1.2 Unbiased and Consistent Estimators

When we estimate something we want to be unbiased. It is desirable that, at least on average, our estimate is correct. For example, assume that based on the data from the previous example, we predict the average accommodation expenditure for 10 specific students to be 500 pounds. Sometimes we will over-predict and sometimes we will under-predict, but we do not want to systematically over-predict or under-predict. In addition, it is desirable that the more data we collect, the closer our estimate will be to the true value of the population characteristic. In our example, if we measure 60 students before we make inferences on the average expenditure, we should be more accurate compared to measuring 6 students.

Mathematically, assume we will obtain observations from random variables $X_1, \ldots, X_n$, and that we wish to estimate a parameter θ, such as their mean or variance. An estimator for θ is a function of the data $T = f(X_1, \ldots, X_n)$. We introduce the following definitions:

Definition

An estimator $T = f(X_1, \ldots, X_n)$ is an unbiased estimator of the parameter θ if,

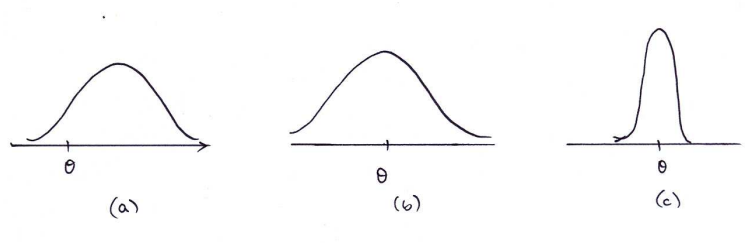
$$\mathbb{E}(T) = \theta.$$

Definition

An estimator T is a consistent estimator of the parameter θ if,

1. $\mathbb{E}(T) = \theta$ (i.e. it is an unbiased estimator of θ) as $n \to \infty$; and
2. $\mathrm{Var}(T) \to 0$ as the sample size $n \to \infty$.

Unbiased estimators are generally desirable, since they provide an accurate estimate of the parameter of interest "on average". Similarly, an estimator with small variability is also desirable, with little uncertainty concerning the precision of the estimation.



1.2.1 Sample Mean and Variance

Let $X_1, \ldots, X_n$ be independent rvs each with mean μ and variance σ^2. The sample mean is defined by $\bar{X}$, where,

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

The sample variance is denoted by S^2 and given by,

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

Aside: Note that we can write,

$$S^2 = \frac{1}{n-1} \left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2 \right) = \frac{1}{n-1} \left(\sum_{i=1}^{n} X_i^2 - \frac{1}{n} \left(\sum_{i=1}^{n} X_i \right)^2 \right)$$

Proof:

$$\begin{aligned}
S^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}(X_i^2 - 2X_i\bar{X} + \bar{X}^2) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2 - 2\bar{X}\sum_{i=1}^{n}X_i + \sum_{i=1}^{n}\bar{X}^2\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2 - 2\bar{X}n\bar{X} + n\bar{X}^2\right) \qquad \text{since } \textstyle\sum_{i=1}^{n}X_i = n\bar{X} \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2 - n\bar{X}^2\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2 - n\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)^2\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n}X_i\right)^2\right)
\end{aligned}$$

This expression greatly simplifies the calculation of the sample variance for a given set of observations.

Theorems

1. The sample mean $\bar{X}$ is an unbiased and consistent estimator of the population mean μ.
2. The sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is an unbiased estimator of σ^2.

Proofs:

1. For $\bar{X}$ to be an unbiased estimator of μ, we need to show $\mathbb{E}(\bar{X}) = \mu$. Now,

$$\begin{aligned}
\mathbb{E}(\bar{X}) &= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) \\
&= \frac{1}{n}n\mu \\
&= \mu.
\end{aligned}$$

   For X to be a consistent estimator of μ, we need that $\mathrm{Var}(\bar{X}) \to 0$ as $n \to \infty$.

$$\begin{aligned}
\mathrm{Var}(\bar{X}) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) \qquad \text{since the } X_i \text{ are independent} \\
&= \frac{1}{n^2}n\sigma^2 \\
&= \frac{\sigma^2}{n}.
\end{aligned}$$

   So, as $n \to \infty$, $\mathrm{Var}(\bar{X}) \to 0$.

2. (Note this proof was asked as part of the May 2009 exam)

   We saw earlier that,

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

   Aside: note that for any rv X,

$$\begin{aligned} \mathrm{Var}(X) &= \mathbb{E}(X^2) - (\mathbb{E}(X))^2 \\ \Rightarrow \mathbb{E}(X^2) &= \mathrm{Var}(X) + (\mathbb{E}(X))^2. \end{aligned}$$

   Then,

$$\begin{aligned} \mathbb{E}(S^2) &= \mathbb{E}\left(\frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right)\right) \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n}\mathbb{E}(X_i^2) - n\mathbb{E}(\bar{X}^2)\right) \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n}(\mathrm{Var}(X_i) + (\mathbb{E}(X_i))^2) - n(\mathrm{Var}(\bar{X}) + (\mathbb{E}(\bar{X}))^2)\right) \\ &= \frac{1}{n-1}\left(n\sigma^2 + n\mu^2 - n\left(\frac{\sigma^2}{n} + \mu^2\right)\right) \\ &= \frac{1}{n-1}(\sigma^2(n-1)) \\ &= \sigma^2. \end{aligned}$$

   This is the reason we define the sample variance with a denominator of $n-1$ (rather than n).

□

Notes

- Note the distinction between the population and sample mean and variance. The population mean and variance ($\mu = \mathbb{E}(X)$ and $\sigma^2 = \mathrm{Var}(X)$) are the moments of the rv X; the sample mean and variance ($\bar{X}$ and S^2) are estimators of μ and σ^2, based on observations from X, and so subject to random variability.
- Returning to the simple example at the start of this Chapter, based on the six observations, an unbiased (but precise for $n = 6$ only?) estimate for the average student accommodation expenditure in St Andrews is, $\bar{x} = (1/6) \times \sum_{i=1}^{6} x_i = 500$.
- S^2 is also a consistent estimator, though we omit the proof here.
- We have that $\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$ so that we can estimate $\mathrm{Var}(\bar{X})$ by $\frac{S^2}{n}$. (The standard deviation of an estimator is also called the 'standard error')
- It is possible that an estimator is consistent, but not unbiased. Consider the following example.

Example

Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ contain independent rvs each with mean μ. Let T be the estimator of μ such that $T = \frac{1}{n-1}\sum_{i=1}^{n} X_i$. Then,

1. T is a biased estimator of μ; but
2. T is a consistent estimator of μ.

For simplicity, we note that $T = \frac{n}{n-1}\bar{X}$.

Then, we have that

$$\mathbb{E}(T) = \frac{n}{n-1}\mathbb{E}(\bar{X}) = \frac{n}{n-1}\mu \quad (\neq \mu).$$

However, $\mathbb{E}(T) \to \mu$ as $n \to \infty$.

Similarly, we have that,

$$\mathrm{Var}(T) = \left(\frac{n}{n-1}\right)^2 \mathrm{Var}(\bar{X}) = \frac{n}{(n-1)^2}\sigma^2,$$

since $\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$. Thus, $\mathrm{Var}(T) \to 0$, as $n \to \infty$.

Example

Let $X_1, \ldots, X_n$ be independent rvs each with mean μ and variance σ^2. Consider the following two estimators of μ, $f(\boldsymbol{X}) = \bar{X}$ and $g(\boldsymbol{X}) = X_n$. We have,

$$\mathbb{E}(\bar{X}) = \mu; \qquad \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}; \qquad \mathbb{E}(X_n) = \mu; \qquad \mathrm{Var}(X_n) = \sigma^2.$$

Thus, $\bar{X}$ would be the preferred estimator since $\mathbb{E}(\bar{X}) = \mathbb{E}(X_n)$ but $\mathrm{Var}(\bar{X}) \leq \mathrm{Var}(X_n)$ for all $n \geq 1$.

Example

(From class test 2009) The pdf of the continuous random variable Y is given by $f(y) = 1/(\theta - 1)$ for $1 \leq y \leq \theta$ and zero elsewhere. Which of the following estimators is unbiased for θ?

(a) $2Y - 1$.

(b) $Y - 1$.

(c) $2Y + 1$.

(d) $(2Y + 1)/2$.

For T to be an unbiased estimator, $\mathbb{E}(T) = \theta$. Now we have,

$$\begin{aligned}
\mathbb{E}(Y) &= \int_{-\infty}^{\infty} y f(y) dy \\
&= \int_1^{\theta} \frac{y}{\theta - 1} dy \\
&= \frac{1}{\theta - 1}\left[\frac{y^2}{2}\right]_1^{\theta} \\
&= \frac{1}{2(\theta - 1)}(\theta^2 - 1) \\
&= \frac{(\theta - 1)(\theta + 1)}{2(\theta - 1)} \\
&= \frac{\theta + 1}{2}.
\end{aligned}$$

Thus, $\mathbb{E}(2Y - 1) = 2\mathbb{E}(Y) - 1 = \theta + 1 - 1 = \theta$. Thus (a) is an unbiased estimator of θ. s

Example

Suppose that $X_1, \ldots, X_n$ are independent rvs from the distribution with pdf,

$$f(x) = \frac{2x}{\theta^2}, \qquad \text{for } 0 \leq x \leq \theta.$$

We wish to find k such that $k\bar{X}$ is an unbiased estimator of θ.

We note that for $i = 1, \ldots, n$, then,

$$\begin{aligned}
\mathbb{E}(X_i) &= \int_{-\infty}^{\infty} x_i f(x_i) dx_i \\
&= \int_0^{\theta} x_i \left(\frac{2x_i}{\theta^2}\right) dx_i \\
&= \frac{2}{\theta^2}\left[\frac{x_i^3}{3}\right]_0^{\theta} \\
&= \frac{2}{3}\theta.
\end{aligned}$$

Additionally, we have that $\mathbb{E}(k\bar{X}) = k\mathbb{E}(\bar{X})$ and we know that $\mathbb{E}(\bar{X}) = \mathbb{E}(X_i) = \frac{2}{3}\theta$, so that,

$$\mathbb{E}(k\bar{X}) = \frac{2}{3}k\theta.$$

So for $k\bar{X}$ to be an unbiased estimator of θ, we have $k = \frac{3}{2}$.

Is this estimator a consistent estimator? We know that it is unbiased for all n, so that it is unbiased as $n \to \infty$. Thus, to show it is a consistent estimator we also need to show that the variance of the estimator tends to zero as $n \to \infty$. To calculate the variance, initially consider,

$$\begin{aligned}
\mathbb{E}(X_i^2) &= \int_0^{\theta} x_i^2 \left(\frac{2x_i}{\theta^2}\right) dx_i \\
&= \frac{2}{\theta^2}\left[\frac{x_i^4}{4}\right]_0^{\theta} \\
&= \frac{\theta^2}{2}.
\end{aligned}$$

Then,

$$\begin{aligned}
\mathrm{Var}(X_i) &= \mathbb{E}(X_i)^2 - (\mathbb{E}(X_i))^2 \\
&= \frac{\theta^2}{2} - \left(\frac{2}{3}\theta\right)^2 \\
&= \theta^2\left(\frac{1}{2} - \frac{4}{9}\right) \\
&= \frac{\theta^2}{18}.
\end{aligned}$$

Now, we know that $\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$ where $\sigma^2 = \mathrm{Var}(X_i)$ for $i = 1, \ldots, n$. Then,

$$\mathrm{Var}(k\bar{X}) = k^2\mathrm{Var}(\bar{X}) = k^2.\frac{1}{n}\left(\frac{\theta^2}{18}\right) = \frac{\theta^2}{8n}.$$

Then, as $n \to \infty$, $\mathrm{Var}(k\bar{X}) \to 0$.

Hence $k\bar{X}$ $(k = 3/2)$ is a consistent estimator of θ.

1.2.2 Sample Covariance and Correlation

Recall that $\mathrm{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$ and the correlation is given by $\rho(X, Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$. Correlation is a measure of the strength of the linear association between two continuous X and Y. For example, if X fully defines Y, but in a quadratic fashion, say $Y = X + X^2$, their correlation will not be one. Correlation is invariant to scale, in contrast to the Covariance.

Suppose that you have paired random variables X and Y, that will provide observations $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$. The covariance of X and Y can be estimated using the *sample covariance*:

$$S_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i Y_i - \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)\right].$$

An estimator for the correlation coefficient, $\rho(X,Y)$, is simply $R_{xy} = \frac{S_{xy}}{S_x S_y}$ where S_x is the sample standard deviation of the X_i (i.e. $\sqrt{S_{xx}}$), and similarly for S_y. For Normally distributed X and Y this is a consistent estimator, although it is not unbiased. (Proof is omitted). It is not a robust estimator, i.e. it is considerably affected by departures from Normality and the presence of outlier observations. An observation is typically considered to be an outlier if it is away from the sample mean by more than three or four sample standard deviations. Once more note the distinction between the population covariance and correlation ($\mathrm{Cov}(X,Y)$ and $\rho(X,Y)$) which are the properties of the rvs X and Y; and the sample covariance and correlation (S_{xy} and R_{xy}) which are estimators of $\mathrm{Cov}(X,Y)$ and $\rho(X,Y)$ based on some observations, and so subject to random variation.

1.3 Interval Estimation

The idea here is not to give a single value for an estimate, but evaluate the uncertainty associated with the population characteristic, providing an indication of the accuracy of the estimation.

Definition

A confidence interval (CI) for a parameter θ is an interval that contains the true value of θ with some (typically high) probability. Let (θ_1, θ_2) be a $100(1-\alpha)\%$ CI for θ. For example, for $\alpha = 0.05$, (θ_1, θ_2) is a 95% CI for θ. If we were to repeat the experiment 100 times, we would expect that $100(1-\alpha)$ of the derived CIs [not necessarily (θ_1, θ_2)] would contain the true value of θ.

Note: A $100(1-\alpha)\%$ CI does NOT contain the true value of θ with probability $(1-\alpha)$. The probability statement relates to the random interval for many different experiments and data sets.

1.3.1 Normal Distribution

Assume independent rvs $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, where σ^2 is known. We wish to estimate μ and obtain a 95% CI for μ. In general, for obtaining a CI theoretically, it is useful to consider some random quantity with a known distribution that is a function of both the data and the parameter of interest. Then, simple probabilistic calculations often provide the CI.

From §1.2 (Theorem 1), we know that $\bar{X}$ is an unbiased estimator of μ. We also know that,

$$\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}.$$

We use the standard result for the Normal distribution (see next chapter) that a linear sum of Normally distributed rvs is also a Normally distributed rv. Then,

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right), \qquad \text{or} \qquad \frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1).$$

If we let $Z \sim N(0,1)$, then we have that

$$\mathbb{P}(z_{1-\alpha/2} < Z < z_{\alpha/2}) = 1-\alpha,$$

where $z_{\alpha/2}$ is such that $\mathbb{P}(Z \geq z_{\alpha/2}) = \frac{\alpha}{2}$. However, the distribution is symmetrical about 0, so that $z_{1-\frac{\alpha}{2}} = -z_{\frac{\alpha}{2}}$) and we can write

$$\mathbb{P}(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1-\alpha.$$

Now,

$$\begin{aligned}
\mathbb{P}\left(-z_{\alpha/2} < \frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} < z_{\alpha/2}\right) &= 1-\alpha \\
\Rightarrow \mathbb{P}\left(\mu - \frac{z_{\alpha/2}\sigma}{\sqrt{n}} < \bar{X} < \mu + \frac{z_{\alpha/2}\sigma}{\sqrt{n}}\right) &= 1-\alpha \\
\Rightarrow \mathbb{P}\left(\bar{X} - \frac{z_{\alpha/2}\sigma}{\sqrt{n}} < \mu < \bar{X} + \frac{z_{\alpha/2}\sigma}{\sqrt{n}}\right) &= 1-\alpha.
\end{aligned}$$

(Note that the random quantity here is $\bar{X}$ and not μ.) Therefore, after we replace the rv $\bar{X}$ with the observed $\bar{x}$, the $(1-\alpha)\%$ CI for μ is given by,

$$\left(\bar{x} - \frac{z_{\alpha/2}\sigma}{\sqrt{n}}, \bar{x} + \frac{z_{\alpha/2}\sigma}{\sqrt{n}}\right);$$

or equivalently,

$$\bar{x} \pm \frac{z_{\alpha/2}\sigma}{\sqrt{n}}.$$

Normal Tables

Published tables are available for probabilities associated with a $N(0,1)$ distribution, however, we will use `R` and the `pnorm` and `qnorm` commands for the c.d.f. and inverse c.d.f. To calculate $z_{\frac{\alpha}{2}}$ in `R` use:

```
> qnorm(1-alpha/2,0,1)
> qnorm(0.975,0,1)          # for 95% interval
[1] 1.959964
> qnorm(0.95,0,1)           # for 90% interval
[1] 1.644845
> qnorm(0.995,0,1)          # for 99% interval
[1] 2.575829
```

Note: Since the $N(0,1)$ distribution is symmetrical around 0, we have that $\mathbb{P}(Z \geq z) = \mathbb{P}(Z \leq -z)$.

Example

In the simple example at the start of this Chapter, it was assumed that $\sigma^2 = 400$ and we observed accommodation expenditures, $\boldsymbol{x} = (x_1, x_2, x_3, x_4, x_5, x_6)$=(400,400,500,500,600,600). Then, $\bar{x} = 500$. For $\alpha = 0.05$, (using `R`), $z_{\alpha/2} = z_{0.025} = 1.96$, and a 95% CI is,

$$\bar{x} \pm \frac{z_{0.025}\sigma}{\sqrt{n}} = (484, 516).$$

So, it is likely the true average accommodation expenditure for students in St Andrews is greater than 484 pounds and less than 516 pounds. (The validity of this statement depends on the validity of the $\sigma^2 = 400$ assumption. See next Chapter for calculating a CI after estimating σ^2 with S^2.)

Example

Suppose that $\sigma^2 = 25$ and we observe $x_1, \ldots, x_{100}$ such that $\sum_{i=1}^{100} x_i = 2430$. Then $\bar{x} = 24.3$ is an estimate of μ. We obtain the 95% CI for μ to be (23.32, 25.28). Similarly we obtain the 90% interval for μ to be (23.48, 25.12) and the 99% interval of (23.01, 25.59).

Alternatively, suppose $\sigma^2 = 25$ and we observed $x_1, \ldots, x_{10}$ such that $\sum_{i=1}^{10} x_i = 243$. Then, again our estimate of μ is $\bar{x} = 24.3$ but the 95% CI for μ is (21.20, 27.40). Similarly we obtain the 90% interval for μ to be (21.70, 26.90) and the 99% interval of (20.23, 28.37).

- the CI is wider for the smaller sample size as we would expect.
- the higher % results in a wider CI, since we have a higher level of confidence that the parameter lies in the interval.
- Consider the 95% CI (21.20,27.40) derived above, and that you are asked to interpret it to somebody with no knowledge of statistics.

  1. The statement 'The true value of μ lies within (21.20,27.40) with probability 95%' is fundamentally wrong, as μ is not a random variable.
  2. The statement 'We are 95% confident that the true value of μ lies within (21.20,27.40)' is not strictly wrong, because the word 'confident' is associated with the formal definition of the CI. However, this statement is misleading because a non-statistician will interpet it exactly as statement (1).
  3. The statement 'The true value of μ is likely to be larger than 21.20 and lower than 27.40' is correct and not misleading.

1.4 Hypothesis Testing

1.4.1 Introduction

Often we wish to test a "hypothesis" about one or more parameters:

Example - Difference in first salary

Let $X \sim N(\mu_x, 100)$ describe describe the first salary for male Maths graduates (in thousands of pounds) and $Y \sim N(\mu_y, 100)$ for female maths graduates.

null hypotheses $\rightarrow H_0$: men and women mathematicians earn the same on average after graduating; $\mu_x - \mu_y = 0$

alternative hypothesis $\rightarrow H_1$: men earn more than women on average; $\mu_x - \mu_y > 0$

1.4.2 Theory

Definitions: Hypotheses

Suppose a data set is drawn from a distribution with a parameter $\theta \in \Theta$. A statistical hypothesis is a statement about the value of θ (a subset of Θ).

A *simple hypothesis* specifies θ as a single point in Θ, i.e. $\theta = \theta_0$.

A *composite hypothesis* specifies a set of more than one value of θ, e.g. $\theta \in [0, 0.5]$.

To formulate a hypothesis test, two hypotheses are needed: the null hypothesis, H_0, and the alternative, H_1 (sometimes H_A is used). The null hypothesis may not be rejected, or it may be rejected in favour of the alternative. Note, H_0 is the hypothesis of no change and is usually (but not always) a simple hypothesis and H_1 is typically a composite hypothesis.

Definitions: Test statistic - Critical region

For data, $\boldsymbol{X}$, we calculate the statistic $T \equiv T(\boldsymbol{X})$, which is a function of $\boldsymbol{X}$ that summarizes the data. A hypothesis test is a rule of the form:

$$\text{Reject } H_0 \text{ in favour of } H_1 \text{ if } T \in C.$$

C is called the *critical region* of the test.

Definitions: Significance Level and Power

Clearly, within hypothesis testing there are two possible errors:

Type I error - reject H_0 when H_0 is true ;

Type II error - fail to reject H_0 when H_0 is false.

Reducing the probability for one of the errors leads to an increased probability for the other. In the classical theory of testing, we choose to keep the probability of the Type I error fixed at a small value α. Formally, let $\mathbb{P}(\text{Type I error}) = \mathbb{P}(T \in C | H_0) = \alpha$. This is called the *significance level* or *size* of the test. It is obvious that the critical region is defined in accordance with the prespecified significance level.

The *power* of the test is the probability of rejecting H_0, when it is false, and is denoted by β.

$$\beta = \mathbb{P}(\text{reject } H_0 | H_0 \text{ is false}).$$

Then,

$$\text{Type II error} = 1 - \beta.$$

When α becomes smaller, the power of the test becomes smaller too. Typically, an investigator is satisfied with $\alpha = 0.05$ and $\beta > 0.7$, but there is nothing special about these exact specifications.

Finally, the *power function* is the power expressed as a function of the parameter θ and is denoted by,

$$\beta(\theta) = \mathbb{P}(\text{reject } H_0 \text{ when the true value is } \theta).$$

Note that since β is a probability,

$$0 \leq \beta(\theta) \leq 1 \qquad \forall \theta \in \Theta.$$

Definitions: p-value

A very important concept is the *p-value*. This is defined as the probability of observing a test statistic T at least as extreme as the observed statistic, t, given that the null hypothesis is true. In other words, the *p-value* is the probability of observing the data we observed (as summarized by the test statistic) or less likely data, given that H_0 is true. This latter interpretation is valid when the distribution of the test statistic T under the H_0 is unimodal.

- If the p-value is "small", it is unlikely that the null hypothesis is true (as the observed data do not appear to be consistent with H_0), so we would reject H_0;
- If the p-value is "large" then the data observed are consistent with H_0, and there is no evidence to reject the null hypothesis.

Because of the manner in which we choose to define the critical region, looking at the tails of the distribution of the test statistic (this is due to optimality arguments with regard to power explained in more advanced modules), a test result is significant (at level α) and we reject H_0 if the p-value $\leq \alpha$. For example, if $\alpha = 0.05$ (often denoted 5%), then we reject H_0 if the p-value is less than 0.05; see also the examples that follow.

1.4.3 Example - Normal Distribution

Let $X_1, \ldots, X_{100}$ be scores on the 1st year Introductory Statistics (MT1007) exam. Assume that $X_i \sim N(\mu, 25)$ independently of each other, so that $\bar{X} \sim N(\mu, 25/100)$. We wish to test whether or not the mean μ is equal to 65. We set up the hypothesis test:

$$H_0 : \mu = 65 \qquad \text{vs} \qquad H_1 : \mu \neq 65.$$

We observe $\bar{x} = 64.0$ and wish to perform a hypothesis test at the 5% significance level. To construct a significance test theoretically, we need to consider a test statistic with a *known* distribution under the H_0. We could choose to use $\bar{X}$ as, under the null it is known that, $\bar{X} \sim N(65, 0.25)$. Another option (historically popular as it allows to use the standard Normal distribution and relevant tables) is to consider the statistic $T = \frac{\bar{X}-65}{\sqrt{0.25}}$, so that if H_0 is true $T \sim N(0,1)$ (see Chapter 2). The observed test statistic is then $(64-65)/0.5 = -2$. This will be our choice for this example.

Calculating the p-value

Note that the distribution of T is symmetrical around 0. The p-value is calculated as the probability of obtaining something as likely as -2, or something less likely.

$$\begin{aligned} \mathbb{P}(T \geq 2 \text{ or } T \leq -2) &= 2\mathbb{P}(T \geq 2) = 2\mathbb{P}(T \leq -2) \\ &= 0.045. \end{aligned}$$

(using `2*pnorm(-2,0,1)`). We reject H_0 at significance level $\alpha = 0.05$ as the p-value is < 0.05.

In practice, statisticians are pragmatic and do not simply "reject" or "not reject" a hypothesis. Instead we typically interpret p-values as follows:

p-value	Interpretation
> 0.1	No evidence against H_0 (not significant)
0.05-0.1	Weak evidence against H_0
0.01-0.05	Moderate evidence against H_0
< 0.01	Strong evidence against H_0

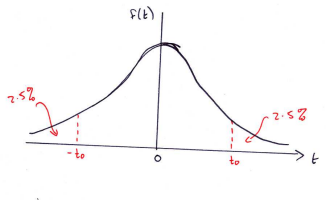
Note: if a p-value is not significant this is not evidence that H_0 is true: "no evidence against H_0" $\neq$ "evidence for H_0".

Calculating the Critical Region

An alternative to calculating the p-value is to calculate the critical region, and see whether the observed test statistic lies within this region.

For size $\alpha = 0.05$, by definition,

$$\begin{aligned} \mathbb{P}(\text{reject } H_0 | H_0 \text{ true}) &= 0.05 \\ \Rightarrow \mathbb{P}(\text{do not reject } H_0 | H_0 \text{ true}) &= 0.95 \\ \Rightarrow \mathbb{P}(-z_{0.05/2} < T < z_{0.05/2} | H_0 \text{ true}) &= 0.95 \end{aligned}$$



If H_0 is true, using `R` we find that $z_{0.025} = 1.96$. For example, we could find this in `R` using the command:

```
> qnorm(0.975, 0, 1)
[1] 1.959964
```

So, the test is to reject H_0 iff,

$$\begin{aligned} T \leq -1.96 \quad \text{or} \quad T &\geq 1.96 \\ \Rightarrow |T| &\geq 1.96 \\ \Rightarrow |(\bar{X} - 65)/0.5| &\geq 1.96, \end{aligned}$$

i.e. if $\bar{X} \leq 64.02$ or $\bar{X} \geq 65.98$. Thus, since we observe $\bar{x} = 64$ we reject the null hypothesis. There is moderate evidence to suggest that the average mark is not 65. Seems that the average mark may be less than 65, as $\bar{x} = 64$. [**Exercise for home:** Calculate a 95% CI for the average mark.]

Power Function: $\beta(\mu)$

The power function $\beta(\mu)$ is the probability that we reject H_0, given some μ. In other words, $\mathbb{P}(|T| \geq 1.96|\mu) = \mathbb{P}(T \leq -1.96|\mu) + \mathbb{P}(T \geq 1.96|\mu)$. These can be easily calculated, (and plotted) using for example `R`. For a true mean μ, we have that $\bar{X} \sim N(\mu, 0.25)$, and $T = \frac{\bar{X}-65}{0.5} \sim N(\frac{\mu-65}{0.5}, 1)$. Thus, to calculate the power function, we would type in `R`,

```
> pnorm(-1.96,(mu-65)/0.5,1) + (1-pnorm(1.96,(mu-65)/0.5,1))
```

for different values of μ.

We obtain:

μ	60	61	62	63	64	65	66	67	68	69	70
$\beta(\mu)$	1	1	0.999	0.979	0.516	0.050	0.516	0.979	0.999	1	1

Increasing sample size

Now, suppose that the number of students taking the course doubles to 200. Let $X_1, \ldots, X_{200}$ be the corresponding scores exam. As before, we assume that $X_i \sim N(\mu, 25)$ independently of each other, so that now $\bar{X} \sim N(\mu, 0.125)$.

What is the critical region at the 5% significance level?

We consider the statistic $T = (\bar{X} - 65)/\sqrt{0.125}$, so that if H_0 is true $T \sim N(0, 1)$

Using `R` we type:

```
> qnorm(0.025, 0, 1)
[1] -1.96
```

So the test is to reject H_0 iff,

$$\begin{aligned} |T| &\geq 1.96 \\ \Rightarrow |\bar{X} - 65|/\sqrt{0.125} &\geq 1.96 \\ \Rightarrow |\bar{X} - 65| &\geq 0.6929646 \end{aligned}$$

i.e. if $\bar{X} \leq 64.30704$ or $\bar{X} \geq 65.69296$.

The power function is calculated as $\mathbb{P}(|T| \geq 1.96|\mu)$, where,

$$T \sim N\left(\frac{\mu - 65}{\sqrt{0.125}}, 1\right).$$

This can be easily checked as before.

So to calculate the power function, in R we could type,

```
> pnorm(-1.96,(mu-65)/sqrt(0.125),1) + (1-pnorm(1.96,(mu-65)/sqrt(0.125),1))
```

for different values of μ. Thus, considering the values $\mu = 60, 61, \ldots, 70$, we obtain the values of the power function:

μ	60	61	62	63	64	65	66	67	68	69	70
$\beta(\mu)$	1	1	1	0.999	0.807	0.050	0.807	0.999	1	1	1

Thus, by increasing the sample size (i.e. increasing the number of student marks considered) we are more likely to reject the null hypothesis, given that the alternative is true, and it is possible to identify smaller deviations from the null hypothesis.

1.4.4 Example - Binomial Distribution (ESP)

We wish to test whether a certain individual possess ESP (Extra Sensory Perception). The subject is asked to name the suit of 20 randomly selected cards (with replacement). Let T be the number of successes. Then, $T \sim Bin(n, p)$. We write the hypotheses in the form:

H_0: $p = 0.25$; H_1: $p \neq 0.25$.

Calculating the p-value

To perform the test, we need to calculate the p-value given by the observed test statistic. Recall the p-value is defined as the probability of observing a result at least as extreme as the observed data, given the null hypothesis is true, i.e. $p = 0.25$. But what is "at least as extreme" in this case? Under the null hypothesis the test statistic $T \sim Bin(20, 0.25)$. This is not a symmetric distribution around the median/mean (0.25*20=5). Consider an observed test statistic of $t <$ median(T). Then the probability of the extreme "left-tail" is $\mathbb{P}(T \leq t)$. We assume that "at least as extreme" corresponds to being in the same upper quantile of the distribution. So that the p-value is $2 \times \mathbb{P}(T \leq t)$. Alternatively, if $t >$ median(T), the probability of the extreme "right-tail" is $\mathbb{P}(T \geq t)$. We assume that "at least as extreme" corresponds to being in the same lower quantile of the distribution. So that the p-value is $2 \times \mathbb{P}(T \geq t)$.

To calculate these probabilities we use R and "`pbinom`" (cdf). To calculate $\mathbb{P}(T \leq t)$ use `pbinom(t,20,0.25)`. Note that

$$\mathbb{P}(T \geq t) = 1 - \mathbb{P}(T < t) = 1 - \mathbb{P}(T \leq t - 1).$$

So, for a given value of t, in R this can be obtained by:

```
> 1 - pbinom(t-1, 20, 0.25)
```

t	0	1	2	3	4	5	6	7
p-value	0.006	0.049	0.18	0.45	0.83	1^*	0.76	0.42
t	8	9	10	11	12	13	14	
p-value	0.20	0.08	0.03	0.008	0.002	0.0004	6×10^{-5}	
t	15	16	17	18	19	20		
p-value	8×10^{-6}	8×10^{-7}	6×10^{-8}	3×10^{-9}	10^{-10}	2×10^{-12}		

* This is given as 1, because $2 \times P(T \leq 5)$ and $2 \times P(T \geq 5)$ are both > 1.

Assuming a significance level $\alpha = 0.05$, we reject the null hypothesis if the p-value is $\leq \alpha = 0.05$. Thus we reject H_0 for $T \leq 1$ or $T \geq 10$ (i.e. the critical region is $T = \{0, 1, 10, 11, \ldots, 20\}$). Observing any of these values would be interpreted as evidence in favour of ESP.

Calculating the Critical Region

Alternatively, we may be interested in calculating the set of values of T for which we would reject H_0 at the 5% significance level, i.e. the critical region C, such that,

$$\mathbb{P}(T \in C) = 0.05.$$

This is most easily done in `R` and the inverse cdf command "`qbinom`". We want 2.5% in each tail so to calculate the right-hand tail:

```
> qbinom(0.975,20,0.25)
[1] 9
```

The `R` command gives the smallest value of t such that $\mathbb{P}(T \leq t) \geq 0.975 \Rightarrow \mathbb{P}(T > t) \leq 0.025$. So, that in other words `R` tells us that $\mathbb{P}(T \geq 10) \leq 0.025$ and $\mathbb{P}(T \geq 9) > 0.025$. Using the `R` command for the cdf "`pbinom` " we can calculate the probabilities exactly:

```
> 1 - pbinom(8,20,0.25)
[1] 0.0409

> 1 - pbinom (9,20,0.25)
[1] 0.0138
```

For the left hand-tail:

```
> qbinom(0.025,20,0.25)
[1] 2
```

This command gives the smallest value of t such that $\mathbb{P}(T \leq t) \geq 0.025$. Check:

```
> pbinom(2,20,0.25)
[1] 0.091
```

whilst,

```
> pbinom(1,20,0.25)
[1] 0.024
```

So, we reject H_0 if $T \leq 1$.

Thus, the critical region is $T = \{0, 1, 10, 11, \ldots, 20\}$).

Note: since T is discrete we cannot choose a critical region such that the size of the test is exactly α in general. We choose the region that has at most $\alpha/2$ in each tail.

Power Function: β

The power function is the probability that we reject H_0. In other words, $\mathbb{P}(T \leq 1 \text{ or } T \geq 10|p)$. These can be easily calculated, (and plotted) using for example `R` and the command:

```
> pbinom(1,20,p) + (1-pbinom(9, 20, p))
```

We obtain:

p	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
$\beta(p)$	0.73	0.39	0.17	0.07	0.04	0.06	0.12	0.24	0.41	0.59
p	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	
$\beta(p)$	0.75	0.87	0.95	0.98	1	1	1	1	1	

Increasing sample size

Now, suppose that we increase the number of test cards to 100. In this case $T \sim Bin(100, p)$. What is the critical region at the 5% significance level (i.e. for which range of values of T would we reject H_0)?

This is most easily done in `R` and the inverse cdf command "`qbinom`". We want (at most) 2.5% in each tail so to calculate the left-hand tail:

```
> qbinom(0.025,100,0.25)
[1] 17
```

Recall that the `R` command gives the smallest value of t such that $\mathbb{P}(T \leq t) \geq 0.025$ and $\mathbb{P}(T \leq t-1) < 0.025$. Thus we would reject all values of $T \leq 16$.

For the right-hand tail:

```
> qbinom(0.975,100,0.25)
[1] 34
```

So that $\mathbb{P}(T \leq 34) \geq 0.975$ and $\mathbb{P}(T \leq 33) < 0.975$. In other words $\mathbb{P}(T \geq 35) \leq 0.025$ and $\mathbb{P}(T \geq 34) > 0.025$. Thus we reject all values of $T \geq 35$.

Using the `R` command for the cdf "`pbinom` " we can calculate the probabilities exactly:

```
> pbinom(16,100,0.25)
[1] 0.021

> 1 - pbinom (34,100,0.25)
[1] 0.016
```

Check:

```
> pbinom(17,100,0.25)
[1] 0.038

> 1 - pbinom (33,100,0.25)
[1] 0.028
```

(which are both > 0.025).

Thus the critical region is $T \leq 16$ and $T \geq 35$.

Decreasing sample size

Now, suppose that we decrease the number of test cards to 10. In this case $T \sim Bin(10, p)$. What is the critical region at the 5% significance level (i.e. for which range of values of T would we reject H_0)?

We want 2.5% in each tail so to calculate the right-hand tail:

```
> qbinom(0.975,10,0.25)
[1] 5
```

The `R` command gives the smallest value of t such that $\mathbb{P}(T \leq t) \geq 0.975 \Rightarrow \mathbb{P}(T > t) \leq 0.025$. So, that in other words `R` tells us that $\mathbb{P}(T \geq 6) \leq 0.025$ and $\mathbb{P}(T \geq 5) > 0.025$. Using the `R` command for the cdf "`pbinom` " we can calculate the probabilities exactly:

```
> 1 - pbinom(4,10,0.25)
[1] 0.078

> 1 - pbinom (5,10,0.25)
[1] 0.020
```

For the left hand-tail:

```
> qbinom(0.025,10,0.25)
[1] 0
```

Check:

```
> pbinom(0,10,0.25)
[1] 0.056
```

Even if we observe $T = 0$, we do not reject H_0.

Thus, the critical region is $\{T : T \geq 6\}$.

1.4.5 One-sided and Two-sided Tests

Suppose θ is the parameter of interest. A two-sided (or two-tailed) test is of the form:

$$H_0 : \theta = \theta_0 \qquad \text{vs} \qquad H_1 : \theta \neq \theta_0.$$

In some cases, it may only be possible for the parameter to take values on one-side of θ_0 or there may only be evidence against H_0 in one tail of the distribution (for example some goodness-of-fit tests - see for example Honours module MT3606). In these cases, one-sided tests are appropriate and are specified as:

$$H_0 : \theta = \theta_0 \qquad \text{vs} \qquad H_1 : \theta > \theta_0,$$

or

$$H_0 : \theta = \theta_0 \qquad \text{vs} \qquad H_1 : \theta < \theta_0.$$

To evaluate the evidence against H_0 in favour of the alternative hypothesis, only consider one tail of the distribution of the test statistic, depending on the direction of the inequality in H_1 (the left tail for $H_1 : \theta < \theta_0$ and the right tail for $H_1 : \theta > \theta_0$). The nature of the problem typically dictates whether a one-sided or two-sided test is most appropriate.

Example

Reconsider Example 1.4.3, with the sample size of 100. Suppose instead that the exam was set such that it contained additional questions on a difficult subject, and we wished to test whether this increased the difficulty of the exam (we assume that it could not have made the exam any easier). Then, we may wish to test,

$$H_0 : \mu = 65 \qquad \text{vs} \qquad H_1 : \mu < 65.$$

Once more we consider the test statistic $T = (\bar{X} - 65)/0.5$, so that under H_0, $T \sim N(0, 1)$. The form of the critical region will be to reject H_0 if,

$$T \leq -z_\alpha.$$

(I.e. we only reject the null hypothesis in favour of the alternative for the test statistic in the "correct" tail). So,

$$\mathbb{P}(T \leq -z_{0.05} | H_0 \text{ true}) = 0.05.$$

Using the following command in R:

```
> qnorm(0.05, 0, 1)
[1] -1.6449
```

Thus, we have that,

$$\mathbb{P}(T \leq -1.645) = 0.05.$$

Thus we reject H_0 (at the 5% significance level) if $T \leq -1.645$, or $\bar{X} \leq 64.18$.

Suppose that $\bar{x}$ is 64, then the observed test statistic is -2. We would reject the H_0 at 5%. The p-value is calculated by $\mathbb{P}(T \leq -2 | H_0 \text{ true})$. Using R, we could calculate this using:

```
> pnorm(-2,0,1)
[1] 0.02275013
```

Thus we would reject H_0 at the 5% level, but not the 1% level.

1.4.6 Confidence Intervals and Hypothesis Tests

Hypothesis testing and confidence intervals are related! In particular, consider the null hypothesis $H_0 : \mu = \mu_0$ against the alternative hypothesis $H_1 : \mu \neq \mu_0$, for the mean of a Normal distribution. The set of values of $\bar{X}$ for which H_0 is not rejected at significance level α, given the observed data, is equivalent to the $100(1-\alpha)\%$ confidence interval for μ. Alternatively, the set of values of $\bar{X}$ for which the null hypothesis is rejected is equivalent to the complement of the equivalent $100(1-\alpha)\%$ confidence interval for μ. However, confidence intervals provide more information than the equivalent hypothesis tests. For example, suppose that the calculated 95% CI for μ is $(\mu_0 - 10^{-100}, \mu_0 + 10^{100})$. Then we know that if we had collect more data, the Null hypothesis may have been rejected. Also, we know that probably we would reject the Null hypothesis at significance level $\alpha = 0.1$, as the corresponding CI would become less wide.