefore, once we find an idempotent and symmetric matrix P, we can use Cochran's theorem with $Q_1 = P$ and $Q_2 = I_n - P$.

With reference to Cochran's theorem, take $Z = \frac{Y - X\beta}{\sigma}$ which was shown to have $\mathcal{N}(0, I_n)$ in Excercise 3.28. Further, consider $m = 2$ with $Q_1 = H$ and $Q_2 = I_n - H$. Obviously, by construction, $Q_1 + Q_2 = I_n$. In order to use Cochran's theorem with those Q_1, Q_2, we also have to make sure that each Q_i is both positive semidefinite. At this point, we show that Q_1 and Q_2 are idempotent. Indeed,

$$\begin{aligned} Q_1 Q_1 = HH &= X(X^TX)^{-1}X^TX(X^TX)^{-1}X^T \\ &= X(X^TX)^{-1}(X^TX)(X^TX)^{-1}X^T \\ &= X(X^TX)^{-1}X^T \\ &= H = Q_1, \end{aligned}$$

so Q_1 is idempotent. By Proposition 3.30, $Q_2 = I - H$ is also idempotent. Furthermore, since Q_1 and Q_2 are symmetric, they are positive semidefinite, by the same Proposition 3.30. Therefore, by Cochrans theorem we have the following result.

Corollary 3.2. For the linear model with normal errors, if X has rank $k+1$, so that the hat matrix $H = X(X^TX)^{-1}X^T$ is well defined and has rank $k+1$, we have

$$\frac{1}{\sigma^2}(Y - X\beta)^T H(Y - X\beta) \sim \chi^2_{k+1}, \quad \text{and} \quad \frac{1}{\sigma^2}(Y - X\beta)^T(I - H)(Y - X\beta) \sim \chi^2_{n-k-1}.$$

Furthermore, the quantities $(Y - X\beta)^T H(Y - X\beta)$ and $(Y - X\beta)^T(I - H)(Y - X\beta)$ are independent.

We have not said anything about the distribution of the RSS yet, but we are close to doing that. It turns out that the second quantity in the above corollary is the RSS in (3.40), scaled by $\frac{1}{\sigma^2}$. Expanding the quantity, we get

$$(Y - X\beta)^T(I - H)(Y - X\beta) = Y^T(I - H)Y - 2\beta^T X^T(I - H)Y + \beta^T X^T(I - H)X\beta. \tag{3.43}$$

At this point, we can show that the second and the third terms in (3.43) are equal to zero, by showing that the matrix $(I - H)X$, that is commonly found in those terms, is a zero matrix.

$$X^T(I - H) = X^T - X^T X(X^T X)^{-1}X^T = X^T - X^T = 0.$$

Therefore, we ended up with

$$(Y - X\beta)^T(I - H)(Y - X\beta) = Y^T(I - H)Y \tag{3.44}$$

But the expression in (3.44) is equal RSS at $\hat{\beta}$ given in (3.40). Check that

$$\begin{aligned}\hat{e}^T\hat{e} = (Y - HY)^T(Y - HY) &= Y^TY - Y^T(H^T + H)Y + Y^TH^THY \\ &= Y^TY - 2Y^THY + Y^THY \\ &= Y^TY - Y^THY \\ &= Y^T(I - H)Y\end{aligned}$$

where we used that facts that H is symmetric and idempotent. As a result, we have the following Corollary of Cochran's theorem for the residual sum of squares.

Corollary 3.3. With the residual vector given in (3.38), we have

$$\frac{S_e^2(n - k - 1)}{\sigma^2} = \frac{\text{RSS}}{\sigma^2} = \frac{\sum_{i=1}^n \hat{e}_i^2}{\sigma^2} \sim \chi^2_{n-k-1}.$$

B.1.3 Independence between $\hat{\beta}$ and S_e^2

Proving the independence claims in Theorem 3.5 is also easy in the general setting. All we need to show is that $\hat{\beta}$ and $\hat{e}$ are independent.

Theorem 3.9. *$\hat{\beta}$ and $\hat{e}$ are independent.*

Proof. First, note that both $\hat{\beta} = DY$, where $D = (X^TX)^{-1}X^T$, and $\hat{e} = Y - HY = (I - H)Y$ are linear combinations of Y, and they jointly have a normal distribution. Therefore, it suffices to show that $\hat{\beta}$ and $\hat{e}$ are uncorrelated, that is, their covariance is a zero matrix. For that, we write

$$\begin{aligned}\text{Cov}(\hat{\beta}, \hat{e}) &= \text{Cov}(DY, (I - H)Y) \\ &= D\text{Cov}(Y, Y)(I - H)^T \\ &= D\sigma^2 I_n(I - H) \\ &= \sigma^2 D(I - H) \\ &= \sigma^2 \left\{[(X^TX)^{-1}X^T] - [(X^TX)^{-1}X^TX(X^TX)^{-1}X^T]\right\} \\ &= \sigma^2 \left\{[(X^TX)^{-1}X^T] - [(X^TX)^{-1}X^T]\right\} \\ &= 0.\end{aligned}$$

Hence, we are done. □

Independence between $\hat{\beta}$ and S_e^2 is merely a direct consequence of Theorem 3.9.

Corollary 3.4. $\hat{\beta}$ and S_e^2 are independent.

Proof. Since S_e^2 is a function of $\hat{e}$, and $\hat{\beta}$ is independent from $\hat{e}$ by Theorem 3.9, $\hat{\beta}$ is also independent from S_e^2. □

We can also show that the 'predicted' (or fitted) values $\hat{Y}$ and $\hat{e}$ are also independent.

Theorem 3.10. *$\hat{Y}$ and $\hat{e}$ are independent.*

Proof. First, note that both $\hat{Y} = HY$ and $\hat{e} = Y - HY = (I - H)Y$ are linear combinations of Y, and they jointly have a normal distribution. Therefore, it suffices to show that $\hat{Y}$ and $\hat{e}$ are uncorrelated, that is, their covariance is a zero matrix. For that, we write

$$\begin{aligned}\mathrm{Cov}(\hat{Y}, \hat{e}) &= \mathrm{Cov}(HY, (I - H)Y) \\ &= H\mathrm{Cov}(Y, Y)(I - H)^T \\ &= H\sigma^2 I(I - H) \\ &= \sigma^2 H(I - H) = \sigma^2(H - H^2) = \sigma^2(H - H) = 0.\end{aligned}$$

Hence, we are done.

Note that an alternative proof based on Theorem 3.9 is available. $\hat{Y} = X\hat{\beta}$ is a function of $\hat{\beta}$, which is shown to be independent of $\hat{e}$ by Theorem 3.9, hence $\hat{Y}$ is also independent from $\hat{e}$. □

B.2 Relation to the simple linear model

I would like the reader to acknowledge (once again) the generality of the normal linear model, in particular, how it accommodates the multiple normal linear regression model with normally distributed error. Many results stated for the multiple normal linear regression model can be stated as a property of the general normal linear model given in (3.35) and (3.36).

The application of the above analysis to the simple normal linear regression model should be clear: The simple normal linear regression model is obtained by taking $k = 1$ (one predictor), and the variables as

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \quad Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \beta = \begin{bmatrix} a \\ b \end{bmatrix}, \quad e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}. \tag{3.45}$$

with $e \sim \mathcal{N}(0, \sigma^2 I_n)$. Clearly, the estimator vector is $\hat{\beta} = \begin{bmatrix} \hat{A} & \hat{B} \end{bmatrix}^T$. Further, the residuals $\hat{e}_i = Y_i - a - bx_i$ can be written in compact form as

$$\hat{e} = Y - HY$$

where $H = X(X^TX)^{-1}X^T$. This rewriting is to emphasise that the simple normal linear regression model can be cast as a linear normal model.

The normal linear model enables us to study both simple and multiple normal linear regression models in one common setting. For example, unbiasedness of $\hat{\beta}$ directly implies unbiasedness of $\hat{A}$ and $\hat{B}$ in the simple normal linear regression model, so we really did not need to prove unbiasedness for $\hat{A}$ and $\hat{B}$ separately. Moreover, the normal distribution of $\hat{\beta}$ stated in equation (3.42) summarises the results that we have shown separately for the simple normal linear regression model (with $k = 1$), namely the results that $\hat{A}$ and $\hat{B}$ are unbiased, normally distributed with covariances in Theorem 3.5. Furthermore, we have generalised the unbiased estimator for the variance to general $k \geq 1$ as $S_e^2 = \frac{1}{n-k-1}\sum_{i=1}\hat{e}_i^2$. We have proven the unbiasedness claim for S_e^2 in Theorem 3.5, even in a more general setting of normal linear models, and showed that $(n-k-1)S_e^2/\sigma^2 \sim \chi^2_{n-k-1}$.

For the simple normal linear regression model, we have $k = 1$ predictor variable, hence the estimator for the variance reduces to

$$S_e^2 = \frac{1}{n-2}\sum_{i=1}\hat{e}_i^2 = \frac{\text{RSS}}{n-2}.$$

with $S_e^2(n-2)/\sigma^2 \sim \chi^2_{n-2}$. The independence of S_e^2 from $\hat{A}$ and $\hat{B}$ separately is also a consequence of a more general result, namely the one that states the independence between $\hat{\beta}$ and S_e^2.

B.3 Tests for β:

B.3.1 Testing for linear regression (at all)

Arguably, the first thing we should test when we obtain the data X and Y should be the existence of a linear relationship whatsoever. That corresponds to the null hypothesis

$$H_0 : \beta_1 = \ldots = \beta_k = 0, \quad H_1 : \text{ at least one } \beta_i \text{ is non-zero}$$

To test this null hypothesis, we make use of an intermediate result, which is related to the partitioning of the total sum of squares. As the following Theorem 3.11 shows, the partitioning of the total sum of squares extends to the multiple linear regression case. We can show that even for the general linear model we have that partitioning provided that $\mathbf{1}$, the vector of all 1's, is in the column space of X.

Theorem 3.11. *Consider the linear model in (3.35). Suppose X a $n \times (k+1)$ matrix of rank $(k+1)$, and there exists a vector $v \in \mathbb{R}^{(k+1)\times 1}$ such that $Xv = \mathbf{1}$, that is, $\mathbf{1}$ is in the column space of X. Then, the sum of squares is partitioned as*

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 \tag{3.46}$$

Furthermore, $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ and $\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$ are independent.

Proof. Define the $n \times n$ matrix $U = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$. This is in fact an operator on a vector v, which returns a vector whose elements are all the same and equal to $\bar{v}$. Observe that U is symmetric, and idempotent, since $U^2 = \frac{1}{n^2}\mathbf{1}_n\mathbf{1}_n^T\mathbf{1}_n\mathbf{1}_n^T = \frac{1}{n^2}\mathbf{1}_n n\mathbf{1}_n^T = U$. As a result, $I - U$ is also idempotent. Then, we can write

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = [(I-U)Y]^T[(I-U)Y] = Y^T(I-U)^T(I-U)Y = Y^T(I-U)Y$$
$$\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 = [(H-U)Y]^T[(H-U)Y] = Y^T(H-U)^T(H-U)Y = Y^T(H-U)^2Y$$
$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = [(I-H)Y]^T[(I-H)Y] = Y^T(I-H)(I-H)Y = Y^T(I-H)Y$$

Using those relations, we check for the equality by considering the difference

$$\begin{aligned}
&\sum_{i=1}^{n}(Y_i - \bar{Y})^2 - \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 - \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \\
&\qquad = Y^T(I-U)Y - Y^T(H-U)^T(H-U)Y - Y^T(I-H)Y \\
&\qquad = Y^T(I-U)Y - Y^T(H-2UH+U)Y - Y^T(I-H)Y \\
&\qquad = Y^T\left[I - 2U + H + 2UH - I + H\right]Y \\
&\qquad = -2Y^TU(I-H)Y = -2Y^T\frac{1}{n}\mathbf{1}\mathbf{1}^T(I-H)Y = -(2Y^T\frac{1}{n}\mathbf{1})(\mathbf{1}^TI - \mathbf{1}^TH)Y
\end{aligned}$$

Now, $H\mathbf{1} = X(X^TX)^{-1}X^T\mathbf{1}$ is the projection of $\mathbf{1}$ onto the space spanned by the columns of X. (Look up for projection matrices). But since $\mathbf{1}$ is in the column space of X, $H\mathbf{1} = \mathbf{1}$. This results in

$$(\mathbf{1}^TI - \mathbf{1}^TH) = (\mathbf{1} - H\mathbf{1})^T = (\mathbf{1} - \mathbf{1})^T = 0^T.$$

Therefore, we have $-(2Y^T\frac{1}{n}\mathbf{1})(\mathbf{1}^TI - \mathbf{1}^TH)Y = 0$, proving the partitioning.

Next, the independence relation: First, note that $(H-U)Y$ and $(I-H)Y$ are both linear combinations of Y and hence they are jointly normal. Also check that their cross-covariance is

$$\begin{aligned}
\mathrm{Cov}((H-U)Y, (I-H)Y) &= (H-U)\mathrm{Cov}(Y,Y)(I-H) \\
&= (H-U)\sigma^2I(I-H) \\
&= \sigma^2(I-H)(H-U) \\
&= \sigma^2(H - U - H^2 + HU) \\
&= \sigma^2(H - U - H + HU) \\
&= \sigma^2(U - HU).
\end{aligned}$$

Since U is a matrix of all 1's, every column of U is in the column space of X, hence $HU = U$, and as a result, we get

$$\mathrm{Cov}((H-U)Y, (I-H)Y) = 0.$$

Therefore, $(H-U)Y$ and $(I-H)Y$ are independent. This implies that $\|(H-U)Y\|_2^2$ and $\|(I-H)Y\|_2^2$, which are the first and second terms on the RHS of the partitioning are also independent. Hence we have proven the claim. □

Theorem 3.11 immediately implies the partitioning for multiple linear regression model, which has an X matrix whose first column is all 1's. In the multiple linear regression model, the terms in the partitioning have specific names, in parallel to the simple linear regression model.

Corollary 3.5. Let X be a design matrix for multiple normal linear regression model as in (3.37), Then, we have (3.46), with specific names for the terms involved,

$$\underbrace{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}_{\text{TSS}}=\underbrace{\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2}_{\text{RSS}}+\underbrace{\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2}_{\text{RegSS}}$$

Furthermore, ResidSS and RegSS are independent.

Theorem 3.12. *Suppose we have a multiple normal linear regression model. Under the null hypothesis*

$$H_0:\beta_1=\ldots=\beta_k=0,$$

that is, there is no regression on any of the predictors (only the intercept parameter is allowed to be non-zero), we have

$$\frac{\text{TSS}}{\sigma^2}\sim\chi^2_{n-1},\quad \frac{\text{RSS}}{\sigma^2}\sim\chi^2_{n-k-1},\quad \frac{\text{RegSS}}{\sigma^2}\sim\chi^2_k.$$

Furthermore, the second and the third terms are independent.

Proof. The first result should be obvious, since the model reduces to $Y_i=\beta_0+e_i$ under H_0, random variables from a simple normal population. In the second one, the quantity in question is RSS/σ^2, which has already been shown to have a χ^2_{n-k-1} distribution, regardless of the value of β. For the last claim, we make use of Theorem 3.11, which states that the terms on the right-hand side are independent. Also by Theorem 3.11, the first quantity is equal to the sum of the second and the third quantities. Hence, by the result in part (b) of Exercise 1.5, we subtract the degrees of freedom to get $n-1-(n-k-1)=k$ to get the degrees of freedom. □

We have already seen in Theorem 3.12, the partitioning in Theorem 3.11 is relevant when we test the existence of a linear relationship whatsoever. Under $H_0:\beta_1=\ldots=\beta_k=0$, the test statistic

$$\frac{\text{RegSS}/k}{\text{RSS}/(n-k-1)}\sim f_{k,n-k-1}.$$

Therefore, the critical region

$$C=\left\{\frac{\text{RegSS}/k}{\text{RSS}/(n-k-1)}\geq f_{\alpha,k,n-k-1}\right\}$$

can be used to test H_0 at a significance level of α.

In the following, we will discuss more general tests; specifically, tests for one, all, or some components of β, as well as any linear combination of β.

B.3.2 Tests for a single component of β:

Recall that $\hat{\beta}$ has a normal distribution, $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(X^TX)^{-1})$. Therefore, for each component, we have

$$\hat{\beta}_i \sim \mathcal{N}(\beta_i, \sigma^2[(X^TX)^{-1}]_{ii}).$$

Combining the $(n-k-1)S_e^2/\sigma^2 \sim \chi^2_{n-k-1}$, we have the statistic

$$\frac{\hat{\beta}_i - \beta_i}{S_e\sqrt{[(X^TX)^{-1}]_{ii}}} \sim t_{n-k-1}.$$

One can use this to have a confidence interval for β_i or conducting a t-test for a given value of β_i. For example, the interval

$$\text{CI}_{\beta_i} = \hat{\beta}_i \pm t_{\alpha/2,n-k-1} S_e \sqrt{[(X^TX)^{-1}]_{ii}} \tag{3.47}$$

is a $100(1-\alpha)$ confidence interval for β_i.

B.3.3 Testing for the whole β

Recall that $\hat{\beta}$ has a normal distribution, $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(X^TX)^{-1})$. An essential result regarding $\hat{\beta}$ is the following:

Theorem 3.13. *We have* $\frac{1}{\sigma^2}(\hat{\beta}-\beta)^T X^T X(\hat{\beta}-\beta) \sim \chi^2_{k+1}$.

The proof of this theorem can be done by following the steps below. Firstly, as for any positive definite matrix, there is a unique square root $L = \sqrt{X^TX}$ such that we can write

$$X^TX = LL$$

for some $(k+1)\times(k+1)$ positive definite matrix L. Since L is positive definite, L^{-1} exists, and $L^{-1}L^{-1} = (L^2)^{-1} = (X^TX)^{-1}$. Furthermore, since X^TX is symmetric, L is symmetric, too, that is, $L^T = L$. The square root allows standardization of β.

Exercise 3.31. Follow the steps below to prove Theorem 3.13.

1. Show that for any $m\times n$ matrix A and a $n\times 1$ random vector U, we have $\text{Cov}(AU) = A\text{Cov}(U)A^T$.

2. Suppose X is of rank $k+1$ and let $L = \sqrt{X^TX}$. Show that $\frac{1}{\sigma}L(\hat{\beta}-\beta) \sim \mathcal{N}(0, I)$. [Hint: Use the first part and write $(X^TX)^{-1} = L^{-1}L^{-1}$.]

3. Using the previous part, prove Theorem 3.13.

Using Theorem 3.13, we can have a confidence region for $\hat{\beta}$. Note that the statistic

$$\frac{(\hat{\beta}-\beta)^T X^T X(\hat{\beta}-\beta)/(k+1)}{S_e^2} \sim f_{k+1,n-k-1} \tag{3.48}$$

Therefore, the set

$$\text{CR}_\beta = \left\{\beta \in \mathbb{R}^{k+1} : (\hat{\beta}-\beta)^T X^T X(\hat{\beta}-\beta) \leq (k+1) f_{\alpha,k+1,n-k-1} S_e^2\right\} \tag{3.49}$$

is a $100(1-\alpha)\%$ confidence region for β.

Exercise 3.32. Prove the claim above about the confidence region, that is, show that $P(\beta \in \text{CR}_\beta) = 1-\alpha$.

The confidence region can be used to derive confidence intervals for all β_i's that simultaneously hold with at least $1-\alpha$ probability. This is thanks to the following lemma.

Lemma 3.4. Let A be a $m \times m$ positive definite matrix and $c > 0$ be a constant. For every $z, u \in \mathbb{R}^{m\times 1}$, we have

$$z^T A z \leq c^2 \Rightarrow |z^T u| \leq c\sqrt{u^T A^{-1} u}$$

Exercise 3.33. Applying Lemma 3.4 with $A = X^T X$, $c^2 = (k+1) f_{\alpha,k+1,n-k-1} S_e^2$, $z = \hat{\beta}-\beta$, and $u_i = (\underbrace{0,\ldots,0}_{i \text{ times}}, 1, \underbrace{0,\ldots,0}_{k-i \text{ times}})^T$ for each $i = 0,\ldots,k$, show that

$$P\left(\left|\beta_i - \hat{\beta}_i\right| \leq \sqrt{(k+1) f_{\alpha,k+1,n-k-1}} \times S_e \sqrt{[(X^T X)^{-1}]_{ii}}, \text{ for } i = 0,\ldots,k\right) \tag{3.50}$$

simultaneously hold for all $i = 0,\ldots,k$ with probability at least $1-\alpha$.

Compare the CI for β_i in (3.47) and the simultaneous confidence intervals implied by (3.50). As expected, the first one is smaller. This is because the first one is designed for only β_i while the latter is designed to hold simultaneously with k other confidence intervals for the other β_j's with no less than $1-\alpha$ probability. If we wanted the confidence intervals in (3.47) for every β_i to hold simultaneously with a probability of at least $1-\alpha$, we would have to adjust the α value in each and have to modify the confidence intervals as

$$\hat{\beta}_i \pm t_{\alpha/2(k+1),n-k-1} S_e \sqrt{[(X^T X)^{-1}]_{ii}}.$$

A comparison between $t_{\alpha/2(k+1),n-k-1}$ and $\sqrt{(k+1) f_{\alpha,k+1,n-k-1}}$ would yield the approach one should go for in order to have simultaneous confidence intervals. (Obviously, we should use the ones yielding shorter intervals.)

B.3.4 Testing a part of β

Here, we will focus on testing null hypotheses which claim that some components of regression parameters β are zero. For example,

$$H_0 : \beta_{q+1} = \ldots = \beta_k = 0, \quad \text{vs} \quad H_1 : \text{not } H_0$$

This null hypothesis corresponds to the claim that the predictor variables $x_{1,j}, \ldots, x_{n,j}$ for $j > q$ do not play any role in the regression model, in other words, they are not 'predictors'. (Typically β_0 is not of much interest.)

A generalised partitioning with extra sum of squares: Consider the normal linear model, $Y \sim \mathcal{N}(X\beta, \sigma^2 I)$. Given the matrix X, let U_1 and U_2 be sub-matrices of X formed by a subset of the columns of X such that, all columns of U_1 reside in U_2. (For example U_1 and U_2 are formed by the first two and three columns of X, respectively.) Let $\hat{\beta}_1 = (U_1^T U_1)^{-1} U_1^T Y$, $\hat{\beta}_2 = (U_2^T U_2)^{-1} U_2^T Y$. Also, define $H_1 = U_1(U_1^T U_1)^{-1} U_1^T$ and $H_2 = U_2(U_2^T U_2)^{-1} U_2^T$, and $\hat{Y}_1 = H_1 Y$, and $\hat{Y}_2 = H_2 Y$. Also, recall the usual estimates $\hat{\beta}$ and $\hat{Y} = X\hat{\beta}$.

The following lemma is useful in the derivations to follow.

Lemma 3.5. For projection matrices constructed as above, we have

$$H_1 = H_2 H_1 = H_1 H_2.$$

Proof. Let v be any vector. $H_1 v$ is the projection of v onto the space spanned by the vectors in U_1. Since the space spanned by the vectors in U_1 is a subset of the space spanned by the vectors in U_2, we have $H_2(H_1 v) = H_1 v$. Hence the first equation is satisfied. Since this is true for any v, we have $H_1 = H_2 H_1$.

For the other equality, note that $v = H_2 v + v_e$, where v_e is the residual from the projection. It can be proven that v_e is orthogonal to the columns of U_2. This implies that v_e is orthogonal to columns of U_1, too. As a result, we have $H_1 v = H_1(H_2 v + v_e) = H_1 H_2 v$. Since this is true for any v, we have $H_1 = H_1 H_2$. □

We can extend the partitioning theorem as follows.

Theorem 3.14 (Partitioning with extra sum of squares)**.** *For U_1, U_2, and the following quantities defined above, we have*

$$\sum_{i=1}^{n} (Y_i - \hat{Y}_{1,i})^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_{2,i})^2 + \sum_{i=1}^{n} (\hat{Y}_{2,i} - \hat{Y}_{1,i})^2$$

Moreover, the terms $\sum_{i=1}^{n} (Y_i - \hat{Y}_{2,i})^2$ and $\sum_{i=1}^{n} (\hat{Y}_{2,i} - \hat{Y}_{1,i})^2$ are independent.

Proof. In the proof we will use the relations $H_1^2 = H_1$, $H_2^2 = H_2$, $(I - H_1)^2 = I - H_1$, $(I - H_2)^2 = I - H_2$, $H_1H_2 = H_2H_1 = H_1$, and $H_1^T = H_1$ and $H_2 = H_2^T$. For the first term, we have

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_{1,i})^2 = (Y - H_1Y)^T(Y - H_1Y) = Y^T(I - H_1)Y.$$

Similarly, for the second term, we have

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_{2,i})^2 = Y^T(I - H_2)Y.$$

For the third term, we have

$$\begin{aligned}\sum_{i=1}^{n}(\hat{Y}_{1,i} - \hat{Y}_{2,i})^2 = (H_1Y - H_2Y)^T(H_1Y - H_2Y) &= Y^T(H_1 + H_2 - 2H_1H_2)Y. \\ &= Y^T(H_1 + H_2 - 2H_1)Y. \\ &= Y^T(H_2 - H_1)Y.\end{aligned}$$

where the second line is due to Lemma 3.5. Checking for the difference between the first and the sum of the second and third terms, we get

$$Y^T(I - H_1)Y - Y^T(I - H_2)Y - Y^T(H_2 - H_1)Y = 0$$

Hence, we have proven the partitioning. For independence,

$$\begin{aligned}\mathrm{Cov}((Y - H_2Y), (H_1Y - H_2Y)) &= \mathrm{Cov}((I - H_2)Y, (H_1 - H_2)Y) \\ &= \sigma^2(I - H_2)(H_1 - H_2) \\ &= \sigma^2(H_1 - H_2 - H_2H_1 + H_2H_2) \\ &= \sigma^2(H_1 - H_2 - H_1 + H_2) = 0.\end{aligned}$$

where the third line is since H_2 is idempotent and $H_1H_2 = H_1$. Since the random variables $(I - H_2)Y$ and $(H_1 - H_2)Y$ are jointly normal, by their uncorrelatedness, they are independent. □

It is possible to partition the left-hand side into even smaller bits and show independence among those bits.

Theorem 3.15. *Let $U_1, \ldots, U_m$ be formed from the columns in such a way that, for every $j > i$, any column of U_i can also be found in U_j. Then*

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_{1,i})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_{m,i})^2 + \sum_{j=2}^{m}\left[\sum_{i=1}^{n}(\hat{Y}_{j,i} - \hat{Y}_{j-1,i})^2\right]$$

and all the sum of squares on the right hand side, $\sum_{i=1}^{n}(Y_i - \hat{Y}_{m,i})^2$, and $\sum_{i=1}^{n}(\hat{Y}_{j,i} - \hat{Y}_{j-1,i})^2$, $j = 2, \ldots, m$ are independent.

Proof. Note that the second term on the right hand side of Theorem 3.14 can be expanded by using the same theorem, and yields

$$\begin{aligned}\sum_{i=1}^{n}(Y_i-\hat{Y}_{1,i})^2 &= \sum_{i=1}^{n}(Y_i-\hat{Y}_{2,i})^2+\sum_{i=1}^{n}(\hat{Y}_{2,i}-\hat{Y}_{1,i})^2\\ &= \sum_{i=1}^{n}(Y_i-\hat{Y}_{3,i})^2+\sum_{i=1}^{n}(\hat{Y}_{3,i}-\hat{Y}_{2,i})^2+\sum_{i=1}^{n}(\hat{Y}_{2,i}-\hat{Y}_{1,i})^2\end{aligned}$$

Continuing until m, we have the recursion. For independence, note that $\hat{Y}_j-\hat{Y}_{j-1}=(H_j-H_{j-1})Y$. Therefore, for $j<j'$,

$$\begin{aligned}\mathrm{Cov}((H_j-H_{j-1})Y,(H_{j'}-H_{j'-1})Y) &= \sigma^2(H_j-H_{j-1})(H_{j'}-H_{j'-1})\\ &= \sigma^2(H_jH_{j'}-H_jH_{j'-1}-H_{j-1}H_{j'}+H_{j-1}H_{j'-1})\\ &= \sigma^2(H_j-H_j-H_{j-1}+H_{j-1})=0.\end{aligned}$$

since $j'-1\geq j$ and we can use either idempotency of H_j or Lemma 3.5. Combining with the fact that $\hat{Y}_j-\hat{Y}_{j-1}$ and $\hat{Y}_{j'}-\hat{Y}_{j'-1}$ are jointly normal, we prove that they are independent, so are their norm squares. Finally, for any j, we need to prove independence between $\sum_{i=1}^{n}(\hat{Y}_{j,i}-\hat{Y}_{j-1,i})^2$ and $\sum_{i=1}^{n}(Y_i-\hat{Y}_{m,i})^2$, which is implied by the independence between $\hat{Y}_j-\hat{Y}_{j-1}=(H_j-H_{j-1})Y$ and $Y-\hat{Y}_m=(I-H_m)Y$. We show it in the same fashion, by checking the covariance

$$\begin{aligned}\mathrm{Cov}((I-H_m)Y,(H_j-H_{j-1})Y) &= \sigma^2(H_j-H_{j-1}-H_mH_j+H_mH_{j-1})\\ &= \sigma^2(H_j-H_{j-1}-H_j+H_{j-1})=0\end{aligned}$$

as desired. □

Application to multiple linear regression: Theorem 3.14 implies several results for the multiple normal linear regression model. We can gather the most important ones into a theorem. Let X be a design matrix for the multiple normal linear regression model as in (3.37). For each $i=1,\ldots,k$, let

$$U_i=\begin{bmatrix}1 & x_{1,1} & \ldots & x_{1,i}\\ 1 & x_{2,1} & \ldots & x_{2,i}\\ \vdots & \vdots & \ddots & \vdots\\ 1 & x_{n,1} & \ldots & x_{n,i}\end{bmatrix},\quad i=1,\ldots,k.$$

Also, let $H_i=U_i(U_i^TU_i)^{-1}U_i^T$. and $\hat{Y}_i=H_iY$. Finally, let $H=X(X^TX)^{-1}X^T$ and $\hat{Y}=HY$ as usual. For any $1\leq m\leq m'\leq k$, define the regression sum of squares and

residual sum of squares based on U_m as

$$\text{RegSS}(1,\ldots,m) = \sum_{i=1}^{n}(\hat{Y}_{m,i} - \bar{Y})^2, \tag{3.51}$$

$$\text{RSS}(1,\ldots,m) = \sum_{i=1}^{n}(Y_i - \hat{Y}_{m,i})^2 \tag{3.52}$$

and note that $\text{RegSS} = \text{RegSS}(1,\ldots,k)$ and $\text{RSS} = \text{RSS}(1,\ldots,k)$. Next, define also the extra sum of squares

$$\text{RegSS}(m+1|1,\ldots,m) = \text{RegSS}(1,\ldots,m+1) - \text{RegSS}(1,\ldots,m) \tag{3.53}$$

$$\text{RegSS}(m+1,\ldots,m'|1,\ldots,m) = \sum_{j=m+1}^{m'} \text{RegSS}(j|1,\ldots,j-1). \tag{3.54}$$

Using those definitions, it is not hard to see that

$$\text{RegSS}(m+1,\ldots,k|1,\ldots,m) = \text{RegSS} - \text{RegSS}(1,\ldots,m) \tag{3.55}$$

Theorem 3.16. *For any $1 \le m \le m' \le k$, we have the following three main results.*

1. $\text{TSS} = \text{RegSS}(1,\ldots,m) + \text{RSS}(1,\ldots,m)$,
2. $\text{RegSS}(m+1,\ldots,m'|1,\ldots,m) = \sum_{i=1}^{n}(\hat{Y}_{m',i} - \hat{Y}_{m,i})^2$.
3. $\text{RegSS}(1), \text{RegSS}(2|1), \ldots, \text{RegSS}(m|1,\ldots,m-1)$ *and* $\text{RSS}(1,\ldots,m')$ *are all independent.*

Corollary 3.6. For any $1 \le m \le k$, we have

$$\text{RegSS} = \text{RegSS}(1,\ldots,m) + \text{RegSS}(m+1,\ldots,k|1,\ldots,m), \tag{3.56}$$

and $\text{RegSS}(m+1,\ldots,k|1,\ldots,m)$, $\text{RegSS}(1,\ldots,m)$, and RSS are independent.

Lemma 3.6. Under the null hypothesis $H_0 : \beta_{m+1} = \ldots = \beta_k = 0$, we have $\frac{1}{\sigma^2}\text{RSS}(1,\ldots,m) \sim \chi^2_{n-m-1}$.

Proof. Under $H_0 : \beta_{m+1} = \ldots = \beta_k = 0$, $X\beta = U_m\eta_m$, where $\eta_m = \begin{bmatrix}\beta_0 & \beta_1 & \ldots & \beta_m\end{bmatrix}^T$. Recall that

$$\text{RSS}(1,\ldots,m) = \sum_{i=1}^{n}(Y_i - \hat{Y}_{m,i})^2 = Y^T(I - H_m)Y$$

Also, we have $Y - U_m\eta_m \sim \mathcal{N}(0, \sigma^2 I)$, and the rank of H_m is $m+1$. Therefore, by Cochran's theorem,

$$\frac{1}{\sigma^2}(Y - U_m\eta_m)^T(I - H_m)(Y - U_m\eta_m) \sim \chi^2_{n-m-1}.$$

All we have to do is to show the equality between $(Y - U_m\eta_m)^T(I - H_m)(Y - U_m\eta_m)$ and $\text{RSS}(1, \ldots, m)$.

$$(Y-U_m\eta_m)^T(I-H_m)(Y-U_m\eta_m) = Y^T(I-H_m)Y-2\eta_m^TU_m^T(I-H_m)Y+\eta_m^TU_m^T(I-H_m)U_m\eta_m$$

We show that the second and third terms are zero, since

$$U_m^T(I - H_m) = U_m^T - U_m^TU_m(U_m^TU_m)^{-1}U_m^T = U_m^T - U_m^T = 0.$$

Hence, we have shown that

$$(Y - U_m\eta_m)^T(I - H_m)(Y - U_m\eta_m) = Y^T(I - H_m)Y$$

and thus proven the claim. □

Back to testing a part of β: Now we can state the result that enables testing a part of β.

Theorem 3.17. *Under the null hypothesis $H_0 : \beta_{m+1} = \ldots = \beta_k = 0$, we have*

$$\frac{1}{\sigma^2}\text{RegSS}(m + 1, \ldots, k|1, \ldots, m) \sim \chi^2_{k-m},$$

independently from RSS*, and therefore*

$$\frac{\text{RegSS}(m + 1, \ldots, k|1, \ldots, m)/(k - m)}{\text{RSS}/(n - k - 1)} \sim f_{k-m,n-k-1}. \tag{3.57}$$

Proof. By the first item of Theorem 3.16, we can write

$$\text{RSS} + \text{RegSS} = \text{RSS}(1, \ldots, m) + \text{RegSS}(1, \ldots, m)$$

for any value of m, or

$$\text{RSS}(1, \ldots, m) = \text{RSS} + \text{RegSS} - \text{RegSS}(1, \ldots, m).$$

Substituting (3.56) in Corollary 3.6 into the relation, we have

$$\text{RSS}(1, \ldots, m) = \text{RSS} + \text{RegSS}(m + 1, \ldots, k|1, \ldots, m). \tag{3.58}$$

Regardless of the null hypothesis, we have $\text{RSS}/\sigma^2 \sim \chi^2_{n-k-1}$. By Lemma 3.6, we also have $\text{RSS}(1, \ldots, m)/\sigma^2 \sim \chi^2_{n-m-1}$. Furthermore, by corollary By Corollary 3.6, RSS and $\text{RegSS}(m+1, \ldots, k|1, \ldots, m)$ are independent. Therefore, $\text{RegSS}(m+1, \ldots, k|1, \ldots, m)/\sigma^2$ must have a chi-sqaured distribution also, with degrees of freedom being the difference $n - m - 1 - (n - k - 1) = k - m$. Hence, we have proven the claim. □

In fact, the test statistic in (3.57) can be shown to be the test statistic of the likelihood ratio test for the null hypothesis $H_0 : \beta_{m+1} = \ldots = \beta_k = 0$. For a likelihood ratio test of size α, the critical region is

$$\left\{\frac{\text{RegSS}(m + 1, \ldots, k|1, \ldots, m)/(k - m)}{\text{RSS}/(n - k - 1)} > f_{\alpha,k-m,n-k-1}\right\}. \tag{3.59}$$

B.3.5 Testing for any linear combination of β

All the tests for β we have seen so far can be written in terms of a $r \times (k+1)$ full rank matrix C with rank r, and a vector $r \times 1$ vector c_0 as

$$H_0 : C\beta = c_0, \quad H_1 : C\beta \neq c_0. \tag{3.60}$$

Note that $C\hat{\beta}$ has a normal distribution with $C\hat{\beta} \sim \mathcal{N}(C\beta, \sigma^2 C(X^T X)^{-1} C^T)$. Using Theorem 3.13 but with the new covariance matrix, it can be shown that

$$\frac{1}{\sigma^2}(C\hat{\beta} - C\beta)^T [C(X^T X)^{-1} C^T]^{-1} (C\hat{\beta} - C\beta) \sim \chi^2_r$$

Combining this with the estimator S_e^2 for σ^2, which is independent from $C\hat{\beta}$, we have

$$\frac{(C\hat{\beta} - C\beta)^T [C(X^T X)^{-1} C^T]^{-1} (C\hat{\beta} - C\beta)/r}{S_e^2} \sim f_{r,n-k-1}.$$

A $100(1-\alpha)\%$ confidence region *for* $C\beta$ is given by

$$\text{CR}_{C\beta} = \left\{ C\beta : (C\hat{\beta} - C\beta)^T [C(X^T X)^{-1} C^T]^{-1} (C\hat{\beta} - C\beta) \leq r S_e^2 f_{\alpha,r,n-k-1} \right\} \tag{3.61}$$

A test based on $\text{CR}_{C\beta}$ would reject H_0 if c_0 is not in $\text{CR}_{C\beta}$, that is, the critical region is

$$\{(C\hat{\beta} - c_0)^T [C(X^T X)^{-1} C^T]^{-1} (C\hat{\beta} - c_0) > r S_e^2 f_{\alpha,r,n-k-1}\}. \tag{3.62}$$

This can be shown to be the likelihood ratio test for the hypotheses in (3.60).

Equivalence of the tests: We will show that the test developed in Section B.3.5 for any linear combination of β is equivalent to the tests we have derived separately for a single, all, or some components of β.

- For the null hypothesis $H_0 : \beta_i = \beta_{i0}$, the corresponding parameters are $r = 1$, $c_0 = \beta_{i0}$, and C is a $1 \times (k+1)$ vector whose $i+1$'th element is 1 and the rest of its elements are zero. It is left to the reader that, the resulting confidence interval $\text{CI}_{C\beta}$ in (3.61) is equivalent to the one in (3.47).

- For a null hypothesis regarding the whole parameter β, in the form of $H_0 : \beta = \vartheta_0$, the corresponding parameters are $r = k+1$, $c_0 = \vartheta_0$, and $C = I_{k+1}$. It is left to the reader that, the resulting confidence region $\text{CR}_{C\beta}$ in (3.61) is equivalent to CR_β in (3.49).

- For a null hypothesis regarding a part of β, such as $H_0 : \beta_{m+1} = \ldots = \beta_k = 0$, the corresponding $r = k - m$, $c_0 = 0_{(k-m)\times 1}$, and C is a $(k-m) \times (k+1)$ matrix given by

  $$C = \begin{bmatrix} \mathbf{0}_{(k-m)\times(m+1)} & I_{k-m} \end{bmatrix} \tag{3.63}$$

  The next theorem states the equivalence of the critical regions in (3.62) and (3.59).

Theorem 3.18. *The critical regions in (3.62) and (3.59), developed for the hypothesis $H_0 : \beta_{m+1} = \ldots = \beta_k = 0$, are exactly the same.*

Proof. Observing (3.62) and (3.59), we can see that the denominators are equal ($\text{RSS}/(n-k-1) = S_e^2$) and the critical values for the ratios are also equal ($r = k-m$). Therefore, noting that $c_0 = 0$ for the H_0 in question, for the equality of the tests it suffices to show that

$$\text{RegSS}(m+1,\ldots,k|1,\ldots,m) = (C\hat{\beta})^T[C(X^TX)^{-1}C^T]^{-1}(C\hat{\beta})$$

with C given in (3.63) . For the term on the left hand side, we have

$$\begin{aligned}\text{RegSS}(m+1,\ldots,k|1,\ldots,m) &= \text{RSS}(1,\ldots,m) - \text{RSS} \\ &= Y^T(I-H_m)^T(I-H_m)Y - Y^T(I-H)^T(I-H)Y \\ &= Y^T(I-H_m)Y - Y^T(I-H)Y \\ &= Y^T(H-H_m)Y.\end{aligned}$$

since the matrices $I-H$ and $I-H_m$ are symmetric and idempotent.

The term on the right-hand side is

$$\begin{aligned}(C\hat{\beta})^T[C(X^TX)^{-1}C^T]^{-1}(C\hat{\beta}) &= Y^TX(X^TX)^{-1}C^T[C(X^TX)^{-1}C^T]^{-1}C(X^TX)^{-1}X^TY. \\ &= Y^TV(V^TV)^{-1}V^TY. \qquad (3.64)\end{aligned}$$

where $V = X(X^TX)^{-1}C^T$. (Check that $V^TV = C(X^TX)^{-1}X^TX(X^TX)^{-1}C^T = C(X^TX)^{-1}C^T$, which is the expression in the brackets as in the first line of the equation above.) The matrix H_m projects onto the space spanned by the first $m+1$ columns of X, and $V(V^TV)^{-1}V^T$ projects onto the space spanned by the last $k-m$ columns of X. In order to show that $V(V^TV)^{-1}V^T + H_m = H$, it suffices to show that $V(V^TV)^{-1}V^T$ and H_m are orthogonal. Noting that $X_m = XC'^T$ with $C' = \begin{bmatrix} I_{m+1} & \mathbf{0}_{(m+1)\times(k-m)}\end{bmatrix}$, we check that

$$V^TX_m = C(X^TX)^{-1}X^TXC'^T = CC'^T = \mathbf{0}$$

hence we are done. □

B.4 Prediction

Given a new predictor vector $x_0 = \begin{bmatrix} x_{00} & \ldots & x_{0k}\end{bmatrix}^T$, a point estimate of the population regression function, which is the expected value of the response Y_0, can be obtained as

$$x_0^T\hat{\beta}.$$

Obviously, this is an unbiased estimator of the population regression function at x_0, since $E(x_0^T\hat{\beta}) = x_0^T\beta$. The distribution of $\hat{\beta}$ allows us to have confidence intervals for $x_0^T\beta$. For

a single x_0 vector, one can consider $C = x_0^T$ and use the confidence interval in (3.61).For multiple x_0 vector, Bonferroni correction can be consulted. In fact, thanks to Scheffe (once again), one can have simultaneous predictions, i.e., simultaneous confidence intervals for $x_0^T\beta$ for all x_0.

Theorem 3.19 (Scheffe's confidence intervals)**.** *The confidence intervals*

$$x_0^T\hat{\beta} \pm \sqrt{(k+1)f_{\alpha,k+1,n-k-1}}S_e\sqrt{x_0^T(X^TX)^{-1}x_0} \tag{3.65}$$

for $x_0^T\beta$ hold simultaneously for all x_0 with $1-\alpha$ probability. In other words,

$$P\left(\left|x_0^T(\beta-\hat{\beta})\right| \le \sqrt{(k+1)f_{\alpha,k+1,n-k-1}}S_e\sqrt{x_0^T(X^TX)^{-1}x_0},\ \textit{for all } x_0\right) = 1-\alpha.$$

Exercise 3.34. Heart catheterization is sometimes performed on children with congenital heart defects. A Teflon tube (catheter) 3 mm in diameter is passed into a major vein or artery in the femoral region and pushed up into the heart to obtain information about the heart's physiology and functional ability. The length of the catheter is typically determined by a physician's educated guess. In a small study involving 12 children, the exact catheter length required was determined by using a fluoroscope to check that the tip of the catheter had reached the pulmonary artery. The patients' heights and weights were recorded. The objective was to see how accurately catheter length could be determined by these two variables. The data are given in the following table:

Height (in.)	Weight (in.)	distance to pulmonary artery (cm)
42.8	40.0	37.0
63.5	93.5	49.5
37.5	35.5	34.5
39.5	30.0	36.0
45.5	52.0	43.0
38.5	17.0	28.0
43.0	38.5	37.0
22.5	8.5	20.0
37.0	33.0	33.5
23.5	9.5	30.5
33.0	21.0	38.5
58.0	79.0	47.0

For the following questions, take $\alpha = 0.1$. A normal multiple linear regression is considered to model the relation between the height and weight of a child (predictors) and the distance (response), in the form of

$$Y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + e_i.$$

(a) Calculate $\hat{\beta}$ and S_e^2, the estimators for the regression parameter vector β and the variance σ^2 of e_i.

(b) Derive a confidence region for β.

(c) Derive confidence intervals for all components of β that hold simultaneously with $(1 - \alpha)$ probability.

(d) Test the null hypothesis that there is no linear regression of the distance variable on the weight or height of a patient.

(e) Test the null hypothesis that the distance does not depend on the weight of the patient.

Chapter 4

Bayesian Inference

Summary: *In this chapter, we provide a brief introduction to Bayesian statistics. Some quantities of interest that are calculated from the posterior distribution will be explained. We will see some examples where one can find the exact form of the posterior distribution. In particular, we will discuss conjugate priors that are useful for deriving tractable posterior distributions. This chapter also introduces a relaxation in the notation to be adopted in the later chapters.*

A Introduction

Bayesian statistics is based on the Bayesian interpretation of probability, in which probability expresses a degree of belief in an event.

In Bayesian statistics, the unknown parameter, which is generically denoted by θ throughout the course, is considered a random variable with a prior distribution. The prior distribution formalises any form of *a priori* belief or information about the parameter before the data are collected. Being a probability distribution, the prior distribution expresses the degree of belief in events regarding the unknown parameter. After obtaining the data, the prior distribution is updated using the Bayes' rule, yielding the posterior distribution. The posterior distribution expresses the *a posteriori belief* about the unknown parameter.

A.1 A review of Bayes' rule

Consider the probability space $(\Omega, \mathcal{F}, P)$. Given two sets $A, B \in \mathcal{F}$, the conditional distribution of A given B is

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B|A)}{P(B)} \tag{4.1}$$

This result is known as Bayes' theorem or Bayes' rule. Here we see some examples where Bayes' rule is in action to calculate posterior probabilities.

Example 4.1 (**Conditional probabilities of sets**). A pair of fair (unbiased) dice are rolled independently. Let the outcomes be X_1 and X_2.

- It is observed that the sum is $S = X_1 + X_2 = 8$. What is the probability that the outcome of at least one of the dice is 3?

  We apply the Bayes rule: Define the sets $A = \{(X_1, X_2) : X_1 = 3 \text{ or } X_2 = 3\}$. $B = \{(X_1, X_2) : S = 8\}$, so that the desired probability is $P(A|B) = P(A \cap B)/P(B)$.

  $$B = \{(2,6), (3,5), (4,4), (5,3), (6,2)\}, \quad A \cap B = \{(3,5), (5,3)\}.$$

  Since the dice are fair, every outcome is equiprobable, having a probability of 1/36. Therefore,

  $$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{2/36}{5/36} = \frac{2}{5}.$$

- It is observed that the sum is even. What is the probability that the sum is smaller than or equal to 4? Similarly, we define the sets $A = \{(X_1, X_2) : X_1 + X_2 \leq 4\}$. $B = \{(X_1, X_2) : X_1 + X_2 \text{ is even}\}$. Explicity, we have

  $$B = \{(X_1, X_2) : X_1, X_2 \text{ are both even}\} \cup \{(X_1, X_2) : X_1, X_2 \text{ are both odd}\}.$$
  $$A \cap B = \{(1,1), (1,3), (3,1), (2,2)\}.$$

  Therefore,

  $$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{4/36}{3/6 \times 3/6 + 3/6 \times 3/6} = \frac{2}{9}.$$

Example 4.2 (**Model selection**). There are two coins in an urn, one fair and one biased with a probability of tail $\rho = 0.3$. Someone picks up one of the coins at random (with half probability for picking up either coin) and tosses it n times and reports the outcomes: $\mathcal{D} = (H, T, H, H, T, H, H, H, T, H)$. Conditional on $\mathcal{D}$, what is the probability that the fair coin was picked up?

We have two hypotheses (models): H_1: The coin picked up was the fair one, H_2: The coin picked was the biased one. The prior probabilities for these models are the same: $P(H_1) = P(H_2) = 0.5$. The likelihood of data, that is the conditional probability of the outcomes is:

$$P(\mathcal{D}|H_i) = \begin{cases} 1/2^{10}, & i = 1, \\ \rho^{n_T}(1-\rho)^{n_H}, & i = 2, \end{cases}$$

where n_T and n_H are the number of times the coin showed tail and head, respectively. From Bayes' rule, we have

$$\begin{aligned} P(H_1|\mathcal{D}) &= \frac{P(\mathcal{D}, H_1)}{P(\mathcal{D})} = \frac{P(H_1)P(\mathcal{D}|H_1)}{P(\mathcal{D}|H_1)P(H_1) + P(\mathcal{D}|H_2)P(H_2)} \\ &= \frac{1/2 \times 1/2^{10}}{1/2 \times 1/2^{10} + 1/2 \times \rho^{n_T}(1-\rho)^{n_H}} \\ &= \frac{1/2^{10}}{1/2^{10} + \rho^{n_T}(1-\rho)^{n_H}} \end{aligned}$$

and, of course, $P(H_2|\mathcal{D}) = 1 - P(H_1|\mathcal{D})$. Substituting $\rho = 0.3$ and $n_T = 3$, we have $P(H_1|\mathcal{D}) = 0.3052$ and $P(H_2|\mathcal{D}) = 0.6948$.

Exercise 4.1. Consider the discrete random variables $X \in \{1, 2, 3\}$ and $Y \in \{1, 2, 3, 4\}$ whose joint probabilities are given in the table below.

$p_{X,Y}(x, y)$	$y = 1$	$y = 2$	$y = 3$	$y = 4$	$p_X(x)$
$x = 1$	1/40	3/40	4/40	2/40	
$x = 2$	5/40	7/40	6/40	5/40	
$x = 3$	1/40	2/40	2/40	2/40	
$p_Y(y)$					

- Find the marginal probabilities $p_X(x)$ and $p_Y(y)$ for all $x = 1, 2, 3$, $y = 1, 2, 3, 4$ and fill in the rest of the table.
- Find the conditional probabilities $p_{X|Y}(x|y)$ and $p_{Y|X}(y|x)$ for all $x = 1, 2, 3$, $y = 1, 2, 3, 4$ and fill in the relevant empty tables.

$p_{X\|Y}(x\|y)$	$y = 1$	$y = 2$	$y = 3$	$y = 4$
$x = 1$				
$x = 2$				
$x = 3$				

$p_{Y\|X}(y\|x)$	$y = 1$	$y = 2$	$y = 3$	$y = 4$
$x = 1$				
$x = 2$				
$x = 3$				

A.2 Posterior distribution

In classical statistics, θ is a fixed non-random variable and inference is largely based on developing confidence intervals or testing hypotheses for θ. The interpretation of probability in classical statistics is the frequentist interpretation. This shows itself in the theoretical guarantees of classical procedures: A test of significance α guarantees that if one conducts the test independently certain many times (each with a different sample independent from the others), he/she is expected to conduct a type-I error in $100\alpha\%$ of the time.

In contrast, in Bayesian statistics, θ is random variable, and inference is based on the posterior distribution of θ conditional on the available data observed so far. The posterior distribution of θ given the data $X = x$ is expressed in terms of a probability mass function or a probability distribution function as

$$p(\theta|x) = \frac{\overbrace{p(\theta)}^{\text{prior}}\ \overbrace{p(x|\theta)}^{\text{likelihood}}}{\underbrace{p(x)}_{\text{evidence}}} = \frac{\overbrace{p(\theta, x)}^{\text{joint distribution}}}{p(x)} \tag{4.2}$$

Here $p(\theta)$ is called the *prior distribution* of θ, which the probability distribution that expresses our belief or information about the θ before we see the data x. The conditional probability density $p(x|\theta)$ is called the *likelihood* and it bears the new information about θ that is brought by the data x and $p(x)$ is called the *evidence*.

The interpretation of probability in Bayesian statistics is called the Bayesian interpretation, where probability quantifies the degree of belief about an event.

Because of the fundamental philosophical difference between classical and Bayesian statistics, the problems they deal with are usually different. However, sometimes they do happen to have to answer similar statistical questions in practice. When the data is ample, the conclusions drawn by classical and Bayesian statistics agree, they can differ quite dramatically with little to moderate amount of data.

Example 4.3 (DNA test). The following example is very commonly used to illustrate the effect of the prior distribution in Bayesian statistics. Suppose that the police find the body of someone who was murdered. The murder weapon was found on the scene as well.

Some DNA evidence was found on the murder weapon. The police compare this DNA to a list of 1000.000 people in their database. Those 1000.000 people in the database have committed crimes previously, so there is a good chance, let us say 0.5 that the guilty person is on the list. Therefore, a randomly selected person on the list has a probability of being guilty with a probability of $1/2 \times 1/1000.000 = 1/2000.000$. This is the prior probability of any person on the list being guilty of this specific crime or murder.

The DNA test has the following accuracy measures:

	Match	No Match
Compared DNAs belong to the same person	1	0
Compared DNAs do not belong to the same person	10^{-6}	$1 - 10^{-6}$

Suppose now that the DNA of a person in the list is compared to the DNA found at the crime scene, and the result is positive, that is, the DNA test resulted in a match. Based on this evidence, can we determine that the person is guilty?

Here, the unknown parameter $\theta \in \{0, 1\}$ represents the true status of the suspect, guilty ($\theta = 1$) or innocent ($\theta = 0$). The available data $x = 1$ is the DNA test result. Note that this is the observed value of a random variable X, whose conditional distribution on θ is specified in the table above. Namely, we have the conditional probabilities

$$\begin{aligned} &p(x = 1|\theta = 1) = 1, \quad p(x = 0|\theta = 1) = 0 \\ &p(x = 0|\theta = 0) = 1 - 10^{-6}, \quad p(x = 1|\theta = 0) = 10^{-6}. \end{aligned} \tag{4.3}$$

Let us tackle this problem with a non-Bayesian approach first. This approach would neglect any a priori belief about the person's true status and base inference only on the conditional distribution of X given θ. The problem can be formulated in the classical sense as a hypothesis testing, with

$$H_0 : \theta = 0 \text{ (person is innocent)}, \quad H_1 : \theta = 1 \text{ (person is guilty)}$$

To perform the hypothesis testing, we need a decision rule based on the single outcome, which is the DNA test result for the suspect. Let us consider the decision

$$\text{Decision} = \begin{cases} H_1 & \text{for } X = 1, \\ H_0 & \text{for } X = 0. \end{cases} \tag{4.4}$$

What are the type I and type II errors of this test? If someone is guilty, there is no chance of missing that, since the DNA test results in a 1 certainly. Therefore, the type II error probability is 0. Type I probability also looks very good. If the suspect is innocent, we convict the suspect with 10^{-6} probability. Therefore, $\alpha = 10^{-6}$.

Both type I and type II errors suggest that the decision rule in (4.4) is quite reliable. Yet, should we *really* convict the suspect based on the positive outcome? The Bayesian approach to the same problem provides a different perspective. This time we do take into account the prior belief about the suspect's true status. The prior probability of the suspect being guilty is 1/2000000 since the database contains 1000000 people and it is believed that the guilty person is on the list with a probability of 1/2. Therefore, the prior probabilities are

$$p(\theta = 1) = 1 - p(\theta = 0) = 5 \times 10^{-7}. \tag{4.5}$$

Applying the Bayes' rule in (4.2) with the prior in (4.5) and likelihood in (4.3), we get

$$\begin{aligned} p(\theta = 1|x = 1) &= \frac{p(\theta = 1)p(x = 1|\theta = 1)}{p(x = 1)} \\ &= \frac{p(\theta = 1)p(x = 1|\theta = 1)}{p(x = 1|\theta = 1)p(\theta = 1) + p(x = 1|\theta = 0)p(\theta = 0)} \\ &= \frac{(1/2000000) \times 1}{(1/2000000) \times 1 + (1999999/2000000) \times 10^{-6}} \\ &\approx 0.33. \end{aligned}$$

While the likelihood $p(x = 1|\theta)$ favours $\theta = 1$ over $\theta = 0$ by far, the suspect is guilty with a posterior probability less than 1/2, leave aside being beyond reasonable doubt.

The conviction of the suspect according to the decision rule is referred to as the *convictor's fallacy*, since the conditional distribution $p(\theta|x)$ is mixed up with the likelihood $p(x|\theta)$.

What happened here? The reason for the dramatical difference between $p(\theta|x)$ and $p(x|\theta)$ is the prior $p(\theta)$ favouring $\theta = 0$ strongly. For example, if, by some preliminary study conducted with other sorts of evidence, the number of suspects was reduced to a dozen of people, say m, then the prior probability of a suspect being guilty would be $1/m$, a larger number than 1/2000000.

Exercise 4.2. Suppose this time that two DNA tests, instead of one, are conducted to compare the suspect's DNA with the DNA found at the crime scene. Let the outcomes of those tests be X_1 and X_2 and assume that X_1 and X_2 are independent. Answer the following questions:

- What is the posterior probability of the suspect being guilty given that $X_1 = 1$ and $X_2 = 0$?
- What is the posterior probability of the suspect being guilty given that $X_1 = 1$ and $X_2 = 1$?

Exercise 4.3. Suppose there are k normal populations, with population distributions $\mathcal{N}(\mu_1, \sigma_1^2), \ldots, \mathcal{N}(\mu_k, \sigma_k^2)$. A random sample X of size 1 is drawn from one of those populations chosen randomly according to the probabilities $w_1, \ldots, w_k$ with $w_1 + \ldots + w_k = 1$. Let $J \in \{1, \ldots, k\}$ be the population number from which X is drawn.

- Express the joint probability distribution of J and X by writing down $p(j, x)$.
- Write down the marginal probability density function $p(x)$ of the marginal distribution of X and show that it is a mixture of normal densities.
- Suppose that $X = x$ is observed. Write down the posterior distribution of J given $X = x$. That is, derive the expression for the conditional probability that $X = x$ is drawn from the i'th population for each $i = 1, \ldots, k$.

Notation: Notice that we used the same letter p to denote four different probability distributions in (4.2). A more rigorous treatment would be to differentiate the distributions with subscripts, such as

$$p_{\Theta|X}(\theta|x) = \frac{p_\Theta(\theta) p_{X|\Theta}(x|\theta)}{p_X(x)}.$$

However, it is common practice to drop the cumbersome subscripts and use $p(x, y)$, $p(x)$, $p(x|y)$, etc. as in in (4.2) whenever it is clear from the context what distribution we mean. We will also adopt that approach in this chapter.

It is also common to use densities as well as distributions to indicate the distribution of a random variable. For example, all the expressions below mean the same thing: X is distributed from the distribution $\mathcal{F}$, whose pdf or pmf is $p(x)$

$$X \sim \mathcal{F}, \quad X \sim p(\cdot), \quad X \sim p(x), \quad x \sim \mathcal{F}, \quad x \sim p(\cdot), \quad x \sim p(x).$$

In the rest of this document, we will use the aforementioned notations interchangeably, choosing the most suitable one depending on the context.

A.3 Prior selection

Consider the variables X, Y and Bayes' theorem for $p(x|y)$ in words,

$$\text{posterior} \propto \text{prior} \times \text{likelihood}.$$

In Bayesian statistics, the usual first step to building a statistical model is to decide on the likelihood, i.e. the conditional distribution of the data given the unknown parameter. The likelihood represents the model choice for the data and it should reflect the real stochastic dynamics/phenomena of the data generation process as accurately as possible.

Bayesian inference for the unknown parameter requires assigning a prior distribution to it. Choosing a prior is the subject of a huge debate in the field of Bayesian statistics. There are several methods to choose a prior, among which one may be preferred over the other depending on the situation.

A.3.1 Informative priors

An informative prior may be used when there exists specific, definite information about θ. This is the case when there is expert knowledge about θ. Informative priors also appear when data accumulate in time. A posterior distribution obtained from pre-existing data can be set as the prior distribution for the new data to come. That is why the terms "prior" and "posterior" are generally relative to a specific piece of data.

If data are collected systematically, the mechanism of using the current posterior as the prior for the future evidence is natural. Suppose $X_1, X_2, \ldots$ are a sequence of random observations that are independent given the parameter θ. Therefore the likelihood function for any $X_1 = x_1, \ldots, X_n = x_n$ is given by

$$p(x_{1:n}|\theta) = \prod_{i=1}^{n} p(x_i|\theta).$$

When the data are collected sequentially as $x_1, x_2, \ldots$, a respective sequential update is available. Observe that

$$\begin{aligned} p(\theta|x_{1:n}) &= \frac{p(\theta)p(x_{1:n}|\theta)}{p(x_{1:n})} \\ &= \frac{p(\theta)p(x_{1:n-1}|\theta)}{p(x_{1:n-1})} \frac{p(x_n|x_{1:n-1},\theta)}{p(x_n|x_{1:n-1})} \\ &= \underbrace{p(\theta|x_{1:n-1})}_{\text{new prior}} \frac{\overbrace{p(x_n|\theta)}^{\text{likelihood (specific to } x_n)}}{\underbrace{p(x_n|x_{1:n-1})}_{\text{conditional evidence}}} \end{aligned}$$

Therefore, $p(\theta|x_{1:n-1})$ can be considered as the prior for the new observation x_n.

In fact, an analogy can be drawn between expert knowledge and $p(\theta|x_{1:n-1})$. Assuming the expert is a Bayesian, the expertise he/she gained from the past is expressed as $p(\theta|x_{1:n-1})$.

Sequential updates of the posterior as above has contributed greatly to the popularity of Bayesian statistics for sequential data. Note that in the previous discussion we have confined to independent observations; however there are similar sequential updates for more complex models. In its most generality, i.e., without the conditional independence assumption, we have the following sequential update

$$p(\theta|x_{1:n}) = \underbrace{p(\theta|x_{1:n-1})}_{\text{new prior}} \frac{\overbrace{p(x_n|x_{1:n-1},\theta)}^{\text{conditional likelihood (given } x_n)}}{\underbrace{p(x_n|x_{1:n-1})}_{\text{conditional evidence}}}$$

Note that, differently from the update under the conditional independence assumption, $p(x_n|\theta)$ is replaced by $p(x_n|x_{1:n-1}, \theta)$ (which, of course reduces to the former under conditional independence).

A.3.2 Weakly informative priors

Weakly informative priors are used when there is not a strong opinion about θ. Such priors express partial knowledge about θ. A prior distribution with high variance, or one that very loosely constraints θ to a wide range, can be named a weakly uninformative prior. For example, if θ is the temperature tomorrow, a normal distribution $N(20, 30^2)$, or a Uniform distribution Unif$(-10, 40)$ could be a weakly informative prior. The distribution Unif$(-10, 40)$ expresses that the temperature tomorrow is not expected below -10 and above 40 degrees. A similar interpretation goes for $N(20, 30^2)$ as well. Gamma and inverse gamma distributions can also be used for weakly informative priors for parameters with a positive range, such as the variance parameter.

A.3.3 Uninformative priors

An uninformative prior expresses the objective information about a parameter, information that everyone has and can agree on. A prior like that can be read as "θ is a positive number", or "θ is between 0 and 1". When θ is discrete, one uninformative prior can be set by assigning equal probability to each value θ can take. For continuous and bounded θ, the continuous uniform distribution is a choice.

However, in general, how to choose an objective prior is a debated subject. There are many priors proposed to be used as an uninformative prior. Those selections aim to ensure some kind of objectivity.

As one issue (among many), think about $\theta \sim$ Unif(a, b). If uniformness is a measure for objectivity, $\theta \sim$ Unif(a, b) is surely an uninformative prior for θ. However, this prior leads to a non-uniform distribution for $\varphi = \log \theta$. Therefore, the objectivity is lost after transforming θ. As a remedy for that, the Jefferys prior is proposed. Jeffery's prior is defined to be proportional to the square of the determinant of the Fisher information matrix with respect to the likelihood. The Jefferys prior can be shown to be invariant under transformations of θ. But the Jefferys prior comes with its own problems, such as violating the likelihood principle.

Another approach is reference priors, which aims to maximize the effect of the likelihood in the posterior distribution, i.e., to minimize the effect of the prior. But it may be hard to derive that choice analytically.

Note that uninformative priors are to be used in the lack of expert knowledge or past data, that is when we do not know anything about the parameter (except objective things such as its natural range). When we do have a data/experience-driven knowledge, it is more appropriate to reflect it in your prior choice.

A.3.4 Improper priors

In the effort of designing an uninformative prior, one may choose a flat prior, such as $p(\theta) \propto 1$. When θ takes values on an infinite range, however, choosing $p(\theta) \propto 1$ is improper. This is because $p(\theta)$ cannot be normalised and therefore it cannot be a distribution. A prior such as that is called an *improper prior*. To give a general definition, a specification of the prior as

$$p(\theta) \propto \eta(\theta)$$

is called an improper prior if $\int_\Theta \eta(\theta) d\theta = \infty$. Examples are

- The uniform distribution on an infinite interval (i.e., $\mathbb{R}^+$ or $\mathbb{R}$).
- Beta$(0, 0)$, the beta distribution for $\alpha = 0$, $\beta = 0$ (uniform distribution on log-odds scale).
- The logarithmic prior on the positive reals, $p(\theta) \propto 1/\theta$ (i.e., flat prior for $\log \theta$).

Although improper, such choices are popularly used. The reason for such choices is to start with prior that has zero or minimal effect on the posterior distribution. The justification for using an improper prior is that the prior need not be a proper distribution to have a proper posterior distribution. Observe that, if $p(\theta) \propto \eta(\theta)$,

$$p(\theta|x) = \frac{\eta(\theta)p(x|\theta)}{\int_\Theta \eta(\theta)p(x|\theta)d\theta}$$

The above expression is a proper distribution if $\int_\Theta \eta(\theta)p(x|\theta)d\theta < \infty$. One should be cautious in using improper priors, though. The likelihood function $p(x|\theta)$ is not required to be integrable over θ, so one must ensure that $\int_\Theta \eta(\theta)p(x|\theta)d\theta < \infty$.

A.3.5 Conjugate priors

For convenience, it is common to choose a family of parametric distributions for the data likelihood. With such choices, θ in $p(x|\theta)$ becomes (some or all of the) parameters of the chosen distribution. For example, $\theta = (\mu, \sigma^2)$ may be the unknown parameters of a normal distribution from which the data $X_1, \ldots, X_n$ are assumed to be distributed, i.e.

$$p(x_{1:n}|\theta) = \prod_{i=1}^{n} \phi(x_i; \mu, \sigma^2)$$

where $\phi(\cdot; \mu, \sigma^2)$ stands for the pdf of the normal distribution $\mathcal{N}(\mu, \sigma^2)$. As another example, let $\theta = (\alpha, \beta)$ be the shape and scale parameters of the gamma distribution Gamma(α, β) and

$$p(x_{1:n}|\theta) = \prod_{i=1}^{n} \frac{e^{-\beta x_i} x_i^{\alpha-1} \beta^\alpha}{\Gamma(\alpha)}.$$

Given the family of distributions for the likelihood, it is sometimes useful to consider a certain family of distributions for the prior distribution so that the posterior distribution has the same form as the prior distribution but with different parameters, i.e. the posterior distribution is in the same family of distributions as the prior. When this is the case, the prior and posterior are then called *conjugate* distributions, and the prior is called a *conjugate prior* for the likelihood $p(x|\theta)$.

Example 4.4 (**Success probability of the Binomial distribution**). A certain coin has $P(\mathrm{T}) = \theta$ where θ is unknown. The prior distribution is $\theta \sim \mathrm{Beta}(a, b)$. The coin is tossed n times, so that if the number of times it brought a tail is X, the conditional distribution for X is $X|\theta \sim \mathrm{Binom}(n, \theta)$. We want to find the posterior distribution of θ given $X = x$ successes out of n trials.

The posterior density is proportional to

$$p(\theta|x) \propto p(\theta)p(x|\theta) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)} \frac{n!}{x!(n-x)!} \theta^x (1-\theta)^{n-x} \tag{4.6}$$

where $B(a, b) = \int u^{a-1}(1-u)^{b-1} du$.

Before continuing with deriving the expression, first note the important remark that our aim here is to *recognise the form* of the density of a parametric distribution for θ in (4.6). Therefore, we can get rid of any multiplicative term that does not depend on θ. That is why we could start with the joint density as $p(\theta|x) \propto p(\theta, x)$. In that way get

$$p(\theta|x) \propto \theta^{a+x-1}(1-\theta)^{b+n-x-1}$$

Since we observe that this has the form of a Beta distribution, we can conclude that the posterior distribution *has to be* a beta distribution

$$\theta|X = x \sim \mathrm{Beta}(a_{\mathrm{post}}, b_{\mathrm{post}})$$

where, from similarity with the prior distribution, we conclude that $a_{\mathrm{post}} = a + x$ and $b_{\mathrm{post}} = b + n - x$.

Example 4.5 (**Mean parameter of the normal distribution**). It is believed that $X_{1:n} = x_{1:n}$ are samples from a normal distribution with unknown μ and known variance σ^2. We want to estimate μ from $x_{1:n}$. The prior for $\theta = \mu$ is chosen as $\mathcal{N}(0, \kappa_0^2)$, the conjugate prior of the normal likelihood for the mean parameter. The joint density can be written

as

$$\begin{aligned}
p(\mu|x) \propto p(\mu, x) &= p(\mu)p(x|\mu) \\
&= \frac{1}{\sqrt{2\pi\kappa_0^2}} \exp\left\{-\frac{1}{2\kappa_0^2}\mu^2\right\} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x_i-\mu)^2\right\} \\
&\propto \exp\left\{-\frac{1}{2\kappa_0^2}\mu^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\} \\
&= \exp\left\{-\frac{1}{2\kappa_0^2}\mu^2 - \frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}x_i^2 + n\mu^2 - 2\mu\sum_{i=1}^{n}x_i\right)\right\} \\
&\propto \exp\left\{-\frac{1}{2\kappa_0^2}\mu^2 - \frac{1}{2\sigma^2}\left(n\mu^2 - 2\mu\sum_{i=1}^{n}x_i\right)\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[\mu^2\left(\frac{1}{\kappa_0^2}+\frac{n}{\sigma^2}\right) - 2\mu\frac{1}{\sigma^2}\sum_{i=1}^{n}x_i\right]\right\}
\end{aligned}$$

Since we observe that this has the form of a normal distribution, we can conclude that the posterior distribution *has to be* normal

$$\mu|X_{1:n} = x_{1:n} \sim \mathcal{N}(m_{\text{post}}, \kappa_{\text{post}}^2)$$

for some m_{post} and κ_{post}^2. In order to find m_{post} and σ_{post}^2, compare the expression above with $\phi(\mu; m, \kappa^2) \propto \exp\left\{-\frac{1}{2}\left[\mu^2\frac{1}{\kappa^2} - 2\mu\frac{m}{\kappa^2} + \frac{m^2}{\kappa^2}\right]\right\}$. Therefore, we must have

$$\kappa_{\text{post}}^2 = \left(\frac{1}{\kappa_0^2}+\frac{n}{\sigma^2}\right)^{-1}, \quad \frac{m_{\text{post}}}{\kappa_{\text{post}}^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}x_i \Rightarrow m_{\text{post}} = \left(\frac{1}{\kappa_0^2}+\frac{n}{\sigma^2}\right)^{-1}\frac{1}{\sigma^2}\sum_{i=1}^{n}x_i$$

Example 4.6 (**Variance of the normal distribution**). Consider the scenario in the previous example above but this time μ is known and the variance σ^2 is unknown. The prior for $\theta = \sigma^2$ is chosen as the conjugate prior of the normal likelihood for the variance parameter, i.e. the inverse gamma distribution $\mathcal{IG}(\alpha, \beta)$ with shape and scale parameters α and β, having the probability density function

$$p(\sigma^2) = \frac{\beta^\alpha}{\Gamma(\alpha)}\sigma^{-2(\alpha+1)}\exp\left(-\frac{\beta}{\sigma^2}\right).$$

The joint density can be written as

$$\begin{aligned}
p(\sigma^2|x) \propto p(\sigma^2, x) &= p(\sigma^2)p(x|\sigma^2) \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)}\sigma^{-2(\alpha+1)}\exp\left(-\frac{\beta}{\sigma^2}\right)\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}(x_i-\mu)^2\right\} \\
&\propto \sigma^{-2(\alpha+n/2+1)}\exp\left\{-\frac{\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2+\beta}{\sigma^2}\right\}
\end{aligned}$$

Comparing this expression to the prior density $p(\sigma^2)$, we observe that they have the same form and therefore,

$$\sigma^2|X_{1:n} = x_{1:n} \sim \mathcal{IG}(\alpha_{\text{post}}, \beta_{\text{post}})$$

for some α_{post} and β_{post}. From similarity, we can conclude

$$\alpha_{\text{post}} = \alpha + \frac{n}{2}, \quad \beta_{\text{post}} = \beta + \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2.$$

Example 4.7 (**Multivariate normal distribution**). Let the likelihood for X given μ is chosen as $X|\mu \sim \mathcal{N}(A\mu, R)$ and the prior for the unknown μ is chosen $\mu \sim \mathcal{N}(m, S)$. The posterior $p(\mu|x)$ is

$$\begin{aligned}
p(\mu|x) \propto p(\mu, x) =& p(\mu)p(x|\mu) \\
=& |2\pi S|^{-1/2} \exp\left\{-0.5(\mu - m)^T S^{-1}(\mu - m)\right\} \\
& |2\pi R|^{-1/2} \exp\left\{-0.5(x - A\mu)^T R^{-1}(x - A\mu)\right\} \\
\propto& \exp\left\{-0.5(\mu^T S^{-1}\mu - 2m^T S^{-1}\mu + \mu^T A^T R^{-1} A\mu - 2x^T R^{-1} A\mu)\right\} \\
=& \exp\left\{-0.5\left[\mu^T (S^{-1} + A^T R^{-1} A)\mu - 2(m^T S^{-1} + x^T R^{-1} A)\mu\right]\right\} \\
\propto& \phi(\mu; m_{\text{post}}, S_{\text{post}}) \propto \exp\left\{-0.5\left[\mu^T S_{\text{post}}^{-1}\mu - 2m_{\text{post}}^T S_{\text{post}}^{-1}\mu\right]\right\}
\end{aligned}$$

where the posterior covariance is

$$S_{\text{post}} = (S^{-1} + A^T R^{-1} A)^{-1} \tag{4.7}$$

and the posterior mean is

$$m_{\text{post}} = S_{\text{post}}(m^T S^{-1} + x^T R^{-1} A)^T = S_{\text{post}}(S^{-1} m + A^T R^{-1} x). \tag{4.8}$$

Exercise 4.4. Show the following conjugacy relations.

(a) Show that the gamma distribution is the conjugate prior of the exponential distribution, i.e. if $\theta \sim \text{Gamma}(\alpha, \beta)$ and $X|\theta \sim \text{Exp}(\theta)$ (with mean $E(X|\theta) = 1/\theta$), then $\theta|X = x \sim \text{Gamma}(\alpha_{\text{post}}, \beta_{\text{post}})$ for some α_{post} and β_{post}. Find α_{post} and β_{post} in terms of α, β, and x.

(b) Show that Dirichlet distribution is the conjugate prior of the multinomial distribution: Let $\theta = (\theta_1, \ldots, \theta_k)$ be a k-dimensional vector of probabilities such that $\theta_1 + \ldots + \theta_k = 1$ and $X = (X_1, \ldots, X_k)|\theta \sim \text{Multinomial}(\theta_1, \ldots, \theta_k)$. That is, show that, if the prior distribution is taken as $\theta \sim \text{Dirichlet}(\rho_1, \ldots, \rho_k)$ for some $\rho_1 > 0, \ldots, \rho_k > 0$, then the posterior distribution of θ can be expressed as $\theta|X = x \sim \text{Dirichlet}(\rho_{\text{post},1}, \ldots, \rho_{\text{post},k})$. Find $\rho_{\text{post},1}, \ldots, \rho_{\text{post},k}$. [The pdf of $(\theta_1, \ldots, \theta_k) \sim \text{Dirichlet}(\rho_1, \ldots, \rho_k)$ is given by

$$p(\theta_1, \ldots, \theta_k) = \frac{\Gamma\left(\sum_{i=1}^k \rho_i\right)}{\prod_{i=1}^k \Gamma(\rho_i)} \prod_{i=1}^k \theta_i^{\rho_i - 1}.$$

with the support $\{\theta_1 + \ldots + \theta_k = 1\}$.]

Computing the evidence: We saw that when conjugate priors are used for the prior, then $p(\theta)$ and $p(\theta|x)$ belong to the same family, i.e. their pdf/pmf have the same form. This is nice: since we know $p(\theta)$, $p(x|\theta)$, and $p(\theta|x)$ exactly, we can compute the evidence $p(x)$ for a given x as

$$p(x) = \frac{p(\theta, x)}{p(\theta|x)} = \frac{p(\theta)p(x|\theta)}{p(\theta|x)}$$

Example 4.8 (**Success probability of the Binomial distribution - ctd**). Consider the setting in Example 4.4. Since we know $p(\theta|x)$ and $p(\theta, x)$ exactly, the evidence $p(x)$ can be found as

$$p(x) = \frac{\frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)} \frac{n!}{x!(n-x)!}\theta^x(1-\theta)^{n-x}}{\frac{\theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1}}{B(\alpha+x,\beta+n-x)}} \tag{4.9}$$

$$= \frac{n!}{x!(n-x)!}\frac{B(\alpha+x, \beta+n-x)}{B(\alpha,\beta)}. \tag{4.10}$$

which is the pmf, evaluated at x, of the Beta-Binomial distribution with trial parameter n and shape parameters α and β.

B Quantities of interest in Bayesian inference

In Bayesian statistics, the ultimate goal is the posterior distribution $p(\theta|x)$ of the unknown variable given the available data $X = x$. There are several quantities one might be interested in; all of these quantities are rooted in $p(\theta|x)$. The following are some examples of such quantities.[1]

B.1 Posterior mean and median

If we want to have a point estimate about θ based on the posterior distribution $p(\theta|x)$, one quantity we can look at is the mean posterior

$$E(\theta|X = x) = \int p(\theta|x)\theta d\theta$$

Other than being an intuitive choice, $E(\theta|X)$, as a random function of X, is justified in the frequentist setting as well, because $E(\theta|X)$ minimises the expected mean-squared error

$$\text{MSE} = E\left([\theta - \hat{\theta}(X)]^2\right) = \int (\theta - \hat{\theta}(x))^2 p(\theta, x) d\theta dx$$

where $\hat{\theta}(X)$ is the estimator for θ and the expectation is taken with respect to the joint distribution of θ, X.

[1] It will be assumed that the random variables involved in the discussion have pdf, so integrals will be used for some definitions. Note that the discussion can be extended to discrete variables as well, by simply replacing integrals with summations.

Theorem 4.1. $\hat{\theta}(X) = E(\theta|X)$ *minimises MSE.*

In general, if we want to estimate a function $\varphi(\theta)$ of θ given $X = x$, we can target the posterior mean of φ

$$E(\varphi(\theta)|X = x) = \int p(\theta|x)\varphi(\theta)d\theta,$$

which minimises the expected mean-squared error for $\varphi(\theta)$

$$E\left([\varphi(\theta) - \hat{\varphi}(X)]^2\right) = \int (\varphi(\theta) - \hat{\varphi}(x))^2 p(\theta, x)d\theta dx.$$

Exercise 4.5. Prove Theorem 4.1 [Hint: write the estimator as $\hat{\theta}(X) = E(\theta|X) + (\hat{\theta}(X) - E(\theta|X))$ and consider conditional expectation of the MSE given $X = x$ first. You should conclude that for any x, $\hat{\theta}(x) - E(\theta|X = x)$ should be zero.]

Although it has nice statistical properties as mentioned above, the posterior mean may not always be a good choice. For example, suppose the posterior is a mixture of normal distributions with pdf $p(\theta|x) = 0.8\phi(\theta; -10, 1) + 0.2\phi(\theta; 40, 1)$. The posterior mean is 0 but density of $p(\theta|x)$ at 0 is almost 0 and the distribution has almost no mass around 0!

An alternative to the posterior mean can be the posterior median. The median for any probability distribution P is defined as any point c (which may not be unique) such that

$$P(\theta \leq c) \geq 1/2, \quad \text{and} \quad P(\theta \geq c) \geq 1/2.$$

While the posterior mean minimizes MSE, the posterior median minimizes the mean absolute error (MAE). For a given estimator $\hat{\theta}(X)$, MAE is defined as

$$\text{MAE} = E\left(\left|\theta - \hat{\theta}(X)\right|\right) = \int \left|\theta - \hat{\theta}(x)\right| p(\theta, x)d\theta dx.$$

B.2 Maximum a posteriori estimation

Another point estimate that is derived from the posterior is the maximum a posteriori (MAP) estimate which is the maximising argument of $p(\theta|x)$

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta\in\Theta} p(\theta|x) = \arg\max_{\theta\in\Theta} p(\theta, x).$$

Note that this procedure is different from maximum likelihood estimation (MLE), which yields the maximising argument of the likelihood

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta\in\Theta} p(x|\theta),$$

since, in the MAP estimation, there is the additional factor due to prior $p(\theta)$.

B.3 Posterior predictive distribution

Assume we are interested in the distribution that a new data point X_{n+1} would have, given a set of n existing observations $X_{1:n} = x_{1:n}$. In a frequentist context, this might be derived by computing the maximum likelihood estimate $\hat{\theta}_{\text{MLE}}$ (or some other point estimate) of θ given $x_{1:n}$, and then plugging it into the distribution function of the new observation X_{n+1} so that the predictive distribution is $p(x_{n+1}|\hat{\theta}_{\text{MLE}})$.

In a Bayesian context, the natural answer to this is the posterior predictive distribution, which is the distribution of unobserved observations (prediction) conditional on the observed data $p(x_{n+1}|x_{1:n})$. To find the posterior predictive distribution, we make use of the entire posterior distribution of the parameter(s) given the observed data to yield a probability distribution rather than simply a point estimate. Specifically, we compute $p(x_{n+1}|x_{1:n})$ by marginalising over the unknown variable θ, using its posterior distribution:

$$\begin{aligned} p(x_{n+1}|x_{1:n}) &= \int p(x_{n+1}, \theta|x_{1:n})d\theta \\ &= \int p(x_{n+1}|\theta, x_{1:n})p(\theta|x_{1:n})d\theta \end{aligned}$$

In many cases, X_{n+1} is independent from $X_{1:n}$ given θ. This happens, for example, when $\{X_i\}_{i\geq 1}$ are i.i.d. given θ, that is $X_i|\theta \sim p(x|\theta)$, $i \geq 1$. In that case, the density above reduces to

$$p(x_{n+1}|x_{1:n}) = \int p(x_{n+1}|\theta)p(\theta|x_{1:n})d\theta$$

Note that this is equivalent to the expected value of the distribution of the new data point when the expectation is taken over the posterior distribution of θ, i.e.:

$$p(x_{n+1}|x_{1:n}) = E_{\theta|x_{1:n}}[p(x_{n+1}|\theta)|X_{1:n} = x_{1:n}].$$

Conjugate priors and posterior predictive density: We saw that when conjugate priors are used for the prior, then $p(\theta)$ and $p(\theta|x)$ belong to the same family, i.e. their pdf/pmf have the same form. This implies that, when X_i's are i.i.d. conditional on θ, the posterior predictive density $p(x_{n+1}|x_{1:n})$ has the same form as the marginal density of a single sample

$$p(x) = \int p(\theta)p(x|\theta)d\theta.$$

Example 4.9 (**Success probability of the Binomial distribution - ctd**). Consider the setting in Example 4.4. Given the prior $\theta \sim \text{Beta}(\alpha, \beta)$ and $X = x$ successes out of n trials, what is the probability of having $Z = z$ successes out of the next m trials?

Here Z is the next sample that is to be predicted. We can employ the posterior predictive probability for Z. We know from the derivation of Example 4.8 that Z will be distributed from the Beta-Binomial distribution with parameters m (trials), $\alpha' = \alpha + x$

and $\beta' = \beta + n - x$ since the prior and the posterior of θ are in the same form and Z given θ and X given θ are both Binomial.

$$p_{Z|X}(z|x) = \frac{m!}{z!(m-z)!} \frac{B(\alpha' + z, \beta' + m - z)}{B(\alpha', \beta')}.$$

Exercise 4.6. Suppose we observe a noisy sinusoid with period T and unknown amplitude θ for n steps: $X|\theta \sim \mathcal{N}(x_t; f(\theta, t), \sigma_x^2)$, for $t = 1, \ldots, n$ where $f(t; \theta) = \theta \sin(2\pi t/T)$ is the sinusoid. The prior for the amplitude is Gaussian: $\theta \sim \mathcal{N}(0, \sigma_\theta^2)$.

1. Find $p(\theta|x_{1:n})$ and $p(x_{1:n})$.
2. What is distribution of $f(n+1, \theta)$ given $X_{1:n} = x_{1:n}$?
3. Find $p(x_{n+1})$ and $p(x_{n+1}|x_{1:n})$. Compare their variances. What can you comment on the difference between the variances?
4. Data given in `sinusoid.txt` are generated with period $T = 40$, $\sigma_x^2 = 1$.
   (a) Calculate the parameters of $p(\theta|x_{1:n})$
   (b) Plot $p(x_{n+1})$ and $p(x_{n+1}|x_{1:n})$ on the same axis. Take $\sigma_\theta^2 = 100$ as the prior parameter.

B.4 Credible Intervals

In Bayesian statistics, $100(1-\alpha)\%$ credible interval is an interval within which an unknown random variable falls with probability $1 - \alpha$. A credible interval for the parameter θ is an interval in the domain of a posterior probability distribution $p(\theta|x)$. Given $X = x$, the interval $(L(x), U(x))$ is a $100(1-\alpha)\%$ credible interval if it satisfies

$$P(L(x) < \theta < U(x)|X = x) = \int_{L(x)}^{U(x)} p(\theta|x) d\theta = 1 - \alpha. \tag{4.11}$$

Note that, to be most general, 'interval' is a loose term, since a suitable 'credible interval' may also be a union of intervals, especially for a multimodal posterior distribution.

Exercise 4.7. Consider the problem in Example 4.5, where we have n observations from a normal population with known variance σ^2 and unknown mean μ with a prior $\mu \sim \mathcal{N}(0, \kappa_0^2)$. Find a 95% credible interval for μ.

B.4.1 Credible intervals and confidence intervals

There is a notable analogy between credible intervals in Bayesian statistics and confidence intervals in frequentist statistics. Both provide an interval for the unknown parameter. However, they differ on a philosophical basis.

For a credible interval, the data, hence the confidence interval itself is considered fixed but θ is random. In other words, the probabilistic statement in (4.11) involves the distribution of θ given x. A 95% credible interval means that conditional the specific data $X = x$, θ is in the credible interval with 0.95 probability.

In contrast, in frequentist statistics, parameter θ is treated as a fixed value and the bounds of the confidence intervals are treated as random. The probabilistic statement that θ is in a confidence interval with probability $1 - \alpha$ is with respect to the population distribution, i.e., the distribution of X. A frequentist 95% confidence interval means that with a large number of repeated samples, around 95% of such calculated confidence intervals would include the true value of θ.

A second major difference is that a credible interval involves a prior distribution whereas a confidence interval does not. For that reason, Bayesian credible intervals can be quite different from frequentist confidence intervals.

B.4.2 Choosing a credible interval

Just like confidence intervals, credible intervals are not unique on a posterior distribution. There are several approaches for defining a suitable credible interval. We will list the most popular three of them.

- One way is to choose the narrowest possible interval. For a unimodal distribution, this corresponds to the highest posterior density interval, that is, the interval with the highest probability density. Such an interval can be formulated as $(L(x), U(x))$ such that for any $\theta \in (L(x), U(x))$ and $\theta' \notin (L(x), U(x))$, we have $p(\theta'|x) \leq p(\theta|x)$. For a unimodal distribution, this interval includes the mode of the posterior distribution, that is, the maximum a posteriori estimate for θ.

- Another way is to choose the credible interval such that the probability of being below the interval is as likely as being above it. If $(L(x), U(x))$ is a $100(1 - \alpha)\%$ credible interval constructed in this way, it satisfies

$$P(\theta \leq L(x)|X = x) = P(\theta \geq U(x)|X = x) = \alpha/2.$$

  This interval will include the median. This is sometimes called the equal-tailed interval.

- Assuming that the mean exists, we can choose the interval for which the mean is the central point.

The generalisation of a credible interval to multivariate random variables is the credible region. More concretely, $R \subseteq \Theta$ is a $100(1 - \alpha)\%$ confidence region if it satisfies

$$P(\theta \in R|X = x) = \int_R p(\theta|x)d\theta = 1 - \alpha.$$

Exercise 4.8. Consider the problem in Example 4.7, where $X|\mu \sim \mathcal{N}(A\mu, R)$ with known R and unknown μ with prior $\mu \sim \mathcal{N}(m, S)$. The posterior distribution of μ given $X = x$ was shown to be $\mathcal{N}(\mu_{\text{post}}, S_{\text{post}})$, with the moments given in (4.8) and (4.7). Show that the region

$$\left\{\mu : (\mu - \mu_{\text{post}})^T S_{\text{post}}^{-1}(\mu - \mu_{\text{post}}) \leq \chi^2_{\alpha,d}\right\}$$

is a $100(1-\alpha)\%$ credible region for μ, where d is the dimension of μ.

It is also possible to talk about credible intervals regarding the posterior predictive distribution as well. Let, $X = x$ be given and X_0 is a future observation that is to be predicted. The conditional distribution of X_0 given $X = x$ is constructed through the posterior distribution as

$$p(x_0|x) = \int_{\theta\in\Theta} p(\theta|x)p(x_0|\theta, x)d\theta.$$

A credible interval for X_0 is an interval in the domain of $p(x_0|x)$. Given $X = x$, a $100(1-\alpha)\%$ credible interval for X_0 is any $(L(x), U(x))$ that satisfies

$$P(L(x) < X_0 < U(x)|X = x) = \int_{L(x)}^{U(x)} p(x_0|x)dx_0 = 1 - \alpha.$$

C Sampling from posterior distributions

A closed-form for the posterior distribution can be obtained if the joint density $p(\theta)p(x|\theta)$ is proportional to the density of a known distribution. It might have already occurred to you that this may not always be the case, especially when θ is multidimensional or the prior distribution is chosen as something other than a conjugate prior.

Posterior distributions that cannot be expressed in a closed-form are called intractable, in a certain sense of the word. Intractable posterior distributions are often encountered in Bayesian statistics.

Example 4.10. Assume we have $X_{1:n} \sim \mathcal{N}(\mu, \sigma^2)$. In the examples we covered previously regarding the normal distribution, either μ or σ^2 were known, and we were able to come up with closed form posterior distributions for the other parameter, thanks to a conjugate prior. When both μ and σ^2 are unknown, we no more have a closed form expression for their joint posterior distribution $p(\mu, \sigma^2|x_{1:n})$.

Example 4.11. This is a simple example that illustrates the source localisation problem. We have a source (or target) on the 2-D plane whose unknown location

$$\theta = (\theta_1, \theta_2) \in \mathbb{R}^2$$

we wish to find. We collect distance measurements for the source using three sensors, located at positions s_1, s_2, and s_3, see Figure 4.1. The measured distances $X = (X_1, X_2, X_3)$,

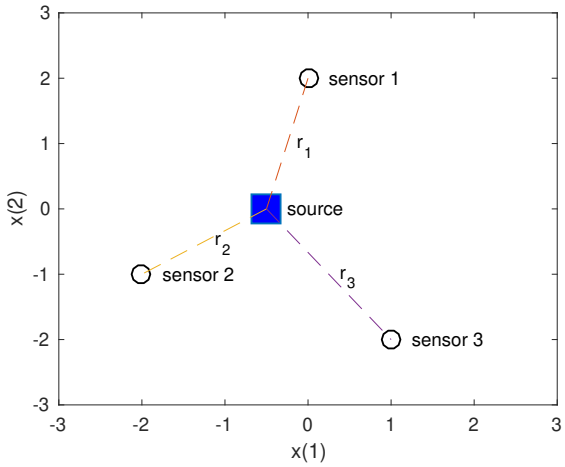

Figure 4.1: Source localisation problem with three sensors and one source

however, are noisy with independent normally distributed noises with equal variance:[2]

$$X_i|\theta \sim \mathcal{N}(||\theta - s_i||, \sigma_x^2), \quad i = 1, 2, 3,$$

where $||\cdot||$ denotes the Euclidean distance. Letting $r_i = ||\theta - s_i||$, the likelihood evaluated at $x = (x_1, x_2, x_3)$ given θ can be written as

$$p(x|\theta) = \prod_{i=1}^{3} \phi(x_i; r_i, \sigma_x^2) \tag{4.12}$$

We do not know much *a priori* information about θ, therefore we take the prior distribution θ as the bivariate normal distribution with zero mean vector and a large diagonal covariance matrix, $\theta \sim \mathcal{N}(0_2, \sigma_\theta^2 I_2)$, so that the density is

$$p(\theta) = \phi(\theta_1; 0, \sigma_\theta^2)\phi(\theta_2; 0, \sigma_\theta^2). \tag{4.13}$$

See Figure 4.2 for an illustration of prior, likelihood, and posterior densities for this problem.

Given noisy measurements, $X = x = (x_1, x_2, x_3)$, we want to locate θ, so we are interested in the posterior distribution

$$p(\theta|x) \propto p(\theta)p(x|\theta).$$

Due to the non-linearities in the likelihood, this posterior distribution is intractable.

[2]In this way we allow negative distances, which makes the normal distribution not the most proper choice. However, for the sake of ease with computations, we overlook that in this example.

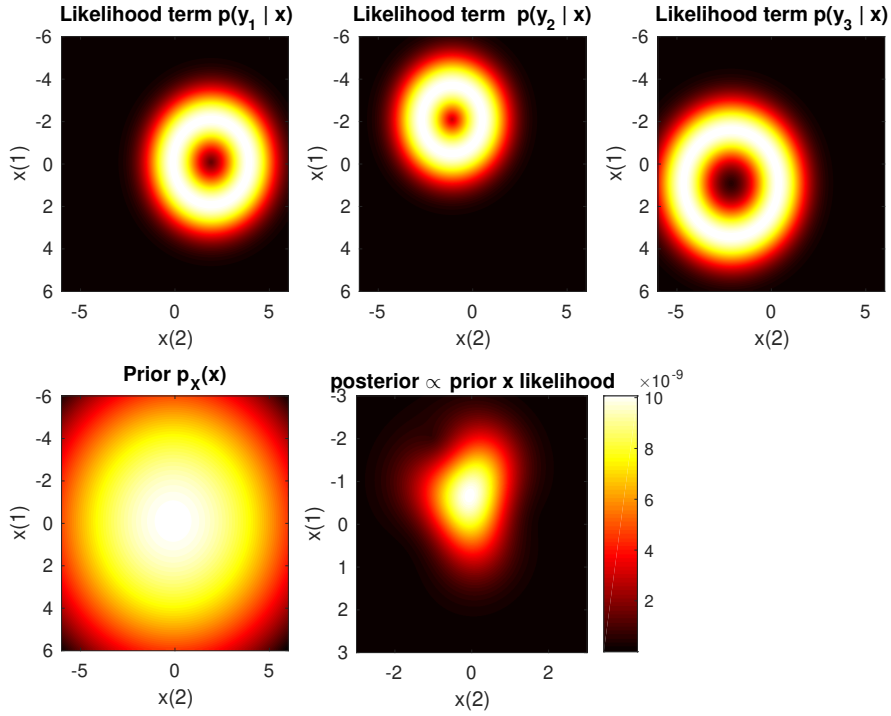

Figure 4.2: Source localisation problem with three sensors and one source: The likelihood terms, prior, and the posterior. The parameters and the variables are $s_1 = (0, 2)$, $s_2 = (-2, -1)$, $s_3 = (1, -2)$, $x_1 = 2$, $x_2 = 1.6$, $x_3 = 2.5$, $\sigma_\theta^2 = 100$, and $\sigma_x^2 = 1$

Example 4.12 (Logistic regression model). In the previous chapter, we studied the linear regression model in detail. Here, we introduce the logistic regression model, a very popular regression model with binary responses, where the regression function is non-linear in its parameters. As before, we have n independent response variables Y_i, $i = 1, \ldots, n$, and for each response variable we have k predictors, $x_{i1}, \ldots, x_{ik}$. Differently than the linear regression model, response variable Y_i is a binary (Bernoulli) variable with a success probability p_i. The logistic regression model assumes a linear relationship between the predictor variables and the log-odds $\log(p_i/(1 - p_i))$ of the event $Y_i = 1$. This linear relationship can be written in the following mathematical form:

$$\log \frac{p_i}{1 - p_i} = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_k x_{i,k}$$

We can rewrite this relation by emphasising on the conditional probability of the response variable as

$$\begin{aligned} p_i = P(Y_i = 1 | x_{i,1}, \ldots, x_{i,k}, \beta) &= \frac{\exp(\beta_0 + \beta_1 x_{i,1} + \ldots + \beta_k x_{i,k})}{1 + \exp(\beta_0 + \beta_1 x_{i,1} + \ldots + \beta_k x_{i,k})} \\ &= \frac{1}{1 + \exp\{-(\beta_0 + \beta_1 x_{i,1} + \ldots + \beta_d x_{i,d})\}}. \end{aligned}$$

The problem of estimating β given a $Y_{1:n} = y_{1:n}$ and $X = x_{1:n,1:k}$ can be formulated in terms of a posterior distribution once we have a prior distribution for β.

$$p(\beta|X, y) \propto \underbrace{p(\beta)}_{\text{prior}} \underbrace{\prod_{i=1}^{n} P(Y_i = y_i|x_{i,1}, \ldots, x_{i,k}, \beta)}_{\text{likelihood}}$$

This posterior distribution does not have a closed-form expression for any reasonable choice for the prior distribution.

The desire to identify the posterior distribution in such difficult cases has led to the development of several methods, many of which can be gathered under the family of *Monte Carlo* methods. While Monte Carlo methods can be used to draw samples from any complex distribution in general; they are most exploited for Bayesian statistics. The idea is to generate samples from the posterior distribution

$$\theta^{(1)}, \ldots, \theta^{(N)} \sim p(\theta|x) \tag{4.14}$$

and use those samples to have an idea about $p(\theta|x)$ (for example, by making a histogram out of those samples), or estimate the expectation of a function $\varphi(\theta)$ of interest with respect to the posterior distribution

$$E(\varphi(\theta)|X = x) = \int_{\theta \in \Theta} p(\theta|x)\varphi(\theta)d\theta.$$

When the samples in (4.14) are provided by some Monte Carlo method, the approximation to the posterior expectation above is

$$E(\varphi(\theta)|X = x) \approx \frac{1}{N} \sum_{i=1}^{N} \varphi(\theta^{(i)}).$$

The variance of the approximation above decreases with N, hence the more samples the better. The main challenge here is the sampling step: How do we generate samples from $p(\theta|x)$ even when it is not in a closed-form? This question is central to Monte Carlo.

In the following, we will make a short review of Monte Carlo methods. We start with an exact sampling method named rejection sampling. We will next discuss the importance sampling method for estimation of posterior expectations with weighted samples. Finally, a larger emphasis will be put on Markov chain Monte Carlo (MCMC) methods, which are the most popular methods among the ones we will have covered.

Remark 4.1 (**Change of notation**). For brevity in the notation, we will denote the distribution we want to sample from by $\pi(\theta)$. When Bayesian inference is concerned, $\pi(\theta) = p(\theta|x)$. Recall that such shorthand notation is suitable since in Bayesian statistics x is fixed.

C.1 Rejection sampling

The method of rejection sampling is used to generate exact i.i.d. samples from the desired distribution, which in our case is $p(\theta|x)$. There are already several methods for i.i.d. sampling. Those methods only rely solely on being able to sample from the uniform distribution Unif$(0, 1)$. Some of those methods are the method of inversion, transformation, and composition; for the interested reader, a brief introduction to those methods is provided in Appendix C. Those methods, however, are typically suitable for well-known distributions that have closed-form expressions. The reason we are particularly interested in rejection sampling is that it is suitable for distributions that are not in a closed form, as we typically encounter in Bayesian statistics as $\pi(\theta) = p(\theta|x) \propto p(\theta)p(x|\theta)$.

Rejection sampling is available when there exists an instrumental distribution with density $q(\theta)$ such that

- $q(\theta) > 0$ whenever $\pi(\theta) > 0$, and
- There exists $M > 0$ such that $\pi(\theta) \leq Mq(\theta)$ for all $\theta \in \Theta$.

The rejection sampling method for obtaining one sample from π can be implemented with any $q(\theta)$ and $M > 0$ that satisfy the conditions above as in Algorithm 4.1.

Algorithm 4.1: Rejection sampling

1 Generate $\theta' \sim q(\theta')$ and $U \sim$ Unif$(0, 1)$.
2 If $U \leq \frac{\pi(\theta')}{Mq(\theta')}$, accept $\theta = \theta'$; else go to 1.

How quickly do we obtain a sample with this method? Noting that the pdf of θ' is $q(\theta')$, the acceptance probability can be derived as

$$\begin{aligned} P(\text{Accept}) &= \int P(\text{Accept}|\theta)q(\theta)d\theta \\ &= \int \frac{\pi(\theta)}{Mq(\theta)}q(\theta)d\theta \\ &= \frac{1}{M}\int \pi(\theta)d\theta \\ &= \frac{1}{M}, \end{aligned} \tag{4.15}$$

which is also the long term proportion of the number accepted samples over the number of trials. Therefore, taking $q(\theta)$ as close to $\pi(\theta)$ as possible to avoid large $\pi(\theta)/q(\theta)$ ratios and taking $M = \sup_\theta \pi(\theta)/q(\theta)$ are sensible choices to make the acceptance probability $P(\text{Accept})$ as high as possible.

The validity of the rejection sampling method can be verified by considering the distribution of the accepted samples. Using Bayes' theorem,

$$p(\theta|\text{Accept}) = \frac{q(\theta)P(\text{Accept}|\theta)}{P(\text{Accept})} = \frac{q(\theta)\frac{1}{M}\frac{\pi(\theta)}{q(\theta)}}{1/M} = \pi(\theta). \tag{4.16}$$

C.1.1 When $\pi(\theta)$ is known up to a normalising constant

One advantage of rejection sampling is that we can implement it even when we know $\pi(\theta)$ only up to some proportionality constants Z_π, that is, when

$$\pi(\theta) = \frac{\widehat{\pi}(\theta)}{Z_\pi}, \quad Z_\pi = \int \widehat{\pi}(\theta) d\theta \tag{4.17}$$

It is easy to check that one can perform the rejection sampling method as in Algorithm 4.2 for any M such that $\widehat{\pi}(\theta) \leq Mq(\theta)$ for all $\theta \in \Theta$.

Algorithm 4.2: Rejection sampling with unnormalised densities

1 Generate $\theta' \sim q(\theta')$ and $u \sim \text{Unif}(0,1)$.

2 If $u \leq \frac{\widehat{\pi}(\theta')}{Mq(\theta')}$, accept $\theta = \theta'$; else go to 1.

Justification of Algorithm 4.2 would follow from similar steps to those in (4.16). Also, in that case, the acceptance probability would be Z_π/M.

Exercise 4.9. Show that the modified rejection sampling method described in Section C.1.1 for unnormalised densities is valid, i.e. the accepted sample $\theta \sim \pi(\theta)$, and it has the acceptance probability Z_π/M as claimed. The derivation is similar to those in (4.15), (4.16).

The unknown normalising constant issue mostly arises in Bayesian inference when we want to sample from a posterior distribution. The posterior density of θ given $X = x$ is proportional to

$$p(\theta|x) \propto p(\theta)p(x|\theta) \tag{4.18}$$

where the normalising constant $p(x) = \int p(\theta)p(x|\theta)d\theta$ is usually intractable. Suppose we want to sample from $p(\theta|x)$. When $p(\theta|x)$ is not the density of a well known distribution, we may be able to use rejection sampling. If we can find $M > 0$ such that $p(x|\theta) \leq M$ for all $\theta \in \Theta$, and the prior distribution with density $p(\theta)$ is easy to sample from, then we can use rejection sampling with $q(\theta) = p(\theta)$.

1. Sample $\theta' \sim p(\theta')$ and $u \sim \text{Unif}(0,1)$,
2. If $u \leq p(x|\theta')/M$, accept $\theta = \theta'$; otherwise go to step 1.

Exercise 4.10. Consider the example in Example 4.11. Write a function that takes the noisy measurements $x = (x_1, x_2, x_3)$, positions of the sensors s_1, s_2, s_3, the prior and likelihood variances σ_θ^2 and σ_x^2, and the number of samples N as inputs, implements rejection sampling to draw i.i.d. samples from the posterior $p(\theta|x)$. Try your code with $s_1 = (0,2)$, $s_2 = (-2,-1)$, $s_3 = (1,-2)$, $x_1 = 2$, $x_2 = 1.6$, $x_3 = 2.5$, $\sigma_\theta^2 = 100$, and $\sigma_x^2 = 1$ which are the values used to generate the plots in Figure 4.2.

C.2 Importance sampling

Consider the expectation with respect to the target distribution

$$E_\pi(\varphi(\theta)) = \int_\Theta \varphi(\theta)\pi(\theta)d\theta.$$

In order to estimate the expectation by a Monte Carlo *plug-in estimator*

$$\frac{1}{N}\sum_{i=1}^{N}\varphi(\theta^{(i)}), \tag{4.19}$$

we need i.i.d. samples from $\pi(\theta)$ and in the previous chapter we covered some exact sampling methods for generating $\theta^{(i)} \sim \pi$, $i = 1, \ldots, N$.

However, there are many cases where $\theta \sim \pi$ is either impossible or too difficult, or wasteful. For example, rejection sampling uses only about $1/M$ of generated random samples to construct an approximation to π. To generate N samples, we need on average NM iterations of rejection sampling. The number M can be very large, especially in high dimensions, and rejection sampling may be wasteful.

In contrast to rejection sampling, *importance sampling* uses every sample but weights each one according to the degree of similarity between the target and instrumental distributions. We describe the importance sampling method assuming that $\pi(\theta)$ is a probability density function - the discrete version should be easy to figure out afterwards.

Suppose there exists a distribution with density $q(\theta)$ such that $q(\theta) > 0$ whenever $\pi(\theta) > 0$. Given $\pi(\theta)$ and $q(\theta)$, define the weight function $w : \Theta \to \mathbb{R}$

$$w(\theta) := \begin{cases} \pi(\theta)/q(\theta), & q(\theta) > 0, \\ 0 & q(\theta) = 0. \end{cases} \tag{4.20}$$

The idea of importance sampling follows from the *importance sampling fundamental identity*: We can rewrite the expectation as

$$\begin{aligned} E_\pi(\varphi(\theta)) &= \int_\theta \varphi(\theta)\pi(\theta)d\theta \\ &= \int_\Theta \varphi(\theta)\frac{\pi(\theta)}{q(\theta)}q(\theta)d\theta \\ &= \int_\Theta \varphi(\theta)w(\theta)q(\theta)d\theta \\ &= E_q(\varphi(\theta)w(\theta)), \end{aligned}$$

This identity can be used with a $q(\theta)$ which is easy to sample from, which leads to importance sampling given in Algorithm 4.3

The weights $w(\theta^{(i)})$ are known as the *importance sampling weights*. Note that (4.21) is another plug-in estimator for the same expectation, but constructed with a different distribution and function. Therefore the estimator in (4.21) is an unbiased estimator of $E_\pi(\varphi(\theta))$.

Algorithm 4.3: Importance sampling

1 **for** $i = 1, \ldots, N$ **do**
2 Sample $\theta^{(i)} \sim q(\theta)$, and calculate $w(\theta^{(i)})$ according to (4.20).
3 Calculate the approximation of the expectation $E_\pi(\varphi(\theta))$ as

$$\frac{1}{N}\sum_{i=1}^{N}\varphi(\theta^{(i)})w(\theta^{(i)}). \tag{4.21}$$

Example 4.13. Suppose we have a joint pdf $p(\theta, x)$ written as

$$p(\theta, x) = p(\theta)p(x|\theta)$$

In the Bayesian framework where θ is the unknown parameter and x is the observed variable (or data), the prior $p(\theta)$ is usually easy to sample from, and the likelihood $p(x|\theta)$ is easy to compute.

In certain applications, we want to compute the *evidence* $p(x)$ at a given value x of the data. We can write $p(x)$ as

$$\begin{aligned} p(x) &= \int_\Theta p(\theta, x)d\theta \\ &= \int_\Theta p(\theta)p(x|\theta)d\theta \\ &= E_{p(\theta)}(p(x|\theta)) \end{aligned}$$

where the last line highlights the crucial observation that given x, the likelihood can be thought as a function of θ, that is, $\varphi(\theta) = p(x|\theta)$, and $p(x)$ can be written as an expectation of $\varphi(\theta)$ with respect to the prior $p(\theta)$. Therefore, $p(x)$ can be estimated using a plug-in estimator where we sample $\theta^{(1)}, \ldots, \theta^{(N)} \sim p(\theta)$ and estimate $p(x)$ as

$$p(x) \approx \frac{1}{N}\sum_{i=1}^{N} p(x|\theta^{(i)}), \quad \theta^{(1)}, \ldots, \theta^{(N)} \sim p(\theta).$$

However, we do not have to sample from $p(\theta)$. We can use importance sampling with an importance density $q(\theta)$.

$$p(x) \approx \frac{1}{N}\sum_{i=1}^{N} \frac{p(\theta^{(i)})}{q(\theta^{(i)})} p(x|\theta^{(i)}), \quad \theta^{(1)}, \ldots, \theta^{(N)} \sim q(\theta).$$

Being able to approximate a marginal distribution as in $p(\theta)$ has an important role in Bayesian model selection, where hypotheses regarding the distributions involved in the statistical model are compared.

C.2.1 Self-normalised importance sampling

Like rejection sampling, the importance sampling method can be modified for the cases when $\pi(\theta) = \frac{\widehat{\pi}(\theta)}{Z_\pi}$ and we only have $\widehat{\pi}(\theta)$. This time, letting

$$w(\theta) := \begin{cases} \frac{\widehat{\pi}(\theta)}{q(\theta)}, & q(\theta) > 0 \\ 0, & q(\theta) = 0, \end{cases}$$

observe that

$$\begin{aligned} E_q(w(\theta)) &= \int \frac{\widehat{\pi}(\theta)}{q(\theta)} q(\theta) d\theta \\ &= \int \frac{\pi(\theta) Z_p}{q(\theta)} q(\theta) d\theta \\ &= Z_\pi. \end{aligned}$$

and

$$\begin{aligned} E_q(w(\theta)\varphi(\theta)) &= \int \frac{\widehat{\pi}(\theta)}{q(\theta)} \varphi(\theta) q(\theta) d\theta \\ &= \int \frac{\pi(\theta) Z_\pi}{q(\theta)} \varphi(\theta) q(\theta) d\theta \\ &= E_\pi(\varphi(\theta)) Z_\pi. \end{aligned}$$

Therefore, we can write the fundamental identity of importance sampling in terms of $\widehat{\pi}$ as

$$E_\pi(\varphi(\theta)) = \frac{E_q(w(\theta)\varphi(\theta))}{E_q(w(\theta))}.$$

The importance sampling method can be modified to approximate both the nominator, the unnormalised estimate, and the denominator, the normalisation constant, by using Monte Carlo. Sampling $\theta^{(1)}, \ldots, \theta^{(N)}$ from q, we have the approximation

$$\frac{\frac{1}{N}\sum_{i=1}^N \varphi(\theta^{(i)}) w(\theta^{(i)})}{\frac{1}{N}\sum_{i=1}^N w(\theta^{(i)})} = \sum_{i=1}^N W^{(i)} \varphi(\theta^{(i)}). \tag{4.22}$$

where

$$W^{(i)} = \frac{w(\theta^{(i)})}{\sum_{j=1}^N w(\theta^{(j)})}$$

are called the *normalised importance weights* as they sum up to 1. The resulting method, which is called *self-normalised importance sampling* is given in Algorithm 4.4: Being the ratio of two unbiased estimators, the estimator of the self-normalised importance sampling is biased for finite N. However, its consistency and stability are provided by a strong law

Algorithm 4.4: Self-normalised importance sampling

for $i = 1, \ldots, N$ **do**
  Generate $\theta^{(i)} \sim Q$, calculate $w(\theta^{(i)}) = \frac{\widehat{\pi}(\theta^{(i)})}{\widehat{q}(\theta^{(i)})}$.
for $i = 1, \ldots, N$ **do**
  Set $W^{(i)} = \frac{w(\theta^{(i)})}{\sum_{j=1}^{N} w(\theta^{(j)})}$.
Calculate the approximation to the expectation

$$E_\pi(\varphi(\theta)) \approx \sum_{i=1}^{N} W^{(i)} \varphi(\theta^{(i)})$$

of large numbers and a central limit theorem. In the same work, the variance of the self-normalised importance sampling estimator is analysed and an approximation is provided, which reveals that self-normalised importance sampling can provide lower variance estimates than the unnormalised importance sampling method. Also, normalised importance sampling has the nice property of estimating a constant by itself, unlike the unnormalised importance sampling method. Therefore, this method can be preferable to its unnormalised version even if it is not the case that π and q are known only up to proportionality constants.

Self-normalised importance sampling is also called Bayesian importance sampling, since, in most Bayesian inference problems, the normalising constant of the posterior distribution is unknown.

Example 4.14. Let us consider a posterior distribution

$$p(\theta|x) \propto p(\theta)p(x|\theta)$$

with the unknown normalising constant is $p(x) = \int p(\theta)p(x|\theta)d\theta$. Given the data $X = x$, we want to calculate the expectation of $\varphi : \Theta \to \mathbb{R}$ with respect to $p(\theta|x)$

$$E(\varphi(\theta)|X = x) = \int p(\theta|x)\varphi(\theta)d\theta.$$

Since we know $p(\theta|x)$ only up to a proportionality constant, we use self-normalised importance sampling. With the choice of $q(\theta)$, self-normalised importance sampling becomes

1. For $i = 1, \ldots, N$; generate $\theta^{(i)} \sim q(\theta)$, calculate

$$w(\theta^{(i)}) = \frac{p(\theta^{(i)})p(x|\theta^{(i)})}{q(\theta^{(i)})}.$$

2. For $i = 1, \ldots, N$; set $W^{(i)} = \frac{w(\theta^{(i)})}{\sum_{j=1}^{N} w(\theta^{(j)})}$.

3. Approximate $E(\varphi(\theta)|X=x) \approx \sum_{i=1}^{N} W^{(i)}\varphi(\theta^{(i)})$.

If we choose $q(\theta) = p(\theta)$, i.e. the prior density, then $w(\theta) = p(x|\theta)$ reduces to the likelihood. But this is not always a good idea as we will see in the next example.

Example 4.15. Suppose we have an unknown mean parameter $\theta \in \mathbb{R}$ whose prior distribution is given by $\theta \sim \mathcal{N}(\mu, \sigma^2)$. Conditional on θ, n observations $X = (X_1, \ldots, X_n) \in \mathbb{R}^n$ are generated independently

$$X_1, \ldots, X_n|\theta \overset{\text{i.i.d.}}{\sim} \text{Unif}(\theta - a, \theta + a).$$

We want to estimate the posterior mean of θ given $X = x = (x_1, \ldots, x_n)$, i.e. $E(\theta|X = x) = \int p(\theta|x)\theta d\theta$, where

$$p(\theta|x) \propto p(\theta)p(x|\theta)$$

The prior density and likelihood are $p(\theta) = \phi(\theta; \mu, \sigma^2)$ and $p(x|\theta) = \prod_{i=1}^{n} \frac{1}{2a}\mathbb{I}_{(\theta-a,\theta+a)}(x_i)$, so the posterior distribution can be written as

$$p(\theta|x) \propto \phi(\theta; \mu, \sigma^2)\frac{1}{(2a)^n}\prod_{i=1}^{n}\mathbb{I}_{(\theta-a,\theta+a)}(x_i)$$

Densities $p(\theta)$ and $p(\theta, x)$ versus θ for a fixed $X = x = (x_1, \ldots, x_n)$ with $n = 10$ generated from the marginal distribution of X with $a = 2$, $\mu = 0$, and $\sigma^2 = 10$ are given in Figure 4.3. Note that the second plot is proportional to the posterior density.

We can use self-normalised importance sampling to estimate $E(\theta|X = x)$. The choice of the importance density is critical here: Suppose we chose $q(\theta)$ to be the prior distribution for θ, i.e. $q(\theta) = \phi(\theta; \mu, \sigma^2)$. This is a valid choice, however if a is small and σ^2 is relatively large, it is likely that the resulting weight function

$$w(\theta) = \frac{1}{(2a)^n}\prod_{i=1}^{n}\mathbb{I}_{(\theta-a,\theta+a)}(x_i).$$

will end up being zero for most of the generated samples from $q(\theta)$ and it will be $\frac{1}{(2a)^n}$ for few samples. This results in a high variance in the importance sampling estimator. What is worse, it is possible to have zero weights for all samples and hence the denominator in (4.22) can be zero. Therefore the estimator is a poor one.

Let $x_{\max} = \max_i x_i$ and $x_{\min} = \min_i x_i$. A careful inspection of $p(\theta|x)$ reveals that given $x = (x_1, \ldots, x_n)$, θ must be contained in $(x_{\max} - a, x_{\min} + a)$. In other words,

$$\theta \in (x_{\max} - a, x_{\min} + a) \Leftrightarrow \theta - a < x_i < \theta + a, \quad \forall i = 1, \ldots, n.$$

Therefore, a better importance density does not waste its time outside the interval $(x_{\max} - a, x_{\min} + a)$ and generate samples in that interval. As an example, we can choose $q(\theta)$ the density of $\text{Unif}(x_{\max} - a, x_{\min} + a)$. With that choice, the weight function will be

$$w(\theta) = \begin{cases} \frac{\phi(\theta;\mu,\sigma^2)\frac{1}{(2a)^n}}{1/(2a+x_{\min}-x_{\max})}, & \theta \in (x_{\max} - a, x_{\min} + a) \\ 0, & \text{else} \end{cases}$$

Note that since we are using the self-normalised importance sampling estimator and hence we normalise the weights $W^{(i)} = w(\theta^{(i)})/\sum_{j=1}^{N} w(\theta^{(j)})$, we do not need to calculate the constant factor $(2a + x_{\min} - x_{\max})/(2a)^n$ for the weights.

Figure 4.4 compares the importance sampling estimators with the two different importance distributions mentioned above. The histograms are generated from 10000 Monte Carlo runs (10000 independent estimates of the posterior mean) for each estimator. Observe that the estimates obtained when the importance distribution is selected the prior distribution is more widespread, exhibiting a higher variance.

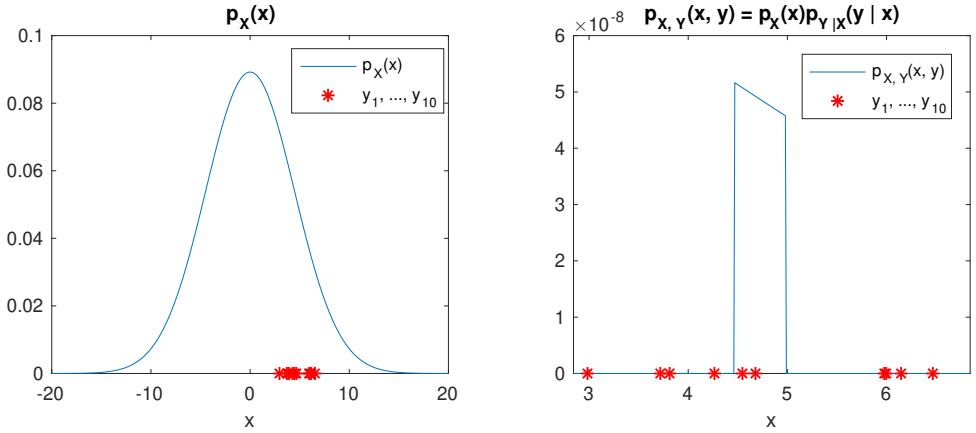

Figure 4.3: $p(\theta)$ and $p(\theta, x)$ vs θ for the problem in Example 4.15 with $n = 10$ and $a = 2$

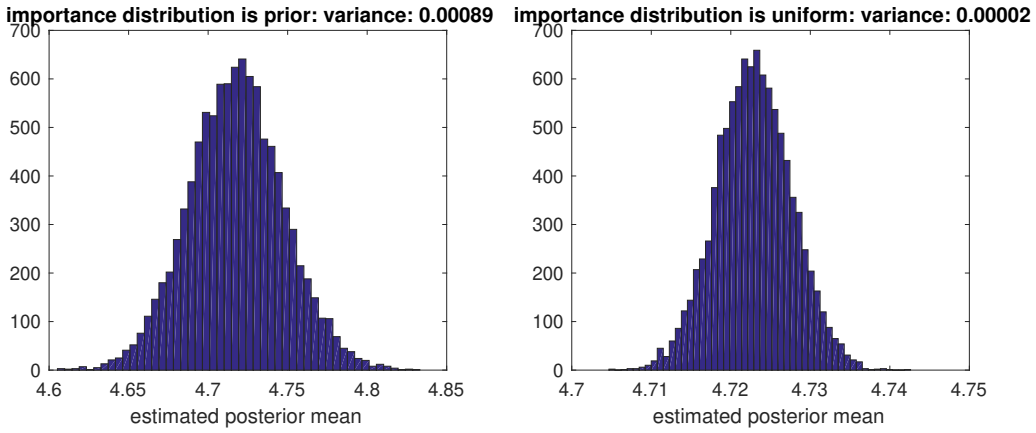

Figure 4.4: Histograms for the estimate of the posterior mean using two different importance sampling methods as described in Example 4.15 with $n = 10$ and $a = 2$.

Exercise 4.11. Consider the example in Example 4.11. Write a function that takes the noisy measurements $x = (x_1, x_2, x_3)$, positions of the sensors s_1, s_2, s_3, the prior and

likelihood variances σ_θ^2 and σ_x^2, and the number of samples N as inputs, implements self-normalised importance sampling (why this version?) in order to approximate

$$E(\theta|X = x) = [E(\theta_1|X = x), E(\theta_2|X = x)].$$

and outputs its estimate. Try your code with $s_1 = (0, 2)$, $s_2 = (-2, -1)$, $s_3 = (1, -2)$, $x_1 = 2$, $x_2 = 1.6$, $x_3 = 2.5$, $\sigma_\theta^2 = 100$, and $\sigma_x^2 = 1$ which are the values used to generate the plots in Figure 4.2.

C.3 Markov chain Monte Carlo

We have already discussed the difficulties of generating a large number of i.i.d. samples from a posterior distribution $\pi(\theta) = p(\theta|x)$. One alternative was importance sampling which involved weighting every generated sample in order not to waste it, but it has its drawbacks mostly due to issues related to controlling the variance of the importance weights. Another alternative is to use *Markov chain Monte Carlo* (MCMC) methods These methods are based on the design of a suitable ergodic Markov chain whose stationary distribution is $\pi(\theta)$. The idea is that if one simulates such a Markov chain, after a long enough time the samples of the Markov chain will approximately be distributed according to $\pi(\theta)$. Although the samples generated from the Markov chain are not i.i.d., their use is justified by convergence results for dependent random variables in the literature.

C.3.1 Metropolis-Hastings

As previously stated, an MCMC method is based on a discrete-time ergodic Markov chain which has its stationary distribution as π. The most widely used MCMC algorithm up to date is the *Metropolis-Hastings* algorithm.

The Metropolis-Hastings algorithm requires a Markov transition kernel on Θ for proposing new values from the old ones. Assume that the pdf/pmf of that transition kernel is $q(\cdot|\theta)$ for any θ. Given the previous sample θ_{n-1} a new value for θ_n is *proposed* as $\theta' \sim q(\theta'|\theta_{n-1})$. The proposed sample θ' is accepted with the acceptance probability $\alpha(\theta_{n-1}, \theta')$, where the function $\alpha : \Theta \times \Theta \to [0, 1]$ is defined as

$$\alpha(\theta, \theta') = \min\left\{1, \frac{\pi(\theta')q(\theta|\theta')}{\pi(\theta)q(\theta'|\theta)}\right\}, \quad \theta, \theta' \in \Theta.$$

If the proposal is accepted, $\theta_n = \theta'$ is taken. Otherwise, the proposal is rejected and $\theta_n = \theta_{n-1}$ is taken.

The ratio in the acceptance probability

$$r(\theta, \theta') = \frac{\pi(\theta')q(\theta|\theta')}{\pi(\theta)q(\theta'|\theta)}$$

is called the acceptance ratio, or the acceptance rate.

Algorithm 4.5: Metropolis-Hastings

1 Begin with some $\theta_1 \in \Theta$.
2 **for** $n = 2, 3, \ldots$ **do**
3 Sample $\theta' \sim q(\theta'|\theta_{n-1})$.
4 Set $\theta_n = \theta'$ with probability

$$\alpha(\theta_{n-1}, \theta') = \min\left\{1, \frac{\pi(\theta')q(\theta_{n-1}|\theta')}{\pi(\theta_{n-1})q(\theta'|\theta_{n-1})}\right\},$$

else set $\theta_n = \theta_{n-1}$.

The invariant distribution of the Metropolis-Hastings algorithm described exists and it is π. To show this, we can check for the detailed balance condition. According to Algorithm 4.5, the transition kernel M of the Markov chain from which the samples are obtained is

$$M(\theta'|\theta) = q(\theta'|\theta)\alpha(\theta, \theta') + p_r(\theta)\delta_\theta(\theta'),$$

where $p_r(\theta)$ is the rejection probability at θ and

$$p_r(\theta) = \left[1 - \int q(\theta'|\theta)\alpha(\theta, \theta')d\theta'\right], \quad \text{or} \quad p_r(\theta) = \left[1 - \sum_{\theta'} q(\theta'|\theta)\alpha(\theta, \theta')\right]$$

depending on the nature of the state-space. For all $\theta, \theta' \in \Theta$, we have

$$\begin{aligned}
\pi(\theta)M(\theta'|\theta) &= \pi(\theta)q(\theta'|\theta)\alpha(\theta, \theta') + \pi(\theta)p_r(\theta)\delta_\theta(\theta') \\
&= \pi(\theta)q(\theta'|\theta)\min\left\{1, \frac{\pi(\theta')q(\theta|\theta')}{\pi(\theta)q(\theta'|\theta)}\right\} + \pi(\theta)p_r(\theta)\delta_\theta(\theta') \\
&= \min\left\{\pi(\theta)q(\theta'|\theta), \pi(\theta')q(\theta|\theta')\right\} + \pi(\theta)p_r(\theta)\delta_\theta(\theta') \\
&= \min\left\{\pi(\theta')q(\theta|\theta'), \pi(\theta)q(\theta'|\theta)\right\} + \pi(\theta')p_r(\theta')\delta_{\theta'}(\theta)
\end{aligned}$$

which is symmetric with respect to θ, θ', so $\pi(\theta)M(\theta'|\theta) = \pi(\theta')M(\theta|\theta')$ and the detailed balance condition holds for π which implies that M is reversible with respect to π and π is invariant for M.

When a symmetric proposal is used, the acceptance probability involves only the ratio of the target distribution evaluated at θ and θ',

$$\alpha(\theta, \theta') = \min\left\{1, \frac{\pi(\theta')}{\pi(\theta)}\right\}, \quad \text{if} \quad q(\theta'|\theta) = q(\theta|\theta').$$

Another version is the independence Metropolis-Hastings algorithm, where, as the name suggests, the proposal kernel Q is chosen to be independent of the current value, i.e. $q(\theta'|\theta) = q(\theta')$, in which case the acceptance probability is

$$\alpha(\theta, \theta') = \min\left\{1, \frac{\pi(\theta')q(\theta)}{\pi(\theta)q(\theta')}\right\}.$$

Toy example: MH for the normal distribution: This is a toy example where $\pi(\theta) = \phi(\theta; \mu, \sigma^2)$ for which we do not need to use MH since we can obviously sample from $\mathcal{N}(\mu, \sigma^2)$ easily. But for the sake of example assume that we have decided to use MH to generate approximate samples from π.

For the proposal kernel, we have several options:

- Symmetric random walk: We can take $q(\theta'|\theta) = \phi(\theta'; \theta, \sigma_q^2)$, that is θ' is proposed from the current value θ by adding a normal random variable with zero mean and variance σ_q^2, or $q(\cdot|\theta) \sim \mathcal{N}(\theta, \sigma_q^2)$. Since

$$q(\theta'|\theta) = \phi(\theta'; \theta, \sigma_q^2) = \phi(\theta; \theta', \sigma_q^2) = q(\theta|\theta'),$$

  this results in the acceptance ratio

$$\begin{aligned} r(\theta, \theta') &= \frac{\pi(\theta')q(\theta|\theta')}{\pi(\theta)q(\theta'|\theta)} \\ &= \frac{\phi(\theta'; \mu, \sigma^2)}{\phi(\theta; \mu, \sigma^2)} \\ &= \frac{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(\theta'-\mu)^2}}{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(\theta-\mu)^2}} \\ &= e^{-\frac{1}{2\sigma^2}\left[(\theta'-\mu)^2-(\theta-\mu)^2\right]} \end{aligned}$$

  The choice of σ_q^2 is important for MH to have good performance. We want the Markov chain generated by the algorithm to *mix* well, that is we want the samples to forget the previous values fast. Consider the acceptance ratio above:

  - A too small value for σ_q^2 will result in the acceptance ratio $r(\theta, \theta')$ being very close to 1, and hence the proposed values will be accepted with high probability. However, the chain will be very slowly mixing, that is the samples will be highly correlated; because any accepted sample θ' will most likely be only slightly different than the current θ due to a small step-size of the random walk.
  - A too large value for σ_q^2 will likely result in the proposed value θ' to be far from the region where π has most of its mass, hence $\pi(\theta')$ will be very small compared to $\pi(\theta)$ and the chain will likely reject the proposed value and stick to the old value θ. This will create a *sticky* chain.

  Therefore, the optimum value for σ_q^2 should be neither too small nor too large. See Figure 4.5 for both bad choices and one in between those. This phenomenon of having to choose the variance of the random walk proposals neither too small nor too big is also valid for most distributions than the normal distribution.

- Another option for the proposal is to sample θ' independently from θ, i.e. $q(\theta'|\theta) = q(\theta')$. For example, suppose we chose $q(\theta) = \phi(\theta; \mu_q, \sigma_q^2)$. Then the acceptance ratio

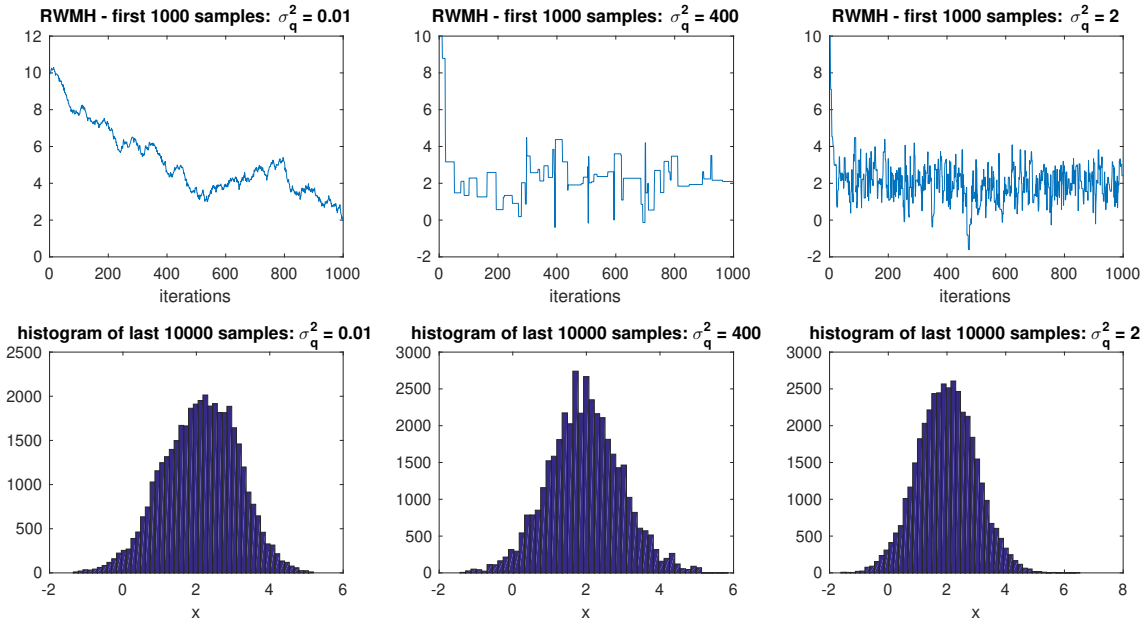

Figure 4.5: Random walk MH for $\pi(\theta) = \phi(\theta; 2, 1)$. The left and middle plots correspond to a too small and a too large value for σ_q^2, respectively. All algorithms are run for 50000 iterations. Both the trace plots and the histograms show that the last choice works the best.

is

$$\begin{aligned} r(\theta, \theta') &= \frac{\pi(\theta')q(\theta)}{\pi(\theta)q(\theta')} \\ &= \frac{\phi(\theta'; \mu, \sigma^2)\phi(\theta; \mu_q, \sigma_q^2)}{\phi(\theta; \mu, \sigma^2)\phi(\theta'; \mu_q, \sigma_q^2)} \\ &= \frac{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(\theta'-\mu)^2} \frac{1}{\sqrt{2\pi\sigma_q^2}} e^{-\frac{1}{2\sigma_q^2}(\theta-\mu_q)^2}}{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(\theta-\mu)^2} \frac{1}{\sqrt{2\pi\sigma_q^2}} e^{-\frac{1}{2\sigma_q^2}(\theta'-\mu_q)^2}} \\ &= e^{-\frac{1}{2\sigma^2}\left[(\theta'-\mu)^2-(\theta-\mu)^2\right]+\frac{1}{2\sigma_q^2}\left[(\theta'-\mu_q)^2-(\theta-\mu_q)^2\right]} \end{aligned}$$

See Figure 4.6 for examples of MH with this choice.

- Another alternative is to use a gradient-guided proposal. We may want to 'guide' the chain towards the high-probability region of $\pi(\theta)$; one proposal that can be chosen for that purpose is

$$q(\theta'|\theta) = \phi(\theta'; g(\theta), \sigma_q^2)$$

where the mean for the proposal $g(\theta)$ is constructed by using the gradient of the

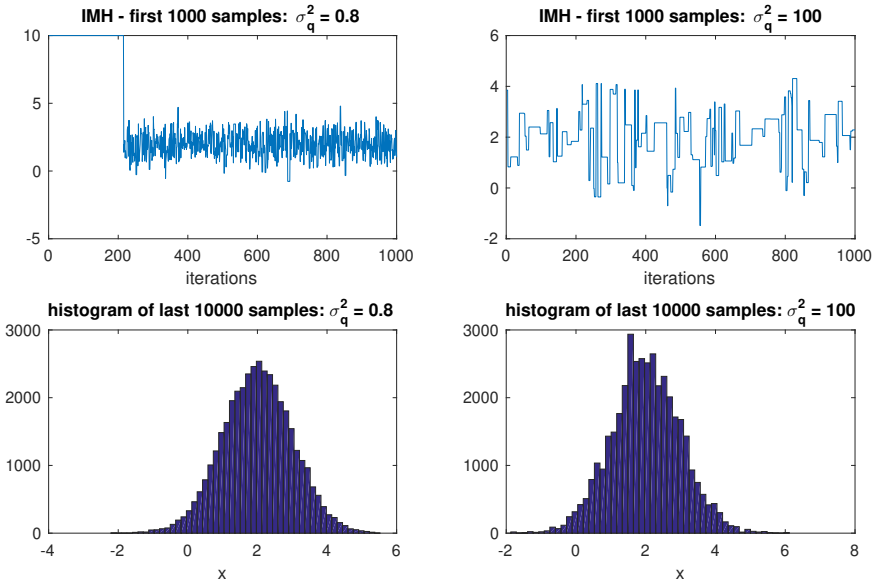

Figure 4.6: Independence MH for $\pi(\theta) = \phi(\theta; 2, 1)$.

logarithm of the target density,

$$g(\theta) = \theta + \gamma \frac{\partial \log \pi(\theta)}{\partial \theta}.$$

Here, γ is a step-size parameter that needs to be adjusted. For $\pi(\theta) = \phi(\theta; \mu, \sigma^2)$, $g(\theta) = \theta - \frac{\gamma}{\sigma^2}(\theta - \mu)$. The acceptance ratio for this choice of proposal becomes

$$r(\theta, \theta') = e^{-\frac{1}{2\sigma^2}\left[(\theta'-\mu)^2-(\theta-\mu)^2\right]+\frac{1}{2\sigma_q^2}\left[\left(\theta'-\theta+\frac{\gamma}{\sigma^2}(\theta-\mu)\right)^2-\left(\theta-\theta'+\frac{\gamma}{\sigma^2}(\theta'-\mu)\right)^2\right]}$$

See Figure 4.7 for examples of MH with this choice.

Example 4.16 (**Normal distribution with unknown mean and variance**). We have observations $X_1, \ldots, X_n \sim \mathcal{N}(z, s)$ and z and s are unknown. The parameters $\theta = (z, s)$ are *a priori* independent with $z \sim \mathcal{N}(m, \kappa^2)$ and $s \sim \mathcal{IG}(\alpha, \beta)$, so that the prior density is

$$p(\theta) = p(z)p(s) = \frac{1}{\sqrt{2\pi\kappa^2}} e^{-\frac{1}{2\kappa^2}(z-m)^2} \frac{\beta^\alpha}{\Gamma(\alpha)} s^{-\alpha-1} e^{-\frac{\beta}{s}}$$

Given the data $X_{1:n} = x_{1:n}$, we want to run the MH algorithm to sample from the posterior distribution $\pi(\theta) = p(\theta|x_{1:n})$, which is given by

$$\pi(\theta) = p(\theta|x_{1:n}) \propto p(\theta)p(x_{1:n}|\theta) = p(z)p(s) \prod_{i=1}^{n} \phi(x_i; z, s)$$

For this problem, $\pi(\theta)$ indeed lacks a well-known form, so it is justified to use a Monte Carlo method for it.

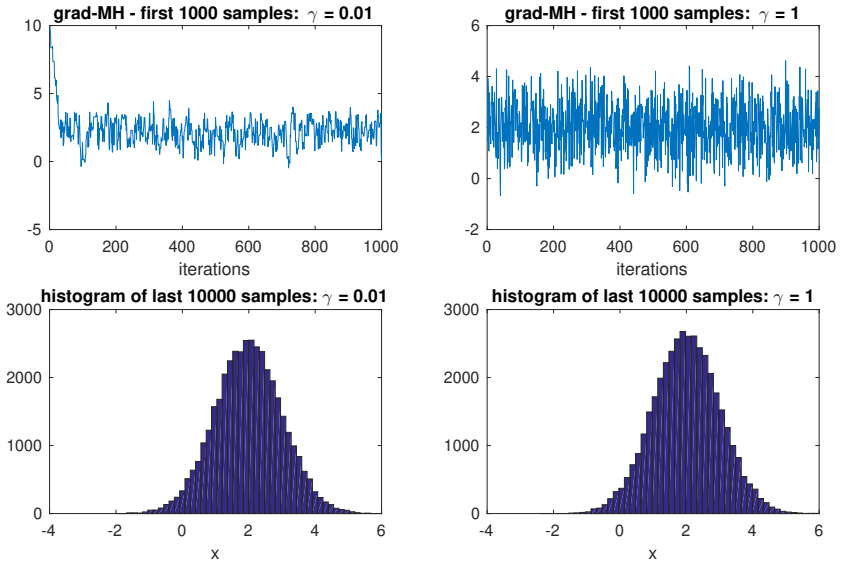

Figure 4.7: Gradient-guided MH for $\pi(\theta) = \phi(\theta; 2, 1)$.

To run the MH algorithm, we need a proposal distribution for proposing $\theta' = (z', s')$. In this example, given $\theta = (z, s)$, we decide to propose $z' \sim \mathcal{N}(z, \sigma_q^2)$ and $s' \sim \mathcal{IG}(\alpha, \beta)$, i.e. we use a random walk for the mean component and the prior distribution for the variance parameter. With this choice, the proposal density becomes

$$\begin{aligned} q(\theta'|\theta) &= \phi(z'; z, \sigma_q^2)p(s') \\ &= \phi(z'; z, \sigma_q^2)\frac{\beta^\alpha}{\Gamma(\alpha)}(s')^{-\alpha-1}e^{-\frac{\beta}{s'}} \end{aligned}$$

The acceptance ratio in this case is

$$\begin{aligned} r(\theta, \theta') &= \frac{\pi(\theta')q(\theta|\theta')}{\pi(\theta)q(\theta'|\theta)} \\ &= \frac{p(z')p(s')\left[\prod_{i=1}^n p(x_i|z', s')\right]\phi(z; z', \sigma_q^2)p(s)}{p(z)p(s)\left[\prod_{i=1}^n p(x_i|z, s)\right]\phi(z'; z, \sigma_q^2)p(s')} \\ &= \frac{\phi(z'; m, \kappa^2)\prod_{i=1}^n \phi(x_i; z', s')}{\phi(z; m, \kappa^2)\prod_{i=1}^n \phi(x_i; z, s)} \end{aligned}$$

See Figure 4.8 for results obtained from this MH algorithm.

Exercise 4.12. Implement the MH algorithm in Example 4.16 on the data given in `normal.txt`. Use $\alpha = 5$ and $\beta = 10$, $m = 0$, and $\kappa^2 = 100$ as the prior parameters, and $\sigma_q^2 = 1$ for the proposal.

Example 4.17 (**A changepoint model**). In this example, we consider a changepoint model. In this model, at each time t we observe the count of an event X_t. All the counts

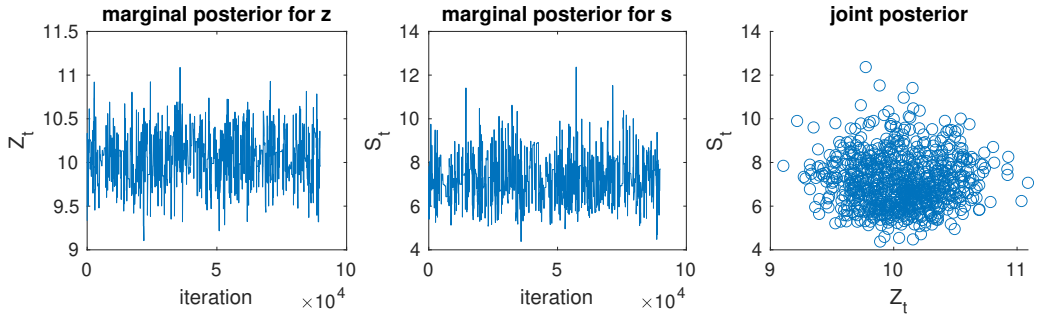

Figure 4.8: MH for parameters of $\mathcal{N}(z, s)$. $\sigma_q^2 = 1$, $\alpha = 5$, $\beta = 10$, $m = 0$, $\kappa^2 = 100$.

up to an unknown time τ come from the same distribution after which the distribution changes. We assume that the changepoint τ is uniformly distributed over $\{1, \ldots, n\}$ where n is the number of time steps. The two different distributional regimes up to τ and after τ are indicated by the random variables λ_i, $i = 1, 2$, which are *a priori* assumed to follow a Gamma distribution

$$\lambda_i \sim \Gamma(\alpha, \beta), \quad i = 1, 2.$$

Under regime i, the counts are assumed to be identically Poisson distributed

$$X_t \sim \begin{cases} \mathcal{PO}(\lambda_1), & 1 \leq t \leq \tau \\ \mathcal{PO}(\lambda_2), & \tau < t \leq n. \end{cases}$$

A typical draw from this model is shown in Figure 4.9. The inferential goal is, given $X_{1:n} = x_{1:n}$, to sample from the posterior distribution of the changepoint location τ and the intensities λ_1, λ_2 given the count data, i.e., letting $\theta = (\tau, \lambda_1, \lambda_2)$, the target distribution is $\pi(\theta) = p(\tau, \lambda_1, \lambda_2 | x_{1:n})$ which is given by

$$\begin{aligned} p(\tau, \lambda_1, \lambda_2 | x_{1:n}) &\propto p(\tau, \lambda_1, \lambda_2, x_{1:n}) \\ &= p(\tau, \lambda_1, \lambda_2) p(x_{1:n} | \tau, \lambda_1, \lambda_2) \\ &= p(\tau) p(\lambda_1) p(\lambda_2) p(x_{1:n} | \tau, \lambda_1, \lambda_2) \\ &= \frac{1}{n} \frac{\beta^\alpha \lambda_1^{\alpha-1} e^{-\beta\lambda_1}}{\Gamma(\alpha)} \frac{\beta^\alpha \lambda_2^{\alpha-1} e^{-\beta\lambda_2}}{\Gamma(\alpha)} \prod_{t=1}^{\tau} \frac{e^{-\lambda_1} \lambda_1^{x_t}}{x_t!} \prod_{t=\tau+1}^{n} \frac{e^{-\lambda_2} \lambda_2^{x_t}}{x_t!} \end{aligned} \tag{4.23}$$

Two choices for the proposal will be considered. Let $\theta' = (\tau', \lambda_1', \lambda_2')$.

- The first one is to use an independent proposal distribution, which is the prior distribution for x

  $$q(\theta' | \theta) = q(\theta') = p(\theta') = p(\tau', \lambda_1', \lambda_2').$$

  This leads to the acceptance ratio being the ratio of the likelihoods

  $$r(\theta, \theta') = \frac{p(x_{1:n} | \tau', \lambda_1', \lambda_2')}{p(x_{1:n} | \tau, \lambda_1, \lambda_2)}$$

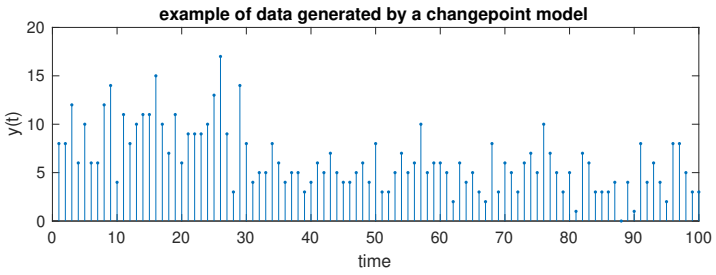

Figure 4.9: An example data sequence of length $n = 100$ generated from the Poisson changepoint model with parameters $\tau = 30$, $\lambda_1 = 10$ and $\lambda_2 = 5$.

- The second choice is a symmetric proposal,

$$q(\theta'|\theta) = \left[\frac{1}{2}\mathbb{I}_{\tau+1}(\tau') + \frac{1}{2}\mathbb{I}_{\tau-1}(\tau')\right] \phi(\lambda_1'; \lambda_1, \sigma_\lambda^2)\phi(\lambda_2'; \lambda_2, \sigma_\lambda^2).$$

  The first factor involving τ indicates that we propose either $\tau' = \tau + 1$ or $\tau' = \tau - 1$ both with probability a half. Since $q(\theta'|\theta) = q(\theta|\theta')$, the acceptance ratio reduces to the ratio of the posteriors

$$\begin{aligned} r(\theta, \theta') =& \frac{p(\tau', \lambda_1', \lambda_2'|y_{1:n})}{p(\tau, \lambda_1, \lambda_2|x_{1:n})} \\ =& e^{-(\tau+\beta)(\lambda_1'-\lambda_1)} e^{-(n-\tau+\beta)(\lambda_2'-\lambda_2)} \left(\frac{\lambda_1'}{\lambda_1}\right)^{\alpha-1+\sum_{t=1}^{\tau} x_t} \left(\frac{\lambda_2'}{\lambda_2}\right)^{\alpha-1+\sum_{t=\tau+1}^{n} x_t} \\ & \times \begin{cases} e^{-\lambda_1'+\lambda_2'} \left(\frac{\lambda_1'}{\lambda_2'}\right)^{x_{\tau+1}}, & \tau' = \tau + 1, \\ e^{-\lambda_2'+\lambda_1'} \left(\frac{\lambda_2'}{\lambda_1'}\right)^{x_\tau}, & \tau' = \tau - 1. \end{cases} \end{aligned}$$

Figure 4.10 illustrates the results obtained from the two algorithms. The initial value for τ is taken $\lfloor n/2 \rfloor$ and for λ_1 and λ_2 we start from the mean of $x_{1:n}$. As we can see, the symmetric proposal algorithm can explore the posterior distribution much more efficiently. This is because the proposal distribution in independence MH, which is chosen as the prior distribution, takes neither the posterior distribution (hence the data) nor the previous sample into account, and as a result, it has a large rejection rate. The independence sampler would become even poorer if n were larger so that the posterior would be more concentrated in contrast to the ignorance of the prior distribution.

Example 4.18 (**MCMC for source localisation**). Consider the source localisation scenario in Example 4.11. From the likelihood and the prior in (4.12) and (4.13), the posterior distribution of the unknown position is

$$p(\theta|x) \propto \phi(\theta_1; 0, \sigma_\theta^2)\phi(\theta_2; 0, \sigma_\theta^2) \prod_{i=1}^{3} \phi(x_i; r_i, \sigma_x^2) \tag{4.24}$$

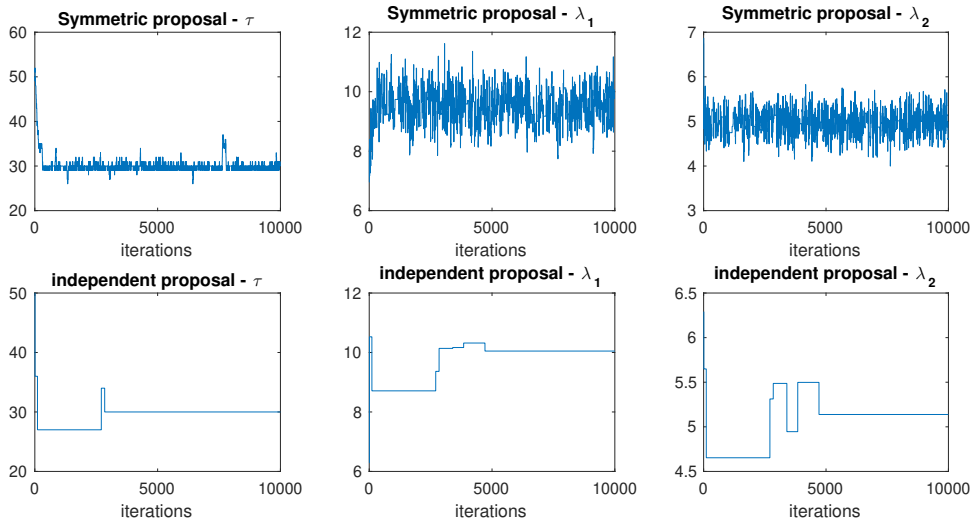

Figure 4.10: MH for parameters of the Poisson changepoint model

Due to the non-linearity in the $r_i = ||\theta - s_i|| = [(\theta_1 - s_i(1))^2 + (\theta_2 - s_i(2))^2]^{1/2}$, $i = 1, 2, 3$, $p(\theta|x)$ does not admit a known distribution. We use the MH algorithm to generate approximate samples from $p(\theta|x)$. We use a symmetric random walk proposal distribution with $q(\theta'|\theta) = \phi(\theta'; \theta, \sigma_q^2 I_2)$, so that $q(\theta'|\theta) = q(\theta|\theta')$. The resulting acceptance rate

$$\begin{aligned} r(\theta, \theta') &= \frac{p(\theta'|x)q(\theta|\theta')}{p(\theta|x)q(\theta'|\theta)} \\ &= \frac{p(\theta'|x)}{p(\theta|x)} \\ &= \frac{\phi(\theta_1'; 0, \sigma_\theta^2)\phi(\theta_2'; 0, \sigma_\theta^2)\prod_{i=1}^3 \phi(x_i; r_i', \sigma_x^2)}{\phi(\theta_1; 0, \sigma_\theta^2)\phi(\theta_2; 0, \sigma_\theta^2)\prod_{i=1}^3 \phi(x_i; r_i, \sigma_x^2)} \end{aligned}$$

where $r_i' = ||\theta' - s_i||$, $i = 1, 2, 3$ is the distance between the proposed value θ' and the location i'th source s_i. Figure 4.11 shows the samples and their histograms obtained from 10000 iterations of the MH algorithm. The chain was started from $\theta = (5, 5)$ and its convergence to the posterior distribution is illustrated in the right pane of the figure where we see the first a few samples of the chain traveling to the high probability region of the posterior distribution.

Exercise 4.13. Implement the MH algorithm for the source localisation problem in Example 4.18. Use $s_1 = (0, 2)$, $s_2 = (-2, -1)$, $s_3 = (1, -2)$, $x_1 = 2$, $x_2 = 1.6$, $x_3 = 2.5$, $\sigma_\theta^2 = 100$, and $\sigma_x^2 = 1$

C.3.2 Gibbs sampling

The *Gibbs sampler* is one of the most popular MCMC methods, which can be used when θ has more than one dimension. If θ has $d > 1$ components (of possibly different dimensions)

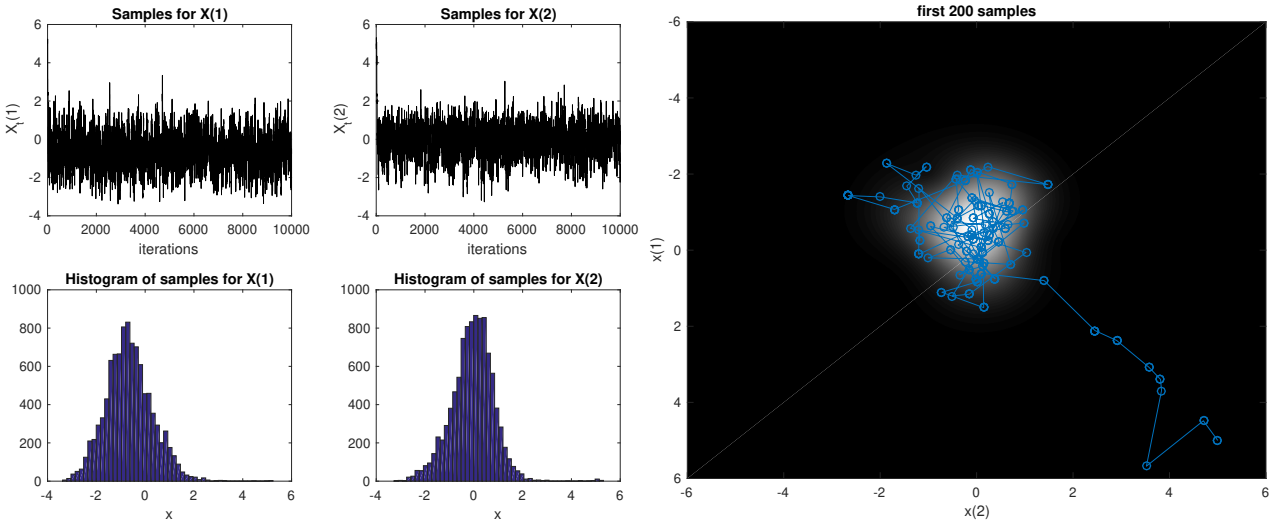

Figure 4.11: MH for the source localisation problem.

such that $\theta = (\theta_1, \ldots, \theta_d)$, and one can sample from each of the *full conditional distributions* $\pi_k(\cdot|\theta_{1:k-1}, \theta_{k+1:d})$, then the Gibbs sampler produces a Markov chain by updating one component at a time using π_k's. One cycle of the Gibbs sampler successively samples from the conditional distributions $\pi_1, \ldots, \pi_d$ by conditioning on the most recent samples.

Algorithm 4.6: The Gibbs sampler:

Begin with some $\theta_1 \in \Theta$.
for $n = 2, 3, \ldots$ **do**
  for $k = 1, \ldots, d$ **do**

$$\theta_{n,k} \sim \pi_k(\cdot|\theta_{n,1:k-1}, \theta_{n-1,k+1:d}).$$

For an $\theta \in \Theta$, let $\theta_{-k} = (\theta_{1:k-1}, \theta_{k+1:d})$ for $k = 1, \ldots, d$ denotes the components of x excluding θ_k, and let us permit ourselves to write $\theta = (\theta_k, \theta_{-k})$. The corresponding MCMC kernel of the Gibbs sampler can be written as $M = M_1 M_2 \ldots M_d$, where each transition kernel M_k for $k = 1, \ldots, d$ can be written as

$$M_k(\theta'|\theta) = \pi_k(\theta'_k|\theta_{-k})\delta_{\theta_{-k}}(\theta'_{-k})$$

where $\theta' = (\theta'_1, \ldots, \theta'_d)$. The justification of the transitional kernel comes from the reversibility of each M_k with respect to π, which can be verified from the detailed balance

condition as follows.

$$\begin{aligned}\pi(\theta)M_k(\theta'|\theta) &= \pi(\theta)\pi_k(\theta'_k|\theta_{-k})\delta_{\theta_{-k}}(\theta'_{-k}) \\ &= \pi(\theta_{-k})\pi_k(\theta_k|\theta_{-k})\pi_k(\theta'_k|\theta_{-k})\delta_{\theta_{-k}}(\theta'_{-k}) \\ &= \pi(\theta'_{-k})\pi_k(\theta'_k|\theta'_{-k})\pi_k(\theta_k|\theta'_{-k})\delta_{\theta'_{-k}}(\theta_{-k}) \\ &= \pi(\theta')M_k(\theta|\theta'), \end{aligned} \tag{4.25}$$

where the third line follows the second since $\delta_{\theta_{-k}}(\theta'_{-k})$ allows the interchange of θ_{-k} and θ'_{-k}. Therefore, the detailed balance condition for M_k is satisfied with π and $\pi M_k = \pi$. If we apply $M_1, \ldots, M_k$ sequentially, we get

$$\pi M = \pi M_1 \ldots M_d = (\pi M_1) M_2 \ldots M_d = \pi M_2 \ldots M_d = \ldots = \pi,$$

so π is indeed the invariant distribution for the Gibbs sampler.

Gibbs sampling as a special Metropolis-Hastings algorithm: An insightful interpretation of (4.25) is that each step of a cycle of Gibbs sampling is a Metropolis-Hastings move whose MCMC kernel is equal to its proposal kernel which results in the acceptance probability being 1 uniformly. Indeed, if the k'th component of θ is to be updated with $Q_k = M_k$, i.e. if we propose the new value θ' as

$$q_k(\theta'|\theta) = M_k(\theta'|\theta) = \pi_k(\theta'_k|\theta_{-k})\delta_{\theta_{-k}}(\theta'_{-k}),$$

the acceptance ratio $\alpha_k(\theta, \theta')$ for this move is

$$\alpha_k(\theta, \theta') = \min\left\{1, \frac{\pi(\theta')q_k(\theta|\theta')}{\pi(\theta)q_k(\theta'|\theta)}\right\} = \min\left\{1, \frac{\pi(\theta')M_k(\theta|\theta')}{\pi(\theta)M_k(\theta'|\theta)}\right\} = 1$$

as shown in (4.25).

Example 4.19. Suppose we wish to sample from a bivariate normal distribution, where

$$\pi(\theta) = \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\left\{-\frac{\theta_1^2 + \theta_2^2 - 2\rho\theta_1\theta_2}{2(1-\rho^2)}\right\}, \quad \rho \in (-1, 1).$$

The full conditionals are

$$\pi(\theta_1|\theta_2) \propto \pi(\theta_1, \theta_2) \propto \exp\left\{-\frac{(\theta_1 - \rho\theta_2)^2}{2(1-\rho^2)}\right\}$$

therefore $\pi(\theta_1|\theta_2) = \phi(\theta_1; \rho\theta_2, (1-\rho^2))$ and $\theta_1|\theta_2 \sim \mathcal{N}(\rho\theta_2, (1-\rho)^2)$. Similarly, we have $\theta_2|\theta_1 \sim \mathcal{N}(\rho\theta_1, (1-\rho)^2)$. So, the iteration $t \geq 2$ of the Gibbs sampling algorithm for this $\pi(\theta)$ is

- Sample $\theta_{t,1} \sim \mathcal{N}(\rho\theta_{t-1,2}, (1-\rho)^2)$,

- Sample $\theta_{t,2} \sim \mathcal{N}(\rho\theta_{t,1}, (1-\rho)^2)$.

Example 4.20 (**ex: Normal distribution with unknown mean and variance**). Let us get back to the problem in Example 4.16 where we want to estimate the mean and the variance of the normal distributions $\mathcal{N}(z, s)$ given samples $x_1, \dots, x_n$ generated from it. Let use the same prior distributions for z and s, namely $z \sim \mathcal{N}(m, \kappa^2)$ and $s \sim \mathcal{IG}(\alpha, \beta)$. Note that these are the conjugate priors for those parameters; and when one of the parameters is given, the posterior distribution of the other one has a known form. Indeed, in Examples 4.5 and 4.6, we derived these full conditional distributions. Example 4.5 can be revisited (but this time with a non-zero prior mean m) to see that

$$z|s, x_{1:n} \sim \mathcal{N}(\mu_{z|s,x}, \sigma^2_{z|s,x})$$

where

$$\sigma^2_{z|s,x} = \left(\frac{1}{\kappa^2} + \frac{n}{s}\right)^{-1}, \quad \mu_{z|s,x} = \left(\frac{1}{\kappa^2} + \frac{n}{s}\right)^{-1}\left(\frac{1}{s}\sum_{i=1}^{n} x_i + \frac{m}{\kappa^2}\right)$$

and from Example 4.6 we can deduce that

$$s|z, x_{1:n} = \mathcal{IG}(\alpha_{s|z,x}, \beta_{s|z,x})$$

where

$$\alpha_{s|z,x} = \alpha + \frac{n}{2}, \quad \beta_{s|z,x} = \beta + \frac{1}{2}\sum_{i=1}^{n}(x_i - z)^2.$$

Therefore, Gibbs sampling for (z, s) given $X_{1:n} = x_{1:n}$ is

- Sample $z_t \sim \mathcal{N}\left(\left(\frac{1}{\kappa^2} + \frac{n}{s_{t-1}}\right)^{-1}\left(\frac{1}{s_{t-1}}\sum_{i=1}^{n} x_i + \frac{m}{\kappa^2}\right), \left(\frac{1}{\kappa^2} + \frac{n}{s_{t-1}}\right)^{-1}\right)$
- Sample $s_t \sim \mathcal{IG}\left(\alpha + \frac{n}{2}, \beta + \frac{1}{2}\sum_{i=1}^{n}(x_i - z_t)^2\right)$.

Exercise 4.14. Implement the Gibbs sampler in Example 4.20 on the data given in `normal.txt`. Use $\alpha = 5$ and $\beta = 10$ as the prior parameters.

Data augmentation: Data augmentation is an application of the Gibbs sampler. It is useful if

1. there is missing data, and/or
2. the likelihood is intractable (hard to compute or does not admit conjugacy, etc), but given some additional unobserved (real or fictitious) data it would be tractable.

Let x_{obs} denote the observed data and x_{mis} the missing data (sometimes x_{mis} is called a latent variable). We suppose we can easily sample θ from the posterior given the augmented data $(x_{\text{obs}}, x_{\text{mis}})$. Also, that we can sample x_{mis}, conditional on x_{obs} and θ (this only involves the sampling distributions). Then we can use the Gibbs sampler of the pair (θ, x_{mis}). Then we perform Monte Carlo marginalisation: If in the resulting joint distribution for θ, x_{mis} given x_{obs} we simply ignore x_{mis}, we shall have our sample from the posterior of θ given x_{obs} alone.

Example 4.21 (**Genetic linkage**). Genetic linkage in an animal can be allocated to one of four categories, coded 1,2, 3, and 4, having respective probabilities

$$(1/2 + \theta/4, (1-\theta)/4, (1-\theta)/4, \theta/4)$$

where θ is an unknown parameter in $(0, 1)$. For a sample of 197 animals, the (multinomial) counts of those falling in the 4 categories are represented by random variables $X = (X_1, X_2, X_3, X_4)$, with observed values $x = (x_1, x_2, x_3, x_4) = (125, 18, 20, 34)$. Suppose we place a Beta(α, β) prior on θ. Then,

$$\begin{aligned}\pi(\theta) = p(\theta|x) &\propto \underbrace{\left(\frac{1}{2} + \frac{\theta}{4}\right)^{125} \left(\frac{1-\theta}{4}\right)^{18+20} \left(\frac{\theta}{4}\right)^{34}}_{\text{Multinomial likelihood}} \theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &\propto (2+\theta)^{125}(1-\theta)^{38+\beta-1}\theta^{34+\alpha-1} \end{aligned} \tag{4.26}$$

How can we sample from this? We can use a rejection sampler (probably with a very high rejection probability) or MH for this posterior distribution. In this example, we seek a suitable Gibbs sampler. Note that the problematic part in (4.26) is the first one; should it be like one of the others, the posterior would lend itself to a Beta distribution.

Suppose we divide category 1, with total probability $1/2 + \theta/4$, into two latent subcategories, a and b, with respective probabilities $\theta/4$ and $1/2$. We regard the number of animals z falling in subcategory a as missing data. If, as well as the observed data x, we are given z, we are in the situation of having observed counts $(z, 125 - z, 18, 20, 34)$ from a multinomial distribution with probabilities $(\theta/4, 1/2, (1-\theta)/4, (1-\theta)/4, \theta/4)$. The resulting joint distribution is

$$p(\theta, z|x) \propto p(\theta, z, x) = \left(\frac{1}{2}\right)^{125-z} \left(\frac{1-\theta}{4}\right)^{18+20} \left(\frac{\theta}{4}\right)^{34+z} \theta^{\alpha-1}(1-\theta)^{\beta-1} \tag{4.27}$$

This easily leads to the posterior distribution

$$\theta|z, x \sim \text{Beta}(z + 34 + \alpha, 38 + \beta). \tag{4.28}$$

Also, simple properties of the multinomial distribution yield

$$z|\theta, x \sim \text{Binom}\left(125, \frac{\theta/4}{1/2 + \theta/4}\right) \tag{4.29}$$

So we can now apply Gibbs sampling, cycling between updates given by (4.28) and (4.29).

Exercise 4.15. Design and implement a symmetric random walk MH algorithm and the Gibbs sampling algorithm for the genetic linkage problem in Example 4.21 with hyperparameters $\alpha = \beta = 2$.

Example 4.22 (**A changepoint model, ctd.**). Consider the changepoint problem in Example 4.17, with the same likelihood and priors. It is possible to run Gibbs sampling algorithm for $\tau, \lambda_1, \lambda_2$. Observing (4.23), where the full posterior distribution is written as proportional to the full joint distribution

$$p(\tau, \lambda_1, \lambda_2, x_{1:n}) = \frac{1}{n} \frac{\beta^\alpha \lambda_1^{\alpha-1} e^{-\beta\lambda_1}}{\Gamma(\alpha)} \frac{\beta^\alpha \lambda_2^{\alpha-1} e^{-\beta\lambda_2}}{\Gamma(\alpha)} \prod_{t=1}^{\tau} \frac{e^{-\lambda_1} \lambda_1^{x_t}}{x_t!} \prod_{t=\tau+1}^{n} \frac{e^{-\lambda_2} \lambda_2^{x_t}}{x_t!},$$

from which we can derive all the full conditionals

$$\begin{aligned}
\lambda_1 | \tau, \lambda_2, x_{1:n} &\sim \Gamma\left(\alpha + \sum_{t=1}^{\tau} x_t, \beta + \tau\right) \\
\lambda_2 | \tau, \lambda_1, x_{1:n} &\sim \Gamma\left(\alpha + \sum_{t=\tau+1}^{n} x_t, \beta + n - \tau\right) \\
\tau | \lambda_1, \lambda_2, x_{1:n} &\sim \text{Categorical}(a_1, \ldots, a_n)
\end{aligned}$$

where the probabilities in the Categorical distribution (which is simply the discrete distribution with probabilities $a_1, \ldots, a_n$, the generalisation of the Bernoulli distribution to the case of multiple (here, n) outcomes) are

$$a_i = \frac{e^{-i\lambda_1} \lambda_1^{\sum_{t=1}^{i} x_t} e^{-(n-i)\lambda_2} \lambda_2^{\sum_{t=i+1}^{n} x_t}}{\sum_{j=1}^{n} \left[e^{-j\lambda_1} \lambda_1^{\sum_{t=1}^{j} x_t} e^{-(n-j)\lambda_2} \lambda_2^{\sum_{t=j+1}^{n} x_t} \right]}$$

Exercise 4.16. Consider the changepoint problem in Example 4.17.

- Download `UK_coal_mining_disaster_days.txt` from SUCourse. The data consists of the day numbers of coal mining disasters between 1851 and 1962, where the first day is the start of the 1851. It is suspected that, due to a policy change, the accident rate over the years is a piecewise constant with a single changepoint time around the time of the policy change.
- From the data, create another data vector of length 112, where the i'th element contains the number of disasters in year i (starting from 1851). Note that some years are 366 days!
- Implement the Gibbs algorithm for the changepoint model given the data that you created. Take the priors for τ, λ_1 and λ_2 the same as in Example 4.17, i.e. with hyperparameters $\alpha = 10$ and $\beta = 1$. All the derivations you need are in Example 4.22.

C.3.3 Metropolis within Gibbs

Having attractive computational properties, the Gibbs sampler is widely used. The requirement for easy-to-sample conditional distributions is the main restriction for the Gibbs sampler. Fortunately, though, replacing an exact simulation $\theta_k \sim \pi_k(\cdot|\theta_{n-1,1:k-1}, \theta_{n-1,k+1:d})$ by a Metropolis-Hastings step in a general MCMC algorithm does not violate its validity as long as the Metropolis-Hastings step has the correct invariant distribution. The most natural alternative to the Gibbs move in step k where sampling from the full conditional distribution $\pi_k(\cdot|\theta_{-k})$ is not directly feasible is to use Metropolis-Hastings move that updates θ_k by using a Metropolis-Hastings kernel that targets $\pi_k(\cdot|\theta_{-k})$.

Exercise 4.17. Suppose we observe a noisy sinusoid with with unknown amplitude a, angular frequency ω, phase z, and noise variance σ_x^2 for n steps. Letting $\theta = (a, \omega, z, \sigma_x^2)$,

$$X|\theta \sim \mathcal{N}(x_t; a\sin(\omega t + z), \sigma_x^2), \quad t = 1, \ldots, n.$$

The unknown parameters are a priori independent with $a \sim \mathcal{N}(0, \sigma_a^2)$, $\omega \sim \Gamma(\alpha, \beta)$, $z \sim \text{Unif}(0, 2\pi)$, $\sigma_x^2 \sim \mathcal{IG}(\alpha, 1/\beta)$.

- Write down the likelihood of $p(x_{1:n}|\theta)$ and the joint density $p(\theta, x_{1:n})$.
- Download the data file `sinusoid.txt` from SUCourse; the observations in the file are your data $x_{1:n}$. Use hyperparameters $\sigma_a^2 = 100$, $\alpha = \beta = 0.01$ and design and implement an MH algorithm for generating samples from the posterior distribution $\pi(\theta) = p(\theta|x_{1:n})$.
- This time, design and implement a MH within Gibbs algorithm where in each loop contains four steps in each of which you update one component only, fixing the others, using an MH kernel that targets the full conditionals.

Exercise 4.18. Download the `logistic_regression.txt` from SUCourse. The data contain several columns, where the last column is the response column and the other columns have the predictor variables. For each variable, we assume a relation given by the logistic regression model

$$P(Y_i = 1|x_{i,1}, \ldots, x_{i,k}, \beta) = \frac{1}{1 + \exp\{-(\beta_0 + \beta_1 x_{i,1} + \ldots + \beta_k x_{i,k})\}}$$

Implement the Metropolis-within-Gibbs algorithm for the posterior distribution of β given the data, where at each iteration a single component of β is updated in turn with a Metropolis-Hastings move. Take the prior distribution as $\beta \sim \mathcal{N}(0, 100 \times I)$. Use a random walk proposal for the each component, where $\beta_k' \sim \mathcal{N}(\beta_k, \sigma_q^2/n)$, where n is the number of rows in the data. Run the algorithm for 100000 iterations. Adjust σ_q^2 to have a good performance. Provide the details of your implementation: report the value of σ_q you used. Plot the trace plot of the samples for β_0 and β_1 (versus iteration). By inspecting those plots, determine a suitable burn-in time.

Appendix A

Some Basics of Probability

Summary: *This chapter provides some basics of probability which is related to the content of this course. The covered concepts are probability, random variables, cumulative distribution function, discrete and continuous distributions, probability mass function, probability density function, expectation, independence, correlation and covariance, Bayes' Theorem, and posterior distribution*

A Axioms and properties of probability

Let Ω be the *sample space* and $\mathcal{F}$ be the *event space*. (In a non-rigorous way, you can think of $\mathcal{F}$ as the set of all subsets of Ω as an example.) A *probability measure* on $(\Omega, \mathcal{F})$ is a function $P : \mathcal{F} \to \mathbb{R}$ that satisfies the following *axioms of probability*.

(A1) The probability of an event is a non-negative and real number:

$$P(E) \in \mathbb{R}, \quad P(E) \geq 0, \quad \forall E \in \mathcal{F}.$$

(A2) Unitarity: The probability that at least one of the elementary events in the entire sample space will occur is 1

$$P(\Omega) = 1.$$

(A3) σ-additivity: A countable sequence of disjoint sets (or *mutually exclusive* sets) $E_1, E_2, \ldots$ ($E_i \cap E_j = \emptyset$ for all $i \neq j$) satisfies

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i)$$

Any function that satisfies those three axioms can be a probability measure. These axioms lead to some useful properties of probability that we are familiar with.

(P1) The probability of the empty set:

$$P(\emptyset) = 0.$$

(P2) Monotonicity:

$$P(A) \leq P(B), \quad \forall A, B \in \mathcal{F} : A \subseteq B.$$

(P3) The numeric bound:

$$0 \leq P(E) \leq 1, \quad \forall E \in \mathcal{F}.$$

(P4) Union of two sets:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B), \quad \forall A, B \in \mathcal{F}.$$

(P5) Completion of a set:

$$P(A^c) = 1 - P(A), \quad \forall A \in \mathcal{F}.$$

B Random variables

Suppose we are given the triple $(\Omega, \mathcal{F}, P)$. A *real-valued* random variable is a function

$$X : \Omega \to \mathbb{R}$$

such that $\{\omega \in \Omega : X(w) \leq x\} \in \mathcal{F}$ for all $x \in \mathbb{R}$. We need this condition since we need the probability of this set in order to construct our cumulative distribution function.

Cumulative distribution function The probability distribution of X is mainly characterised by its cumulative distribution function (cdf) denoted as F, which is defined as

$$F(x) := P(X \leq x) = P(\{\omega \in \Omega : X(\omega) \leq x\}), \quad x \in \mathbb{R}.$$

There are three points to note here:

- The probability distribution of X is *induced* by P: There is always an implicit reference to $(\Omega, \mathcal{F}, P)$ when one calculates $P(X \leq x)$, but we tend to forget it once we have out cumulative distribution function F for X. This is because once we know F, we know everything about the probability distribution of X and usually we do not need to go back to the lower level and work with $(\Omega, \mathcal{F}, P)$ in practice. However, it may be useful to know what a random variable is in general.
- The use of $\leq$ (and not $<$) is important. Especially for discrete random variables, this matters a lot.
- Note that X, written in capital letter, represents the randomness in the probability statement while x is a given certain value in $\mathbb{R}$.

By definition, F has the following properties:

(P1) F is a non-decreasing function: For any $a, b \in \mathbb{R}$, if $a < b$, then $F(a) \leq F(b)$.

(P2) F is right continuous (no jumps occur when the limit point is approached from the right).

(P3) $\lim_{x\to-\infty} F(x) = 0$.

(P4) $\lim_{x\to\infty} F(x) = 1$.

Any function that satisfies those four properties can be a cdf. Therefore, the definition and the properties have an if and only if relation.

All the probability questions about X can be answered in terms of F. Examples:

- $P(X \in (a, b]) = P(X \le b) - P(X \le a) = F(b) - F(a)$
- $P(X = a) = F(a) - \lim_{x\to a^-} F(x)$. (the second term is a limit from the left)
- $P(X \in [a, b]) = P(X \in (a, b]) + P(X = a) = F(b) - F(a) + [F(a) - \lim_{x\to a^-} F(x)]$
- $P(X \in (a, b)) = P(X \in (a, b]) - P(X = b) = F(b) - F(a) - [F(b) - \lim_{x\to b^-} F(x)]$

Depending on the nature of set of values X takes, it can be called a discrete or a continuous random variable (sometimes neither of them!).

B.1 Discrete random variables

If X takes finite or countably infinite number of possible values in $\mathbb{R}$, then X is called a discrete random variable. The possible values of X may be listed as $x_1, x_2, \ldots$, where the sequence terminates in the finite case but continues indefinitely in the countably infinite case.

Let $p(x_i) := P(X = x_i)$, $i = 1, 2, \ldots$ The function $p(\cdot)$ is called the *probability mass function (pmf)* of X and has the following properties: $p(x_i) \ge 0$, $i = 1, 2, \ldots$ and $\sum_i p(x_i) = 1$.

It can be shown that, for any $x \in \mathbb{R}$,

$$F(x) = \sum_{i:x_i \le x} p(x_i).$$

Hence, the cdf F of X is a step function where jumps occur at points x_i with jump height being $p(x_i) = P(X = x_i) = F(x_i) - F(x_{i-1})$.

Some discrete distributions: Some well known distributions with a pmf (hence the cdf is a step function): Bernoulli $\mathcal{B}(\rho)$, Geometric distribution Geo(ρ), Binomial distribution Binom(n, ρ) Negative binomial NB(r, ρ), Poisson distribution $\mathcal{PO}(\lambda)$.

B.2 Continuous random variables

If X takes values on a continuous subset R_X of $\mathbb{R}$ (such as $\mathbb{R}$ itself, an interval $[a, b]$ or union of such intervals), then X is said to be a continuous random variable. Furthermore, if F for X is continuous (i.e. no jumps), we have

$$P(X \in (a, b)) = P(X \in (a, b]) = P(X \in [a, b)) = P(X \in [a, b]) = F(b) - F(a).$$

Also, if F is right differentiable, we can define the *probability density function (pdf)* for X

$$p(x) := \lim_{h \to 0} \frac{F(x+h) - F(x)}{h} = \frac{\partial_+ F(x)}{\partial x}, \quad x \in \mathbb{R}.$$

Since F is monotonic, we have $p(x) \geq 0$ for all $x \in \mathbb{R}$. Also, p integrates to 1 i.e. $\int_{-\infty}^{\infty} p(x)dx = \int_{R_X} p(x)dx = 1$. All probability statements for X can be calculated using f, such as

$$P(X \in [a, b]) = F(b) - F(a) = \int_a^b p(x)dx,$$
$$P(X \leq a) = F(a) = \int_{-\infty}^a p(x)dx.$$

From the above equation, we can conclude that $P(X = x) = 0$ for any $x \in \mathbb{R}$, because

$$\int_x^x p(x)dx = F(x) - F(x) = 0.$$

B.2.1 Some continuous distributions

The following are some well known distributions with a continuous cdf (hence admitting a pdf): Uniform distribution Unif(a, b), exponential distribution Exp(μ), gamma distribution $\Gamma(\alpha, \beta)$, inverse gamma distribution $\mathcal{IG}(\alpha, k)$, normal (Gaussian) distribution $\mathcal{N}(\mu, \sigma^2)$, Beta distribution Beta$(\alpha, \beta)$.

B.3 Moments, expectation and variance

If X is a random variable, the n'th *moment* of X, $n \geq 1$, denoted by $E(X^n)$, is defined for discrete and continuous random variables as follows:

$$E(X^n) := \begin{cases} \sum_i x_i^n p(x_i), & \text{if } X \text{ is discrete,} \\ \int_{-\infty}^{\infty} x^n p(x)dx, & \text{if } X \text{ is continuous.} \end{cases} \tag{A.1}$$

The first moment ($n = 1$) is called the *expectation* of X, also sometimes referred to as the mean of X.

If $|E(X)| < \infty$, the n'th central moments of X, $n \geq 1$, is defined for discrete and continuous random variables as follows:

$$E([X - E(X)]^n) := \begin{cases} \sum_i [x_i - E(X)]^n p(x_i), & \text{if } X \text{ is discrete,} \\ \int_{-\infty}^{\infty} [x - E(X)]^n p(x)dx, & \text{if } X \text{ is continuous.} \end{cases} \tag{A.2}$$

The second central moment is the most notable of them and is called the variance of X and denoted by $V(X)$:

$$\mathbb{V}(X) := E([X - E(X)]^2).$$

A useful identity relating $\mathbb{V}(X)$ to the expectation and the second moment of X is

$$\mathbb{V}(X) = E(X^2) - E(X)^2.$$

Finally, the standard deviation of X is

$$\sigma_X := \sqrt{\mathbb{V}(X)}.$$

B.4 More than one random variables

Suppose we have two real valued random variables, $X : \Omega \to \mathbb{R}$ and $Y : \Omega \to \mathbb{R}$, both defined on the same proabability space $(\Omega, \mathcal{F}, P)$.[1] The joint distribution of X and Y is characterised by the joint cdf $F_{X,Y}$ which is defined as

$$F_{X,Y}(x, y) := P(X \leq x, Y \leq y) = P(\{\omega \in \Omega : X(\omega) \leq x, Y(\omega) \leq y\}).$$

The marginal cdf's for X and Y can be deduced from $F_{X,Y}(x, y)$:

$$F_X(x) = \lim_{y\to\infty} F_{X,Y}(x, y), \quad F_Y(y) = \lim_{x\to\infty} F_{X,Y}(x, y).$$

Discrete variables: For discrete X and Y taking values x_i, $i = 1, 2, \ldots$ and y_j, $j = 1, 2, \ldots$, we can define a joint pmf $p_{X,Y}$ for X and Y such that

$$p_{X,Y}(x_i, y_j) := P(X = x_i, Y = y_j)$$

so that for any $x, y \in \mathbb{R}$, we have

$$F_{X,Y}(x, y) = \sum_{i,j:x_i\leq x, y_i\leq y} p_{X,Y}(x_i, y_j).$$

Expectation of any function g of X, Y can be evaluated using the joint pmf, for example

$$E(g(X, Y)) = \sum_{i,j} p_{X,Y}(x_i, y_j)g(x_i, y_j).$$

The *marginal pmf*'s for X and Y are given as follows:

$$p_X(x_i) = \sum_j p_{X,Y}(x_i, y_j), \quad p_Y(y_j) = \sum_i p_{X,Y}(x_i, y_j),$$

[1] (X, Y) together can be called a bivariate random variable. A generalisation of this is a multivariate random variable of dimension m, such as $(X_1, X_2, \ldots, X_m)$.

Continuous variables: Similar to the joint pmf defined for discrete X and Y, one can define the joint pdf for continuous X and Y, assuming F is right-differentiable,

$$p_{X,Y}(x,y) := \frac{\partial_+^2 F(x,y)}{\partial x \partial y}$$

so that for any a, b, we have

$$F_{X,Y}(a,b) = \int_{-\infty}^{b}\int_{-\infty}^{a} p_{X,Y}(x,y)dxdy$$

Expectation of any function g of X, Y can be evaluated using the joint pdf,

$$E(g(X,Y)) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p_{X,Y}(x,y)g(x,y)dxdy.$$

The *marginal pdf*'s for X and Y can be obtained

$$p_X(x) = \int_{-\infty}^{\infty} p_{X,Y}(x,y)dy, \quad p_Y(y) = \int_{-\infty}^{\infty} p_{X,Y}(x,y)dx,$$

Independence: We say random variables X and Y are independent if for all pairs of sets $A \subseteq \mathbb{R}$, $B \subseteq \mathbb{R}$ we have

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B).$$

If X and Y are discrete variables taking x_i, $i = 1, 2, \ldots$ and y_j, $j = 1, 2, \ldots$, then independence between X and Y can be expressed as

$$p_{X,Y}(x_i, y_j) = P(X = x_i, Y = y_j) = p_X(x_i)p_Y(y_j), \quad \forall i, j$$

If X and Y are continuous variables, then independence between X and Y can be expressed as

$$p_{X,Y}(x,y) = p_X(x)p_Y(y), \quad \forall x, y \in \mathbb{R}.$$

Covariance and Correlation: Covariance between two random variables X and Y, $\mathrm{Cov}(X,Y)$ is given as

$$\begin{aligned}\mathrm{Cov}(X,Y) &:= E([X - E(X)][Y - E(Y)]) \\ &= E(XY) - E(X)E(Y)\end{aligned}$$

A normalised version of covariance is *correlation* $\rho(X,Y)$. Provided that $\mathbb{V}(X) \geq 0$ and $\mathbb{V}(Y) \geq 0$,

$$\rho(X,Y) := \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y};$$

When one of $\mathbb{V}(X)$ and $\mathbb{V}(Y)$ is 0, we set $\rho(X,Y) = 1$ if $X = Y$ and $\rho(X,Y) = 0$ if $X \neq Y$. One can show that

$$-1 \leq \rho(X,Y) \leq 1.$$

Absolute value of $\rho(X,Y)$ indicates the level of correlation. We say two random variables X, Y are uncorrelated if $\mathrm{Cov}(X,Y) = 0$ (hence $\rho(X,Y) = 0$).

Note: Independence implies uncorrelatedness, but the reverse is not always true.

C Conditional probability and Bayes' rule

Consider the probability space $(\Omega, \mathcal{F}, P)$ again. Given two sets $A, B \in \mathcal{F}$, the conditional distribution of A given B is denoted by $P(A|B)$ and is defined as

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

The Bayes' rule is derived from this definition and it relates the two conditional probabilities $P(A|B)$ and $P(B|A)$:

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} \tag{A.3}$$

This relation can be written in terms of two random variables. Suppose X, Y are discrete random variables with joint pmf $p_{X,Y}(x_i, y_j)$, where $x \in \mathcal{X} = \{x_1, x_2, \ldots\}$ and $y \in \mathcal{Y} = \{y_1, y_2, \ldots\}$ so that the marginal pmf's are

$$p_X(x) = \sum_{y \in \mathcal{Y}} p_{X,Y}(x, y), \quad p_Y(y) = \sum_{x} p_{X,Y}(x, y), \quad x \in \mathcal{X}, y \in \mathcal{Y}.$$

Then the conditional pmf's $p_{X|Y}(x|y)$ and $p_{Y|X}(y|x)$ are defined as

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x, y)}{p_Y(y)}, \quad p_{Y|X}(y|x) = \frac{p_{X,Y}(x, y)}{p_X(x)} \tag{A.4}$$

and Bayes' rule relating them together is

$$p_{X|Y}(x|y) = \frac{p_X(x)p_{Y|X}(y|x)}{p_Y(y)} \tag{A.5}$$

When X, Y are continuous random variables taking values from $\mathcal{X}$ and $\mathcal{Y}$, respectively, with a joint pdf $p_{X,Y}(x, y)$, similar definitions follow: The marginal pdf's are

$$p_X(x) = \int_{-\infty}^{\infty} p_{X,Y}(x, y)dy, \quad p_Y(y) = \int_{-\infty}^{\infty} p_{X,Y}(x, y)dx.$$

The conditional pdf's are defined exactly the same way as in (A.4) and (A.5).

Appendix B

Discrete time Markov chains

Summary: *This chapter provides some basics of discrete time discrete space Markov chains.*

The review made here is very brief and limited by the relation of Markov chains to the topics covered in the course.

A Definition

Definition B.1 (**Markov chain**)**.** A stochastic process $\{X_n\}_{n\geq 1}$ on $\mathcal{X}$ is called a Markov chain if its probability law is defined from the initial distribution $\eta(x)$ and a sequence of Markov transition (or transition, state transition) kernels (or probabilities, densities) $\{M_n(x'|x)\}_{n\geq 2}$ by finite dimensional joint distributions as

$$p(x_1, \ldots, x_n) = \eta(x_1)M_2(x_2|x_1)\ldots M_n(x_n|x_{n-1})$$

for all $n \geq 1$.

The random variable X_t is called the *state* of the chain at time t and $\mathcal{X}$ is called the *state-space* of the chain. For uncountable $\mathcal{X}$, we have a discrete-time continuous-state Markov chain, and $\eta(\cdot)$ and $M_n(\cdot|x_{n-1})$ are pdf's[1]. Similarly, $\mathcal{X}$ is countable (finite or infinite), then the chain is a discrete-time discrete-state Markov chain and $\eta(\cdot)$ and $M_n(\cdot|x_{n-1})$ are pmf's. Moreover, when $\mathcal{X} = \{x_1, \ldots, x_m\}$ is finite with m states, the transition kernel can sometimes be expressed in terms of an $m \times m$ transition matrix $M_n(i, j) = P(X_n = j|X_{n-1} = i)$.

The definition of the Markov chain leads to the characteristic property of a Markov chain, which is also referred to as the *weak Markov property*: The current state of the chain at time n depends only on the previous state at time $n - 1$.

$$p(x_n|x_{1:n-1}) = p(x_n|x_{n-1}) = M_n(x_{n-1}, x_n)$$

From now on, we will consider *time-homogenous* Markov chains where $M_n = M$ for all $n \geq 2$, and we will denote them as Markov(η, M).

[1] In fact, there are exceptions where the transition kernels do not have a probability density; and this is indeed the case for the transition kernel of the Markov chain of the Metropolis-Hastings algorithm which we will see in Section C.3.1. However, for the sake of brevity we ignore this technical issue and with abuse of notation pretend as if we always have a density for $M_n(\cdot|x_{n-1})$ for continuous states

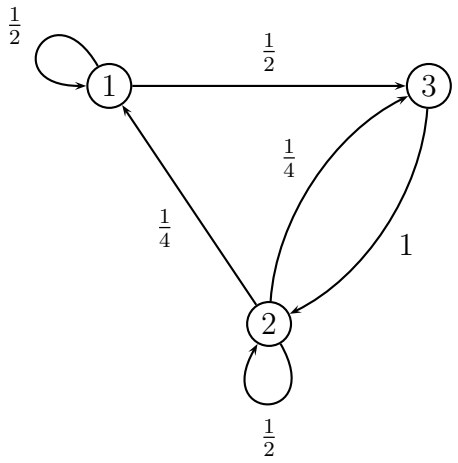

Figure B.1: State transition diagram of a Markov chain with 3 states, $1, 2, 3$.

Example B.1. The simplest examples of a Markov chain are those with a finite state-space, say of size m. Then, the transition rule can be expressed by an $m \times m$ transition probability matrix M, which in this example is the following

$$M = \begin{bmatrix} 1/2 & 0 & 1/2 \\ 1/4 & 1/2 & 1/4 \\ 0 & 1 & 0 \end{bmatrix}$$

Also, the state-transition diagram of such a Markov chain with $m = 3$ states is given in Figure B.1, where the state-space is simply $\{1, 2, 3\}$.

Example B.2. Let $\mathcal{X} = \mathbb{Z}$ be the set of integers, $X_1 = 0$, and for $n > 1$ define X_n as

$$X_n = X_{n-1} + V_n,$$

where $V_n \in \{-1, 1\}$ with $p = P(V_n = 1) = 1 - P(V_n = -1) = 1 - q$. This is a random walk (of step-size 1) on $\mathbb{Z}$ and it is a time homogenous discrete-time discrete state Markov chain with $\eta(x_1) = \delta_0(x_1)$ and

$$M(x'|x) = \begin{cases} p, & x' = x + 1 \\ q, & x' = x - 1 \end{cases}$$

When $p = q$, the process is called a symmetric random walk.

Example B.3. Let $\mathcal{X} = \mathbb{R}$, $X_1 = 0$, and for $n > 1$ define X_n as

$$X_n = X_{n-1} + V_n,$$

but this time $V_n \in \mathbb{R}$ with $V_n \sim \mathcal{N}(0, \sigma^2)$. This is a Gaussian random walk process on $\mathbb{R}$ with normally distributed step sizes, and it is a time homogenous discrete-time continuous state Markov chain with $\eta(x_1) = \delta_0(x_1)$ and

$$M(x'|x) = \phi(x'; x, \sigma^2).$$

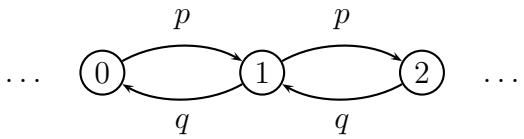

Figure B.2: State transition diagram of the symmetric random walk on $\mathbb{Z}$.

Example B.4. A generalisation of the Gaussian random walk is the first order autoregressive process, or shortly AR(1). Let $\mathcal{X} = \mathbb{R}$ the set of integers, $X_1 = 0$, and for $n > 1$ define X_n as

$$X_n = aX_{n-1} + V_n,$$

for some $a \in \mathbb{R}$, and $V_n \in \mathbb{R}$ with $V_n \sim \mathcal{N}(0, \sigma^2)$. AR(1) is a time homogenous discrete-time continuous state Markov chain with $\eta(x_1) = \delta_0(x_1)$ and

$$M(x'|x) = \phi(x'; ax, \sigma^2).$$

When $|a| < 1$, another choice for the initial distribution is $X_1 \sim \mathcal{N}(0, \frac{\sigma^2}{1-a^2})$, which is the stationary distribution of $\{X_t\}_{t\geq 1}$. We will see more on the stationary distributions below.

B Properties of Markov(η, M)

For MCMC, we require the Markov chain to have a unique invariant distribution π and to converge to π. Before discussing that, we need to review some fundamental properties of a discrete time Markov chain to understand when the existence of an invariant distribution and convergence to it are ensured. Those properties will be discussed in specific to discrete-state Markov chains only, for sake of simplicity and delivering the intuition behind the concepts. Although for general state-space Markov chains similar concepts also exist, they are more complicated and with less intuition, due to which we mostly omit them from our review.

B.1 Irreducibility

In a discrete state Markov chain, for two states $x, x' \in \mathcal{X}$, we say x *leads to* x' and show it by $x \to x'$ if the chain can travel from x to x' with a positive probability, i.e.

$$\exists n > 1 \text{ s.t. } P(X_n = x'|X_1 = x) > 0$$

If both $x \to x'$ and $x' \to x$, we say x and x' *communicate* and we show it by $x \leftrightarrow x'$.

A subset of states $C \subseteq \mathcal{X}$ is called a *communicating class*, or simply *class*, if (i) all $x, x' \in C$ communicate, and (ii) $x \in C$, $x \leftrightarrow y$ together imply $y \in C$, too (that is, there is no such $y \notin C$ such that $x \leftrightarrow y$ for some $x \in C$).

A communicating class is *closed* if $x \in C$ and $x \to y$ imply $y \in C$, that is there is no path with positive probability from outside the class to any of the states of the class.

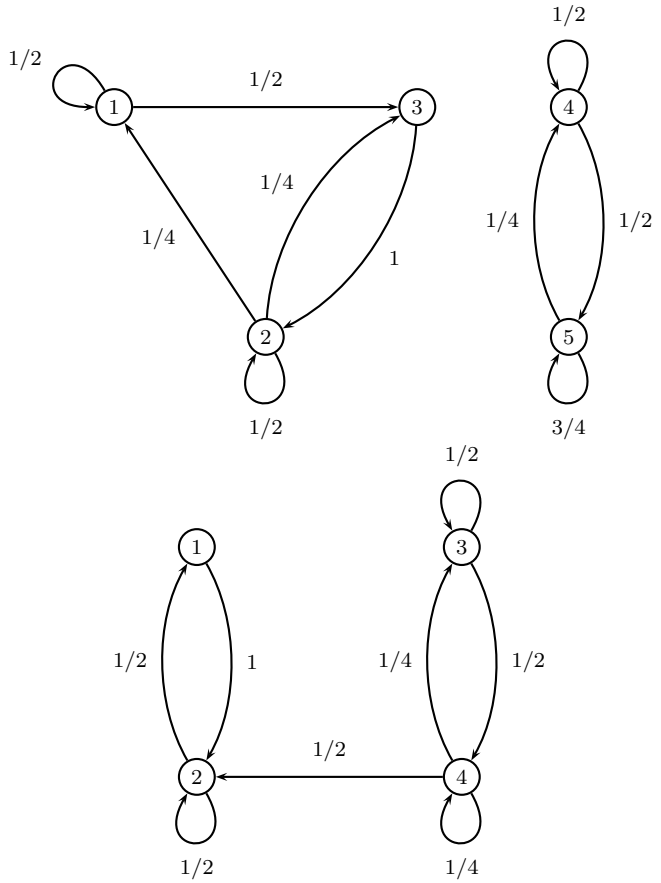

Figure B.3: State transition diagrams of two Markov chains that are not irreducible.

Definition B.2 (**Irreducibiliy**)**.** A discrete state Markov chain is called irreducible if the whole $\mathcal{X}$ is a communication class, i.e. all its states communicate.

For general state-spaces, we need to generalise the concept of irreducibility to ϕ-*irreducibility*.

Example B.5. Figure B.3 shows two chains that are not irreducible. In the first chain, the communication classes are $\{1, 2, 3\}$ and $\{4, 5\}$; both are closed. In the second chain, the communication classes are $\{1, 2\}$ and $\{3, 4\}$; the first one is closed and the second one is not.

B.2 Recurrence and Transience

In the discrete state-space, we say that a Markov chain is *recurrent* if every of its states is expected to be visited by the chain infinitely often, otherwise it is *transient*. More precisely, define the return time

$$\tau_x = \min\{n \geq 1 : X_{n+1} = x\}$$

Definition B.3 (**Recurrence**). We say the state $x \in \mathcal{X}$ is recurrent if

$$P(\tau_x < \infty | X_1 = x) = 1 \tag{B.1}$$

or equivalently $\sum_{n=1}^{\infty} P(X_n = x | X_1 = x) = \infty$. If a state is not recurrent, it is called transient.

If M is irreducible, then either every state is recurrent (and M is said to be recurrent) or every state is transient (and M is said to be transient).

Example B.6. The random walk on integers in Example B.2 is an irreducible chain. It can be shown that, in the symmetric case when $p = q = 1/2$, the chain is recurrent; if $p \neq q$, the chain is transient.

Definition B.4 (**Positive recurrence and null recurrence**). We say a state $x \in \mathcal{X}$ is positive recurrent if

$$E(\tau_x | X_1 = x) < \infty \tag{B.2}$$

(Note that (B.2) is a stronger condition than (B.1).) If a recurrent state is not positive recurrent, it is called null recurrent.

If M is irreducible and recurrent, then either every state is positive recurrent (and M is said to be positive recurrent) or every state is null recurrent (and M is said to be null recurrent).

To talk about recurrence in general state-space chains, instead of states we consider *accessible sets* in relation to ϕ-irreducibility.

Example B.7. It can be shown that the random walk on integers in Example B.2 is a null recurrent chain for $p = q = 1/2$.

B.3 Invariant distribution

A probability distribution π is called M-invariant if

$$\pi(x) = \int \pi(x') M(x|x') dx'$$

where we have assumed that $\{X_t\}_{t \geq 1}$ is continuous (hence π is a pdf). When $\{X_t\}_{t \geq 1}$ is discrete (hence π is a pmf), this relation is written as

$$\pi(x) = \sum_{x'} \pi(x') M(x|x')$$

The expressions on the RHS of the two equations above are short-handedly written as πM, so that for invariant π we have $\pi = \pi M$. In fact, when $\mathcal{X} = \{x_1, \ldots, x_m\}$ is finite with $M(i,j) = P(X_n = j | X_{n-1} = i)$ and $\pi = \begin{bmatrix} \pi(1) & \ldots & \pi(m) \end{bmatrix}$, we can indeed write $\pi = \pi M$ using notation for vector matrix multiplication.

Theorem B.1 (**Existence and uniqueness of invariant distribution**). *Suppose M is irreducible. M has a unique invariant distribution if and only if it is positive recurrent.*

Example B.8. The chain in Example B.1 has the invariant distribution $\pi = \begin{bmatrix}1/4 & 1/2 & 1/4\end{bmatrix}$. By solving $\mu = \mu M$, it can be shown that π is the only invariant distribution, so the chain is positive recurrent.

Example B.9. The random walk on integers in Example B.2 is irreducible. Therefore, it does not have an invariant distribution since it is not positive recurrent for any choice of $p = 1 - q$.

Example B.10. The Markov chain on top of Figure B.3 has two invariant distributions $\pi = \begin{bmatrix}1/4 & 1/2 & 1/4 & 0 & 0\end{bmatrix}$ and $\pi = \begin{bmatrix}0 & 0 & 0 & 1/3 & 2/3\end{bmatrix}$ although every state is positive recurrent. Note that the chain is not irreducible with two isolated communication classes, that is why Theorem B.1 is not applicable and uniqueness may not follow.

Example B.11. The Markov chain at the bottom of Figure B.3 is neither irreducible nor all of its states are positive recurrent (the states of the second class are transient). However, it has a unique invariant distribution, namely $\pi = \begin{bmatrix}1/3 & 2/3 & 0 & 0\end{bmatrix}$. Note that for this chain Theorem B.1 is not applicable since the chain is not irreducible.

B.4 Reversibility and detailed balance

One useful way for spotting the existence of an invariant probability measure for a Markov chain is to check for its *reversibility*, which is a sufficient (but not necessary) condition for existence of a stationary distribution.

Definition B.5 (**reversibility**). Let M be a transitional kernel having an invariant distribution and assume the associated Markov chain is started from π. We say that M is reversible if the reversed process $\{X_{n-m}\}_{0\leq m<n}$ is also $Markov(\pi, M)$ for all $n \geq 1$.

According to the definition above, M is reversible with respect to π if the backward transition density of the process $\{X_n\}_{n\geq 1}$ with $X_1 \sim \pi$ is the same as its forward transition density, i.e.

$$p(x_{n-1}|x_n) = \frac{p(x_{n-1})p(x_n|x_{n-1})}{p(x_n)} = \frac{p(x_{n-1})M(x_n|x_{n-1})}{\int p(x_{n-1})M(x_n|x_{n-1})dx_{n-1}} = M(x_{n-1}|x_n).$$

This immediately leads to the necessary and sufficient condition for reversibility of M is the detailed balance condition.

Proposition B.1 (**detailed balance**). *We say a Markov kernel M is reversible with respect to a probability distribution π if and only if the following condition, known as the detailed balance condition, holds:*

$$\pi(x)M(y|x) = \pi(y)M(x|y), \quad x, y \in \mathcal{X}.$$

Also, then π is an invariant distribution for M.

Being a sufficient condition for stationarity, the detailed balance condition is quite useful for designing transition kernels for MCMC algorithms.

ṣubsectionErgodicity Let π_n be the distribution of X_n of a Markov chain $\{X_n\}_{n\geq 1}$ with initial distribution η and transition kernels M. We have $\pi_1(x_1) = \eta(x_1)$ and the rest can be written recursively as $\pi_n = \pi_{n-1}M$, or explicitly

$$\pi_n(x_n) = \int \pi_{n-1}(x_{n-1})M(x_n|x_{n-1})dx_{n-1}$$

for continuous state chains, or

$$\pi_n(x_n) = \sum_{x_{n-1}\in\mathcal{X}} \pi_{n-1}(x_{n-1})M(x_n|x_{n-1}),$$

for discrete state chains, which reduces to

$$\pi_n = \pi_{n-1}M$$

when the state space is finite and π and M are considered as a vector and a matrix, respectively.

In MCMC methods that aim to approximately sample from π, we generate a Markov chain $\{X_n\}_{n\geq 1}$ with invariant distribution π and hope that for n large enough X_n is approximately distributed from π. This relies on the hope that π_n converges to π.

We have shown the conditions for a unique stationary distribution of a Markov chain. Note that having a unique invariant distribution does not mean that the chain will converge to its stationary distribution. For that to happen the Markov chain is required to have *aperiodicity*, a property which restricts the chain from getting trapped in cycles.

Definition B.6 (**aperiodicity**)**.** In a discrete state Markov chain, a state $x \in \mathcal{X}$ is called *aperiodic* if the set

$$\{n > 0 : P(X_{n+1} = x|X_1 = x)\}$$

has no common divisor other than 1. Otherwise, the state is *periodic* and its period is the greatest common divisor of state x. The Markov chain is said to be aperiodic if all of its states are aperiodic.

If the Markov chain is irreducible, then aperiodicity of one state implies the aperiodicity of all the states.

Definition B.7 (**ergodic state**)**.** A state is called ergodic if it is positive recurrent and aperiodic.

Finally, the definition of ergodicity for a Markov chain follows.

Definition B.8 (**ergodic Markov chain**)**.** An irreducible Markov chain is called ergodic if it is positive recurrent and aperiodic.

Ergodic chains ensure that the sequence of distributions $\{\pi_n\}_{n\geq 1}$ for $\{X_n\}_{n\geq 1}$ converge to the invariant distribution π.

Theorem B.2. *Suppose $\{X_n\}_{n\geq 1}$ is a discrete-state ergodic Markov chain with any initial distribution η and Markov transition kernel M with invariant distribution π. Then,*

$$\lim_{n\to\infty} \pi_n(x) = \pi(x) \tag{B.3}$$

In particular, for all $x, x' \in \mathcal{X}$,

$$\lim_{n\to\infty} P(X_n = x | X_1 = x') = \pi(x)$$

Example B.12. The Markov chain illustrated in Figure B.4 is irreducible and positive recurrent; so it has a unique invariant distribution, which is $\pi = \begin{bmatrix} 1/3 & 1/3 & 1/3 \end{bmatrix}$. However, it is periodic with period 3, and as a result π_n does not converge to π unless $\eta = \pi$. Indeed, one can show that for $\eta = \begin{bmatrix} \eta(1) & \eta(2) & \eta(3) \end{bmatrix}$, we have

$$\pi_n = \eta M^{n-1} = \begin{bmatrix} \eta(\mathrm{mod}(n-1,3)+1) & \eta(\mathrm{mod}(n-1,3)+2) & \eta(\mathrm{mod}(n-1,3)+3) \end{bmatrix}.$$

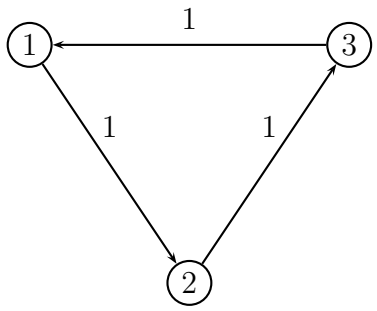

Figure B.4: An irreducible, positive recurrent, and periodic Markov chain.

Appendix C

Exact sampling methods

A Pseudo-random number generation

"The generation of random numbers is too important to be left to chance" and truly random numbers are impossible to generate on a deterministic computer. Published tables or other mechanical methods such as throwing dice, flipping coins, shuffling cards or turning the roulette wheels are clearly not very practical for generating the random numbers that are needed for computer simulations. Other techniques rely on chaotic behaviour, such as the thermal noise in Zener diodes or other analog circuits as well as the atmospheric noise (see, e.g., `www.Random.org`) or running a hash function against a frame of a video stream. Still, the vast amount of random numbers are obtained from pseudo-random number generators. Apart from being very efficient, one additional advantage of these techniques is that the sequences are reproducible by setting a *seed*, this property is key for debugging a Monte Carlo code.

Today, in most applications the task of random variable generation is performed on computers. In fact, a computer is mainly responsible for generating pseudo-random numbers that *look as if* they are independent and distributed uniformly from between 0 and 1, so goes the name "pseudo-random". That is, any sequence of pseudo-random numbers that are produced by the pseudo-random number generator should look like a sequence of i.i.d. uniformly distributed random numbers between 0 and 1, showing no correlation and spreading over the $(0, 1)$ interval uniformly.

There already exist highly sophisticated numerical methods to generate such pseudo-random numbers that pass certain tests for uniformity and independence. The most well known method for generating random numbers is based on a Linear Congruential Generator (LCG). The theory is well understood, and the method is easy to implement and fast. A LCG is defined by the recurrence relation:

$$x_{n+1} = (ax_n + c)(\text{mod } M)$$

If the coefficients a and c are chosen carefully (e.g. relatively prime to M), x_n will be roughly uniformly distributed between 0 and $M - 1$ (and with normalisation by M they can be shrunk between 0 and 1). By "roughly uniformly" we mean that the sequence of numbers x_n will pass many reasonable tests for randomness. One such test suite are the so called DIEHARD tests, developed by George Marsaglia, that are a battery of statistical tests for measuring the quality of a random number generator.

A more recently proposed generator is the Mersenne Twister algorithm, by Matsumoto and Nishimura, 1997. It has several desirable features such as a long period and being very fast. Many public domain implementations of the algorithm exist and it is the preferred random number generator for statistical simulations and Monte Carlo computations.

These random number generators provide uniformly distributed numbers on an interval. Hence, provided we have a good random number generator that can generate uniformly random integers (say between 0 and $2^{64}-1$), the resulting integers can be used to generate double precision numbers almost uniformly distributed in $(0,1)$.

B Some exact sampling methods

In the sequel, we will assume that a computer can produce for us an independent variable

$$U \sim \text{Unif}(0,1)$$

every time we ask it to do so. The crucial part is how to transform one or more copies of U such that the resulting number is distributed according to a particular distribution that we want to sample from. In a more general context, how can one exploit the ability of the computer to generate uniform random variables so that we can obtain random numbers from any desired distribution?

In the following we will see some exact sampling methods.

B.1 Method of inversion

Suppose $X \sim P$ taking values in $\mathcal{X} \subseteq \mathbb{R}$ with cdf F as defined above: $F(x) = \mathbb{P}(X \leq x)$, $x \in \mathbb{R}$. Recall that F takes values in $[0,1]$. Define the *generalised inverse cdf* $G : (0,1) \to \mathbb{R}$ as

$$G(u) := \inf\{x \in \mathcal{X} : F(x) \geq u\}. \tag{C.1}$$

Remark C.1. Define the set $S(u) = \{x \in \mathcal{X} : F(x) \geq u\}$. We can show that, by right-continuity of F, $S(u)$ actually attains its infimum, that is the minimum of $S(u)$ exists and hence $\inf S(u) = \min S(u)$, or $S(u) = [G(u), \infty)$[1].

If X is discrete taking values $x_1, x_2, \ldots$, this definition reduces to $G(u) = x_{i^*}$ where $i^* = \min\{i : F(x_i) \geq u\}$. In other words, $G(u) = x_{i^*}$ such that

$$F(x_{i^*-1}) < u \leq F(x_{i^*}). \tag{C.2}$$

If X is continuous with a pdf $p(x) > 0$ for all $x \in \mathcal{X}$, (i.e. F has no jumps and no flat parts in $\mathcal{X}$), then F is strictly monotonic in $\mathcal{X}$, its inverse $G = F^{-1}$ can be defined on $\mathcal{X}$, and we simply have $G(u) = F^{-1}(u)$.

The following Lemma enables the method of inversion.

[1]Proof: If $x < G(u)$, $x \notin S(u)$ by definition. If $x > G(u)$, then there exists $x' < x$ with $x' \in S(u)$; since $F$is non-decreasing, $F(x) \geq F(x') \geq u$, so $x \in S(u)$. Finally, by the right-continuity of F, we have $F(G(u)) = \inf F(y) : y > G(u) \geq u$. Therefore $G(u) \in S(u)$ and $S(u) = [G(u), \infty)$

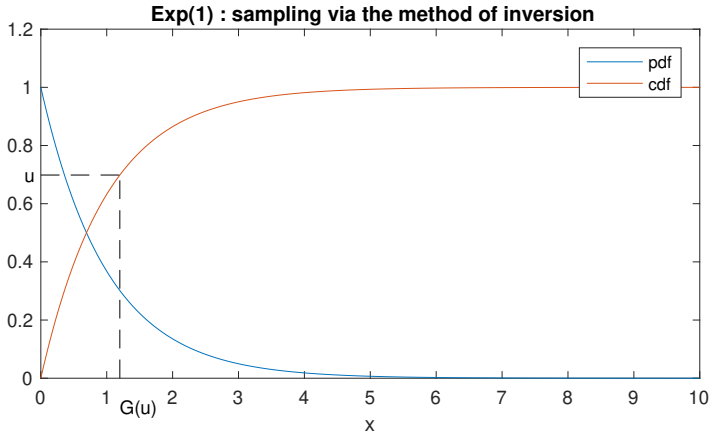

Figure C.1: Method of inversion for the exponential distribution

Lemma C.1. If $U \sim \text{Unif}(0,1)$, $G(U) \sim P$

Proof. Since $S(u) = [G(u), \infty)$ (see Remark C.1), we have $x \geq G(u)$ if and only if $F(x) \geq u$. Hence, $\mathbb{P}(X \leq x) = \mathbb{P}(G(U) \leq x) = \mathbb{P}(U \leq F(x)) = F(x)$. □

Lemma C.1 suggests we can sample $X \sim P$ by first sampling $U \in \text{Unif}(0,1)$ and then transforming $X = G(U)$. This approach is called the method of inversion.

Corollary C.1. Suppose F is continuous. If $X \sim P$, then $F(X) \sim \text{Unif}(0,1)$.

Proof. Since we have $S(u) = [G(u), \infty)$, $x \geq G(u)$ implies $F(x) \geq u$. Moreover, if $x < G(u)$ then $F(x) < u$ by definition of G. By continuity of F, we have $F(G(u)) = u$, so $F(x) \leq u$ if and only if $x \leq G(u)$. Hence $\mathbb{P}(F(X) \leq u) = \mathbb{P}(X \leq G(u)) = F(G(u)) = u$, and we conclude that the cdf of $F(X)$ is the cdf of $\text{Unif}(0,1)$. □

Example C.1. Suppose we want to sample $X \sim P = \text{Exp}(\lambda)$ from the exponential distribution with rate parameter $\lambda > 0$. The pdf of $\text{Exp}(\lambda)$ is

$$p(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & \text{else} \end{cases}.$$

The cdf is

$$u = F(x) = \begin{cases} \int_0^x \lambda e^{-\lambda t} dt = 1 - e^{-\lambda x}, & x \geq 0 \\ 0, & \text{else} \end{cases}.$$

Therefore, we have $x = -\log(1-u)/\lambda$. So, we can generate $U \sim \text{Unif}(0,1)$ and transform $X = -\log(1-U)/\lambda \sim \text{Exp}(\lambda)$. See Figure C.1 for an illustration.

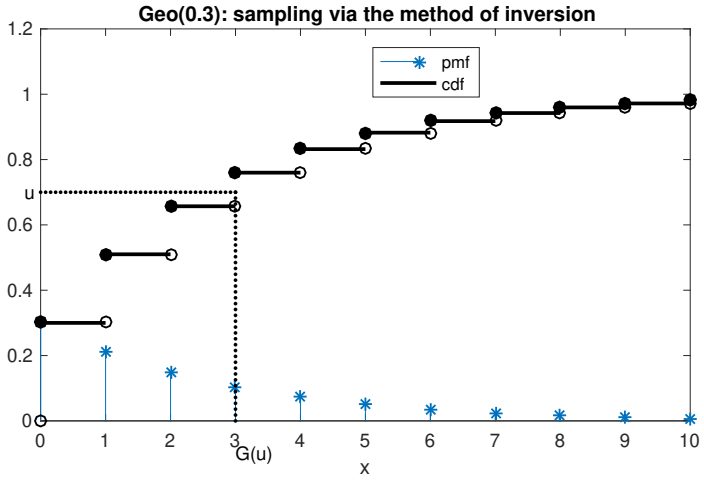

Figure C.2: Method of inversion for the geometric distribution

Example C.2. Suppose we want to sample $X \sim P = \text{Geo}(\rho)$ from the geometric distribution on $\mathcal{X} = \mathbb{N}$ with success rate parameter $\rho \in (0,1)$ and pmf[2]

$$p(x) = (1-\rho)^x \rho, \quad x = 0, 1, 2 \ldots.$$

Making use of $\sum_{i=0}^{x} \alpha^i = \frac{1-\alpha^{x+1}}{1-\alpha}$ with $\alpha = 1-\rho$, the cdf at the support points is given by

$$F(x) = 1 - (1-\rho)^{x+1}.$$

Given $U = u$ sampled from $\text{Unif}(0,1)$, the rule in (C.2) implies

$$1 - (1-\rho)^x < u \leq 1 - (1-\rho)^{x+1}$$

Solving the inequality for x we arrive at

$$\frac{\log(1-u)}{\log(1-\rho)} - 1 \leq x < \frac{\log(1-u)}{\log(1-\rho)}.$$

This is nothing but the round-up function written explicitly:

$$x = \left\lceil \frac{\log(1-u)}{\log(1-\rho)} - 1 \right\rceil.$$

See Figure C.2 for an illustration.

[2]This distribution is used for the number of trials prior to the first success in a Bernoulli process with success rate ρ. Another convention is to take the support range as $1, 2, \ldots$ rather than $0, 1, 2$ and interpret X as the number of trials until the successful trial, including the successful one. Then the pmf changes to $p(x) = (1-\rho)^{x-1}\rho, x \geq 1$

B.2 Transformation (change of variables)

The method of inversion can be seen as a transformation from U to $X = G(U)$. In fact, one can use transformation in a more general sense than using G by considering a change of variables via a suitable function g.

Example C.3. If we want to sample from $X \sim \text{Unif}(a, b)$, $a < b$, we can sample $U \sim \text{Unif}(0, 1)$ and use the transformation

$$X = g(U) := (b - a)U + a. \tag{C.3}$$

Transformation can also be used for more complicated situations than in Example C.3. Suppose we have an m-dimensional random variable $X \in \mathcal{X} \subseteq \mathbb{R}^m$ with pdf $p_X(x)$ and we apply a transform to X using an invertible function $g : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{Y} \subseteq \mathbb{R}^m$ to obtain

$$Y = (Y_1, \ldots, Y_m) = g(X_1, \ldots, X_m)$$

Since g is invertible, we have $X = g^{-1}(Y)$. What is the pdf of Y, $p_Y(y)$? This density can be found as follows: Define the Jacobian determinant (or simply Jacobian) of the inverse transformation g^{-1} as

$$J(y) = \det \frac{\partial g^{-1}(y)}{\partial y} \tag{C.4}$$

The usual practice to ease the notation is to introduce the short hand notation $(y_1, \ldots, y_m) = g(x_1, \ldots, x_m)$ and write $J(y)$ by making implicit reference to g as

$$J(y) = \det \frac{\partial x}{\partial y} = \det \frac{\partial(x_1, \ldots, x_m)}{\partial(y_1, \ldots, y_m)} = \det \begin{bmatrix} \partial x_1/\partial y_1 & \ldots & \partial x_1/\partial y_m \\ \vdots & \ddots & \vdots \\ \partial x_m/\partial y_1 & \ldots & \partial x_m/\partial y_m \end{bmatrix}$$

The Jacobian is useful for integration: If we make a change of variables from $x \to y$, we have to substitute $dx = |J(y)|dy$. When we apply this for the integral of any function $\varphi : \mathcal{X} \to \mathbb{R}$ with respect to $p_X(x)$, we have

$$\begin{aligned} \int p_X(x)\varphi(x)dx &= \int p_X(g^{-1}(y))\varphi(g^{-1}(y)) \, |J(y)| \, dy \\ &= \int p_X(g^{-1}(y)) \, |J(y)| \, \varphi(g^{-1}(y))dy \\ &= \int p_Y(y)\varphi(g^{-1}(y))dy \end{aligned}$$

where

$$p_Y(y) := p_X(g^{-1}(y)) \, |J(y)| \tag{C.5}$$

is the pdf of Y.

Change of variables can be useful when P is difficult to sample from using the method of inversion but $X \sim P$ can be performed by a certain transformation of random variables that are easier to generate, such as uniform random variables.

Example C.4. We describe the Box-Muller method for generating random variables from the standard normal (Gaussian) distribution $\mathcal{N}(0,1)$. The pdf for $\mathcal{N}(\mu,\sigma^2)$ is

$$\phi(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

The method of inversion is not an easy option to sample from $\mathcal{N}(0,1)$ since the cdf of $\mathcal{N}(0,1)$ is not easy to invert. Instead we use transformation.

The Box-Muller method generates a pair of independent standard normal random variables $X_1, X_2 \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ as follows: First we generate

$$R \sim \text{Exp}(1/2), \quad \Theta \sim \text{Unif}(0, 2\pi).$$

and then apply the transformation

$$X_1 = \sqrt{R}\cos(\Theta), \quad X_2 = \sqrt{R}\sin(\Theta)$$

If we wanted to start off from uniform random numbers, we could consider generating $U_1, U_2 \overset{i.i.d.}{\sim} \text{Unif}(0,1)$ and setting $R = -2\log(U_1)$ and $\Theta = 2\pi U_2$ so that R, Θ are distributed as desired. In other words,

$$X_1 = \sqrt{-2\log(U_1)}\cos(2\pi U_2), \quad X_2 = \sqrt{-2\log(U_1)}\sin(2\pi U_2)$$

One way to see why this works is to use change of variables. Note that[3]

$$(R, \Theta) = (X_1^2 + X_2^2, \arctan(X_2/X_1))) \tag{C.6}$$

Then the Jacobean at $(x_1, x_2) = (\sqrt{r}\cos\theta, \sqrt{r}\sin\theta)$ is

$$J(x_1,x_2) = \begin{vmatrix} \partial r/\partial x_1 & \partial r/\partial x_2 \\ \partial\theta/\partial x_1 & \partial\theta/\partial x_2 \end{vmatrix} = \begin{vmatrix} 2x_1 & 2x_2 \\ \frac{1}{1+(y_2/y_1)^2}\frac{-y_2}{y_1^2} & \frac{1}{1+(y_2/y_1)^2}\frac{1}{y_1} \end{vmatrix} = 2 \tag{C.7}$$

Therefore, we can apply (C.5) to get

$$\begin{aligned} p_{X_1,X_2}(x_1,x_2) &= p_R(r)p_\Theta(\theta)|J(x_1,x_2)| \\ &= p_R(x_1^2+x_2^2)p_\Theta(\arctan(x_2/x_1))|J(x_1,x_2)| \\ &= \frac{1}{2}e^{-\frac{1}{2}(x_1^2+x_2^2)}\frac{1}{2\pi}2 \\ &= \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x_1^2}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x_2^2} \\ &= \phi(x_1;0,1)\phi(x_2;0,1) \end{aligned} \tag{C.8}$$

which is the product of pdf of $\mathcal{N}(0,1)$ evaluated at x_1 and x_2. Therefore, we conclude that $X_1, X_2 \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$.

[3]To be precise, $\Theta = \arctan(X_2/X_1) + \pi\mathbb{I}(X_1 < 0)$ since $\Theta \in [0, 2\pi]$, but omitting the extra term $\pi\mathbb{I}(X_1 < 0)$ does not change the results.

Multivariate normal distribution: Another important transformation that we should be familiar with is a linear transformation of a multivariate normal random variable. We denote the distribution of an $n \times 1$ multivariable normal random variable as $X \sim \mathcal{N}(\mu, \Sigma)$ where $\mu = \mathbb{E}(X)$ is an $n \times 1$ *mean vector* and

$$\Sigma = \text{Cov}(X) = \mathbb{E}[(X - \mu)(X - \mu)^T]$$

is an $n \times n$ symmetric positive definite[4] *covariance matrix* The (i, j)'th element of Σ is

$$\sigma_{ij} = \text{Cov}(X_i, X_j) = \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)] = \mathbb{E}(X_i X_j) - \mu_i \mu_j$$

The pdf of this distribution is (using the same letter as for the pdf of the univariate normal distribution)

$$\phi(x; \mu, \Sigma) = \frac{1}{|2\pi\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)\right\} \tag{C.9}$$

where $|\cdot|$ stands for determinant.

Suppose $X = (X_1, \ldots, X_n)^T \sim \mathcal{N}(\mu, \Sigma)$ and we have the transformation

$$Y = AX + \eta$$

where A is an $m \times n$ matrix with $m \leq n$ with rank m[5], and η is an $m \times 1$ vector. We know for a fact that a linear transformation of X has to be normally distributed as well. Also, the normal distribution is completely characterised by its mean and covariance. Therefore, we can work out the mean and the variance of Y in order to identify its distribution.

$$\begin{aligned}
\mathbb{E}(Y) &= \mathbb{E}(AX + \eta) \\
&= A\mathbb{E}(X) + \eta \\
&= A\mu + \eta \\
\text{Cov}(Y) &= \mathbb{E}([Y - \mathbb{E}(Y)][Y - \mathbb{E}(Y)]^T) \\
&= \mathbb{E}([AX + \eta - (A\mu + \eta)][AX + \eta - (A\mu + \eta)]^T) \\
&= \mathbb{E}(A(X - \mu)(X - \mu)^T A^T) = A\text{Cov}(X)A^T \\
&= A\Sigma A^T
\end{aligned}$$

Therefore, $Y \sim \mathcal{N}(A\mu + \eta, A\Sigma A^T)$.

Example C.5. The above derivation suggests a way to generate an n-dimensional multivariate sample $X \sim \mathcal{N}(\mu, \Sigma)$. We can first generate i.i.d. normal random variables $R_1, \ldots, R_n \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$ so that $R = (R_1, \ldots, R_n) \sim \mathcal{N}(0_n, I_n)$ where 0_n is the $n \times 1$ vector of zeros and I_n is the identity matrix of size n. Then, we decompose $\Sigma = AA^T$ using the Cholesky decomposition. Finally, we let $X = AR + \mu$. Observe that the mean of X is $A0_n + \mu = \mu$ and covariance matrix of X is $AI_nA^T = AA^T = \Sigma$, so we are done.

[4] In fact, positive semi-definite covariance matrices are also allowed, however the distribution is called degenerate and it does not have a pdf.

[5] We constraint A to full row rank matrices since otherwise the resulting covariance matrix for $A\Sigma A^T$ is no longer positive definite and Y is degenerate.

B.3 Composition

Let a random variable $Z \sim \Pi$ taking values from the set $\mathcal{Z}$ and Π has a pdf or pmf shown as $\pi(z)$. Suppose also that given z, $X|z \sim P_z$ where each P_z admits either a pmf or a pdf shown as $p_z(x)$. Then the marginal distribution P is a *mixture distribution* and in the presence of pdf's or pmf's, we have

$$p(x) = \begin{cases} \int p_z(x)\pi(z)dz, & \text{if } \pi(z) \text{ is a pdf} \\ \sum_z p_z(x)\pi(z), & \text{if } \pi(z) \text{ is a pmf} \end{cases} \tag{C.10}$$

Whether $p(x)$ is pmf or a pdf depends on whether $p_z(x)$ is pmf or pdf. The integral/sum may be hard to evaluate, and the mixture distribution may be hard to sample directly. But if we can easily sample from Π and from each P_z, then we can just

1. sample $Z \sim \Pi$,
2. sample $X \sim P_Z$, and
3. ignore Z and return X.

The random number we produce in this way will be an exact sample from P, i.e. $X \sim P$. This is the method of *composition*. Ignoring Z is also called *marginalisation*, by which we overcome the difficulty of dealing with the tough integral/sum in (C.10).

Example C.6. The density of a mixture of Gaussian distribution with K components with means and variance values $(\mu_1, \sigma_1^2), \ldots, (\mu_K, \sigma_K^2)$, and probability weights $w_1, \ldots, w_K$ for its components (such that $w_1 + \cdots + w_K = 1$) is given by

$$p(x) = \sum_{k=1}^{K} w_k \phi(x; \mu_k, \sigma_k^2).$$

To sample from $p(x)$, we first sample the component number k with probability w_k (for example using the method of inversion), and given k, we sample $X \sim \mathcal{N}(\mu_k, \sigma_k^2)$

Example C.7. A sales company decides to reveal the demand D for a product over a month. However, for privacy reasons, it shares this average by adding some noise to D, which results in the shared value X. It is given that the distribution of the revealed demand X has the pdf

$$p(x) = \sum_d \left[\frac{e^{-\lambda}\lambda^d}{d!}\right] \left[\frac{1}{2b} \exp\left(-\frac{|x-d|}{b}\right)\right]$$

We want to perform a Monte Carlo simulation for this data sharing process. How do we sample $X \sim P$?

Although $p(x)$ looks hard, observe that the first term in the sum is the pmf of $\mathcal{PO}(\lambda)$ evaluated at d (can be viewed as the demand) and the second term in the sum is the pdf of

Laplace(d, b) evaluated at x (can be viewed as the noisy demand)[6]. Therefore, generation of X is possible by the method of composition as

1. Sample $D \sim \mathcal{PO}(\lambda)$,
2. Sample $X \sim$ Laplace(D, b) (equivalent to $V \sim$ Laplace$(0, b)$ and $X = D + V$.).
3. Ignore D and return X.

It is an exercise for you to figure out how one can sample from the Poisson and Laplace distributions.

[6]The pmf of $\mathcal{PO}(\lambda)$ evaluated at k is $\frac{e^{-\lambda}\lambda^k}{k!}$, and the pdf of Laplace$(\mu, b)$ evaluated at x is $\frac{1}{2b}\exp\left(-\frac{|x-\mu|}{b}\right)$

Appendix D

Term project: A toolbox for ANOVA and Linear regression

In your term project, you will submit a small toolbox for ANOVA and linear regression, which is package of functions, written in a language that you are comfortable with. The toolbox will contain the functions named and described below, as well as a main script that demonstrates the use of the functions with examples. A readme text file describing the functions and explaining how to use the toolbox in general should also be added.

ANOVA

- `ANOVA1_partition_TSS`

  Write a function that partitions the sum of squares in a one way ANOVA layout. The function should take a data set $X_{i,j}$ for $j = 1, \ldots, n_i$ and $i = 1, \ldots, I$, and return SS_{total}, SS_w, and SS_b in (2.2).

- `ANOVA1_test_equality`

  Write a function that tests the equality of the means in a one way ANOVA layout. The input is a data set $X_{i,j}$ for $j = 1, \ldots, n_i$ and $i = 1, \ldots, I$, and the significance level α. As the output, the function prints all the quantities in the table at the end of Section A.2, as well as the critical value, the p-value, and the decision.

- `ANOVA1_is_contrast`

  Write a function that takes $c_1, c_2, \ldots, c_I$ as an input and determines whether or not a linear combination defined by $c_1, c_2, \ldots, c_I$ is a contrast.

- `ANOVA1_is_orthogonal`

  Write a function that takes group sizes $n_1, \ldots, n_I$ and coefficients $c_{1,1}, c_{1,2}, \ldots, c_{1,I}$ and $c_{2,1}, c_{2,2}, \ldots, c_{2,I}$ and determines whether or not the corresponding contrasts are orthogonal. Your function should return a warning if any of the linear combinations is not a contrast.

- `Bonferroni_correction`

  Write a function that takes a FWER α and the number of tests m and determines the significance level that each test must have with Bonferroni correction.

- `Sidak_correction`

  Write a function that takes a FWER α and the number of tests m and determines the significance level of each test with Sidak's correction.

- `ANOVA1_CI_linear_combs`

  Write a function whose inputs and outputs are follows:

  Input:

  - a data set $X_{i,j}$ for $j = 1, \ldots, n_i$ and $i = 1, \ldots, I$,
  - Significance level α,
  - An $m \times I$ matrix C, where each row defines linear combination of the group means.
  - Method: This may be "Scheffe", "Tukey", "Bonferroni", "Sidak", "best"

  Output: As the output, the function should return simultaneous confidence intervals for those linear combinations.

  Your function should determine whether the chosen method is valid for the inputted linear combinations. If not, it should give a warning and not return anything. (For example, Tukey's confidence intervals are valid for pairwise comparisons only.)

  If "Scheffe" is chosen, the function should choose between Theorem 2.7 or Theorem 2.8 depending on whether the linear combination is a contrast or not.

  If "best" is chosen, your function should check the number and the nature of the linear combinations and choose the best method among the valid ones, i.e. the method with narrowest confidence intervals.

  Here are some example cases:

  - If all linear combinations are contrasts, you have several options:
    - If those contrasts are orthogonal as well, you should compare Theorem 2.8 with Sidak's correction.
    - If all linear combinations are contrasts but not orthogonal, then you should compare Theorem 2.8 with Bonferroni's correction.
    - If all linear combinations are pairwise differences, then you should compare Tukey's confidence intervals and Bonferroni's correction.
    - If all linear combinations are pairwise differences and orthogonal, then you should compare Tukey's confidence intervals with Sidak's correction.
  - If not all linear combinations are contrasts, you have Theorem 2.7 and Bonferroni's correction:
  - etc.

- `ANOVA1_test_linear_combs`

  Write a function whose inputs and outputs are follows:

  Input:

  - a data set $X_{i,j}$ for $j = 1, \ldots, n_i$ and $i = 1, \ldots, I$,
  - FWER α,
  - An $m \times I$ matrix C and a $m \times 1$ vector d, where each row of C defines linear combination of the group means and each element of d is the hypothesized value for the corresponding combination.

    $$H_0 : c_{i,1}\mu_1 + \ldots + c_{i,I}\mu_I = d_i, \quad i = 1, \ldots, m.$$

  - Method: This may be "Scheffe", "Tukey", "Bonferroni", "Sidak", "best".

  Output: As the output, the function should return the test outcomes, with p-values in such a way that FWER is kept at α.

  The comments for the previous questions apply to this one, too.

- `ANOVA2_partition_TSS`

  Write a function that takes $X_{i,j,k}$ for $i = 1, \ldots, I$, $j = 1, \ldots, J$, and $k = 1, \ldots, K$ in a two way ANOVA layout and returns SS_{total}, SS_A, SS_B, SS_{AB}, and SS_E.

- `ANOVA2_MLE`

  Write a function that takes $X_{i,j,k}$ for $i = 1, \ldots, I$, $j = 1, \ldots, J$, and $k = 1, \ldots, K$ in a two way ANOVA layout and returns the maximum likelihood estimates for the parameters μ, a_i, b_j, δ_{ij}.

- `ANOVA2_test_equality`

  Write a function that performs one of the basic three tests in the two-way ANOVA layout. The function takes $X_{i,j,k}$ for $i = 1, \ldots, I$, $j = 1, \ldots, J$, and $k = 1, \ldots, K$, and a significance level α and performs one of the following (depending on the choice)

  - The hypothesis that $a_1 = \ldots = a_I = 0$.
  - The hypothesis that $b_1 = \ldots = b_J = 0$.
  - The hypothesis that all δ_{ij}'s are equal to 0.

  The choice for the test should also be inputted as an input as either "A", "B", or "AB". The function should print the relevant rows of the table below, depending on

the test to be run:

Source	degrees of freedom	SS	MS	F
A	$I-1$	SS_A	MS_A	MS_A/MS_E
B	$J-1$	SS_B	MS_B	MS_B/MS_E
$A \times B$	$(I-1)(J-1)$	SS_{AB}	MS_{AB}	MS_{AB}/MS_E
within	$IJ(K-1)$	SS_E	MS_E	
Total	$IJK-1$	SS_{total}		

Linear regression

- `Mult_LR_Least_squares`

  Write a function that finds the least squares solution according to the multiple linear regression model: The function takes X and y, the design matrix and the response vector, and produces the maximum likelihood estimators for β and σ^2 as well as the unbiased estimate for σ^2.

- `Mult_LR_partition_TSS`

  Write a function that takes an $n \times (k+1)$ matrix X, $n \times 1$ vector y as inputs and returns the total sum of squares, regression sum of squares, and residual sum of squares.

  The rest of the questions will be answered according to the normal multiple linear regression model.

- `Mult_norm_LR_simul_CI`

  Write a function that takes X and y, the design matrix and the response vector, and a significance parameter α, and produces confidence intervals for β_i's that simultaneously hold with probability $1-\alpha$.

- `Mult_norm_LR_CR`

  Write a function that takes an $n \times (k+1)$ matrix X, $n \times 1$ vector y, an $r \times (k+1)$ matrix C with rank r, and a significance level α as inputs, and returns the specifications (that is, parameters of the ellipsoid) of the $100(1-\alpha)\%$ confidence region for $C\beta$ according to the normal multiple linear regression model.

- `Mult_norm_LR_is_in_CR`

  Write a function that takes an $n \times (k+1)$ matrix X, $n \times 1$ vector y, an $r \times (k+1)$ matrix C with rank r, a $r \times 1$ vector c_0, and a significance level α as inputs, and answers whether c_0 is in the $100(1-\alpha)\%$ confidence region for $C\beta$ according to the normal multiple linear regression model.

- `Mult_norm_LR_test_general`

  Write a function that takes an $n \times (k+1)$ matrix X, $n \times 1$ vector y, an $r \times (k+1)$ matrix C with rank r, a $r \times 1$ vector c_0, and a significance level α as inputs, and tests the null hypothesis $H_0 : C\beta = c_0$ vs $H_1 : C\beta \neq c_0$ at a significance level of α. [Note that you can use this function to test any hypothesis regarding linear regression]

- `Mult_norm_LR_test_comp`

  Write a function that takes an $n \times (k+1)$ matrix X, $n \times 1$ vector y, a significance level α, and $j_1, \ldots, j_r \in \{0, \ldots, k\}$ as inputs, and returns the outcome of testing $H_0 : \beta_{j_1} = \ldots = \beta_{j_r} = 0$ vs H_1 : not H_0. You can use the previous function with a suitable C and c_0.

- `Mult_norm_LR_test_linear_reg`

  Write a function that takes an $n \times (k+1)$ matrix X, $n \times 1$ vector y, a significance level α as inputs, and returns the outcome of testing the existence of linear regression at all, i.e., $H_0 : \beta_1 = \ldots = \beta_k = 0$ vs H_1 : not H_0. You can use the previous function with a suitable $j_1 = 1, \ldots, j_k = k$.

- `Mult_norm_LR_pred_CI`

  Write a function that takes a $n \times (k+1)$ matrix X, $n \times 1$ vector y, a $m \times (k+1)$ matrix D, a significance level α, and a method as inputs, and returns simultaneous confidence bounds for $d_i\beta$ for all $i = 1, \ldots, m$ according to the normal multiple linear regression model, where d_i is the i'th row of the matrix. The bounds should hold simultaneously with probability $1 - \alpha$ at least. The possible choices for the method are "Bonferroni", "Scheffe", and "best", where "best" is the best of the two.

Main script

Finally, write a main script that demonstrates all the functions in action. The script should have an example for each function whose output for a sample input is printed on the screen. You can generate your own data sets for the examples, or obtain suitable sample datasets from other sources. I will start checking your code by running this main script.