Chapter 3: A Review of Statistics

Instructor: Sergi Quintana Garcia
Notes obtained from Manuel V. Montesinos (UAB and IDEA)

Statistics Brush-Up Course

Fall 2022

Random Sampling

Random Sampling

The sequence of random variables $\{X_n\}_{n=1}^{\infty}$ is **independently** and **identically distributed (i.i.d.)** if they are independent and have the same distribution.

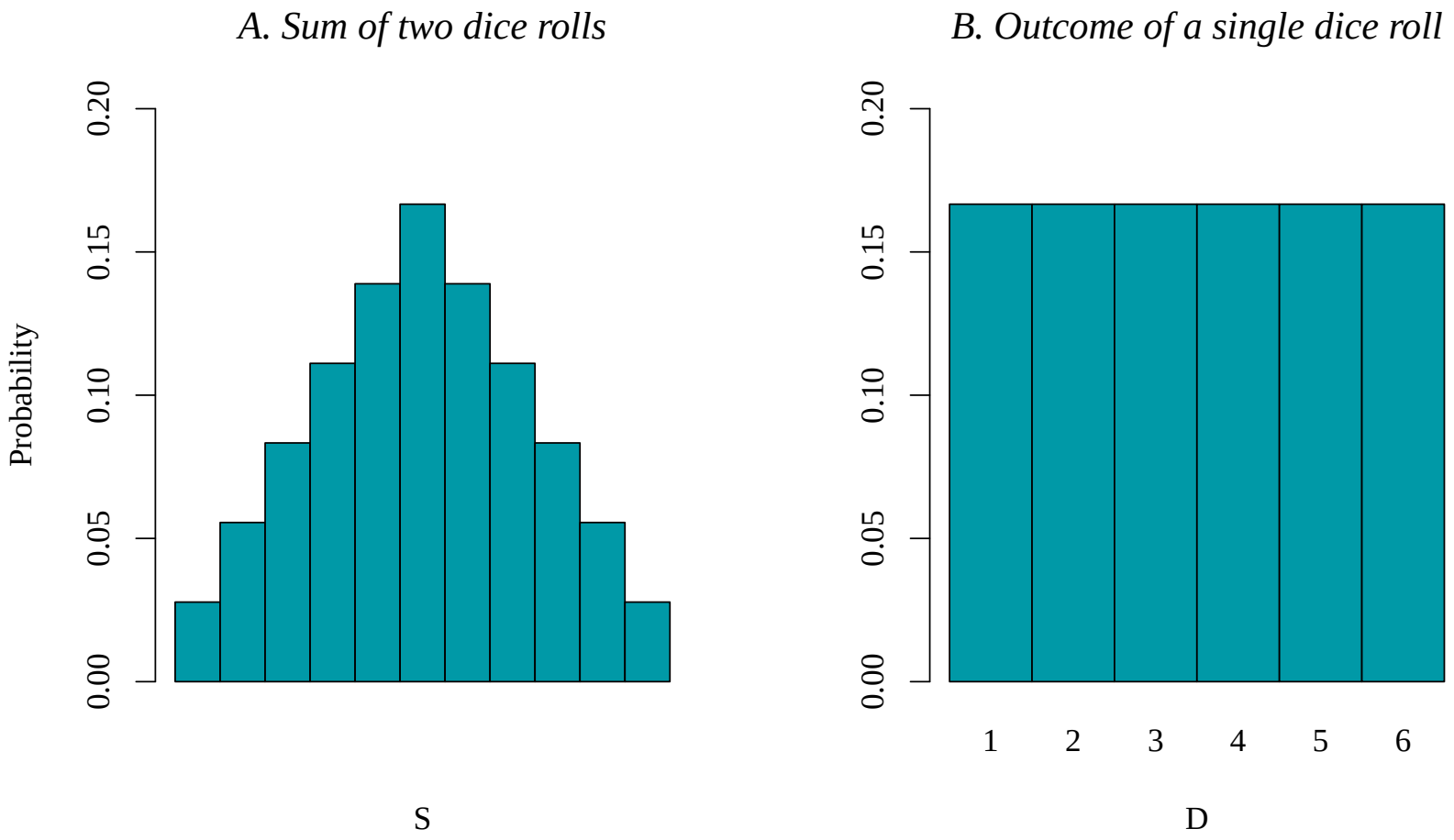
Let $Y = (X_1, \ldots, X_n)$ be a random vector. Then, a **sample** of the random vector is a realization (value) $y = (x_1, \ldots, x_n)$ of the random vector.

A **random sample** of size n is a realization of a collection of random variables $\{X_i\}_{i=1}^{n}$ that are i.i.d. Its common distribution is called the **population distribution**. We say that $\{x_i\}_{i=1}^{n}$ is a random sample of X with sample size n.

A random vector $u = h(X_1, \ldots, X_n)$ is called a **statistic** of the sample $\{X_i\}_{i=1}^{n}$. Statistics are random variables and their distribution is called the **sampling distribution**. An example of a statistic is the sample mean.

Random Sampling

Figure 1: Probability Distributions of the Sum of Two Dice Rolls and a Single Dice Roll



Distribution of the Sample Mean

Distribution of the Sample Mean

The **sample mean** of a random sample $\{x_i\}_{i=1}^n$ is given by

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}. \tag{1}$$

If $\{x_i\}_{i=1}^n$ is a random sample from a population with mean μ and variance σ^2 ($\sigma^2 < \infty$), then

$$\mathbb{E}(\bar{x}) = \mu,$$
$$\mathrm{Var}(\bar{x}) = \frac{\sigma^2}{n}.$$

Distribution of the Sample Mean

The fact that a sample is random implies that the collection $\{X_i\}_{i=1}^n$ is i.i.d., and thus the expected value of each of the X_i is $\mathbb{E}(X_i) = \mu$. Then:

$$\begin{aligned}
\mathbb{E}\left[\bar{x}\right] &= \mathbb{E}\left[\frac{X_1 + X_2 + \ldots + X_n}{n}\right] \\
&= \frac{1}{n}\mathbb{E}\left[X_1 + X_2 + \ldots + X_n\right] \\
&= \frac{1}{n}\left[\mathbb{E}(X_1) + \mathbb{E}(X_2) + \ldots + \mathbb{E}(X_n)\right] \\
&= \frac{1}{n}\left[\mu + \mu + \ldots + \mu\right] \\
&= \frac{1}{n}\left[n\mu\right] = \mu.
\end{aligned}$$

Distribution of the Sample Mean

To obtain the variance of the sample mean, we use the fact that if two random variables X and Y are independent, then $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$ and the identical distribution implies that $\mathrm{Var}(X_i) = \sigma^2$ for all i:

$$\begin{aligned}
\mathrm{Var}(\bar{x}) &= V\left[\frac{X_1 + X_2 + \ldots + X_n}{n}\right] \\
&= \frac{1}{n^2}\mathrm{Var}\left[X_1 + X_2 + \ldots + X_n\right] \\
&= \frac{1}{n^2}\left[\mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \ldots + \mathrm{Var}(X_n)\right] \\
&= \frac{1}{n^2}\left[\sigma^2 + \sigma^2 + \ldots + \sigma^2\right] \\
&= \frac{1}{n^2}\left[n\sigma^2\right] = \frac{\sigma^2}{n}.
\end{aligned}$$

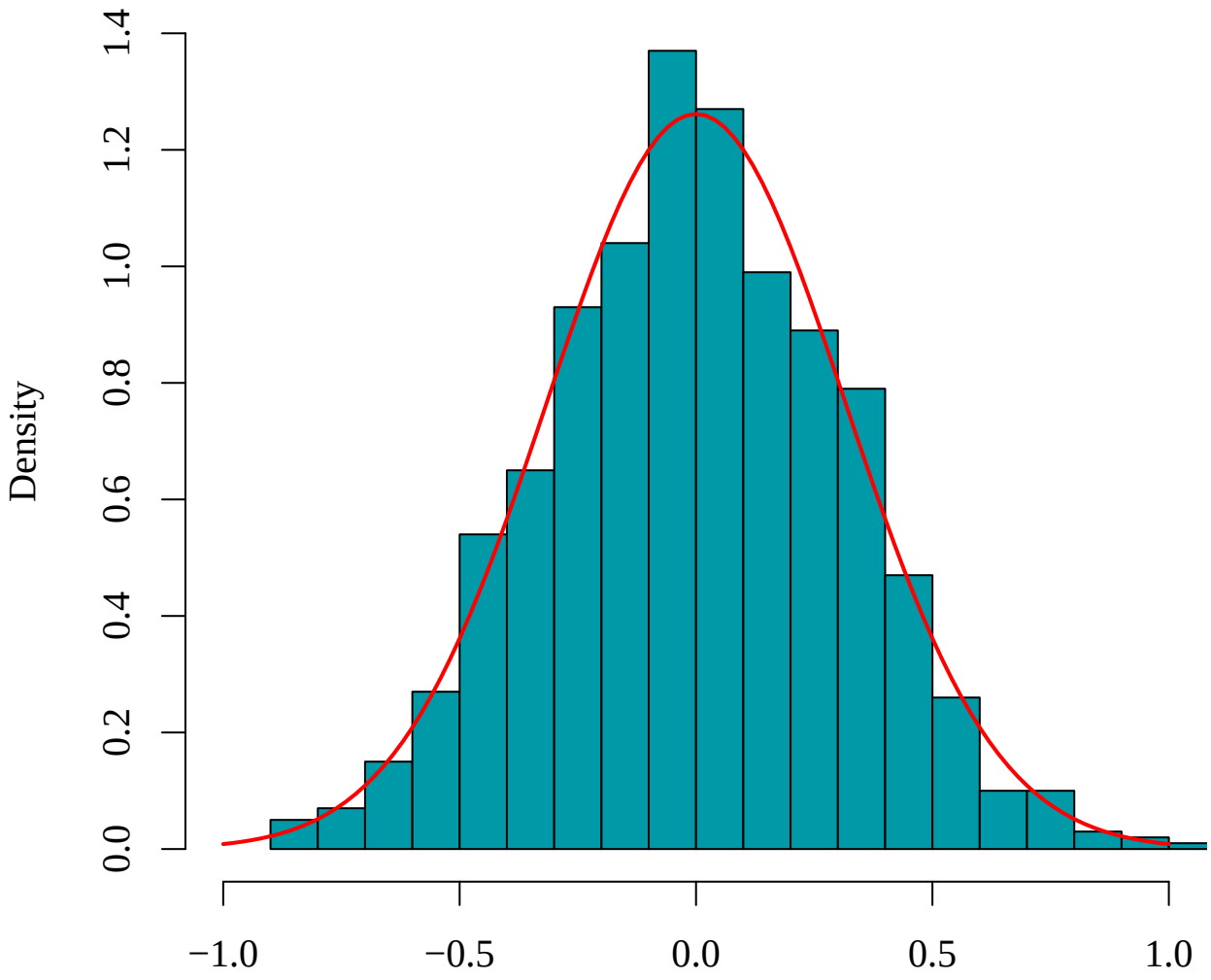
Distribution of the Sample Mean

The previous results hold for any distribution of X_i.

For instance, if $\{x_i\}_{i=1}^n$ is a random sample of size n, and X_i is normally distributed with mean μ and variance σ^2 for all i, then the sampling distribution of $\bar{x}$ is $\mathcal{N}(\mu, \sigma^2/n)$.

Distribution of the Sample Mean

Figure 2: Histogram of $\overline{X}$ when $X \sim \mathcal{N}(0, 1)$



Distribution of the Sample Variance

Distribution of the Sample Variance

The **sample variance** is defined as

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}.$$

If $x_1, \ldots, x_n$ is a sample of $\{X_i\}_{i=1}^{n}$, and X_i is normally distributed with mean μ and variance σ^2 for all i, then

$$\frac{(n-1)s^2}{\sigma^2} \equiv \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Student's-t and F Distributions

Student's-t and F Distributions

If $\bar{x}$ and s^2 are the mean and the variance of a random sample of $\{X_i\}_{i=1}^n$, and X_i is normally distributed with mean μ and variance σ^2 for all i, then

$$\frac{\bar{x} - \mu}{s/\sqrt{n}} \sim t_{n-1}.$$

If s_1^2 and s_2^2 are the variances of two independent random samples of $\{X_{1i}\}_{i=1}^{n_1}$ and $\{X_{2i}\}_{i=1}^{n_2}$, and X_{1i} and X_{2i} are normally distributed with variances σ_1^2 and σ_1^2 for all i, then

$$y = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \sim F_{n_1-1,n_2-1}.$$

Large Sample Approximations

Large Sample Approximations

Convergence in probability and convergence in distribution

The sequence of random variables $\{X_n\}_{n=1}^{\infty}$ **converges in probability** to the random variable X if

$$\lim_{n\to\infty} \Pr(|X_n - X| \geq \epsilon) = 0$$

for all $\epsilon > 0$, or

$$\lim_{n\to\infty} \Pr(|X_n - X| < \epsilon) = 1.$$

The notation we use is $X_n \xrightarrow{p} X$ or $\operatorname*{plim}_{n\to\infty} X_n = X$.

Large Sample Approximations

Convergence in probability and convergence in distribution

The sequence of random variables $\{X_n\}_{n=1}^{\infty}$ **converges in distribution** to the random variable X if the distribution function F_n of X_n converges pointwise to the distribution F of X at every point of F. We write that $X_n \xrightarrow{d} X$.

Convergence in probability implies convergence in distribution.

Large Sample Approximations

Law of large numbers

Let $X_1, \ldots, X_n$ be iid random variables with mean μ and standard deviation σ, and let $\bar{x} = \frac{X_1 + \ldots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$. The **law of large numbers** states that

$$\bar{x} - \mathbb{E}(\bar{x}) \xrightarrow{p} 0,$$

and for any small number ϵ,

$$\lim_{n \to \infty} \Pr(|\bar{x}_n - \mu| < \epsilon) = 1.$$

This means that, by choosing an n large enough, we can make $\bar{x}_n$ as close to μ as we want with probability close to one.

Example

Repeated coin tossing: let Y_i be the result of tossing a coin for the ith time. Y_i is a Bernoulli distributed random variable with p being the probability of observing head:

$$\Pr(Y_i) = \begin{cases} p, \ Y_i = 1 \\ 1 - p, \ Y_i = 0, \end{cases}$$

with $p = 0.5$, as we assume a fair coin. We know that $\mu_Y = p = 0.5$. Let R_n denote the proportion of heads in the first n tosses,

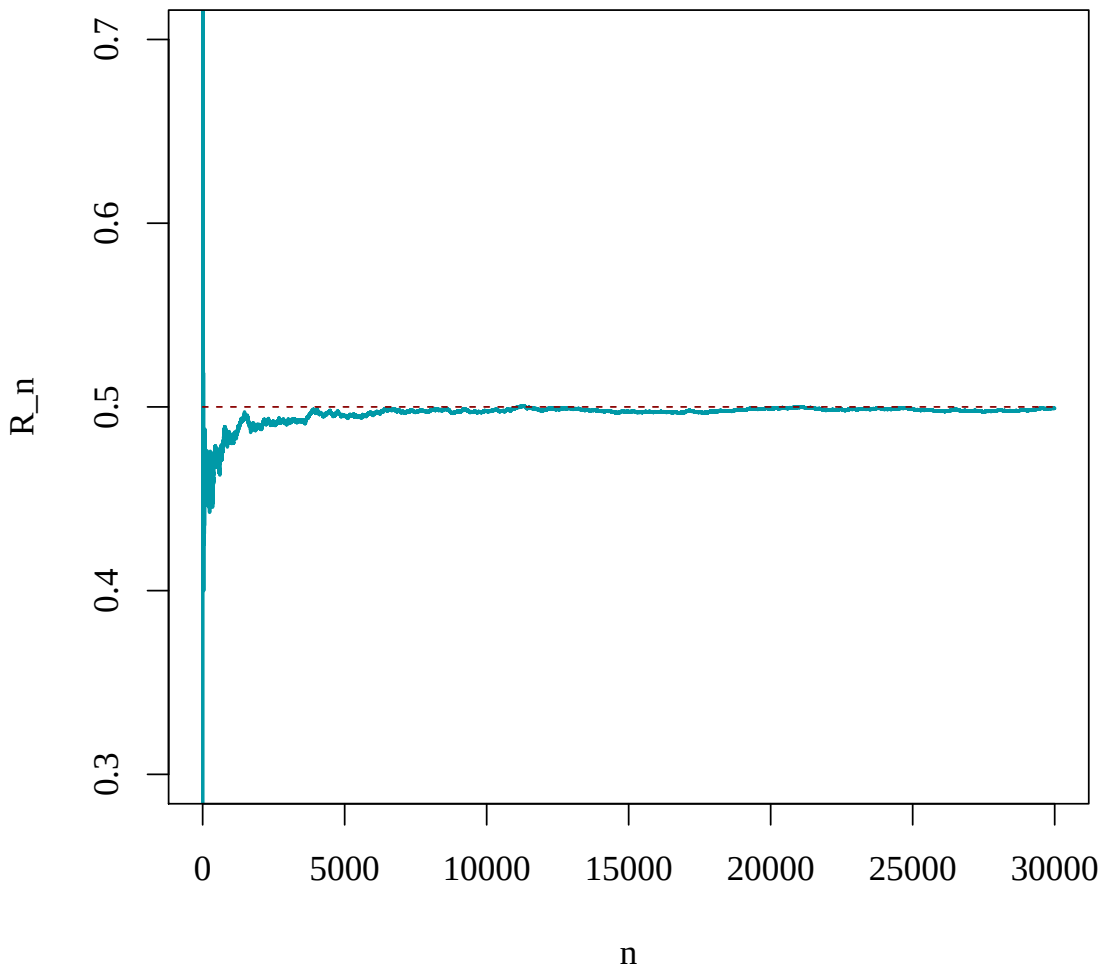
$$R_n = \frac{1}{n} \sum_{i=1}^{n} Y_i.$$

By the law of large numbers:

$$R_n \xrightarrow{p} \mu_Y = 0.5 \text{ as } n \to \infty. \tag{2}$$

Example

Figure 3: Converging Share of Heads in Repeated Coin Tossing



Large Sample Approximations

Central limit theorem

The **central limit theorem** states that, under general conditions, the distribution of $\overline{X}$ is well approximated by a normal distribution when the sample size is large.

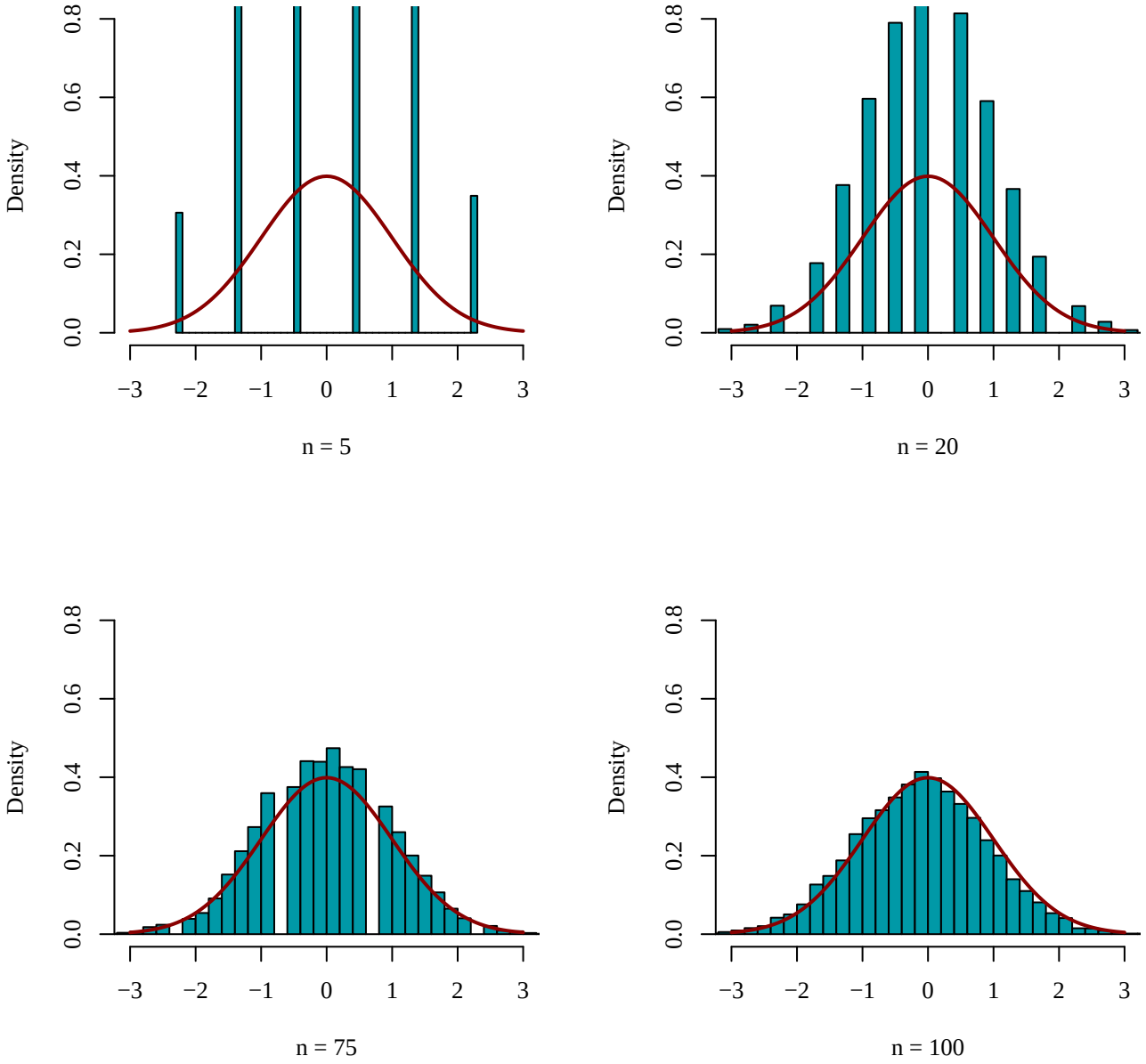
Let $X_1, \ldots, X_n$ be iid random variables with $\mathbb{E}(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$. Then,

$$Z = \frac{\bar{x} - \mathbb{E}(\bar{x})}{\sqrt{\mathrm{Var}(\bar{x})}} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, 1).$$

This means that for large n,

$$\bar{x}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

Figure 4: Converging Standardized Sample Mean of the Bernoulli Distribution



Exercise 1

Suppose that X_j for $j = 1, \ldots, n$ are iid random variables with $\mathbb{E}(X_j) = \mu$ and $\mathrm{Var}(X_j) = \sigma^2$. Use the central limit theorem to give an approximation to $\Pr(a < S_n \leq b)$ for $-\infty < a < b < \infty$, where $S_n = \sum_{j=1}^{n} X_j$.

Estimators and their Properties

Estimators and their Properties

Let's assume that the population can be modeled as a random variable X with a distribution function F which is known except for a vector of parameters θ.

A **point estimator** is a statistic $\hat{\theta} = \hat{\theta}(X_1, X_2, ..., X_n)$ that approximates the value of the unknown vector θ. If we observe a particular random sample $X_1 = x_1, X_2 = x_2, ..., X_n = x_n$, then $\hat{\theta} = \hat{\theta}(x_1, x_2, ..., x_n)$ is called a **point estimate**.

Interval estimation proposes two statistics, $T_1(X_1, X_2, ..., X_n)$ and $T_2(X_1, X_2, ..., X_n)$, such that the probability that the random interval

$$[T_1(X_1, X_2, \ldots, X_n), T_2(X_1, X_2, \ldots, X_n)]$$

contains the unknown vector of parameters can be calculated.

Estimators and their Properties

Unbiasedness

Let X be a random variable with CDF $F(\cdot, \theta)$, which depends on the unknown parameters θ (or vector of unknown parameters) and let $X_1, X_2, ..., X_n$ be a random sample and $\hat{\theta} = \hat{\theta}(X_1, X_2, ..., X_n)$ an estimator of θ.

$\hat{\theta}$ is **unbiased** if

$$\mathbb{E}(\hat{\theta}) = \theta,$$

i.e., an estimator is unbiased if its mean coincides with the unknown parameter.

Example

The sample mean, which is an estimator of the unknown population mean μ, is unbiased, since $\mathbb{E}(\bar{x}) = \mu$.

Another example of an unbiased estimator is the sample variance:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Let's prove that s^2 is an unbiased estimator of the population variance σ^2, so $\mathbb{E}(s^2) = \sigma^2$. For this purpose, remind that $\mathrm{Var}(x_i) = \mathbb{E}(x_i^2) - \mu^2 = \sigma^2$ and $\mathrm{Var}(\bar{x}) = \mathbb{E}(\bar{x}^2) - \mu^2 = \frac{\sigma^2}{n}$.

Example

$$
\begin{aligned}
\mathbb{E}(s^2) &= \mathbb{E}\left(\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}\right) = \frac{1}{n-1}\mathbb{E}\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right) \\
&= \frac{1}{n-1}\frac{n}{n}\mathbb{E}\left(\sum_{i=1}^{n}\left(x_i^2-2x_i\bar{x}+\bar{x}^2\right)\right) \\
&= \frac{n}{n-1}\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}x_i^2-2\frac{1}{n}\sum_{i=1}^{n}x_i\bar{x}+\bar{x}^2\right) \\
&= \frac{n}{n-1}\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}x_i^2-2\bar{x}^2+\bar{x}^2\right) \\
&= \frac{n}{n-1}\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}x_i^2-\bar{x}^2\right)
\end{aligned}
$$

Example

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

Estimators and their Properties

Consistency

Let $X_1, X_2, ..., X_n$ be a sequence of iid random variables with common CDF $F(\cdot, \theta)$ that depends on an unknown parameter θ and let $\hat{\theta}_n = \hat{\theta}_n(X_1, X_2, ..., X_n)$ be an estimator of θ based on a random sample of size n.

We can consider the sequence of random variables $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_n$ and study some limiting properties of this sequence as the sample size increases. These limiting properties are called **asymptotic** or **large sample properties** of the estimator $\hat{\theta}_n$.

Estimators and their Properties

Consistency

The estimator $\hat{\theta}_n$ is said to be a **consistent** estimator of θ if

$$\hat{\theta}_n \xrightarrow{p} \theta,$$

or

$$\lim_{n\to\infty} \Pr(|\hat{\theta}_n - \theta| \geq \epsilon) = 0.$$

An unbiased estimator is consistent, but the converse is not true.

Example

Let $X_1, X_2, ..., X_n$ be a sequence of iid random variables with finite mean μ and variance σ^2. By the law of large numbers, we know that $\bar{x} \xrightarrow{p} \mathbb{E}(\bar{x})$, and $\mathbb{E}(\bar{x}) = \mu$. Thus, $\bar{x}$ is a consistent estimator of the population mean.

Example

Let $x_1, x_2, ..., x_n$ be a random sample of $\{X\}_{i=1}^n$. Let's consider two alternative estimators of σ^2:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

and

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

We have shown that s^2 is an unbiased estimator of σ^2. Meanwhile, $\hat{\sigma}^2$ is biased:

$$\hat{\sigma}^2 = \frac{n-1}{n}s^2 \Rightarrow \mathbb{E}(\hat{\sigma}^2) = \mathbb{E}\left(\frac{n-1}{n}s^2\right) = \frac{n-1}{n}\sigma^2,$$

so the bias of $\hat{\sigma}^2$ is

$$\text{bias}(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{\sigma^2}{n}.$$

Example

However, both estimators are consistent. Let's prove consistency of s^2:

$$
\begin{aligned}
s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{n}{n(n-1)}\sum_{i=1}^{n}\left(x_i^2+\bar{x}^2-2x_i\bar{x}\right) \\
&= \frac{n}{n(n-1)}\left(\sum_{i=1}^{n}x_i^2+n\bar{x}^2-2\sum_{i=1}^{n}x_i\bar{x}\right) \\
&= \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n}x_i^2-\bar{x}^2\right).
\end{aligned}
$$

By the law of large numbers, we have

$$
\frac{1}{n}\sum_{i=1}^{n}x_i^2 \xrightarrow{p} \mathbb{E}(X_i^2),
$$

$$
\frac{1}{n}\sum_{i=1}^{n}x_i \xrightarrow{p} \mathbb{E}(X_i) = \mu.
$$

Example

Therefore,

$$\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2 = \bar{x}^2 \xrightarrow{p} \mathbb{E}(X_i)^2 = \mu^2,$$

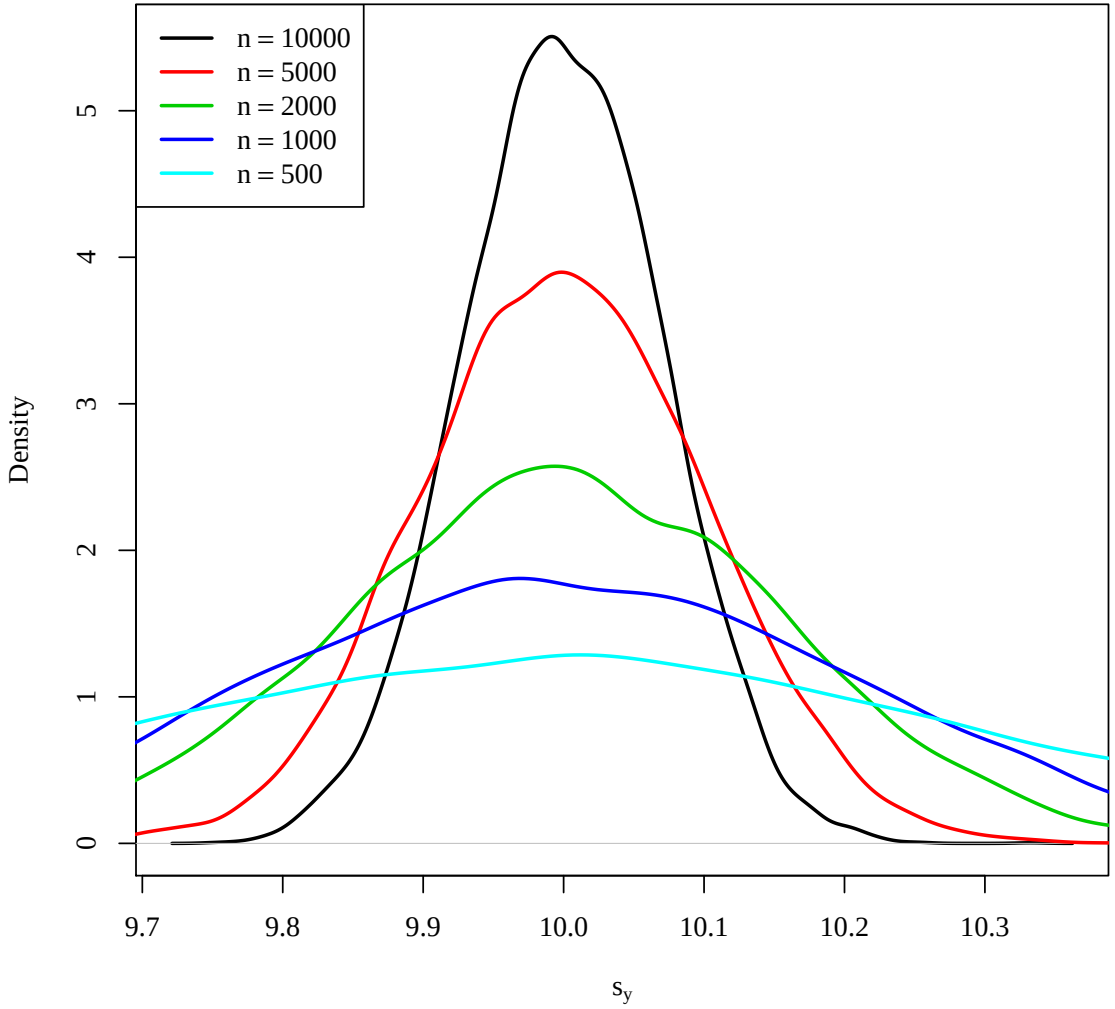
and

$$\lim_{n\to\infty} \frac{n}{n-1} = 1,$$

so $s^2 \xrightarrow{p} \mathbb{E}(X^2) - \mu^2 = \mathrm{Var}(x) = \sigma^2$. In a similar way, we can prove that $\hat{\sigma}^2$ is consistent.

Example

Figure 5: Sampling Distribution of s_X for $X \sim \mathcal{N}(10, 10)$



Estimators and their Properties

Efficiency

The **Mean Squared Error (MSE)** of an estimator $\hat{\theta} = \hat{\theta}(X_1, X_2, ..., X_n)$ is defined as

$$\mathrm{MSE}(\hat{\theta}) = \mathbb{E}\left[(\hat{\theta} - \theta)^2\right] = \mathrm{Var}(\hat{\theta}) + \left(\mathrm{bias}(\hat{\theta})\right)^2,$$

where $\mathrm{bias}(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta$.

Estimators and their Properties

Efficiency

Let X be a random variable with cdf $F(\cdot, \theta)$, $x_1, x_2, \ldots, x_n$ be a random sample, and $\hat{\theta}_1 = \hat{\theta}_1(x_1, x_2, \ldots, x_n)$ and $\hat{\theta}_2 = \hat{\theta}_2(x_1, x_2, \ldots, x_n)$ be two estimators of θ.

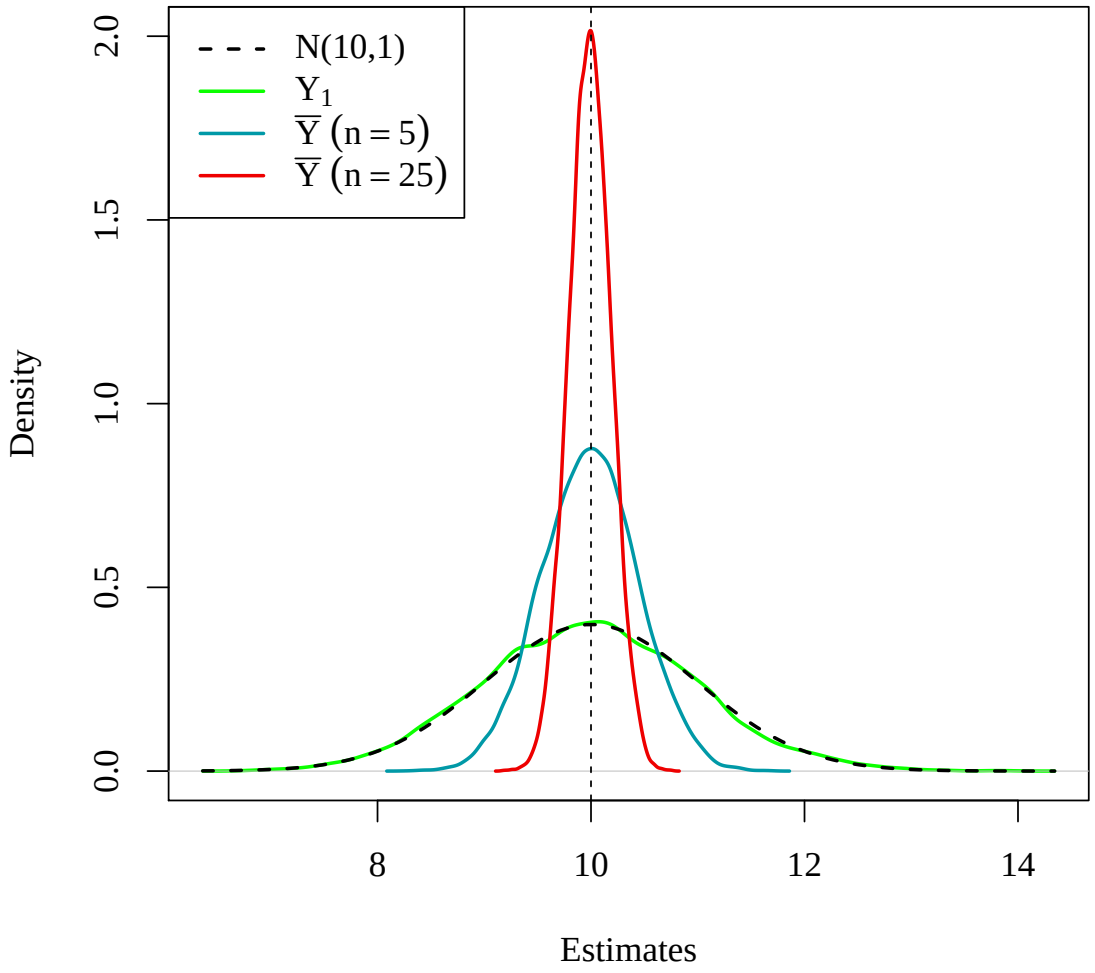
We say that $\hat{\theta}_1$ is more **efficient** than $\hat{\theta}_2$ if

$$\text{MSE}(\hat{\theta}_1) < MSE(\hat{\theta}_2).$$

If we are comparing two unbiased estimators, their MSE is equal to their variance, as the bias is equal to zero for both. Thus, the estimator with smaller variance is more efficient.

Estimators and their Properties

Figure 6: Sampling Distributions of Unbiased Estimators of the Population Mean for $\mathcal{N}(10, 1)$



Hypothesis Testing

Hypothesis Testing

The methodology used to analyze the validity of hypotheses about the characteristics of a population is called **hypothesis testing**.

A statistical **hypothesis** is a statement about the distribution of one or more random variables.

The aim of a **hypothesis test** is to decide, based on a sample from the population, which of two complementary hypotheses is true.

The two complementary hypotheses are called the **null hypothesis**, denoted by H_0, and the **alternative hypothesis**, denoted by H_1.

Example

We have been using $\bar{x}$ as an estimator of μ. However, given that we have a sample of size n, $\bar{x}$ will not be exactly equal to μ.

Suppose we want to test whether the mean wage in the US is 50 (thousand dollars per year), i.e. whether $\mu = 50$. Suppose we have a random sample and $\bar{x} = 52$. Should we reject our hypothesis?

The null hypothesis H_0 is the hypothesis to be tested:

$$H_0 : \mu = 50.$$

The alternative hypothesis H_1 is what we contrast our hypothesis with:

$$H_1 : \mu \neq 50$$

In this case, we are doing a **two-sided** test.

Hypothesis Testing

When we decide on a test, there are four possible scenarios:

	H_0 true	H_1 true
decision: H_0	Correct	Type 2 error
decision: H_1	Type 1 error	Correct

The probability of each event is:

	H_0 true	H_1 true
decision: H_0	1-α	β
decision: H_1	α	1-β

Hypothesis Testing

A **type 1 error** occurs when the null hypothesis is true but is rejected, and a **type 2 error** occurs when the alternative hypothesis is true but the null is not rejected.

The **significance level** α is the probability of a type 1 error, while β is the probability of a type 2 error.

We call $1 - \beta$ the **power** of the test. It is the probability of rejecting the null and being right about it.

Ideally, the significance level should be close to 0 and the power close to 1.

Hypothesis Testing

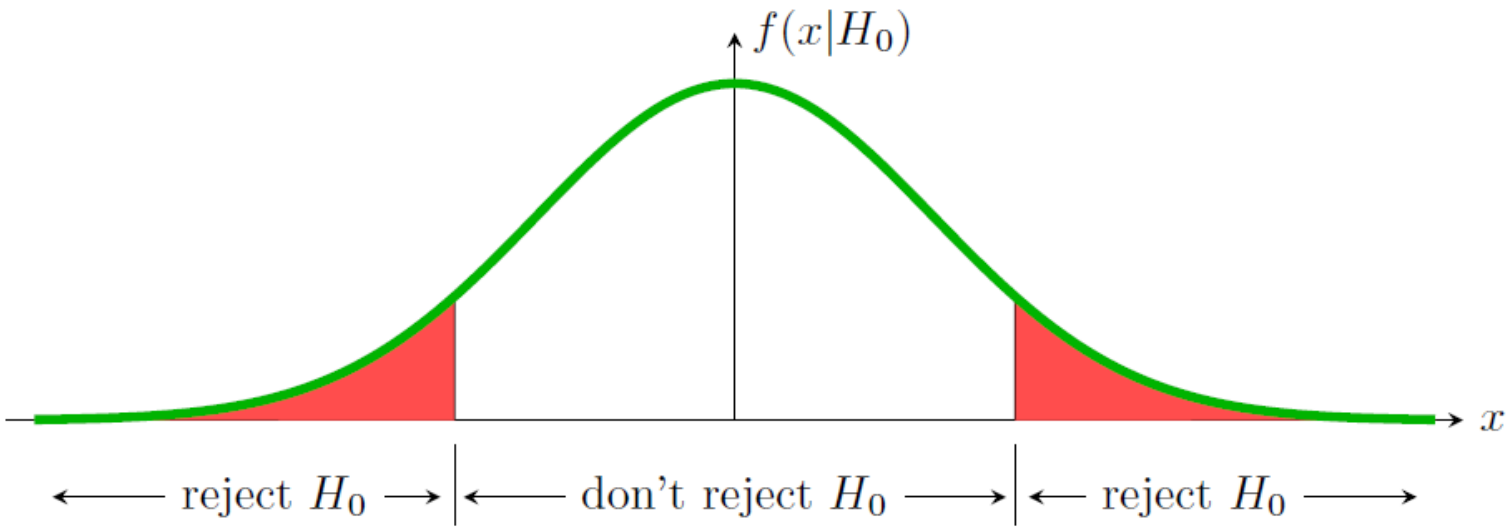
The **steps** of hypothesis testing are:

1. Choose H_0 and H_1.
2. Choose a significance level α.
3. Choose a test statistic and determine the null distribution.
4. Determine how to compute a p-value and/or the rejection region. Find the critical value.
5. Using your data, compute the test statistic.
6. Reject or fail to reject the null hypothesis.

Typically, a hypothesis test is based on a **test statistic** $T = T(X_1, X_2, ..., X_n)$ that is a function of the sample and whose behavior is different under the null than under the alternative.

Hypothesis Testing

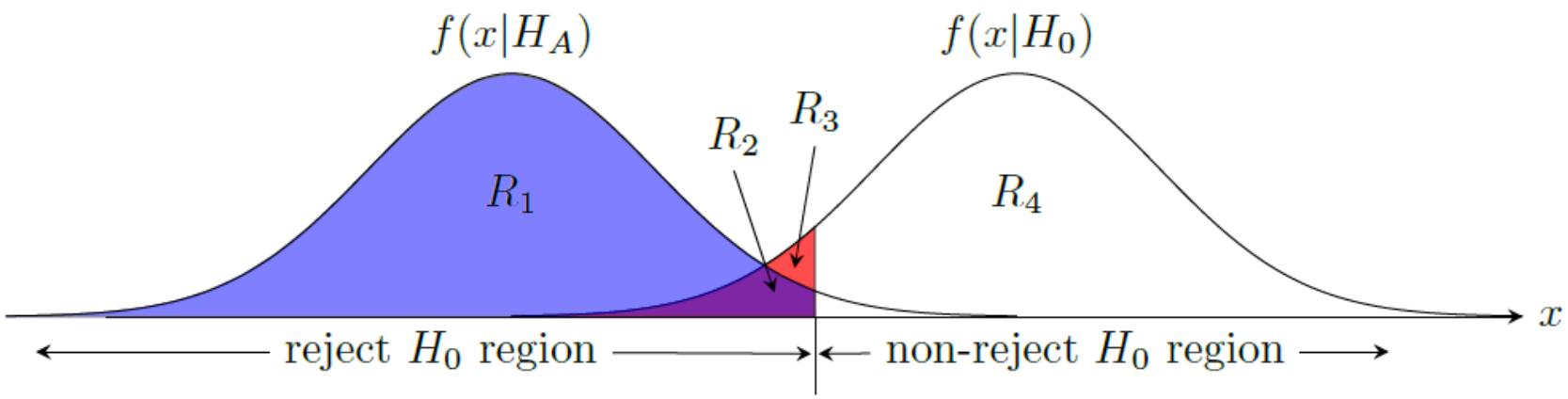
In this graph:



- x is a test statistic.
- $f(x|H_0)$ is the pdf of the null distribution (green curve).
- The rejection region is a proportion of the x-axis.
- The level of significance is the probability contained in the rejection region (red area).

Hypothesis Testing

In this graph:



- Significance: $\Pr(\text{rejection region}|H_0) = R_2 + R_3$.
- Power: $\Pr(\text{rejection region}|H_A) = R_1 + R_2$.

Hypothesis Testing

The **critical values** are the boundaries of the rejection region. Critical values are labeled by the probability left to their right. An example for the standard normal distribution is

$$c_{0.025} = 1.96$$
$$c_{0.975} = -1.96.$$

Some critical values for the standard normal distribution:

Level of significance α	**0.10**	**0.05**	**0.01**
One-tailed test	-1.28 or 1.28	-1.645 or 1.645	-2.33 or 2.33
Two-tailed test	-1.645 and 1.645	-1.96 and 1.96	-2.58 and 2.58

Hypothesis testing

Tests concerning means when the variance is known

The height of adult men, denoted Y, is normally distributed. We want to test whether $\mu = 70$ inches. Assume that we have a random sample, and the sample mean is 74. For now, suppose that we know the variance of Y and it is 900. The sample size is 100.

Let's follow the steps for testing this hypothesis:

1. H_0: $\mu = 70$, H_1: $\mu \neq 70$.
2. $\alpha = 5\%$.
3. Test statistic/decision variable: we must be able to find its distribution. Let's take the standardized mean.

Hypothesis testing

Tests concerning means when the variance is known

4. The critical value is given by $z_{\alpha/2}$ and $z_{1-\alpha/2}$:

$$z_{\alpha/2} = z_{2.5\%} = 1.96 \text{ and } z_{1-\alpha/2} = z_{97.5\%} = -1.96$$

5. The actual value of the decision variable/test statistic is

$$Z^{act} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{74 - 70}{30/\sqrt{100}} = \frac{4}{3} = 1.33$$

6. Decision: we cannot reject the null, as 1.33 is smaller than the critical value of 1.96, i.e., the test statistic does not lie in the rejection region.

Hypothesis testing

Tests concerning means when the variance is unknown

Assume now that we do not know the variance. We can use the sample variance s^2 to estimate it:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

We already know that this is an unbiased and consistent estimator of σ^2.

The **standard error** of $\bar{x}$ is given by

$$SE(\bar{x}) = \hat{\sigma}_{\bar{x}} = \frac{s}{\sqrt{n}}.$$

The **t-statistic** or **t-ratio** has a Student's-t distribution with $n-1$ degrees of freedom

$$t = \frac{\bar{x} - \mu_0}{SE(\bar{x})} \sim t_{n-1}.$$

Hypothesis testing

Tests concerning means when the variance is unknown

The height of adult men, denoted Y, is normally distributed. We want to test whether $\mu = 70$ inches. Assume that we have a random sample, and the sample mean is 74.

Now the variance is not known, but assume that we have computed the standard error of x in our sample, which turns out to be $SE(\bar{x}) = 2$.

There are two cases in which we will perform our test: a **small** and a **large sample**.

Our hypotheses are the same as before, H_0: $\mu = 70$ and H_1: $\mu \neq 70$, and we choose a significance level $\alpha = 5\%$.

Hypothesis testing

Tests concerning means when the variance is unknown

Assume that we have a sample with $n = 20$ (**small**):

1. Let's compute the t-statistic: $t = \frac{\bar{x}-\mu_0}{SE(\bar{x})} = \frac{74-70}{2} = 2$.
2. Our t-statistic follows the Student's-t distribution with degrees of freedom $n - 1 = 19$.
3. We have to use the table for the Student's-t distribution to find the critical values.
4. The critical values are $t_{\alpha/2} = 2.093$ and $t_{1-\alpha/2} = -2.093$. In absolute value, this is larger than the t-statistic.
5. Thus, now we cannot reject the null hypothesis. The small sample does not provide enough statistical evidence to reject H_0.

Hypothesis testing

Tests concerning means when the variance is unknown

Assume now that our sample is **large** ($n > 30$). Then, $t \sim N(0, 1)$ approximately:

1. The only thing that changes from the example before is that now our test statistic follows a standard normal distribution, so we need to get critical values for $\alpha = 0.05$ from the standard normal, which are -1.96 and 1.96.
2. **Decision**: we reject the null, as the actual value of the statistic is larger than the critical value (1.96).

Two-sided tests:

1. Choose α.

2. Based on the chosen α, determine the critical value c such that $\Pr(-c \leq t \leq c) = 1 - \alpha$.

3. Compute the test statistic t in the sample:
   - If $t < -c$, reject the null.
   - If $t > c$, reject the null.
   - If $-c \leq t \leq c$, do not reject the null.

Hypothesis testing

Hypothesis testing with critical values

One-sided tests: suppose that we want to test the null $H_0 : \mu = 50$ against an alternative $H_1 : \mu > 50$ or $H_1 : \mu < 50$. The test statistic does not change, only the critical values do:

1. Choose α.

2. If the alternative is such that $H_1 : \mu > \mu_0$, determine the critical value c such that $\Pr(t \leq c) = 1 - \alpha$.

3. Compute the test statistic t in the sample:
   - If $t > c$, reject the null.
   - If $t \leq c$, do not reject the null.

Hypothesis testing

The p-value

Given a sample, we define the **p-value** as the smallest size of the test at which we can reject the null hypothesis. P-values provide an alternative way to do hypothesis testing.

Notice that the size is something we choose and it does not depend on the sample we observe, whereas the p-value is not chosen and it depends on the observed sample.

The smaller the p-value is, the more statistical evidence there is against H_0.

Assume that we have a large sample so that our test statistic is distributed as standard normal.

Hypothesis testing

The p-value

Two-sided test:

1. Compute the actual value of the test statistic t.
2. Compute the p-value (instead of the critical values):

$$\text{p-value} = 2\Phi(-|t|),$$

   where $\Phi(.)$ is the standard normal cdf.

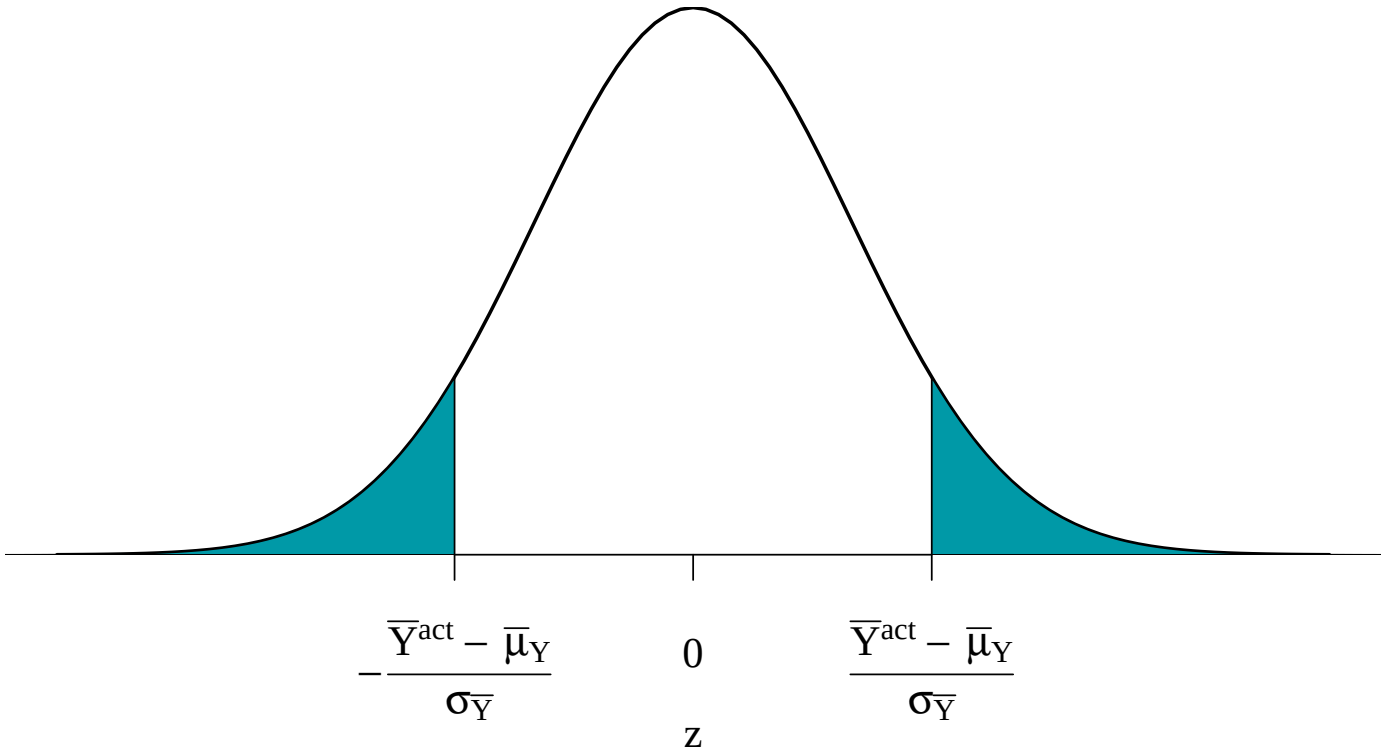
3. Decision: reject H_0 if $p \leq \alpha$.

Going back to the previous example, our test statistic was $t = 2$. Let's compute the p-value:

$$\begin{aligned} p &= 2 \times \Phi(-|2|) = 2 \times (1 - \Phi(2)) \\ &= 2 \times 0.0288 = 0.0456 = 4.56\% < 5\% = \alpha. \end{aligned}$$

Given that the p-value is smaller than the significance level α, we reject H_0.

Figure 7: Calculating a P-value



Exercise

Assume that you have as data points 2, 4, 4 and 10, coming from a normal distribution $\mathcal{N}(\mu, \sigma^2)$. For $\alpha = 0.05$, test the hypothesis

$$H_0 : \mu = 0$$

versus

$$H_1 : \mu \neq 0,$$

assuming

1. $\sigma^2 = 16$ is known.
2. σ^2 is unknown.

Hypothesis testing

The p-value

One-sided test: the only thing that changes in computing the p-value is that we only consider one of the tails. If the alternative H_1 is such that $\mu > \mu_0$:

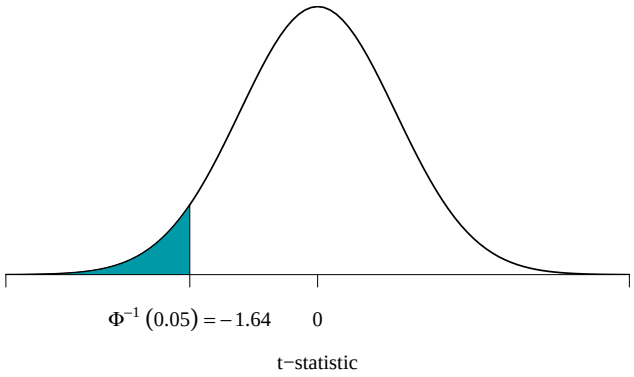
- Compute the actual value of the test statistic t.
- Compute the p-value (instead of the critical values). This is the probability that we would observe a test statistic greater than the one obtained from the data if μ were actually equal to μ_0:

$$\text{p-value} = \Phi(-|t|) = 1 - \Phi(t)$$

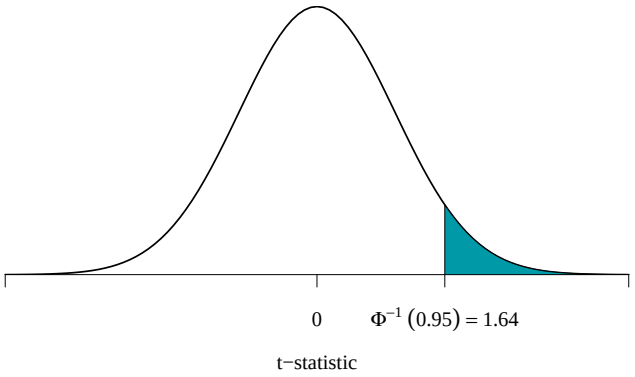
- Decision: reject H_0 if $p \leq \alpha$, i.e., if it is unlikely that we would observe such an extreme test statistic as the one computed in the direction of H_1 if the the null hypothesis were true (same decision rule in case $H_1 : \mu < \mu_0$).

Figure 8: Rejection Region of a One-Sided Test

(a) Left-sided test ($\mu_Y < \mu_{Y,0}$)



(b) Right-sided test ($\mu_Y > \mu_{Y,0}$)



Hypothesis testing

Tests concerning differences between means

Use the test statistic

$$\frac{\bar{x}_1 - \bar{x}_2 - \delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \overset{a}{\sim} \mathcal{N}(0,1)$$

where $\bar{x}_1$ and $\bar{x}_2$ are the means of two independent random samples and $H_0 = \mu_1 - \mu_2 = \delta$.

If $n_1 \geq 30$ and $n_2 \geq 30$, we can replace σ_1^2 and σ_2^2 by s_1^2 and s_2^2, respectively.

Hypothesis testing

Tests concerning differences between means

If $n_1 < 30$ or $n_2 < 30$, σ_1^2 and σ_2^2 are equal but unknown, and the two random samples are independent and come from a normal population, use the test statistic

$$\frac{\bar{x}_1 - \bar{x}_2 - \delta}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

where

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

Hypothesis testing

Tests concerning variances

Assume that you have two random samples $\{x_{1i}\}_{i=1}^{n_1}$ and $\{x_{2i}\}_{i=1}^{n_2}$ of $\{X_{1i}\}_{i=1}^{n_1}$ and $\{X_{2i}\}_{i=1}^{n_1}$ respectively, with X_{1i} and X_{2i} following the same normal distribution for all i.

- To test $H_0 : \sigma_j^2 = \sigma_0$, use the statistic

$$\frac{(n_j - 1)s^2}{\sigma_0^2} \sim \chi^2_{n_j - 1}.$$

- To test $H_0 : \sigma_1^2 = \sigma_2^2$, use the statistic

$$\frac{s_1^2}{s_2^2} \sim F_{n_1 - 1, n_2 - 1}.$$

Confidence Intervals

Confidence intervals

In general, a point estimate $\bar{x}$ of μ will not equal μ. Therefore, we might want to know the range that μ will be in. Typically, we want to determine an interval that will contain the true population mean μ, for instance, 95% of the time.

The **confidence interval** for μ is

$$[\bar{x} + z_{\alpha/2} SE(\bar{x}), \bar{x} + z_{1-\alpha/2} SE(\bar{x})].$$

Replacing the critical values for $\alpha = 0.05$, the 95% confidence interval is

$$[\bar{x} - 1.96 \times SE(\bar{x}), \bar{x} + 1.96 \times SE(\bar{x})].$$

Example

Using the values from the previous example, $\bar{x} = 74$ and $SE(\bar{x}) = 2$, we get

$$[74 - 1.96 \times 2, 74 + 1.96 \times 2] = [70.08, 77.92].$$

This is a numerical interval, **not a random interval**. This interval will or will not contain (data will tell) the true value of the parameter, but it is wrong to say that "this interval contains the true value of the parameter x% of the times".

What is right to say is that, if we calculate this interval for a very large number of random samples, approximately $100(1 - \alpha)\%$ of the resulting intervals will contain the true value of the parameter.

Simple Linear Regression

Simple Linear Regression

Conditional expectation

If X and Y are two random variables, the **conditional expectation** of Y given X is $\mathbb{E}(Y|X)$.

Properties:

- $\mathbb{E}(X|X) = X$.
- The **law of iterated expectations** tells us that

$$\mathbb{E}(Y) = \mathbb{E}\left[\mathbb{E}(Y|X)|X\right],$$

  where the outer expectation is computed using the marginal distribution of X.

Simple Linear Regression

In the simple linear regression model, our goal is to study the relation between two random variables x and y. Let's assume that in the population

$$y = f(x) + u,$$

where u is a random **error**. This model allows us to predict y for any given value of x.

- x is called the **independent variable** (or explanatory variable, control variable, predictor, regressor, or covariate).
- y is called the **dependent variable** (or explained variable, response variable, predicted variable, or regressand).

Simple Linear Regression

The following equation defines the **simple linear regression model**, and we assume that it holds in the population of interest:

$$y = \beta_0 + \beta_1 x + u.$$

A linear regression is linear in the **parameters**, but not necessarily in the explanatory variables. Some examples are:

- $y_i = \beta_0 + \beta_1 x_i^2 + u$
- $y_i = \beta_0 + \beta_1 \log(x_i) + u$

Simple Linear Regression

In the linear regression model, we assume that, if there is an intercept, then $\mathbb{E}(u) = 0$, which implies that the **error** has **zero mean** in the population.

To make sure that we are isolating the effect of x on y, we need to establish how u is related to x. We want these to be as unrelated as possible, but we do not want to impose the assumption of independence. We thus assume that the **conditional mean is zero**:

$$\mathbb{E}(u|x) = 0,$$

i.e., for any given value of x, the mean of u (anything unobserved) is the same and therefore must equal the average value of u in the entire population.

Simple Linear Regression

The conditional independence assumption implies that the covariance between x and u is zero:

$$\mathrm{Cov}(x, u) = \mathbb{E}(xu) - \mathbb{E}(x)\mathbb{E}(u) = \mathbb{E}(xu).$$

We can obtain $\mathbb{E}(xu) = 0$ by the law of iterated expectations:

$$\mathbb{E}(xu) = \mathbb{E}\left[\mathbb{E}(xu)|x\right] = \mathbb{E}[x\,\underbrace{\mathbb{E}(u|x)}_{=0}] = 0.$$

Simple Linear Regression

Given the above assumptions, we can get the **population regression function** as

$$\begin{aligned}\mathbb{E}(y|x) &= \mathbb{E}(\beta_0 + \beta_1 x + u|x) \\ &= \beta_0 + \beta_1 \mathbb{E}(x|x) + \underbrace{\mathbb{E}(u|x)}_{=0} = \beta_0 + \beta_1 x,\end{aligned}$$

which is linear in x. Thus, we know that a one-unit change in x changes $\mathbb{E}(y|x)$ by β_1 units:

$$\frac{\partial \mathbb{E}(y|x)}{\partial x} = \beta_1.$$

Simple Linear Regression

Ordinary least squares

Imagine that we have a random sample from the population of X and Y, $\{(x_i, y_i)\}_{i=1}^n$, and we want to use this sample to make inference about the population parameters β_0 and β_1. We have to come up with estimators of β_0 and β_1, which are denoted by $\hat{\beta}_0$ and $\hat{\beta}_1$.

We use this equation to model the relationship between Y and X:

$$y_i = \beta_0 + \beta_1 x_i + u_i.$$

We would like to find the values of β_0 and β_1 that give the **best fitting line**, that is, the values that minimize the sum of squared errors. These are given by the **ordinary least squares (OLS) estimator**.

Simple Linear Regression

Ordinary least squares

The **sum of squared errors** is given by

$$\sum_{i=1}^{n} u_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

To obtain the OLS estimator of β_0 and β_1, we have to minimize the above function with respect to β_0 and β_1. The first order conditions of that problem are

$$\frac{\partial \sum_{i=1}^{n} u_i^2}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\frac{\partial \sum_{i=1}^{n} u_i^2}{\partial \beta_1} = -2 \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0.$$

Simple Linear Regression

Ordinary least squares

From the first order condition of β_0:

$$\begin{aligned}
\sum_{i=1}^{n} y_i &= \sum_{i=1}^{n} \hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i \\
\frac{1}{n}\sum_{i=1}^{n} y_i &= \frac{1}{n}\sum_{i=1}^{n} \hat{\beta}_0 - \hat{\beta}_1 \frac{1}{n}\sum_{i=1}^{n} x_i \\
\bar{y} &= \frac{1}{n} n\hat{\beta}_0 - \hat{\beta}_1 \bar{x} \\
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}.
\end{aligned}$$

Simple Linear Regression

Ordinary least squares

We can substitute for β_0 in the first order condition of β_1 to solve for β_1:

$$\sum_{i=1}^{n} x_i \left(y_i - \left(\bar{y} - \hat{\beta}_1 \bar{x} \right) - \hat{\beta}_1 x_i \right) = 0$$

$$\sum_{i=1}^{n} x_i (y_i - \bar{y}) = \hat{\beta}_1 \sum_{i=1}^{n} x_i (x_i - \bar{x})$$

Simple Linear Regression

Ordinary least squares

Given the definition of $\bar{x}$, $\sum_{i=1}^{n} x_i = n\bar{x}$. We want to show that $\sum_{i=1}^{n} x_i(x_i - \bar{x}) = \sum_{i=1}^{n}(x_i - \bar{x})^2$:

$$\sum_{i=1}^{n} x_i(x_i - \bar{x}) = \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i\bar{x} = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2.$$

Note that

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \bar{x})^2 &= \sum_{i=1}^{n} x_i^2 - 2\sum_{i=1}^{n} x_i\bar{x} + \sum_{i=1}^{n} \bar{x}^2 \\ &= \sum_{i=1}^{n} x_i^2 - 2n\bar{x}\bar{x} + n\bar{x}^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2,\end{aligned}$$

so we see these two are equivalent.

Simple Linear Regression

Ordinary least squares

Second, we want to show that
$\sum_{i=1}^{n} x_i(y_i - \bar{y}) = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$:

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) &= \sum_{i=1}^{n} x_i(y_i - \bar{y}) - \sum_{i=1}^{n} \bar{x}y_i + n\bar{x}\bar{y} \\ &= \sum_{i=1}^{n} x_i(y_i - \bar{y}) - n\bar{y}\bar{x} + n\bar{x}\bar{y} \\ &= \sum_{i=1}^{n} x_i(y_i - \bar{y}).\end{aligned}$$

and thus

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

which is equal to $s_{x,y}$ divided by s_x^2.

Simple Linear Regression

Ordinary least squares

Having computed the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, it is possible to obtain the **fitted values** of y as

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i.$$

Having the fitted values of y, we can compute the **residuals** $\hat{u}$:

$$\hat{u}_i = y_i - \hat{y}_i.$$

Simple Linear Regression

Ordinary least squares

There are two important properties of OLS concerning the residuals:

1. A direct implication of the first order condition for β_0 is that $\sum_{i=1}^{n} \hat{u}_i = 0$.
2. We know that $\mathrm{Cov}(x, u) = \mathbb{E}(xu) - \mathbb{E}(x)\mathbb{E}(u)$. Its sample counterpart is

$$\frac{1}{n}\sum_{i=1}^{n}(x_i\hat{u}_i) - \frac{1}{n}\sum_{i=1}^{n} x_i \frac{1}{n}\underbrace{\sum_{i=1}^{n}\hat{u}_i}_{=0} = \frac{1}{n}\underbrace{\sum_{i=1}^{n}(x_i\hat{u}_i)}_{=0} = 0.$$

Simple Linear Regression

Goodness of fit

The **total sum of squares (TSS)** is given by $\sum_{i=1}^{n}(y_i - \bar{y})^2$, and it represents the total variation of the dependent variable.

The **residual sum of squares (RSS)** is given by $\sum_{i=1}^{n}(y_i - \hat{y})^2$, and tells us how much of the variation of the dependent variable is not explained by the model.

We can use the TSS and the RSS to compute a measure of the **goodness of fit** of the model called the **R-squared**:

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}.$$

This is the fraction of the variance of y explained by the model. The R-squared is bounded by 0 and 1 ($0 \leq R^2 \leq 1$) and can only be equal to 1 if all the data points lie on the fitted curve, i.e. $y_i = \hat{y}_i$ for all i.

Simple Linear Regression

Assumptions of the simple linear regression model

1. The population is modeled as $y = \beta_0 + \beta_1 x + u$.
2. We have a random sample of size n: $\{(x_i, y_i)\}_{i=1}^n$.
3. Zero conditional mean: $\mathbb{E}(u|x) = 0$. For a random sample, this implies $\mathbb{E}(u_i|x_i) = 0$ for all i.
4. Sample variation in the independent variable x (we need it so that the denominator of $\hat{\beta}$ is not close to zero).

If these assumptions hold, the OLS estimator is unbiased:

$$\mathbb{E}(\hat{\beta}_0) = \beta_0$$
$$\mathbb{E}(\hat{\beta}_1) = \beta_1.$$

Hypothesis Testing in OLS

Hypothesis Testing in OLS

Variance of the OLS estimator

In order to run tests on $\hat{\beta}$, we need to know its variance, which depends on the variance of the error term. The standard error of $\hat{\beta}$ can be computed using an estimate of the variance.

In order to derive it, we need to assume **homoskedasticity**:

$$\mathrm{Var}(u|x) = \sigma^2.$$

Now we need to derive $\mathrm{Var}(\hat{\beta}|x)$. We know that $\hat{\beta}$ is a function of x and y, so:

$$\mathrm{Var}(y|x) = \mathrm{Var}(\beta_0 + \beta_1 x + u|x) = \mathrm{Var}(u|x) = \sigma^2,$$

and we can write

$$\mathrm{Var}(\hat{\beta}_0) = \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2} \qquad \mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Hypothesis Testing in OLS

Variance of the OLS estimator

The variance of $\hat{\beta}_0$ and $\hat{\beta}_1$ depends on the variance of the error term, σ^2, which is unknown. However, we can estimate σ^2.

The variance of u is

$$\begin{aligned}\mathrm{Var}(u|x) &= \mathbb{E}(u^2|x) - (\underbrace{\mathbb{E}(u|x)}_{=0})^2 = \mathbb{E}(u^2|x) \\ &= \mathbb{E}(\mathbb{E}(u^2|x)|x) = \mathbb{E}(u^2) = \sigma^2,\end{aligned}$$

so the conditional variance of u is equal to the unconditional expectation of u^2.

Hypothesis Testing in OLS

Variance of the OLS estimator

An unbiased estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n-k-1} \sum_{i=1}^{n} \hat{u}_i^2,$$

where k is the number of explanatory variables.

We can substitute this estimate into the variance of both $\hat{\beta}$s and obtain their standard error:

$$SE\left(\hat{\beta}_0\right) = \sqrt{\frac{\hat{\sigma}^2 \sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

$$SE\left(\hat{\beta}_1\right) = \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}.$$

Hypothesis Testing in OLS

t-statistic

We are usually interested in finding out if β_1 is **statistically significant**, i.e., significantly different from zero. We pose the null hypothesis $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$.

The test statistic is

$$\frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} \sim t_{n-k-1},$$

which is approximately distributed as standard normal when n is large.

Hypothesis Testing in OLS

Units of the variables

When we obtain the results of a linear regression, we would like to interpret the values of $\hat{\beta}$s. First we need to know the units used to measure each variable.

Interpretation: a one unit increase in x is associated with a β_1 units change in y.

If we are measuring x in thousands of dollars, for instance, then we would say that a thousand dollar increase in x is associated with a β_1 units change in y.

Hypothesis Testing in OLS

Non-linearities

As we have said, linearity of the model does not mean that x has to appear linearly. For instance, it is common to take natural logarithms, which changes the way to interpret the coefficients.

Model	Dep. Var.	Control	Interpretation of $\hat{\beta}$
Level-Level	y	x	1 unit increase in x results in $\hat{\beta}$ units change in y.
Level-Log	y	$\log(x)$	1 percent increase in x results in $\hat{\beta}/100$ units change in y.
Log-Level	$\log(y)$	x	1 unit increase in x results in $100\hat{\beta}\%$ units change in y.
Log-Log	$\log(y)$	$\log(x)$	1 percent increase in x results in a $\hat{\beta}\%$ change in y.

Hypothesis Testing in OLS

Dummy variables

Dummy variables are 0/1 variables. They are the result of a yes/no question or variables that only have two outcomes.

Suppose that x is a dummy variable taking value 0 if the observation corresponds to a man, and 1 when the observation corresponds to a woman. Then, if we have the model

$$y = \beta_0 + \beta_1 x + u$$

and compute $\mathbb{E}(y|x = 1) - \mathbb{E}(y|x = 0)$, we obtain

$$\mathbb{E}(y|x = 1) - \mathbb{E}(y|x = 0) = (\beta_0 + \beta_1) - \beta_0 = \beta_1.$$

Interpretation: β_1 is the effect of being woman on y.

Bonus: More on Estimation Methods

Method of Moments

Let $\theta \in \Theta \subset \mathbb{R}^K$ be the vector of parameters to estimate. The kth **sample moment** of the value $y = (x_1, x_2, ..., x_n)$ taken by a random sample is defined as

$$m'_k(y) = \frac{\sum_{i=1}^n x_i^k}{n}.$$

Let $\mu'_k(\theta)$ be the kth **population moment**.

Estimation of θ by the **method of moments (MM)** consists of solving the system of equations

$$m'_k(y) = \mu'_k(\theta) \qquad \text{for} \qquad k = 1, 2, ..., K.$$

The solution is the estimate $\hat{\theta}_{\text{MM}}$.

Method of Moments

The starting point for the MM estimator is the **analogy principle**: we can estimate a parameter by replacing a population moment condition with its sample analogue.

For example, the population mean μ solves the first (central) population moment:

$$\mathbb{E}(y - \mu) = 0.$$

The analogy principle tells us to replace the population-expectations operator with its sample analogue:

$$\mathbb{E}(y - \mu) = 0 \rightarrow \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mu}) = 0 \rightarrow \hat{\mu}_{\text{MM}} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Method of Moments

The OLS estimation can also be viewed as an MM estimator. In the model

$$y = \boldsymbol{x}'\boldsymbol{\beta} + u$$

we assume that u has mean zero conditional on x: $\mathbb{E}(u|\boldsymbol{x}) = 0$. We know that, by the law of iterated expectations, $\mathbb{E}(\boldsymbol{x}u) = 0$.

Applying the analogy principle,

$$\mathbb{E}\left[\boldsymbol{x}(y - \boldsymbol{x}'\boldsymbol{\beta})\right] \to \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x_i}\left(y_i - \boldsymbol{x_i'}\boldsymbol{\beta}\right) = \mathbf{0}$$

and solving for $\boldsymbol{\beta}$,

$$\hat{\boldsymbol{\beta}}_{\text{MM}} = \left(\sum_i \boldsymbol{x_i}\boldsymbol{x_i'}\right)^{-1} \sum_i \boldsymbol{x_i} y_i,$$

which is equivalent to $\hat{\boldsymbol{\beta}}_{\text{OLS}}$.

Method of Moments

We can generate sample moments of order larger than K,

$$m'_k(y) = \mu'_k(\theta) \text{ for } k = 1, 2, ..., J, \text{ where } J \geq K.$$

However, the previous system of J equations and K unknowns does not have solution generically when $J > K$.

In this case, the **Generalized Method of Moments (GMM)** estimate is given by

$$\hat{\theta}_{\text{GMM}} = \arg\min_{\theta \in \Theta} \left\{ \left[m'(y) - \mu'(\theta) \right]' W \left[m'(y) - \mu'(\theta) \right] \right\},$$

where $[m'(y) - \mu'(\theta)]' = \Big(..., m'_j(y) - \mu'_j(\theta), ... \Big)_{1 \times J}$ and $W_{J \times J}$ is a weighting matrix.

Note that if $J = K$ and $W = I$, then $\hat{\theta}_{\text{MM}} = \hat{\theta}_{\text{GMM}}$.

Maximum Likelihood Estimation

Let $\theta \in \Theta \subset \mathbb{R}^K$ be the vector of parameters to estimate and $y = (x_1, x_2, ..., x_n)$ the value of the random vector taken by a random sample, with $f(y; \theta)$ being the value of its pdf.

Define the **likelihood function** $L(\theta; y) = f(y; \theta)$, which maps θ and y into the probability (or density) that the sample y is obtained from the specified model, characterized by θ.

The **maximum likelihood (ML)** estimate of θ is given by

$$\hat{\theta}_{\mathrm{ML}} = \arg\max_{\theta \in \Theta} L(\theta; y).$$

Maximum Likelihood Estimation

The ML estimator is based on the **likelihood principle**: the estimate of the *true* parameter θ_0 is given by the value of θ that maximizes the chance of observing our sample y.

In the case of **discrete** data, this "chance" or "likelihood" is given by the probability $\Pr[y;\theta]$ of drawing the sample. In the case of **continuous** data, it is given by its probability density function $f(y;\theta)$.

Maximum Likelihood Estimation

If $y = (x_1, x_2, ..., x_n)$ is a **random sample** from a $Y = (X_1, X_2, ..., X_n)$ with underlying pdf $f(x_i; \theta)$, then

$$L(\theta; y) = f(y; \theta) = \prod_{i=1}^{n} f(x_i; \theta).$$

In this case, maximizing the likelihood is equivalent to maximizing the **log-likelihood**:

$$\hat{\theta}_{\mathrm{ML}} = \arg\max_{\theta \in \Theta} L(\theta; y) = \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \ln f(x_i; \theta).$$

Maximum Likelihood Estimation

Example

Suppose that we are flipping a coin that may be biased, so that the probability of getting heads may not be 0.5. We are interested in estimating this probability.

Let $Y = \mathbb{1}\{\text{heads}\}$ be a binary variable that indicates whether or not we get heads in a toss. Then, this is a Bernoulli random variable with probability function

$$f_Y(y; p_0) = \begin{cases} p_0^y (1 - p_0)^{1-y} & \text{for } y \in \{0, 1\} \\ 0 & \text{otherwise.} \end{cases}$$

Thus, the likelihood function in this case is

$$L(p; y) = \prod_{i=1}^{n} p^{y_i} (1 - p)^{1-y_i}.$$

Maximum Likelihood Estimation

Example

Taking logs, we can obtain $\hat{p}_{\mathrm{ML}}$ by solving

$$\hat{p}_{\mathrm{ML}} = \arg\max_{p} \frac{1}{n} \sum_{i=1}^{n} y_i \ln p + (1 - y_i) \ln(1 - p),$$

where the scalar $1/n$ is added for convenience.

Taking derivatives in a representative term yields

$$\frac{\partial \ln L(p; y_i)}{\partial p} = \frac{y_i}{p} - \frac{(1 - y_i)}{(1 - p)} = \frac{y_i - p}{p(1 - p)}.$$

Maximum Likelihood Estimation

Example

For the n observations in the sample,

$$\frac{\partial \ln L(p; y)}{\partial p} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i - p}{p(1-p)}.$$

Setting to zero and solving for $\hat{p}_{\mathrm{ML}}$ gives

$$\hat{p}_{\mathrm{ML}} = \bar{y}.$$

Maximum Likelihood Estimation

Example

The OLS estimator can also be seen as a ML estimator. Suppose that $y \sim \mathcal{N}\left(\mu_0, \sigma_0^2\right)$. Then the pdf of y is

$$f\left(y; \mu_0, \sigma_0^2\right) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{\left(y-\mu_0\right)^2}{2\sigma_0^2}\right).$$

In the model

$$y = \boldsymbol{x}'\boldsymbol{\beta} + u$$

the mean μ_0 depends on the regressors $\boldsymbol{x}$. Thus, the density of y conditional on $\boldsymbol{x}$ is

$$f\left(y|\boldsymbol{x}; \beta_0, \sigma_0^2\right) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{\left(y-\boldsymbol{x}'\boldsymbol{\beta_0}\right)^2}{2\sigma_0^2}\right).$$

Maximum Likelihood Estimation

Example

With an i.i.d. sample of size n, the overall conditional density is the product of the conditional density of each observation:

$$f\left(y_1, ..., y_n | \boldsymbol{x_1}, ..., \boldsymbol{x_n}; \beta_0, \sigma_0^2\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{\left(y_i - \boldsymbol{x_i'}\boldsymbol{\beta_0}\right)^2}{2\sigma_0^2}\right).$$

Taking logs, our objective function becomes

$$\ln L(\boldsymbol{\beta}, \sigma^2; y|\boldsymbol{x}) = -n \ln \sqrt{2\pi} - n \ln \sigma - \sum_{i=1}^{n} \frac{\left(y_i - \boldsymbol{x_i'}\boldsymbol{\beta}\right)^2}{2\sigma^2}.$$

The solution to the maximization problem with respect to $\boldsymbol{\beta}$ and σ is given by

$$\hat{\boldsymbol{\beta}}_{\mathrm{ML}} = \left(\sum_i \boldsymbol{x_i}\boldsymbol{x_i'}\right)^{-1} \sum_i \boldsymbol{x_i} y_i$$

$$\hat{\sigma}^2_{\mathrm{ML}} = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \boldsymbol{x_i'}\boldsymbol{\beta}\right)^2.$$

Bayesian Estimation

So far, we have adopted a **frequentist approach**: we have assumed that data is generated by a model characterized by a parameter vector, and the observed data is generated from that model at a particular value of the parameter vector θ_0.

From a **Bayesian approach**, we view data as given, and update **beliefs** about the parameters using the information contained in the data:

- The density $\pi(\theta)$, known as the **prior**, reflects current beliefs about the parameters before observing the sample. The statistician is responsible of providing this density.
- The density of the sample $y = (y_1, y_2, ..., y_n)$, given a parameter value θ, is given by the **likelihood function** $f(y|\theta)$.

Bayesian Estimation

Using these two pieces, we can write the **joint density** of the sample and the beliefs about θ:

$$f(y, \theta) = f(y|\theta)\pi(\theta).$$

We can obtain the **marginal likelihood** by integrating out θ over its support Θ:

$$f(y) = \int_{\Theta} f(y, \theta) \mathrm{d}\theta.$$

Using **Bayes' theorem**,

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)},$$

we can obtain the **posterior** of θ:

$$f(\theta|y) = \frac{f(y, \theta)}{f(y)} = \frac{f(y|\theta)\pi(\theta)}{f(y)}.$$

Bayesian Estimation

While the **frequentist approach** suggests estimating θ by maximizing the likelihood of observing the actual sample, the **Bayesian approach** updates the prior knowledge about θ given the data.

The posterior distribution $f(\theta|y)$ reflects the statistician's beliefs about θ after this updating process.

Some reasons to be Bayesian:

- In some contexts, Bayesian estimators may be easier to compute reliably than analogous frequentist estimators.
- Possibility of using information about θ from previous work.

Bayesian Estimation

Example

Suppose $y \sim \mathcal{N}(\theta, \sigma^2)$, where σ^2 is known but the scalar parameter θ is not. Given a random sample $\boldsymbol{y} = (y_1, ..., y_n)$, the joint density of $\boldsymbol{y}$ is

$$f(\boldsymbol{y}|\theta) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_i - \theta)^2}{2\sigma^2}\right).$$

Suppose that our prior for θ is $\theta \sim \mathcal{N}(\mu, \tau^2)$, with specified values of the prior mean μ and prior variance τ^2. Then the prior density is

$$\pi(\theta) = \frac{1}{\tau\sqrt{2\pi}} \exp\left(-\frac{(\theta - \mu)^2}{2\tau^2}\right).$$

Bayesian Estimation

Example

Using this information, we can obtain the posterior density

$$f(\theta|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\theta)\pi(\theta)}{\int_{-\infty}^{\infty} f(\boldsymbol{y}|\theta)\pi(\theta)\mathrm{d}\theta}, \ -\infty < \theta < \infty.$$

This gives us the **posterior mean**

$$\mu_1 = \tau_1^2 \left(n\frac{\bar{y}}{\sigma^2} + \frac{\mu}{\tau^2} \right)$$

and **posterior precision** τ_1^{-2}, inverse of the precision parameter

$$\tau_1^2 = \frac{1}{\left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right)}.$$

Therefore, when n is small, the information from the sample has less weight, and more information comes from the prior.