

Probability and Statistics

Imperial College London
(based on Din-Houn Lau's material)

Contents

1	Introduction	1
1.1	Introduction to Uncertainty	1
1.2	Introduction to Statistics	1
1.2.1	Population vs. Sample	1
1.3	Probability AND Statistics	2
1.4	Statistical Modelling	3
2	Mathematical Methods	4
2.1	Notation	4
2.2	Log and Exponential	4
2.3	Arithmetic and Geometric Progressions	5
2.4	Calculus	5
2.5	Function images and inverses	6
2.6	Interpolation	7
3	Numerical Summaries	8
3.1	Summary Statistics	8
3.1.1	Measures of Location	8
3.1.2	Measures of Dispersion	11
3.1.3	Skewness	12
3.1.4	Covariance and Correlation	13
3.2	Related Graphical Displays	13
3.2.1	Box-and-Whisker Plots	13
3.2.2	Empirical CDF	14
4	Elementary Set Theory	15
4.1	Sets, subsets and complements	15
4.1.1	Sets and notation	15
4.1.2	Subsets, Complements and Singletons	15
4.2	Set operations	15
4.2.1	Unions and Intersections	15
4.2.2	Cartesian Products	17
4.3	Cardinality	18
5	Probability	19
5.1	Sample Spaces and Events	19
5.1.1	Sample Spaces	19
5.1.2	Events	19
5.1.3	Combinations of Events	20
5.2	The σ -algebra	20
5.3	Probability Measure	21
5.3.1	Properties of $P(\cdot)$: The Axioms of Probability	22

5.3.2	Interpretations of Probability	22
5.3.3	Independent Events	24
5.4	More Examples	25
5.4.1	Joint events	26
5.5	Conditional Probability	27
5.5.1	Conditional Independence	29
5.5.2	Bayes' Theorem	29
5.7.1	More Examples	31
6	Discrete Random Variables	35
6.1	Random Variables	35
6.1.1	Cumulative Distribution Function	38
6.2	Discrete Random Variables	39
6.2.1	Properties of Mass Function p_X	39
6.2.2	Discrete Cumulative Distribution Function	39
6.2.3	Connection between F_X and p_X	39
6.2.4	Properties of Discrete CDF F_X	40
6.3	Mean and Variance	42
6.3.1	Expectation	42
6.4.1	Sums of Random Variables	45
6.5	Some Important Discrete Random Variables	45
6.5.1	Bernoulli Distribution	45
6.5.2	Binomial Distribution	46
6.5.3	Geometric Distribution	48
6.5.4	Poisson Distribution	50
6.5.5	Discrete Uniform Distribution	52
7	Continuous Random Variables	53
7.0.1	Continuous Cumulative Distribution Function	53
7.0.2	Properties of Continuous F_X and f_X	53
7.0.3	Transformations	57
7.1	Mean, Variance and Quantiles	58
7.1.1	Expectation	58
7.1.2	Variance	58
7.1.3	Quantiles	59
7.2	Some Important Continuous Random Variables	60
7.2.1	Continuous Uniform Distribution	60
7.2.2	Exponential Distribution	61
7.2.3	Normal Distribution	62
8	Jointly Distributed Random Variables	68
8.0.1	Joint Cumulative Distribution Function	68
8.0.2	Properties of Joint CDF F_{XY}	69

8.0.3	Joint Probability Mass Functions	69
8.0.4	Joint Probability Density Functions	70
8.1	Independence and Expectation	70
8.1.1	Independence	70
8.1.2	Expectation	71
8.1.3	Conditional Expectation	72
8.2	Examples	73
9	Estimation	74
9.1	Estimators	75
9.1.1	Point Estimates	75
9.1.2	Bias, Efficiency and Consistency	76
9.1.3	Maximum Likelihood Estimation	78
9.2	Confidence Intervals	81
9.2.1	Normal Distribution with Known Variance	82
9.2.2	Normal Distribution with Unknown Variance	83
10	Hypothesis Testing	85
10.0.1	Error Rates and Power of a Test	86
10.1	Testing for a population mean	87
10.1.1	Normal Distribution with Known Variance	87
10.1.2	Normal Distribution with Unknown Variance	89
10.2	Testing for differences in population means	90
10.2.1	Two Sample Problems	90
10.2.2	Normal Distributions with Known Variances	91
10.2.3	Normal Distributions with Unknown Variances	91
10.3	Goodness of Fit	94
10.3.1	Count Data and Chi-Square Tests	94
10.3.2	Proportions	95
10.3.3	Model Checking	96
10.3.4	Independence	96

Course Information

Course Material

All the material for the course will be made available on CATE. Lecture notes will be revealed in batches over the term. Also, Piazza will be available for this course.

Lectures

The lecture notes will be made available ahead of the lectures — you will be expected to print out the notes. During the lecture, I will go through the notes and work through examples and fill in the gaps. You are encouraged to ask questions during the lecture. Lectures should be available on Panopto – please do not rely on these recordings, as they may fail.

Chapter 1. Introduction

1.1 Introduction to Uncertainty

This course is about uncertainty, measuring and quantifying uncertainty. Loosely speaking, by uncertainty we mean the condition when results, outcomes, the nearest and remote future are not completely determined; their development depends on a number of factors and just on a pure chance.

Simple examples of uncertainty appear when you bet on the outcome of a football match, turn a wheel of fortune, or toss a coin to make a choice.

Uncertainty appears in virtually all areas of Computer Science and Software Engineering. Installation of software requires uncertain time and often uncertain disk space. A newly released software contains an uncertain number of defects. When a computer program is executed, the amount of required memory may be uncertain. When a job is sent to a printer, it takes uncertain time to print, and there is always a different number of jobs in a queue ahead of it. Electronic components fail at uncertain times, and the order of their failures cannot be predicted exactly. Viruses attack a system at unpredictable times and affect an unpredictable number of files and directories. Uncertainty surrounds us in everyday life, at home, at work, in business, and in leisure.

This course is about measuring and dealing with uncertainty and randomness. It teaches you that probability is a language used to describe and quantify uncertainty. But what about Statistics?

1.2 Introduction to Statistics

Definition of "Statistics"

- **Statistics** is the science and practice of developing human knowledge through the use of empirical data.
- **Statistical theory** is a branch of mathematics using probability theory to model randomness and uncertainty in data.
- **Statistical inference** is inference made from the sample data to the defined population using inductive methods and statistical theory.
- A **statistic** is a numerical summary of data.

1.2.1 Population vs. Sample

The previous definitions suggested an important distinction between a sample and a population.

Loosely, we can think of a population as being a large, perhaps infinite, collection of individuals or objects or quantities in which we are interested. For reasons of generality we would wish to make inferences about the entire population.

Example Suppose a new treatment for headaches was being developed by a pharmaceutical company. The target population for this new drug is potentially everybody in the world. To truly, fully understand the efficacy of the new treatment with respect to this population, we would have to administer treatments to every living individual, just after they get a headache, and measure their response. ■

Often it will be impractical or impossible to exhaustively observe every member of a population. So instead we observe what we hope is a representative sample from the population.

To best ensure the sample is representative and not biased in some way, where possible we draw the sample at random from the population.

Example Returning to the drug discovery example, we perform a clinical trial testing efficacy on a small subset of the population, randomising allocation of the new drug or a control treatment. ■

Statistical methods are then used to relate the measurements of the sample to the characteristics of the entire population.

1.3 Probability AND Statistics

A typical Probability problem sounds like this:

Example A folder contains 50 executable files. When a computer virus attacks a system, each file is affected with probability 0.2. Compute the probability that during a virus attack, more than 15 files get affected. ■

Notice that the situation is rather clearly described, in terms of the total number of files and the chance of affecting each file. The only uncertain quantity is the number of affected files, which cannot be predicted for sure.

A typical Statistics problem sounds like this:

Example A folder contains 50 executable files. When a computer virus attacks a system, each file is affected with the same probability p . It has been observed that during a virus attack, 15 files got affected. Estimate p . Is there a strong indication that p is greater than 0.2? ■

This is a practical situation. A user only knows the objectively observed data: the number of files in the folder and the number of files that got affected. Based on that, she needs to estimate p , the proportion of all the files, including the ones in her system and any similar systems. One may provide a point estimator of p , a real number, or may choose to construct a confidence interval of “most probable” values of p .

1.4 Statistical Modelling

Modern statistical methods are largely driven by the notion of a **model**. In contrast with Machine Learning, which focuses on algorithms.

A model is a postulated structure, or an approximation to a structure, which could have led to the data. (Berthold & Hand, 2002)

- Commonly models are **parametric**.
- \Rightarrow Problem of learning about the underlying population is reduced to learning about a finite set of parameters.

Besides machine learning techniques, computing advances have enabled the fitting of complex parametric/**non-parametric** models (e.g. Bayesian methods, Monte Carlo simulation, ...).

In this course, we will consider very simple parametric statistical models to represent our populations of interest.

- Statistical inference will thus mean estimating model parameters using our observed sample.
- Likelihood methods will be our main tool for this task. We will learn to calculate the likelihood of a particular parameter solution given our observed sample.

Chapter 2. Mathematical Methods

Before discussing probability and statistics, we first review basic mathematical methods that we shall use in this course.

2.1 Notation

The following conventions and set notation will be used throughout the course:

Notation	Set	Description
\mathbb{R}	$(-\infty, \infty)$	The real numbers
\mathbb{R}^+	$(0, \infty)$	The positive real numbers
\mathbb{Z}	$\{\dots, -2, -1, 0, 1, 2, \dots\}$	The integers
\mathbb{Z}^+	$\{1, 2, 3, \dots\}$	The positive integers
\mathbb{N}	$\{0, 1, 2, 3, \dots\}$	The natural numbers

2.2 Log and Exponential

Where there is any ambiguity, by “log” we will mean the natural logarithm, sometimes written “ln” elsewhere. For any other base b (e.g. $b = 10$), we will write \log_b . So

$$\log \equiv \log_e \equiv \ln.$$

Rules:

- $\log(xy) = \log(x) + \log(y) \implies \log\left(\prod_i x_i\right) = \sum_i \log(x_i)$
- $\log(x^y) = y \log(x)$
- $\log(e^x) = x$
- $\lim_{x \rightarrow 0} \log(x) = -\infty$

For exponential, we will use the notations “ e ” or “exp” interchangeably. So

$$e^x \equiv \exp(x).$$

Rules:

- $\exp(x + y) = \exp(x) \exp(y) \implies \exp\left(\sum_i x_i\right) = \prod_i \exp(x_i)$
- $\exp(x)^y = \exp(xy)$
- $\exp\{\log(x)\} = x$
- $\exp(0) = 1$

2.3 Arithmetic and Geometric Progressions

Consider the infinite sequence of numbers

$$a, a + d, a + 2d, a + 3d, \dots$$

The first term in this sequence is a , and then each subsequent term is equal to the previous term plus d , the common difference. Any such sequence is known as an arithmetic progression.

Formulae:

- n^{th} term $= a + (n - 1)d$
- Sum of first n terms, $S_n = \frac{n}{2}\{2a + (n - 1)d\}$
- (Infinite sum, $S_\infty = \pm\infty$, unless $a = d = 0$)

Consider the infinite sequence of numbers

$$a, ar, ar^2, ar^3, \dots$$

The first term in this sequence is a , and then each subsequent term is equal to the previous term multiplied by r , the common ratio. Any such sequence is known as a geometric progression.

Formulae:

- n^{th} term $= ar^{n-1}$
- Sum of first n terms, $S_n = \frac{a(1 - r^n)}{1 - r}$, if $r \neq 1$
- Infinite sum, $S_\infty = \frac{a}{1 - r}$, if $|r| < 1$; (S_∞ diverges otherwise for $a \neq 0$)

2.4 Calculus

Let f, g be functions of a variable x . Then the derivative of f with respect to x , denoted as $\frac{df}{dx}$ or $f'(x)$, is

$$\frac{df}{dx} \equiv f'(x) \equiv \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h}.$$

Formulae:

- Chain Rule: $\frac{d}{dx}f(g(x)) = f'(g(x))g'(x)$
- Product Rule: $\frac{d}{dx}f(x)g(x) = f'(x)g(x) + f(x)g'(x)$
- Quotient Rule: for $g(x) \neq 0$, $\frac{d}{dx} \frac{f(x)}{g(x)} = \frac{f'(x)g(x) - f(x)g'(x)}{\{g(x)\}^2}$

Fundamental Theorem of Calculus:

$$\frac{d}{dx} \int_{u=a}^x f(u)du = f(x),$$

so differentiation and integration are inverse operations.

Formulae:

- Change of variable: if $y = g(x)$, $\int_a^b f(x)dx = \int_{g(a)}^{g(b)} f(g^{-1}(y))g^{-1'}(y)dy$
- By parts: $\int_a^b f(x)g'(x)dx = [f(x)g(x)]_a^b - \int_a^b f'(x)g(x)dx$

Both differentiation and integration are additive. That is, for functions f, g ,

$$\begin{aligned}\frac{d}{dx}\{f(x) + g(x)\} &= \frac{df(x)}{dx} + \frac{dg(x)}{dx}, \\ \int \{f(x) + g(x)\}dx &= \int f(x)dx + \int g(x)dx.\end{aligned}$$

And for any constant $c \in \mathbb{R}$,

$$\begin{aligned}\frac{d}{dx}\{c \cdot f(x)\} &= c \cdot \frac{df(x)}{dx}, \\ \int \{c \cdot f(x)\}dx &= c \cdot \int f(x)dx.\end{aligned}$$

2.5 Function images and inverses

Suppose f is a function $f : X \rightarrow Y$. For $A \subseteq X$, the image of A under f , written $f(A)$ is the subset of Y given by

$$f(A) = \{y \in Y | f(x) = y \text{ for some } x \in A\}.$$

The image of X under f can be referred to simply as the *image of f* . Recall the inverse of f , should it exist, is denoted f^{-1} and has the property

$$f^{-1}(f(x)) = x$$

for any value $x \in X$. The inverse image of f could therefore be defined as the image of Y under f^{-1} .

More generally, for $B \subseteq Y$, the inverse image of B under (possibly non-invertible) f is given by

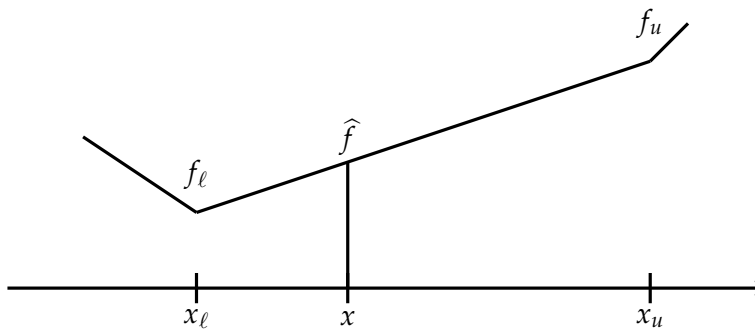
$$f^{-1}(B) = \{x \in X | f(x) \in B\}.$$

2.6 Interpolation

Consider a function $f : \mathbb{R} \rightarrow \mathbb{R}$ of unknown form, where all that is known about f is the value it takes at each of a pre-determined discrete set of points $\mathcal{X} = \{a = x_1 < x_2 < \dots < x_k = b\}$; for each of these values x , denote the corresponding function value $f_x = f(x)$.

Interpolation is the task of finding an approximate value of f for a general point x in the interval $[a, b]$, say $\hat{f}(x)$. (Extrapolation would be the task of finding an approximate value of f for x outside the interval $[a, b]$.)

The most commonly used approximation is linear interpolation, which assumes the underlying function f can be considered approximately piecewise linear between the set of known points.



Let x_ℓ, x_u be nearest pair of points in \mathcal{X} on either side of x . Then $f(x)$ is linearly approximated by

$$\hat{f}(x) = f_{x_\ell} + (x - x_\ell) \frac{(f_{x_u} - f_{x_\ell})}{(x_u - x_\ell)}.$$

Chapter 3. Numerical Summaries

Once a sample of data has been drawn from the population of interest, the first task of the statistical analyst might be to calculate various numerical summaries of these data.

This procedure serves two purposes:

- The first is exploratory. Calculating statistics which characterise general properties of the sample, such as location, dispersion, or symmetry, helps us to understand the data we have gathered. This aim can be greatly aided by the use of graphical displays representing the data.
- The second, as we shall see later in Chapters 9 and 10, is that these summaries will commonly provide the means for relating the sample we have learnt about to the wider population in which we are truly interested.

In this chapter we introduce some common numerical summaries used in statistics.

3.1 Summary Statistics

3.1.1 Measures of Location

The **arithmetic mean** (or just mean for short) of a sample of real values (x_1, \dots, x_n) is the sum of the values divided by their number. That is,

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

This is often colloquially referred to as the average.

Example The mean of $(8, 3, 2, 12, 5)$ is

$$\frac{8 + 3 + 2 + 12 + 5}{5} = \frac{30}{5} = 6.$$

■

For a sample of real values (x_1, \dots, x_n) , define the i^{th} **order statistic** $x_{(i)}$ to be the i^{th} smallest value of the sample.

So

- $x_{(1)} \equiv \min(x_1, \dots, x_n)$ is the smallest value;
- $x_{(2)}$ is the next smallest, and so on, up to
- $x_{(n)} \equiv \max(x_1, \dots, x_n)$ being the largest value.

Example For the sample $(8, 3, 2, 12, 5)$ we have

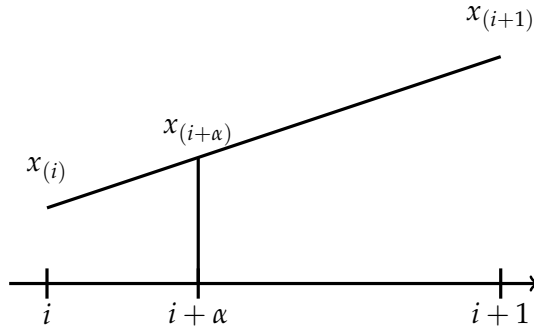
$$x_{(1)} = 2, \quad x_{(2)} = 3, \quad x_{(3)} = 5, \quad x_{(4)} = 8, \quad x_{(5)} = 12.$$

■

Furthermore, in an abuse of notation it will be useful to define $x_{(i+\alpha)}$ for integer $1 \leq i < n$ and non-integer $\alpha \in (0, 1)$ as the linear interpolant

$$x_{(i+\alpha)} = (1 - \alpha) x_{(i)} + \alpha x_{(i+1)},$$

where the order statistics $x_{(i)}$ are defined as before.



Example

$$x_{(4.2)} = 0.8 \times x_{(4)} + 0.2 \times x_{(5)}.$$

■

The **median** of a sample of real values (x_1, \dots, x_n) is the middle value of the order statistics. That is, using our extended notation,

$$\text{median} = x_{(\{n+1\}/2)} = \begin{cases} x_{(\{n+1\}/2)} & \text{if } n \text{ is odd} \\ \frac{x_{(n/2)} + x_{(n/2+1)}}{2} & \text{if } n \text{ is even} \end{cases}$$

Example

- The median of $(7, 2, 4, 12, 5)$ is 5.
- The median of $(7, 2, 4, 12, 5, 15)$ is 6.

■

The mean is sensitive to outlying points, whilst the median is not.

Example

$(1, 2, 3, 4, 5)$ has median = mean = 3

$(1, 2, 3, 4, 40)$ has median = 3, but now mean = 10

■

The arithmetic mean is the most commonly used *location* statistic, followed by the median.

The **mode** of a sample of real values (x_1, \dots, x_n) is the value of the x_i which occurs most frequently in the sample.

Example The mode of $(3, 5, 7, 2, 10, 14, 12, 2, 5, 2)$ is 2. ■

Note Some data sets are *multimodal*.

Two other useful measures of location (other *averages*) are the geometric and harmonic mean.

For positive data, the **geometric mean** is given by

$$x_G = \sqrt[n]{\prod_{i=1}^n x_i}.$$

Note It is easy to show that $x_G = \exp \left\{ \frac{1}{n} \sum_{i=1}^n \log x_i \right\}$, the exponential of the arithmetic mean of the logs of the data. This implies that geometric mean is less severely affected by exceptionally large values.

The **harmonic mean** is given by

$$x_H = \left\{ \frac{1}{n} \sum_{i=1}^n \frac{1}{x_i} \right\}^{-1}.$$

which is most useful when averaging rates.

Note For positive data (x_1, \dots, x_n) ,

$$\text{Arithmetic mean} \geq \text{geometric mean} \geq \text{harmonic mean}.$$

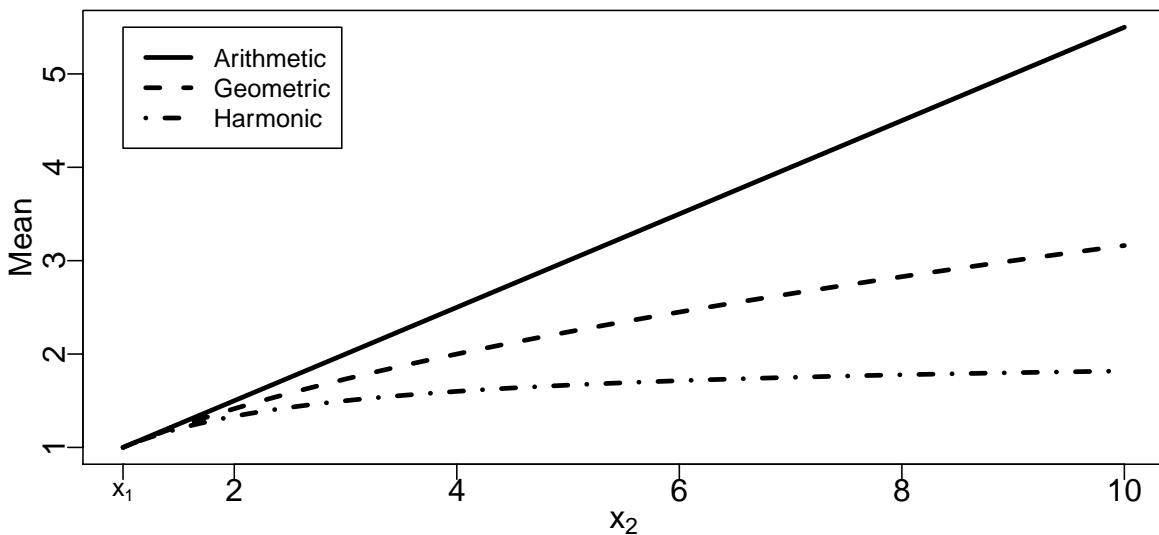


Figure 3.1: Arithmetic, geometric and harmonic means for two data points (x_1, x_2) , where $x_1 = 1$.

3.1.2 Measures of Dispersion

The **range** of a sample of real values (x_1, \dots, x_n) is the difference between the largest and the smallest values. That is

$$\text{range} = x_{(n)} - x_{(1)}$$

Example The range of $(7, 1, 4, 15, 5)$ is $15 - 1 = 14$. ■

Consider again the order statistics of a sample, $(x_{(1)}, \dots, x_{(n)})$.

We defined the *median* so that it lay approximately $\frac{1}{2}$ of the way through the ordered sample — not necessarily exactly or uniquely since there may be tied values or n even.

Similarly, we can define the **first** and **third quartiles** respectively as being values $\frac{1}{4}$ and $\frac{3}{4}$ of the way through the ordered sample:

$$\text{first quartile} = x_{(\{n+1\}/4)}$$

$$\text{third quartile} = x_{(3\{n+1\}/4)}$$

and thus we define the **interquartile range** as the range of the data lying between the first and third quartiles,

$$\text{interquartile range} = \text{third quartile} - \text{first quartile}$$

$$= x_{(3\{n+1\}/4)} - x_{(\{n+1\}/4)}$$

The five point summary of a set of data lists, in order:

- The minimum value in the sample
- The lower quartile
- The sample median
- The upper quartile
- The maximum value

The most widely used measure of dispersion is based on the squared differences between the data points and their mean, $(x_i - \bar{x})^2$. The average (the mean) of these squared differences is the **mean square** or **sample variance**

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

Equivalently, it is often more convenient to rewrite this formula as

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

That is, the mean of the squares minus the square of the mean.

The square root of the variance is the **root mean square** or **sample standard deviation**

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

Unlike the variance, the standard deviation is in the same units as the x_i .

We can see analogies between the numerical summaries for location and dispersion, and their robustness properties are comparable.

	Least Robust	More Robust	Most Robust
Location	$\frac{x_{(1)} + x_{(n)}}{2}$	\bar{x}	$x_{(\{n+1\}/2)}$
Dispersion	$x_{(n)} - x_{(1)}$	s^2	$x_{(3\{n+1\}/4)} - x_{(\{n+1\}/4)}$

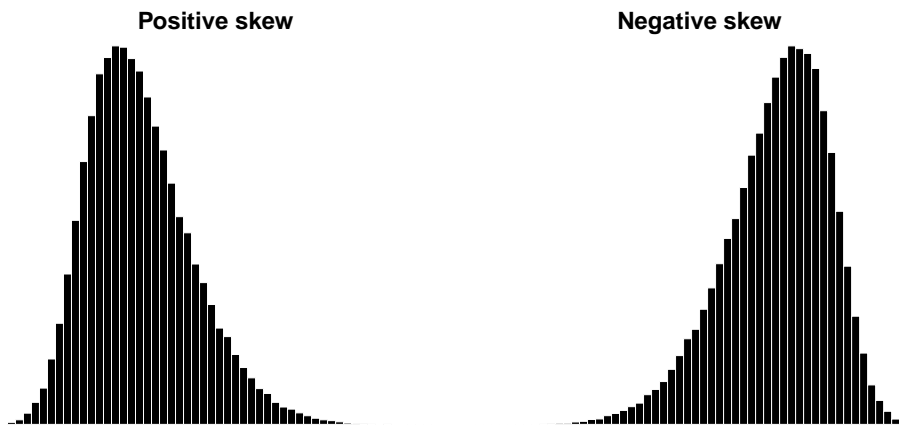
(where $\frac{x_{(1)} + x_{(n)}}{2}$ would be the midpoint of our data halfway between the minimum and maximum values in the sample, which provides another alternative descriptor of location.)

3.1.3 Skewness

Skewness is a measure of asymmetry. The **skewness** of a sample of real values (x_1, \dots, x_n) is given by

$$\text{skewness} = \frac{1}{n} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s} \right)^3.$$

A sample is **positively (negatively) or right (left) skewed** if the upper tail of the histogram of the sample is longer (shorter) than the lower tail.



Since the mean is more sensitive to outlying points than is the median, one might choose the median as a more suitable measure of 'average value' if the sample is skewed.

We expect skewness for example when the data can only take positive (or only negative values) and if the values are not far from zero.

We can remove skewness by transforming the data. In the case above, we need a transformation which has larger effect on the larger values: e.g. square root, log (though beware 0 values).

Note For a positively skewed sample the mean is greater than the median.

3.1.4 Covariance and Correlation

Suppose we have a sample made up of ordered pairs of real values $((x_1, y_1), \dots, (x_n, y_n))$. The value x_i might correspond to the measurement of one quantity x of individual i , and y_i to another quantity y of the same individual e.g. an individual's weight and height.

The **covariance** between the samples of x and y is given by

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

It gives a measurement of relatedness between the two quantities x and y .

The covariance can be rewritten equivalently as

$$s_{xy} = \frac{\sum_{i=1}^n x_i y_i}{n} - \bar{x}\bar{y}.$$

Note that the magnitude of s_{xy} varies according to the scale on which the data have been measured. The **correlation** between the samples of x and y is defined to be

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{n s_x s_y}.$$

where s_x and s_y are the sample standard deviations of (x_1, \dots, x_n) and (y_1, \dots, y_n) respectively.

Unlike covariance, correlation gives a measurement of relatedness between the two quantities x and y which is scale-invariant.

3.2 Related Graphical Displays

3.2.1 Box-and-Whisker Plots

Based on the five point summary.

- Median – middle line in the box
- 3rd & 1st Quartiles – top and bottom of the box
- 'Whiskers' – extend out to any points which are within $(\frac{3}{2} \times \text{interquartile range})$ from the box
- Any extreme points out to the maximum and minimum which are beyond the whiskers are plotted individually.

Example Figure 3.2 are box plots of the counts of insects found in agricultural experimental units treated with six different insecticides (A-F).

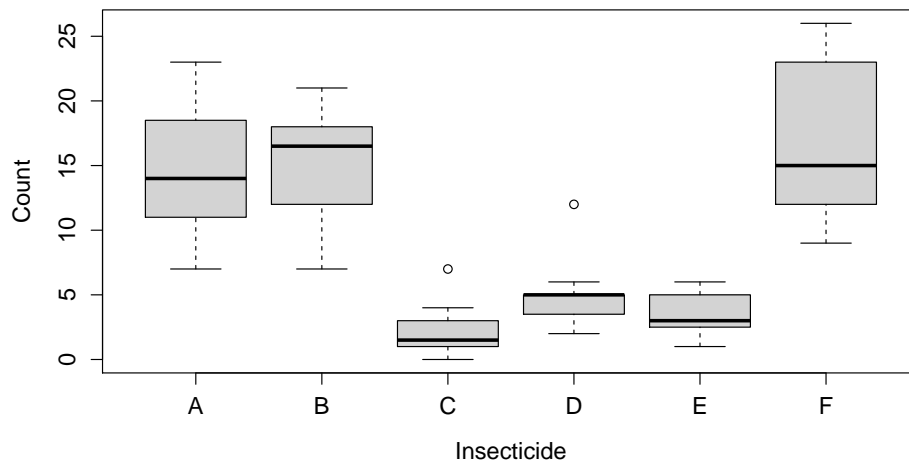


Figure 3.2

3.2.2 Empirical CDF

The **empirical cumulative distribution function (CDF)** of a sample of real values (x_1, \dots, x_n) is given by

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n I(x_i \leq x)$$

For any real number x , $F_n(x)$ returns the proportion of the data having values which do not exceed x . Note this is a step function, with change points at the sampled values.

Example Figure 3.3 is a plot of the empirical CDF of the insecticide data across the different treatments.

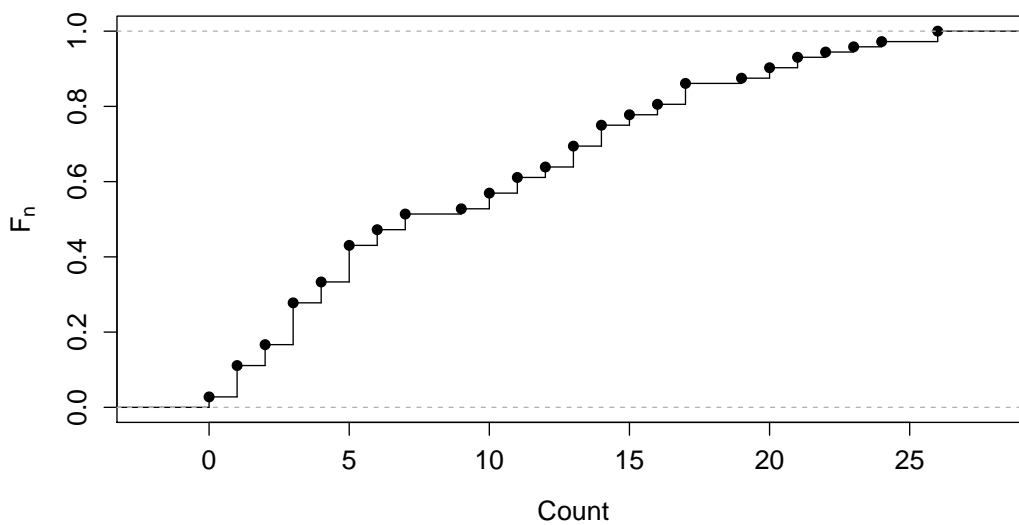


Figure 3.3

Chapter 4. Elementary Set Theory

This chapter develops the formal mathematical model for events, the theory of sets. We introduce several important notions such as elements, empty sets, intersection, union and disjoint sets.

4.1 Sets, subsets and complements

4.1.1 Sets and notation

A **set** is any collection of distinct objects, and is a fundamental object of mathematics.

$\emptyset = \{\}$, the empty set containing no objects, is included.

The objects in a set can be anything, for example integers, real numbers, or the objects may even themselves be sets.

\in	-	"is an element of" (set membership)
\iff	-	"if and only if" (equivalence)
\implies	-	"implies"
\exists	-	"there exists"
\forall	-	"for all"
s.t. or $ $	-	"such that"
wrt	-	"with respect to"

Table 4.1: Notations

4.1.2 Subsets, Complements and Singletons

If a set B contains all of the objects contained in another set A , and possibly some other objects besides, we say A is a **subset** of B and write $A \subseteq B$.

Suppose $A \subseteq B$ for two sets A and B . If we also have $B \subseteq A$ we write $A = B$, whereas if we know $B \not\subseteq A$ we write $A \subset B$.

The **complement** of a set A wrt a universal set Ω (say, of all "possible values") is $\overline{A} = \{\omega \in \Omega | \omega \notin A\}$.

A **singleton** is a set with exactly one element — $\{\omega\}$ for some $\omega \in \Omega$.

4.2 Set operations

4.2.1 Unions and Intersections

Consider two sets A and B .

The **union** of A and B , $A \cup B = \{\omega \in \Omega | \omega \in A \text{ or } \omega \in B\}$.

The **intersection** of A and B , $A \cap B = \{\omega \in \Omega | \omega \in A \text{ and } \omega \in B\}$.

More generally, for sets A_1, A_2, \dots we define

$$\bigcup_i A_i = \{\omega \in \Omega \mid \exists i \text{ s.t. } \omega \in A_i\}$$

$$\bigcap_i A_i = \{\omega \in \Omega \mid \forall i, \omega \in A_i\}$$

If $A \cap B = \emptyset$, then we say the sets are **disjoint**, that is, the sets have no common element.

The sets $A_1, \dots, A_k \subseteq \Omega$ form a partition of event $B \subseteq \Omega$ if

(a) $A_i \cap A_j = \emptyset$, for $i \neq j$, $i, j = 1, \dots, k$ (disjoint)

(b) $\bigcup_{i=1}^k A_i = B$

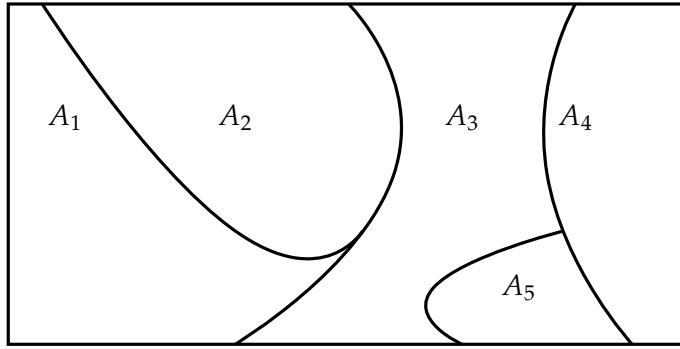


Figure 4.1: Partition of Ω

In Figure 4.1, we have $\Omega = \bigcup_{i=1}^5 A_i$.

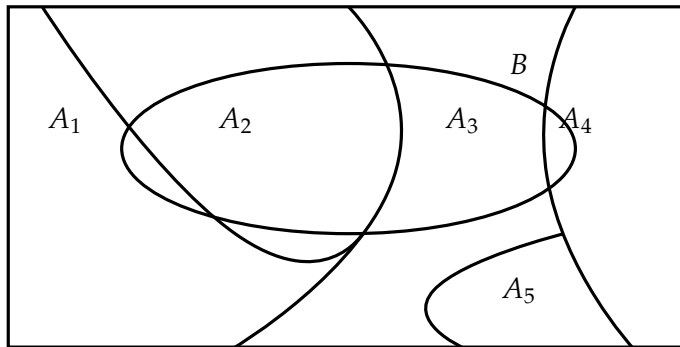


Figure 4.2: Partition of $B \subset \Omega$

In Figure 4.2, we have $B = \bigcup_{i=1}^5 (A_i \cap B)$, but, for example, $B \cap A_5 = \emptyset$.

Properties of Union and Intersection Operators

Consider the sets $A, B, C \subseteq \Omega$

$$\begin{array}{ll} \text{COMMUTATIVITY} & A \cup B = B \cup A \\ & A \cap B = B \cap A \end{array}$$

$$\begin{array}{ll} \text{ASSOCIATIVITY} & A \cup (B \cup C) = (A \cup B) \cup C \\ & A \cap (B \cap C) = (A \cap B) \cap C \end{array}$$

$$\begin{array}{ll} \text{DISTRIBUTIVITY} & A \cup (B \cap C) = (A \cup B) \cap (A \cup C) \\ & A \cap (B \cup C) = (A \cap B) \cup (A \cap C) \end{array}$$

$$\begin{array}{ll} \text{DE MORGAN'S LAWS} & \overline{(A \cup B)} = \overline{A} \cap \overline{B} \\ & \overline{(A \cap B)} = \overline{A} \cup \overline{B} \end{array}$$

The **difference** of A and B is $A \setminus B = A \cap \overline{B} = \{\omega \in \Omega \mid \omega \in A \text{ and } \omega \notin B\}$.

Example Let $\Omega = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ be our universal set and $A = \{1, 2, 3, 4, 5, 6\}$ and $B = \{5, 6, 7, 8, 9\}$ be two sets of elements of Ω .

- $A \cup B = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$
- $A \cap B = \{5, 6\}$
- $A \setminus B = \{1, 2, 3, 4\}$
- $\overline{(A \cup B)} = \{10\}$
- $\overline{(A \cap B)} = \{1, 2, 3, 4, 7, 8, 9, 10\}$
- $\overline{A} = \{7, 8, 9, 10\}$, $\overline{B} = \{1, 2, 3, 4, 10\}$
- $\overline{A} \cap \overline{B} = \{10\}$
- $\overline{A} \cup \overline{B} = \{1, 2, 3, 4, 7, 8, 9, 10\}$

■

4.2.2 Cartesian Products

For two sets Ω_1, Ω_2 , their **Cartesian product** is the set of all ordered pairs of their elements. That is,

$$\Omega_1 \times \Omega_2 = \{(\omega_1, \omega_2) \mid \omega_1 \in \Omega_1, \omega_2 \in \Omega_2\}$$

More generally, the Cartesian product for sets $\Omega_1, \Omega_2, \dots$ is written $\prod_i \Omega_i$.

4.3 Cardinality

A useful *measure* of a set is the size, or **cardinality**. The cardinality of a finite set is simply the number of elements it contains. For infinite sets, there are again an *infinite* number of different cardinalities they can take. However, amongst these there is a most important distinction: Between those which are **countable** and those which are not.

A set Ω is countable if \exists a function $f : \mathbb{N} \rightarrow \Omega$ s.t. $f(\mathbb{N}) \supseteq \Omega$. That is, the elements of Ω can satisfactorily be written out as a possibly unending list $\{\omega_1, \omega_2, \omega_3, \dots\}$. Note that all finite sets are countable.

A set is **countably infinite** if it is countable but not finite. Clearly \mathbb{N} is countably infinite. So is $\mathbb{N} \times \mathbb{N}$.

A set which is not countable is **uncountable**. For instance, \mathbb{R} is uncountable.

The empty set \emptyset has zero cardinality,

$$|\emptyset| = 0.$$

For finite sets A and B , if A and B are disjoint (that is $A \cap B = \emptyset$), then

$$|A \cup B| = |A| + |B|$$

otherwise,

$$|A \cup B| = |A| + |B| - |A \cap B|.$$

Summary of Notation

Ω	universal set
\emptyset	empty set
$A \subseteq \Omega$	Subset of Ω
\overline{A}	Complement of A
$ A $	Cardinality (or size) of A
$A \cup B$	union (A or B)
$A \cap B$	intersection (A and B)
$A \subset B$	set inclusion (elements of A are also in B)
A/B	set difference (elements in A that are not in B)

Chapter 5. Probability

Probability is a mathematical language for quantifying uncertainty. In this chapter we introduce the basic ideas underlying probability theory.

5.1 Sample Spaces and Events

5.1.1 Sample Spaces

We consider a random **experiment** whose range of possible outcomes can be described by a set S , called the **sample space**.

We use S as our universal set (Ω).

Example

- Coin tossing: $S = \{H, T\}$.
- Die rolling: $S = \{\square, \square, \square, \square, \square, \square\}$.
- 2 coins: $S = \{(H, H), (H, T), (T, H), (T, T)\}$.



5.1.2 Events

An **event** E is any subset of the sample space, $E \subseteq S$; it is a collection of some of the possible outcomes.

Example

- Coin tossing: $E = \{H\}$, $E = \{T\}$.
- Die rolling: $E = \{\square\}$, $E = \{\text{Even numbered face}\} = \{\square, \square, \square\}$.
- 2 coins: $E = \{\text{Head on the first toss}\} = \{(H, H), (H, T)\}$.



Extreme possible events are \emptyset (the *null* event) or S .

The *singleton* subsets of S , those subsets which contain exactly one element from S , are known as the **elementary events** of S .

Suppose we now perform this random experiment; the outcome will be a single element $s^* \in S$. Then for any event $E \subseteq S$, we will say E has **occurred** if and only if $s^* \in E$.

If E has not occurred, it must be that $s^* \notin E \iff s^* \in \bar{E}$, so \bar{E} has occurred; so \bar{E} can be read as the event *not* E .

First notice that the smallest event which will have occurred will be the singleton $\{s^*\}$. For any other event E , E will occur if and only if $\{s^*\} \subset E$. Thus we can immediately draw two conclusions before the experiment has even been performed.

Remark 1. For any sample space S , the following statements will always be true:

1. the null event \emptyset will never occur;
2. the universal event S will always occur.

Hence it is only for events E in between these extreme events, $\emptyset \subset E \subset S$ for which we have uncertainty about whether E will occur. It is precisely for quantifying this uncertainty over these events that we require the notion of probability.

5.1.3 Combinations of Events

Consider a set of events $\{E_1, E_2, \dots\}$.

- The event $\bigcup_i E_i = \{s \in S \mid \exists i \text{ s.t. } s \in E_i\}$ will occur if and only if *at least one* of the events $\{E_i\}$ occurs. So $E_1 \cup E_2$ can be read as event E_1 or E_2 or both.
- The event $\bigcap_i E_i = \{s \in S \mid \forall i, s \in E_i\}$ will occur if and only if *all* of the events $\{E_i\}$ occur. So $E_1 \cap E_2$ can be read as events E_1 and E_2 .
- The events are said to be **mutually exclusive** if $\forall i, j, E_i \cap E_j = \emptyset$ (i.e. they are disjoint). At most one of the events can occur.

5.2 The σ -algebra

Henceforth, we shall think of events as subsets of the sample space S . Thus events are subsets of S , but need all subsets of S be events? The answer is *no*, for reasons that are beyond the scope of this course. But it suffices to think of the collection of events as a subcollection \mathcal{F} of the sets of all subsets of S . This subcollection should have the following properties:

- a) if $E, F \in \mathcal{F}$ then $E \cup F \in \mathcal{F}$ and $E \cap F \in \mathcal{F}$;
- b) if $E \in \mathcal{F}$ then $\bar{E} \in \mathcal{F}$;
- c) $\emptyset \in \mathcal{F}$.

A collection \mathcal{F} of subsets of S which satisfies these three conditions is called a **field**. It follows from the properties of a field that if $E_1, E_2, \dots, E_k \in \mathcal{F}$, then

$$\bigcup_{i=1}^k E_i \in \mathcal{F}.$$

So, \mathcal{F} is closed under finite unions and hence under finite intersections also. To see this note that if $E_1, E_2 \in \mathcal{F}$, then

$$\bar{E}_1, \bar{E}_2 \in \mathcal{F} \implies \underbrace{\bar{E}_1 \cup \bar{E}_2 \in \mathcal{F}}_{\text{by a)}} \implies \underbrace{\overline{(\bar{E}_1 \cup \bar{E}_2)} \in \mathcal{F}}_{\text{by b)}} \implies \underbrace{E_1 \cap E_2 \in \mathcal{F}}_{\text{by De Morgans Law}}.$$

This is fine when S is a finite set, but we require slightly more to deal with the common situation when S is infinite, as the following example indicates.

Example A coin is tossed repeatedly until the first head turns up; we are concerned with the number of tosses before this happens. The set of all possible outcomes is the set $S = \{0, 1, 2, \dots\}$. We may seek to assign a probability to the event E , that the first head occurs after an even number of tosses, that is $E = \{1, 3, 5, \dots\}$. This is an infinite countable union of members of S and we require that such a set belongs to \mathcal{F} in order that we can discuss its probability. ■

Thus we also require that the collection of events to be closed under the operation of taking countable unions, not just finite unions.

Definition 5.2.1. A collection \mathcal{F} of subsets of S is called a **σ -field** (or **σ -algebra**) if it satisfies the following conditions:

- a) $\emptyset \in \mathcal{F}$;
- b) if $E_1, E_2, \dots \in \mathcal{F}$ then $\bigcup_{i=1}^{\infty} E_i \in \mathcal{F}$;
- c) if $E \in \mathcal{F}$ then $\bar{E} \in \mathcal{F}$.

Here are some examples of σ -algebras:

Example

- The smallest σ -algebra associated with S is the collection $\mathcal{F} = \{\emptyset, S\}$.
- If E is any subset of S then $\mathcal{F} = \{\emptyset, E, \bar{E}, S\}$ is a σ -algebra. ■

To recap, with any experiment we may associate a pair (S, \mathcal{F}) , where S is the set of all possible outcomes (or elementary events) and \mathcal{F} is a σ -algebra of subsets of S , which contains all the events in whose occurrences we may be interested. So, from now on, to call a set E an event is equivalent to asserting that E belongs to the σ -algebra in question.

5.3 Probability Measure

For an event $E \subseteq S$, the probability that E occurs will be written as $P(E)$.

Interpretation: P is a set-function or measure that assigns “weight” to collections of possible outcomes of an experiment. There are many ways to think about precisely how this assignment is achieved;

CLASSICAL : “Consider equally likely sample outcomes ...”

FREQUENTIST : “Consider long-run *relative frequencies*...”

SUBJECTIVE : “Consider personal degree of belief...”

Formally, we have the following definition:

Definition 5.3.1. A **probability measure** P on (S, \mathcal{F}) is a mapping $P : \mathcal{F} \rightarrow [0, 1]$ satisfying

a) $P(S) = 1$;

b) if E_1, E_2, \dots is a collection of disjoint members of \mathcal{F} , so that $E_i \cap E_j = \emptyset$ for all pairs i, j with $i \neq j$, then

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

The triple (S, \mathcal{F}, P) , consisting of a set S , a σ -algebra \mathcal{F} and probability measure P on (S, \mathcal{F}) is called a **probability space**.

5.3.1 Properties of $P(\cdot)$: The Axioms of Probability

The triple (S, \mathcal{F}, P) denotes a typical probability space. We now give some of its simple but important properties. For events $E, F \subseteq S$,

1. $P(\bar{E}) = 1 - P(E)$
2. if $E \subseteq F$, then $P(E) \leq P(F)$.
3. In general, $P(E \cup F) = P(E) + P(F) - P(E \cap F)$.
4. $P(E \cap \bar{F}) = P(E) - P(E \cap F)$.
5. $P(E \cup F) \leq P(E) + P(F)$.
6. $P(E \cap F) \geq P(E) + P(F) - 1$.

5.3.2 Interpretations of Probability

Classical

Suppose that the sample space $S = \{s_1, \dots, s_n\}$ is finite. For example, if we toss a die twice, then S has 36 elements: $S = \{(i, j); i, j \in \{1, \dots, 6\}\}$. If each outcome is equally likely, then $P(E) = |E|/36$ where $|E|$ denotes the number of elements in E . The probability that the sum of the dice is 11 is $2/36$ since there are two outcomes that correspond to this event.

In general, if S is finite and the *elementary events* are considered “equally likely”, then the probability of an event E is the proportion of all outcomes in S in which lie inside E ,

$$P(E) = \frac{|E|}{|S|}.$$

Example Rolling a die: Elementary events are $\{\square\}, \{\square\cdot\}, \dots, \{\square\cdot\cdot\cdot\}$.

- $P(\{\square\}) = P(\{\square\cdot\}) = \dots = P(\{\square\cdot\cdot\cdot\}) = \frac{1}{6}$.
- $P(\text{Odd number}) = P(\{\square\cdot, \square\cdot\cdot, \square\cdot\cdot\cdot\}) = \frac{3}{6} = \frac{1}{2}$.

■

The “equally likely” (uniform) idea can be extended to infinite spaces, by apportioning probability to sets not by their cardinality but by other standard *measures*, like volume or mass.

Example If a meteorite were to strike Earth, the probability that it will strike land rather than sea would be given by

$$\frac{\text{Total area of land}}{\text{Total area of Earth}}.$$

■

Frequentist

Observation shows that if one takes repeated observations in “identical” random situations, in which event E may or may not occur, then the proportion of times in which E occurs tends to some limiting value – called the probability of E .

Example Proportion of heads in tosses of a coin: $H, H, T, H, T, T, H, T, T, \dots \rightarrow \frac{1}{2}$.

■

Subjective

Probability is a degree of belief held by an individual.

For example, De Finetti (1937/1964) suggested the following: Suppose a random experiment is to be performed, where an event $E \subseteq S$ may or may not happen. Now suppose an individual is entered into a game regarding this experiment where he has two choices, each leading to monetary consequences:

1. Gamble: If E occurs, he wins £1; if \bar{E} occurs, he wins £0;
2. Stick: Regardless of the outcome of the experiment, he receives $\mathbb{E}P(E)$ for some real number $P(E)$.

The critical value of $P(E)$ for which the individual is *indifferent* between options 1 and 2 is defined to be the individual’s probability for the event E occurring.

This procedure can be repeated for all possible events E in S .

Suppose after this process of *elicitation* of the individual’s preferences under the different events, we can simultaneously arrange an arbitrary number of monetary bets with the individual based on the outcome of the experiment.

If it is possible to choose these bets in such a way that the individual is certain to lose money (this is called a “Dutch Book”), then the individuals degrees of belief are said to be **incoherent**.

To be **coherent**, it is easily seen, for example, that we must have $0 \leq P(E) \leq 1$ for all events E , $E \subseteq F \implies P(E) \leq P(F)$, etc.

5.3.3 Independent Events

If we flip a fair coin twice, then the probability of two heads is $\frac{1}{2} \times \frac{1}{2}$. We multiply the probabilities because we regard the two tosses as independent. The definition of independence is as follows:

Definition 5.3.2. Two events E and F are **independent** if and only if

$$P(E \cap F) = P(E)P(F).$$

Extension: The events E_1, \dots, E_k are independent if, for **every** subset of events of size $\ell \leq k$, indexed by $\{i_1, \dots, i_\ell\}$, say,

$$P\left(\bigcap_{j=1}^{\ell} E_{i_j}\right) = \prod_{j=1}^{\ell} P(E_{i_j})$$

Independence can arise in two distinct ways. Sometimes, we explicitly assume that two events are independent. For example, in tossing a coin twice we usually assume the tosses are independent. In other instances, we derive independence by verifying that $P(E \cap F) = P(E)P(F)$ holds true.

Example Toss a fair coin 10 times. Let A be the event {at least one head}. Let T_j be the event {tails occurs on the j th toss}. Then

$$\begin{aligned} P(A) &= 1 - P(\bar{A}) \\ &= 1 - P(\text{all tails}) \\ &= 1 - P(T_1 \cap T_2 \cap \dots \cap T_{10}) \\ &= 1 - P(T_1)P(T_2) \cdots P(T_{10}) \quad \text{by indep.} \\ &= 1 - \left(\frac{1}{2}\right)^{10} \end{aligned}$$

■

Example Suppose that the events E and F are independent. Show that \bar{E} and F are also independent.

$$\text{So } P(\bar{E} \cap F) = \underbrace{P(F) - P(E \cap F)}_{\text{by Axiom (4)}} = \underbrace{P(F) - P(E)P(F)}_{\text{by indep}} = (1 - P(E))P(F) = P(\bar{E})P(F). \quad \blacksquare$$

Example Suppose that E and F are disjoint events, and that $P(E) > 0$ and $P(F) > 0$. Can the events E and F be independent?

No. As $P(E) > 0$ and $P(F) > 0$ we have that $P(E)P(F) > 0$. However, since $E \cap F = \emptyset$ we also have that $P(E \cap F) = P(\emptyset) = 0$. Since $P(E \cap F) \neq P(E)P(F)$ we conclude that the events are not independent.



Summary of Independence

1. E and F are independent if and only if $P(E \cap F) = P(E)P(F)$.
2. Independence is sometimes assumed and sometimes derived.
3. Disjoint events with positive probability are not independent.

5.4 More Examples

Example Which of these two events is more likely?

$E = \{4 \text{ rolls of a die yield at least one } 1\}$; or

$F = \{24 \text{ rolls of two dice yield at least one pair of } (1,1)\}$.

We calculate $P(E)$ and $P(F)$.

1. Each roll of the die is independent from the previous rolls, and so there are 6^4 equally likely outcomes. Of these, 5^4 show no 1s.

So the probability of no 1 showing is $\frac{5^4}{6^4} \approx 0.4823$.

So $P(E)$, the probability of at least one 1 showing, is $= 1 - \frac{5^4}{6^4} \approx 1 - 0.4823 = 0.5177$.

2. There are 36^{24} equally likely outcomes here. Of these, 35^{24} don't show a (1,1).

So the probability of no (1,1) is $\frac{35^{24}}{36^{24}} \approx 0.5086$

So $P(F)$, the probability of at least one (1,1), is $\approx 1 - 0.5086 = 0.4914$

Hence $P(E) \approx 0.5177 > \frac{1}{2} > 0.4914 \approx P(F)$.



Example There is a 1% probability for a hard drive to crash. Therefore, it has two backups, each having a 2% probability to crash, and all three components are independent of each other. The stored information is lost only in the event that all three devices crash. What is the probability that the information is saved?

Start by organising and labelling the events. Denote

$H = \{\text{hard drive crashes}\}$

$B_1 = \{\text{first backup crashes}\}$

$B_2 = \{\text{second backup crashes}\}$

In the wording, we are given that H, B_1 and B_2 are independent and

$$P(H) = 0.01, \quad P(B_1) = 0.02, \quad P(B_2) = 0.02$$

Then, applying rules of complements and intersection for independent events we have

$$\begin{aligned} P(\text{saved}) &= 1 - P(\text{lost}) = 1 - P(H \cap B_1 \cap B_2) \\ &= 1 - P(H)P(B_1)P(B_2) \\ &= 1 - (0.01)(0.02)(0.02) = 1 - 0.000004 = 0.999996 \end{aligned}$$







■

5.4.1 Joint events

So far we have only considered a single outcome or event. We can extend the ideas seen so far for joint events.

For example, consider tossing a coin and rolling a die. We would consider each of the 12 possible combinations of Head/Tail and die value as equally likely.

So we can construct a **probability table**:

						
H	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$
T	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$
	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

From this table we can calculate the probability of *any* event we might be interested in, simply by adding up the probabilities of all the elementary events it contains.

For example, the event of getting a head on the coin

$$\{H\} = \{(H, \text{die face 1}), (H, \text{die face 2}), \dots, (H, \text{die face 6})\}$$




has probability

$$\begin{aligned} P(\{H\}) &= P(\{(H, \text{die face 1})\}) + P(\{(H, \text{die face 2})\}) + \dots + P(\{(H, \text{die face 6})\}) \\ &= \frac{1}{12} + \frac{1}{12} + \dots + \frac{1}{12} \\ &= \frac{1}{2}. \end{aligned}$$

Notice the two experiments satisfy our probability definition of independence, since for example

$$P(\{(H, \text{die face 6})\}) = \frac{1}{12} = \frac{1}{2} \times \frac{1}{6} = P(\{H\}) \times P(\{\text{die face 6}\}).$$

A crooked die called a *top* has the same faces on opposite sides.

Suppose we have two dice, one normal and one which is a top with opposite faces numbered , , or .

Now suppose we first flip the coin. If it comes up heads, we roll the normal die; tails, and we roll the top.

To calculate the probability table easily, we notice that this is equivalent to the previous game using one normal die except with the change after tails that a roll of a \square is relabelled as a \square , $\square \rightarrow \square$, $\square \rightarrow \square$. So we can just merge those probabilities in the tails row.

	\square	\square	\square	\square	\square	\square
H	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$
T	$\frac{1}{6}$	0	$\frac{1}{6}$	0	$\frac{1}{6}$	0
	$\frac{1}{4}$	$\frac{1}{12}$	$\frac{1}{4}$	$\frac{1}{12}$	$\frac{1}{4}$	$\frac{1}{12}$

The probabilities of the different outcomes of the dice change according to the outcome of the coin toss. And note, for example,

$$P(\{(H, \square)\}) = \frac{1}{12} \neq \frac{1}{24} = \frac{1}{2} \times \frac{1}{12} = P(\{H\}) \times P(\{\square\}).$$

So the two experiments are now *dependent*.

5.5 Conditional Probability

Assuming that $P(F) > 0$, we define the conditional probability of E given that F has occurred as follows:

Definition 5.5.1. If $P(F) > 0$ then the **conditional probability** of E given F is

$$P(E|F) = \frac{P(E \cap F)}{P(F)}.$$

Note If E and F are independent, then

$$P(E|F) = \frac{P(E \cap F)}{P(F)} = \frac{P(E)P(F)}{P(F)} = P(E).$$

Example Suppose a normal die is rolled once.

Questions

Q1) What is the probability of $E = \{\text{the die shows a } \square\}$?

Q2) What is the probability of $E = \{\text{the die shows a } \square\}$ given we know $F = \{\text{the die shows an odd number}\}$?

Solutions

$$S1) P(E) = \frac{\text{Number of ways a } \square \text{ can come up}}{\text{Total number of possible outcomes}} = \frac{1}{6}.$$

S2) Now the set of possible outcomes is just $F = \{\square, \square, \square\}$.

$$\text{So } P(E|F) = \frac{\text{Number of ways a } \square \text{ can come up}}{\text{Total number of possible outcomes}} = \frac{1}{3}.$$

$$\text{Note } P(F) = \frac{1}{2} \text{ and } E \cap F = E, \text{ and hence we have } P(E|F) = \frac{P(E \cap F)}{P(F)}.$$

Now suppose we roll two normal dice, one from each hand. Then the sample space comprises all of the ordered pairs of dice values

$$S = \{(\square, \square), (\square, \square), \dots, (\square, \square)\}.$$

Let E be the event that the die thrown from the left hand will show a larger value than the die thrown from the right hand.

$$P(E) = \frac{\# \text{ outcomes with left value} > \text{right}}{\text{total \# outcomes}} = \frac{15}{36}.$$

Suppose we are now informed that an event F has occurred, where

$$F = \{\text{the value of the left hand die is } \square\}$$

How does this change the probability of E occurring?

Well since F has occurred, the only sample space elements which could have possibly occurred are exactly those elements in $F = \{(\square, \square), (\square, \square), (\square, \square), (\square, \square), (\square, \square), (\square, \square)\}$.

Similarly the only sample space elements in E that could have occurred now must be in $E \cap F = \{(\square, \square), (\square, \square), (\square, \square), (\square, \square)\}$.

So our revised probability is

$$\frac{\# \text{ outcomes with left value } \square > \text{right}}{\text{total \# outcomes } (\square, \cdot)} = \frac{4}{6} = \frac{P(E \cap F)}{P(F)} \equiv P(E|F).$$

■

In both examples, we considered the probability of an event E , and then reconsidered what this probability would be if we were given the knowledge that F had occurred. What we did was replace the sample space S by F , and the event E was replaced by $E \cap F$. So originally, we had

$$P(E) = P(E|S) = \frac{P(E \cap S)}{P(S)} \text{ (since } E \cap S = E, \text{ and } P(S) = 1 \text{ by def of a prob measure.)}$$

So we can think of probability conditioning as a shrinking of the sample space, with events replaced by their intersections with the reduced space and a consequent rescaling of probabilities.

Warning! It is generally the case that $P(E|F) \neq P(F|E)$. People are confused by this all the time. For example, the probability of spots given you have measles is 1, but the probability

that you have measles given that you have spots is not 1. In this case, the difference between $P(E|F)$ and $P(F|E)$ is obvious but there are cases where it is less obvious.

Example A medical test for a disease D has outcomes $+$ and $-$. The probabilities are

	D	\bar{D}
$+$	0.009	0.099
$-$	0.001	0.891

Using the definition of conditional probability, we have

$$P(+|D) = \frac{P(+ \cap D)}{P(D)} = \frac{0.009}{0.009 + 0.001} = 0.9$$

and

$$P(-|\bar{D}) = \frac{P(- \cap \bar{D})}{P(\bar{D})} = \frac{0.891}{0.891 + 0.099} \approx 0.9$$

Apparently, the test is fairly accurate. Sick people yield a positive 90% of the time and healthy people yield a negative about 90% of the time.

Now suppose you go for a test and get a positive. What is the probability you have the disease? Most people would answer 0.9. The correct answer is

$$P(D|+) = \frac{P(+ \cap D)}{P(+)} = \frac{0.009}{0.009 + 0.099} \approx 0.08$$

which is much less than 0.9.

The lesson here is that you need to compute the answer numerically. Do not trust your intuition!

Also note that $P(+|D) \neq P(D|+)$ ■

5.5.1 Conditional Independence

Earlier we met the concept of independence of events according to a probability measure P . We can now extend that idea to conditional probabilities since $P(\cdot|F)$ is itself a measure obeying the axioms of probability.

For three events E_1 , E_2 and F , the event pair E_1 and E_2 are said to be **conditionally independent** given F if and only if $P(E_1 \cap E_2|F) = P(E_1|F)P(E_2|F)$.¹

5.5.2 Bayes' Theorem

Before we introduce Bayes' theorem, we require a preliminary result.

¹This is sometimes written $E_1 \perp E_2|F$.

Theorem 5.6 (The Theorem of Total Probability). Let E_1, \dots, E_k be a partition on S . Then, for any event $F \subseteq S$, we have

$$P(F) = \sum_{i=1}^k P(F|E_i)P(E_i)$$

Proof. Define $C_j = F \cap E_j$. Note that C_1, \dots, C_k are disjoint and that $F = \bigcup_{j=1}^k C_j$. Hence,

$$P(F) = \underbrace{P\left(\bigcup_{j=1}^k C_j\right)}_{\text{by def. of probability measure}} = \sum_{j=1}^k P(C_j) = \sum_{j=1}^k P(F \cap E_j) = \sum_{j=1}^k P(F|E_j)P(E_j)$$

□

A simple use of this theorem is as follows: for any events E and F in S , note that $\{F, \bar{F}\}$ form a partition of S . So by the law of total probability we have

$$\begin{aligned} P(E) &= P(E \cap F) + P(E \cap \bar{F}) \\ &= P(E|F)P(F) + P(E|\bar{F})P(\bar{F}). \end{aligned}$$

Theorem 5.7 (Bayes' Theorem). Let E_1, \dots, E_k be a partition on S such that $P(E_i) > 0$ for each i . If $P(F) > 0$ then, for each $i = 1, \dots, k$, we have

$$P(E_i|F) = \frac{P(F|E_i)P(E_i)}{P(F)} = \frac{P(F|E_i)P(E_i)}{\sum_{j=1}^k P(F|E_j)P(E_j)}$$

Proof. Starting with the LHS, we have

$$P(E_i|F) = \frac{P(E_i \cap F)}{P(F)} = \frac{P(F|E_i)P(E_i)}{P(F)} = \frac{P(F|E_i)P(E_i)}{\sum_{j=1}^k P(F|E_j)P(E_j)}$$

□

Note We will often deal with probabilities of single events, intersection of events and conditional events. To this end we will refer to

- probabilities of the form $P(E|F)$ as **conditional probabilities**;
- probabilities of the form $P(E \cap F)$ as **joint probabilities**;
- probabilities of the form $P(E)$ as **marginal probabilities**.

5.7.1 More Examples

Example A box contains 5000 VLSI chips, 1000 from company X and 4000 from Y. 10% of the chips made by X are defective and 5% of those made by Y are defective. If a randomly chosen chip is found to be defective, find the probability that it came from company X.

Let E = “chip was made by X”;

let F = “chip is defective”.

First of all, which probabilities have we been given?

The statement

“A box contains 5000 VLSI chips, 1000 from company X and 4000 from Y.”

$$\implies P(E) = \frac{1000}{5000} = 0.2, \quad \text{and} \quad P(\bar{E}) = \frac{4000}{5000} = 0.8.$$

and

“10% of the chips made by X are defective and 5% of those made by Y are defective.”

$$\implies P(F|E) = 10\% = 0.1, \quad \text{and} \quad P(F|\bar{E}) = 5\% = 0.05.$$

We have enough information to construct the probability table

	E	\bar{E}	
F	0.02	0.04	0.06
\bar{F}	0.18	0.76	0.94
	0.2	0.8	

The law of total probability has enabled us to extract the marginal probabilities $P(F)$ and $P(\bar{F})$ as 0.06 and 0.94 respectively.

So by Bayes Theorem we can calculate the conditional probabilities. In particular, we want

$$P(E|F) = \frac{P(E \cap F)}{P(F)} = \frac{0.02}{0.06} = \frac{1}{3}.$$

■

Example Kidney stones are small (< 2cm diam) or large (> 2cm diam). Treatment can succeed or fail. The following data were collected from a sample of 700 patients with kidney stones.

	Success (S)	Failure (\bar{S})	
Large (L)	247	96	343
Small (\bar{L})	315	42	357
Total	562	138	700

For a patient randomly drawn from this sample, what is the probability that the outcome of treatment was successful, given the kidney stones were large?

Clearly we can get the answer directly from the table by ignoring the small stone patients

$$P(S|L) = \frac{247}{343}$$

or we can go the long way round:

$$P(L) = \frac{343}{700}, \quad P(S \cap L) = \frac{247}{700},$$

$$P(S|L) = \frac{P(S \cap L)}{P(L)} = \frac{\frac{247}{700}}{\frac{343}{700}} = \frac{247}{343}.$$

■

Example A multiple choice question has c available choices. Let p be the probability that the student knows the right answer, and $1 - p$ that he does not. When he doesn't know, he chooses an answer at random. Given that the answer the student chooses is correct, what is the probability that the student knew the correct answer?

Let A be the event that the question is answered correctly;

let K be the event that the student knew the correct answer.

Then we require $P(K|A)$.

By Bayes Theorem

$$P(K|A) = \frac{P(A|K)P(K)}{P(A)}$$

and we know $P(A|K) = 1$ and $P(K) = p$, so it remains to find $P(A)$.

By the partition rule, we have

$$P(A) = P(A|K)P(K) + P(A|\bar{K})P(\bar{K})$$

and since $P(A|\bar{K}) = \frac{1}{c}$, this gives

$$P(A) = 1 \times p + \frac{1}{c} \times (1 - p).$$

Hence

$$P(K|A) = \frac{p}{p + \frac{1-p}{c}} = \frac{cp}{cp + 1 - p}.$$

Notice that the larger c is, the greater the probability that the student knew the answer, given that they answered correctly. ■

Example Measurements at the North Carolina Super Computing Center (NCSC) on a certain day showed that 15% of the jobs came from Duke, 35% from UNC, and 50% from NC State

University. Suppose that the probabilities that each of these jobs is a multitasking job is 0.01, 0.05, and 0.02 respectively.

Questions

- Q1) Find the probabilities that a job chosen at random is a multitasking job.
- Q2) Find the probability that a randomly chosen job comes from UNC, given that it is a multitasking job.

Solutions Let

U_i = "job is from university i ", $i = 1, 2, 3$ for Duke, UNC, NC State respectively; and

M = "job uses multitasking".

- S1) We want to find $P(M)$. Since U_1, U_2, U_3 form a partition we have

$$\begin{aligned} P(M) &= P(M|U_1)P(U_1) + P(M|U_2)P(U_2) + P(M|U_3)P(U_3) \\ &= 0.01 \times 0.15 + 0.05 \times 0.35 + 0.02 \times 0.5 = 0.029. \end{aligned}$$

- S2) We want to find the conditional probability $P(U_2|M)$.

$$P(U_2|M) = \frac{P(M|U_2)P(U_2)}{P(M)} = \frac{0.05 \times 0.35}{0.029} = 0.603.$$



Example A new HIV test is claimed to correctly identify 95% of people who are really HIV positive and 98% of people who are really HIV negative. Is this acceptable?

If only 1 in a 1000 of the population are HIV positive, what is the probability that someone who tests positive actually has HIV?

Solution: Let

H = "has the HIV virus"; and

T = "test is positive"

We have been given $P(T|H) = 0.95$, $P(\bar{T}|\bar{H}) = 0.98$ and $P(H) = 0.001$. We wish to find $P(H|T)$.

$$\begin{aligned} P(H|T) &= \frac{P(T|H)P(H)}{P(T|H)P(H) + P(T|\bar{H})P(\bar{H})} \\ &= \frac{0.95 \times 0.001}{0.95 \times 0.001 + 0.02 \times 0.999} \\ &= 0.045 \end{aligned}$$

That is, less than 5% of those who test positive really have HIV.

If the HIV test shows a positive result, the individual might wish to retake the test. Suppose that the results of a person retaking the HIV test are conditionally independent given HIV status (clearly two results of the test would certainly not be *unconditionally* independent).

If the test again gives a positive result, what is the probability that the person actually has HIV?

Solution: Let $T_i = "i^{\text{th}} \text{ test is positive}"$.

$$\begin{aligned} P(H|T_1 \cap T_2) &= \frac{P(T_1 \cap T_2|H)P(H)}{P(T_1 \cap T_2)} \\ &= \frac{P(T_1 \cap T_2|H)P(H)}{P(T_1 \cap T_2|H)P(H) + P(T_1 \cap T_2|\bar{H})P(\bar{H})} \\ &= \frac{P(T_1|H)P(T_2|H)P(H)}{P(T_1|H)P(T_2|H)P(H) + P(T_1|\bar{H})P(T_2|\bar{H})P(\bar{H})} \end{aligned}$$

by conditional independence.

Since $P(T_i|H) = 0.95$ and $P(T_i|\bar{H}) = 0.02$,

$$P(H|T_1 \cap T_2) = \frac{0.95 \times 0.95 \times 0.001}{0.95 \times 0.95 \times 0.001 + 0.02 \times 0.02 \times 0.999} \approx 0.693.$$

So almost a 70% chance after taking the test twice and both times showing as positive. For three times, this goes up to 99%. ■

Example *Question:* I divide my emails into 3 categories: $A_1 = \text{"spam"}$, $A_2 = \text{"reply today"}$ and $A_3 = \text{"reply later"}$. From previous experience I find that $P(A_1) = 0.5$, $P(A_2) = 0.1$ and $P(A_3) = 0.4$. Let B be the event that the email contains the word "trial". From previous experience, I find that $P(B|A_1) = 0.9$, $P(B|A_2) = 0.05$ and $P(B|A_3) = 0.05$. I receive an email with the word "trial". What is the probability that it is spam?

Solution: Using Bayes' theorem we have

$$\begin{aligned} P(A_1|B) &= \frac{P(B|A_1)P(A_1)}{P(B)} = \frac{P(B|A_1)P(A_1)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + P(B|A_3)P(A_3)} \\ &= \frac{(0.9)(0.5)}{(0.9)(0.5) + (0.05)(0.1) + (0.05)(0.4)} = \frac{0.45}{0.45 + 0.005 + 0.02} = \frac{0.45}{0.475} \end{aligned}$$

■

Summary of Conditional Probability

1. If $P(F) > 0$ then

$$P(E|F) = \frac{P(E \cap F)}{P(F)}$$

2. $P(\cdot|F)$ satisfies the axioms of probability, for fixed F . However, in general, $P(E|\cdot)$ does not satisfy the axioms of probability, for fixed E .
3. In general, $P(E|F) \neq P(F|E)$.
4. E and F are independent if and only if $P(E|F) = P(E)$.

Chapter 6. Discrete Random Variables

6.1 Random Variables

We are not always interested in an experiment itself, but rather in some consequence of its random outcome. Such consequences, when real valued, may be thought of as functions which map S to \mathbb{R} , and these functions are called random variables.

Definition 6.1.1. A **random variable** is a (measurable) mapping

$$X : S \rightarrow \mathbb{R}$$

with the property that $\{s \in S : X(s) \leq x\} \in \mathcal{F}$ for each $x \in \mathbb{R}$.

If we denote the unknown outcome of the random experiment as s^* , then the corresponding unknown outcome of the random variable $X(s^*)$ will be generically referred to as X .

The probability measure P already defined on S induces a **probability distribution** on the random variable X in \mathbb{R} : For each $x \in \mathbb{R}$, let $S_x \subseteq S$ be the set containing just those elements of S which are mapped by X to numbers no greater than x . Then we see

$$P_X(X \leq x) \equiv P(S_x).$$

Definition 6.1.2. The image of S under X is called the **range** of the random variable:

$$\mathbb{X} \equiv X(S) = \{x \in \mathbb{R} | \exists s \in S \text{ s.t. } X(s) = x\}$$

So as S contains all the possible outcomes of the experiment, \mathbb{X} contains all the possible outcomes for the random variable X .

Example Let our random experiment be tossing a fair coin, with sample space $S = \{H, T\}$ and probability measure $P(\{H\}) = P(\{T\}) = \frac{1}{2}$.

We can define a random variable $X : \{H, T\} \rightarrow \mathbb{R}$ taking values, say,

$$X(T) = 0 \quad \text{and} \quad X(H) = 1$$

In this case, we have

$$S_x = \begin{cases} \emptyset & \text{if } x < 0; \\ \{T\} & \text{if } 0 \leq x < 1; \\ \{H, T\} & \text{if } x \geq 1. \end{cases}$$

This defines a range of probabilities P_X on the continuum \mathbb{R}

$$P_X(X \leq x) = P(S_x) = \begin{cases} P(\emptyset) = 0 & \text{if } x < 0; \\ P(\{T\}) = \frac{1}{2} & \text{if } 0 \leq x < 1; \\ P(\{H, T\}) = 1 & \text{if } x \geq 1. \end{cases}$$

Random variables are important because they provide a compact way of referring to events via their numerical attributes.

Example Consider counting the number of heads in a sequence of 3 coin tosses. The underlying sample space is

$$S = \{TTT, TTH, THT, HTT, THH, HTH, HHT, HHH\}$$

which contains the 8 possible sequences of tosses. However, since we are only interested in the number of heads in each sequence, we define the random variable X by

$$X(s) = \begin{cases} 0, & s = TTT, \\ 1, & s \in \{TTH, THT, HTT\}, \\ 2, & s \in \{HHT, HTH, THH\}, \\ 3, & s = HHH. \end{cases}$$

This mapping is illustrated in Figure 6.1 below

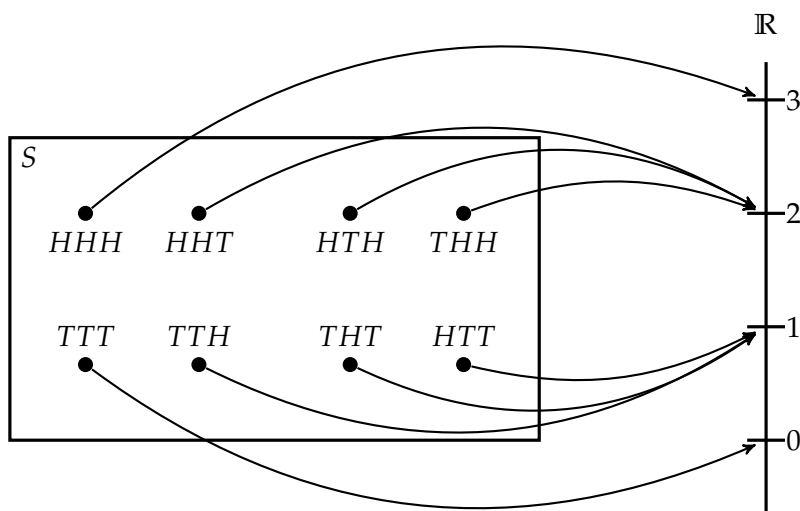


Figure 6.1: Illustration of a random variable X that counts the number of heads in a sequence of 3 coin tosses.

Continuing this example, let us assume that the sequences are equally likely. Now let's find the probability that the number of heads X is less than 2. In other words, we want to find $P_X(X < 2)$... but what does this precisely mean? $P_X(X < 2)$ is the shorthand for

$$P_X(\{s \in S : X(s) < 2\}).$$

The first step in calculating the probability is therefore to identify the event $\{s \in S : X(s) < 2\}$. In Figure 6.1, the only lines pointing to the numbers less than 2 are 0 and 1. Tracing these lines backwards from \mathbb{R} into S , we see that

$$\{s \in S : X(s) < 2\} = \{TTT, TTH, THT, HTT\}.$$

Since we have assumed that the sequences are equally likely

$$P_X(\{TTT, TTH, THT, HTT\}) = \frac{|\{TTT, TTH, THT, HTT\}|}{|S|} = \frac{4}{8} = \frac{1}{2}$$

On the same sample space, we can define another random variable able to describe the event that the number of heads in 3 tosses is even. Define this random variable, Y , as

$$Y(s) = \begin{cases} 0, & s \in \{TTT, THH, HTH, HHT\} \\ 1, & s \in \{TTH, THT, HTT, HHH\} \end{cases}.$$

The probability that the number of heads is less than two and odd is $P(X < 2, Y = 1)$, by which we mean the probability of the event

$$\{s \in S : X(s) < 2 \text{ and } Y(s) = 1\}.$$

This event is equal to

$$\{s \in S : X(s) < 2\} \cap \{s \in S : Y(s) = 1\}$$

which is

$$\{TTT, TTH, THT, HTT\} \cap \{TTH, THT, HTT, HHH\} = \{TTH, THT, HTT\}.$$

The probability of this event, assuming all sequences are equally likely, is $3/8$.

The shorthand introduced above is standard in probability theory. In general, if $B \subset \mathbb{R}$,

$$\{X \in B\} := \{s \in S : X(s) \in B\}$$

and

$$P_X(X \in B) := P_X(\{X \in B\}) = P_X(\{s \in S : X(s) \in B\}).$$

If B is an interval such as $B = (a, b]$,

$$\{X \in (a, b]\} := \{a < X \leq b\} := \{s \in S : a < X(s) \leq b\}$$

and

$$P_X(a < X \leq b) = P_X(\{s \in S : a < X(s) \leq b\}).$$

Analogous notation applies to intervals such as $[a, b]$, $[a, b)$, (a, b) , $(-\infty, b)$, $(-\infty, b]$, (a, ∞) and $[a, \infty)$.

6.1.1 Cumulative Distribution Function

Given a random variable X , we define the cumulative distribution function (CDF or just distribution function) as follows:

Definition 6.1.3. *The cumulative distribution function (CDF) of a random variable X is the function $F_X : \mathbb{R} \rightarrow [0, 1]$, defined by*

$$F_X(x) = P_X(X \leq x)$$

For any random variable X , F_X is right-continuous, meaning if a decreasing sequence of real numbers $x_1, x_2, \dots \rightarrow x$, then $F_X(x_1), F_X(x_2), \dots \rightarrow F_X(x)$.

For a given function $F_X(x)$, to check this is a valid CDF, we need to make sure the following conditions hold.

- i) $0 \leq F_X(x) \leq 1, \forall x \in \mathbb{R}$;
- ii) Monotonicity: $\forall x_1, x_2 \in \mathbb{R}, x_1 < x_2 \Rightarrow F_X(x_1) \leq F_X(x_2)$;
- iii) $F_X(-\infty) = 0, F_X(\infty) = 1$.

Some results regarding CDFs.

- For finite intervals $(a, b] \subseteq \mathbb{R}$, it is easy to check that

$$P_X(a < X \leq b) = F_X(b) - F_X(a).$$

- Unless there is any ambiguity, we generally suppress the subscript of $P_X(\cdot)$ in our notation and just write $P(\cdot)$ for the probability measure for the random variable.
 - That is, we forget about the underlying sample space and use random variable directly and its probabilities.
 - Often, it will be most convenient to work this way and consider the random variable directly from the very start, with the range of X being our sample space.

6.2 Discrete Random Variables

Definition 6.2.1. A random variable X is **discrete** if the range of X , denoted by \mathbb{X} , is countable, that is

$$\mathbb{X} = \{x_1, x_2, \dots, x_n\} \quad (\text{FINITE}) \quad \text{or} \quad \mathbb{X} = \{x_1, x_2, \dots\} \quad (\text{INFINITE}).$$

The even numbers, the odd numbers and the rational numbers are countable; the set of real numbers between 0 and 1 is not countable.

Definition 6.2.2. For a discrete random variable X , we define the **probability mass function** (or **probability function**) as

$$p_X(x) = P(X = x), \quad x \in \mathbb{X}.$$

Note For completeness, we define

$$p_X(x) = 0, \quad x \notin \mathbb{X}.$$

for that p_X is defined for all $x \in \mathbb{R}$. Furthermore, we will refer to \mathbb{X} as the support of random variable X , that is, the set of $x \in \mathbb{R}$ such that $p_X > 0$.

6.2.1 Properties of Mass Function p_X

A function p_X is a probability mass function for a discrete random variable X with range \mathbb{X} of the form $\{x_1, x_2, \dots\}$ if and only if

- i) $p_X(x_i) \geq 0$;
- ii) $\sum_{x \in \mathbb{X}} p_X(x) = 1$.

6.2.2 Discrete Cumulative Distribution Function

The cumulative distribution function, or CDF, F_X of a discrete random variable X is defined by

$$F_X(x) = P(X \leq x), \quad x \in \mathbb{R}.$$

6.2.3 Connection between F_X and p_X

Let X be a discrete random with range $\mathbb{X} = \{x_1, x_2, \dots\}$, where $x_1 < x_2 < \dots$, and probability mass function p_x and CDF F_X . Then, for any real value x , if $x < x_1$, then $F_X(x) = 0$ and for $x \geq x_1$.

$$F_X(x) = \sum_{x_i \leq x} p_X(x_i) \iff p_X(x_i) = F_X(x_i) - F_X(x_{i-1}), \quad i = 2, 3, \dots$$

with, for completeness, $p_X(x_1) = F_X(x_1)$.

6.2.4 Properties of Discrete CDF F_X

i) In the limiting cases,

$$\lim_{x \rightarrow -\infty} F_X(x) = 0, \quad \lim_{x \rightarrow \infty} F_X(x) = 1.$$

ii) F_X is continuous from the right on \mathbb{R} , that is, for $x \in \mathbb{R}$,

$$\lim_{h \rightarrow 0^+} F_X(x + h) = F_X(x)$$

iii) F_X is non-decreasing, that is,

$$a < b \implies F_X(a) \leq F_X(b).$$

iv) For $a < b$

$$P(a < X \leq b) = F_X(b) - F_X(a).$$

Note The key idea is that the functions p_X and/or F_X can be used to describe the probability distribution of the random variable X . A graph of the function p_X is non-zero only at the elements of \mathbb{X} . A graph of the function F_X is a step-function which takes the value zero at minus infinity, the value one at infinity, and is non-decreasing with points of discontinuity at the elements of \mathbb{X} .

Example Consider a coin tossing experiment where a fair coin is tossed repeatedly under identical experimental conditions, with the sequence of tosses independent, until a Head is obtained. For this experiment, the sample space, S , consists of the set of sequences $(\{H\}, \{TH\}, \{TTH\}, \dots)$ with associated probabilities $1/2, 1/4, 1/8, \dots$.

Define the discrete random variable $X : S \rightarrow \mathbb{R}$, by $X(s) = x \iff$ first Head on toss x . Then

$$p_X(x) = P(X = x) = \left(\frac{1}{2}\right)^x, \quad x = 1, 2, 3, \dots$$

and zero otherwise. For $x \geq 1$, let $k(x)$ be the largest integer not greater than x , then

$$F_X(x) = \sum_{x_i < x} p_X(x_i) = \sum_{i=1}^{k(x)} p_X(i) = 1 - \left(\frac{1}{2}\right)^{k(x)}$$

and $F_X(x) = 0$ for $x < 1$.

Figure 6.2 displays the probability mass function (left) and cumulative distribution function (right). Note that the mass function is only non-zero at points that are elements of \mathbb{X} and the CDF is defined for all real values of x , but is only continuous from the right. F_X is therefore a step-function.

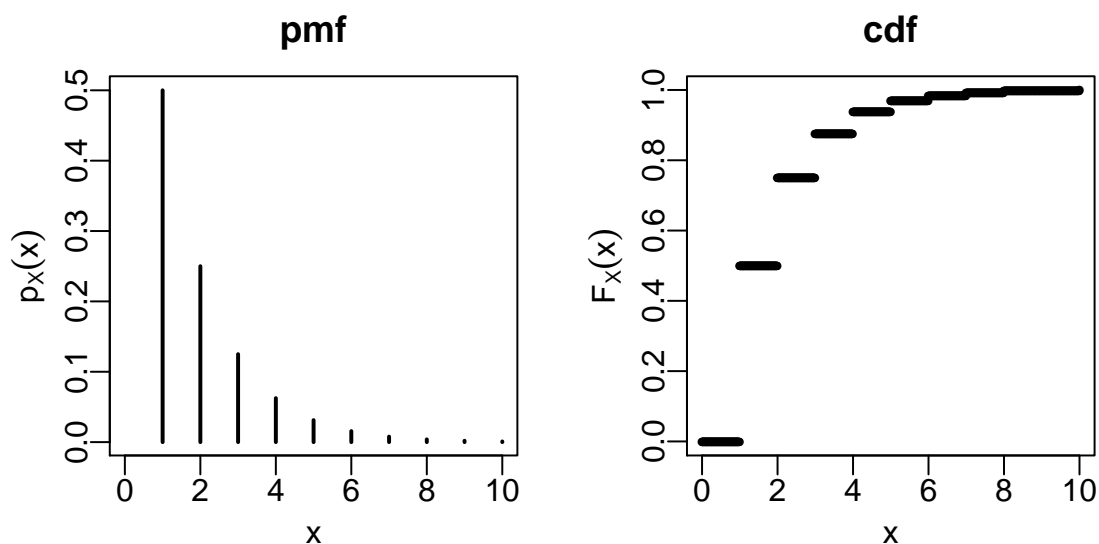


Figure 6.2: pmf $p_X(x) = (\frac{1}{2})^x$, $x = 1, 2, \dots$, and CDF $F_X(x) = 1 - (\frac{1}{2})^x$.

■

We are now starting to see the connections between the numerical summaries and graphical displays we saw in earlier lectures and probability theory.

We can often think of a set of data (x_1, x_2, \dots, x_n) as n realisations of a random variable X defined on an underlying population for the data.

- Recall the frequency counts we considered for a set of data and their histogram plot. This can be seen as an *empirical estimate* for the pmf of their underlying population.
- Also recall the empirical cumulative distribution function. This too is an empirical estimate, but for the CDF of the underlying population.

6.3 Mean and Variance

6.3.1 Expectation

The mean, or expectation, of a discrete random variable is the average value of X .

Definition 6.3.1. The **expected value**, or **mean** of a discrete random variable X is defined to be

$$E_X(X) = \sum_x x p_X(x)$$

The expectation is a one-number summary of the distribution and is often just written $E(X)$ or even μ_X .

$E(X)$ gives a weighted average of the possible values of the random variable X , with the weights given by the probability of that particular outcome.

1. If X is a random variable taking the integer value scored with a single roll of a fair die, then

$$\begin{aligned} E(X) &= \sum_{x=1}^6 x p_X(x) \\ &= 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = \frac{21}{6} = 3.5. \end{aligned}$$

2. If X is a score from a student answering a single multiple choice question with four options, with 3 marks awarded for a correct answer, -1 for a wrong answer and 0 for no answer, what is the expected value if they answer at random?

$$E(X) = 3.P_X(\text{Correct}) + (-1).P_X(\text{Incorrect}) = 3 \cdot \frac{1}{4} - 1 \cdot \frac{3}{4} = 0.$$

Extension: Let $g : \mathbb{R} \rightarrow \mathbb{R}$ be a real-valued (measurable) function of interest of the random variable X ; then we have the following result:

Theorem 6.4.

$$E(g(X)) = \sum_x g(x) p_X(x)$$

Properties of Expectations

Let X be a random variable with pmf p_X . Let g and h be real-valued functions, $g, h : \mathbb{R} \rightarrow \mathbb{R}$, and let a and b be constants. Then

$$E(ag(X) + bh(X)) = aE(g(X)) + bE(h(X))$$

Special Cases:

(i) For a linear function, $g(X) = aX + b$ for constants, we have (from Theorem 6.4) that

$$\begin{aligned} E(g(X)) &= \sum_x (ax + b)p_X(x) \\ &= a \sum_x xp_X(x) + b \sum_x p_X(x) \end{aligned}$$

and since $\sum_x xp_X(x) = E(X)$ and $\sum_x p_X(x) = 1$ we have

$$E(aX + b) = aE(X) + b$$

(ii) Consider $g(x) = (x - E(X))^2$. The expectation of this function wrt P_X gives a measure of spread or variability of the random variable X around its mean, called the **variance**.

Definition 6.4.1. Let X be a random variable. The **variance** of X , denoted by σ^2 or σ_X^2 or $\text{Var}_X(X)$ is defined by

$$\text{Var}_X(X) = E_X[\{X - E_X(X)\}^2].$$

We can expand the expression $\{X - E(X)\}^2$ and exploit the linearity of expectation to get an alternative formula for the variance.

$$\begin{aligned} \{X - E(X)\}^2 &= X^2 - 2E(X)X + \{E(X)\}^2 \\ \implies \text{Var}(X) &= E[X^2 - \{2E(X)\}X + \{E(X)\}^2] \\ &= E(X^2) - 2E(X)E(X) + \{E(X)\}^2 \end{aligned}$$

and hence

$$\text{Var}(X) = E(X^2) - \{E(X)\}^2.$$

It is easy to show that the corresponding result is

$$\text{Var}(aX + b) = a^2\text{Var}(X), \quad \forall a, b \in \mathbb{R}$$

Related to the variance is the standard deviation, which is defined as follows:

Definition 6.4.2. The **standard deviation** of a random variable X , written $sd_X(X)$ (or sometimes σ_X), is the square root of the variance.

$$sd_X(X) = \sqrt{\text{Var}_X(X)}.$$

Lastly, we can define the skewness of a discrete random variable as follows:

Definition 6.4.3. The **skewness** (γ_1) of a discrete random variable X is given by

$$\gamma_1 = \frac{E_X[\{X - E_X(X)\}^3]}{sd_X(X)^3}.$$

Example If X is a random variable taking the integer value scored with a single roll of a fair die, then

$$\begin{aligned} \text{Var}(X) &= E(X^2) - (E(X))^2 \\ &= \sum_{x=1}^6 x^2 p_X(x) - 3.5^2 \\ &= 1^2 \cdot \frac{1}{6} + 2^2 \cdot \frac{1}{6} + \dots + 6^2 \cdot \frac{1}{6} - 3.5^2 = 1.25. \end{aligned}$$

■

Example If X is a score from a student answering a single multiple choice question with four options, with 3 marks awarded for a correct answer, -1 for a wrong answer and 0 for no answer, what is the standard deviation if they answer at random?

$$\begin{aligned} E(X^2) &= 3^2 \cdot P_X(\text{Correct}) + (-1)^2 \cdot P_X(\text{Incorrect}) = 9 \cdot \frac{1}{4} + 1 \cdot \frac{3}{4} = 3 \\ \Rightarrow \text{sd}(X) &= \sqrt{3 - 0^2} = \sqrt{3}. \end{aligned}$$

■

Note We have met three important quantities for a random variable, defined through expectation – the mean μ , the variance σ^2 and the standard deviation σ .

Again we can see a duality with the corresponding numerical summaries for data which we met – the sample mean \bar{x} , the sample variance s^2 and the sample standard deviation s .

The duality is this: If we were to consider the data sample as the *population* and draw a random member from that sample as a *random variable*, this random variable would have CDF $F_n(x)$, the empirical CDF. The mean of the random variable $\mu = \bar{x}$, variance $\sigma^2 = s^2$ and standard deviation $\sigma = s$.

6.4.1 Sums of Random Variables

Let X_1, X_2, \dots, X_n be n random variables, perhaps with different distributions and not necessarily independent.

Let $S_n = \sum_{i=1}^n X_i$ be the sum of those variables, and $\frac{S_n}{n}$ be their average.

Then the mean of S_n is given by

$$E(S_n) = \sum_{i=1}^n E(X_i), \quad E\left(\frac{S_n}{n}\right) = \frac{\sum_{i=1}^n E(X_i)}{n}.$$

However, for the variance of S_n , only if X_1, X_2, \dots, X_n are **independent**, we have

$$\text{Var}(S_n) = \sum_{i=1}^n \text{Var}(X_i), \quad \text{Var}\left(\frac{S_n}{n}\right) = \frac{\sum_{i=1}^n \text{Var}(X_i)}{n^2}.$$

So if X_1, X_2, \dots, X_n are independent and identically distributed with $E(X_i) = \mu_X$ and $\text{Var}(X_i) = \sigma_X^2$ we get

$$E\left(\frac{S_n}{n}\right) = \mu_X, \quad \text{Var}\left(\frac{S_n}{n}\right) = \frac{\sigma_X^2}{n}.$$

6.5 Some Important Discrete Random Variables

6.5.1 Bernoulli Distribution

Consider an experiment with only two possible outcomes, encoded as a random variable X taking value 1, with probability p , or 0, with probability $(1 - p)$, accordingly.

Example Tossing a coin, $X = 1$ for a head, $X = 0$ for tails, $p = \frac{1}{2}$. ■

Then we say $X \sim \text{Bernoulli}(p)$ and note the pmf to be

$$p_X(x) = p^x(1 - p)^{1-x}, \quad x \in \mathbb{X} = \{0, 1\}, \quad 0 \leq p \leq 1$$

Note Using the formulae for mean and variance, it follows that

$$\mu \equiv E(X) = p, \quad \sigma^2 \equiv \text{Var}(X) = p(1 - p).$$

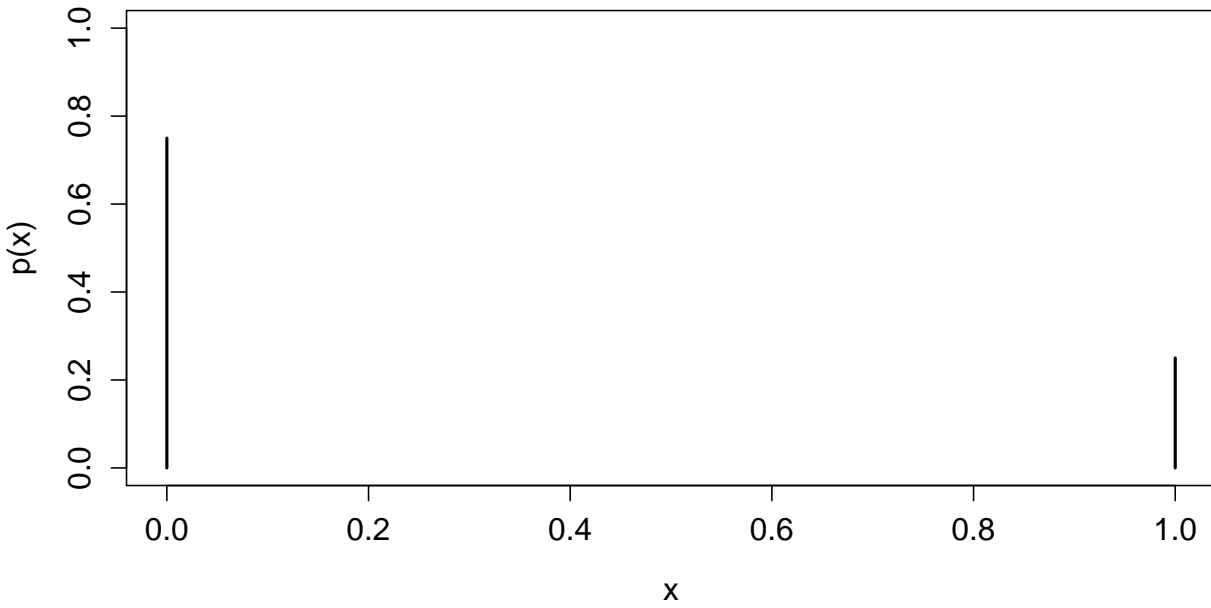


Figure 6.3: Example: pmf of Bernoulli(1/4).

6.5.2 Binomial Distribution

Now consider n identical, independent Bernoulli(p) trials X_1, \dots, X_n . Let $X = \sum_{i=1}^n X_i$ be the total number of 1s observed in the n trials.

Example Tossing a coin n times, X is the number of heads obtained, $p = \frac{1}{2}$. ■

Then X is a random variable taking values in $\{0, 1, 2, \dots, n\}$, and we say $X \sim \text{Binomial}(n, p)$.

From the Binomial Theorem we find the pmf to be

$$p_X(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x \in \mathbb{X} = \{0, 1, 2, \dots, n\}, \quad n \geq 1, \quad 0 \leq p \leq 1.$$

Notes

- To calculate the Binomial pmf we recall that $\binom{n}{x} = \frac{n!}{x!(n-x)!}$ and $x! = \prod_{i=1}^x i$. (Note $0! = 1$.)
- It can be shown, either directly from the pmf or from the results for sums of random variables, that the mean and variance are

$$\mu \equiv E(X) = np, \quad \sigma^2 \equiv \text{Var}(X) = np(1-p).$$

- The skewness is given by

$$\gamma_1 = \frac{1-2p}{\sqrt{np(1-p)}}.$$

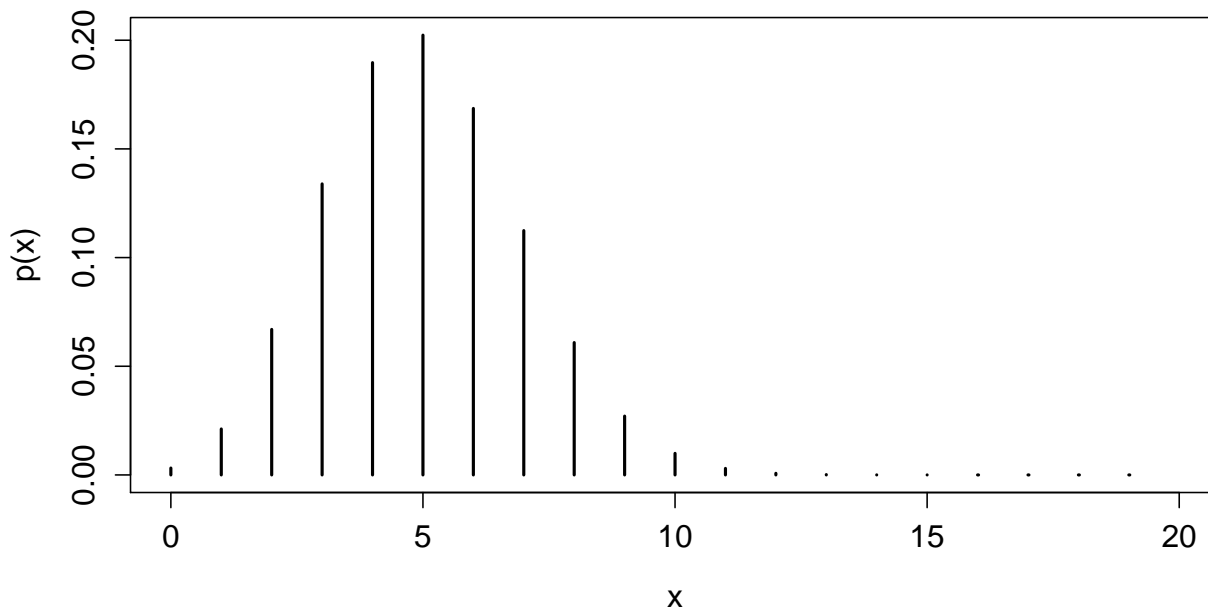


Figure 6.4: Example: pmf of Binomial(20, 1/4).

Example Suppose that 10 users are authorised to use a particular computer system, and that the system collapses if 7 or more users attempt to log on simultaneously. Suppose that each user has the same probability $p = 0.2$ of wishing to log on in each hour.

Question: What is the probability that the system will crash in a given hour?

Solution

The probability that exactly x users will want to log on in any hour is given by $\text{Binomial}(n, p) = \text{Binomial}(10, 0.2)$.

Hence the probability of 7 or more users wishing to log on in any hour is

$$\begin{aligned}
 & p_X(7) + p_X(8) + p_X(9) + p_X(10) \\
 &= \binom{10}{7} 0.2^7 0.8^3 + \dots + \binom{10}{10} 0.2^{10} 0.8^0 \\
 &= 0.00086.
 \end{aligned}$$

- A manufacturing plant produces chips with a defect rate of 10%. The quality control procedure consists of checking samples of size 50. Then the distribution of the number of defectives is expected to be Binomial(50, 0.1).
- When transmitting binary digits through a communication channel, the number of digits received correctly out of n transmitted digits, can be modelled by a Binomial(n, p), where p is the probability that a digit is transmitted incorrectly.

Note The independence condition necessary for these models to be reasonable. ■

6.5.3 Geometric Distribution

Consider a potentially infinite sequence of independent Bernoulli(p) random variables X_1, X_2, \dots . Suppose we define a quantity X by

$$X = \min\{i | X_i = 1\}$$

to be the index of the first Bernoulli trial to result in a 1.

Example Tossing a coin, X is the number of tosses until the first head is obtained, $p = \frac{1}{2}$. ■

Then X is a random variable taking values in $\mathbb{Z}^+ = \{1, 2, \dots\}$, and we say $X \sim \text{Geometric}(p)$.

Clearly the pmf is given by

$$p_X(x) = p(1-p)^{x-1}, \quad x \in \mathbb{X} = \{1, 2, \dots\}, \quad 0 \leq p \leq 1.$$

Notes

- The mean and variance are

$$\mu \equiv E(X) = \frac{1}{p}, \quad \sigma^2 \equiv \text{Var}(X) = \frac{1-p}{p^2}.$$

- The skewness is given by

$$\gamma_1 = \frac{2-p}{\sqrt{1-p}},$$

and so is always positive.

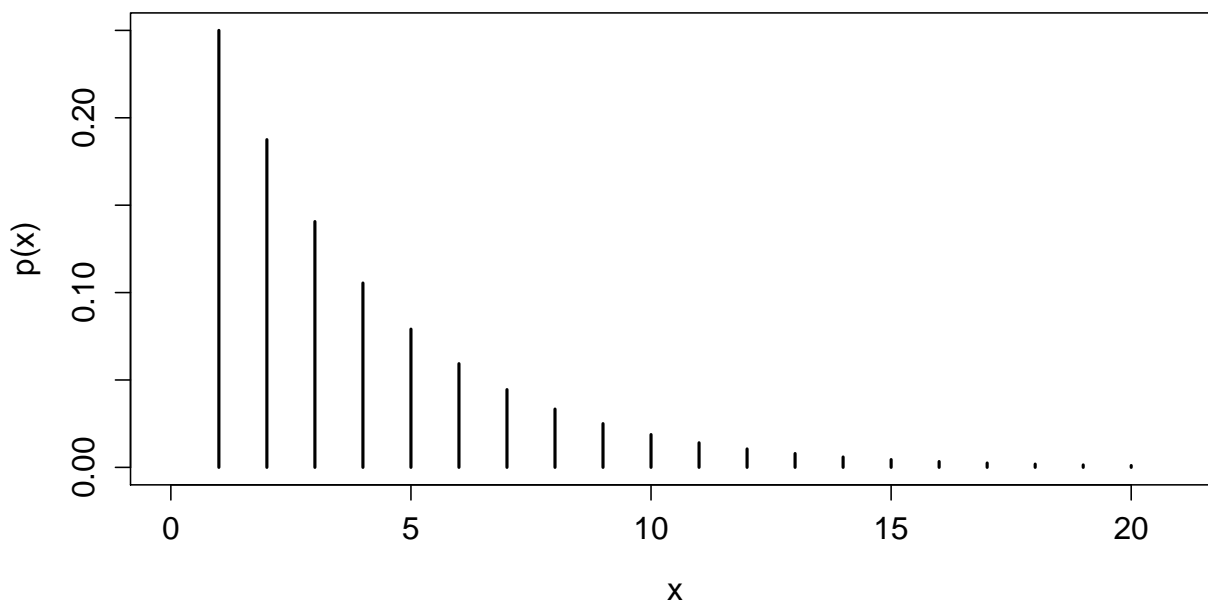


Figure 6.5: Example: pmf of Geometric(1/4).

Alternative Formulation

If $X \sim \text{Geometric}(p)$, let us consider $Y = X - 1$.

Then Y is a random variable taking values in $\mathbb{N} = \{0, 1, 2, \dots\}$, and corresponds to the number of independent Bernoulli(p) trials *before* we obtain our first 1. (Some texts refer to *this* as the Geometric distribution.)

Note we have pmf

$$p_Y(y) = p(1-p)^y, \quad y = 0, 1, 2, \dots,$$

and the mean becomes

$$\mu_Y \equiv E_Y(Y) = \frac{1-p}{p}.$$

while the variance and skewness are unaffected by the shift.

Example Suppose people have problems logging onto a particular website once every 5 attempts, on average.

1. Assuming the attempts are independent, what is the probability that an individual will not succeed until the 4th?

$$p = \frac{4}{5} = 0.8. \quad p_X(4) = (1-p)^3 p = 0.2^3 0.8 = 0.0064.$$

2. On average, how many trials must one make until succeeding?

$$\text{Mean} = \frac{1}{p} = \frac{5}{4} = 1.25.$$

3. What's the probability that the first successful attempt is the 7th or later?

$$p_X(7) + p_X(8) + p_X(9) + \dots = \frac{p(1-p)^6}{1-(1-p)} = (1-p)^6 = 0.2^6.$$

Again suppose that 10 users are authorised to use a particular computer system, and that the system collapses if 7 or more users attempt to log on simultaneously. Suppose that each user has the same probability $p = 0.2$ of wishing to log on in each hour.

Using the Binomial distribution we found the probability that the system will crash in any given hour to be 0.00086.

Using the Geometric distribution formulae, we are able to answer questions such as: On average, after how many hours will the system crash?

$$\text{Mean} = \frac{1}{p} = \frac{1}{0.00086} = 1163 \text{ hours.} \quad \blacksquare$$

Example A dictator, keen to maximise the ratio of males to females in his country (so he could build up his all male army) ordered that each couple should keep having children until a boy was born and then stop.

Calculate the number expected number of boys that a couple will have, and the expected number of girls, given that $P(\text{boy}) = \frac{1}{2}$.

Assume for simplicity that each couple can have arbitrarily many children (although this is not necessary to get the following results). Then since each couple stops when 1 boy is born, the expected number of boys per couple is 1.

On the other hand, if Y is the number of girls given birth to by a couple, Y clearly follows the alternative formulation for the Geometric($\frac{1}{2}$) distribution.

So the expected number of girls for a couple is $\frac{1 - \frac{1}{2}}{\frac{1}{2}} = 1$. ■

6.5.4 Poisson Distribution

Let X be a random variable on $\mathbb{N} = \{0, 1, 2, \dots\}$ with pmf

$$p_X(x) = \frac{e^{-\lambda} \lambda^x}{x!}, \quad x \in \mathbb{X} = \{0, 1, 2, \dots\}, \quad \lambda > 0.$$

Then X is said to follow a Poisson distribution with *rate* parameter λ and we write $X \sim \text{Poisson}(\lambda)$.

Notes

- Poisson random variables are concerned with the number of random events occurring per unit of time or space, when there is a constant underlying probability ‘rate’ of events occurring across this unit.
 - the number of minor car crashes per day in the U.K.;
 - the number of potholes in each mile of road;
 - the number of jobs which arrive at a database server per hour;
 - the number of particles emitted by a radioactive substance in a given time.
- An interesting property of the Poisson distribution is that it has equal mean and variance, namely

$$\mu \equiv E(X) = \lambda, \quad \sigma^2 \equiv \text{Var}(X) = \lambda.$$

- The skewness is given by

$$\gamma_1 = \frac{1}{\sqrt{\lambda}},$$

so is always positive but decreasing as λ increases.

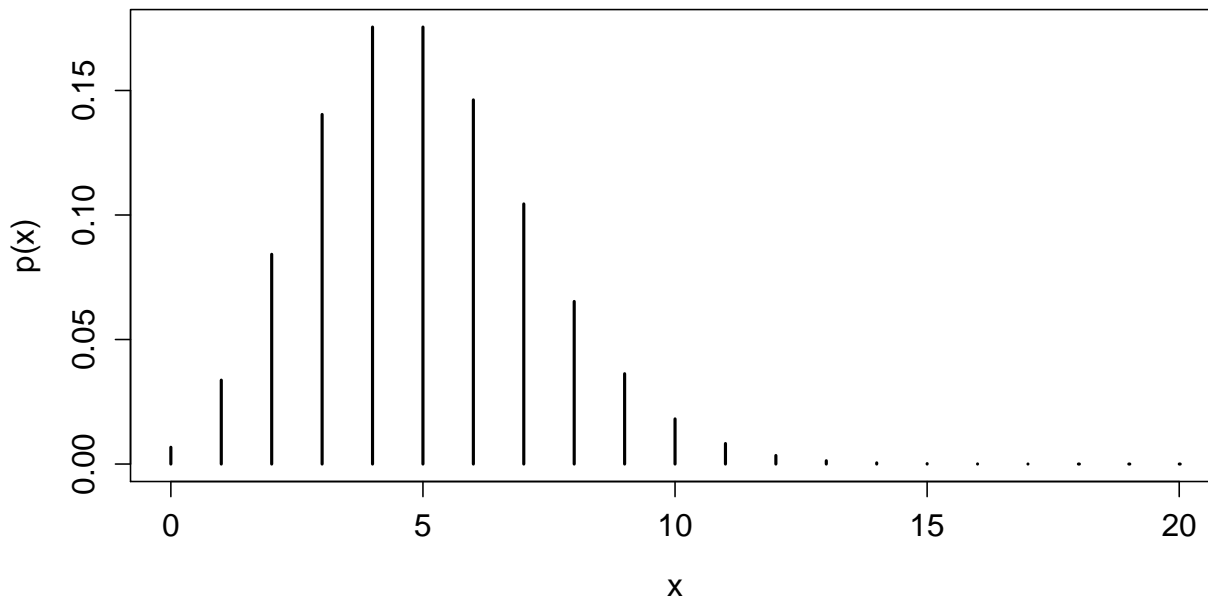


Figure 6.6: Example: pmf of Poisson(5).

Notice the similarity between the pmf plots for Binomial(20, 1/4) and Poisson(5) (Figures 6.4 and 6.6).

It can be shown that for Binomial(n, p), when p is small and n is large, this distribution can be well approximated by the Poisson distribution with rate parameter np , Poisson(np).

The value of p in the above is not small, we would typically prefer $p < 0.1$ for the approximation to be useful.

The usefulness of this approximation is in using probability tables; tabulating a single Poisson(λ) distribution encompasses an infinite number of possible corresponding Binomial distributions, Binomial($n, \frac{\lambda}{n}$).

Example A manufacturer produces VLSI chips, of which 1% are defective. Find the probability that in a box of 100 chips none are defective.

We want $p_X(0)$ from Binomial(100, 0.01). Since n is large and p is small, we can approximate this distribution by Poisson(100×0.01) \equiv Poisson(1).

$$\text{Then } p_X(0) \approx \frac{e^{-1}\lambda^0}{0!} = 0.3679. \quad \blacksquare$$

Example The number of particles emitted by a radioactive substance which reached a Geiger counter was measured for 2608 time intervals, each of length 7.5 seconds.

The (real) data are given in the table below:

x	0	1	2	3	4	5	6	7	8	9	≥ 10
n_x	57	203	383	525	532	408	273	139	45	27	16

Do these data correspond to 2608 independent observations of an identical Poisson random variable?

The total number of particles, $\sum_x xn_x$, is 10,094, and the total number of intervals observed, $n = \sum_x n_x$, is 2608, so that the average number reaching the counter in an interval is $\frac{10094}{2608} = 3.870$.

Since the mean of $\text{Poisson}(\lambda)$ is λ , we can try setting $\lambda = 3.87$ and see how well this fits the data.

For example, considering the case $x = 0$, for a single experiment interval the probability of observing 0 particles would be $p_X(0) = \frac{e^{-3.87}3.87^0}{0!} = 0.02086$. So over $n = 2608$ repetitions, our (Binomial) expectation of the number of 0 counts would be $n \times p_X(0) = 54.4$.

Similarly for $x = 1, 2, \dots$, we obtain the following table of expected values from the $\text{Poisson}(3.87)$ model:

x	0	1	2	3	4	5	6	7	8	9	≥ 10
$O(n_x)$	57	203	383	525	532	408	273	139	45	27	16
$E(n_x)$	54.4	210.5	407.4	525.5	508.4	393.5	253.8	140.3	67.9	29.2	17.1

(O=Observed, E=Expected).

The two sets of numbers appear sufficiently close to suggest the Poisson approximation is a good one. Later, when we come to look at *hypothesis testing*, we will see how to make such judgements quantitatively. ■

6.5.5 Discrete Uniform Distribution

Let X be a random variable on $\{1, 2, \dots, n\}$ with pmf

$$p_X(x) = \frac{1}{n}, \quad x \in \mathbb{X} = \{1, 2, \dots, n\}.$$

Then X is said to follow a discrete uniform distribution and we write $X \sim U(\{1, 2, \dots, n\})$.

Note The mean and variance are

$$\mu \equiv E(X) = \frac{n+1}{2}, \quad \sigma^2 \equiv \text{Var}(X) = \frac{n^2-1}{12}.$$

and the skewness is clearly zero.

Chapter 7. Continuous Random Variables

Suppose again we have a random experiment with sample space S and probability measure P .

Recall our definition of a random variable as a mapping $X : S \rightarrow \mathbb{R}$ from the sample space S to the real numbers inducing a probability measure $P_X(B) = P\{X^{-1}(B)\}, B \subseteq \mathbb{R}$.

Definition 7.0.1. A random variable X is (absolutely) **continuous** if $\exists f_X : \mathbb{R} \rightarrow \mathbb{R}$ (measurable) such that

$$P_X(B) = \int_{x \in B} f_X(x) dx, \quad B \subseteq \mathbb{R},$$

in which case f_X is referred to as the **probability density function**, or **pdf**, of X .

7.0.1 Continuous Cumulative Distribution Function

Definition 7.0.2. The **cumulative distribution function** of CDF, F_X of a continuous random variable X is defined as

$$F_X(x) = P(X \leq x), \quad x \in \mathbb{R}.$$

Note From now on, when we speak of a continuous random variable, we will implicitly assume the absolutely continuous case.

7.0.2 Properties of Continuous F_X and f_X

By analogy with the discrete case, let \mathbb{X} be the range of X , so that $\mathbb{X} = \{x : f_X(x) > 0\}$.

i) For the cdf of a continuous random variable,

$$\lim_{x \rightarrow -\infty} F_X(x) = 0, \quad \lim_{x \rightarrow \infty} F_X(x) = 1.$$

ii) At values of x where F_X is differentiable

$$f_X(x) = \left. \frac{d}{dt} F_X(t) \right|_{t=x} \equiv F'_X(x).$$

iii) If X is continuous,

$$f_X(x) \neq P(X = x) = \lim_{h \rightarrow 0^+} [P(X \leq x) - P(X \leq x - h)] = \lim_{h \rightarrow 0^+} [F_X(x) - F_X(x - h)] = 0$$

Warning! People usually forget, that $P(X = x) = 0$ for all x , when X is a continuous random variable.

iv) The pdf $f_X(x)$ is not itself a probability, then unlike the pmf of a discrete random variable we do not require $f_X(x) \leq 1$.

v) For $a < b$,

$$P(a < X \leq b) = P(a \leq X < b) = P(a \leq X \leq b) = P(a < X < b) = F_X(b) - F_X(a).$$

vi) From Definition 7.0.1 it is clear that the pdf of a continuous random variable X completely characterises its distribution, so we often just specify f_X .

It follows that a function f_X is a pdf for a continuous random variable X if and only if

i) $f_X(x) \geq 0$,

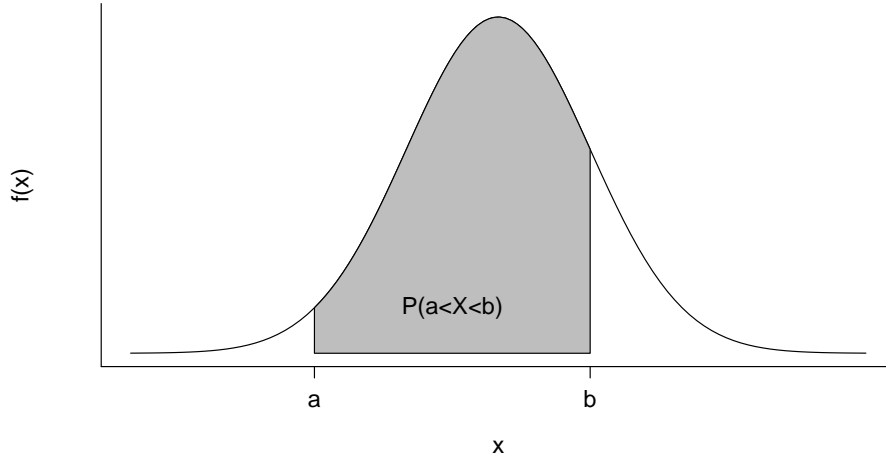
ii) $\int_{-\infty}^{\infty} f_X(x) dx = 1$

This result follows direct from the definitions and properties of F_X .

Suppose we are interested in whether a continuous random variable X lies in an interval $(a, b]$. Well, $P_X(a < X \leq b) = P_X(X \leq b) - P_X(X \leq a)$, which in terms of the cdf and pdf gives

$$\begin{aligned} P_X(a < X \leq b) &= F_X(b) - F_X(a) \\ &= \int_a^b f_X(x) dx. \end{aligned}$$

That is, the area under the pdf between a and b .



Example Consider an experiment to measure the length of time that an electrical component functions before failure. The sample space of outcomes of the experiment, S is \mathbb{R}^+ and if A_x is the event that the component functions longer than $x > 0$ time units, suppose that $P(A_x) = \exp\{-x^2\}$.

Define continuous random variable $X : S \rightarrow \mathbb{R}^+$, by $X(s) = x \iff$ component fails at time x . Then, if $x > 0$

$$F_X(x) = P(X \leq x) = 1 - P(A_x) = 1 - \exp\{-x^2\}$$

and $F_X(x) = 0$ if $x \leq 0$. Hence if $x > 0$,

$$f_X(x) = \left. \frac{d}{dt} F_X(t) \right|_{t=x} = 2x \exp\{-x^2\}$$

and zero otherwise.

Figure 7.1 displays the probability density function (left) and cumulative distribution function (right). Note that both the PDF and CDF are defined for all real values of x , and that both are continuous functions.

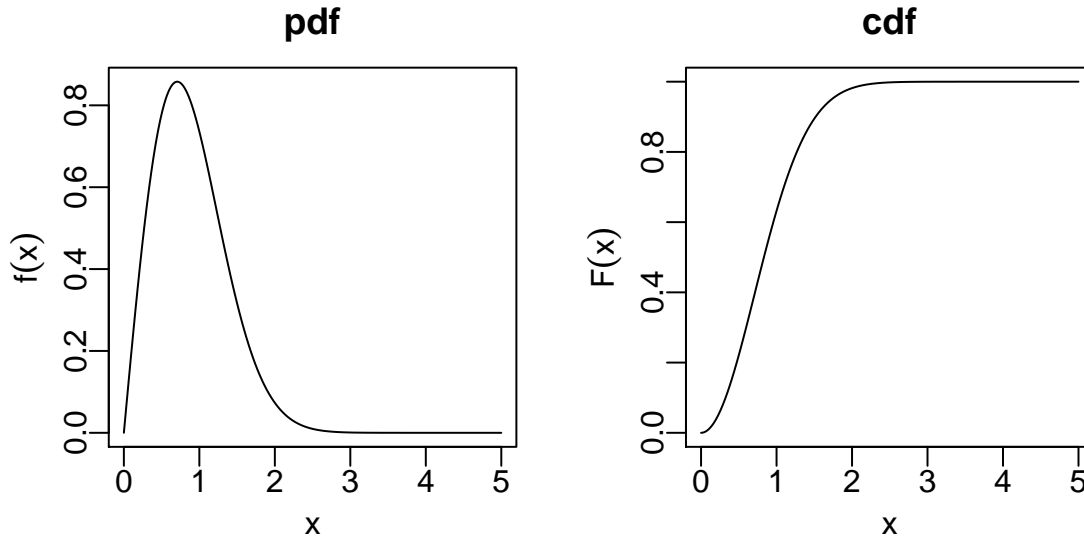


Figure 7.1: PDF $f_X(x) = 2x \exp\{-x^2\}$, $x > 0$, and CDF $F_X(x) = 1 - \exp\{-x^2\}$.

Also note that here

$$F_X(x) = \int_{-\infty}^x f_X(t) dt = \int_0^x f_X(t) dt$$

as $f_X(x) = 0$ for $x \leq 0$, and also that

$$\int_{-\infty}^{\infty} f_X(x) dx = \int_0^{\infty} f_X(x) dx = 1$$

■

Example Suppose we have a continuous random variable X with probability density function given by

$$f_X(x) = \begin{cases} cx^2, & 0 < x < 3 \\ 0, & \text{otherwise} \end{cases}$$

for some unknown constant c .

Questions

Q1) Determine c .

Q2) Find the cdf of X .

Q3) Calculate $P(1 < X < 2)$.

Solutions

S1) We must have

$$1 = \int_0^3 cx^2 dx = c \left[\frac{x^3}{3} \right]_0^3 = 9c$$

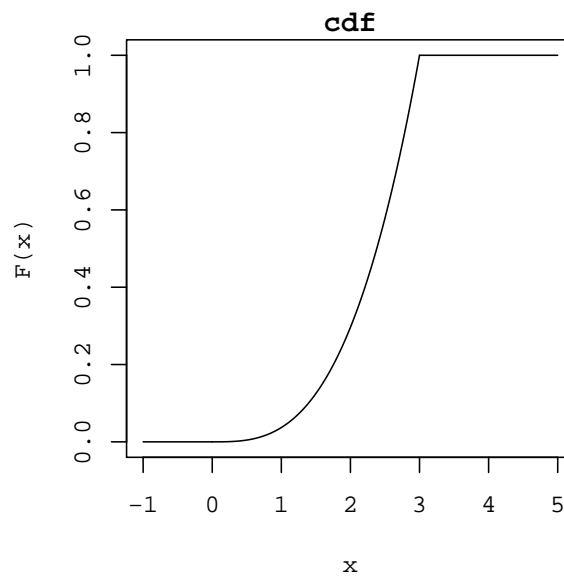
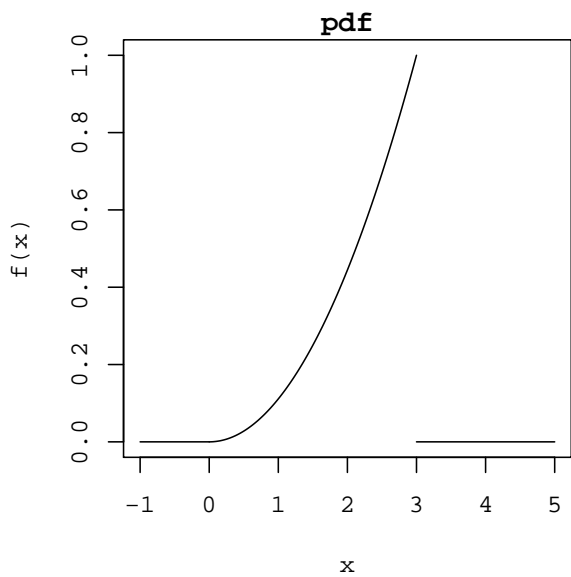
$$\implies c = \frac{1}{9}.$$

S2)

$$F_X(x) = \begin{cases} 0 & x < 0 \\ \int_{-\infty}^x f(u) du = \int_0^x \frac{u^2}{9} du = \frac{x^3}{27} & 0 \leq x \leq 3 \\ 1 & x > 3 \end{cases}$$

S3)

$$P(1 < X < 2) = F(2) - F(1) = \frac{8}{27} - \frac{1}{27} = \frac{7}{27} = 0.2593.$$



7.0.3 Transformations

Suppose that X is a continuous random variable X with pdf f_X and cdf F_X . Let $Y = g(X)$ be a function of X for some (measurable) function $g : \mathbb{R} \rightarrow \mathbb{R}$ s.t. g is continuous and strictly monotonic (so g^{-1} exists). We call $Y = g(X)$ a transformation of X .

Suppose g is monotonic increasing. We can compute the pdf and cdf of $Y = g(X)$ as follows:

The cdf of Y is given by

$$F_Y(y) = P_Y(Y \leq y) = P_Y(g(X) \leq y) = P_X(X \leq g^{-1}(y)) = F_X(g^{-1}(y))$$

The pdf of Y is given by using the chain rule of differentiation:

$$f_Y(y) = F'_Y(y) = f_X\{g^{-1}(y)\}g^{-1'}(y)$$

Note $g^{-1'}(y) = \frac{d}{dy}g^{-1}(y)$ is positive since we assumed g was increasing.

If g monotonic decreasing, we have that

$$F_Y(y) = P_Y(Y \leq y) = P_Y(g(X) \leq y) = P_X(X \geq g^{-1}(y)) = 1 - F_X(g^{-1}(y))$$

By comparison with before, we would have

$$f_Y(y) = F'_Y(y) = -f_X\{g^{-1}(y)\}g^{-1'}(y)$$

with $g^{-1'}(y)$ always negative.

Therefore, for $Y = g(X)$ we have

$$f_Y(y) = f_X\{g^{-1}(y)\}|g^{-1'}(y)|. \quad (7.1)$$

Example Let $f_X(x) = e^{-x}$ for $x > 0$. Hence, $F_X(x) = \int_0^x f_X(u)du = 1 - e^{-x}$. Let $Y = g(X) = \log(X)$. Then

$$g^{-1}(y) = e^y \quad \text{and} \quad g^{-1'}(y) = e^y$$

Then, using (7.1), the pdf of Y is

$$f_Y(y) = e^{-e^y} e^y \quad \text{for } y \in \mathbb{R}.$$

■

7.1 Mean, Variance and Quantiles

7.1.1 Expectation

Definition 7.1.1. For a continuous random variable X we define the **mean or expectation** of X ,

$$\mu_X \text{ or } E_X(X) = \int_{-\infty}^{\infty} x f_X(x) dx.$$

Extension: More generally, for a (measurable) function of interest of the random variable $g : \mathbb{R} \rightarrow \mathbb{R}$ we have

$$E_X\{g(X)\} = \int_{-\infty}^{\infty} g(x) f_X(x) dx.$$

Properties of Expectations

Clearly, for continuous random variables we again have linearity of expectation

$$E(aX + b) = aE(X) + b, \quad \forall a, b \in \mathbb{R},$$

and that for two functions $g, h : \mathbb{R} \rightarrow \mathbb{R}$, we have additivity of expectation

$$E\{g(X) + h(X)\} = E\{g(X)\} + E\{h(X)\}.$$

7.1.2 Variance

Definition 7.1.2. The **variance** of a continuous random variable X is given by

$$\sigma_X^2 \text{ or } \text{Var}_X(X) = E\{(X - \mu_X)^2\} = \int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x) dx.$$

and again it is easy to show that

$$\text{Var}_X(X) = \int_{-\infty}^{\infty} x^2 f_X(x) dx - \mu_X^2 = E(X^2) - \{E(X)\}^2.$$

For a linear transformation $aX + b$ we again have

$$\text{Var}(aX + b) = a^2 \text{Var}(X), \quad \forall a, b \in \mathbb{R}.$$

7.1.3 Quantiles

Recall we defined the lower and upper quartiles and median of a sample of data as points (1/4, 3/4, 1/2)-way through the ordered sample. This idea can be generalised as follows:

Definition 7.1.3. For a (continuous) random variable X we define the α -quantile $Q_X(\alpha)$, $0 \leq \alpha \leq 1$ to satisfy $P(X \leq Q_X(\alpha)) = \alpha$,

$$Q_X(\alpha) = F_X^{-1}(\alpha).$$

In particular the **median** of a random variable X is $F_X^{-1}\left(\frac{1}{2}\right)$. That is, the solution to the equation $F_X(x) = \frac{1}{2}$.

Example Suppose we have a continuous random variable X with probability density function given by

$$f_X(x) = \begin{cases} x^2/9, & 0 < x < 3 \\ 0, & \text{otherwise.} \end{cases}$$

Questions

Q1) Calculate $E(X)$.

Q2) Calculate $\text{Var}(X)$.

Q3) Calculate the median of X .

Solutions

S1)

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx = \int_0^3 x \cdot \frac{x^2}{9} dx = \frac{x^4}{36} \Big|_0^3 = \frac{3^4}{36} = 2.25$$

S2)

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x) dx = \int_0^3 x^2 \cdot \frac{x^2}{9} dx = \frac{x^5}{45} \Big|_0^3 = \frac{3^5}{45} = 5.4$$

$$\text{So } \text{Var}(X) = E(X^2) - \{E(X)\}^2 = 5.4 - 2.25^2 = 0.3375$$

S3) From earlier, $F(x) = \frac{x^3}{27}$, for $0 < x < 3$.

Setting $F(x) = \frac{1}{2}$ and solving, we get

$$\frac{x^3}{27} = \frac{1}{2} \iff x = \sqrt[3]{\frac{27}{2}} = \frac{3}{\sqrt[3]{2}} \approx 2.3811$$

for the median.

■

7.2 Some Important Continuous Random Variables

7.2.1 Continuous Uniform Distribution

Suppose X is a continuous random variable with probability density function

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & a < x < b \\ 0, & \text{otherwise,} \end{cases}$$

Then X is said to follow a uniform distribution on the interval (a, b) and we write $X \sim U(a, b)$.

Notes

- The cdf is

$$F_X(x) = \begin{cases} 0 & x \leq a \\ \frac{x-a}{b-a} & a < x < b \\ 1 & x \geq b \end{cases}$$

- The case $a = 0$ and $b = 1$ is referred to as the **Standard uniform**.

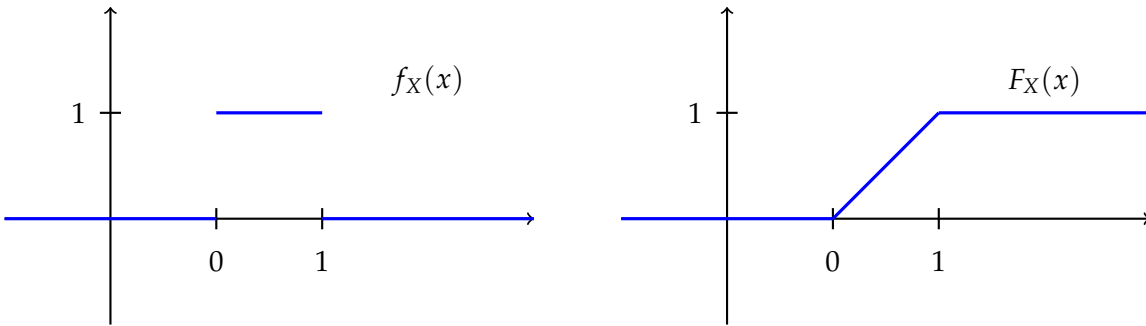


Figure 7.2: PDF and CDF of a standard uniform distribution.

- Suppose $X \sim U(0,1)$, so $F_X(x) = x$, $0 \leq x \leq 1$. We wish to map the interval $(0,1)$ to the general interval (a,b) , where $a < b \in \mathbb{R}$. So we define a new random variable $Y = a + (b-a)X$, so $a < Y < b$.

We first observe that for any $y \in (a,b)$,

$$Y \leq y \iff a + (b-a)X \leq y \iff X \leq \frac{y-a}{b-a}.$$

From this we find $Y \sim U(a,b)$, since

$$F_Y(y) = P(Y \leq y) = P\left(X \leq \frac{y-a}{b-a}\right) = F_X\left(\frac{y-a}{b-a}\right) = \frac{y-a}{b-a}.$$

- To find the mean of $X \sim U(a,b)$,

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx = \int_a^b x \cdot \frac{1}{b-a} dx = \left[\frac{x^2}{2(b-a)} \right]_a^b$$

$$= \frac{b^2 - a^2}{2(b-a)} = \frac{(b-a)(b+a)}{2(b-a)} = \frac{a+b}{2}.$$

Similarly we get $\text{Var}(X) = E(X^2) - E(X)^2 = \frac{(b-a)^2}{12}$, so

$$\mu = \frac{a+b}{2}, \quad \sigma^2 = \frac{(b-a)^2}{12}.$$

7.2.2 Exponential Distribution

Suppose now X is a random variable taking value on $\mathbb{R}^+ = [0, \infty)$ with pdf

$$f_X(x) = \lambda e^{-\lambda x}, \quad x \geq 0,$$

for some $\lambda > 0$.

Then X is said to follow an exponential distribution with *rate* parameter λ and we write $X \sim \text{Exp}(\lambda)$.

Notes

- The cdf is

$$F_X(x) = 1 - e^{-\lambda x}, \quad x > 0.$$

- An alternative representation uses $\theta = 1/\lambda$ as the parameter of the distribution. This is sometimes used because the expectation and variance of the Exponential distributions are

$$E(X) = \frac{1}{\lambda} = \theta, \quad \text{Var}(X) = \frac{1}{\lambda^2}.$$

- If $X \sim \text{Exp}(\lambda)$, then, for all $x, t > 0$,

$$P(X > x+t | X > t) = \frac{P(X > x+t \cap X > t)}{P(X > t)} = \frac{P(X > x+t)}{P(X > t)} = \frac{e^{-\lambda(x+t)}}{e^{-\lambda t}} = e^{-\lambda x} = P(X > x).$$

Thus, for all $x, t > 0$, $P(X > x+t | X > t) = P(X > x)$ — this is known as the **Lack of Memory Property**, and is unique to the exponential distribution amongst continuous distributions.

Interpretation: So if we think of the exponential variable as the time to an event, then knowledge that we have waited time s for the event tells us nothing about how much longer we will have to wait — the process has *no memory*.

- Exponential random variables are often used to model the time until occurrence of a random event where there is an assumed constant risk (λ) of the event happening over time, and so are frequently used as a simplest model, for example, in reliability analysis. So examples include:

- the time to failure of a component in a system;
- the time until we find the next mistake on my slides;
- the distance we travel along a road until we find the next pothole;

- the time until the next jobs arrives at a database server;

Notice the duality between some of the exponential random variable examples and those we saw for a Poisson distribution. In each case, “number of events” has been replaced with “time between events”.

Claim:

If events in a random process occur according to a Poisson distribution with rate λ then the time between events has an Exponential distribution with rate parameter λ .

Proof. Suppose we have some random event process such that $\forall x > 0$, the number of events occurring in $[0, x]$, N_x , follows a Poisson distribution with rate parameter λ , so $N_x \sim \text{Poisson}(\lambda x)$. Such a process is known as an *homogeneous Poisson process*. Let X be the time until the first event of this process arrives.

Then we notice that

$$\begin{aligned} P(X > x) &\equiv P(N_x = 0) \\ &= \frac{(\lambda x)^0 e^{-\lambda x}}{0!} \\ &= e^{-\lambda x}. \end{aligned}$$

and hence $X \sim \text{Exp}(\lambda)$. The same argument applies for all subsequent inter-arrival times. \square

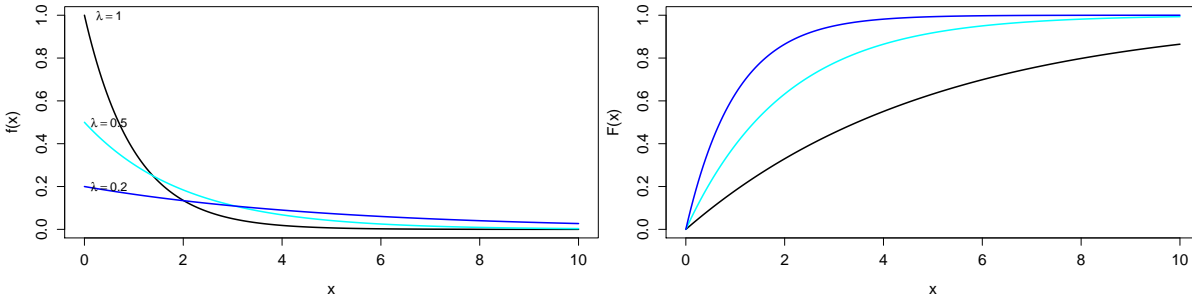


Figure 7.3: PDFs and CDFs for Exponential distribution with different rate parameters.

7.2.3 Normal Distribution

Suppose X is a random variable taking value on \mathbb{R} with pdf

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

for some $\mu \in \mathbb{R}, \sigma > 0$. Then X is said to follow a Gaussian or normal distribution with mean μ and variance σ^2 , and we write $X \sim N(\mu, \sigma^2)$.

Notes

- The cdf of $X \sim N(\mu, \sigma^2)$ is not analytically tractable for any (μ, σ) , so we can only write

$$F_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x \exp\left\{-\frac{(t-\mu)^2}{2\sigma^2}\right\} dt.$$

- Special Case: If $\mu = 0$ and $\sigma^2 = 1$, then X has a **standard** or **unit** normal distribution. The pdf of the standard normal distribution is written as $\phi(x)$ and simplifies to

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}x^2\right\}.$$

Also, the cdf of the standard normal distribution is written as $\Phi(x)$. Again, for the cdf, we can only write

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt.$$

- If $X \sim N(0, 1)$, and

$$Y = \sigma X + \mu$$

then $Y \sim N(\mu, \sigma^2)$. Re-expressing this result: if $X \sim N(\mu, \sigma^2)$ and $Y = (X - \mu)/\sigma$, then $Y \sim N(0, 1)$. This is an important result as it allows us to write the cdf of any normal distribution in terms of Φ : If $X \sim N(\mu, \sigma^2)$ and we set $Y = \frac{X - \mu}{\sigma}$, then since $\sigma > 0$ we can first observe that for any $x \in \mathbb{R}$,

$$\begin{aligned} X \leq x &\iff \frac{X - \mu}{\sigma} \leq \frac{x - \mu}{\sigma} \\ &\iff Y \leq \frac{x - \mu}{\sigma}. \end{aligned}$$

Therefore we can write the cdf of X in terms of Φ ,

$$\begin{aligned} F_X(x) &= P(X \leq x) = P\left(Y \leq \frac{x - \mu}{\sigma}\right) \\ &= \Phi\left(\frac{x - \mu}{\sigma}\right). \end{aligned}$$

- Since the cdf, and therefore any probabilities, associated with a normal distribution are not analytically available, numerical integration procedures are used to find approximate probabilities. In particular, statistical tables contain values of the standard normal cdf $\Phi(z)$ for a range of values $z \in \mathbb{R}$, and the quantiles $\Phi^{-1}(\alpha)$ for a range of values $\alpha \in (0, 1)$. Linear interpolation is used for approximation between the tabulated values. As seen in the point above, all normal distribution probabilities can be related back to probabilities from a standard normal distribution.
- Table 7.1 is an example of a statistical table for the standard normal distribution.

z	$\Phi(z)$	z	$\Phi(z)$	z	$\Phi(z)$	z	$\Phi(z)$
0	0.5	0.9	0.816	1.8	0.964	2.8	0.997
0.1	0.540	1.0	0.841	1.9	0.971	3.0	0.998
0.2	0.579	1.1	0.864	2.0	0.977	3.5	0.9998
0.3	0.618	1.2	0.885	2.1	0.982	1.282	0.9
0.4	0.655	1.3	0.903	2.2	0.986	1.645	0.95
0.5	0.691	1.4	0.919	2.3	0.989	1.96	0.975
0.6	0.726	1.5	0.933	2.4	0.992	2.326	0.99
0.7	0.758	1.6	0.945	2.5	0.994	2.576	0.995
0.8	0.788	1.7	0.955	2.6	0.995	3.09	0.999

Table 7.1

Notice that $\Phi(z)$ has been tabulated for $z > 0$.

This is because the standard normal pdf ϕ is *symmetric* about 0, that is, $\phi(-z) = \phi(z)$. For the cdf Φ , this means

$$\Phi(z) = 1 - \Phi(-z).$$

So, for example, $\Phi(-1.2) = 1 - \Phi(1.2) \approx 1 - 0.885 = 0.115$.

Similarly, if $Z \sim N(0,1)$ and we want, for example, $P(Z > 1.5) = 1 - P(Z \leq 1.5) = 1 - \Phi(1.5) (= \Phi(-1.5))$.

So more generally we have

$$P(Z > z) = \Phi(-z).$$

We will often have cause to use the 97.5% and 99.5% quantiles of $N(0,1)$, given by $\Phi^{-1}(0.975)$ and $\Phi^{-1}(0.995)$.

$$\Phi(1.96) \approx 97.5\%.$$

So with 95% probability an $N(0,1)$ random variable will lie in $[-1.96, 1.96]$ ($\approx [-2, 2]$).

$$\Phi(2.58) = 99.5\%.$$

So with 99% probability an $N(0,1)$ random variable will lie in $[-2.58, 2.58]$.

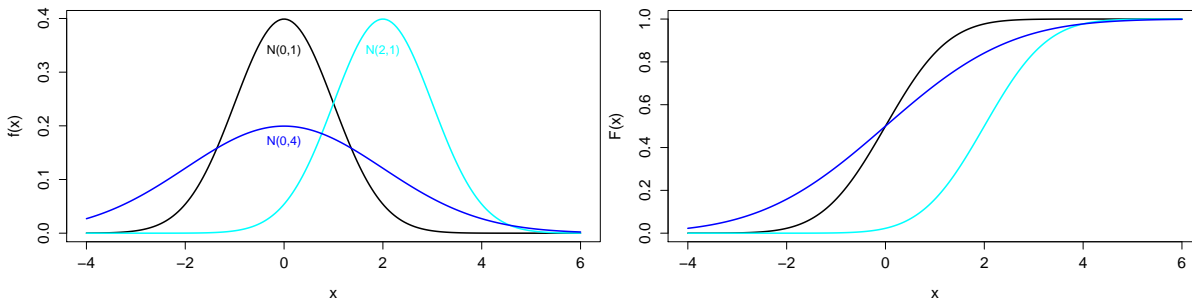


Figure 7.4: PDFs and CDFs of normal distributions with different means and variances.

Example An analogue signal received at a detector (measured in microvolts) may be modelled as a Gaussian random variable $X \sim N(200, 256)$.

Questions

Q1) What is the probability that the signal will exceed $240\mu V$?

Q2) What is the probability that the signal is larger than $240\mu V$ given that it is greater than $210\mu V$?

Solutions

S1)

$$P(X > 240) = 1 - P(X \leq 240) = 1 - \Phi\left(\frac{240 - 200}{\sqrt{256}}\right) = 1 - \Phi(2.5) \approx 0.00621.$$

S2)

$$P(X > 240 | X > 210) = \frac{P(X > 240)}{P(X > 210)} = \frac{1 - \Phi\left(\frac{240 - 200}{\sqrt{256}}\right)}{1 - \Phi\left(\frac{210 - 200}{\sqrt{256}}\right)} \approx 0.02335.$$

■

Let X_1, X_2, \dots, X_n be n independent and identically distributed (i.i.d.) random variables from **any** probability distribution, each with mean μ and variance σ^2 .

From before we know

$$E\left(\sum_{i=1}^n X_i\right) = n\mu, \quad \text{Var}\left(\sum_{i=1}^n X_i\right) = n\sigma^2.$$

First notice

$$E\left(\sum_{i=1}^n X_i - n\mu\right) = 0, \quad \text{Var}\left(\sum_{i=1}^n X_i - n\mu\right) = n\sigma^2.$$

Dividing by $\sqrt{n}\sigma$, we obtain

$$E\left(\frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma}\right) = 0, \quad \text{Var}\left(\frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma}\right) = 1.$$

Theorem 7.3 (Central Limit Theorem or CLT).

$$\lim_{n \rightarrow \infty} \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma} \sim \Phi.$$

This can also be written as

$$\lim_{n \rightarrow \infty} \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \Phi, \quad \text{where } \bar{X} = \frac{\sum_{i=1}^n X_i}{n}.$$

Or finally, for large n we have approximately

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right),$$

or

$$\sum_{i=1}^n X_i \sim N(n\mu, n\sigma^2).$$

We note here that although all these approximate distributional results hold irrespective of the distribution of the $\{X_i\}$, in the special case where $X_i \sim N(\mu, \sigma^2)$ these distributional results are, in fact, exact. This is because the sum of independent normally distributed random variables is also normally distributed.

Example Consider the most simple example, that X_1, X_2, \dots are i.i.d. Bernoulli(p) discrete random variables taking value 0 or 1.

Then the $\{X_i\}$ have mean $\mu = p$ and variance $\sigma^2 = p(1 - p)$. Then, by definition, we know that for any n we have

$$\sum_{i=1}^n X_i \sim \text{Binomial}(n, p).$$

which has mean np and variance $np(1 - p)$.

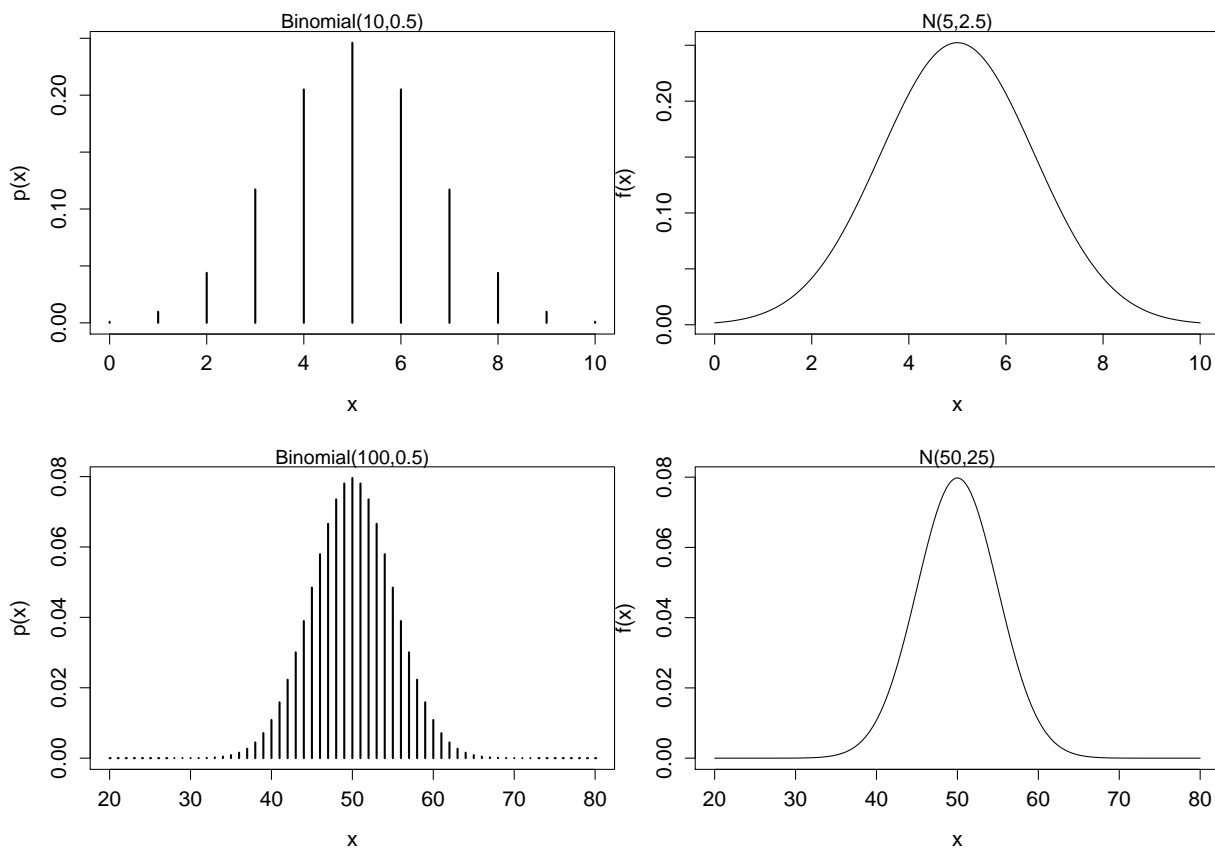
But now, by the Central Limit Theorem (CLT), we also have for large n that approximately:

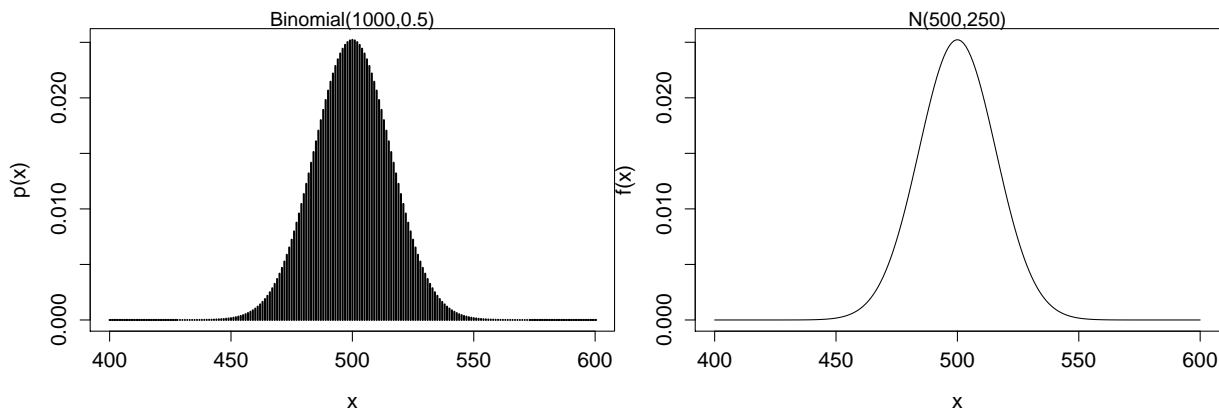
$$\sum_{i=1}^n X_i \sim N(n\mu, n\sigma^2) \equiv N(np, np(1 - p)).$$

So for large n

$$\text{Binomial}(n, p) \approx N(np, np(1 - p)).$$

Notice that the LHS is a discrete distribution, and the RHS is a continuous distribution.





■

Example Suppose X was the number of heads found on 1000 tosses of a fair coin, and we were interested in $P(X \leq 490)$.

Using the binomial distribution pmf, we would need to calculate

$$P(X \leq 490) = p_X(0) + p_X(1) + p_X(2) + \dots + p_X(490) (\approx 0.27).$$

However, using the CLT we have approximately $X \sim N(500, 250)$ and so

$$P(X \leq 490) \approx \Phi\left(\frac{490 - 500}{\sqrt{250}}\right) = \Phi(-0.632) = 1 - \Phi(0.632) \approx 0.26.$$

■

Example Suppose $X \sim N(\mu, \sigma^2)$, and consider the transformation $Y = e^X$.

Then if $g(x) = e^x$, $g^{-1}(y) = \log(y)$ and $g^{-1}'(y) = \frac{1}{y}$.

Then by (7.1) we have

$$f_Y(y) = \frac{1}{\sigma y \sqrt{2\pi}} \exp\left[-\frac{\{\log(y) - \mu\}^2}{2\sigma^2}\right], \quad y > 0,$$

and we say Y follows a **log-normal** distribution.

■

Chapter 8. Jointly Distributed Random Variables

Suppose we have two random variables X and Y defined on a sample space S with probability measure $P(E), E \subseteq S$.

Note that S could be the set of outcomes from two ‘experiments’, and the sample space points be two-dimensional; then perhaps X could relate to the first experiment, and Y to the second.

Then from before we know to define the *marginal* probability distributions P_X and P_Y by, for example,

$$P_X(B) = P(X^{-1}(B)), \quad B \subseteq \mathbb{R}.$$

We now define the **joint probability distribution**:

Definition 8.0.1. *Given a pair of random variables, X and Y , we define the **joint probability distribution** P_{XY} as follows:*

$$P_{XY}(B_X, B_Y) = P\{X^{-1}(B_X) \cap Y^{-1}(B_Y)\}, \quad B_X, B_Y \subseteq \mathbb{R}.$$

So $P_{XY}(B_X, B_Y)$, the probability that $X \in B_X$ **and** $Y \in B_Y$, is given by the probability of the set of all points in the sample space that get mapped **both** into B_X by X **and** into B_Y by Y .

More generally, for a single region $B_{XY} \subseteq \mathbb{R}^2$, find the collection of sample space elements

$$S_{XY} = \{s \in S \mid (X(s), Y(s)) \in B_{XY}\}$$

and define

$$P_{XY}(B_{XY}) = P(S_{XY}).$$

8.0.1 Joint Cumulative Distribution Function

We define the joint cumulative distribution as follows:

Definition 8.0.2. *Given a pair of random variables, X and Y , the **joint cumulative distribution function** is defined as*

$$F_{XY}(x, y) = P_{XY}(X \leq x, Y \leq y), \quad x, y \in \mathbb{R}.$$

It is easy to check that the marginal cdfs for X and Y can be recovered by

$$F_X(x) = F_{XY}(x, \infty), \quad x \in \mathbb{R},$$

$$F_Y(y) = F_{XY}(\infty, y), \quad y \in \mathbb{R},$$

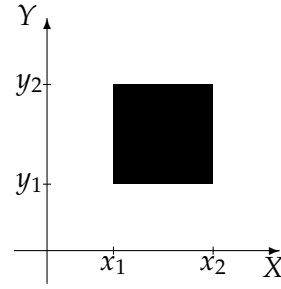
and that the two definitions will agree.

8.0.2 Properties of Joint CDF F_{XY}

For F_{XY} to be a valid cdf, we need to make sure the following conditions hold.

1. $0 \leq F_{XY}(x, y) \leq 1, \forall x, y \in \mathbb{R};$
2. Monotonicity: $\forall x_1, x_2, y_1, y_2 \in \mathbb{R},$
 $x_1 < x_2 \Rightarrow F_{XY}(x_1, y_1) \leq F_{XY}(x_2, y_1)$ and $y_1 < y_2 \Rightarrow F_{XY}(x_1, y_1) \leq F_{XY}(x_1, y_2);$
3. $\forall x, y \in \mathbb{R},$
 $F_{XY}(x, -\infty) = 0, F_{XY}(-\infty, y) = 0$ and $F_{XY}(\infty, \infty) = 1.$

Suppose we are interested in whether the random variable pair (X, Y) lie in the interval cross product $(x_1, x_2] \times (y_1, y_2]$; that is, if $x_1 < X \leq x_2$ and $y_1 < Y \leq y_2$.



First note that $P_{XY}(x_1 < X \leq x_2, Y \leq y) = F_{XY}(x_2, y) - F_{XY}(x_1, y).$

It is then easy to see that $P_{XY}(x_1 < X \leq x_2, y_1 < Y \leq y_2)$ is given by

$$F_{XY}(x_2, y_2) - F_{XY}(x_1, y_2) - F_{XY}(x_2, y_1) + F_{XY}(x_1, y_1).$$

8.0.3 Joint Probability Mass Functions

Definition 8.0.3. If X and Y are both discrete random variables, then we can define the **joint probability mass function** as

$$p_{XY}(x, y) = P_{XY}(X = x, Y = y), \quad x, y \in \mathbb{R}.$$

We can recover the marginal pmfs p_X and p_Y since, by the law of total probability, $\forall x, y \in \mathbb{R}$

$$p_X(x) = \sum_y p_{XY}(x, y), \quad p_Y(y) = \sum_x p_{XY}(x, y).$$

Properties of Joint PMFs

For p_{XY} to be a valid pmf, we need to make sure the following conditions hold.

1. $0 \leq p_{XY}(x, y) \leq 1, \forall x, y \in \mathbb{R};$
2. $\sum_y \sum_x p_{XY}(x, y) = 1.$

8.0.4 Joint Probability Density Functions

On the other hand, if $\exists f_{XY} : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ s.t.

$$P_{XY}(B_{XY}) = \int_{(x,y) \in B_{XY}} f_{XY}(x,y) dx dy, \quad B_{XY} \subseteq \mathbb{R} \times \mathbb{R},$$

then we say X and Y are **jointly continuous** and we refer to f_{XY} as the **joint probability density function** of X and Y .

In this case, we have

$$F_{XY}(x,y) = \int_{t=-\infty}^y \int_{s=-\infty}^x f_{XY}(s,t) ds dt, \quad x,y \in \mathbb{R},$$

Definition 8.0.4. By the Fundamental Theorem of Calculus we can identify the joint pdf as

$$f_{XY}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{XY}(x,y).$$

Furthermore, we can recover the marginal densities f_X and f_Y :

$$\begin{aligned} f_X(x) &= \frac{d}{dx} F_X(x) = \frac{d}{dx} F_{XY}(x, \infty) \\ &= \frac{d}{dx} \int_{y=-\infty}^{\infty} \int_{s=-\infty}^x f_{XY}(s,y) ds dy. \end{aligned}$$

By the Fundamental Theorem of Calculus, and through a symmetric argument for Y , we thus get

$$f_X(x) = \int_{y=-\infty}^{\infty} f_{XY}(x,y) dy, \quad f_Y(y) = \int_{x=-\infty}^{\infty} f_{XY}(x,y) dx.$$

Properties of Joint PDFs

For f_{XY} to be a valid pdf, we need to make sure the following conditions hold.

1. $f_{XY}(x,y) \geq 0, \forall x,y \in \mathbb{R}$;
2. $\int_{y=-\infty}^{\infty} \int_{x=-\infty}^{\infty} f_{XY}(x,y) dx dy = 1$.

8.1 Independence and Expectation

8.1.1 Independence

Two random variables X and Y are **independent if and only if** $\forall B_X, B_Y \subseteq \mathbb{R}$,

$$P_{XY}(B_X, B_Y) = P_X(B_X)P_Y(B_Y).$$

More specifically,

Definition 8.1.1. Two random variables X and Y are independent **if and only if**

$$f_{XY}(x, y) = f_X(x)f_Y(y), \quad \forall x, y \in \mathbb{R}.$$

Definition 8.1.2. For two random variables X, Y we define the **conditional probability distribution** $P_{Y|X}$ by

$$P_{Y|X}(B_Y|B_X) = \frac{P_{XY}(B_X, B_Y)}{P_X(B_X)}, \quad B_X, B_Y \subseteq \mathbb{R}.$$

This is the revised probability of Y falling inside B_Y given that we now know $X \in B_X$.

Then we have X and Y are independent $\iff P_{Y|X}(B_Y|B_X) = P_Y(B_Y), \forall B_X, B_Y \subseteq \mathbb{R}$.

Definition 8.1.3. For random variables X, Y we define the **conditional probability density function** $f_{Y|X}$ by

$$f_{Y|X}(y|x) = \frac{f_{XY}(x, y)}{f_X(x)}, \quad x, y \in \mathbb{R}.$$

Note The random variables X and Y are independent $\iff f_{Y|X}(y|x) = f_Y(y), \forall x, y \in \mathbb{R}$.

8.1.2 Expectation

Suppose we have a (measurable) bivariate function of interest of the random variables X and Y , $g: \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$.

Definition 8.1.4. If X and Y are discrete, we define $E\{g(X, Y)\}$ by

$$E_{XY}\{g(X, Y)\} = \sum_y \sum_x g(x, y) p_{XY}(x, y).$$

Definition 8.1.5. If X and Y are jointly continuous, we define $E\{g(X, Y)\}$ by

$$E_{XY}\{g(X, Y)\} = \int_{y=-\infty}^{\infty} \int_{x=-\infty}^{\infty} g(x, y) f_{XY}(x, y) dx dy.$$

Immediately from these definitions we have the following:

- If $g(X, Y) = g_1(X) + g_2(Y)$,

$$E_{XY}\{g_1(X) + g_2(Y)\} = E_X\{g_1(X)\} + E_Y\{g_2(Y)\}.$$

- If $g(X, Y) = g_1(X)g_2(Y)$ and X and Y are **independent**,

$$E_{XY}\{g_1(X)g_2(Y)\} = E_X\{g_1(X)\}E_Y\{g_2(Y)\}.$$

In particular, considering $g(X, Y) = XY$ for independent X, Y we have

$$E_{XY}(XY) = E_X(X)E_Y(Y).$$

8.1.3 Conditional Expectation

Warning! In general $E_{XY}(XY) \neq E_X(X)E_Y(Y)$.

Suppose X and Y are discrete random variables with joint pmf $p(x, y)$. If we are given the value x of the random variable X , our revised pmf for Y is the conditional pmf $p(y|x)$, for $y \in \mathbb{R}$.

Definition 8.1.6. The **conditional expectation** of Y given $X = x$ is therefore

$$E_{Y|X}(Y|X = x) = \sum_y y p(y|x).$$

Similarly,

Definition 8.1.7. If X and Y were continuous,

$$E_{Y|X}(Y|X = x) = \int_{y=-\infty}^{\infty} y f(y|x) dy.$$

In either case, the conditional expectation is a function of x but not the unknown Y .

For a single variable X we considered the expectation of $g(X) = (X - \mu_X)(X - \mu_X)$, called the variance and denoted σ_X^2 .

The bivariate extension of this is the expectation of $g(X, Y) = (X - \mu_X)(Y - \mu_Y)$. We define the **covariance** of X and Y by

$$\sigma_{XY} = \text{Cov}(X, Y) = E_{XY}[(X - \mu_X)(Y - \mu_Y)].$$

Covariance measures how the random variables move in tandem with one another, and so is closely related to the idea of correlation.

Definition 8.1.8. We define the **correlation** of X and Y by

$$\rho_{XY} = \text{Cor}(X, Y) = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}.$$

Unlike the covariance, the correlation is invariant to the scale of the random variables X and Y .

It is easily shown that if X and Y are independent random variables, then $\sigma_{XY} = \rho_{XY} = 0$.

8.2 Examples

Example Suppose that the lifetime, X , and brightness, Y of a light bulb are modelled as continuous random variables. Let their joint pdf be given by

$$f(x, y) = \lambda_1 \lambda_2 e^{-\lambda_1 x - \lambda_2 y}, \quad x, y > 0.$$

Question Are lifetime and brightness independent?

Solution If the lifetime and brightness are independent we would have

$$f(x, y) = f(x)f(y) \quad \text{for all } x, y \in \mathbb{R}.$$

The marginal pdf for X is

$$\begin{aligned} f(x) &= \int_{y=-\infty}^{\infty} f(x, y) dy = \int_{y=0}^{\infty} \lambda_1 \lambda_2 e^{-\lambda_1 x - \lambda_2 y} dy \\ &= \lambda_1 e^{-\lambda_1 x}. \end{aligned}$$

Similarly $f(y) = \lambda_2 e^{-\lambda_2 y}$. Hence $f(x, y) = f(x)f(y)$ and X and Y are independent. ■

Example Suppose continuous random variables $(X, Y) \in \mathbb{R}^2$ have joint pdf

$$f(x, y) = \begin{cases} 1, & |x| + |y| < 1/\sqrt{2} \\ 0, & \text{otherwise.} \end{cases}$$

Question Determine the marginal pdfs for X and Y .

Solution

We have $|x| + |y| < 1/\sqrt{2} \iff |y| < 1/\sqrt{2} - |x|$. So

$$f(x) = \int_{y=-(\frac{1}{\sqrt{2}}-|x|)}^{\frac{1}{\sqrt{2}}-|x|} dy = \sqrt{2} - 2|x|.$$

Similarly $f(y) = \sqrt{2} - 2|y|$. Hence $f(x, y) \neq f(x)f(y)$ and X and Y are not independent. ■

Chapter 9. Estimation

In statistics we typically analyse a set of data by considering it as a random sample from a larger, underlying population about which we wish to make inference.

1. Chapter 3 on numerical summaries considered various summary sample statistics for describing a particular sample of data. We defined quantities such as the sample mean \bar{x} , and sample variance s^2 .
2. Chapters 6 and 7 on random variables, on the other hand, were concerned with characterising the underlying population. We defined corresponding population parameters such as the population mean $E(X)$, and population variance $\text{Var}(X)$.

We noticed a duality between the two sets of definitions of statistics and parameters.

In particular, we saw that they were equivalent in the extreme circumstance that our sample exactly represented the entire population. Away from this extreme circumstance, the sample statistics can be seen to give approximate values for the corresponding population parameters. We can use them as **estimates**.

For convenient modelling of populations, we met several simple parameterised probability distributions e.g. $\text{Poisson}(\lambda)$, $\text{Exp}(\lambda)$, $U(a, b)$, $N(\mu, \sigma^2)$. There, population parameters such as mean and variance are functions of the distribution parameters. So more generally, we may wish to use the data, or just their sample statistics, to estimate distribution parameters.

For a sample of data $\underline{x} = (x_1, \dots, x_n)$, we can consider these observed values as realisations of corresponding random variables $\underline{X} = (X_1, \dots, X_n)$.

If the underlying population, from which the sample has been drawn, is such that the distribution of a single random draw X has probability distribution $P_{X|\theta}(\cdot|\theta)$, where θ is a generic parameter or vector of parameters, we typically then assume that our n data point random variables \underline{X} are i.i.d. $P_{X|\theta}(\cdot|\theta)$.

Suppose I was interested in the ages of people on this course. Then I could ask every person in this population their age, and thus calculate the population mean μ , variance σ^2 , etc.

More realistically I might just, randomly select people, say 20, in the class and ask them their age. That is, collect a sample of 20 observations from the population, and thus calculate a sample mean \bar{x} , variance s^2 , etc.

If this sampling had to be done *with replacement*, then we know by the CLT that the sample mean \bar{x} is a random variable whose distribution is well approximated by a normal distribution with mean equal to the population mean μ and variance one twentieth of the population variance σ^2 .

Note If the sampling were done *without replacement* (which is preferable), the samples would be (marginally) identically distributed but not independent.

Suppose we have a random variable X (representing a random draw from our underlying population), s.t. $X \sim \text{Binomial}(10, p)$, where the probability parameter p is unknown.

Then suppose we are able to draw a sample of size 100 from the population, that is, observe 100 independent $\text{Binomial}(10, p)$ random variables, and that we observe the data in the table below.

x	0	1	2	3	4	5	6	7	8	9	10
Freq.	2	16	35	22	21	3	1	0	0	0	0

Using these data, how might we estimate p ?

9.1 Estimators

Consider a sequence of random variables $\underline{X} = (X_1, \dots, X_n)$ corresponding to n i.i.d. data samples to be drawn from a population with distribution P_X . Let $\underline{x} = (x_1, \dots, x_n)$ be the corresponding realised values we observe for these random variables.

Definition 9.1.1. A **statistic** is a function $T = T(X_1, \dots, X_n) = T(\underline{X})$, and is itself a random variable.

For example, $\bar{X} = \sum_{i=1}^n X_i / n$ is a statistic. The corresponding realised value of a statistic, e.g. \bar{x} , is written $t = t(\underline{x})$.

If a statistic $T(\underline{X})$ is to be used to approximate parameters of the distribution $P_{X|\theta}(\cdot|\theta)$, we say T is an estimator for those parameters; we call the actual realised value of the estimator for a particular data sample, $t(\underline{x})$, an estimate.

9.1.1 Point Estimates

A **point estimate** is a statistic estimating a single parameter or characteristic of a distribution.

For a running example which we will return to, consider a sample of data (x_1, \dots, x_n) from an $\text{Exponential}(\lambda)$ distribution with unknown λ ; we might construct a point estimate for either λ itself, or perhaps for the mean of the distribution ($= \lambda^{-1}$), or the variance ($= \lambda^{-2}$).

Concentrating on the mean of the distribution in this example, we could propose simply the first data point we observed, X_1 as our point estimator; or we might use the sample mean, \bar{X} ; or, if the data had been given to us already ordered we might (lazily) suggest the median, $X_{(\{n+1\}/2)}$.

Suppose for a moment we actually knew the parameter values θ of our population distribution $P_{X|\theta}(\cdot|\theta)$ (so we know λ in our exponential example).

Then since our sampled data are considered to be i.i.d. realisations from this distribution (so each $X_i \sim \text{Exp}(\lambda)$), it follows that any statistic $T = T(X_1, \dots, X_n)$ is also a random variable with some distribution which also only depends on these parameters.

If we are able to (approximately) identify this sampling distribution of our statistic, call it $P_{T|\theta}$, we can then find the expectation, variance, etc of our statistic.

Sometimes $P_{T|\theta}$, will have a convenient closed-form expression which we can derive, but in other cases it will not.

In those other cases, provided that our sample size n is large, we can at least use the CLT to give us an approximate distribution for $P_{T|\theta}$ if T is the sample mean. Whatever the form of $P_{X|\theta}$, we know that approximately $\bar{X} \sim N(E[X], \text{Var}[X]/n)$.

For our $X_i \sim \text{Exp}(\lambda)$ example, it can be shown that our statistic $T = \bar{X}$ is a continuous random variable with pdf

$$f_{T|\lambda}(t|\lambda) = \frac{(n\lambda)^n t^{n-1} e^{-n\lambda t}}{(n-1)!}, \quad t > 0.$$

This is the pdf of a $\text{Gamma}(n, n\lambda)$ random variable, a well known continuous random variable distribution, $T \sim \text{Gamma}(n, n\lambda)$.

So using the fact that $\text{Gamma}(\alpha, \beta)$ has expectation $\frac{\alpha}{\beta}$, we have

$$E(\bar{X}) = E_{T|\lambda}(T|\lambda) = \frac{n}{n\lambda} = \frac{1}{\lambda},$$

the same as the mean of our population distribution, $E(X)$.

9.1.2 Bias, Efficiency and Consistency

The previous result suggests that \bar{X} is, at least one respect, a good statistic for estimating the unknown mean of an exponential distribution.

Formally, we define the bias of an estimator T for a parameter θ ,

$$\text{bias}(T) = E(T|\theta) - \theta.$$

If, as in the exponential distribution example above (where $\theta = \lambda^{-1}$), our estimator has zero bias we say the estimator is unbiased. So in our example, \bar{x} gives an unbiased estimate of the mean of an exponential distribution.

In fact, this is true for any distribution; the sample mean \bar{x} will always be an unbiased estimate for the population mean μ :

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

Similarly, there is an estimator for the population variance σ^2 which is unbiased, irrespective of the population distribution. This estimator is not the sample variance

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

as this has one too many degrees of freedom.

Note If we knew the population mean μ , then $\frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2$ would be unbiased for σ^2 .

However, we can instead define the **bias-corrected sample variance**,

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

which is then always an unbiased estimator of the population variance σ^2 .

Warning! Because of its usefulness as an unbiased estimate of σ^2 , many statistical text books and software packages (and indeed your formula sheet for the exam) refer to $s_{n-1}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$ as the sample variance.

Suppose we have two unbiased estimators for a parameter θ , which we will call $\hat{\Theta}(\underline{X})$ and $\tilde{\Theta}(\underline{X})$. And again suppose we have the corresponding sampling distributions for these estimators, $P_{\hat{\Theta}|\theta}$ and $P_{\tilde{\Theta}|\theta}$, and so can calculate their means, variances, etc.

Then we say $\hat{\Theta}$ is more efficient than $\tilde{\Theta}$ if:

1. $\forall \theta, \text{Var}_{\hat{\Theta}|\theta}(\hat{\Theta}|\theta) \leq \text{Var}_{\tilde{\Theta}|\theta}(\tilde{\Theta}|\theta);$
2. $\exists \theta \text{ s.t. } \text{Var}_{\hat{\Theta}|\theta}(\hat{\Theta}|\theta) < \text{Var}_{\tilde{\Theta}|\theta}(\tilde{\Theta}|\theta).$

That is, the variance of $\hat{\Theta}$ is never higher than that of $\tilde{\Theta}$, no matter what the true value of θ is; and for some value of θ , $\hat{\Theta}$ has a strictly lower variance than $\tilde{\Theta}$.

If $\hat{\Theta}$ is more efficient than any other possible estimator, we say $\hat{\Theta}$ is efficient.

Suppose we have a population with mean μ and variance σ^2 , from which we are to obtain a random sample X_1, \dots, X_n . Consider two estimators for μ , $\hat{M} = \bar{X}$, the sample mean, and $\tilde{M} = X_1$, the first observation in the sample.

We have seen $E(\bar{X}) = \mu$ always, and certainly $E(X_1) = \mu$, so both estimators are unbiased. We also know $\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$, and of course $\text{Var}(X_1) = \sigma^2$, independent of μ . So if $n \geq 2$, \hat{M} is more efficient than \tilde{M} as an estimator of μ .

In the previous example, the worst aspect of the estimate $\tilde{M} = X_1$ is that it does not change, let alone improve, no matter how large a sample n of data is collected. In contrast, the variance of $\hat{M} = \bar{X}$ gets smaller and smaller as n increases.

Technically, we say an estimator $\hat{\Theta}$ is a consistent estimator for the parameter θ if $\hat{\Theta}$ converges in probability to θ . That is, $\forall \epsilon > 0, P(|\hat{\Theta} - \theta| > \epsilon) \rightarrow 0$ as $n \rightarrow \infty$.

This is hard to demonstrate, but if $\hat{\Theta}$ is unbiased we do have:

$$\lim_{n \rightarrow \infty} \text{Var}(\hat{\Theta}) = 0 \Rightarrow \hat{\Theta} \text{ is consistent.}$$

So returning to our example, we see \bar{X} is a consistent estimator of μ for any underlying population.

9.1.3 Maximum Likelihood Estimation

Recall the Binomial(10, p) example, where p was unknown.

We asked how we might propose an estimator for p . Since then, we have met different criteria for measuring the relative quality of rival estimators, but no principled manner for deriving these estimates.

There are many ways of deriving estimators, but we shall concentrate on just one — maximum likelihood estimation.

If the underlying population is a discrete distribution with an unknown parameter θ , then each of the samples X_i are i.i.d. with probability mass function $p_{X|\theta}(x_i)$.

Since the n data samples are independent, the joint probability of all of the data, $\underline{x} = (x_1, \dots, x_n)$, is

$$L(\theta|\underline{x}) = \prod_{i=1}^n p_{X|\theta}(x_i).$$

The function $L(\theta|\underline{x})$ is called the likelihood function and is considered as a function of the parameter θ for a fixed sample of data \underline{x} . $L(\theta|\underline{x})$ is the probability of observing the data we have, \underline{x} , if the true parameter were θ .

If, on the other hand, the underlying population is a continuous distribution with an unknown parameter θ , then each of the samples X_i are i.i.d. with probability density function $f_{X|\theta}(x_i)$, and the likelihood function is defined by

$$L(\theta|\underline{x}) = \prod_{i=1}^n f_{X|\theta}(x_i).$$

Clearly, for a fixed set of data, varying the population parameter θ would give different probabilities of observing these data, and hence different likelihoods.

Maximum likelihood estimation seeks to find the parameter value $\hat{\theta}_{MLE}$ which maximises the likelihood function,

$$\hat{\theta}_{MLE} = \underset{\theta}{\operatorname{argmax}} L(\theta|\underline{x}).$$

This value $\hat{\theta}_{MLE}$ is known as the maximum likelihood estimate (MLE).

For maximising the likelihood function, it often proves more convenient to consider the log-likelihood, $\ell(\theta|\underline{x}) = \log\{L(\theta|\underline{x})\}$. Since $\log(\cdot)$ is a monotonic increasing function, the argument θ maximising ℓ maximises L .

The log-likelihood function can be written as

$$\ell(\theta|\underline{x}) = \sum_{i=1}^n \log\{p_{X|\theta}(x_i)\} \quad \text{or} \quad \ell(\theta|\underline{x}) = \sum_{i=1}^n \log\{f_{X|\theta}(x_i)\},$$

for discrete and continuous distributions respectively.

In either case, finding $\hat{\theta}$ that solves $\frac{\partial}{\partial \theta} \ell(\hat{\theta}) = 0$ yields the MLE if $\frac{\partial^2}{\partial \theta^2} \ell(\hat{\theta}) < 0$.

Example Continuing the Binomial question... each of our Binomial(10, p) samples X_i have pmf

$$p_X(x_i) = \binom{10}{x_i} p^{x_i} (1-p)^{10-x_i}, \quad i = 1, 2, \dots, 100.$$

Since the $n = 100$ data samples are assumed independent, the likelihood function for p for all of the data is

$$\begin{aligned} L(p|\underline{x}) &= L(p) = \prod_{i=1}^n p_X(x_i) = \prod_{i=1}^n \left\{ \binom{10}{x_i} p^{x_i} (1-p)^{10-x_i} \right\} \\ &= \left\{ \prod_{i=1}^n \binom{10}{x_i} \right\} p^{\sum_{i=1}^n x_i} (1-p)^{10n - \sum_{i=1}^n x_i}. \end{aligned}$$

So the log-likelihood is given by

$$\ell(p) = \log \left\{ \prod_{i=1}^n \binom{10}{x_i} \right\} + \log(p) \sum_{i=1}^n x_i + \log(1-p) \left(10n - \sum_{i=1}^n x_i \right).$$

Next, we differentiate $\ell(p)$

$$\frac{\partial}{\partial p} \ell(p) = 0 + \frac{\sum_{i=1}^n x_i}{p} - \frac{10n - \sum_{i=1}^n x_i}{1-p}.$$

Setting this derivative equal to zero, we get

$$\begin{aligned} \frac{\sum_{i=1}^n x_i}{\hat{p}} - \frac{10n - \sum_{i=1}^n x_i}{1 - \hat{p}} &= 0 \Rightarrow (1 - \hat{p}) \sum_{i=1}^n x_i = \hat{p} \left(10n - \sum_{i=1}^n x_i \right) \\ &\Rightarrow \sum_{i=1}^n x_i = \hat{p} \left(10n - \sum_{i=1}^n x_i + \sum_{i=1}^n x_i \right) \\ &\Rightarrow \hat{p} = \frac{\sum_{i=1}^n x_i}{10n} = \frac{\bar{x}}{10}. \end{aligned}$$

To check this point is a maximum of ℓ , we find the second derivative

$$\frac{\partial^2}{\partial p^2} \ell(p) = -\frac{\sum_{i=1}^n x_i}{p^2} - \frac{10n - \sum_{i=1}^n x_i}{(1-p)^2} = -\frac{n\bar{x}}{p^2} - \frac{10n - n\bar{x}}{(1-p)^2} = -n \left(\frac{\bar{x}}{p^2} + \frac{10 - \bar{x}}{(1-p)^2} \right)$$

(which is in fact $< 0 \forall p$, the likelihood is *log concave*).

Substituting $\hat{p} = \frac{\bar{x}}{10}$, this gives

$$-100n \left(\frac{1}{\bar{x}} + \frac{1}{10 - \bar{x}} \right) = -\frac{1000n}{(10 - \bar{x})\bar{x}},$$

which is clearly < 0 . So the MLE for p is $\hat{p} = \frac{\bar{x}}{10} = 0.257$.

Example If $X \sim N(\mu, \sigma^2)$, then

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp \left\{ -\frac{(x - \mu)^2}{2\sigma^2} \right\}$$

For an i.i.d. sample $\underline{x} = (x_1, \dots, x_n)$, the likelihood function for (μ, σ^2) for all of the data is

$$L(\mu, \sigma^2) = \prod_{i=1}^n f_X(x_i) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp \left\{ -\frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2} \right\}.$$

The log likelihood is

$$\ell(\mu, \sigma^2) = -\frac{n}{2} \log(2\pi) - n \log(\sigma) - \frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2}.$$

For the MLE for μ , we can take the partial derivative wrt μ and set this equal to zero.

$$0 = \frac{\sum_{i=1}^n (x_i - \hat{\mu})}{\sigma^2} \iff 0 = \sum_{i=1}^n (x_i - \hat{\mu}) = \sum_{i=1}^n x_i - n\hat{\mu} \iff \hat{\mu} = \frac{\sum_{i=1}^n x_i}{n} = \bar{x}.$$

To check this is a maximum, we look at the second derivative.

$$\frac{\partial^2}{\partial \mu^2} \ell(\mu, \sigma^2) = -\frac{n}{\sigma^2},$$

which is negative everywhere, so \bar{x} is the MLE for μ , independently from the value of σ^2 . ■

Finding the MLE

In general, we have the following procedure to find MLEs.

1. Write down the likelihood function, $L(\theta)$ where

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

that is, the product of the n mass/density functions viewed as a function of θ .

2. Take the natural log of the likelihood, and collect terms involving θ .
3. Find the value of θ for which log-likelihood is maximised. This is typically done by finding $\hat{\theta}$ that solves

$$\frac{\partial}{\partial \theta} \ell(\hat{\theta}) = \frac{\partial}{\partial \theta} \log(L(\hat{\theta})) = 0$$

4. Check that the estimate $\hat{\theta}$ obtained in step 3 corresponds to a maximum in the (log) likelihood function by inspecting the second derivative of $\ell(\hat{\theta})$ wrt θ . If

$$\frac{\partial^2}{\partial \theta^2} \ell(\hat{\theta}) < 0$$

at $\theta = \hat{\theta}$, then $\hat{\theta}$ is confirmed as the MLE of θ .

CLT significance

We have already seen that \bar{X} is always an unbiased estimator for the population mean μ . And now, using the CLT, we also have that for large n that \bar{X} is always approximately the MLE for the population mean μ , irrespective of the distribution of X .

So how good an estimator of θ is the MLE?

- The MLE is not necessarily unbiased.
- + For large n , the MLE is approximately normally distributed with mean θ .
- + The MLE is consistent.
- + The MLE is always asymptotically efficient, and if an efficient estimator exists, it is the MLE.
- + Because it is derived from the likelihood of the data, it is well-principled. This is the “likelihood principle”, which asserts that all the information about a parameter from a set of data is held in the likelihood.

9.2 Confidence Intervals

In most circumstances, it will not be sufficient to report simply a point estimate $\hat{\theta}$ for an unknown parameter θ of interest. We would usually want to also quantify our degree of uncertainty in this estimate.

If we were again to suppose we had knowledge of the true value of our unknown parameter θ , or at least had access to the (approximate) true sampling distribution of our statistic, $P_{T|\theta}$, then the variance of this distribution would give such a measure.

The solution we consider is to plug in our known estimated value of θ , $\hat{\theta}$, into $P_{T|\theta}$ and hence use the (maybe further) approximated sampling distribution, $P_{T|\hat{\theta}}$.

In particular, we know by the CLT that for any underlying distribution (mean μ , variance σ^2) for our sample, the sample mean \bar{X} is approximately normally distributed with mean μ and variance $\frac{\sigma^2}{n}$. We now further approximate this by imagining $\bar{X} \sim N(\bar{x}, \frac{\sigma^2}{n})$.

Then, if we knew the true population variance σ^2 , we would be able to say that *had the true mean parameter μ been \bar{x}* , then for large n , from our standard normal tables, with 95% probability we would have observed our statistic \bar{X} within the interval

$$\left[\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}} \right].$$

This is known as the 95% confidence interval for μ .

More generally, for any desired *coverage* probability level $1 - \alpha$ we can define the the $100(1 - \alpha)\%$ **confidence interval** for μ by

$$\left[\bar{x} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right]$$

where z_α is the α -quantile of the standard normal (so before we used $\alpha = 0.05$ and hence $z_{0.975}$ to obtain our 95% C.I.).

Loose interpretation: Amongst all the possible intervals $\bar{x} \pm z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ we might have observed (that is, from samples we might have drawn), 95% would have contained the unknown true parameter value μ .

Example A corporation conducts a survey to investigate the proportion of employees who thought the board was doing a good job. 1000 employees, randomly selected, were asked, and 732 said they did. Find a 99% confidence interval for the value of the proportion in the population who thought the board was doing a good job.

Clearly each observation $X_i \sim \text{Bernoulli}(p)$ for some unknown p , and we want to find a C.I. for p , which is also the mean of X_i .

We have our estimate $\hat{p} = \bar{x} = 0.732$ for which we can use the CLT. Since the variance of $\text{Bernoulli}(p)$ is $p(1-p)$, we can use $\bar{x}(1-\bar{x}) = 0.196$ as an approximate variance.

So an approximate 99% C.I. is

$$\left[0.732 - 2.576 \times \sqrt{\frac{0.196}{1000}}, 0.732 + 2.576 \times \sqrt{\frac{0.196}{1000}} \right]$$

■

9.2.1 Normal Distribution with Known Variance

The confidence interval given in the $\text{Bernoulli}(p)$ example was only an approximate interval, relying on the Central Limit Theorem, and also assuming the population variance σ^2 was known.

However, if we in fact know that X_1, \dots, X_n are an i.i.d. sample from $N(\mu, \sigma^2)$, then we have exactly

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

In which case,

$$\left[\bar{x} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right]$$

is an *exact* confidence interval for μ , assuming we know σ^2 .

9.2.2 Normal Distribution with Unknown Variance

In any applied example where we are aiming to fit a normal distribution model to real data, it will usually be the case that both μ and σ^2 are unknown.

However, if again we have X_1, \dots, X_n as an i.i.d. sample from $N(\mu, \sigma^2)$ but with σ^2 now unknown, then we have exactly

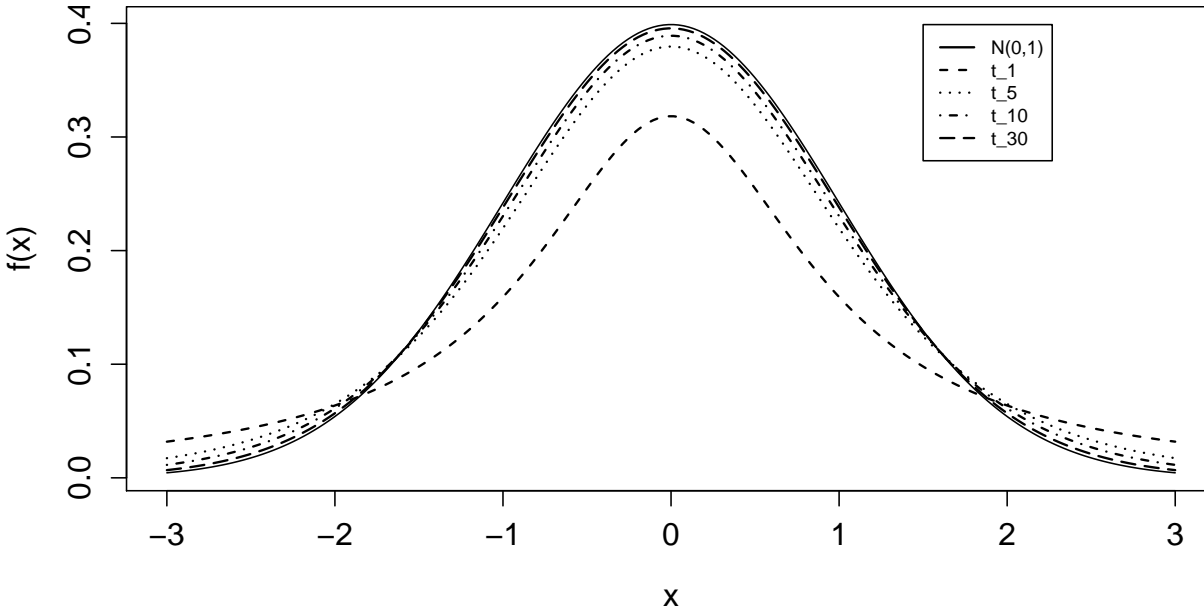
$$\frac{\bar{X} - \mu}{s_{n-1}/\sqrt{n}} \sim t_{n-1}$$

where $s_{n-1} = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}}$ is the bias-corrected sample standard deviation, and t_ν is the Student's t -distribution with ν degrees of freedom.

Then it follows that an exact $100(1 - \alpha)\%$ confidence interval for μ is

$$\left[\bar{x} - t_{n-1, 1-\frac{\alpha}{2}} \frac{s_{n-1}}{\sqrt{n}}, \bar{x} + t_{n-1, 1-\frac{\alpha}{2}} \frac{s_{n-1}}{\sqrt{n}} \right]$$

where $t_{\nu, \alpha}$ is the α -quantile of t_ν .



Notes

- t_ν is heavier tailed than $N(0,1)$ for any number of degrees of freedom ν .
- Hence the t -distribution CI will always be wider than the Normal distribution CI. So if we know σ^2 , we should use it.
- $\lim_{\nu \rightarrow \infty} t_\nu \equiv N(0,1)$.
- For $\nu > 40$ the difference between t_ν and $N(0,1)$ is so insignificant that the t distribution is not tabulated beyond this many degrees of freedom, and so there we can instead revert to $N(0,1)$ tables.

ν	α				ν	α			
	0.95	0.975	0.99	0.995		0.95	0.975	0.99	0.995
1	6.31	12.71	31.82	63.66	9	1.83	2.26	2.82	3.25
2	2.92	4.30	6.96	9.92	10	1.81	2.23	2.76	3.17
3	2.35	3.18	4.54	5.84	12	1.78	2.18	2.68	3.05
4	2.13	2.78	3.75	4.60	15	1.75	2.13	2.60	2.95
5	2.02	2.57	3.36	4.03	20	1.72	2.09	2.53	2.85
6	1.94	2.45	3.14	3.71	25	1.71	2.06	2.48	2.78
7	1.89	2.36	3.00	3.50	40	1.68	2.02	2.42	2.70
8	1.86	2.31	2.90	3.36	∞	1.645	1.96	2.326	2.576

Example A random sample of 100 observations from a normally distributed population has sample mean 83.2 and bias-corrected sample standard deviation of 6.4.

1. Find a 95% confidence interval for μ .
2. Give an interpretation for this interval.

Solution:

1. An exact 95% confidence interval would given by $\bar{x} \pm t_{99,0.975} \frac{s_{n-1}}{\sqrt{100}}$.

Since $n = 100$ is large, we can approximate this by

$$\bar{x} \pm z_{0.975} \frac{s_{n-1}}{\sqrt{100}} = 83.2 \pm 1.96 \times \frac{6.4}{10} = [81.95, 84.45].$$

2. With 95% confidence, we can say that the population mean lies between 81.95 and 84.45



Chapter 10. Hypothesis Testing

Suppose we want to know if exposure to asbestos is associated with lung disease. We take some rat and randomly divide them into two groups. We expose one group to asbestos and leave the second group unexposed. Then we compare the disease rate in the two groups. Consider the following two hypotheses:

The Null Hypothesis : The disease rate is the same in the two groups.

The Alternative Hypothesis : The disease rate is not the same in the two groups.

If the exposed groups has a much higher rate of disease than the unexposed group then we will reject the null hypothesis and conclude that the evidence favours the alternative hypothesis. This is an example of hypothesis testing.

More specifically, we may fix upon a parametric family $P_{X|\theta}$ and then test whether hypothesised parameter values for θ are plausible; that is, test whether we could reasonably assume $\theta = \theta_0$ for some particular value θ_0 . For example, if $X \sim N(\mu, \sigma^2)$ we may wish to test whether $\mu = 0$ is plausible in light of the data.

Formally, suppose that we partition the parameter space Θ into two disjoint sets Θ_0 and Θ_1 and that we wish to test

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

We call the H_0 the **null hypothesis** and H_1 the **alternative hypothesis**.

To test the validity of H_0 , we first choose a test statistic $T(\underline{X})$ of the data for which we can find the distribution, P_T , under H_0 . One of the difficulties in hypothesis testing is to find an appropriate testing statistic T .

Then, we identify a rejection region $R \subset \mathbb{R}$ of low probability values of T under the assumption that H_0 is true, so

$$P(T \in R | H_0) = \alpha$$

for some small probability α (typically 5%).

A well chosen rejection region will have relatively high probability under H_1 , whilst retaining low probability under H_0 .

Finally, we calculate the observed test statistic $t(\underline{x})$ for our observed data \underline{x} .

- If $t \in R$ we “reject the null hypothesis at the $100\alpha\%$ level”.
- If $t \notin R$ we “retain (do not reject) the null hypothesis at the $100\alpha\%$ level”.

For each possible **significance level** $\alpha \in (0, 1)$, a hypothesis test at the $100\alpha\%$ level will result in either rejecting or not rejecting H_0 .

- As $\alpha \rightarrow 0$ it becomes less and less likely that we will reject our null hypothesis, as the rejection region is becoming smaller and smaller.
- Similarly, as $\alpha \rightarrow 1$ it becomes more and more likely that we will reject our null hypothesis.

For any given data, we might, therefore, be interested in identifying the critical significance level which marks the threshold between us rejecting and not rejecting the null hypothesis. This is known as the p -value of the data. Smaller p -values suggest stronger evidence against H_0 .

Interpretation

Hypothesis testing is like a legal trial. We assume someone is innocent unless the evidence strongly suggests that the person is guilty. Similarly, we retain H_0 unless there is strong evidence to reject H_0 .

10.0.1 Error Rates and Power of a Test

There are two types of error in the outcome of a hypothesis test:

- Type I: Rejecting H_0 when in fact H_0 is true. By construction, this happens with probability α . For this reason, the significance level of a hypothesis test is also referred to as the Type I error rate.
- Type II: Not rejecting H_0 when in fact H_1 is true i.e. $\beta = P(T \notin R | \theta \in \Theta_1)$.

Definition 10.0.1. *The power of a hypothesis test is defined as*

$$1 - \beta = P(T \in R | \theta \in \Theta_1).$$

A hypothesis of the form $\theta = \theta_0$ is called a simple hypothesis. A hypothesis of the form $\theta > \theta_0$ or $\theta < \theta_0$ is called a composite hypothesis. A test of the form

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

is called a two-sided test. A test of the form

$$H_0 : \theta \leq \theta_0 \quad \text{versus} \quad H_1 : \theta > \theta_0$$

or

$$H_0 : \theta \geq \theta_0 \quad \text{versus} \quad H_1 : \theta < \theta_0$$

is called a one-sided test.

Note It would be desirable to find the test with highest power under H_1 , among all size α tests. Such a test, if it exists, is called most powerful. Finding most powerful tests is hard and, in many cases, most powerful test don't even exist. Instead of going into detail about when most powerful tests exists, we shall just consider some commonly used test.

10.1 Testing for a population mean

10.1.1 Normal Distribution with Known Variance

Suppose X_1, \dots, X_n are i.i.d. $N(\mu, \sigma^2)$ with σ^2 known and μ unknown. We may wish to test if $\mu = \mu_0$ for some specific value μ_0 (e.g. $\mu_0 = 0$, $\mu_0 = 9.8$).

Then we can state our null and alternative hypotheses as

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu \neq \mu_0.$$

Under $H_0 : \mu = \mu_0$, we then know both μ and σ^2 . So for the sample mean \bar{X} we have a known distribution for the test statistic

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim \Phi.$$

So if we define our rejection region R to be the $100\alpha\%$ tails of the standard normal distribution,

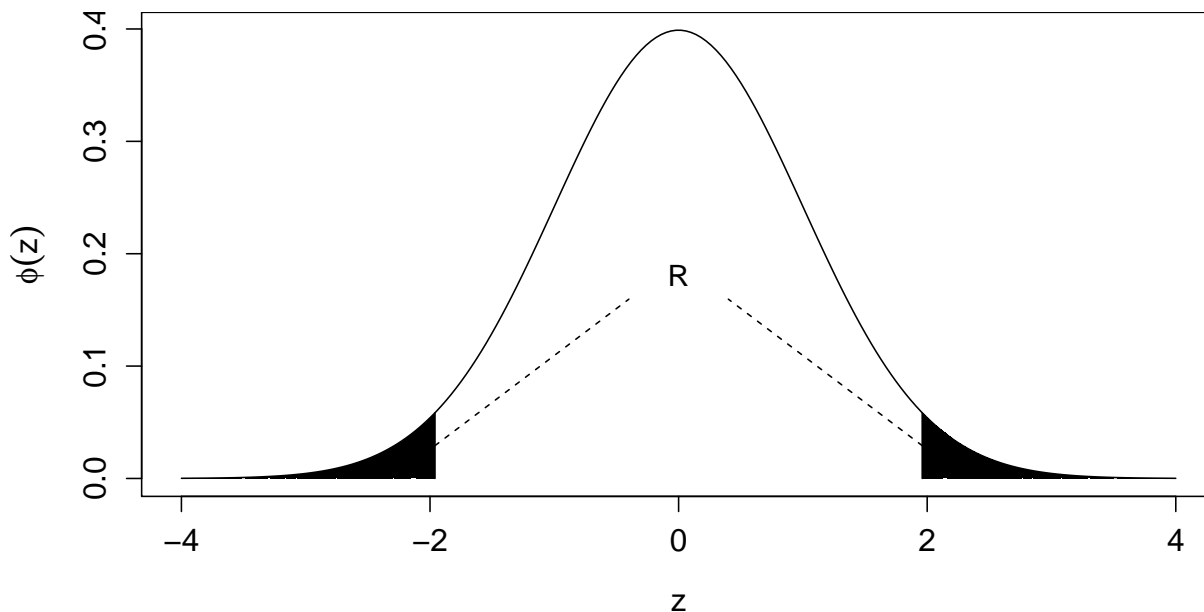
$$\begin{aligned} R &= \left(-\infty, -z_{1-\frac{\alpha}{2}}\right) \cup \left(z_{1-\frac{\alpha}{2}}, \infty\right) \\ &\equiv \left\{z \mid |z| > z_{1-\frac{\alpha}{2}}\right\}, \end{aligned}$$

we have $P(Z \in R | H_0) = \alpha$.

We thus reject H_0 at the $100\alpha\%$ significance level \iff our observed test statistic

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \in R.$$

The p-value is given by $2 \times \{1 - \Phi(|z|)\}$.



Example A company makes packets of snack foods. The bags are labelled as weighing 454g; of course they won't all be exactly 454g, and let's suppose the variance of bag weights is known to be 70g^2 . The following data show the mass in grams of 50 randomly sampled packets.

464, 450, 450, 456, 452, 433, 446, 446, 450, 447, 442, 438, 452, 447, 460, 450, 453, 456,
446, 433, 448, 450, 439, 452, 459, 454, 456, 454, 452, 449, 463, 449, 447, 466, 446, 447,
450, 449, 457, 464, 468, 447, 433, 464, 469, 457, 454, 451, 453, 443

Are these data consistent with the claim that the mean weight of packets is 454g?

1. We wish to test $H_0 : \mu = 454$ vs. $H_1 : \mu \neq 454$. So set $\mu_0 = 454$.
2. Although we have not been told that the packet weights are individually normally distributed, by the CLT we still have that the mean weight of the sample of packets is approximately normally distributed, and hence we still *approximately* have $Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim \Phi$.
3. Compute the realised value of the test statistic: $\bar{x} = 451.22$ and $n = 50 \Rightarrow z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = -2.350$.
4. For a 5%-level significance test, we compare the statistic $z = -2.350$ with the rejection region $R = (-\infty, -z_{0.975}) \cup (z_{0.975}, \infty) = (-\infty, -1.96) \cup (1.96, \infty)$. Clearly we have $z \in R$, and so at the 5%-level we reject the null hypothesis that the mean packet weight is 454g. We conclude the test by stating: **there is sufficient evidence to reject the null hypothesis at the 5% level.**
5. At which significance levels would we have not rejected the null hypothesis?

- For a 1%-level significance test, the rejection region would have been

$$R = (-\infty, -z_{0.995}) \cup (z_{0.995}, \infty) = (-\infty, -2.576) \cup (2.576, \infty).$$

In which case $z \notin R$, and so at the 1%-level we would not have rejected the null hypothesis.

- The p -value is

$$\begin{aligned} 2 \times \{1 - \Phi(|z|)\} &= 2 \times \{1 - \Phi(|-2.350|)\} \approx 2(1 - 0.9906) \\ &= 0.019, \end{aligned}$$

and so we would only reject the null hypothesis for $\alpha > 1.9\%$.



Note There is a strong connection between hypothesis testing and confidence intervals. Suppose we have constructed a $100(1 - \alpha)\%$ confidence interval for a parameter θ . Then this is precisely the set of values θ_0 for which there would be not be sufficient evidence to reject a null hypothesis $H_0 : \theta = \theta_0$ at the $100\alpha\%$ -level.

10.1.2 Normal Distribution with Unknown Variance

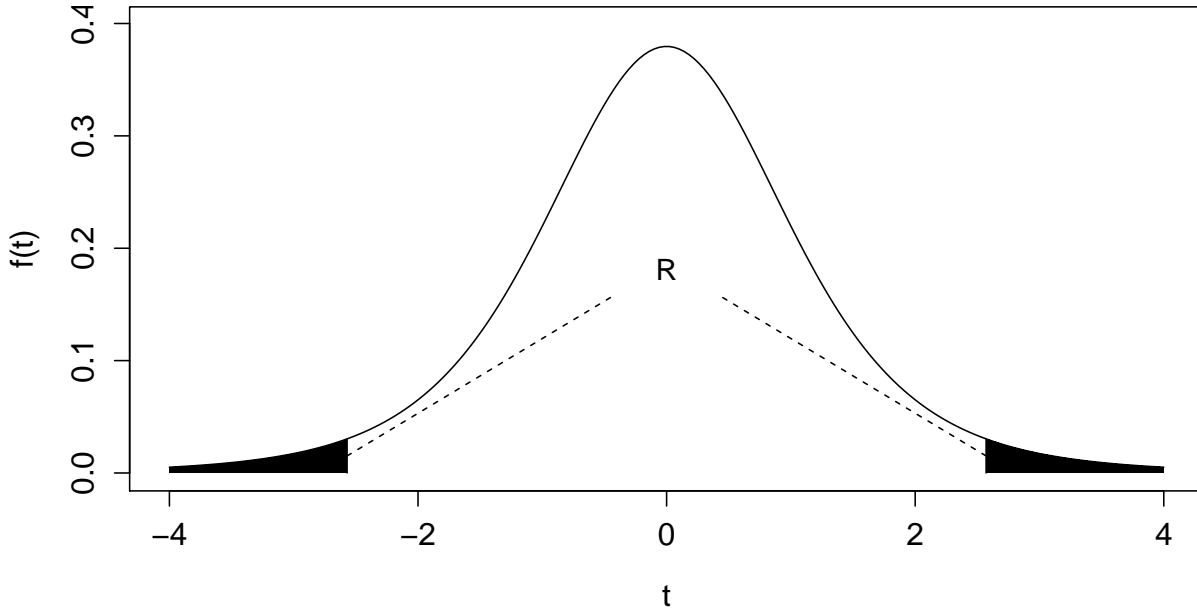
Similarly, if σ^2 in the previous example were unknown, we still have that

$$T = \frac{\bar{X} - \mu_0}{s_{n-1}/\sqrt{n}} \sim t_{n-1}.$$

So for a test of $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$ at the α level, the rejection region of our observed test statistic $t = \frac{\bar{x} - \mu_0}{s_{n-1}/\sqrt{n}}$ is

$$\begin{aligned} R &= \left(-\infty, -t_{n-1, 1-\frac{\alpha}{2}}\right) \cup \left(t_{n-1, 1-\frac{\alpha}{2}}, \infty\right) \\ &\equiv \left\{t \mid |t| > t_{n-1, 1-\frac{\alpha}{2}}\right\}. \end{aligned}$$

Again, we have that $P(T \in R | H_0) = \alpha$.



Example Consider again the snack food weights example. There, we assumed the variance of bag weights was known to be 70. Without this, we could have estimated the variance by

$$s_{n-1}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = 70.502.$$

Then the corresponding t -statistic becomes

$$t = \frac{\bar{x} - \mu_0}{s_{n-1}/\sqrt{n}} = -2.341,$$

very similar to the z -statistic of before.

And since $n = 50$, we compare with the t_{49} distribution which is approximately $N(0, 1)$. So the hypothesis test results and p -value would be practically identical. ■

Example A particular piece of code takes a random time to run on a computer, but the average time is known to be 6 seconds. The programmer tries an alternative optimisation in compilation and wishes to know whether the mean run time has changed. To explore this, he runs the re-optimised code 16 times, obtaining a sample mean run time of 5.8 seconds and bias-corrected sample standard deviation of 1.2 seconds. Is the code any faster?

1. We wish to test $H_0 : \mu = 6$ vs. $H_1 : \mu \neq 6$. So set $\mu_0 = 6$.
2. By the CLT, $T = \frac{\bar{X} - \mu_0}{s_{n-1}/\sqrt{n}} \sim t_{n-1}$. That is, $\frac{\bar{X} - 6}{s_{n-1}/\sqrt{16}} \sim t_{15}$. So we reject H_0 at the $100\alpha\%$ level if $|t| > t_{15, 1-\alpha/2}$.
3. Compute the realised value of the test statistic: $\bar{x} = 5.8$, $s_{n-1} = 1.2$ and $n = 16$
 $\Rightarrow t = \frac{\bar{x} - \mu_0}{s_{n-1}/\sqrt{n}} = -0.657$.
4. We have $|t| = 0.657 \ll 2.13 = t_{15, 975}$, so **we have insufficient evidence to reject H_0 at the 5% level.**
5. In fact, the p -value for these data is 51.51%, so there is very little evidence to suggest the code is now any faster.

■

10.2 Testing for differences in population means

10.2.1 Two Sample Problems

Suppose, as before, we have a random sample $\underline{X} = (X_1, \dots, X_{n_1})$ from an unknown population distribution P_X .

But now, suppose we have a further random sample $\underline{Y} = (Y_1, \dots, Y_{n_2})$ from a second, different population P_Y .

Then we may wish to test hypotheses concerning the similarity of the two distributions P_X and P_Y .

In particular, we are often interested in testing whether P_X and P_Y have equal means. That is, to test

$$H_0 : \mu_X = \mu_Y \quad \text{vs} \quad H_1 : \mu_X \neq \mu_Y.$$

A special case is when the two samples \underline{X} and \underline{Y} are *paired*. That is, if $n_1 = n_2 = n$ and the data are collected as pairs $(X_1, Y_1), \dots, (X_n, Y_n)$ so that, for each i , X_i and Y_i are possibly dependent.

For example, we might have a random sample of n individuals and X_i represents the heart rate of the i^{th} person before light exercise and Y_i the heart rate of the same person afterwards.

In this special case, for a test of equal means we can consider the sample of differences $Z_1 = X_1 - Y_1, \dots, Z_n = X_n - Y_n$ and test $H_0 : \mu_Z = 0$ using the single sample methods we have seen.

In the above example, this would test whether light exercise causes a change in heart rate.

10.2.2 Normal Distributions with Known Variances

Suppose

- $\underline{X} = (X_1, \dots, X_{n_1})$ are i.i.d. $N(\mu_X, \sigma_X^2)$ with μ_X unknown;
- $\underline{Y} = (Y_1, \dots, Y_{n_2})$ are i.i.d. $N(\mu_Y, \sigma_Y^2)$ with μ_Y unknown;
- the two samples \underline{X} and \underline{Y} are independent.

Then we still have that, independently,

$$\bar{X} \sim N\left(\mu_X, \frac{\sigma_X^2}{n_1}\right), \quad \bar{Y} \sim N\left(\mu_Y, \frac{\sigma_Y^2}{n_2}\right)$$

From this it follows that the difference in sample means,

$$\bar{X} - \bar{Y} \sim N\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}\right),$$

and hence

$$\frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\sigma_X^2/n_1 + \sigma_Y^2/n_2}} \sim \Phi.$$

So under the null hypothesis $H_0 : \mu_X = \mu_Y$, we have

$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\sigma_X^2/n_1 + \sigma_Y^2/n_2}} \sim \Phi.$$

So if σ_X^2 and σ_Y^2 are known, we immediately have a test statistic

$$z = \frac{\bar{x} - \bar{y}}{\sqrt{\sigma_X^2/n_1 + \sigma_Y^2/n_2}}$$

which we can compare against the quantiles of a standard normal.

That is,

$$R = \left\{ z \mid |z| > z_{1-\frac{\alpha}{2}} \right\},$$

gives a rejection region for a hypothesis test of $H_0 : \mu_X = \mu_Y$ vs. $H_1 : \mu_X \neq \mu_Y$ at the $100\alpha\%$ level.

10.2.3 Normal Distributions with Unknown Variances

On the other hand, suppose σ_X^2 and σ_Y^2 are unknown. Then if we know $\sigma_X^2 = \sigma_Y^2 = \sigma^2$ but σ^2 is unknown, we can still proceed.

We have

$$\frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sigma \sqrt{1/n_1 + 1/n_2}} \sim \Phi,$$

and so, under $H_0 : \mu_X = \mu_Y$,

$$\frac{\bar{X} - \bar{Y}}{\sigma \sqrt{1/n_1 + 1/n_2}} \sim \Phi.$$

but with σ unknown.

We need an estimator for the variance using samples from two populations with different means. Just combining the samples together into one big sample would over-estimate the variance, since some of the variability in the samples would be due to the difference in μ_X and μ_Y .

So we define the **bias-corrected pooled sample variance**

$$S_{n_1+n_2-2}^2 = \frac{\sum_{i=1}^{n_1} (X_i - \bar{X})^2 + \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2}{n_1 + n_2 - 2},$$

which is an unbiased estimator for σ^2 .

We can immediately see that $s_{n_1+n_2-2}^2$ is indeed an unbiased estimate of σ^2 by noting

$$S_{n_1+n_2-2}^2 = \frac{n_1 - 1}{n_1 + n_2 - 2} S_{n_1-1}^2 + \frac{n_2 - 1}{n_1 + n_2 - 2} S_{n_2-1}^2;$$

That is, $s_{n_1+n_2-2}^2$ is a weighted average of the bias-corrected sample variances for the individual samples \underline{x} and \underline{y} , which are both unbiased estimates for σ^2 .

Then substituting $S_{n_1+n_2-2}$ in for σ we get

$$\frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{S_{n_1+n_2-2} \sqrt{1/n_1 + 1/n_2}} \sim t_{n_1+n_2-2},$$

and so, under $H_0 : \mu_X = \mu_Y$,

$$T = \frac{\bar{X} - \bar{Y}}{S_{n_1+n_2-2} \sqrt{1/n_1 + 1/n_2}} \sim t_{n_1+n_2-2}.$$

So we have a rejection region for a hypothesis test of $H_0 : \mu_X = \mu_Y$ vs. $H_1 : \mu_X \neq \mu_Y$ at the $100\alpha\%$ level given by

$$R = \left\{ t \mid |t| > t_{n_1+n_2-2, 1-\frac{\alpha}{2}} \right\},$$

for the statistic

$$t = \frac{\bar{x} - \bar{y}}{s_{n_1+n_2-2} \sqrt{1/n_1 + 1/n_2}}.$$

Example The same piece of C code was repeatedly run after compilation under two different C compilers, and the run times under each compiler were recorded. The sample mean and bias-corrected sample variance for Compiler 1 were 114s and 310s respectively, and the corresponding figures for Compiler 2 were 94s and 290s. Both sets of data were each based on 15 runs.

Suppose that Compiler 2 is a refined version of Compiler 1, and so if μ_1, μ_2 are the expected run times of the code under the two compilations, we might fairly assume $\mu_2 \leq \mu_1$.

Conduct a hypothesis test of $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 > \mu_2$ at the 5% level.

Until now we have exclusively considered *two-sided* tests. That is tests of the form $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$.

Here we need to consider *one-sided* tests, which differ by the alternative hypothesis being of the form $H_1 : \theta < \theta_0$ or $H_1 : \theta > \theta_0$.

This presents no extra methodological challenge and requires only a slight adjustment in the construction of the rejection region.

We still use the t -statistic

$$t = \frac{\bar{x} - \bar{y}}{s_{n_1+n_2-2} \sqrt{1/n_1 + 1/n_2}},$$

where \bar{x}, \bar{y} are the sample mean run times under Compilers 1 and 2 respectively. But now the one-sided rejection region becomes

$$R = \{t \mid t > t_{n_1+n_2-2, 1-\alpha}\}.$$

First calculating the bias-corrected pooled sample variance, we get

$$s_{n_1+n_2-2}^2 = \frac{14 \times 310 + 14 \times 290}{28} = 300.$$

(Note that since the sample sizes n_1 and n_2 are equal, the pooled estimate of the variance is the average of the individual estimates.)

$$\begin{aligned} \text{So } t &= \frac{\bar{x} - \bar{y}}{s_{n_1+n_2-2} \sqrt{1/n_1 + 1/n_2}} = \frac{114 - 94}{\sqrt{300} \sqrt{1/15 + 1/15}} \\ &= \sqrt{10} = 3.162. \end{aligned}$$

For a 1-sided test we compare $t = 3.162$ with $t_{28, 0.95} = 1.701$ and conclude that we reject the null hypothesis at the 5% level; the second compilation is significantly faster. ■

10.3 Goodness of Fit

10.3.1 Count Data and Chi-Square Tests

The results in the previous sections relied upon the data being either normally distributed, or at least through the CLT having the sample mean being approximately normally distributed. Tests were then developed for making inference on population means under those assumptions. These tests were very much *model-based*.

Another important but very different problem concerns *model checking*, which can be addressed through a more general consideration of count data for simple (discrete and finite) distributions.

The following ideas can then be trivially extended to infinite range discrete and continuous random variables by *binning* observed samples into a finite collection of predefined intervals.

Let X be a simple random variable taking values in the range $\{x_1, \dots, x_k\}$, with probability mass function $p_j = P(X = x_j|\theta)$, $j = 1, \dots, k$ depending on an unknown parameter p -vector θ .

Then a random sample of size n from the distribution of X can be summarised by the *observed* frequency counts $\underline{O} = (O_1, \dots, O_k)$ at the points x_1, \dots, x_k (so $\sum_{j=1}^k O_j = n$).

Suppose we have a null hypothesis $H_0 : \theta = \theta_0$ for the value of the unknown parameter(s). Then under H_0 we know the pmf $\{p_j\}$, and so we are able to calculate the expected frequency counts $\underline{E} = (E_1, \dots, E_k)$ by $E_j = np_j$. (Note again we have $\sum_{j=1}^k E_j = n$.)

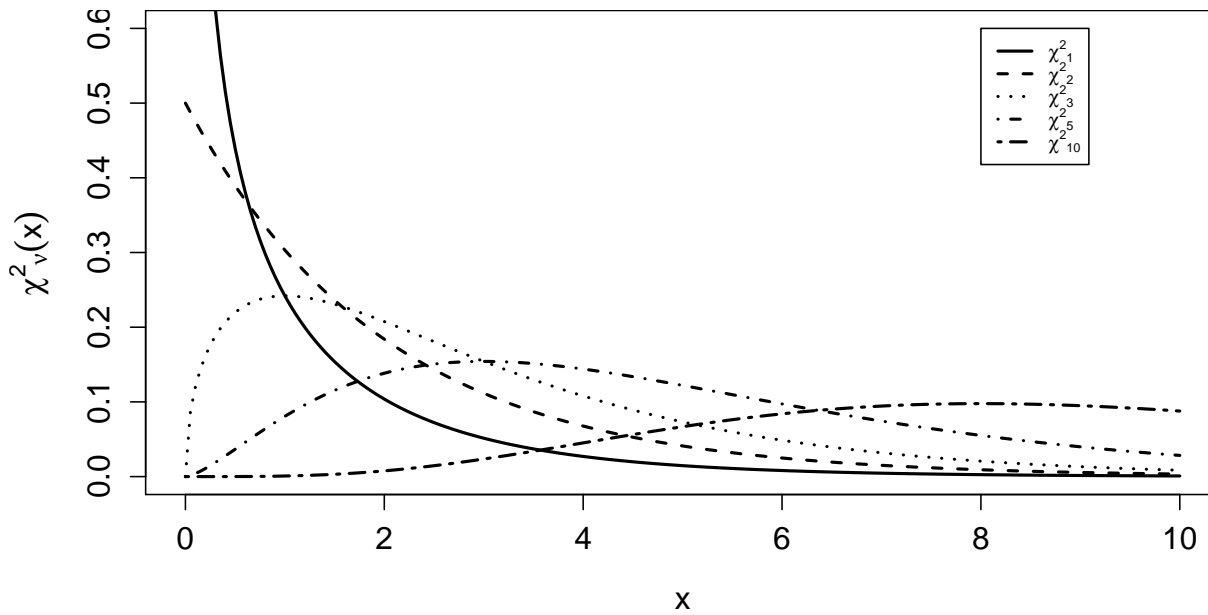
We then seek to compare the observed frequencies with the expected frequencies to test for **goodness of fit**.

To test $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$ we use the **chi-square statistic**

$$X^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}.$$

If H_0 were true, then the statistic X^2 would approximately follow a **chi-square distribution** with $\nu = k - p - 1$ degrees of freedom.

- k is the number of values (categories) the simple random variable X can take.
- p is the number of parameters being estimated ($\dim(\theta)$).
- For the approximation to be valid, we should have $\forall j, E_j \geq 5$. This may require some merging of categories.



Clearly larger values of X^2 correspond to larger deviations from the null hypothesis model. That is, if $X^2 = 0$ the observed counts exactly match those expected under H_0 .

For this reason, we always perform a one-sided goodness of fit test using the χ^2 statistic, looking only at the upper tail of the distribution.

Hence the rejection region for a goodness of fit hypothesis test at the $100\alpha\%$ level is given by

$$R = \left\{ x^2 \mid x^2 > \chi^2_{k-p-1, 1-\alpha} \right\}.$$

10.3.2 Proportions

Each year, around 1.3 million people in the USA suffer adverse drug effects (ADEs). A study in the *Journal of the American Medical Association* (July 5, 1995) gave the causes of 95 ADEs below.

Cause	Number of ADEs
Lack of knowledge of drug	29
Rule violation	17
Faulty dose checking	13
Slips	9
Other	27

Test whether the true percentages of ADEs differ across the 5 causes.

Under the null hypothesis that the 5 causes are equally likely, we would have expected counts of $\frac{95}{5} = 19$ for each cause.

So our χ^2 statistic becomes

$$\begin{aligned} x^2 &= \frac{(29-19)^2}{19} + \frac{(17-19)^2}{19} + \frac{(13-19)^2}{19} + \frac{(9-19)^2}{19} + \frac{(27-19)^2}{19} \\ &= \frac{100}{19} + \frac{4}{19} + \frac{36}{19} + \frac{100}{19} + \frac{64}{19} = \frac{304}{19} = 16. \end{aligned}$$

We have not estimated any parameters from the data, so we compare x^2 with the quantiles of the $\chi^2_{5-1} = \chi^2_4$ distribution.

Well $16 > 9.49 = \chi^2_{4,0.95}$, so we reject the null hypothesis at the 5% level; we have reason to suppose that there is a difference in the true percentages across the different causes.

10.3.3 Model Checking

Recall the example from Chapter 6 (Discrete random variable) the number of particles emitted by a radioactive substance which reached a Geiger counter was measured for 2608 time intervals, each of length 7.5 seconds.

We fitted a $\text{Poisson}(\lambda)$ distribution to the data by plugging in the sample mean number of counts (3.870) for the rate parameter λ . (Which we now know to be the MLE.)

x	0	1	2	3	4	5	6	7	8	9	≥ 10
$O(n_x)$	57	203	383	525	532	408	273	139	45	27	16
$E(n_x)$	54.4	210.5	407.4	525.5	508.4	393.5	253.8	140.3	67.9	29.2	17.1

(O=Observed, E=Expected).

Whilst the fitted $\text{Poisson}(3.87)$ expected frequencies looked quite convincing to the eye, at that time we had no formal method of quantitatively assessing the fit. However, we now know how to proceed.

x	0	1	2	3	4	5	6	7	8	9	≥ 10
O	57	203	383	525	532	408	273	139	45	27	16
E	54.4	210.5	407.4	525.5	508.4	393.5	253.8	140.3	67.9	29.2	17.1
$O - E$	2.6	-7.5	-24.4	-0.5	23.6	14.5	19.2	-1.3	22.9	2.2	1.1
$\frac{(O-E)^2}{E}$	0.124	0.267	1.461	0.000	1.096	0.534	1.452	0.012	7.723	0.166	0.071

The statistic $x^2 = \sum \frac{(O-E)^2}{E} = 12.906$ should be compared with a $\chi^2_{11-1-1} = \chi^2_9$ distribution.

Well $\chi^2_{9,0.95} = 16.91$, so at the 5% level we do not reject the null hypothesis of a $\text{Poisson}(3.87)$ model for the data.

10.3.4 Independence

Suppose we have two discrete random variables X and Y that can each take finite values which are jointly distributed with unknown probability mass function p_{XY} .

We are often interested in trying to ascertain whether X and Y are independent. That is, determine whether $p_{XY}(x, y) = p_X(x)p_Y(y)$.

Let the ranges of the random variables X and Y be $\{x_1, \dots, x_k\}$ and $\{y_1, \dots, y_\ell\}$ respectively. Then an i.i.d. sample of size n from the joint distribution of (X, Y) can be

represented by a list of counts n_{ij} ($1 \leq i \leq k; 1 \leq j \leq \ell$) of the number of times we observe the pair (x_i, y_j) .

Tabulating these data in the following way gives what is known as a $k \times \ell$ **contingency table**.

	y_1	y_2	\dots	y_ℓ	
x_1	n_{11}	n_{12}		$n_{1\ell}$	$n_{1\bullet}$
x_2	n_{21}	n_{22}		$n_{2\ell}$	$n_{2\bullet}$
\vdots					
x_k	n_{k1}	n_{k2}		$n_{k\ell}$	$n_{k\bullet}$
	$n_{\bullet 1}$	$n_{\bullet 2}$	\dots	$n_{\bullet \ell}$	n

Note the row sums $(n_{1\bullet}, n_{2\bullet}, \dots, n_{k\bullet})$ represent the frequencies of x_1, x_2, \dots, x_k in the sample (that is, ignoring the value of Y). Similarly for the column sums $(n_{\bullet 1}, n_{\bullet 2}, \dots, n_{\bullet \ell})$ and y_1, \dots, y_ℓ .

Under the null hypothesis

$$H_0 : X \text{ and } Y \text{ are independent,}$$

the expected values of the entries of the contingency table, conditional on the row and column sums, are given by

$$\hat{n}_{ij} = \frac{n_{i\bullet} \times n_{\bullet j}}{n}, \quad 1 \leq i \leq k, 1 \leq j \leq \ell.$$

To see this, consider the marginal distribution of X ; we could approximate $p_X(x_i)$ by $\hat{p}_{i\bullet} = \frac{n_{i\bullet}}{n}$. Similarly for $p_Y(y_j)$ we get $\hat{p}_{\bullet j} = \frac{n_{\bullet j}}{n}$.

Then under independence $p_{XY}(x_i, y_j) = p_X(x_i)p_Y(y_j)$, and so we can estimate $p_{XY}(x_i, y_j)$ by

$$\hat{p}_{ij} = \hat{p}_{i\bullet} \times \hat{p}_{\bullet j} = \frac{n_{i\bullet} \times n_{\bullet j}}{n^2}.$$

Now that we have a set of expected frequencies to compare against our $k \times \ell$ observed frequencies, a χ^2 test can be performed.

We are using both the row and column sums to estimate our probabilities, and there are k and ℓ of these respectively. So we compare our calculated x^2 statistic against a χ^2 distribution with $k\ell - (k-1) - (\ell-1) - 1 = (k-1)(\ell-1)$ degrees of freedom.

Hence the rejection region for a hypothesis test of independence in a $k \times \ell$ contingency table at the $100\alpha\%$ level is given by

$$R = \left\{ x^2 \mid x^2 > \chi_{(k-1)(\ell-1), 1-\alpha}^2 \right\}.$$

Example An article in *International Journal of Sports Psychology* (July-Sept 1990) evaluated the relationship between physical fitness and stress. 549 people were classified as good, average, or poor fitness, and were also tested for signs of stress (yes or no). The data are shown in the table below.

	Poor Fitness	Average Fitness	Good Fitness	
Stress	206	184	85	475
No stress	36	28	10	74
	242	212	95	549

Question Is there any relationship between stress and fitness?

Under independence we would estimate the expected values to be

	Poor Fitness	Average Fitness	Good Fitness	
Stress	209.4	183.4	82.2	475
No stress	32.6	28.6	12.8	74
	242	212	95	549

Hence the χ^2 statistic is calculated to be

$$x^2 = \sum_i \frac{(O_i - E_i)^2}{E_i} = \frac{(206 - 209.4)^2}{209.4} + \dots + \frac{(10 - 12.88)^2}{12.8} = 1.1323.$$

This should be compared with a χ^2 distribution with $(2 - 1) \times (3 - 1) = 2$ degrees of freedom. We then have $\chi^2_{2,0.95} = 5.99$, so we have insufficient evidence to reject to null i.e. no significant evidence to suggest there is any relationship between fitness and stress. ■