EN.553.431 Honors Mathematical Statistics

Daniel Yao

Fall 2024

Contents

0.0 Introduction.

Introduction.

0.1 Remark.

The following notes follow the material presented in EN.553.431 Honors Mathematical Statistics taught by Professor Avanti Athreya during the semester of Fall 2024 at The Johns Hopkins University. The content of lectures is presented along with selected homework exercises.

0.2 Notation.

Rice shall refer to 'Mathematical Statistics and Data Analysis' 3rd edition (US) by John A. Rice.

0.3 Notation.

The following abbreviations shall be observed.

1. The term 'rv' shall denote 'random variable'.
2. The term 'pmf' shall denote 'probability mass function'.
3. The term 'pdf' shall denote 'probability density function'.
4. The term 'cdf' shall denote 'cumulative distribution function'.
5. The term 'iid' shall denote 'independent and identically distributed'.
6. The term 'mom' shall denote 'method of moments'.
7. The term 'mle' shall denote 'maximum likelihood estimator'.
8. The term 'df' shall denote degrees of freedom.
9. The term 'glr' shall denote 'generalized likelihood ratio'.
10. The term 'logarithm log' shall denote the 'natural logarithm ln'.

0.4 Notation

A finite population sample shall be as follows. We are given a finite bivariate population of N distinct objects, and associated to each object k is a pair of measurements (x_k, y_k). Suppose our population of measurements is represented by $\{(x_1, y_1), \ldots, (x_N, y_N)\}$. We assume $N > 1$. Let τ_x and τ_y be the population totals of the x- and y-measurements, respectively; let μ_x and μ_y be the population means of the x- and y-measurements, respectively; let σ_x^2 and σ_y^2 denote the population variances of the x- and y-measurements, respectively. Let σ_{xy} denote the population covariance. The population 3rd and 4th moments, $\mu_3(x)$ and $\mu_4(x)$, respectively, of the x-values are

$$\mu_3(x) = \frac{1}{N}\sum_{k=1}^{N} x_k^3, \quad \mu_4(x) = \frac{1}{N}\sum_{k=1}^{N} x_k^4$$

Similarly, the population 3rd and 4th moments are, respectively, $\mu_3(y)$ and $\mu_4(y)$. Let $\sigma_{x^2y^2}$ denote

$$\sigma_{x^2y^2} = \left(\frac{1}{N}\sum_{k=1}^{N} x_k^2 y_k^2\right) - (\sigma_x^2 + \mu_x^2)(\sigma_y^2 + \mu_y^2)$$

Let M_x and M_y represent the population maximum of the x- and y-values, respectively, so that M_x and M_y are defined by

$$M_x = \max\{x_k : 1 \leq k \leq N\}, \quad M_y = \max\{y_k : 1 \leq k \leq N\}$$

Let m_x and m_y denote the population minimum of the x- and y-measurements, respectively, so that

$$m_x = \min\{x_k : 1 \leq k \leq N\}, \quad m_y = \min\{y_k : 1 \leq k \leq N\}$$

All sample sizes n satisfy $n \geq 1$, and in some cases, we specify if $n > 1$ or we give an explicit value for n. In what follows below, $\bar{X}$ denotes the sample mean of the x-measurements in the sample, and $\bar{Y}$ denotes the sample mean of the y-measurements in the sample. The letter E represents expected value; Var represents variance; and Cov represents covariance.

1.0 Introduction.

Probability.

1.1 Remark.

This course EN.553.431 Honors Mathematical Statistics is taken after the prerequisite EN.553.420 Probability or EN.553.420 Honors Probability. The following notes shall be brief.

1.2 Definition.

If X and Y are rvs with respective pdfs f_x and f_y and joint pdf f_{xy}, then the conditional pdf of Y given $X = x$ is

$$f_{y|x}(y|x) = \frac{f_{xy}(x,y)}{f_x(x)}.$$

1.3 Definition.

If X and Y are rvs with respective pdfs f_x and f_y, then X and Y are independent if and only if the joint pdf

$$f_{xy}(x,y) = f_x(x)f_y(y).$$

This implies that the conditional pdf of Y given $X = x$ is

$$f_{y|x}(y|x) = f_y(y),$$

i.e., the marginal pdf of Y.

1.4 Definition.

If X and Y are rvs with joint pdf f_{xy}, then X and Y are exchangeable if and only if

$$f_{xy}(x,y) = f_{xy}(y,x).$$

That is, the joint pdf is symmetric under permutations of its arguments.

1.5 Theorem.

(Markov's Inequality.) Let X be a nonnegative rv. Then, for any $a > 0$,

$$P(X \geq a) \leq \frac{\mathrm{E}(X)}{a}.$$

1.6 Proof.

(Continued.) Observe that X may be written as

$$X = XI_{\{X<a\}} + XI_{\{X\geq a\}}.$$

Then, since $X \geq 0$ and expectation is monotonic, we have that

$$\begin{aligned} \mathrm{E}(X) &\leq \mathrm{E}(XI_{\{X\geq a\}}) \\ &\leq aP(X \geq a) \\ P(X \geq a) &\leq \frac{\mathrm{E}(X)}{a}. \end{aligned}$$

1.7 Theorem.

(Chebyshev's Inequality.) Let X be a rv with finite mean μ and finite variance σ^2. Then, for any $\delta > 0$,

$$P(|X - \mu| \geq \delta) \leq \frac{\sigma^2}{\delta^2}.$$

1.8 Proof.

(Continued.) From Markov's Inequality, take

$$X = (X - \mu)^2$$

and

$$\mathrm{E}\,(X - \mu)^2 = \sigma^2$$

to obtain Chebyshev's Inequality.

1.9 Theorem.

(Cauchy-Schwarz Inequality.) Let X and Y be rvs with finite second moments. Then,

$$|\mathrm{Cov}(X, Y)| \leq \sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}.$$

1.10 Theorem.

(Jensen's Inequality.) Let f be a convex function and X be a rv. Then,

$$f(\mathrm{E}(X)) \leq \mathrm{E}(f(X)).$$

If f is strictly convex, then the inequality is strict if X is not degenerate.

1.11 Definition.

For two random k-vectors X and Y, the covariance matrix is a matrix $\Sigma \in \mathbb{R}^{k \times k}$ such that

$$(\Sigma)_{ij} = \mathrm{Cov}(X_i, Y_j).$$

1.12 Theorem.

Suppose that X, Y are random k-vectors and that $A, B \in \mathbb{R}^{m \times k}$. Then, the covariance matrix between AX and BY is an $m \times m$ matrix given by

$$\mathrm{Cov}(AX, BY) = A\Sigma B^T.$$

1.13 Definition.

For $Z_1, Z_2 \sim \mathrm{N}(0, 1)$ iid, the rvs

$$X = \sigma_x Z_1 + \mu_x$$
$$Y = \sigma_y(\rho Z_1 + \sqrt{1 - \rho^2} Z_2) + \mu_y$$

form a bivariate normal distribution with means (μ_x, μ_y), variances (σ_x^2, σ_y^2), and covariance $\sigma_{xy} = \rho\sigma_x\sigma_y$. We may write the joint pdf of X and Y in matrix form as

$$f_{x_1,x_2}(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1 - \rho^2}} \exp(-A(x_1, x_2)/2)$$

where

$$A(x_1, x_2) = (x - \mu)^T \Sigma^{-1} (x - \mu)$$

where

$$\Sigma = \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix}$$

is the covariance matrix.

1.14 Theorem.

For X, Y bivariate normal and $D \in \mathbb{R}^{2\times 2}$, the covariance matrix of

$$\begin{bmatrix} U \\ V \end{bmatrix} = D \begin{bmatrix} X \\ Y \end{bmatrix}$$

where

$$D = \begin{bmatrix} a_1 & a_2 \\ b_1 & b_2 \end{bmatrix}$$

is

$$\Sigma_{uv} = D\Sigma_{xy}D^T.$$

From the method of Jacobians, we have that the bivariate normal pdf of U, V is therefore

$$f_{uv}(u,v) = \frac{1}{2\pi\sqrt{\det(\Sigma_{uv})}} \exp\left(A(u,v)/2\right)$$

where

$$A(u,v) = (u-\mu_u)^T \Sigma_{uv}^{-1}(u-\mu_u).$$

and

$$\mu_u = D\mu_x.$$

2.0 Introduction.

Finite Population Samples.

2.1 Definition.

Let $k = 1, 2, \ldots, N$ index the objects in a finite population of size N. The objects z_k are represented as measurement-index tuples

$$z_k = (x_k, k)$$

where $x_k \in \mathbb{R}^d$ is the numerical d-tuple of measurements for object k.

2.2 Definition.

For an object z_k in a finite population, let the operators

$$\begin{aligned} \pi(z_k) &= x_k \\ \pi_i(z_k) &= x_{k,i} \\ \pi_{\text{index}}(z_k) &= k \end{aligned}$$

where π is the projection operator, π_i is the ith component projection operator, and π_{index} is the index projection operator.

2.3 Definition.

For a bivariate population of size N such that

$$z_k = (x_k, y_k),$$

the population parameters are

$$\begin{aligned} \mu_x &= \frac{1}{N} \sum_{k=1}^{N} x_k \\ \sigma_x^2 &= \frac{1}{N} \sum_{k=1}^{N} (x_k - \mu_x)^2 \\ \sigma_{xy} &= \frac{1}{N} \sum_{k=1}^{N} (x_k - \mu_x)(y_k - \mu_y) \end{aligned}$$

along with the corresponding parameters for the y-measurements. Note the similarities between μ_x and expectation, σ_x^2 and variance, and σ_{xy} and covariance. We often refer to these respective parameters as the population mean, the population variance, and the population covariance.

2.4 Definition.

The population total is

$$\begin{aligned} \tau_x &= \sum_{k=1}^{N} x_k \\ &= N\mu_x. \end{aligned}$$

2.5 Example.

An easier way to compute the population variance is

$$\begin{aligned}\sigma_x^2 &= \frac{1}{N}\sum_{k=1}^{N}(x_k-\mu)^2 \\ &= \frac{1}{N}\sum_{k=1}^{N}\left(x_k^2-2x_k\mu+\mu^2\right) \\ &= \frac{1}{N}\sum_{k=1}^{N}x_k^2-2\mu\frac{1}{N}\sum_{k=1}^{N}x_k+\mu^2 \\ &= \frac{1}{N}\sum_{k=1}^{N}x_k^2-\mu^2,\end{aligned}$$

i.e., the second population moment minus the square of the population mean. Similarly, the population covariance is

$$\sigma_{xy} = \frac{1}{N}\sum_{k=1}^{N}x_k y_k - \mu_x\mu_y.$$

2.6 Definition.

A uniform sample of size n is a sample of n objects drawn from a finite population of size N such that each object has an equal probability of being selected. We consider a sample as a collection of random objects

$$\{Z_1, Z_2, \ldots Z_n.\}$$

We consider the special cases:

(1) *With Replacement.* An object may be drawn from the population multiple times. In this case, the sample points are independent.

(2) *Without Replacement.* An object may be drawn at most once from the population. In this case, the sample points are not independent. Note that a sample without replacement can only have sample size $n \leq N$.

2.7 Lemma.

A uniform sample with replacement of size n produces iid sample points. A uniform sample without replacement of size n produces identically distributed but not independent sample points. In either case, the probability that the ith sample point is z_k is

$$P(X_i = z_k) = \frac{1}{N}$$

for all $i = 1, 2, \ldots, n$ and $k = 1, 2, \ldots, N$ and the sample points are exchangeable.

2.8 Definition.

For a finite bivariate population of size N with objects

$$z_k = (x_k, y_k),$$

the (distinct) numerical values of the x- and y-measurements are

$$\xi_l \text{ and } \eta_r$$

for $l = 1, 2, \ldots, L$ and $r = 1, 2, \ldots, R$ where L and R are the respective number of distinct x- and y-measurements. The probability that the ith sample point has $X_i = \xi_l$ is

$$P(X_i = \xi_l) = \frac{n_l}{N}$$

and similarly for the y-measurements.

2.9 Remark.

Observe the distinction between x_k and ξ_l. x_k is the x-measurement, which need not be distinct, for the kth object. That is

$$X_1 = x_k \text{ iff } Z_1 = z_k.$$

On the other hand, ξ_l is a distinct x-measurement. That is

$$X_1 = \xi_l \text{ iff } Z_1 \in \{Z_k : x_k = \xi_l\}.$$

2.10 Example.

If Z_i is the ith object in a uniform sample, then

$$\begin{aligned} \mathrm{E}(X) &= \sum_{l=1}^{L} \xi_l P(X = \xi_l) \\ &= \sum_{l=1}^{L} \xi_l \frac{n_l}{N} \\ &= \mu_x. \end{aligned}$$

2.11 Definition.

For a sample of size n, the sample mean is defined as

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

2.12 Theorem.

For a uniform sample with replacement, the sample mean has the following properties:

(1) The expectation is

$$\mathrm{E}(\bar{X}) = \mu_x.$$

(2) The variance is

$$\mathrm{Var}(\bar{X}) = \frac{1}{n}\sigma^2.$$

2.13 Proof.

(Continued.) Observe that our sample points are iid. We have that

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

since expectation is linear. We also have that

$$\begin{aligned}\mathrm{Var}(\bar{X}) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) \\ &= \frac{1}{n}\sigma^2\end{aligned}$$

since variance is bilinear.

2.14 Theorem.

If we sample uniformly with replacement, the covariance between X_i and X_j is

$$\mathrm{Cov}(X_i, X_j) = 0.$$

If we sample uniformly without replacement, the covariance between X_i and X_j is

$$\mathrm{Cov}(X_i, X_j) = -\frac{\sigma^2}{N-1}.$$

2.15 Proof.

(Continued.) If we sample uniformly with replacement, then the sample points are iid, so the covariance is zero. If we sample uniformly without replacement, then the sample points are not independent. Suppose that our sample is of size $n = N$. Then,

$$\mathrm{Var}(\bar{X}) = 0$$

because we exhaust the population. We have that

$$\begin{aligned}0 &= \mathrm{Var}(\bar{X}) \\ &= \frac{1}{N^2}\mathrm{Cov}\left(\sum_{i=1}^{N} X_i, \sum_{j=1}^{N} X_j\right) \\ &= \frac{1}{N^2}\left(\sum_{i=1}^{N}\mathrm{Var}(X_i) + \sum\sum_{i\neq j}\mathrm{Cov}(X_i, X_j)\right) \\ &= \frac{1}{N^2}\left(N\sigma^2 + N(N-1)\mathrm{Cov}(X_i, X_j)\right) \\ &= \frac{1}{N}\sigma^2 + \frac{N-1}{N}\mathrm{Cov}(X_i, X_j) \\ \mathrm{Cov}(X_i, X_j) &= -\frac{\sigma^2}{N-1}.\end{aligned}$$

2.16 Theorem.

If we sample from a bivariate population with replacement, the covariance between X_i and Y_j is

$$\mathrm{Cov}(X_i, Y_j) = 0.$$

If we sample from a bivariate population without replacement, the covariance between X_i and Y_j is

$$\mathrm{Cov}(X_i, Y_j) = -\frac{\sigma_{xy}}{N-1}.$$

2.17 Theorem.

For a uniform sample without replacement, the sample mean has the following properties:

(1) The expectation is

$$\mathrm{E}(\bar{X}) = \mu_x.$$

(2) The variance is

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

where $(N-n)/(N-1)$ is the finite population correction factor.

2.18 Proof.

(Continued.) The expectation is the same as before since the sample points are exchangeable. We have that the variance is

$$\begin{aligned}
\mathrm{Var}(\bar{X}) &= \frac{1}{n^2}\mathrm{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} X_j\right) \\
&= \frac{1}{n}\sigma^2 + \frac{n-1}{n}\mathrm{Cov}(X_i, X_j) \\
&= \frac{1}{n}\sigma^2 - \frac{n-1}{n}\frac{\sigma^2}{N-1} \\
&= \frac{1}{n}\left(1 - \frac{n-1}{N-1}\right)\sigma^2 \\
&= \frac{1}{n}\left(\frac{N-n}{N-1}\right)\sigma^2.
\end{aligned}$$

3.0 Introduction.

Estimators for Finite Population Samples.

3.1 Definition.

An estimator $\hat{\theta}$ is a function of the sample points $Z_1, Z_2, \ldots, Z_n$ that estimates a population parameter θ.

3.2 Definition.

The bias of an estimator $\hat{\theta}$ for θ is

$$\text{Bias}(\hat{\theta}) = \text{E}(\hat{\theta}) - \theta.$$

$\hat{\theta}$ is unbiased for θ if

$$\text{E}(\hat{\theta}) = \theta.$$

3.3 Definition.

An estimator $\hat{\theta}$ for θ is consistent if for all $\delta > 0$,

$$P(|\hat{\theta} - \theta| > \delta) \to 0 \text{ as } n \to \infty.$$

3.4 Definition.

For an estimator $\hat{\theta}$ of a parameter θ, the mean squared error is

$$\begin{aligned}\text{MSE}(\hat{\theta}) &= \text{E}\left((\hat{\theta} - \theta)^2\right) \\ &= \text{Var}(\hat{\theta}) + \left(\text{E}(\hat{\theta}) - \theta\right)^2 \\ &= \text{Var}(\hat{\theta}) + \text{Bias}(\hat{\theta})^2.\end{aligned}$$

3.5 Definition.

The sample mean $\bar{X}$ is an estimator of the population mean μ_x defined as

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

3.6 Theorem.

The sample mean $\bar{X}$ is an unbiased estimator of the population mean μ_x.

3.7 Proof.

(Continued.) Consider the expectation

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

3.8 Example.

$\bar{X}^2$ is not an unbiased estimator of μ_x^2. Observe that

$$\begin{aligned} \mathrm{E}(\bar{X}^2) &= \mathrm{Var}(\bar{X}) + \mu_x^2 \\ &= \frac{1}{n}\sigma^2 + \mu_x^2 \\ \mathrm{Bias}(\bar{X}^2) &= \frac{1}{n}\sigma^2 \end{aligned}$$

in the case of sampling with replacement and

$$\mathrm{Bias}(\bar{X}^2) = \frac{1}{n}\frac{N-n}{N-1}\sigma^2$$

in the case of sampling without replacement.

3.9 Definition.

The mean squared deviation $\hat{\sigma}^2$ is an estimator for the population variance σ^2 defined as

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

3.10 Definition.

The sample variance s^2 is an estimator for the population variance σ^2 defined as

$$s^2 = \frac{n}{n-1}\hat{\sigma}^2.$$

3.11 Remark.

For n large,

$$s^2 \approx \hat{\sigma}^2.$$

3.12 Theorem.

Suppose that we sample uniformly with replacement. The sample variance s^2 is an unbiased estimator of the population variance σ^2.

3.13 Proof.

(Continued.) Consider the expectation

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

We can correct this bias to obtain the unbiased estimator s^2 such that

$$\begin{aligned} \mathrm{E}(s^2) &= \mathrm{E}\left(\frac{n}{n-1}\hat{\sigma}^2\right) \\ &= \sigma^2. \end{aligned}$$

3.14 Theorem.

Suppose that we sample uniformly without replacement. The unbiased estimator $\hat{\theta}$ for σ^2 is

$$\begin{aligned}\hat{\theta} &= \frac{n}{n-1}\frac{N-1}{N}\hat{\sigma}^2 \\ &= \frac{N-1}{N}s^2.\end{aligned}$$

3.15 Proof.

(Continued.) Consider the expectation

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

We then have that

$$\begin{aligned}\mathrm{E}(s^2) &= \mathrm{E}\left(\frac{n}{n-1}\hat{\sigma}^2\right) \\ &= \frac{n}{n-1}\left(1 - \frac{1}{n}\frac{N-n}{N-1}\sigma^2\right) \\ &= \frac{n}{n-1}\left(\frac{N(n-1)}{n(N-1)}\right)\sigma^2 \\ &= \frac{N}{N-1}\sigma^2.\end{aligned}$$

Hence, the unbiased estimator $\hat{\theta}$ for σ^2 is

$$\hat{\theta} = \frac{n}{n-1}\frac{N}{N-1}\hat{\sigma}^2.$$

3.16 Theorem.

The sample variance s^2 is a consistent estimator of the population variance σ^2.

3.17 Example.

s is not an unbiased estimator of σ. Observe that $f : t \to \sqrt{t}$ is a strictly concave function. Therefore, by Jensen's Inequality, we have that

$$\begin{aligned}\mathrm{E}(\sqrt{s^2}) &\leq \sqrt{\mathrm{E}(s^2)} \\ \mathrm{E}(s) &\leq \sigma\end{aligned}$$

and the inequality is strict if the sample points X are not degenerate.

4.0 Introduction.

Dichotomous Populations.

4.1 Definition.

A population is dichotomous if at least one measurement is binary, that is

$$x_k \in \{0, 1\}$$

for all $k = 1, 2, \ldots, N$ for some measurement x.

4.2 Remark.

Dichotomous populations have a number of nice properties. These properties arise from the Bernoulli nature of the sample points.

4.3 Definition.

The population proportion is

$$p = \frac{1}{N} \sum_{k=1}^{N} x_k.$$

Note that p is also the population mean μ.

4.4 Theorem.

The population variance of a dichotomous population is bounded by

$$0 \leq \sigma^2 \leq \frac{1}{4}.$$

4.5 Proof.

(Continued.) Since our sample points are Bernoulli, we have that

$$\mathrm{Var}(X) = p(1 - p),$$

which is maximized for $p = 1/2$.

4.6 Definition.

The sample proportion $\hat{p}$ is an estimator of the population proportion p defined as

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Note that $\hat{p}$ is also the sample mean $\bar{X}$.

4.7 Theorem.

The sample proportion $\hat{p}$ is an unbiased estimator of the population proportion p.

4.8 Proof.

(Continued.) The sample mean is an unbiased estimator of the population mean, so the sample proportion is an unbiased estimator of the population proportion.

4.9 Example.

The estimator

$$\frac{1}{n-1}\hat{p}(1-\hat{p})$$

is unbiased for $\sigma^2_{\hat{p}}$.

4.10 Proof.

(Continued.) Recall that

$$\sigma^2_{\hat{p}} = \frac{1}{n}\sigma^2.$$

We have that

$$\begin{aligned} E\left(\frac{\hat{p}(1-\hat{p})}{n-1}\right) &= \frac{1}{n-1}\left(\mathrm{E}(\hat{p}) - \mathrm{E}(\hat{p}^2)\right) \\ &= \frac{1}{n-1}(p - \sigma^2_{\hat{p}} - p^2) \\ &= \frac{1}{n-1}(p(1-p) - \sigma^2_{\hat{p}}) \\ &= \frac{1}{n-1}(n\sigma^2_{\hat{p}} - \sigma^2_{\hat{p}}) \\ &= \sigma^2_{\hat{p}}. \end{aligned}$$

4.11 Theorem.

Suppose we take a uniform sample of size n from a dichotomous population. Then, for any $\varepsilon > 0$ and $\delta > 0$,

$$n \geq \frac{1}{4\varepsilon\delta^2}$$

such that

$$P(|\hat{p} - p| > \delta) \leq \varepsilon.$$

4.12 Proof.

(Continued.) Recall that the population variance is bounded by $0 \leq \sigma^2 \leq 1/4$. We therefore have that the variance of the sample proportion is

$$\mathrm{Var}(\hat{p}) \leq \frac{1}{4n}.$$

By Chebyshev's Inequality, we have that

$$\begin{aligned} P(|\hat{p} - p| > \delta) &\leq \frac{\mathrm{Var}(\hat{p})}{\delta^2} \\ &\leq \frac{1}{4n\delta^2}. \end{aligned}$$

We want

$$\frac{1}{4n\delta^2} \leq \varepsilon,$$

so we choose

$$n \geq \frac{1}{4\varepsilon\delta^2}.$$

4.13 Proof.

(Continued.) The sample mean is a consistent estimator of the population mean, so the sample proportion is a consistent estimator of the population proportion.

4.14 Theorem.

For a bivariate population of two dichotomous measurements, if the population covariance

$$\sigma_{xy} = 0,$$

then the measurements X and Y are independent.

4.15 Proof.

(Continued.) Since X and Y are Bernoulli, if they are of covariance zero, then they are independent.

5.0 Introduction.

Confidence Intervals.

5.1 Theorem.

(Central Limit Theorem.) Let $U_1, U_2, \ldots, U_n$ be iid rvs with $\mathrm{E}(U_i) = \mu$ and $\mathrm{Var}(U_i) = \sigma^2$. For

$$\bar{U} = \frac{1}{n}\sum_{i=1}^{n} U_i,$$

we have that for all $t \in \mathbb{R}$,

$$P\left(\left|\frac{\bar{U} - \mu}{\sigma/\sqrt{n}}\right| \leq t\right) \to \Phi(t) \text{ as } n \to \infty.$$

5.2 Definition.

An α-critical value z_α for an rv Z is such that

$$P(Z > z_\alpha) = \alpha.$$

That is, α is the upper-tail probability of Z.

5.3 Example.

For $\bar{X}$ approximately normal, we have that

$$P\left(-z_\alpha \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_\alpha\right) \approx 1 - 2\alpha,$$

so

$$P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \approx 1 - \alpha$$

where the (random) interval is the $1 - \alpha$ confidence interval for μ.

5.4 Note.

The population standard deviation σ may be unknown, but we may substitute the sample standard deviation s in its place.

5.5 Example.

By Chebyshev's Inequality, we have that for a sample of size n,

$$P(|s_n^2 - \sigma^2| \geq \delta) \leq \frac{\mathrm{Var}(s_n^2)}{\delta^2} \to 0 \text{ as } n \to \infty,$$

so

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} \approx \mathrm{N}(0, 1)$$

for n large.

5.6 Example.

If we sample without replacement and $n << N$, then

$$\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}} \approx \mathrm{N}(0, 1).$$

5.7 Theorem.

For the sample total

$$T_n = \sum_{i=1}^{n} X_i,$$

the CLT says that

$$P\left(\frac{T_n - n\mu}{\sigma\sqrt{n}} \leq t\right) \to \Phi(t) \text{ as } n \to \infty.$$

5.8 Note.

The sample mean and the sample total are related in that

$$\frac{n}{n}\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{T_n - n\mu}{\sigma\sqrt{n}}.$$

6.0 Introduction.

Delta Methods.

6.1 Theorem.

(Mean Value Theorem.) Suppose that $g : \mathbb{R} \to \mathbb{R}$ is differentiable on (a, b) and that $a < x < y < b$. Then there exists a $\xi \in [x, y]$ such that

$$g'(\xi) = \frac{g(y) - g(x)}{y - x}.$$

6.2 Theorem.

(Taylor's Theorem with Remainder.) Let $g : \mathbb{R} \to \mathbb{R}$ be n times differentiable on (a, b) and let $a < x < y < b$. Then there exists a $\xi \in [x, y]$ such that

$$g(y) = \sum_{k=0}^{n} \frac{g^{(k)}(x)}{k!}(y - x)^k + R_n(y)$$

where

$$R_n(y) = \frac{g^{(n+1)}(\xi)}{(n+1)!}(y - x)^{n+1}.$$

We may bound this remainder by

$$|R_n(y)| \leq \frac{M}{(n+1)!}|y - x|^{n+1}$$

where $M = \max |g^{(n+1)}(\xi)|$ on the interval (x, y).

6.3 Theorem.

For $g : \mathbb{R}^n \to \mathbb{R}$ twice-differentiable on a closed ball B containing x and y, we have that the first-order Taylor polynomial with remainder is

$$g(y) = g(x) + \nabla g(x)^T (y - x) + \frac{1}{2}(y - x)^T H(\xi)(y - x)$$

where

$$\nabla g(x) = \left(\frac{\partial g}{\partial x_1}(\xi) \quad \cdots \quad \frac{\partial g}{\partial x_n}(\xi) \right)^T$$

for some $\xi \in \mathbb{R}^n$ between x and y where is the gradient of g at x and

$$H(\xi) = \begin{pmatrix} \frac{\partial^2 g}{\partial x_1^2}(\xi) & \cdots & \frac{\partial^2 g}{\partial x_1 \partial x_n}(\xi) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 g}{\partial x_n \partial x_1}(\xi) & \cdots & \frac{\partial^2 g}{\partial x_n^2}(\xi) \end{pmatrix}$$

is the Hessian of g at ξ.

6.4 Example.

For $g : \mathbb{R}^2 \to \mathbb{R}$, the second order taylor polynomial is

$$g(y) \approx g(x) + \nabla g(x)^T (y - x) + \frac{1}{2}(y - x)^T H(x)(y - x).$$

Written in polynomial form, this is

$$\begin{aligned} g(y_1, y_2) \approx g(x_1, x_2) &+ \frac{\partial g}{\partial x_1}(y_1 - x_1) + \frac{\partial g}{\partial x_2}(y_2 - x_2) \\ &+ \frac{1}{2}\left(\frac{\partial^2 g}{\partial x_1^2}(y_1 - x_1)^2 + 2\frac{\partial^2 g}{\partial x_1 \partial x_2}(y_1 - x_1)(y_2 - x_2) + \frac{\partial^2 g}{\partial x_2^2}(y_2 - x_2)^2 \right) \end{aligned}$$

where the partial derivatives are evaluated at (x_1, x_2).

6.5 Definition.

For a random vector $X = (X_1, X_2)$ and a function $g : \mathbb{R}^2 \to \mathbb{R}$, the delta method is a method to approximate the distribution of $g(X)$ with a Taylor polynomial about μ_x.

6.6 Example.

For $h : \mathbb{R}^2 \to \mathbb{R}$, the second-order taylor polynomial about (μ_x, μ_y) is

$$\begin{aligned} h(x,y) \approx h(\mu_x, \mu_y) &+ \frac{\partial h}{\partial x}(x-\mu_x) + \frac{\partial h}{\partial y}(y-\mu_y) \\ &+ \frac{1}{2}\left(\frac{\partial^2 h}{\partial x^2}(x-\mu_x)^2 + 2\frac{\partial^2 h}{\partial x \partial y}(x-\mu_x)(y-\mu_y) + \frac{\partial^2 h}{\partial y^2}(y-\mu_y)^2\right) \end{aligned}$$

where the partial derivatives are evaluated at (μ_x, μ_y). Therefore, the expected value of $h(X, Y)$ (under certain conditions) is

$$\mathrm{E}(h(X,Y)) \approx h(\mu_x, \mu_y) + \frac{1}{2}\frac{\partial^2 h}{\partial x^2}\sigma_X^2 + \frac{\partial^2 h}{\partial x \partial y}\sigma_{XY} + \frac{1}{2}\frac{\partial^2 h}{\partial y^2}\sigma_Y^2$$

because the first-order terms vanish when

$$\mathrm{E}(X - \mu_x) = 0.$$

6.7 Example.

For $h : \mathbb{R}^2 \to \mathbb{R}$, the first-order taylor polynomial about (μ_X, μ_Y) is

$$h(X,Y) \approx h(\mu_X, \mu_Y) + \frac{\partial h}{\partial x}(X-\mu_X) + \frac{\partial h}{\partial y}(Y-\mu_Y)$$

where the partial derivatives are evaluated at (μ_X, μ_Y). Therefore, the variance of $h(X, Y)$ (under certain conditions) is

$$\mathrm{Var}(h(X,Y)) \approx \left(\frac{\partial h}{\partial x}\right)^2 \mathrm{Var}(X) + \left(\frac{\partial h}{\partial y}\right)^2 \mathrm{Var}(Y) + 2\frac{\partial h}{\partial x}\frac{\partial h}{\partial y}\mathrm{Cov}(X,Y).$$

To approximate the variance, the second-order terms become small quickly, so a first-order approximation is appropriate.

7.0 Introduction.

Sample Ratio.

7.1 Definition.

For a bivariate population and a sample of size n, the sample ratio is

$$\bar{R} = \frac{\bar{Y}}{\bar{X}}.$$

7.2 Example.

We then have that

$$g(x, y) = \frac{y}{x}$$

with partial derivatives

$$\frac{\partial g}{\partial x} = -\frac{y}{x^2}, \quad \frac{\partial g}{\partial y} = \frac{1}{x}$$
$$\frac{\partial^2 g}{\partial x^2} = \frac{2y}{x^3}, \quad \frac{\partial^2 g}{\partial x \partial y} = -\frac{1}{x^2}, \quad \frac{\partial^2 g}{\partial y^2} = 0.$$

Therefore,

$$\begin{aligned}\bar{R} \approx \frac{\mu_y}{\mu_x} - \frac{\mu_y}{\mu_x^2}(\bar{X} - \mu_x) + \frac{1}{\mu_x}(\bar{Y} - \mu_y) \\ + \frac{\mu_y}{\mu_x^3}(\bar{X} - \mu_x)^2 - \frac{1}{\mu_x^2}(\bar{X} - \mu_x)(\bar{Y} - \mu_y).\end{aligned}$$

7.3 Example.

(Continued.) The expected value of $\bar{R}$ is

$$\begin{aligned}\mathrm{E}(\bar{R}) &\approx \frac{\mu_y}{\mu_x} + \frac{1}{\mu_x^2}\left(\mathrm{Var}(\bar{X})\frac{\mu_y}{\mu_x} - \mathrm{Cov}(\bar{X}, \bar{Y})\right) \\ &= r + \frac{1}{\mu_x^2}(\mathrm{Var}(\bar{X})r - \mathrm{Cov}(\bar{X}, \bar{Y}))\end{aligned}$$

where

$$r = \frac{\mu_y}{\mu_x}$$

In the case of sampling with replacement, we have that

$$\mathrm{E}(\bar{R}) = \frac{\mu_y}{\mu_x} + \frac{1}{\mu_x^2}\left(\frac{\mu_y}{\mu_x}\frac{\sigma_x^2}{n} - \frac{\sigma_{xy}}{n}\right).$$

In the case of sampling without replacement, we have that

$$\mathrm{E}(\bar{R}) = \frac{\mu_y}{\mu_x} + \frac{1}{\mu_x^2}\left(\frac{\mu_y}{\mu_x}\frac{\sigma_x^2}{n}\frac{N-n}{N-1} - \frac{\sigma_{xy}}{n}\frac{N-n}{N-1}\right).$$

7.4 Example.

(Continued.) The variance of $\bar{R}$ is

$$\mathrm{Var}(\bar{R}) = \frac{\mu_y^2}{\mu_x^4}\sigma_{\bar{x}}^2 + \frac{1}{\mu_x^2}\sigma_{\bar{y}}^2 - \frac{2\mu_y}{\mu_x^3}\sigma_{\bar{x}\bar{y}}.$$

In the case of sampling with replacement, we have that

$$\begin{aligned}\mathrm{Var}(\bar{R}) &= \frac{\mu_y^2}{\mu_x^4}\frac{\sigma_x^2}{n} + \frac{1}{\mu_x^2}\frac{\sigma_y^2}{n} - \frac{2\mu_y}{\mu_x^3}\frac{\sigma_{xy}}{n} \\ &= \frac{1}{\mu_x^2}\frac{1}{n}\left(\frac{\mu_y^2}{\mu_x^2}\sigma_x^2 + \sigma_y^2 - \frac{2\mu_y}{\mu_x}\sigma_{xy}\right).\end{aligned}$$

and in the case of sampling without replacement, we have that

$$\begin{aligned}\mathrm{Var}(\bar{R}) &= \frac{\mu_y^2}{\mu_x^4}\frac{\sigma_x^2}{n}\frac{N-n}{N-1} + \frac{1}{\mu_x^2}\frac{\sigma_y^2}{n}\frac{N-n}{N-1} - \frac{2\mu_y}{\mu_x^3}\frac{\sigma_{xy}}{n}\frac{N-n}{N-1} \\ &= \frac{1}{\mu_x^2}\frac{1}{n}\frac{N-n}{N-1}\left(\frac{\mu_y^2}{\mu_x^2}\sigma_x^2 + \sigma_y^2 - \frac{2\mu_y}{\mu_x}\sigma_{xy}\right).\end{aligned}$$

where n, N are the sample size and population size, respectively.

7.5 Theorem.

We have the following propositions from Rice, Chapter (7), Section (7.4). Consider the case of sampling without replacement. Taking

$$r = \frac{\mu_x}{\mu_y},$$

we arrive at Theorem B, that the approximate expectation of $R = \bar{Y}/\bar{X}$ is

$$\mathrm{E}(R) \approx r + \frac{1}{n}\left(1 - \frac{n-1}{N-1}\right)\frac{1}{\mu_x^2}(r\sigma_x^2 - \rho\sigma_x\sigma_y)$$

where ρ is the correlation

$$\rho = \frac{\sigma_{xy}}{\sigma_x\sigma_y}$$

as well as Corollary B, that the approximate bias of the ratio estimate $\bar{Y}_R = \mu_x R$ of μ_y is

$$\mathrm{E}(\bar{Y}_R) - \mu_y \approx \frac{1}{n}\left(1 - \frac{n-1}{N-1}\right)\frac{1}{\mu_x}\sigma_x^2.$$

7.6 Theorem.

We have the following propositions from Rice, Chapter (7), Section (7.4). Consider the case of sampling without replacement. Taking

$$r = \frac{\mu_x}{\mu_y},$$

we arrive at Theorem A, that the approximate variance of $R = \bar{Y}/\bar{X}$ is

$$\mathrm{Var}(R) \approx \frac{1}{n}\left(1 - \frac{n-1}{N-1}\right)\frac{1}{\mu_x^2}(r^2\sigma_x^2 + \sigma_y^2 - 2r\sigma_{xy}),$$

and Corollary A, that the the estimated variance of the ratio estimate $\bar{Y}_R = \mu_x R$ of μ_y is

$$\mathrm{Var}(\bar{Y}_R) \approx \frac{1}{n}\left(1 - \frac{n-1}{N-1}\right)\left(r^2\sigma_y^2 + \sigma_x^2 - 2r\sigma_{xy}\right),$$

and Corollary C, that the variance of $\bar{Y}_R$ can be estimated by

$$s_{\bar{Y}_R}^2 \approx \frac{1}{n}\left(1 - \frac{n-1}{N-1}\right)(R^2 s_x^2 + s_y^2 - 2Rs_{xy}).$$

7.7 Theorem.

(Rice, Chapter (7), Exercise (50).) Hartley and Ross (1954) derived the following exact bound on the relative size of the bias and standard error of a ratio estimate:

$$\frac{|\mathrm{E}(R) - r|}{\sigma_R} \leq \frac{\sigma_{\bar{X}}}{\mu_x}.$$

7.8 Proof.

(Continued.) Consider the relation

$$\mathrm{Cov}(R, \bar{X}) = \mathrm{E}(R\bar{X}) - \mathrm{E}(R)\mathrm{E}(\bar{X}).$$

We have that

$$\begin{aligned} E(R\bar{X}) - E(R)E(\bar{X}) &= E(\bar{Y}) - E(R)E(\bar{X}) \\ &= \mu_y - \mu_x E(R) \end{aligned}$$

so by the Cauchy-Schwarz inequality, we have that

$$\begin{aligned} |\mu_y - \mu_x E(R)| &\leq \sigma_{\bar{x}}\sigma_R \\ |\mu_x E(R) - \mu_y| &\leq \sigma_{\bar{x}}\sigma_R \\ |E(R) - r| &\leq \frac{\sigma_{\bar{x}}\sigma_R}{\mu_x} \\ \frac{|E(R) - r|}{\sigma_R} &\leq \frac{\sigma_{\bar{x}}}{\mu_x}. \end{aligned}$$

Remark. For the ratio estimate of the population total

$$T_R = \tau_x R,$$

the squared standard error for T_R is

$$s^2_{T_R} = N^2 \frac{1}{n}\left(\frac{N-n}{N-1}\right)(R^2 s_x^2 + s_y^2 - 2Rs_{xy}).$$

Compare that to the standard error for the direct estimate T in part (c), which is

$$s^2_T = N^2 \frac{1}{n}\left(\frac{N-n}{N-1}\right) s_y^2.$$

If R is small or if s_x is small, then

$$\begin{aligned} R^2 s_x^2 + s_y^2 - 2Rs_{xy} &< s_y^2 \\ s^2_{T_R} &< s^2_T. \end{aligned}$$

The same argument holds for the variance of the ratio estimate

$$\bar{Y}_R = \mu_x R.$$

This is an example of a biased estimator possessing a smaller variance than the unbiased estimator.

8.0 Introduction.

Real Analysis.

8.1 Definition.

A field is a set F equipped with two operations: addition and multiplication. The field axioms are as follows.

(A) Addition.

(A1) If $x, y \in F$, then $x + y \in F$. (Closure.)

(A2) If $x, y \in F$, then $x + y = y + x$. (Commutativity.)

(A3) If $x, y, z \in F$, then $(x + y) + z = x + (y + z)$. (Associativity.)

(A4) There exists an element $0 \in F$ such that $0 + x = x$ for all $x \in F$. (Identity.)

(A5) To every $x \in F$ there corresponds an element $-x \in F$ such that $x + (-x) = 0$. (Inverse.)

(M) Multiplication.

(M1) If $x, y \in F$, then $xy \in F$. (Closure.)

(M2) If $x, y \in F$, then $xy = yx$. (Commutativity.)

(M3) If $x, y, z \in F$, then $(xy)z = x(yz)$. (Associativity.)

(M4) There exists an element $1 \in F, 1 \neq 0$, such that $1x = x$ for all $x \in F$. (Identity.)

(M5) If $x \in F, \neq 0$, then there corresponds an element $1/x \in F$ such that $x(1/x) = 1$. (Inverse.)

(D) Distribution.

(D1) If $x, y, z \in F$, then $x(y + z) = xy + xz$. (Left distribution.)

8.2 Definition.

An ordered set is a set S equipped with a relation $<$ such that for all $x, y, z \in S$,

(1) If $x, y \in S$, then one and only one of

$$x < y, \quad x = y, \quad y < x$$

is true. (Trichotomy.)

(2) If $x, y, z \in S$ and $x < y$ and $y < z$, then $x < z$. (Transitivity.)

8.3 Definition.

An ordered field is a field F equipped with an order relation $<$ such that for all $x, y, z \in F$,

(1) If $x, y, z \in F$ and $y < z$, then $x + y < x + z$.

(2) If $x, y \in F$ and $x, y > 0$, then $xy > 0$.'

8.4 Remark.

From these axioms, we may derive the familiar properties of $\mathbb{Q}, \mathbb{R}, \mathbb{C}$.

8.5 Definition.

A subset D of an ordered field F is said to be bounded above if there exists an element $M \in F$ such that

$$x \leq M, \quad \forall x \in D.$$

The element M is called an upper bound of D. M is a least upper bound of D if

(1) $\forall x \in D, M \leq x$.

(2) $\forall m < M, \exists x \in D$ s.t. $m < x$.

8.6 Definition.

The least upper bound property states that every nonempty subset D of F that is bounded above has a least upper bound

$$\sup D.$$

8.7 Theorem.

There exists an ordered field $\mathbb{R}$ with the least upper bound property. Moreover, $\mathbb{Q} \subset \mathbb{R}$.

8.8 Definition.

A metric space is a set X equipped with a metric $d : X \times X \to \mathbb{R}$ such that for all $x, y, z \in X$,

(1) $d(x, y) \geq 0$. (Non-negativity.)

(2) $d(x, y) = 0$ if and only if $x = y$. (Positive definiteness.)

(3) $d(x, y) = d(y, x)$. (Symmetry.)

(4) $d(x, y) \leq d(x, z) + d(z, y)$. (Triangle inequality.)

Unless otherwise specified, we assume that the standard metric is

$$d(x, y) = |x - y|.$$

8.9 Definition.

A sequence $\{x_n\}$ in $\mathbb{R}$ is the indexed output of a map

$$\phi : \mathbb{N} \to \mathbb{R}$$

and we denote the sequence as

$$\{a_n : n \in \mathbb{N}\}.$$

or simply as $\{a_n\}$ or even more simply as a_n.

8.10 Definition.

A sequence is Cauchy if

$$\forall \varepsilon > 0, \exists N_\varepsilon \in \mathbb{N} \text{ s.t. } d(a_n, a_m) < \varepsilon, \forall n, m \geq N_\varepsilon.$$

8.11 Definition.

A sequence $\{a_n\}$ in $\mathbb{R}$ is convergent if

$$\exists L \in \mathbb{R} \text{ s.t. } \forall \varepsilon > 0, \exists N_\varepsilon \in \mathbb{N} \text{ s.t. } d(a_n - L) < \varepsilon, \forall n \geq N_\varepsilon.$$

L is said to be the limit of the sequence $\{a_n\}$.

8.12 Lemma.

A sequence is Cauchy if it is convergent.

8.13 Proof.

(Continued.) Suppose that $\varepsilon > 0$. Since

$$a_n \to L,$$

there exists an $N_{\varepsilon/2} \in \mathbb{N}$ such that for all $n > N_{\varepsilon/2}$,

$$\begin{aligned} d(a_n, L) &< \frac{\varepsilon}{2} \\ d(a_m, L) &< \frac{\varepsilon}{2} \\ d(a_n, a_m) &< \varepsilon \end{aligned}$$

for all $m, n > N_{\varepsilon/2}$ by the triangle inequality. Hence, a sequence if convergent if it is Cauchy.

8.14 Definition.

A set D is complete if every Cauchy sequence in D is convergent to a limit $L \in D$.

8.15 Theorem.

$\mathbb{R}$ is complete.

8.16 Proof.

(Continued.) Suppose that a_n is a Cauchy sequence in $\mathbb{R}$ that is not bounded. Then, for all $\varepsilon > 0$, there exists an $N_\varepsilon \in \mathbb{N}$ such that

$$d(a_n, a_m) < \varepsilon, \forall n, m \geq N_\varepsilon.$$

But a_n is not bounded, so for any $\varepsilon > 0$ and $m \in \mathbb{N}$, there exists an $n \geq m$ such that

$$d(a_n, a_m) \geq \varepsilon$$

because $a_m \pm \varepsilon$ would otherwise be a bound for a_n. Hence, we have a contradiction, so $\mathbb{R}$ is complete.

8.17 Lemma.

x such that $x^2 = 2$ is irrational.

8.18 Proof.

(Continued.) Suppose that x such that $x^2 = 2$ is rational, i.e., that $x = p/q$ for some $p, q \in \mathbb{Z}$ coprime. Then,

$$\begin{aligned} \frac{p^2}{q^2} &= 2 \\ p^2 &= 2q^2. \end{aligned}$$

Hence, p^2 is even, which means that p is even. Let $p = 2k$ for some $k \in \mathbb{Z}$. Then,

$$\begin{aligned} 4k^2 &= 2q^2 \\ 2k^2 &= q^2. \end{aligned}$$

Hence, q^2 is even, which means that q is even. But p and q are coprime, so we have a contradiction. Therefore, x such that $x^2 = 2$ is irrational.

8.19 Theorem.

$\mathbb{Q}$ is not complete.

8.20 Proof.

(Continued.) Recall that a field is complete if every Cauchy sequence in the field converges to a limit in the field We shall construct a Cauchy sequence in $\mathbb{Q}$ that does not converge to a limit in $\mathbb{Q}$.

The Sequence. Consider the function $f(x) = x^2 - 2$. Recall Newton's method for finding roots of $y = f(x)$, wherein iterates are defined as

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}.$$

We have that

$$\begin{aligned} x_{n+1} &= x_n - \frac{x_n^2 - 2}{2x_n} \\ &= \frac{1}{2}x_n + \frac{1}{x_n}. \end{aligned}$$

Boundedness. Suppose that $x_n = \sqrt{2} + \varepsilon$ for some $\varepsilon > 0$. We then have that

$$\begin{aligned} x_{n+1} &= \frac{1}{2}(\sqrt{2} + \varepsilon) + \frac{1}{\sqrt{2} + \varepsilon} \\ &= \sqrt{2} + \frac{\varepsilon^2}{2(\sqrt{2} + \varepsilon)} \\ x_{n+1} &> \sqrt{2}. \end{aligned}$$

For $\varepsilon > 0$, the inequality always holds, i.e., if $x_n > \sqrt{2}$, then $x_{n+1} > \sqrt{2}$.

Monotonacity. We also have that

$$\begin{aligned} x_{n+1} - x_n &= \sqrt{2} + \frac{\varepsilon^2}{2(\sqrt{2} + \varepsilon)} - \sqrt{2} - \varepsilon \\ x_{n+1} - x_n &\leq 0. \end{aligned}$$

For $\varepsilon > 0$, the inequality always holds, i.e., if $x_n > \sqrt{2}$, then $x_{n+1} \leq x_n$. In fact, the sequence is strictly decreasing such that $x_{n+1} < x_n$.

Convergence. The sequence x_n is nonempty and bounded below, so it must have a greatest lower bound L. Suppose that x_n does not converge to L. Then,

$$\exists \varepsilon > 0 \text{ s.t. } \neg \exists N \in \mathbb{N} \text{ s.t. } |x_n - L| < \varepsilon, \quad \forall n \geq N.$$

Recall that L is a lower bound, so the only way for this statement to hold is if $x_n \geq L + \varepsilon$ for all $n \geq N$. But then $L + \varepsilon$ is a lower bound, which means that L is not the greatest lower bound, hence a contradiction. Therefore, the sequence x_n converges to L.

Observe that L cannot be strictly less than $\sqrt{2}$ because $\sqrt{2}$ is a lower bound of x_n. Observe also that L cannot be strictly greater than $\sqrt{2}$ because we define $x_1 = \sqrt{2} + \varepsilon$ for some $\varepsilon > 0$ and the sequence is strictly decreasing. Hence, $L = \sqrt{2}$.

Rationality. Suppose that $x_1 = 3/2$. Then, $x_n \in \mathbb{Q}$ for all $n \in \mathbb{N}$. We shall prove this fact by induction.

(1) *Base Case.* $x_1 = 3/2 \in \mathbb{Q}$.

(2) *Inductive Hypothesis.* Suppose that $x_n \in \mathbb{Q}$ for some $n \in \mathbb{N}$, i.e., that $x_n = p/q$ for some $p, q \in \mathbb{Z}$ coprime.

(3) *Inductive Step.* We have that

$$\begin{aligned} x_{n+1} &= \frac{1}{2}x_n + \frac{1}{x_n} \\ &= \frac{1}{2}\frac{p}{q} + \frac{q}{p} \in \mathbb{Q} \end{aligned}$$

because the sums and products of rationals are rational. Therefore, each element $x_n, n \in \mathbb{N}$, for $x_1 = 3/2$, is in $\mathbb{Q}$.

Conclusion. We have hence constructed a sequence x_n, for $x_1 = 3/2$, that is in $\mathbb{Q}$ but converges to a limit $L = \sqrt{2}$ that is not in $\mathbb{Q}$. Therefore, $\mathbb{Q}$ is not complete.

8.21 Definition.

A series is a sequence s_n of partial sums

$$s_n = \sum_{k=1}^{n} a_k.$$

for some sequence $\{a_n\}$.

8.22 Lemma.

Suppose that $a_n \in \mathbb{R}, a_n \geq 0$, i.e., that the series s_n is monotone increasing. Then, if s_n is bounded above, then s_n converges to its least supper bound, i.e.,

$$\lim_{n\to\infty} s_n = \sup s_n.$$

We often denote this limit as

$$S = \sup s_n.$$

9.0 Introduction.

Probability Spaces.

9.1 Definition.

A $\sigma-$algebra $\mathcal{F}$ is a collection of subsets of Ω such that

(1) $\emptyset \in \mathcal{F}$.

(2) If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$. (Closure under complement.)

(3) If $A_1, A_2, \ldots \in \mathcal{F}$, then

$$\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}.$$

(Closure under countable union.)

9.2 Definition.

The σ-algebra generated by a collection of subsets $\mathcal{A}$ is the smallest σ-algebra containing $\mathcal{A}$. We denote the σ-algebra generated by $\mathcal{A}$ as

$$\sigma(\mathcal{A}).$$

9.3 Example.

The σ-algebra generated by A is

$$\sigma(A) = \{\emptyset, A, A^c, \Omega\}.$$

9.4 Lemma.

If $A_1, A_2, \ldots, A_n$ partition Ω, then

$$\sigma(A_1, A_2, \ldots, A_n) = \bigcup_{i \in S} A_i$$

for all $S \subseteq \{1, 2, \ldots, n\}$. That is,

$$\sigma(A_1, A_2, \ldots, A_n) = 2^{\Omega}.$$

9.5 Example.

Consider the coin-flip space

$$\Omega = \{\omega = (\omega_1, \omega_2, \ldots) : \omega_i \in \{0, 1\}\},$$

that is, the sample space consisting of countable-length strings of 0 and 1. Denote

$$A_x = \{\omega \in \Omega : \omega_1 = x\}$$
$$A_{xy} = \{\omega \in \Omega : \omega_1 = x, \omega_2 = y\}$$

We claim that the σ-algebra $\mathcal{F}$ generated by the events

$$A_1, \quad A_{11}, \quad A_{01}$$

has cardinality $|\mathcal{F}| = 2^4$. Observe that we can form a partition of Ω by

$$P = \{A_{00}, A_{01}, A_{10}, A_{11}\}.$$

Therefore,

$$\mathcal{F} = 2^P$$

with cardinality $|\mathcal{F}| = 2^4$.

9.6 Definition.

A probability measure P is a function $P : \mathcal{F} \to [0, 1]$ such that

(1) $P : \mathcal{F} \to [0, 1]$. (Non-negativity.)

(2) If $A_1, A_2, \ldots \in \mathcal{F}$ are disjoint, then

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

(Countable additivity.)

9.7 Definition.

A probability space is a triple $(\Omega, \mathcal{F}, P)$ where the event space F is a σ-algebra of subsets of the sample space Ω and $P : \mathcal{F} \to [0, 1]$ is a probability measure on $\mathcal{F}$.

9.8 Definition.

A real-valued random variable is a function

$$X : \Omega \to \mathbb{R}$$

with a $(F, \mathcal{B}(\mathbb{R}))$-measurability requirement. That is, for all $B \in \mathcal{B}(\mathbb{R})$,

$$X^{-1}(B) \in \mathcal{F}$$

where

$$X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\}$$

is the preimage of B under X and where $\mathcal{B}(\mathbb{R})$ is the Borel σ-algebra on $\mathbb{R}$, i.e., the σ-algebra generated by the open intervals of $\mathbb{R}$.

9.9 Example.

Consider the random variable $X = I_A$, that is, indicator function of the event A. Then, for all $B \in \mathcal{B}(\mathbb{R})$,

(1) If $0 \notin B$ and $1 \notin B$, then $X^{-1}(B) = \emptyset$.

(2) If $0 \notin B$ and $1 \in B$, then $X^{-1}(B) = A$.

(3) If $0 \in B$ and $1 \notin B$, then $X^{-1}(B) = A^c$.

(4) If $0 \in B$ and $1 \in B$, then $X^{-1}(B) = \Omega$.

Therefore, X is a random variable for

$$\mathcal{F} = \{\emptyset, A, A^c, \Omega\}.$$

9.10 Example.

Consider the sample space of two die rolls, that is

$$\Omega = \{(i, j) : i, j \in \{1, 2, 3, 4, 5, 6\}\}.$$

Let $X(\omega) = \omega_1 + \omega_2$ be the sum of the two die rolls. Consider B such that $B \cap \Omega = 3$ with preimage

$$X^{-1}(B) = \{(1, 2), (2, 1)\}.$$

So $\mathcal{F}$ must include $\{(1, 2), (2, 1)\}$. But if A includes $(1, 2)$, then it must also include $(2, 1)$. Therefore,

$$\sigma(X^{-1}(\mathcal{B}(\mathbb{R}))) \neq 2^{\Omega}.$$

In fact, for any $B \in \mathcal{B}(\mathbb{R})$,

$$X^{-1}(B) = X^{-1}(B \cap \text{image}(X)).$$

In particular, $\mathcal{F}$ is the σ-algebra generated by

$$A = \{X^{-1}(\{2\}), X^{-1}(\{3\}), \ldots, X^{-1}(\{12\})\},$$

i.e.,

$$\mathcal{F} = 2^A.$$

9.11 Example.

(Continued.) Recall that

$$\mathcal{F} = 2^A.$$

Consider the random variable

$$W_1 = \omega_1$$

for $\omega \in \Omega$. But the preimage

$$W_1^{-1}(1) = \{(1,1), (1,2), \ldots, (1,6)\},$$

is not in $\mathcal{F}$, so W_1 is not a random variable defined on the specified probability space.

9.12 Definition.

If $X : \Omega \to \mathbb{R}$ is a random variable, then the σ-algebra generated by X is

$$\sigma(X) = \sigma\left(\{X^{-1}(B) : B \in \mathcal{B}(\mathbb{R})\}\right).$$

10.0 Introduction.

Convergence of Random Variables.

10.1 Definition.

Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let

$$\{X_n : n \in \mathbb{N}\}$$

be a sequence of rvs on Ω and let X also be an rv on Ω. Then, X_n converges pointwise everywhere to X if for all $\omega \in \Omega$,

$$X_n(\omega) \to X(\omega).$$

That is, X_n converges pointwise everywhere to X if

$$\forall \omega \in \Omega, \forall \varepsilon > 0, \exists N_{\omega,\varepsilon} \in \mathbb{N} \text{ s.t. } |X_n(\omega) - X(\omega)| < \varepsilon, \forall n > N_{\omega,\varepsilon}.$$

10.2 Definition.

Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let

$$\{X_n : n \in \mathbb{N}\}$$

be a sequence of rvs on Ω and let X also be an rv on Ω. Then, X_n converges pointwise with probability one to X if there exists a set A with $P(A) = 0$ such that

$$\forall \omega \in A^c, X_n(\omega) \to X(\omega).$$

10.3 Definition.

Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let

$$\{X_n : n \in \mathbb{N}\}$$

be a sequence of rvs on Ω and let X also be an rv on Ω. Then, X_n converges uniformly everywhere to X if

$$\forall \varepsilon > 0, \exists N_\varepsilon \in \mathbb{N} \text{ s.t. } |X_n(\omega) - X(\omega)| < \varepsilon, \forall n > N_\varepsilon, \forall \omega \in \Omega.$$

10.4 Definition.

Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let

$$\{X_n : n \in \mathbb{N}\}$$

be a sequence of rvs on Ω and let X also be an rv on Ω. Then, X_n converges in probability to X if for all $\delta > 0$, the n, δ-problematic set

$$A_{n,\delta} = \{\omega \in \Omega : |X_n(\omega) - X(\omega)| \geq \delta\}$$

satisfies

$$P(A_{n,\delta}) \to 0 \text{ as } n \to \infty.$$

We denote this type of convergence as

$$X_n \xrightarrow{P} X.$$

10.5 Note.

For $\delta_1 < \delta_2$, we have that

$$A_{n,\delta_1} \supseteq A_{n,\delta_2}.$$

10.6 Remark.

Consistency

$$P\left(|\bar{X}_n - \mu| \geq \varepsilon\right) \to 0 \text{ as } n \to \infty$$

is a statement about convergence in probability to the degenerate random variable $X = \mu$.

10.7 Definition.

Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let

$$\{X_n : n \in \mathbb{N}\}$$

be a sequence of rvs on Ω and let X also be an rv on Ω. Then, X_n converges in L^p to X if

$$\mathrm{E}\left(|X_n - X|^p\right) \to 0 \text{ as } n \to \infty.$$

10.8 Remark.

Consider Markov's Inequality

$$P(|X_n - X| \geq \delta) \leq \frac{\mathrm{E}(|X_n - X|)}{\delta}.$$

Then, for $p > 0$, we have that

$$P(|X_n - X|^p \geq \delta^p) \leq \frac{\mathrm{E}(|X_n - X|^p)}{\delta^p},$$

so if X_n converges in L^p to X, then X_n converges in probability to X.

10.9 Definition.

A cumulative distribution function is a function that satisfies

(1) $F : \mathbb{R} \to [0, 1]$.

(2) F is non-decreasing.

(3) $\lim_{x \to -\infty} F(x) = 0$.

(4) $\lim_{x \to \infty} F(x) = 1$.

(5) F has left limits.

(6) F is right-continuous.

If F is continuous, then F is a continuous cumulative distribution function.

10.10 Definition.

A continuity point of a function $f : \mathbb{R} \to \mathbb{R}$ is a point t such that

$$\forall \varepsilon > 0, \exists \delta > 0 \text{ s.t. if } |x - t| < \delta \text{ then } |f(x) - f(t)| < \varepsilon.$$

10.11 Definition.

A sequence of rvs X_n converges in distribution to an rv X if for all continuity points t of F,

$$\lim_{n \to \infty} F_n(t) = F(t)$$

where F_n is the cdf of X_n and F is the cdf of X. We denote this type of convergence as

$$X_n \xrightarrow{D} X.$$

Note that $X_1, X_2, \ldots$ and X need not be defined on a common probability space.

10.12 Theorem.

Suppose that $g : \mathbb{R} \to \mathbb{R}$ is continuous. Then, if

$$X_n \xrightarrow{D} X,$$

then

$$g(X_n) \xrightarrow{D} g(X).$$

10.13 Lemma.

If a sequence of rvs X_n is defined on a common probability space Ω and

$$X_n \xrightarrow{D} c$$

for some degenerate random variable $c \in \mathbb{R}$, then

$$X_n \xrightarrow{P} c.$$

10.14 Theorem.

(Slutsky's Theorem.) Suppose that

$$X_n \xrightarrow{D} X, \quad Y_n \xrightarrow{D} c$$

for some degenerate random variable $c \in \mathbb{R}$. If X_n, Y_n are defined on a common Ω such that $X+Y$ and XY are well-defined, then

(1) $X_n + Y_n \xrightarrow{D} X + c$.

(2) $X_n Y_n \xrightarrow{D} cX$.

(3) $X_n / Y_n \xrightarrow{D} X/c$ if $c \neq 0$.

10.15 Example.

Let Ω be the unit interval and P be the uniform probability measure. Define the sequence of rvs X_n such that

$$X_n = I_{[0,1/n]}.$$

Define also the degenerate rv $X = 0$.

We claim that X_n does not converge pointwise everywhere, does converge pointwise with probability one, does not converge uniformly, does converge uniformly with probability one, does converge in probability, does converge in L^2, and does converge in distrubtion to X.

Non-Convergence Pointwise Everywhere. Recall that X_n converges pointwise everywhere to X if

$$X_n(\omega) \to X(\omega)$$

for all $\omega \in \Omega$. Consider $\omega = 0$. For all $n \in \mathbb{N}$, $X_n(0) = 1$ but $X(0) = 0$, so X_n does not converge pointwise everywhere to X.

Convergence Pointwise with Probability One. Consider the set $A^c = \Omega \setminus 0$. Consider an arbitary $\omega \in A^c$. Then, for all $\varepsilon > 0$, we can choose

$$N_\varepsilon = \left\lceil \frac{1}{\omega} \right\rceil + 1$$

such that for all $n > N_\varepsilon$,

$$|X_n(\omega) - 0| = 0.$$

Hence, X_n converges pointwise with probability one to X.

Non-Convergence Uniformly. Recall that X_n converges uniformly to X if

$$\forall \varepsilon > 0, \exists N_\varepsilon \in \mathbb{N} \text{ s.t. } |X_n(\omega) - X(\omega)| < \varepsilon, \forall n > N_\varepsilon, \forall \omega \in \Omega.$$

Uniform convergence is a stronger condition than pointwise convergence, so if X_n does not converge pointwise everywhere to X, then X_n does not converge uniformly to X.

Convergence in Probability. Recall that X_n converges in probability to X if for all $\delta > 0$, the set

$$A_{n,\delta} = \{\omega \in \Omega : |X_n(\omega) - X(\omega)| \geq \delta\}$$

satisfies

$$P(A_{n,\delta}) \to 0 \text{ as } n \to \infty.$$

For any $\delta > 0$, the problematic set is the set for which the indicator is one, i.e.,

$$A_{n,\delta} = [0, 1/n)$$

with probability

$$P(A_{n,\delta}) = 1/n \to 0 \text{ as } n \to \infty.$$

Hence, X_n converges in probability to X.

Convergence in L2. Recall that X_n converges in L^2 to X if

$$\mathrm{E}(|X_n - X|^2) \to 0 \text{ as } n \to \infty.$$

We have that

$$\begin{aligned} \mathrm{E}(|X_n - X|^2) &= \mathrm{E}(I_{[0,1/n]}) \\ &= 1/n \to 0 \text{ as } n \to \infty. \end{aligned}$$

Hence, X_n converges in L^2 to X.

Convergence in Distribution. Recall that X_n converges in distribution to X if for all continuity points t of F, the limit

$$\lim_{n\to\infty} F_n(t) = F(t).$$

The cdf of X_n is

$$F_n(t) = \begin{cases} 0 & t < 0 \\ 1 - 1/n & 0 \leq t < 1 \\ 1 & t \geq 1 \end{cases}$$

and the cdf of X is

$$F(t) = \begin{cases} 0 & t < 0 \\ 1 & t \geq 0 \end{cases}.$$

For all points $t \neq 0$, we have that

$$\lim_{n\to\infty} F_n(t) = F(t).$$

But $t = 0$ is not a continuity point of F because

$$\begin{aligned} F(0) &= 1 \\ \lim_{t\to 0^-} F(t) &= 0. \end{aligned}$$

Therefore, X_n converges in distribution to X.

10.16 Example.

Let Ω be the unit interval and P be the uniform probability measure. Define the sequence of rvs X_n such that

$$X_n = 2^n I_{[0,2^{-n}]}.$$

We claim that X_n converges in probability to X but that X_n does not converge in L^2 to X.

Convergence in Probability. For any $\delta > 0$, the problematic set is the set for which the indicator is one, i.e.,

$$A_{n,\delta} = [0, 2^{-n})$$

with probability

$$P(A_{n,\delta}) = 2^{-n} \to 0 \text{ as } n \to \infty.$$

Hence, X_n converges in probability to X.

Non-Convergence in L2. We have that the second absolute moment is

$$\begin{aligned} \mathrm{E}(|X_n - X|^2) &= \mathrm{E}(2^{2n} I_{[0,2^{-n}]}) \\ &= 2^{2n} 2^{-n} \\ &= 2^n. \end{aligned}$$

Therefore, X_n does not converge in L^2 to X.

10.17 Example.

Let Ω be the unit interval and P be the uniform probability measure. For any $\omega \in \Omega$, let $Y_n(\omega) = g_n(\omega)$, defined as follows. First, for any $n \geq 1$, let m be the largest integer such that $2^m \leq n$. Then note that n can be written uniquely as $n = 2^m + j$, where $0 \leq j < 2^m$. For such n, define

$$g_n(\omega) = I_{\left[\frac{j}{2^m}, \frac{j+1}{2^m}\right]}(\omega)$$

where I denotes the indicator function.

We claim that $g_n(\omega)$ converges in probability to zero. But for any value of $\omega \in \Omega$, there exists an infinite sequence of integers n_k where $g_{n_k}(\omega) = 0$ and an infinite sequence of integers n_l where $g_{n_l}(\omega) = 1$, so $g_n(\omega)$ does not converge with probabilty one to zero.

Convergence in Probability. For all $\delta \in (0, 1]$, the set for which $|g_n(\omega) - 0| \geq \delta$ is $\omega \in [j/2^m, (j+1)/2^m]$. We thus have that

$$\begin{aligned} P(A_{n,\delta}) &= P\left(\left[\frac{j}{2^m}, \frac{j+1}{2^m}\right]\right) \\ &= \frac{1}{2^m} \to 0 \text{ as } n \to \infty \end{aligned}$$

where m is defined as the largest integer such that $2^m \leq n$. Observe that the problematic set $A_{n,\delta}$ does not depend on the particular value of δ. So, for all $\delta > 0$, we have that

$$P(A_n, \delta) = \frac{1}{2^m} = 2^{-\lfloor \log_2 n \rfloor}.$$

Suppose that $P(A_{n,\delta})$ converges to 0. Then, for all $\varepsilon > 0$, we can choose

$$N_\varepsilon = \left\lceil \frac{1}{\varepsilon} \right\rceil + 1$$

such that for all $n > N_\varepsilon$,

$$\begin{aligned} |2^{-\lfloor \log_2 n \rfloor} - 0| &\leq 2^{-\log_2 n - 1} \\ &= \frac{1}{2n} \\ &< \frac{1}{2} \frac{1}{1/\lceil \varepsilon \rceil + 1} \\ &< \varepsilon. \end{aligned}$$

Therefore, for all δ, the problematic set $A_{n,\delta}$ converges to probability zero as n. Thus, $Y_n = g_n(\omega)$ converges in probability to zero.

Non-Convergence with Probability One. Observe that any number $\omega \in [0,1]$ can be written in binary form as

$$\omega = 0.\omega_1\omega_2\omega_3\ldots$$

where $\omega_i \in \{0,1\}$. For all m, we truncate ω at the mth digit to get

$$\omega' = 0.\omega_1\omega_2\ldots\omega_m.$$

For all such ω', we can write

$$\omega' = \frac{j}{2^m}$$

for some unique $0 \leq j < 2^m$. Clearly, then,

$$\omega \in [j/2^m, (j+1)/2^m].$$

Thus, for all $m = 1, 2, \ldots$, let $n_l = 2^m + j$, where j is the unique integer described above. Therefore,

$$\begin{aligned} g_{n_l}(\omega) &= I_{\left[\frac{j}{2^m}, \frac{j+1}{2^m}\right]}(\omega) \\ &= 1 \end{aligned}$$

for all $\omega \in \Omega$. Similarly, we can choose $n_k = 2^m + j - 1$ where $m = m_1, m_2, \ldots$ such that m_1 is the smallest $2^k, k \in \mathbb{N}$, such that the special j described above is not 0. Then,

$$\begin{aligned} g_{n_k}(\omega) &= I_{\left[\frac{j-1}{2^m}, \frac{j}{2^m}\right]}(\omega) \\ &= 0 \end{aligned}$$

for all $\omega \in \Omega$. Thus, for any value of $\omega \in \Omega$, there exists an infinite sequence of integers n_k where $g_{n_k}(\omega) = 0$ and an infinite sequence of integers n_l where $g_{n_l}(\omega) = 1$.

For any $\omega \in [0,1]$, there exists an unbounded sequence of integers n_l where $g_{n_l}(\omega) = 1$. Therefore, for any $\omega \in [0,1]$ and any $\varepsilon > 0$, there exists no N_ε such that for all $n > N_\varepsilon$, $|g_n(\omega) - 0| < \varepsilon$. Thus, $g_n(\omega)$ does not converge to zero with probability one.

11.0 Introduction.

Parametric Estimation.

11.1 Remark.

The motivation for parametric estimation is the following problem. Suppose that we have a collection of iid rvs X_n with a common cdf F_θ for some $\theta \in \Theta \subseteq \mathbb{R}^k$. Suppose that we know the parametric family of F_θ but not the exact value of θ. Then, we would like to estimate θ from the observations of X_n.

11.2 Remark.

We may ask two types of questions about $\hat{\theta}$. If we investigate finite sample properties of $\hat{\theta}$, then we are interested in the properties of $\hat{\theta}$ for a fixed n. Examples of these properties are

(1) Distribution of $\hat{\theta}$

(2) $\mathrm{E}(\hat{\theta})$

(3) $\mathrm{Var}(\hat{\theta})$

If we investigate asymptotic properties of $\hat{\theta}$, then we are interested in the properties of $\hat{\theta}$ as $n \to \infty$. Examples of these properties are

(1) Consistency of $\hat{\theta}$ for θ.

(2) Limiting Distribution of $\hat{\theta}$.

11.3 Definition.

For an rv X with pdf f, the support of f is the set

$$\mathrm{supp}(f) = \{x \in \mathbb{R} : f(x) > 0\}.$$

For an rv X with a pmf p, the support of p is the set

$$\mathrm{supp}(p) = \{x \in \mathbb{R} : p(x) > 0\}.$$

11.4 Example.

The support of f_θ can depend on θ. Consider $X_n \sim \mathrm{unif}(0, \theta)$ for $\theta > 0$. Then, the support of f_θ is

$$\mathrm{supp}(f_\theta) = [0, \theta].$$

11.5 Definition.

Suppose that we have a collection of random variables X_n iid with a common cdf F_θ for some $\theta \in \Theta \subseteq \mathbb{R}^k$. We may write the parameter θ as the solution to the system of equations

$$\mu_r(\theta) = \mathrm{E}_\theta(X^r)$$

for $r = 1, 2, \ldots, k$ where $\mathrm{E}(X^r)$ is the rth moment of X. We denote the solution to this system of equations as the inverse function

$$\theta = g(\mu_1, \mu_2, \ldots, \mu_k).$$

Suppose that we do not know the exact value of θ but we observe a collection of iid rvs X_n with a common cdf F_θ. Then, we would like to estimate θ from the observations of X_n. In particular,

$$\hat{\theta}_{MOM} = g(\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_k)$$

where $\hat{\mu}_r$ is the rth sample moment of X_n. We call $\hat{\theta}_{MOM}$ the method of moments (mom) estimator of θ.

11.6 Example.

If $X_n \sim \mathrm{N}(\theta_1, \theta_2)$ iid, then

$$\mu_1 = \theta_1$$
$$\mu_2 = \theta_1^2 + \theta_2,$$

so

$$\hat{\theta}_{1_{MOM}} = \hat{\mu}_1$$
$$\hat{\theta}_{2_{MOM}} = \hat{\mu}_2 - \hat{\mu}_1^2.$$

Note that this mom estimator is simply the sample mean and sample variance, which is expected for a distribution $X \sim \mathrm{N}(\mu, \sigma^2)$. Note that $\hat{\theta}_{1_{MOM}} = \bar{X}$ is unbiased for θ_1 but that $\hat{\theta}_{2_{MOM}} = \hat{\sigma}^2$ is biased for θ_2.

11.7 Example.

Sometimes, the distribution of $\hat{\theta}$ can be determined exactly. For $X_n \sim \mathrm{Exp}(\lambda)$ iid, we have that

$$\hat{\lambda}_{MOM} = \frac{1}{\bar{X}}.$$

But we know that

$$\bar{X} \sim \mathrm{Gamma}(n, n\lambda).$$

11.8 Example.

(Continued.) We may also determine the asymptotic properties of $\hat{\lambda}_{MOM} = 1/\bar{X}$. We have that

$$P\left(\frac{1}{\bar{X}} \leq t\right) = P(\bar{X} \geq \frac{1}{t}).$$

Since $f : t \to 1/t$ is continuous for $t > 0$ and

$$\bar{X} \xrightarrow{P} \frac{1}{\lambda},$$

then

$$\hat{\lambda}_{MOM} \xrightarrow{P} \lambda.$$

11.9 Definition.

For the likelihood function

$$f(x_1, x_2, \ldots, x_n|\theta),$$

the maximum likelihood estimator (mle) of θ given the observations $X_i = x_i$ is the value of θ that maximizes the likelihood function. That is,

$$\hat{\theta}_{MLE} = \arg\max_{\theta \in \Theta} f(x_1, x_2, \ldots, x_n|\theta).$$

11.10 Example.

Suppose that X_i are iid. Then, the arg max of the likelihood function occurs at

$$\begin{aligned}\arg\max_{\theta \in \Theta} f(x_1, x_2, \ldots, x_n|\theta) &= \arg\max_{\theta \in \Theta} \log f(x_1, x_2, \ldots, x_n|\theta) \\ &= \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \log f(x_i|\theta)\end{aligned}$$

since $f : t \to \log t$ is monotone increasing. Under certain conditions, we maximize the likelihood function by setting the gradient of the log-likelihood function equal to zero. That is, setting

$$\nabla \sum_{i=1}^{n} \log f(x_i|\theta) = \left(\frac{\partial}{\partial \theta_j} \sum_{i=1}^{n} \log f(x_i|\theta) \right)$$

for $j = 1, 2, \ldots k$ equal to zero and solving for θ.

11.11 Example.

For $X_i \sim \text{Exp}(\lambda)$ iid for $\lambda > 0$, the likelihood function is

$$f(x_1, x_2, \ldots, x_n|\lambda) = \lambda^n \exp\left(-\lambda \sum_{i=1}^{n} x_i \right).$$

Then, the log-likelihood function is

$$\log f(x_1, x_2, \ldots, x_n|\lambda) = n \log \lambda - \lambda \sum_{i=1}^{n} x_i,$$

which is maximized at

$$0 = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i$$
$$\lambda = n \left(\sum_{i=1}^{n} x_i \right)^{-1}.$$

Therefore,

$$\hat{\lambda}_{MLE} = \frac{1}{\bar{X}}.$$

Note that this agrees with the mom estimator for λ.

11.12 Example.

If $X_n \sim \text{unif}(0, \theta)$ iid, then

$$\mu_1 = \frac{\theta}{2},$$

so

$$\hat{\theta}_{MOM} = 2\hat{\mu}_1.$$

But some observations X_i may be greater than $\hat{\theta}_{MOM}$, which means that those observations are beyond the support of the estimated f_θ.

11.13 Example.

(Continued.) The likelihood function of $X_n \sim \text{unif}(0, \theta)$ iid is

$$f(x_1, x_2, \ldots, x_n|\theta) = \theta^{-n} I_{\{x_i \in [0,\theta] \forall i\}}.$$

Since the support of f_θ depends on θ, then $f(x|\theta)$ is not differentiable in θ, so we cannot set the derivative equal to zero. But we observe that if $\theta \leq \max\{x_i\}$, i.e., if θ is feasible, then

$$f(x_1, x_2, \ldots, x_n|\theta) = \theta^{-n},$$

which is maximized when θ is small. Therefore, the mle of θ is the smallest feasible θ, i.e.,

$$\hat{\theta}_{MLE} = \max\{x_i\}.$$

Note that $\hat{\theta}_{MLE}$ is always feasible unlike $\hat{\theta}_{MOM}$.

11.14 Definition.

The Fisher Information of a random variable X with pdf $f(x|\theta)$ is

$$I(\theta) = \mathrm{E}\left(\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2\right)$$

where the expected value is taken over the support of $f(x|\theta)$. Under sufficient regularity conditions, the Fisher information is equivalent to

$$I(\theta) = -\mathrm{E}\left(\frac{\partial^2}{\partial\theta^2}\log f(X|\theta)\right).$$

The Fisher information $I(\theta)$ of an rv X is a measure of how much information an observation of X contains about its parameter θ.

11.15 Theorem.

Suppose that $X_1, X_2, \ldots, X_n$ are iid and $f(x|\theta)$ satisfies sufficient regularities conditions, i.e.,

(1) $f(x|\theta)$ is sufficiently regular (sufficiently smooth).

(2) supp$f(x|\theta)$ does not depend on θ.

Let $\hat{\theta}_n$ be the mle for $\theta \in \Theta$. Then,

$$\forall\delta > 0, P(|\hat{\theta}_n - \theta| \geq \delta) \to 0 \text{ as } n \to \infty,$$

i.e., $\hat{\theta}_n$ is consistent for θ, and

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{D} \mathrm{N}(0, I(\theta)^{-1})$$

and

$$\mathrm{Var}(\hat{\theta}_n) = \frac{1}{nI(\theta)}.$$

11.16 Theorem.

(Cramer-Rao Bound.) Suppose that $X_1, X_2, \ldots, X_n$ are iid with common pdf $f(x|\theta)$ that satisfies sufficient regularity conditions. Let $T(X_1, X_2, \ldots, X_n)$ be an estimator for θ and let

$$\phi(\theta) = \mathrm{E}(T(X_1, X_2, \ldots, X_n)).$$

Then, the variance of T satisfies the lower bound

$$\mathrm{Var}(T) \geq \frac{(\phi'(\theta))^2}{nI(\theta)}$$

where $I(\theta)$ is the Fisher Information of X with pdf $f(x|\theta)$.

11.17 Definition.

Suppose that $\hat{\theta}_1$ and $\hat{\theta}_2$ are two estimators for θ. The relative efficiency of $\hat{\theta}_1$ with respect to $\hat{\theta}_2$ is

$$\mathrm{RE}(\hat{\theta}_1, \hat{\theta}_2) = \frac{\mathrm{MSE}(\hat{\theta}_2)}{\mathrm{MSE}(\hat{\theta}_1)}.$$

If $\hat{\theta}_1$ and $\hat{\theta}_2$ are unbiased for θ, then

$$\mathrm{RE}(\hat{\theta}_1, \hat{\theta}_2) = \frac{\mathrm{Var}(\hat{\theta}_2)}{\mathrm{Var}(\hat{\theta}_1)}.$$

11.18 Definition.

$T(X_1, X_2, \ldots, X_n)$ is sufficient for θ if the conditional distribution of $X_1, X_2, \ldots, X_n$ given $T(X_1, X_2, \ldots, X_n)$ does not depend on θ.

11.19 Theorem.

(Factorization Theorem.) Suppose that $X_1, X_2, \ldots, X_n$ are iid with common pdf $f(x|\theta)$. T is sufficient for θ if and only if

$$f(x_1, x_2, \ldots, x_n|\theta) = g(T(x_1, x_2, \ldots, x_n), \theta)h(x_1, x_2, \ldots, x_n).$$

11.20 Definition.

T is minimally sufficient for θ if for all sufficient statistics S for θ, there exists a function l such that

$$T = l(S).$$

11.21 Example.

Suppose that $X_1, X_2, \ldots, X_n \sim \text{Bern}(\theta)$ iid. Define $X = (X_1, X_2, \ldots, X_n$ and consider the pdf

$$f(x|\theta)) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}.$$

Define the statistic

$$T = \sum_{i=1}^{n} X_i \sim \text{binom}(n, \theta).$$

Then, the pdf is

$$\begin{aligned} f(x|\theta) &= \binom{n}{T}\theta^T(1-\theta)^{n-T} \\ &= g(T, \theta)h(x). \end{aligned}$$

Therefore, T is sufficient for θ by the factorization theorem.

11.22 Example.

(Continued.) Suppose that we have two values $\theta_1 \neq \theta_2$. The likelihood ratio is

$$\frac{f(x|\theta_1)}{f(x|\theta_2)} = \left(\frac{\theta_1}{\theta_2}\right)^T \left(\frac{1-\theta_1}{1-\theta_2}\right)^{n-T}.$$

Suppose also that S is another sufficient statistic for θ. By the factorization theorem, the likelihood ratio is

$$\frac{f(x|\theta_1)}{f(x|\theta_2)} = \frac{g(S, \theta_1)}{g(S, \theta_2)}.$$

Suppose that $\theta_1 = 1 - \theta_2$. We then have that

$$\begin{aligned} \frac{g(S, \theta_1)}{g(S, \theta_2)} &= \left(\frac{\theta_1}{\theta_2}\right)^T \left(\frac{1-\theta_1}{1-\theta_2}\right)^{n-T} \\ \frac{g(S, \theta_1)}{g(S, \theta_2)} &= \left(\frac{\theta_1}{1-\theta_1}\right)^{2T-n} \\ \log\left(\frac{g(S, \theta_1)}{g(S, \theta_2)}\right) &= 2T\log\left(\frac{\theta_1}{1-\theta_1}\right) - n\log\left(\frac{\theta_1}{1-\theta_1}\right) \\ T &= \log\left(\frac{g(S, \theta_1)}{g(S, \theta_2)}\right)\left(2\log\left(\frac{\theta_1}{1-\theta_1}\right)\right)^{-1} + \frac{n}{2}. \end{aligned}$$

We choose $\theta_1 = 1/3, \theta_2 = 2/3$ to see that T is a function of S, so T is minimally sufficient for θ.

12.0 Introduction.

Bayesian Estimation.

12.1 Definition.

Suppose that we have the observations $X_1, X_2, \ldots, X_n$ iid with common pdf $f_{X|\theta}(x|\theta)$. Suppose that we want to estimate a paramter $\theta \in \Theta$ with a prior distribution $f_\theta(\theta)$. The posterior distribution of θ given the observations $X_1, X_2, \ldots, X_n$ is

$$f_{\theta|X}(x|\theta) = \frac{f_{X|\theta}(x|\theta) f_\theta(\theta)}{f_X(x)}$$

where the denominator is a normalizing constant equal to

$$f_X(x) = \int_\Theta f_{X|\theta}(x|\theta) f_\theta(\theta)\, d\theta.$$

12.2 Lemma.

Suppose that we have a vector of iid observations $X = (X_1, X_2, \ldots, X_n)$ with common pdf $f_{X|\theta}(x|\theta)$. Suppose also that $T(X)$ is a sufficient statistic for θ. Then, the posterior distribution of θ given X only depends on $T(X)$. For a sketch of the proof, observe that

$$\begin{aligned} f_{\theta|X}(\theta|X) &= \frac{f_{X|\theta}(X|\theta) f_\theta(\theta)}{f_X(X)} \\ &= (g(T,\theta)h(X) f_\theta(\theta)) \left(\int_\Theta g(T,\theta)h(X) f_\theta(\theta)\, d\theta \right)^{-1} \\ &= g(T,\theta) f_\theta(\theta) \left(\int_\Theta g(T,\theta) f_\theta(\theta)\, d\theta \right)^{-1} \end{aligned}$$

by the factorization theorem.

12.3 Note.

Suppose that θ has a distribution of $f(\theta)$. Estimates for θ include

(1) The mean of θ.

(2) The median of θ.

(3) The mode of θ.

12.4 Definition.

A loss function is $l : \Theta \times A \to \mathbb{R}^+ \cup 0$ where A is the action space.

12.5 Definition.

Suppose that we have a vector of observations $(X_1, X_2, \ldots, X_n)$. Then, a decision rule is

$$\delta : (\text{supp} f_X)^n \to A.$$

12.6 Definition.

Suppose that we have a vector of iid observations $X = (X_1, X_2, \ldots, X_n)$ with common pdf $f_X(x)$. The risk of a decision rule δ at parameter value θ is

$$\begin{aligned} R(\theta, \delta) &= \mathrm{E}(l(\theta, \delta(X))) \\ &= \int_{\mathbb{R}^n} l(\theta, \delta(x)) f_X(x)\, dx_1 dx_2 \ldots dx_n. \end{aligned}$$

12.7 Definition.

The Bayes risk of a decision rule δ is

$$\begin{aligned} r(\delta) &= \mathrm{E}(\mathrm{E}(l(\theta, \delta(X))|\theta)) \\ &= \int_\Theta \left[\int_{\mathbb{R}^n} l(\theta, \delta(x)) f_{X|\theta}(x|\theta)\, dx_1 dx_2 \ldots dx_n \right] f_\theta(\theta)\, d\theta. \end{aligned}$$

12.8 Definition.

The Bayes estimator of θ is the decision rule δ that minimizes the Bayes risk $r(\delta)$, i.e.,

$$\delta = \arg\min_{\delta \in \Delta} r(\delta).$$

12.9 Definition.

The minimax estimator of θ is the decision rule δ that minimizes the maximum risk, i.e.,

$$\delta = \arg\min_{\delta \in \Delta} \max_{\theta \in \Theta} R(\theta, \delta).$$

12.10 Definition.

Suppose that our prior θ belongs to a family G and that our likelihood $f(x|\theta)$ belongs to a family H. If the posterior θ also belongs to G, then G and H are said to be conjugate families.

12.11 Example.

Suppose that we observe a vector $X = (X_1, X_2, \ldots, X_n)$ of iid Bern(θ) rvs. Suppose that our prior for θ is Beta(α, β) with pdf

$$f_\theta(\theta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha - 1} (1 - \theta)^{\beta - 1}.$$

Observe also that

$$T = \sum_{i=1}^{n} X_i \sim \mathrm{binom}(n, \theta)$$

is a sufficient statistic for θ. The posterior distribution of θ given X is

$$\begin{aligned} f(\theta|X) &= A f_{X|\theta}(X|\theta) f_\theta(\theta) \\ &= A\theta^T (1 - \theta)^{n - T} \theta^{\alpha - 1} (1 - \theta)^{\beta - 1} \\ &= A\theta^{T + \alpha - 1} (1 - \theta)^{n - T + \beta - 1} \sim \mathrm{Beta}(T + \alpha, n - T + \beta) \end{aligned}$$

where A is the appropriate normalizing constant. Therefore, the binomial distribution and the Beta distribution are conjugate families.

13.0 Introduction.

Hypothesis Testing.

13.1 Definition.

Suppose that we are in the Bayesian paradigm. Suppose that we observe a vector of iid data $X = (X_1, X_2, \ldots, X_n)$ with common pdf $f_{X|\theta}(x|\theta)$. Suppose that we want to test the null hypothesis H_0 against the alternative hypothesis H_1 where

$$H_0 : f = f_0 \quad \text{vs.} \quad H_1 : f = f_1.$$

Then, our posterior probabilities are

$$P(H_0|X) = \frac{f(x|H_0)p_0}{f(x|H_0)p_0 + f(x|H_1)p_1}$$
$$P(H_1|X) = \frac{f(x|H_1)p_1}{f(x|H_0)p_0 + f(x|H_1)p_1}$$

where p_0 and p_1, respectively, are the prior probabilities of H_0 and H_1. The posterior odds are

$$\frac{P(H_0|X)}{P(H_1|X)} = \frac{f(x|H_0)p_0}{f(x|H_1)p_1}.$$

We may say that our decision rule rejects H_0 if and only if the posterior odds are less than some threshold $c \in \mathbb{R}^+$ (often one), i.e., if and only if

$$\frac{f(x|H_0)p_0}{f(x|H_1)p_1} < c$$
$$\frac{f_0(x)}{f_1(x)} < c\frac{p_1}{p_0}.$$

13.2 Definition.

Suppose that we are in the frequentist paradigm. Suppose that we observe a vector of iid data $X = (X_1, X_2, \ldots, X_n)$ with common pdf $f_{X|\theta}(x|\theta)$. Suppose that we want to test the null hypothesis H_0 against the alternative hypothesis H_1 where

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

Then, our decision rule δ is

$$\delta : (\text{supp} f_X)^n \to \{0, 1\}$$

where $\delta(X) = 0$ means that we fail to reject H_0 and $\delta(X) = 1$ means that we reject H_0. Note that the frequentist paradigm assigns probabilities to the data X and not to the hypotheses H_0 and H_1.

13.3 Definition.

The likelihood ratio test is the decision rule δ that rejects H_0 if and only if the likelihood ratio

$$\text{LR}(X) = \frac{f(X|H_0)}{f(X|H_1)}$$

is less than some threshold $c \in \mathbb{R}^+$, i.e., if and only if

$$\frac{f_0(x)}{f_1(x)} < c.$$

13.4 Definition.

Type I error is if we reject H_0 when H_0 is true. Type II error is if we fail to reject H_0 when H_0 is false.

13.5 Definition.

The Type I error probability, i.e., the significance level α, for a decision rule δ is

$$\alpha = P(\delta(X) = 1|H_0)$$

if $H_0 \cap H_1 = \emptyset$ and $H_0 \cup H_1 = \mathcal{H}$ where $\mathcal{H}$ is the hypothesis space.

13.6 Definition.

The power of a decision rule δ is

$$\text{Power}(\delta) = P(\delta(X) = 1|H_1)$$

where $H_0 \cap H_1 = \emptyset$ and $H_0 \cup H_1 = \mathcal{H}$ where $\mathcal{H}$ is the hypothesis space.

13.7 Definition.

The p-value for an observed statistic T_0 is the smallest α for which we reject H_0 given T_0 at significance level α with the likelihood ratio test. Under the Neyman-Pearson framework, the p-value is the probability that we observe a test statistic at least as extreme as T_0 given that H_0 is true.

13.8 Definition.

A hypothesis H is simple if it completely specifies the distribution from which the data are taken, i.e., if the the hypothesis space under H, i.e., $\mathcal{H}_0$ is a singleton set. If H is not simple, then it is composite.

13.9 Lemma.

(Neyman-Pearson Lemma.) Suppose that we have a vector of iid data $X = (X_1, X_2, \ldots, X_n)$ with common pdf $f_{X|\theta}(x|\theta)$. Suppose that we want to test the null hypothesis H_0 against the alternative hypothesis H_1 where

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta = \theta_1$$

such that H_0, H_1 are simple and that $\theta_0 \neq \theta_1$ and that $H_0 \cup H_1 = \mathcal{H}$. Then, the likelihood ratio test δ is the most powerful test for H_0 against H_1 at significance level α.

That is, for any test δ^* with significance level $\alpha^* \leq \alpha$, i.e, with

$$P(\delta^*(X) = 1|H_0) \leq \alpha,$$

then

$$P(\delta^*(X) = 1|H_1) \leq P(\delta(X) = 1|H_1).$$

If $\alpha^* < \alpha$, then the above inequality is strict.

13.10 Definition.

Suppose that we have the hypotheses

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta \in \Theta_1.$$

A test δ is uniformly most powerful for the composite hypothesis H_1 if for any $\theta_1 \in \Theta_1$, the test δ is most powerful for the hypotheses

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta = \theta_1.$$

13.11 Definition.

Suppose that we observe a vector of iid data $X = (X_1, X_2, \ldots, X_n)$ with common pdf $f_{X|\theta}(x|\theta)$. Suppose that we want to test the null hypothesis H_0 against the alternative hypothesis H_1 where

$$H_0 : \theta = \Theta_0 \quad \text{vs.} \quad H_1 : \theta \notin \Theta_0.$$

The generalized likelihood ratio (glr) is

$$\Lambda = \max_{\theta \in \Theta_0} f(x|\theta) (\max_{\theta \in \Theta} f(x|\theta))^{-1}.$$

The glr test is to reject H_0 if and only if $\Lambda < c$ where $c \in \mathbb{R}^+$.

13.12 Theorem.

(Wilks's Theorem.) Suppose that we have a vector of iid data $X = (X_1, X_2, \ldots, X_n)$ with common pdf $f_{X|\theta}(x|\theta)$. Suppose that $f(x|\theta)$ satisfies sufficient regularity conditions. Suppose that we have the hypotheses

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

and that

$$\dim \mathcal{H} - \dim \mathcal{H}_0 = k \in \mathbb{N}.$$

Then,

$$-2 \log \Lambda \xrightarrow{D} \chi^2_k \text{ as } n \to \infty$$

where Λ is the glr.

13.13 Lemma.

Suppose that $X_1, X_2, \ldots, X_n \sim \mathrm{N}(\mu, \sigma^2)$ iid. Then, the sample mean $\bar{X}$ is independent of the vector $W = (X_1 - \bar{X}, X_2 - \bar{X}, \ldots, X_n - \bar{X})$.

13.14 Corollary.

(Continued.) If $X_1, X_2, \ldots, X_n \sim \mathrm{N}(\mu, \sigma^2)$ iid, then the sample mean $\bar{X}$ is independent of the sample variance

$$S^2 = \frac{1}{n-1} ||W||_2^2.$$

13.15 Definition.

A rv W has a chi-square distribution with ν df if it has a Gamma distribution with shape parameter $\nu/2$ and rate parameter $1/2$, i.e., $W \sim \chi^2_\nu$ if $W \sim \text{Gamma}(\nu/2, 1/2)$ with pdf

$$f(w) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} w^{\nu/2-1} e^{-w/2}$$

for $w > 0$.

13.16 Lemma.

If $Z \sim \mathrm{N}(0, 1)$, then $Z^2 \sim \chi^2_1$.

13.17 Lemma.

If $W_1 \sim \chi^2_{\nu_1}$ and $W_2 \sim \chi^2_{\nu_2}$, then $W_1 + W_2 \sim \chi^2_{\nu_1+\nu_2}$.

13.18 Corollary.

(Continued.) If $Z_1, Z_2, \ldots, Z_n \sim \mathrm{N}(0,1)$ iid, then

$$\sum_{i=1}^{n} Z_i^2 \sim \chi^2_n.$$

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

13.19 Definition.

A rv T has a Student's t distribution with ν df, i.e., $T \sim \mathrm{t}_\nu$, if T has the pdf

$$f(t) = \frac{\Gamma((\nu+1)/2)}{\sqrt{\nu\pi}\Gamma(\nu/2)} \left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2}.$$

13.20 Lemma.

Suppose that $Z \sim \mathrm{N}(0,1)$ and $V \sim \chi^2_\nu$ are independent. Then,

$$T = \frac{Z}{\sqrt{V/\nu}} \sim \mathrm{t}_\nu.$$

13.21 Lemma.

Suppose that $X_1, X_2, \ldots, X_2 \sim \mathrm{N}(\mu, \sigma^2)$ iid. Then,

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim \mathrm{t}_{n-1}.$$

13.22 Proof.

(Continued.) Observe that

$$\begin{aligned}
\frac{\bar{X} - \mu}{S/\sqrt{n}} &= \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \Big/ \frac{S/\sqrt{n}}{\sigma/\sqrt{n}} \\
&= \frac{Z}{s/\sigma} \\
&= Z \Big/ \sqrt{\frac{s^2(n-1)}{\sigma^2(n-1)}} \\
&= \frac{Z}{\sqrt{V/(n-1)}} \sim \mathrm{t}_{n-1}
\end{aligned}$$

because we know from Corollary (13.13) that $\bar{X}$ is independent of S^2 and therefore Z is independent of V.

13.23 Definition.

An rv X has an F distribution with ν_1 and ν_2 df, i.e., $X \sim \mathrm{F}_{\nu_1,\nu_2}$ if X has the pdf

$$f(x) = \frac{\Gamma((\nu_1+\nu_2)/2)}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)} \left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2} \frac{x^{\nu_1/2-1}}{(1 + \nu_1 x/\nu_2)^{(\nu_1+\nu_2)/2}}.$$

13.24 Lemma.

If $V_1 \sim \chi^2_{\nu_1}$ and $V_2 \sim \chi^2_{\nu_2}$ are independent, then

$$\frac{V_1/\nu_1}{V_2/\nu_2} \sim \mathrm{F}_{\nu_1,\nu_2}.$$

If $Y \sim \mathrm{F}_{\nu_1,\nu_2}$, then

$$1/Y \sim \mathrm{F}_{\nu_2,\nu_1}.$$

If $T \sim \mathrm{t}_\nu$, then

$$T^2 \sim \mathrm{F}_{1,\nu}.$$

because

$$\frac{Z^2/1}{V/\nu} \sim \mathrm{F}_{1,\nu}.$$

13.25 Example.

Suppose that $X_1, X_2, \ldots, X_n \sim \mathrm{N}(\mu, \sigma^2)$ iid and that we have the hypotheses

$$H_o : \mu = \mu_0, \sigma^2 > 0 \quad \text{vs.} \quad H_1 : \mu \neq \mu_0, \sigma^2 > 0.$$

Then, we have that the glr is

$$\Lambda = \frac{f(x|\mu_0, \hat{\sigma}_0^2)}{f(x|\hat{\mu}, \hat{\sigma}^2)}$$

where

$$\hat{\mu}_0 = \mu_0, \quad \hat{\sigma}_0^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \mu_0)^2$$

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

The likelihood ratio simplifies to

$$\Lambda = \left(\frac{\hat{\sigma}}{\hat{\sigma}_0}\right)^n.$$

So we reject H_0 if the ratio $\hat{\sigma}_0^2/\hat{\sigma}^2 > c$. Observe that

$$\sum_{i=1}^n (x_i - \mu_0)^2 / \sum_{i=1}^n (x_i - \bar{x})^2 = 1 + n(\bar{x} - \mu_0)^2 / \sum_{i=1}^n (x_i - \bar{x})^2.$$

So we reject H_0 if the ratio

$$n(\bar{x} - \mu_0)^2 / \sum_{i=1}^n (x_i - \bar{x})^2 > c.$$

But observe that

$$n(n-1)(\bar{X} - \mu_0)^2 / \sum_{i=1}^n (X_i - \bar{X})^2 = \left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}\right)^2 / \frac{S^2(n-1)}{\sigma^2(n-1)} \sim F_{1,n-1}.$$

Define the test statistic W as

$$\begin{aligned} W &= \left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}\right)^2 / \frac{S^2(n-1)}{\sigma^2(n-1)} \\ &= \frac{(\bar{X} - \mu_0)^2}{S^2/n} \sim F_{1,n-1}. \end{aligned}$$

Then, we reject H_0 if $W > c$ where c is the $1-\alpha$ quantile of the $F_{1,n-1}$ distribution. Observe also that

$$T = \sqrt{W} \sim t_{n-1},$$

so we reject H_0 if $|T| > c$ where c is the $1 - \alpha/2$ quantile of the t_{n-1} distribution.

13.26 Example.

Suppose that we have I vectors of n_i data such that

$$Y_{ij} \sim \mathrm{N}(\mu_i, \sigma^2)$$

for $i = 1, 2, \ldots, I$ and $j = 1, 2, \ldots, n_i$. Define

$$\mu = \frac{1}{I}\sum_{i=1}^{I} \mu_i$$

and

$$\alpha_i = \mu_i - \mu$$

such that $\alpha_1 + \alpha_2 + \ldots + \alpha_I = 0$. We have then that the data Y_{ij} can be written as

$$Y_{ij} \sim \mathrm{N}(\mu + \alpha_i, \sigma^2)$$

such that

$$Y_{ij} = \alpha_i + \mu + X_i$$

where $X_{ij} \sim \mathrm{N}(0, \sigma^2)$ iid. Suppose that we don't know the value of σ^2 but we know that it is the same for all I vectors. Then, the likelihood of our data is

$$\begin{aligned} f(y|\mu, \alpha, \sigma^2) &= \prod_{i=1}^{I}\prod_{j=1}^{n_i} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_{ij} - (\alpha_i + \mu))^2\right) \\ &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij} - (\alpha_i + \mu))^2\right) \end{aligned}$$

where $N = n_1 + n_2 + \ldots + n_I$. Notice that our degrees of freedom are $\mu, \sigma^2, \alpha_1, \alpha_2, \ldots, \alpha_{I-1}$ because α_I is uniquely determined by the other α_i.

13.27 Example.

Suppose that we have two vectors $Y_{1\cdot}$ and $Y_{2\cdot}$ such that

$$\begin{aligned} Y_{11}, Y_{12}, \ldots, Y_{1n_1} &\sim \mathrm{N}(\mu_1, \sigma^2) \\ Y_{21}, Y_{22}, \ldots, Y_{2n_2} &\sim \mathrm{N}(\mu_2, \sigma^2) \end{aligned}$$

where all observations are independent. Suppose that σ^2 is known and that we have the hypotheses

$$H_0 : \mu_1 = \mu_2 \quad \text{vs.} \quad H_1 : \mu_1 \neq \mu_2.$$

Note that H_0 is not simple because the set of all possible values of $\mu_1 = \mu_2$ is of dimension one. We have that the glr is

$$\Lambda = \frac{f(y|\mu_1, \mu_2, \sigma^2)}{f(y|\mu, \mu, \sigma^2)}.$$

Under H_0,

$$\hat{\mu} = \frac{n_1\bar{Y}_{1\cdot} + n_2\bar{Y}_{2\cdot}}{n_1 + n_2}$$

and under H_1, our maximizer is

$$(\hat{\mu}_1, \hat{\mu}_2) = (\bar{Y}_{1\cdot}, \bar{Y}_{2\cdot}).$$

We have then that the glr is

$$\Lambda = \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{j=1}^{n_1}(Y_{1j} - \hat{\mu})^2 + \sum_{j=1}^{n_2}(Y_{2j} - \hat{\mu})^2\right)\right) / \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{j=1}^{n_1}(Y_{1j} - \hat{\mu}_1)^2 + \sum_{j=1}^{n_2}(Y_{2j} - \hat{\mu}_2)^2\right)\right).$$

We reject H_0 if $\Lambda < c$, so we reject H_0 if $\log \Lambda < c$, i.e., if

$$\sum_{j=1}^{n_1}(Y_{1j} - \hat{\mu}_1)^2 + \sum_{j=1}^{n_2}(Y_{2j} - \hat{\mu}_2)^2 - \sum_{j=1}^{n_1}(Y_{1j} - \hat{\mu})^2 - \sum_{j=1}^{n_2}(Y_{2j} - \hat{\mu})^2 < c.$$

But observe that

$$\sum_{j=1}^{n_1}(Y_{1j} - \hat{\mu}_1 + \hat{\mu}_1 - \hat{\mu})^2 = \sum_{j=1}^{n_2}(Y_{1j} - \hat{\mu}_1)^2 + n_1(\hat{\mu}_1 - \hat{\mu})^2$$

and similarly for Y_2.. It follows then that we reject H_0 if

$$n_1(\hat{\mu}_1 - \hat{\mu})^2 + n_2(\hat{\mu}_2 - \hat{\mu})^2 > c.$$

13.28 Example.

Suppose that we have two vectors Y_1. and Y_2. such that

$$Y_{11}, Y_{12}, \ldots, Y_{1m} \sim \mathrm{N}(\mu_1, \sigma^2)$$
$$Y_{21}, Y_{22}, \ldots, Y_{2n} \sim \mathrm{N}(\mu_2, \sigma^2)$$

where all observations are independent. Suppose that σ^2 is unknown and that we have the hypotheses

$$H_0 : \mu_1 = \mu_2, \sigma^2 > 0 \quad \text{vs.} \quad H_1 : \mu_1 \neq \mu_2, \sigma^2 > 0.$$

It can be shown that the glr is

$$\Lambda = \left(\frac{\hat{\sigma}}{\hat{\sigma}_0}\right)^{m+n}.$$

where

$$\sigma^2 = \frac{1}{m+n}\left(\sum_{j=1}^{m}(Y_{1j} - \hat{\mu}_1)^2 + \sum_{j=1}^{n}(Y_{2j} - \hat{\mu}_2)^2\right)$$
$$\sigma_0^2 = \frac{1}{m+n}\left(\sum_{j=1}^{m}(Y_{1j} - \hat{\mu}_0)^2 + \sum_{j=1}^{n}(Y_{2j} - \hat{\mu}_0)^2\right)$$

where μ_1, μ_2 are respectivey the sample mean of group one and group two and μ_0 is the grand sample mean. We reject H_0 for small values of Λ, i.e, for large values of

$$\frac{\hat{\sigma}_0^2}{\hat{\sigma}^2} = 1 + \frac{m(\hat{\mu}_1 - \hat{\mu}_0)^2 + n(\hat{\mu}_2 - \hat{\mu}_0)^2}{(m+n-2)s_p^2}$$

where s_p^2 is the pooled sample variance. This is equivalent to if we reject H_0 for large values of

$$\frac{mn(\hat{\mu}_1 - \hat{\mu}_2)^2/(m+n)}{(m+n-2)s_p^2}.$$

Observe that under H_0,

$$\frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{(m+n)\sigma^2/(mn)}} \sim \mathrm{N}(0,1)$$

and

$$\frac{(m+n-2)s_p^2}{\sigma^2} \sim \chi^2_{m+n-2}$$

are independent. We see then that we reject H_0 for large values of the statistic

$$\frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{(1/m + 1/n)s_p^2}} \sim \mathrm{t}(m+n-2).$$

14.0 Introduction.

Analysis of Variance.

14.1 Theorem.

(Cochran's Theorem.) Suppose that $Z_1, \ldots, Z_n \sim \mathrm{N}(0,1)$ iid. Then,

$$\sum_{i=1}^{n} (Z - \bar{Z})^2 \sim \chi^2(n-1).$$

14.2 Corollary.

(Continued.) Suppose that $X_1, \ldots, X_n \sim \mathrm{N}(0, \sigma^2)$. Then, the sample variance is such that

$$\frac{s^2}{\sigma^2} \sim \chi^2(n-1).$$

14.3 Definition.

Suppose that we have I groups of data Y_{ij} where $j = 1, \ldots, J_i$. Define the group sample mean of the ith group as

$$\bar{Y}_{i\cdot} = \frac{1}{J_i} \sum_{j=1}^{J_i} Y_{ij}$$

and the grand sample mean as

$$\begin{aligned} \bar{Y}_{\cdot\cdot} &= \frac{1}{N} \sum_{i=1}^{I} \sum_{j=1}^{J_i} Y_{ij} \\ &= \frac{1}{N} \sum_{i=1}^{I} J_i \bar{Y}_{i\cdot}. \end{aligned}$$

14.4 Definition.

Define the group sample variance of the ith group as

$$S_i^2 = \frac{1}{J_i - 1} \sum_{j=1}^{J_i} (Y_{ij} - \bar{Y}_{i\cdot})^2.$$

Define the pooled sample variance as

$$\begin{aligned} S_p^2 &= \frac{1}{N - I} \sum_{i=1}^{I} \sum_{j=1}^{J_i} (Y_{ij} - \bar{Y}_{i\cdot})^2 \\ &= \frac{1}{N - I} \sum_{i=1}^{I} (J_i - 1) S_i^2. \end{aligned}$$

14.5 Lemma.

If the data Y_{ij} are iid, then the pooled sample variance is such that

$$\frac{(N-I)S_p^2}{\sigma^2} \sim \chi^2(N-I).$$

14.6 Definition.

Define the total sum of squares as

$$TSS = \sum_{i=1}^{I} \sum_{j=1}^{J_i} (Y_{ij} - \bar{Y}_{..})^2$$

and the sum of squares within as

$$\begin{aligned} SSW &= \sum_{i=1}^{I} \sum_{j=1}^{J_i} (Y_{ij} - \bar{Y}_{i.})^2 \\ &= \sum_{i=1}^{I} (J_i - 1) S_i^2 \end{aligned}$$

and the sum of squares between as

$$\begin{aligned} SSB &= \sum_{i=1}^{I} \sum_{j=1}^{J_i} (\bar{Y}_{i.} - \bar{Y}_{..})^2 \\ &= \sum_{i=1}^{I} J_i (\bar{Y}_{i.} - \bar{Y}_{..})^2. \end{aligned}$$

Observe that

$$TSS = SSW + SSB.$$

14.7 Theorem.

The expectation of SSW is

$$\mathrm{E}(SSW) = (N - I)\sigma^2.$$

The expectation of SSB is

$$(I - 1)\sigma^2 + \sum_{i=1}^{I} J_i \alpha_i^2.$$

14.8 Theorem.

If the errors $\varepsilon_{ij} \sim \mathrm{N}(0, \sigma^2)$ iid and furthermore $\alpha_i = 0$ for all i, then

$$\frac{SSW}{\sigma^2} \sim \chi^2(N - I)$$

and

$$\frac{SSB}{\sigma^2} \sim \chi^2(I - 1).$$

Furthermore, SSW and SSB are independent, so the ratio

$$\frac{SSB/(I-1)}{SSW/(N-I)} \sim \mathrm{F}(I - 1, N - I).$$

14.9 Definition.

The one-way ANOVA layout is such that suppose that we have I groups of data such that

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

where $i = 1, \ldots, I$ and $j = 1, \ldots, J_i$. Suppose also that $\varepsilon_{ij} \sim \mathrm{N}(0, \sigma^2)$ iid and $\alpha_1 + \ldots + \alpha_I = 0$. Our unknown variables are therefore $\mu, \sigma^2, \alpha_1, \ldots, \alpha_{I-1}$ (because α_I is uniquely determined by the other α_i). We have then that the likelihood of our data is

$$f(y|\mu, \alpha, \sigma^2) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{I}\sum_{j=1}^{J_i}(y_{ij} - \mu - \alpha_i)^2\right)$$

where $N = J_1 + \ldots + J_I$ and $\alpha = (\alpha_1, \ldots, \alpha_I)$. Suppose that we want to test the hypotheses

$$H_0 : \alpha = 0 \quad \text{vs.} \quad H_1 : \alpha \neq 0.$$

The glr is therefore

$$\Lambda = \left(\frac{\hat{\sigma}}{\hat{\sigma}_0}\right)^N$$

where

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{I}\sum_{j=1}^{J_i}(Y_{ij} - \bar{Y}_{i\cdot})^2$$

$$\hat{\sigma}_0^2 = \frac{1}{N}\sum_{i=1}^{I}\sum_{j=1}^{J_i}(Y_{ij} - \bar{Y}_{\cdot\cdot})^2$$

are respectively the mles under H_1 and H_0. We reject H_0 for small values of Λ, i.e., for large values of

$$\begin{aligned}\hat{\frac{\sigma_0^2}{\sigma^2}} &= \left(\sum_{i=1}^{I}\sum_{j=1}^{J_i}(Y_{ij} - \bar{Y}_{\cdot\cdot})^2\right) / \left(\sum_{i=1}^{I}\sum_{j=1}^{J_i}(Y_{ij} - \bar{Y}_{i\cdot})^2\right) \\ &= \left(\sum_{i=1}^{J}\sum_{j=1}^{J_i}(Y_{ij} - \bar{Y}_{i\cdot})^2 + \sum_{i=1}^{I}\sum_{j=1}^{J_i}(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2\right) / \left(\sum_{i=1}^{I}\sum_{j=1}^{J_i}(Y_{ij} - \bar{Y}_{i\cdot})^2\right) \\ &= \frac{SSB + SSW}{SSW}\end{aligned}$$

or equivalently for large values of

$$\frac{SSB/(I-1)}{SSW/(N-I)} \sim \mathrm{F}(I-1, N-I)$$

where the threshold is chosen such that our test is level α.

14.10 Definition.

Tukey's test is such that suppose that we have the one-way ANOVA layout where the group sample sizes are all equal to J. Consider the distribution of the statistic

$$W = \max_{i \neq j} \frac{|(\bar{Y}_{i\cdot} - \bar{Y}_{j\cdot})|}{\sqrt{S_p^2/J}}.$$

The distribution of W is the studentized range distribution with I and $I(J-1)$ df. The $1-\alpha$-quantiles of the studentized range distribution are denoted by

$$q_\alpha(I, I(J-1)).$$

We reject $H_0 : \mu_1 = \ldots = \mu_I$ at level α if

$$W > q_\alpha(I, I(J-1))\frac{S_p^2}{\sqrt{J}}.$$

The $100(1-\alpha)\%$-confidence interval for any pair $i \neq j$ is

$$(\bar{Y}_{i\cdot} - \bar{Y}_{j\cdot}) \pm q_\alpha(I, I(J-1))\frac{S_p}{\sqrt{J}}.$$

For any pair $i \neq j$, we reject $H_0 : \mu_i = \mu_j$ at level α if

$$|\bar{Y}_{i\cdot} - \bar{Y}_{j\cdot}| > q_\alpha(I, I(J-1))\frac{S_p}{\sqrt{J}}.$$

14.11 Definition.

The two-way ANOVA layout is such that suppose that we have I groups of data such that

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \delta_{ij} + \varepsilon_{ijk}$$

where $i = 1, \ldots, I$ and $j = 1, \ldots J$ and $k = 1, \ldots, K$. Suppose also that $\varepsilon_{ijk} \sim \mathrm{N}(0, \sigma^2)$ iid and that

$$\sum_{i=1}^{I} \alpha_i = 0$$
$$\sum_{j=1}^{J} \beta_j = 0$$
$$\sum_{i=1}^{I} \delta_{ij} = 0$$
$$\sum_{j=1}^{J} \delta_{ij} = 0.$$

The likelihood is then

$$f(y|\mu, \alpha, \beta, \delta, \sigma^2) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(y_{ijk} - \mu - \alpha_i - \beta_j - \delta_{ij})^2\right).$$

It follows that the mles are

$$\begin{aligned}
\hat{\mu} &= \bar{Y}_{\cdots} \\
\hat{\alpha}_i &= \bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdots} \\
\hat{\beta}_j &= \bar{Y}_{\cdot j\cdot} - \bar{Y}_{\cdots} \\
\hat{\delta}_{ij} &= \bar{Y}_{ij\cdot} - \bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdot j\cdot} + \bar{Y}_{\cdots} \\
\hat{\sigma}^2 &= \frac{1}{N}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{ijk} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j - \hat{\delta}_{ij})^2.
\end{aligned}$$

Define the total sum of squares as

$$TSS = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{ijk} - \bar{Y}_{\cdots})^2$$

and note that it has $IJK - 1$ df. Define the sum of squares for the main effect of α as

$$SSA = JK\sum_{i=1}^{I}(\bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdots})^2$$

and the sum of squares for the main effect of β as

$$SSB = IK \sum_{j=1}^{J} (\bar{Y}_{\cdot j \cdot} - \bar{Y}_{\cdots})^2$$

and the sum of squares for the interaction effect as

$$SSAB = K \sum_{i=1}^{I} \sum_{j=1}^{J} (\bar{Y}_{ij \cdot} - \bar{Y}_{i \cdot \cdot} - \bar{Y}_{\cdot j \cdot} + \bar{Y}_{\cdots})^2.$$

and the sum of squares for the error as

$$SSE = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} (Y_{ijk} - \bar{Y}_{ij \cdot})^2.$$

Note that $TSS = SSA + SSB + SSAB + SSE$. Observe also that

$$\begin{aligned}
\frac{SSA}{\sigma^2} &\sim \chi^2(I-1) \\
\frac{SSB}{\sigma^2} &\sim \chi^2(J-1) \\
\frac{SSAB}{\sigma^2} &\sim \chi^2((I-1)(J-1)) \\
\frac{SSE}{\sigma^2} &\sim \chi^2(IJ(K-1)).
\end{aligned}$$

We may then choose the appropriate F-statistic to test our desired hypothesis.

15.0 Introduction.

Simple Linear Regression.

15.1 Definition.

Simple linear regression is such that suppose that we have data $(x_1, y_1), \ldots, (x_n, y_n)$ and that the conditional distribution of Y_i given $X_i = x_i$ is

$$Y(x) = \beta_0 + \beta_1 X + \varepsilon_i$$

where $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$ iid. The least-squares estimators are

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

The solutions are

$$\hat{\beta}_1 = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) / \sum_{i=1}^{n} (x_i - \bar{x})^2$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

15.2 Definition.

Define

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

Observe that

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}.$$

15.3 Lemma.

S_{xx} is such that

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

S_{xy} is such that

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

15.4 Theorem.

The estimator $\hat{\beta}_1$ for β_1 is such that $\hat{\beta}_1 \sim \mathrm{N}(\beta_1, \sigma^2 / S_{xx})$.

15.5 Proof.

(Continued.) Observe that

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

We model the x data as nonrandom, so $\hat{\beta}_1$ is a linear combination of the normal y data and is therefore normal. Furthermore, the expectation of $\hat{\beta}_1$ is

$$\begin{aligned}
\mathrm{E}(\hat{\beta}_1) &= \mathrm{E}\left(\frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) y_i \right) \\
&= \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) \mathrm{E}(y_i) \\
&= \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x})(\beta_0 + \beta_1 x_i) \\
&= \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) \beta_1 x_i \\
&= \beta_1.
\end{aligned}$$

The variance of $\hat{\beta}_1$ is

$$\begin{aligned}
\mathrm{Var}(\hat{\beta}_1) &= \mathrm{Var}\left(\frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) y_i \right) \\
&= \frac{1}{S_{xx}^2} \sum_{i=1}^{n} (x_i - \bar{x})^2 \mathrm{Var}(y_i) \\
&= \frac{\sigma^2}{S_{xx}}.
\end{aligned}$$

15.6 Theorem.

The estimator $\hat{\beta}_0$ for β_0 is such that $\hat{\beta}_0 \sim \mathrm{N}(\beta_0, \sigma^2 S_{xx}/n)$.

15.7 Proof.

(Continued.) Observe that

$$\begin{aligned}
\hat{\beta}_0 &= \frac{1}{n} \sum_{i=1}^{n} y_i - \bar{x} \sum_{i=1}^{n} \frac{(x_i - \bar{x}) y_i}{S_{xx}} \\
&= \sum_{i=1}^{n} \left(\frac{1}{n} + \frac{x_i - \bar{x}}{S_{xx}} \right) y_i.
\end{aligned}$$

We model the x data as nonrandom, so $\hat{\beta}_0$ is a linear combination of the normal y data and is therefore normal. Furthermore, the expectation of $\hat{\beta}_0$ is

$$\begin{aligned}
\mathrm{E}(\hat{\beta}_0) &= \mathrm{E}(\bar{y} - \hat{\beta}_1 \bar{x}) \\
&= \frac{1}{n} \sum_{i=1}^{n} (\beta_0 + \beta_1 x_i) - \hat{\beta}_1 \bar{x} \\
&= \beta_0.
\end{aligned}$$

The variance of $\hat{\beta}_0$ is

$$\begin{aligned}
\mathrm{Var}(\hat{\beta}_0) &= \mathrm{Var}(\bar{y} - \hat{\beta}_1 \bar{x}) \\
&= \mathrm{Var}\left(\frac{1}{n} \sum_{i=1}^{n} (\beta_0 + \beta_1 x_i + \varepsilon_i) - \hat{\beta}_1 \bar{x} \right) \\
&= \frac{1}{n} \sigma^2 + \bar{x}^2 \frac{\sigma^2}{S_{xx}} \\
&= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right) \\
&= \frac{\sigma^2}{n S_{xx}} \left(\sum_{i=1}^{n} (x_i - \bar{x}) x_i + n \bar{x}^2 \right) \\
&= \frac{\sigma^2}{n S_{xx}} \sum_{i=1}^{n} x_i^2 .
\end{aligned}$$

15.8 Theorem.

The covariance between $\hat{\beta}_0$ and $\hat{\beta}_1$ is

$$\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\bar{x} \sigma^2}{S_{xx}} .$$

15.9 Proof.

(Continued.) Observe that

$$\begin{aligned}
\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) &= \mathrm{Cov}\left(\bar{y} - \hat{\beta}_1 \bar{x}, \hat{\beta}_1 \right) \\
&= \mathrm{Cov}(\bar{y}, \hat{\beta}_1) - \bar{x} \mathrm{Var}(\hat{\beta}_1) \\
&= \mathrm{Cov}\left(\frac{1}{n} \sum_{i=1}^{n} y_i, \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) y_i \right) - \bar{x} \frac{\sigma^2}{S_{xx}} \\
&= \frac{1}{n S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) \sigma^2 - \bar{x} \frac{\sigma^2}{S_{xx}} \\
&= -\frac{\bar{x} \sigma^2}{S_{xx}} .
\end{aligned}$$

15.10 Theorem.

The estimator

$$\begin{aligned}
\hat{Y}(x) &= \hat{\beta}_0 + \hat{\beta}_1 x \\
&= \sum_{i=1}^{n} \left(\frac{1}{n} + \frac{(x_i - \bar{x})}{S_{xx}} (x - \bar{x}) \right) y_i
\end{aligned}$$

for $Y(x)$ is such that

$$\hat{Y}(x) \sim \mathrm{N}\left(\beta_0 + \beta_1 x, \sigma^2 \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}} \right) \right) .$$

15.11 Proof.

(Continued.) Observe that $Y(x)$ is a linear combination of the normal y data and is therefore normal. The expectation of $\hat{Y}(x)$ is

$$\begin{aligned}\mathrm{E}(\hat{Y}(x)) &= \mathrm{E}(\hat{\beta}_0 + \hat{\beta}_1 x) \\ &= \beta_0 + \beta_1 x.\end{aligned}$$

The variance of $\hat{Y}(x)$ is

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

15.12 Definition.

Define the mean squared error as

$$MSE = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{Y}(x_i))^2.$$

15.13 Lemma.

The estimator MSE for σ^2 is such that

$$\frac{(n-2)MSE}{\sigma^2} \sim \chi^2(n-2).$$

15.14 Definition.

The question of model utility is to test the hypotheses

$$H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_1 : \beta_1 \neq 0.$$

Under H_0, we have that $\hat{\beta}_1 \sim \mathrm{N}(0, \sigma^2/S_{xx})$. We therefore reject H_0 for large absolute values of

$$\frac{\hat{\beta}_1}{\sqrt{MSE/S_{xx}}} \sim \mathrm{t}(n-2)$$

where the threshold is chosen such that our test is level α.

15.15 Lemma.

The mean of the predicted values $\bar{\hat{y}}$ is equal to the mean of the observed values $\bar{y}$.

15.16 Definition.

Define the total sum of squares

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

and the sum of squares regression

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

and the sum of squares error

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

15.17 Lemma.

TSS, SSR, and SSE are such that

$$\begin{aligned} TSS &= S_{yy} \\ SSR &= \frac{S_{xy}^2}{S_{xx}} \\ SSE &= S_{yy} - \frac{S_{xy}^2}{S_{xx}}. \end{aligned}$$

15.18 Proof.

(Continued.) $TSS = S_{yy}$ is trivial. We may write

$$\begin{aligned} SSR &= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \\ &= \sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y})^2 \\ &= \sum_{i=1}^{n} (\bar{y} + \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i - \bar{y})^2 \\ &= \sum_{i=1}^{n} \hat{\beta}_1 (x_i - \bar{x})^2 \\ &= \hat{\beta}_1^2 S_{xx} \\ &= \frac{S_{xy}^2}{S_{xx}} \end{aligned}$$

and

$$\begin{aligned}
SSE &= \sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \\
&= \sum_{i=1}^{n}(\hat{\beta}_0 + \hat{\beta}_1 x_i - y_i)^2 \\
&= \sum_{i=1}^{n}(\bar{y} + \hat{\beta}_1\bar{x} + \hat{\beta}_1 x_i - y_i)^2 \\
&= \sum_{i=1}^{n}(y_i - \bar{y} - \hat{\beta}_1(x_i - \bar{x}))^2 \\
&= \sum_{i=1}^{n}((y_i - \bar{y})^2 - 2\beta_1(x_i - \bar{x})(y_i - \bar{y}) + \beta_1^2(x_i - \bar{x})^2) \\
&= S_{yy} - 2\frac{S_{xy}}{S_{xx}}S_{xy} + \frac{S_{xy}^2}{S_{xx}^2}S_{xx} \\
&= S_{yy} - \frac{S_{xy}^2}{S_{xx}}.
\end{aligned}$$

15.19 Theorem.

$TSS = SSR + SSE$.

15.20 Proof.

(Continued.) It is clear from the previous lemma that $TSS = SSR + SSE$. Alternatively, we may first observe that

$$\begin{aligned}
\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) &= \sum_{i=1}^{n}(y_i - \hat{y}_i)\hat{y}_i - \sum_{i=1}^{n}(y_i - \hat{y}_i)\bar{y} \\
&= \sum_{i=1}^{n}(y_i - \hat{y}_i)\hat{y}_i \\
&= \sum_{i=1}^{n}(y_i - \bar{y} - \hat{\beta}_1(x_i - \bar{x}))(\hat{\beta}_0 + \hat{\beta}_1 x_i) \\
&= \sum_{i=1}^{n}(y_i - \bar{y})(\hat{\beta}_0 + \hat{\beta}_1 x_i) - \hat{\beta}_1\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \\
&= \sum_{i=1}^{n}(y_i - \bar{y})\hat{\beta}_1 x_i - \frac{S_{xy}^2}{S_{xx}} \\
&= 0.
\end{aligned}$$

We see then that

$$\begin{aligned}
TSS &= \sum_{i=1}^{n}(y_i - \bar{y})^2 \\
&= \sum_{i=1}^{n}(y_i - \hat{y})_i^2 + 2\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \\
&= SSE + SSR.
\end{aligned}$$

15.21 Definition.

ANOVA for simple linear regression is such that suppose that we want to test the hypotheses

$$H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_1 : \beta_1 \neq 0.$$

We then reject H_0 for large values of

$$\frac{SSR}{SSE/(n-2)} \sim \mathrm{F}(1, n-2)$$

where the threshold is chosen such that our test is level α.

15.22 Example.

The distribution of $\hat{Y}(x) - E(Y(x))$ is such that

$$\frac{\hat{Y}(x) - \beta_0 - \beta_1 x}{\sqrt{\sigma^2(1/n + (x-\bar{x})^2/S_{xx})}} \sim \mathrm{N}(0,1).$$

If σ^2 is not known, then we may use the estimator MSE to obtain the distribution

$$\frac{\hat{Y}(x) - \beta_0 - \beta_1 x}{\sqrt{MSE(1/n + (x-\bar{x})^2/S_{xx})}} \sim \mathrm{t}(n-2).$$

It follows that the $100(1-\alpha)\%$ confidence interval for $\mathrm{E}(Y(x))$ is

$$\hat{Y}(x) \pm \mathrm{t}_{\alpha/2}(n-2)\sqrt{MSE(1/n + (x-\bar{x})^2/S_{xx})}.$$

15.23 Example.

The distribution of $\hat{Y}(x) - Y(x)$ is such that

$$\frac{\hat{Y}(x) - Y(x)}{\sqrt{\sigma^2(1 + 1/n + (x-\bar{x})^2/S_{xx})}} \sim \mathrm{N}(0,1)$$

because $Y(x)$ itself is a rv with variance σ^2 that is furthermore independent of $\hat{Y}(x)$ (because $\hat{Y}(x)$ is a function of the data $(x_1, y_1), \ldots, (x_n y_n)$ which is independent of the new data (x, y)). If σ^2 is not known, then we may use the estimator MSE to obtain the distribution

$$\frac{\hat{Y}(x) - Y(x)}{\sqrt{MSE(1 + 1/n + (x-\bar{x})^2/S_{xx})}} \sim \mathrm{t}(n-2).$$

It follows that the $100(1-\alpha)\%$ prediction interval for $Y(x)$ is

$$\hat{Y}(x) \pm \mathrm{t}_{\alpha/2}(n-2)\sqrt{MSE(1 + 1/n + (x-\bar{x})^2/S_{xx})}.$$

15.24 Definition.

Define the coefficient of determination r^2 as

$$\begin{aligned} r^2 &= \frac{SSR}{TSS} \\ &= \frac{S_{xy}^2}{S_{xx} S_{yy}}. \end{aligned}$$