1.1 Introduction

Subjects of Statistical Studies

Definition

Data is a mathematical object, usually a collection of measurements represented by numbers or names, but it could also take on more abstract forms such as trees, graphs, and sets.

Definition

Statistics is the study of the collection, organization, analysis, interpretation and presentation of data. It deals with all aspects of data, including the planning of data collection in terms of the design of surveys and experiments.

Definition

A **population** is the entire group to be studied. An **individual** is a member of a population. A **sample** is a subset of a population.

Descriptive Versus Inferential

Definition

Descriptive Statistics consists of organizing and summarizing the data. It describes the data with numerical summaries, tables, and graphs.

Definition

Inferential Statistics is the practice of drawing conclusions about populations based on sample data, and measuring the reliability of the results.

Remark

Inferential Statistics is needed because sampling is difficult and expensive. If every population of interest could be surveyed cheaply and reliably, there would be no point to Inferential Statistics.

Definition

A (**population**) **parameter** is a numerical summary derived from the population data.

Definition

A (**sample**) **statistic** is a numerical summary derived from the sample data.

Remark

One major goal of Inferential Statistics is to estimate population parameters from sample statistics.

Variable Types

- **Qualitative (categorical)**
  classify individuals based on some attribute
- **Quantitative**
  measure a numerical quantity associated with individuals
  - **Discrete**
    finite or countable number of values, is not expected to take a value between any two possible values
  - **Continuous**
    uncountably infinite number of values, may always take on a value in between any two possible values

Levels of Measurement

- **Nominal**
  values name, label, categorize; cannot be meaningfully ordered or ranked
- **Ordinal**
  like nominal, but values can be ordered
- **Interval**
  like ordinal, but with meaningful differences between values, addition, and subtraction
- **Ratio**
  like interval, but with meaningful ratios of values, multiplication, division, and zero representing the absence of quantity

Levels of Measurement

Determine the level of measurement for the following variables:

1. Last name
2. Social Security Number
3. Mass
4. Temperature in °F
5. Military rank

1.2 Types of Statistical Studies

Variables Types in Relation to Studies

- **Explanatory variables** are controlled or manipulated by statisticians
- **Response variables** are the measured outcomes of the studies

Types of Studies

Definition

An **observational study** measures the value of the response variable without attempting to influence the outcome

Definition

In a **designed experiment**, the individuals are divided into several groups, and the study intentionally changes the value of an explanatory variable before recording the response variable

Examples of Studies

1. Measuring the amount of cellphone usage by drivers in cars
2. Measuring the likelihood of an accident depending on the amount of cellphone usage by car drivers
3. Measuring the consumption of saturated fat
4. Measuring the risk factor for cardiovascular disease (CVD) depending on the consumption rate of saturated fat

Pitfalls

- **Confounding** occurs when a relationship between explanatory variables is not accounted for
- A **lurking variable** is an explanatory variable that was not considered in a study
- It may be unethical to measure a variable via a designed experiment

Census

Definition

A **census** is a list of all individuals in a population along with certain characteristics of each individual

1.3 Simple Random Sampling

Sampling

Definition

Random sampling is the process of using chance (randomness) to select individuals from a population to be included in the sample.

Obtaining a Simple Random Sample

Definition

A sample of size n from a population of size N is obtained through **simple random sampling** if every possible sample of size n is equally likely.

Sampling Procedure

To draw a simple random sample of size n out of a population of size N,

1. Make a list of all individuals in a population, enumerated by natural numbers from 1 to N
2. Select randomly n pairwise distinct natural numbers from the interval $[1, N]$ (sampling *without replacement*)
3. Administer the study to the individuals whose numbers were selected

Sampling Difficulties

- The population size may be unknown
- Individuals picked for the sample may be inaccessible

Generating Random Numbers

- Pull scraps of paper out of a hat
- Run a pseudo-random number generator
- Sample a “noisy” source and have a computer “extract the entropy”
- Read the table of “random” numbers

Using R For Sampling

R code

```
sample(x, size, ...)

sample(1:100, 10)
[1] 26 90 86 73 4 39 38 37 81 34

sample(c("red", "green", "blue"), 1)
[1] "red"
```

1.4 Other Sampling Methods

Sampling Frame

Definition

A **sampling frame** is the source material or device from which a sample is drawn. It is a list of all those within a population who *can be sampled*, and may include individuals, households or institutions.

Example

If the population of interest consists of adult US citizens, then following frames can be used:

- List of names obtained from a census
- List of Social Security numbers
- List of phone numbers
- List of registered voters

Systematic Sample

Definition

A **systematic sample** is obtained by picking every k-th individual out of the sampling frame. The first individual is selected at random out of the first k.

Example

One can use systematic sampling for giving a service satisfaction survey to customers who visit a retail store on a given day.

Remark

The results may be biased if the frame is periodic, and the period happens to be related to k.

Obtaining a Systematic Sample

1. Estimate the population size N
2. Determine the desired sample size n
3. We want to obtain the sample by taking every k-th individual, so let $k \approx N/n$
4. Let p be a random number between 1 and k
5. Given a sampling frame, the sample will include individuals indexed by

$$p,\ p+k,\ p+2k,\ p+3k,\ \ldots$$

Stratified Sample

Definition

A **Stratified Sample** is obtained by partitioning the sampling frame into pairwise disjoint, exhaustive, and homogeneous subsets called **strata**, and taking either a simple random or a systematic sample out of each stratum.

Example

One can use stratified sampling to measure the mass of an adult spider of a given species. It makes sense to split this frame into males and females, since each group is homogeneous, but the respective sub-population parameters and the sampling procedures may be different.

Cluster Sample

Definition

A **Cluster Sample** is obtained by partitioning the sampling frame into pairwise disjoint, exhaustive, and heterogeneous subsets called **clusters**, and taking all the individuals within one or several randomly selected clusters. Ideally, each cluster is the population in miniature, with a lot of homogeneity between the clusters.

Example

One can use cluster sampling to measure the parameters of very young galaxies by partitioning the sky into many tiny patches and obtaining a high resolution photo of a few of them, each similar to the Hubble Ultra-Deep Field, which represents one thirteen-millionth of the total area of the sky, and yet required 11.5 days of exposure. It would take Hubble more than 400000 years to survey the sky at this resolution.

Multistage Sampling

Example

Which of the following is more suitable if a nation-wide grocery chain wants to sample out of the population consisting of all customers in a given year?

- Simple random sample
- Cluster sample (pick a few stores at random), followed by a simple random sample (pick a few dates at random), followed by a systematic sample of customers who visit chosen stores on chosen dates.

2.1 Organizing Qualitative Data

Frequency Distribution

Definition

An **(absolute) frequency distribution** lists each category of data and the number of occurrences in that category.

A **relative frequency** is the proportion of observations within a category and is found by dividing the corresponding frequency by the sum of all frequencies.

A **relative frequency distribution** lists each category of data and the corresponding relative frequency.

Using R for Tabular Display

R can construct absolute frequency distributions out of the data with `table(x)` and relative frequency distributions with

```
table(x) / length(x)
```

where `length(x)` returns the total number of data points.

Bar Graphs

Definition

A **bar graph** or **bar chart** is a chart with rectangular bars with lengths proportional to the values that they represent. The bars can be plotted vertically or horizontally. They must all have the same width, and be evenly spaced.

A bar graph can be used to display frequency tables, with each bar corresponding to a category, and lengths of bars proportional to either absolute or relative frequency.

Definition

A **Pareto chart** is a bar graph whose bars are drawn in decreasing order of frequency.

Using R for Bar Graphs

A qualitative data column `x` can be displayed as a bar plot with `barplot(table(x))`. A Pareto chart can be produced with

```
barplot(sort(table(x), decreasing=T))
```

If the list of categories is too long, a horizontal bar graph can be used. Adjust the margins if needed

```
par(mar=c(5.1,10.1,4.1,2.1))
```

and plot with `las=1` to rotate the labels

```
barplot(table(x), horiz=T, las=1)
```

Pie Charts

Definition

A **pie chart** (or a **circle graph**) is a circular chart divided into sectors, illustrating numerical proportion. In a pie chart, the arc length of each sector (and consequently its central angle and area), is proportional to the quantity it represents.

Pie charts are used for the same purpose as bar charts, but are considered inferior by many statisticians. They get overcrowded easily, and some studies have shown that comparing angles of sectors is harder than comparing lengths of bars.

Remark

In R, pie charts can be created with `pie(x)`.

2.2 Organizing Quantitative Data

Organizing Discrete Data

1. Determine the full range of the data
2. Partition the range into several intervals of equal length, called **classes**
3. Make a frequency table by treating classes as categories; the frequency of each class is the number of data points within.

Remark

Sometimes a data point falls on the boundary between two adjacent classes. (These boundaries are called **breaks**.) Such a point will belong to the class on the right. In other words, a class A–B corresponds to the interval $[A, B)$. We call such classes left-closed (or right-open), and they are far from conventional.

Class Frequency Table

Example

Outcomes of 12 tosses of a 6-sided die are

$$3, 2, 2, 1, 6, 4, 5, 6, 1, 1, 1, 4.$$

Let us assign a class to each possible die roll. For integer-valued data, it pays to assign breaks in between the possible values. This results in 6 classes: $[0.5, 1.5)$, $[1.5, 2.5)$, ... , $[6.5, 7.5)$, and the following frequency table:

Class	Frequency	Relative Frequency
0.5-1.5	4	33%
1.5-2.5	2	17%
2.5-3.5	1	8%
3.5-4.5	2	17%
4.5-5.5	1	8%
5.5-6.5	2	17%

Histograms of Discrete Data

Definition

A **histogram** is constructed by drawing rectangles for each **class** of data. The height of each rectangle is the frequency or relative frequency of the corresponding class. All rectangles must have the same width and, unlike in the bar chart, must be adjacent.

R code

```
hist(x)
```

Determining Number And Width of Classes

1. To determine the range of the histogram, find minimum and maximum observations in the data and set the lowest break somewhat below the min, and the highest break somewhat above the max, rounded for convenience. The range is the distance between these breaks.
2. Pick the number of classes between 5 and 20. Smaller numbers work better for small data sets, and big numbers for big ones.
3. Class width is equal to the range divided by the number of classes. We can round it for convenience, which will change the number of classes.

Remark

This approach works equally well for continuous data.

Stem-And-Leaf Plot

Definition

A **stem-and-leaf plot** (or **stem-and-leaf display**, or **stem plot**) is a device for presenting quantitative data in a graphical format, similar to a histogram. Assuming that the data has uniform precision, digits to the left of the right-most digit form a **stem**, and the right-most digit forms a **leaf**. The plot consists of two columns: the left column lists the stems in ascending order, and the right column lists sequences of corresponding leaves.

Making Stem Plot

1. Sort the data.
2. Decide on where to split the leaves from the stem in order to get the total of 5-20 classes. If the stems are too few, they may be **split**.
3. Write the stems vertically in ascending order and draw a vertical line to the right of the stems.
4. To the right of each stem, list the leaves in ascending order.
5. Indicate the position of the decimal point with respect to the vertical line.

Remark

Unlike histograms, stem-and-leaf displays retain the original data to at least two significant digits, and put the data in order

R code

```
stem(x)
```

Dot Plot

A **dot plot** is similar to a bar chart and a histogram, but now the individual observations are represented by little circles stacked on top of each other, so that the number of observations in a given category or class is proportional to the height of the stack.

R code

```
stripchart(x, method="stack", pch=20, at=0)
```

Distribution Shape

A histogram may reveal the general shape of the distribution.

- A **uniform** distribution will have all bars of about the same height.
- A **bell-shaped** distribution is a symmetric distribution with a prominent bulge in the middle and thin tails.
- A bell-shaped distribution that is not symmetric is said to be **skewed left** if the bulge is on the right, and **skewed right** if the bulge is on the left. This terminology is somewhat unconventional, since the skew has a strict definition, and the shape of the histogram is often misleading.

2.3 Additional Displays of Quantitative Data

Frequency Polygons

Definition

A **frequency polygon** is a line graph with a segment connecting each consecutive pair of midpoints of the top sides of a histogram.

R code

```
h <- hist(x)
lines(h$mids, h$counts, type="o", pch=20)
```

Cumulative Frequency Tables

Definition

A **cumulative frequency distribution** lists each category of data and the number of observations less than or equal to the corresponding category. For continuous data, it lists each class and the number of observations less than or equal to the corresponding upper class limit.

Definition

A **cumulative relative frequency distribution** is just like a cumulative frequency distribution, but lists the proportion rather than the number of observations.

R code

```
cumsum(table(x))
```

Frequency Ogives

Definition

A **frequency ogive** is a frequency polygon for a histogram representing a cumulative (absolute or relative) frequency distribution.

Time-series Graphs

Definition

A **time-series graph** or **plot** consists of points and segments and looks just like a frequency polygon, but now each point with coordinates (x, y) corresponds to a measurement y at time x.

Time-series graphs are useful for observing trends over time.

Example

Several time-series can be constructed using the data about terrorism patterns published by the United States Department of State.

3.1 Measures of Central Tendency

Mean

Definition

An (**arithmetic**) **mean** of a data set is the sum of all observations divided by the number of observations.

If $x_1, x_2, \ldots, x_N$ are the observations for each individual in a population of size N, then the **population mean**

$$\mu = \frac{x_1 + x_2 + \ldots + x_N}{N} = \frac{1}{N}\sum_{k=1}^{N} x_k$$

If $x_1, x_2, \ldots, x_n$ are the observations for each individual in a sample of size n, then the **sample mean**

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{k=1}^{n} x_k$$

More About Mean

Example

For a sample of size $n = 5$

$$10 \quad 30 \quad 45 \quad -15 \quad 20$$

the sample mean

$$\bar{x} = \frac{10 + 30 + 45 - 15 + 20}{5} = 24.$$

Remark

A population mean μ is computed using all the population members, and is a parameter. A sample mean $\bar{x}$ is computed from the sample data and is a statistic.

R code

```
mean(x)
```

Median

Definition

A **median** of a data set is a numerical value such that the number of data set values below it is the same as the number of data set values above it.

To compute the median of a finite data set, arrange it in ascending order. For a data set of odd size, the median is the value in the middle of the list. For a data set of even size, there is no middle value, and no single choice for the median, but it is usually defined as the mean of the middle two values.

Computing Median For Odd n

Remark

There is no widely accepted notation for median, but there is another summary called *second quartile*, or Q_2, which is defined to be equal to the median.

Example

Given a sample of size $n = 5$

$$10 \quad 30 \quad 45 \quad -15 \quad 20,$$

order it first

$$-15 \quad 10 \quad 20 \quad 30 \quad 45,$$

and then pick the middle value $Q_2 = 20$.

Computing Median For Even n

Example

Given a sample of size $n = 6$

$$10 \quad 30 \quad 45 \quad -15 \quad 20 \quad -10,$$

order it first

$$-15 \quad -10 \quad 10 \quad 20 \quad 30 \quad 45,$$

and then compute the mean of the two middle values

$$Q_2 = \frac{10 + 20}{2} = 15.$$

R code

```
median(x)
```

Resistant Statistics

Definition

A numerical summary of data is said to be **resistant** if extreme values (very large or very small) do not affect its value substantially.

Definition

A **breakdown point** of a numerical summary is the proportion of "incorrect" observations a summary can handle before giving an "incorrect" (say, arbitrarily large) result. A summary is considered resistant if it has a high breakdown point.

Mean Versus Median

Example

Given a data set of size n, adding a single very large observation is sufficient to make the mean arbitrarily large, so the breakdown point for the mean is 0%.

$$-15 \quad -10 \quad 10 \quad 20 \quad 30 \quad 10^6$$

On the other hand, to make the median arbitrarily large, one needs to add at least n very large observations, so the breakdown point for the median is 50%.

$$-15 \quad -10 \quad 10 \quad 20 \quad 30 \quad 10^6 \quad 10^6 \quad 10^6 \quad 10^6 \quad 10^6$$

Mode

Definition

The **mode** of a data set is the most frequent observation that occurs in that data set.

Remark

The mode can be computed for both quantitative and qualitative variables.

Remark

Two or more different values may occur with the same highest frequency, in which case the data set is called **bimodal** or **multimodal** respectively. In the latter case, the mode is usually not reported.

3.2 Measures of Dispersion

Dispersion Versus Central Tendency

Measures of central tendency describe the typical value of a variable. Often, we would also like to know the amount of dispersion in the variable. Dispersion is the degree to which the data are spread out.

Range

Definition

The **range** of a data set is the difference between the largest and the smallest value.

To compute the range, find the minimum value x_m, the maximum value x_M, and then the range $x_M - x_m$.

Naive Approach to Deviation

Naively, we could find the typical absolute deviation from some measure of central tendency, such as the median.

Definition

If the **absolute deviation** of a data point x is $|Q_2 - x|$, then the **median absolute deviation** of a data set is the median of the set of all absolute deviations.

Example

For a data set $\{1, 2, 3\}$, the median is 2, the set of corresponding absolute deviations is $\{1, 0, 1\}$, and the median of the latter is 1.

This measure, however, is not very popular, mainly because the absolute value function is not differentiable.

Standard Deviation

The traditional approach is to use a differentiable function with non-negative values.

Definition

The **population standard deviation** is the square root of the sum of squared deviations from the population mean, divided by the population size.

$$\sigma = \sqrt{\frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \ldots + (x_N - \mu)^2}{N}} = \sqrt{\frac{\sum_{k=1}^{N} (x_k - \mu)^2}{N}},$$

where $x_1, \ldots, x_N$ are the observations, N is the population size, and μ is the population mean.

Computing Standard Deviation

Example

Consider the population of size $N = 4$ with measurements $\{3, 4, 2, 11\}$. The population mean is

$$\mu = (3 + 4 + 2 + 11)/4 = 5,$$

and so the population standard deviation is

$$\sigma = \sqrt{\frac{(3-5)^2 + (4-5)^2 + (2-5)^2 + (11-5)^2}{4}} = \sqrt{\frac{50}{4}} = \frac{5}{\sqrt{2}}$$

Meaning of Standard Deviation

Remark

We take a square root in the end because without it we compute a typical *square* of the deviation from the mean. While it doesn't make the standard deviation an average deviation, it gives us a measure of spread in the same units as the individual measurements.

Sample Standard Deviation

Definition

The **sample standard deviation** of a sample is the square root of the sum of squared deviations from the sample mean, divided by $n-1$, where n is the sample size.

$$s = \sqrt{\frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\ldots+(x_N-\bar{x})^2}{n-1}} = \sqrt{\frac{\sum_{k=1}^{n}(x_k-\bar{x})^2}{n-1}},$$

where $x_1, \ldots, x_n$ are the observations, n is the sample size, and $\bar{x}$ is the sample mean.

Remark

The factor $(n-1)$ in the sample standard deviation formula is called *degrees of freedom*.

Meaning of Sample Standard Deviation

When we compare two populations, the one with the larger standard deviation is the one where the distribution of values is more dispersed, while the other one has the values more bunched up around the mean.

Remark

The primary purpose of computing the sample standard deviation is obtaining an estimate of the population standard deviation, and this is the key to understanding why we divide by $n - 1$ instead of n. For example, if the sample size is $n = 1$, then no meaningful estimate of spread can be made, so leaving s undefined is better than accepting an arbitrary value. On a deeper level, it can be proven that if we keep taking samples, then s averages out to be σ in every population, as long as the sampling is perfectly random.

Variance

Definition

The **variance** is the square of the standard deviation. The **population variance** is σ^2 and the **sample variance** is s^2.

Remark

While we defined variance as derived from the standard deviation, formally it is the other way around. Computing and manipulating the variance is easier (there is no square root), and so it plays a more fundamental role in theoretical Statistics.

Empirical Rule

The **Empirical Rule** states that if a distribution is approximately *normal* (symmetric and bell-shaped), then

- 68% of data lie within 1 standard deviation of the mean.
- 95% of data lie within 2 standard deviations of the mean.
- 99.7% of data lie within 3 standard deviation of the mean.

Remark

It may be useful to remember a different set of numbers: 2%, 13.5%, and 34%.

Example

If the population mass distribution is approximately normal with the mean of 60 kg and standard deviation of 10 kg, what proportion of the population has mass between 40 and 70 kg?

Chebyshev's Inequality

Theorem (Chebyshev's Inequality)

For any data set or distribution, at least $\left(1 - \frac{1}{k^2}\right) 100\%$ *of all observations fall within k standard deviations of the mean, where k is any real number greater than 1.*

Example

- At least 75% of all data is within 2 standard deviations of the mean ($k = 2$).
- At least 88% of all data is within 3 standard deviations of the mean ($k = 3$).
- At least 96% of all data is within 5 standard deviations of the mean ($k = 5$).

ER versus CI

Remark

The Empirical Rule may seem more powerful because it gives a higher bound on the proportion of the population close to the mean. It also copes well with intervals that are not symmetric with respect to the population mean. But unlike the Empirical Rule, the Chebyshev's Inequality makes no assumptions about the shape of the distribution, and so it provides an accurate bound in every possible situation.

3.4 Measures of Position

z-score

Definition

The **population z-score** of a measurement x is $z = \dfrac{x - \mu}{\sigma}$,

while the **sample z-score** of a measurement x is $z = \dfrac{x - \bar{x}}{s}$.

Remark

The z-score is negative if the measurement is to the left of the mean, and positive if it is to the right. Its absolute value is great for observations far away from the mean, and close to zero otherwise.

Example

The z-score tells us how "typical" a particular measurement is. Only a small proportion of the population falls far away from the mean, so a z-score of 5 is pretty atypical.

Percentiles and Quartiles

Definition

The k-**th percentile**, denoted P_k, is a value such that k percent of observations fall at or below it. The median, in particular, happens to be P_{50}.

Definition

The **first quartile**, denoted Q_1, is the 25-th percentile, the **second quartile** Q_2 is the 50-th percentile, and the **third quartile** Q_3 is the 75-th percentile.

Interquartile Range

Definition

The **interquartile range** or **IQR** is the distance from Q_1 to Q_3:

$$\text{IQR} = Q_3 - Q_1.$$

Computing Quartiles

1. For a data set $\{x\}$ of size one, $Q_1 = Q_3 = x$.
2. For a large data set, arrange the data in ascending order.
3. Determine the median Q_2.
4. Divide the data into two halves: below and above Q_2 respectively. Q_1 is the median of the lower half, and Q_3 is the median of the upper half.

Example

10	12	13	13	17	18	29
	Q_2		Q_2		Q_3	

Outliers

Definition

An **outlier** is an observation that is numerically distant from the rest of the data.

Remark

Outliers should be investigated. They could indicate a measurement error or a heavy-tailed distribution. In the former case, one can use statistics that are robust to outliers. In the latter case, one should be careful not to assume a normal distribution.

Checking for Outliers Using Quartiles

1. Determine IQR.
2. Determine the **lower fence** $\text{LF} = Q_1 - 1.5(\text{IQR})$ and the **upper fence** $\text{UF} = Q_3 + 1.5(\text{IQR})$.
3. A data point below LF or above UF is considered an outlier.

Example

$$\begin{array}{ccccccc} 10 & 12 & 13 & 13 & 17 & 18 & 29 \\ & Q_2 & & Q_2 & & Q_3 & \end{array}$$

$$\begin{aligned} \text{IQR} &= 18 - 12 = 6, \\ \text{LF} &= 12 - 1.5 \cdot 6 = 3, \\ \text{UF} &= 18 + 1.5 \cdot 6 = 27, \end{aligned}$$

so 29 is the only outlier.

3.5 Five Number Summary And Box Plots

Obtaining Five Number Summary

Definition

The **five number summary** of a data set consists of the minimal value, Q_1, Q_2, Q_3, and the maximum value.

Example

9	10	12	12	13	13	17	17	18	29	31
min		Q_2			Q_2			Q_3		max

Constructing Box Plot

1. Above a labeled coordinate axis, draw a box staring at Q_1 and ending at Q_3.
2. Draw a vertical line through the box at Q_2.
3. Draw a horizontal line from the lowest data point at or above the lower fence to Q_1.
4. Draw another line from Q_3 to the highest data point at or below the upper fence.
5. Label outliers with asterisks $*$.

Remark

The horizontal lines are sometimes called **whiskers**, and the plot a **box-and-whisker plot**.

4.1 Scatter Diagrams And Correlation

Scatter Diagrams

Definition

A **scatter diagram** is a graph that shows a relationship between two variables. Each individual in the sampling frame is represented by a point (x,y), where x is the measurement of the explanatory variable, and y is the measurement of the response variable.

Definition

Two variables are **positively associated** when larger values of one variable tend to correspond to larger values of the other variable. They are **negatively associated** when larger values of one variable tend to correspond to smaller values of the other variable.

Linear Correlation Coefficient

Definition

The **linear correlation coefficient** (or Pearson product-moment correlation coefficient, or PCC) is a measure of the strength and the direction of the linear relation between two quantitative variables. ρ represents the population linear correlation coefficient, and r represents the corresponding sample statistic.

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E((X-\mu_X)(Y-\mu_Y))}{\sigma_X \sigma_Y}$$

$$r = \frac{1}{n-1} \sum_{k=1}^{n} \left(\frac{x_k - \bar{x}}{s_x} \right) \left(\frac{y_k - \bar{y}}{s_y} \right) = \frac{\sum_{k=1}^{n} (x_k - \bar{x})\,(y_k - \bar{y})}{(n-1) s_x s_y}$$

Properties of Linear Correlation Coefficient

1. $r \in [-1, 1]$
2. $r = 1$ implies a perfect linear relation
3. $r = -1$ implies a perfect negative linear relation
4. if r is far away from zero when a linear relation (positive or negative) is strong, and close to zero when it is weak
5. r is unitless
6. r is not resistant

Correlation Versus Causation

An observational study may show a high degree of correlation, but won't allow us to conclude that high values of one variable *cause* the high values of the other variable. In a designed experiment, we could attempt to remove all lurking variables and claim that correlation implies causation.

Simple Linear Regression

Definition

The **simple linear regression** is the least squares estimator of a linear regression model with a single explanatory variable. In other words, simple linear regression fits a straight line through the set of n points in such a way that makes the sum of squared residuals of the model (that is, vertical distances between the points of the data set and the fitted line) as small as possible.

$$\hat{y} = b_1 x + b_0, \text{ where}$$

$$b_1 = r\frac{s_y}{s_x},$$

$$b_0 = \bar{y} - b_1\bar{x}$$

5.1 Probability Axioms

Definitions and Notation

Definition

A **probability experiment** is any process with uncertain results that can be repeated.

Definition

The **sample space** S of a probability experiment is the collection of all possible outcomes. An **event** is any collection of possible outcomes (in other words, a subset of the sample space S).

Definition

The **probability** of an event E is the real-valued measure of how likely the event to occur, denoted $P(E)$. Higher probability values correspond to events that are more likely to occur.

Definition and Notation

Definition

A **union** of events E and F is the event which occurs if either E or F or both occur:

$$E \cup F = E \text{ or } F.$$

An **intersection** of events E and F is the event which occurs if both E and F occur:

$$E \cap F = E \text{ and } F.$$

Definition

An event E is **impossible** or **null** if it never occurs: $P(E) = 0$.

Definition

Events E and F are **mutually exclusive** or **disjoint** if they have no outcomes in common. This implies $P(E \cap F) = 0$.

Examples of Experiments

Example

Look outside to see whether it's raining.

$$S = \{R, R^c\}$$

Example

Roll a six-sided die, let it come to a stop, and record the number X on the top face. Let E_1 be the event "$X > 4$" and E_2 be the event "X is odd".

$$S = \{1, 2, 3, 4, 5, 6\}$$

$$E_1 = \{5, 6\}$$

$$E_2 = \{1, 3, 5\}$$

Kolmogorov Axioms

1. $P(E) \geq 0$
   The probability of any event is non-negative.
2. $P(S) = 1$
   The probability that some event will occur is 1. Here S is the sample space, or the union of all possible events.
3. For mutually exclusive events E_1, E_2, E_3, ...

$$P(E_1 \cup E_2 \cup E_3 \cup \ldots) = \sum_{k=1}^{\infty} P(E_k)$$

   The probability of the union of mutually exclusive events is the sum of probabilities of individual events.

Probability Model

Definition

A **probability model** lists the possible outcomes of an experiment and the associated probabilities. A probability model must be consistent with the probability axioms.

Example

Look outside and determine whether it's raining, snowing, neither, or both. A corresponding sample space could look like

$$S = \{E_R, E_S, E_N, E_B\},$$

where all events are mutually exclusive. A model may assign probabilities as follows: $P(E_R) = 0.2$, $P(E_S) = 0.1$, $P(E_N) = 0.6$. Note that $P(E_B) = 0.1$ necessarily, since the probabilities of the four disjoint events have to add up to $P(S) = 1$ (axioms 2 and 3).

Empirical Probability

Definition

The **empirical probability** model is built up from the observed sample proportions. An experiment is conducted n times, and the probability of an event E is approximated by the number of times E occurred, divided by n.

Example

A random sample of 10000 new-born children is selected out of the population of children born in 2012. 5125 of the children are male, and 4875. A corresponding empirical probability model for sex ratio at birth will have the sample space $S = \{M, F\}$, with

$$P(M) = \frac{5125}{10000} = 0.5125 \quad \text{and} \quad P(F) = \frac{4875}{10000} = 0.4875.$$

Classical Probability

Definition

In the classical interpretation of probability, the event space consists of n equally likely, mutually exclusive elementary events. The axioms imply that the probability of any event E is the number of ways E can occur divided by n.

Example

The experiment consisting of rolling a fair six-sided die and recording the value X on the top face can be modeled by a probability space with $n = 6$ equally likely outcomes. If E is the event "X is even", then there are 3 elementary events which make E happen: $X = 2$, $X = 4$, and $X = 6$, so

$$P(E) = P(X \text{ is even}) = P(\{2, 4, 6\}) = \frac{3}{6} = \frac{1}{2}$$

Law of Large Numbers

Theorem (LLN Preview)

If we keep repeating a probability experiment, then the empirical probability of an event E will tends to its true value.

Example

Tossing a fair coin n times may yield any number of Heads between 0 and n. If we let n be very large, though, we expect to see about $n/2$ Heads in the sample, and the proportion should get better as n gets larger.

R code

```
barplot(table(round(runif(n))))
```

5.2 The Addition Rule And Complements

The General Addition Rule

Theorem

For any two events E and F,

$$P(E \cup F) = P(E \text{ or } F) = P(E) + P(F) - P(E \cap F).$$

Proof.

Let E' be the event when E occurs and F doesn't, and let F' be the event when F occurs and E doesn't. Using the axiom 3 we can write

$$P(E) = P(E') + P(E \cap F),$$

$$P(F) = P(F') + P(E \cap F),$$

and hence

$$P(E \cup F) = P(E') + P(F') + P(E \cap F) = P(E) + P(F) - P(E \cap F).$$

The Complement Rule

Definition

If S is the sample space and E is any event, then the **complement of** E is the event E^c which consists of all outcomes that are not in E.

Theorem

For all events E in the sample space S, $P(E^c) = 1 - P(E)$.

Proof.

Since E and E^c are mutually exclusive, axioms 2 and 3 imply

$$P(E) + P(E^c) = P(E \cup E^c) = P(S) = 1.$$

□

Elaborate Example

Example

A deck of playing cards consists of 52 cards. There are four suits (spades, hearts, diamonds, and clubs), with 13 cards in each suit, from ace to king. Our experiment is to draw one card at random. How likely is each of the following draws?

- E_1 Any spade
- E_2 Any king
- E_3 The king of spades
- E_4 Either a king or a spade
- E_5 Neither a spade nor a king

Contingency Table

Definition

A **contingency table** (or cross tabulation, or cross tab) is a type of table that displays the (multivariate) frequency distribution of the variables. The row events form a partition of the event space: they are pairwise mutually exclusive and their union is the entire event space. Ditto for the column events.

Example

	Right-handed (R)	Left-handed (L)	Total
Males (M)	43	9	52
Females (F)	44	4	48
Total	87	13	100

$P(M) = 0.52$, $P(L) = 0.13$, $P(F \cap L) = 0.04$

$P(F \cup L) = (44 + 4 + 9)/100 = P(F) + P(L) - P(F \cap L) = 0.57$

5.3 Independence and the multiplication rule

Independence and the Multiplication Rule

Definition

Two events A and B are **independent** if and only if

$$P(A \cap B) = P(A)P(B).$$

Remark

Independent events are used to model situations when the occurrence of an event A does not affect the probability of the occurrence of B.

Example

Toss a six-sided die and let $E = \{1, 3, 5\}$ and $M = \{1, 6\}$.

Example

What can we say about probabilities of events A and B if they are both independent and mutually exclusive?

Generalizations of Independence

Definition

Events $E_1, E_2, E_3, \ldots, E_n$ are **pairwise independent** if and only if E_i and E_j are independent for all i and j from 1 to n.
The same events are **mutually independent** if and only if the multiplication rule holds for all subsets of $\{E_1, E_2, \ldots, E_n\}$. In particular,

$$P(E_1 \cap E_2 \cap \ldots \cap E_n) = P(E_1)P(E_2)\ldots P(E_n).$$

Remark

Mutual independence implies pairwise independence, but not the other way around.

Pairwise Independence

Example

	R	L	Total
M	0.25	0.25	0.5
F	0.25	0.25	0.5
Total	0.5	0.5	1.00

Let $E = (F \cap R) \cup (M \cap L)$. Are events M, R, and E pairwise independent? Are they mutually independent?

5.4 Conditional Probability and the General Multiplication Rule

Conditional Probability

Definition

The probability of the event F, given that the event E has already occurred is

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

Remark

$P(F|E)$ is not defined if $P(E) = 0$.

Conditional Probability from Table

Example

1. How likely a freshman is to run GNU/Linux?
2. Given that Alice is running Windows or OS X, what is the probability of her being a sophomore?
3. How likely is Bob to run GNU/Linux or OS X if he is not a senior?

	W	X	G	Total
E_1	10	10	0	20
E_2	8	8	3	19
E_3	5	10	4	19
E_4	12	17	13	42
Total	35	45	20	100

Table 1 : Columns are **W**indows, OS **X**, and **G**NU/Linux, and rows are Freshman, Sophomore, Junior, and Senior.

Answers

1. $P(G|E_1) = \dfrac{0}{0.2} = 0$
2. $P(E_2|W \cup X) = \dfrac{0.16}{0.8} = 0.2$
3. $P(X \cup G|E_4^c) = \dfrac{0.35}{0.58} \approx 0.60$

General Multiplication Rule

Theorem

For any two events E and F with $P(E) \neq 0$,

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

Example

A shopping bag contains 13 apples: 3 of them are Macoun and 10 are MacIntosh. Pull two apples out of the bag without looking. What is the probability that

1. both are Macoun?
2. one is Macoun and the other one is MacIntosh?

5.5 Counting Techniques

Motivation

While working with a classical model for a given experiment, it is often necessary to count all the ways a given event can happen.

Example (Seven-card Stud)

In the game of seven-card stud each player gets seven cards out of the standard 52-card deck. If one believes that every possible arrangement (or **hand**) is equally likely, then a classical model can be built. But when we think about concrete events, difficult questions arise.

1. How many different arrangements are there? In other words, what is the size of the sampling space?
2. How many hands will have at least one ace?
3. How many will make a flush? (Five or more suited cards.)
4. How many will make four of a kind?
5. How many will make a royal flush? (suited TJQKA.)
6. What are the probabilities of the events listed above?

Factorial

Definition

For any non-negative integer n, the **factorial** of n, or "n factorial" is

$$n! = 1 \cdot 2 \cdot 3 \cdot \ldots \cdot (n-1) \cdot n,$$

while $0! = 1$.

Remark

Factorials quickly become large and unruly, but can cancel really well in fractions:

$$\frac{100!}{98!} = \frac{1 \cdot 2 \cdot \ldots \cdot 98 \cdot 99 \cdot 100}{1 \cdot 2 \cdot \ldots \cdot 98} = 99 \cdot 100 = 9900,$$

$$\frac{(n+1)!}{(n-1)!} = \frac{1 \cdot 2 \cdot \ldots \cdot (n-1) \cdot n \cdot (n+1)}{1 \cdot \ldots \cdot (n-1)} = n(n+1).$$

Multiplication Rule of Counting

The **multiplication rule of counting** or the **rule of product** is a basic counting principle. It is the idea that if there are a ways of completing the first part of the task and b ways of completing the second part of the task, then there are ab ways of performing the task. It generalizes readily to situations with more than two parts to the task.

Example (Salad)

Suppose a salad must be made with greens, meat, and dressing. If there are 3 choices for greens, 2 for meat, and 7 for dressing, then how many different salads can be made?

Definition

To obtain a random sample of size n **with replacement**, do the following n times: pick a random individual from the population, record his name and measurements, and place him back into the population.

To obtain a random sample of size n **without replacement**, simply pick n distinct individuals out of the population.

Permutations

Example (Race Stats)

The final outcome of a race is the list containing names of runners who finished 1st, 2nd, and 3rd, in that order. If there are ten runners in a race, how many different outcomes can the race have?

Proposition

The number of arrangements of r distinct objects chosen from n distinct objects, where the order of the arrangement is important, is called the number of ***permutations****, and is given by*

$$ {}_nP_r = \frac{n!}{(n-r)!} = (n-r+1)\dots(n-2)(n-1)n. $$

It follows that number of ways to ***permute*** *n distinct objects, or to arrange them in order, is $n!$.*

Combinations

Example (Race Sample)

Out of ten runners in a race, three will be randomly chosen to undergo a drug test. How many different ways are there to make this choice?

Proposition

The number of arrangements of r distinct objects chosen from n distinct objects, where the order of the arrangement is not important, is called the number of ***combinations****, and is given by*

$$ {}_nC_r = \frac{n!}{r!(n-r)!} $$

Combinations and Rule of Product

Example (Sack of Apples)

A burlap sack contains 42 apples: 10 of them are Macoun, 10 are MacIntosh, and 22 are Fuji. Pull 15 apples out of the sack without looking. What is the probability that

1. exactly five are Macoun?
2. there are exactly five of each variety?

Counting with Repetition

Example (Combination Lock)

A combination lock on a leather briefcase has 6 wheels with digits 0 through 5 on each wheel. How many different combinations of digits are there?

Example (Password)

An online retailer website is asking users to create passwords at least 6 characters long. If a password string is only allowed to have Latin letters (both cases) and Arabic digits, then how many different 6-character passwords are there?

Proposition

The number of literal strings of length r, where characters are chosen from n distinct characters, is n^r.

Counting Techniques Summary

	with replacement	without replacement
order important	n^r	${}_nP_r = \dfrac{n!}{(n-r)!}$
order not important	$\dfrac{(n+r-1)!}{(n-1)!r!}$	${}_nC_r = \dfrac{n!}{r!(n-r)!}$

Table 2 : The number of ways to sample r objects out of n objects.

More Counting Techniques

- The number of ways to put n distinct things into r distinct baskets.
- The number of ways to put n distinct things into r indistinct baskets (the number of partitions of a set, given by the Bell number).
- The number of ways to arrange in order n_1 indistinct objects of the first type, n_2 indistinct objects of the second type, and so on to n_r indistinct objects of the r-th type.

Example

Constrained Password If a password has to consist of 3 letters "a", 4 letters "b", and 5 letters "c", then there are $\frac{(3+4+5)!}{3!4!5!}$ different passwords.

6.1 Discrete Random Variables

Random Variables

Definition

A **real-valued random variable** X is a function from the sample space into the real line such that the probability $P(X \leq a)$ can be computed for all real numbers a. We will think of random variables as of experiments they model, and use notation such as $P(a < X < b)$ to express the probability of the event when the outcome of the experiment falls between a and b.

Discrete and Continuous Random Variables

Definition

A random variable is **discrete** if it has finitely or countably many possible outcomes. A random variable is **continuous** if it has a probability density function (pdf).

Remark

While the definition does not require it, most of the discrete random variables in use have isolated outcomes, with an empty neighborhood about each one. In this case, all the possible values of a discrete random variable can be listed in a single table, in the ascending order.

Remark

Most useful continuous random variables take possible values from an interval (or intervals) on the real line. which is how we can identify them until we learn about pdf.

Discrete Random Variables

Let X be a discrete random variable.

Definition

The **probability mass function** (or **pmf**) of X is the table (or a formula) which determines the probability of X assuming each possible value:

$$f_X(a) = P(X = a).$$

The **cumulative distribution function** (or **cdf**) of X is the table (or a formula) which determines the probability of X being less than or equal to each possible value:

$$\begin{aligned} F_X(x_k) &= P(X \leq x_k) \\ &= P(X = x_1) + P(X = x_2) + \ldots + P(X = x_k), \end{aligned}$$

where $x_1, \ldots, x_k$ are all the possible values of X at or below x_k, in ascending order.

Discrete RV Example

Example

Suppose a game involves tossing a fair coin, and the player receives \$20 if it shows Tails, and has to pay \$10 if it shows Heads. We can model the outcome by X with the following pmf:

x	$P(X = x)$
-10	0.5
20	0.5

or the corresponding cdf:

x	$P(X \leq x)$
-10	0.5
20	1

Discrete RV Example

Example

Suppose a game involves tossing a fair six-sided die, and the player pays \$50 for one dot, receives \$100 for six dots, and breaks even otherwise. We can model the outcome by Y with the following pmf:

y	$P(Y = y)$
-50	1/6
0	4/6
100	1/6

or the corresponding cdf:

y	$P(Y \leq y)$
-50	1/6
0	5/6
100	6/6

More on Discrete Distributions

- For any random variable X and any a,

$$0 \leq f_X(a) = P(X = a) \leq 1.$$

- The sum of probabilities in the pmf table (or the last entry in the cdf table) is always 1.

Definition

A **probability histogram** for a discrete random variable is a relative frequency histogram corresponding to the pdf.

Mean

Definition

Suppose that X is a discrete random variable with the list of all possible values

$$x_1, x_2, \ldots, x_k, \ldots$$

The **mean** (or **expected value**) of X is

$$\mu_X = EX = \sum_k x_k P(X = x_k).$$

Example

Using the variables from the examples just above,

$$EX = -10 \cdot 0.5 + 20 \cdot 0.5 = 5,$$

$$EY = -50 \cdot \frac{1}{6} + 0 \cdot \frac{4}{6} + 100 \cdot \frac{1}{6} = 25/3 = 8 + \frac{1}{3}$$

Standard Deviation

Definition

Suppose that X is a discrete random variable with the list of all possible values

$$x_1, x_2, \ldots, x_k, \ldots$$

The **standard deviation** of X is

$$\begin{aligned}
\sigma_X &= \sqrt{\sum_k (x_k - \mu_X)^2 \cdot P(X = x_k)} \\
&= \sqrt{\sum_k x_k^2 P(X = x_k) - \mu_X^2} \\
&= \sqrt{E(X^2) - (EX)^2}
\end{aligned}$$

6.2 Binomial Distribution

Binomial Experiment

Definition

The **Bernoulli experiment** (or **Bernoulli trial**) is an experiment with exactly two disjoint outcomes labeled as "success" and "failure" (or 1 and and 0 respectively), with $P(1) = p$ and $P(0) = 1 - p$.

Definition

An experiment is a **binomial experiment** if it consists of a fixed number of mutually independent Bernoulli experiments, each with probability of success p.

Recognizing Binomial Experiment

Remark

To establish whether or not a given experiment is binomial, check for these properties:

- consists of n identical trials,
- trials are mutually independent,
- each trial has exactly two mutually exclusive outcomes, "success" and "failure", and
- the probability of *success* is the same in each trial.

Binomial Examples

Example

Are these binomial experiments?

1. Toss 100 fair coins.
2. Toss 100 fair six-sided dice.
3. Take a random sample of Bostonians (with replacement) and write down their genders.
4. Look out of the window every 5 seconds for 10 minutes and establish whether it's raining.

Binomial pmf

When we use a random variable (say, X) to model a binomial experiment, we have X count the number of successful trials. That is, if n is the total number of trials, then possible values of X are $0, 1, 2, \ldots, n$.

Definition

X is a **binomial random variable** if its pmf is

$$f_X(x) = P(X = x) = {}_nC_x p^x (1-p)^{n-x},$$

where n is the total number of trials, p is the probability of success, and x is the number of successful trials.

$$X \sim \text{Binom}(n, p)$$

Computing Binomial Probabilities

Example (Disk Drives)

In order to back up sensitive data, a journalist purchased 10 identical disk drives, uploaded an encrypted copy of the data onto each of them, and placed them into a safe deposit box for one year. Suppose that a disk drive fails within one year with probability 0.1, and that different drives do so independently. After one year, the drives will be taken out of the bank and inspected. What is the probability that

1. none of the drives fail
2. two or fewer drives fail
3. at least one drive survives
4. every drive fails

Mean and Standard Deviation

Theorem

A binomial random variable $X \sim \text{Binom}(n, p)$ *has mean*

$$\mu_X = EX = np$$

and standard deviation

$$\sigma_X = \sqrt{np(1-p)}.$$

Example (Disk Drives)

For $X \sim \text{Binom}(n = 10, p = 0.1)$, $EX = 1$, and

$$\sigma_X = \sqrt{10 \cdot 0.1 \cdot 0.9} = \sqrt{0.9}.$$

Shape of the Binomial Distribution

For large n, the binomial distribution histogram looks a lot like the normal distribution: it becomes bell-shaped and symmetric about the mean. In fact, it also obeys the Empirical rule, and as n grows, a binomial distribution converges to a corresponding normal distribution.

R code

```
barplot(dbinom(0:n,n,p))
```

7.1 Properties of the Normal Distribution

Probability Density Function

Definition (Informal)

Probability Density Function, or **pdf**, is used to compute the probabilities associated with continuous random variables. It must be non-negative, integrable, and the total area under the graph of a pdf must be equal to one. If $F_X(x)$ is a pdf associated with the random variable X, then

$$P(X \in [a,b]) = \int_a^b f_X(x)dx.$$

Example

Let $f_X(x) = x/2$ for $x \in [0,2]$, and let $f_X(x) = 0$ otherwise.

Uniform Distribution

Definition

X has a **uniform** probability distribution, or is **uniformly distributed** between a and b, written as

$$X \sim U(a,b),$$

iff

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a,b] \\ 0 & \text{if } x \notin [a,b] \end{cases}$$

Uniform Example

Example

Suppose that the large bowl is used to mix the pancake batter, and then the pancakes are baked by using 1/2 cup of batter for each. The volume of batter X that is left over is distributed uniformly between 0 and 0.5 cup. What is the probability that

1. more than 0.1 cup of batter is left over?
2. between 1/3 and 2/3 cups of batter is left over?
3. no batter is left over?

Lemma

If $X \sim U(a,b)$, *and* $c,d \in [a,b]$, *then*

$$P(X \in [c,d]) = \frac{d-c}{b-a}.$$

Cumulative Distribution Function

Definition

For any random variable X, the **cumulative distribution function**, or **cdf**, written as $F_X(x)$, is the function such that

$$P(X \leq x) = F_X(x).$$

It is immediate that if X has a pdf, then

$$F_X(x) = \int_{-\infty}^{x} f_X(u)du.$$

Example

A uniform X has a triangular cdf.

Definition

The **mean** (or **expected value**) of a continuous random variable X with pdf $f_X(x)$ is

$$\mu_X = EX = \int_{\mathbb{R}} x f_X(x) dx.$$

The **variance** and the **standard deviation** are defined as before:

$$\sigma_X = \sqrt{E(X^2) - (EX)^2}.$$

The **median** m leaves half of the area under the pdf on its left:

$$F_X(m) = 1/2.$$

Normal Distribution

Definition

A random variable X has **normal distribution**, or is **normally distributed** with mean μ and standard deviation σ, written as

$$X \sim N(\mu, \sigma),$$

iff it has a pdf $f_X(x)$, and

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

A random variable Z has the **standard normal distribution** iff

$$Z \sim N(\mu = 0, \sigma = 1).$$

Properties of the Normal Distribution

For every $X \sim N(\mu, \sigma)$,

1. mean = median = mode,
2. the pdf $f_X(x)$ is symmetric about the mean,
3. the pdf has inflection points at $\mu - \sigma$ and $\mu + \sigma$,
4. the area under the pdf is one (duh),
5. the area under the pdf to the left (or the right) of the mean is $1/2$,
6. the tails get thin fast, but never touch the x-axis,
7. X obeys the empirical rule,
8. the pdf is insensitive to the inequality strictness.

Motivation

Approximately normal distributions tend arise in nature wherever the outcome of an experiment is a sum of many outcomes of mutually independent experiments, or when the measured variable is affected by many mutually independent factors. The ultimate insight into why this happens is provided by the *Central Limit Theorem* (CLT).

7.2 Applications of Normal Distribution

Standardizing a Normal Random Variable

Theorem

If X is a normal random variable with mean μ and standard deviation σ, then

$$Z = \frac{X - \mu}{\sigma}$$

has the standard normal distribution $N(\mu = 0, \sigma = 1)$.

It follows that

$$P(X \leq x) = P\left(Z \leq \frac{x - \mu}{\sigma}\right),$$

which gives us a way to compute normal probabilities with the help of a single pdf: that of a standard normal distribution.

Normal Probabilities

Corollary

If $X \sim N(\mu_X, \sigma_X)$, and F_Z is the cdf for the standard normal distribution, then

$$\begin{aligned} P(X < x) &= F_Z\left(\frac{x-\mu_X}{\sigma_X}\right), \\ P(X > x) &= 1 - F_Z\left(\frac{x-\mu_X}{\sigma_X}\right), \\ P(a < X < b) &= F_Z\left(\frac{b-\mu_X}{\sigma_X}\right) - F_Z\left(\frac{a-\mu_X}{\sigma_X}\right), \end{aligned}$$

and the k-th percentile of the distribution of X is

$$x_k = \mu_X + \sigma_X z_k,$$

where z_k is the k-th percentile of the standard normal distribution.

Application of Normal Probabilities

Example

Suppose that a test grade is approximately normally distributed with mean 85 and standard deviation 11. Find the proportion of students with grades

1. above 85,
2. between 85 and 96,
3. between 70 and 80.

Also, find the test scores corresponding to the 50-th, the 90-th, and the 95-th percentiles of this distribution.

7.4 Normal Approximation to the Binomial

Normal Approximation

For large enough values of $np(1-p)$, a binomial variable

$$X \sim \text{Binom}(n, p)$$

can be approximated by a normally distributed

$$Y \sim N(\mu, \sigma),$$

where $\mu = np$ and $\sigma = \sqrt{np(1-p)}$.

Remark

Historically, this is the first practical application of the normal distribution (1738). The proof of its effectiveness relies on the CLT, since the outcome of a binomial experiment is precisely the sum of n mutually independent Bernoulli trials.

Continuity Correction

In practice, to compute $P(a \leq X \leq b)$, where $X \sim \text{Binom}(n, p)$, $\mu = np$, $\sigma = \sqrt{np(1-p)}$, and $Y \sim N(\mu, \sigma)$, we can calculate

$$P(a - 0.5 < Y < b + 0.5).$$

Example

If 26000 people participate in Boston Marathon in 2014, and the historical finishing rate is 80% (that is, 4 out of 5 participants are expected to reach the finish line by running or walking), then what is the probability that

1. between 20700 and 20900 participants will finish the race?
2. exactly 20800 participants will finish the race?

8.1 Distribution of the Sample Mean

Sampling Distribution

Definition

The **sampling distribution** of a statistic is the probability distribution for all the possible values of the statistic computed from a sample of size n.

Definition

The **sampling distribution of the sample mean** $\overline{X}$ is the probability distribution of all possible values of the random variable $\overline{X}$ computed from a sample of size n from a population with mean μ and standard deviation σ.

Sampling a Normal Distribution

Theorem

If a random variable X is normally distributed with mean μ_X and standard deviation σ_X, then the sampling distribution of the sample mean is also normally distributed with mean $\mu_{\overline{X}} = \mu_X$ and standard deviation

$$\sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{n}}$$

Definition

The standard deviation of the sampling distribution of $\overline{x}$

$$\sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{n}}$$

is called the **standard error of the mean.**

Sampling Distribution Example

Example

Suppose that the the waiting time at an emergency room is approximately normally distributed with mean 50 minutes and standard deviation 10 minutes. Find the probability that

1. a random patient has to wait more 60 minutes.
2. patients in a random sample of size 4 have to wait more than 60 minutes on average.
3. patients in a random sample of size 40 have to wait more than 60 minutes on average.

Law of Large Numbers

Theorem (Law of Large Numbers)

Let $X_1, \ldots, X_n$ be random variables, each with the same mean μ, and let $E|X_i|$ be finite for each $i = 1, \ldots, n$. We can think of $X_1, \ldots, X_n$ as of a random sample of size n from a distribution with mean μ. Then the sample average

$$\overline{X} = \frac{1}{n}(X_1 + X_2 + \ldots + X_n)$$

tends to μ as n tends to infinity.

Central Limit Theorem

Theorem

Let $\{X_1, \ldots, X_n\}$ be a random sample of size n; that is, a sequence of mutually independent and identically distributed random variables drawn from a distribution with mean μ_X and standard deviation σ_X. Then, as n tends to infinity, the sampling distribution of the sample mean

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

tends to the normal distribution with mean μ_X and standard deviation $\sigma_X/\sqrt{n}$. Put in different terms, the distribution of

$$\frac{\overline{X} - \mu_X}{\sigma_X/\sqrt{n}}$$

tends to that of the standard normal.

Using CLT

Remark

How big does the sample size n have to be before the distribution of the sample mean $\overline{X}$ is approximately normal? If X is normal, then 1 is enough, but if X is suspected to be very asymmetric and heavy-tailed, then samples of size about 30 are suggested.

A practical statistical question often sounds like this: if a population has mean μ_X and standard deviation σ_X, and samples of size n are taken, then what is the probability that the sample mean is at or below a certain value a? Well, the z-score of a particular sample mean $\overline{x}$ is

$$z = \frac{\overline{x} - \mu_X}{\sigma_X / \sqrt{n}}$$

and so

$$P(\overline{X} \leq a) \approx P\left(Z \leq \frac{a - \mu_X}{\sigma_X / \sqrt{n}}\right)$$

CLT Application

Example

Suppose that at the end of each business day a cash register at a spice shop shows a discrepancy with mean -1 USD and standard deviation 14 USD (that is, 1 dollar is missing on average, but extra money appears quite often too). Assuming that discrepancies on different days are mutually independent, and that the sampling distribution of the sampling mean is approximately normal, what is the probability that after 100 business days,

1. more than 100 USD is missing.
2. no money is missing (0 or more USD are gained).
3. the discrepancy is between -140 and 140 USD.
4. more than 600 USD is missing.

Is there a way to estimate the probability of a single-day discrepancy with absolute value above 140 USD?

Hypothesis Testing Conventions

Remark

While the following numbers do not represent a consensus among statisticians, we will use them in this class to make informal decisions.

1. The sampling distribution of the sample mean is approximately normal for samples of size 30 or more.
2. For a binomial distribution Binom(n, p) to be well approximated by a normal distribution, we demand $np(1-p) \geq 10$.
3. An event is "unlikely" or "surprising" if its probability is below 5%.

8.2 Distribution of the Sample Proportion

Sample Proportion

Definition

Suppose a random sample is drawn from a population where each individual does or does not have a certain characteristic. Then the **sample proportion** $\hat{p}$ is

$$\hat{p} = \frac{x}{n}$$

where n is the sample size and x is the number of the individuals in the sample with the given characteristic.

Remark

$\hat{p}$ is an *unbiased estimator* of the population proportion p, meaning that $E(\hat{p} - p) = 0$.

Standard Deviation Scaling

Theorem (Linearity of Expected Value)

If X and Y are random variables with finite means and variances, and a and b are real constants, then

$$E(aX + bY) = aEX + bEY.$$

Corollary

Let X be a random variable with mean μ_X and standard deviation σ_X, and let $Y = cX$, where c is a real number. Then $\mu_Y = c\mu_X$ and $\sigma_Y = c\sigma_X$.

Proof.

$\mathrm{Var}(cX) = E((cX)^2) - (E(cX))^2 = c^2(E(X^2) - (EX)^2) = c^2 \mathrm{Var}(X).$ □

Distribution of the Sample Proportion

Theorem

For a simple random sample (with replacement) of size n drawn out of a population with proportion p,

1. *the mean of the sampling distribution of $\hat{p}$ is $\mu_{\hat{p}} = p$.*
2. *the standard deviation of the sampling distribution of $\hat{p}$ is*

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

Proof.

As a random variable, $\hat{p} = X/n$, where X is binomial with n trials, probability of success p, $\mu_X = np$, and $\sigma_X = \sqrt{np(1-p)}$. □

Sample Proportion Remarks

Remark

1. Sampling without replacement is acceptable when the population is large, the sample size is small, and the population proportion is not too close to 0 or 1, since in this case the population proportion remains nearly constant.
2. The shape of the sampling distribution is approximately normal when

$$np(1-p) \geq 10,$$

which means that we can approximate proportion probabilities by

$$P(\hat{p} \leq p_0) \approx P\left(Z \leq \frac{p_0 - \mu_{\hat{p}}}{\sigma_{\hat{p}}}\right)$$

Sample Proportion Probabilities

Example

Suppose that about 32% of math PhDs granted to US citizens belong to females, and the rest to males.[1] Obtain a random sample of US citizens with a doctorate in mathematics.

1. What is the probability that the proportion of females in a random sample of size 64 is above 40%?
2. What is the probability that the proportion of females in a random sample of size 100 is above 40%?
3. Suppose that a particular mathematics department has 50 faculty members who are US citizens with a doctorate in mathematics, and only 6 of them are females. Are there grounds to suspect that the hiring process is biased?

[1] This is about right for degrees granted in years 1999-2003, according to AMS.

9.1 Estimating A Population Proportion

Point Estimate

Definition

A **point estimate** is the value of a statistic that estimated the value of a population parameter.

Example

1. $\hat{p} = x/n$ estimates the population proportion p.
2. The sample mean $\bar{x}$ estimates the population mean μ.
3. The sample standard deviation s estimates the population standard deviation σ.

Confidence Interval

Definition

A **confidence interval** for an unknown population parameter is an interval of numbers based on a point estimate, and is used to indicate the reliability of an estimate.

Definition

The **level of confidence** is the proportion of confidence intervals that will contain the population parameter if a large number of different samples is obtained. 95% confidence level can be written as $\alpha = 0.05$, where the confidence level is $(1 - \alpha) \cdot 100\%$.

CI for Population Proportion

For large enough sample sizes n, the sample proportion $\hat{p}$ is approximately normally distributed with mean $\mu_{\hat{p}} = p$ and standard deviation

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

If we were to keep taking random samples of size n, we would expect $\hat{p}$ to obey the empirical rule. We could say things like "$\hat{p}$ is within one standard deviation from the population proportion about 68% of the time", or give it as a confidence interval

$$\left[\hat{p} - \sigma_{\hat{p}}, \hat{p} + \sigma_{\hat{p}}\right]$$

with $\alpha = 0.32$.

Constructing the Interval

Definition

Suppose that a simple random sample of size n is drawn from a population of size $N \geq 20n$, and that $n\hat{p}(1-\hat{p}) \geq 10$. Then the $(1-\alpha)\cdot 100\%$ confidence interval for the population proportion is given by

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where $z_{\alpha/2}$ is the z-score such that $P(Z > z_{\alpha/2}) = \alpha/2$. The radius of the confidence interval is called the **margin of error**.

Critical Values

Definition

A **critical value** of a distribution is the number which represents the number of standard deviations the sample statistic can be away from the parameter, and still result in a confidence interval which includes the parameter.

Table 3 : Critical Values for Confidence Intervals

Level of confidence	α	Area in each tail	Critical value $z_{\alpha/2}$
90%	0.1	0.05	1.645
95%	0.05	0.025	1.96
99%	0.01	0.005	2.575

To find $z_{\alpha/2}$, we can look up the probability $\alpha/2$ in the Z-table and reverse the sign of the z-score.

Interpreting the Confidence Interval

Remark

The *confidence* is in the method, not in the interval. Having a 90% confidence interval means having an interval which was obtained via a method that places the interval around the true population parameter 90% of the time, and misses 10% of the time. It is not OK to say that the probability of the interval containing the true mean is 90%, since the interval either contains the mean or it doesn't, so that probability is either zero or one for every confidence interval.

CI Example

Suppose that we took a random sample of 42 stars in the Milky Way galaxy and established that 17 of them have planets.

1. Find the 99% confidence interval for the proportion of stars with planets.
2. Find the 95% confidence interval for the proportion of stars with planets.
3. Find the 80% confidence interval for the proportion of stars with planets.
4. Find the minimal sample size sufficient for constructing the 80% confidence interval with margin of error 0.1.
5. Find the minimal sample size sufficient for constructing the 99.9% confidence interval with margin of error 0.01.

Minimal Sample Size

Theorem

The minimal sample size required to obtain a $(1-\alpha)\cdot 100\%$ *confidence interval for* p *with a margin of error* E *is given by*

$$n = \left\lceil \hat{p}(1-\hat{p}) \left(\frac{z_{\alpha/2}}{E} \right)^2 \right\rceil$$

or, if the estimate $\hat{p}$ *is unavailable,*

$$n = \left\lceil \frac{1}{4} \cdot \left(\frac{z_{\alpha/2}}{E} \right)^2 \right\rceil$$

where $\lceil x \rceil$ *is the ceiling function: the smallest integer at or above* x*.*

Level of Confidence and Margin of Error

Remark

Recall that the margin of error

$$E = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

If we ask for a higher level of confidence, then we have to settle for a wider interval (higher margin of error). If we increase the sample size, then we can improve the level of confidence, or reduce the margin of error, or both.

Relevant News

According to a November 4 2013 press release from NASA, one in five "Sun-like" stars is orbited by an "Earth-like" planet within the star's habitable zone, meaning that the surface temperature is just right for the liquid water to persist.
Given 10^9 or so stars in our galaxy, this estimate, if correct, means that Milky Way is likely home to billions of Goldilocks planets, where the temperature and the pressure are neither too high nor too low for bacterial life to exist on the surface.

9.2 Estimating a Population Mean

Obtaining a Point Estimate

Recall that the sample mean $\overline{X}$ is normal if the samples are drawn from a normal distribution, or approximately normal if the sample size is large enough. Either way, if we draw samples from a population with mean μ and standard deviation σ, then $\overline{X}$ is distributed with mean $\mu_{\overline{X}} = \mu$ and standard deviation $\sigma_{\overline{X}} = \sigma / \sqrt{n}$.

Naive Approach

A naive approach to obtaining a confidence interval would be to take the point estimate as the center, and the margin of error as the radius:

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

In practice, however, we rarely know the population standard deviation, especially when we do not know the population mean, so we may attempt to use the sample standard deviation instead:

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

This is exactly what William Gosset did, as he was analyzing small samples of barley for the Guinness Brewery in Dublin, Ireland. To his surprise, the intervals he obtained did not cover the mean with the predicted frequency.

Student's t-distribution

Gosset realized that using s as an estimate for the population standard deviation introduced additional uncertainty to the estimate of the population mean, and that it could be accounted for by replacing the z-score with a t-score obtained from a slightly different distribution. Out of respect for trade secrets, he published his results under a pseudonym "Student".

Theorem

If a simple random sample (with replacement) of size n is taken from a population that follows the normal distribution with mean μ and standard deviation σ, then the distribution of

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

is the Student's t-distribution with $n - 1$ degrees of freedom.

t-distribution Remarks

Remark

t-score delivers the correct estimate only if the population is normal, which real life populations are not. The CLT, however, assures us that it will work for non-normal populations as long as the sample size is large enough.

Remark

There are infinitely many t-distributions, with one corresponding to each sample size. The shape of a t-distribution is similar to that of the standard normal, but the tails are heavier, resulting in higher spread and wider confidence intervals. As the sample size n tends to infinity, the shape of the corresponding t-distribution tends to that of standard normal.

CI for Population Mean

If the sample data is

1. obtained from a simple random sample,
2. the sample size is small relative to the population ($n \leq 0.05N$), and
3. the population is normally distributed or the sample size is large,

then a $(1-\alpha)\cdot 100\%$ confidence interval for the population mean is given by

$$\overline{x} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}},$$

where $t_{\alpha/2,n-1}$ is the appropriate score from a t-distribution with $n-1$ degrees of freedom.

An Application of t-distribution

Example

Suppose that a random sample of size 4 is taken out of a normally distributed population, the sample mean $\overline{x} = 10$, and the sample standard deviation $s = 2$.

1. Find the 99% confidence interval for the population mean.
2. Do it again, assuming that the sample size $n = 40$.
3. Do it again for samples of size 4 and 40, but this time assuming that the population standard deviation $\sigma = 2$. This will produce over-confident intervals just like the ones that alerted Gosset.

Determining Sample Size

As with the sample size estimation for finding the population proportion, we can naively write that the margin of error

$$E = t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}$$

and attempt to solve for n. But we cannot! The value of t depends on n. We can, however, hope that using a z-score instead will not affect the estimate too much. So the minimal sample size required to estimate the population mean with the level of confidence $(1-\alpha)\cdot 100\%$ and with the margin of error E, is given by

$$n = \left\lceil \left(\frac{z_{\alpha/2}\cdot s}{E}\right)^2 \right\rceil$$

where $\lceil x \rceil$ rounds up to an integer.

Sample Size Example

Example

Suppose that a random sample of size 4 is taken out of a normally distributed population, the sample mean $\overline{x} = 10$, and the sample standard deviation $s = 2$. Find the minimal sample size required to produce

1. a 95% confidence interval for the population mean, with the margin of error $E = 1$.
2. a 99% confidence interval for the population mean, with the margin of error $E = 1$.
3. a 95% confidence interval for the population mean, with the margin of error $E = 0.1$.

9.3 Estimating a Population Standard Deviation

Chi-Squared Distribution

Theorem

If a simple random sample of size n is obtained from a normally distributed population with mean μ and standard deviation σ, then

$$\chi^2[n-1] = \frac{(n-1)s^2}{\sigma^2}$$

has a ***chi-squared distribution*** *with $n-1$ degrees of freedom.*

Properties of Chi-Squared Distribution

1. The pdf of $\chi^2[n-1]$ depends on the number of degrees of freedom.
2. The pdf is not symmetric.
3. The mean of $\chi^2[n-1]$ is $n-1$.
4. The pdf is zero for all negative numbers.

Theorem

Suppose that $\chi^2_{1-\alpha/2}$ *and* $\chi^2_{\alpha/2}$ *are the critical values of* χ^2*, which is the chi-squared distribution with* $n-1$ *degrees of freedom. That is,*

$$\begin{aligned} P\left(\chi^2_{1-\alpha/2} \leq \chi^2 \leq \chi^2_{\alpha/2}\right) &= 1-\alpha \\ P\left(\chi^2_{1-\alpha/2} \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi^2_{\alpha/2}\right) &= 1-\alpha \\ P\left(\frac{(n-1)s^2}{\chi^2_{\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}\right) &= 1-\alpha \end{aligned}$$

meaning that $(1-\alpha)\cdot 100\%$ *of all such intervals will contain the true population standard deviation.*

Confidence Interval for Population Variance

Theorem

If a simple random sample of size n is taken out of a normally distributed population with mean μ and standard deviation σ, then a $(1-\alpha)\cdot 100\%$ confidence interval about the population variance σ^2 is given by

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

Remark

This estimate is sensitive to a departure from normality, so it is extremely important to check that the population is sufficiently normal before computing it.

Application

Example

Suppose that the following sample was obtained from a normally distributed population: $\{2, 8, 14, 12\}$.

1. Find the sample mean.
2. Find the sample variance.
3. Find a 90% confidence interval for the population variance.
4. Find a 90% confidence interval for the population standard deviation.
5. Find a 80% confidence interval for the population standard deviation.

9.4 Which Procedure to Use?

When to Use Continuity Correction?

Example (1)

Suppose that a colony of ants has two castes: workers and drones, with one drone per nine workers. A random sample of 10 ants is drawn from the colony. What are the chances that there are three or more drones in the sample?

Example (2)

Suppose that a colony of ants has two castes: workers and drones, with one drone per 174 workers. A random sample of 1000 ants is drawn from the colony. What are the chances that there are three or more drones in the sample?

Remark

The continuity correction is a but a tool to approximate a binomial probability.

Which Distribution to Use?

Example (1)

Suppose that the mass of stars in a galaxy is approximately normally distributed with mean μ and standard deviation σ. What are the chances that a randomly chosen star has mass above M?

Example (2)

Suppose that 3% of all emergency calls are hoaxes. What are the chances that in a randomly drawn sample of 50 emergency phone calls, there is at most one hoax?

Example (3)

Suppose that 3% of all emergency calls are hoaxes. What are the chances that in a randomly drawn sample of 500 emergency phone calls, more than 5% are hoaxes?

Interval Estimates

Example (1)

Suppose that a random sample of 40 polar bears is drawn from the world-wide population, and 17 of them are male and 23 are female. Construct a 95% confidence interval for the proportion of male polar bears.

Example (2)

Suppose that a random sample of 40 polar bears is drawn from the world-wide population, and their weights are measured. The mean weight in the sample is 265 kg and the standard deviation of weights in the sample is 15 kg. Construct a 95% confidence interval for the mean weight of a polar bear.

Example (3)

Using the sample data from the previous example, construct a 95% confidence interval for the variance of the weight of a polar bear.

10.1 The Language of Hypothesis Testing

Statistical Inference

Remark

We will study but one approach to statistical inference: the so called **frequentist inference**. Another popular approach, which arises from a different interpretation of probability, is known as **Bayesian inference**.

Frequentist and Bayesian Inference

Remark

A frequentist statistician treats population parameters as constants, and attempts to estimate them and to draw conclusions about them by taking large samples. A Bayesian statistician has an option of treating population parameters as random variables, will typically sample individuals one at a time, and adjust the probabilities every time new information is obtained.

The result of a frequentist approach is either a "true or false" conclusion from a significance test or a conclusion in the form that a given sample-derived confidence interval covers the true value: either of these conclusions has a given probability of being correct. In contrast, the Bayesian approach can yield a distribution.

Null and Alternative Hypotheses

Definition

The **null hypothesis**, denoted H_0, is a statement to be tested. H_0 is assumed to be true until the evidence indicates otherwise.

The **alternative hypothesis**, denoted H_1, is the statement that we are trying to support by evidence.

Example

H_0: the average amount of soda in a 12 oz can is 12 oz.
H_1: the average amount of soda in a 12 oz can is less than 12 oz.

Example

H_0: the average weight of a US resident is 81 kg.
H_1: the average weight of a US resident is different from 81 kg.

Stating the Hypotheses

We will consider three ways to set up our hypotheses.

Definition

1. Equal versus not equal (**two-tailed test**)
   H_0: parameter $=$ value
   H_1: parameter $\neq$ value
2. Equal versus less than (**left-tailed test**)
   H_0: parameter $=$ value
   H_1: parameter $<$ value
3. Equal greater less than (**right-tailed test**)
   H_0: parameter $=$ value
   H_1: parameter $>$ value

The last two are known collectively as **one-tailed tests**.

Outcomes of Hypothesis Testing

Definition

The hypothesis testing procedure can be seen as a commitment to take certain actions depending on the outcome of the test. Specifically, a statistician will either reject or fail to reject H_0 based on the sample data. Each test has four possible outcomes:

1. Rejecting H_0 when H_1 is true (correct).
2. Failing to reject H_0 when H_0 is true (correct).
3. Rejecting H_0 when H_0 is true (**Type I error**).
4. Failing to reject H_0 when H_1 is true (**Type II error**).

Courtroom Analogy

Example

If the justice system is similar to that in US, then the defendant is assumed innocent until proven guilty. So we can let H_0 stand for the former, and H_1 for the latter. The purpose of the court is to make a correct decision: either to convict a criminal, or to let an innocent person go free. Life is not perfect, though, so sometimes an innocent person will be convicted (Type I error), and a criminal will be set free (Type II error).

Which Error is Worse?

Remark

Applications of hypothesis testing will typically attempt to minimize one or both types of error. Which error is "worse" should be determined within the context of a specific application, and may even be subjective.

Example

1. Prosecuting for an offense which demands capital punishment. (H_0: innocent, H_1: guilty)
2. Testing the efficacy of a symptom-releaving drug with a life-threatening side-effect. (H_0: ineffective, H_1: effective)
3. Testing the efficacy of a life-saving drug with annoying but survivable side-effects. (H_0: ineffective, H_1: effective)
4. Testing the economical impact of the World calendar, compared to that of the Gregorian calendar. (H_0: same as Gregorian, H_1: more efficient than Gregorian)

Type I and Type II Errors

Definition

$\alpha = P(\text{Type I error}) = P(\text{rejecting } H_0 \text{ when it is true})$

$\beta = P(\text{Type II error}) = P(\text{failing to reject } H_0 \text{ when it is false})$

Definition

The **level of significance** α is the probability of making a Type I error.

Stating the Conclusion

Example

When a frequentist statistician states the conclusion of a testing procedure, she must abide by her philosophical and mathematical commitments. For example, suppose that H_0 states that $\mu = 81$ kg, H_1 states that $\mu > 81$ kg, and the test is run with $\alpha = 0.05$.

1. If the evidence leads her to reject H_0, then she can say "There is sufficient evidence to conclude that the population mean is greater than 81, $\alpha = 0.05$".
2. If the evidence does not allow to reject H_0, then she can say "There is not sufficient evidence to conclude that the population mean is greater than 81, $\alpha = 0.05$".

Interpretation Remark

Remark

Unlike a Bayesian statistician, a frequentist can never "prove" the null hypothesis. No amount of evidence allows one to conclude that H_0 is true, even though a large enough sample may produce a very narrow confidence interval about the assumed value of a population parameter.

Testing Methods

There are three ways to conduct a hypothesis testing procedure, and the difference between them is purely algebraic. That is, all three methods are guaranteed to produce the same conclusion from a given sample data; only the way the conclusion is reached is different.

1. Confidence interval approach
2. Classical approach
3. *P*-value approach

Remark

While the methods are entirely interchangeable, the confidence interval method is more straightforward when applied to a two-tailed testing procedure.

10.2 Tests for Population Proportion

Classical Approach

1. Determine the null and the alternative hypotheses. The null always takes the form of $p = p_0$. Determine if the test is one-tailed or two-tailed.
2. Select the significance level α depending on the perceived seriousness of Type I error.
3. Compute the test statistic $z_0 = \dfrac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$
4. Based on α, determine the critical value(s), and compare them with the test statistic.
5. State the conclusion.

Finding Critical Value(s)

1. For a two-tailed test, the critical values are $-z_{\alpha/2}$ and $z_{\alpha/2}$.
2. For a left-tailed test, the critical value is $-z_{\alpha}$.
3. For a right-tailed test, the critical value is z_{α}.

Example

Find the critical value(s) for a two-sided test with $\alpha = 0.02$.

Example

Find the critical value(s) for a left-tailed test with $\alpha = 0.04$.

Test Statistic and Critical Value(s)

Suppose we found the test statistic z_0 and the critical value(s).

1. For a two-tailed test, reject H_0 if and only if $z_0 < -z_{\alpha/2}$ or $z_0 > z_{\alpha/2}$.
2. For a left-tailed test, reject H_0 if and only if $z_0 < -z_\alpha$.
3. For a right-tailed test, reject H_0 if and only if $z_0 > z_\alpha$.

In other words, we reject H_0 just in case when the test statistic is more "extreme" than the accepted critical value(s). Otherwise we fail to reject H_0.

Example

Example (Quitting Aid)

A pharmaceutical company claims that a new drug treatment for aiding people who are trying to quit smoking is 25% effective: that is, one in four people undergoing the treatment will successfully kick the nicotine addiction. In order to test this claim, a statistician draws a random sample of 80 smokers and has them go through the treatment. In the end, 14 out of 80 patients report that they successfully quit smoking. Can this be interpreted as the evidence that the effectiveness of the treatment is different from 25%? Run an appropriate test at 95% significance level.

Example

Example (Corked Wine)

Suppose that we are attempting to show that the proportion of corked wine bottles in a given batch is below 7%. A random sample of 160 wine bottles is drawn, and 8 of them are found to be corked. Run an appropriate test with $\alpha = 0.05$.

P-value Approach

Remark

The P-value approach is just a different way to arrive to the same conclusion. Instead of comparing the test statistic z_0 with the critical value(s), we compute the exact significance of z_0 and compare it with our acceptable significance level α. Only the step (4) of the testing procedure [211] is different.

Step 4: Analyze the P-value

Definition

Given the assumed value of the population parameter and a test statistic z_0, the P-**value** is the probability that a more "extreme" value is observed in a random sample of the same size.

1. For a two-tailed test, the P-value is $2 \cdot P(Z < -|z_0|)$.
2. For a left-tailed test, the P-value is $P(Z < z_0)$.
3. For a right-tailed test, the P-value is $P(Z > z_0)$.

Once the P-value is obtained, compare it with α. Reject H_0 if and only if the P-value is less than α.

10.3 Tests for Population Mean

Classical Approach

1. Determine the null and the alternative hypotheses. The null always takes the form of $\mu = \mu_0$. Determine if the test one-tailed or two-tailed.
2. Select the significance level α depending on the perceived seriousness of Type I error.
3. Compute the test statistic $t_0 = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$
4. Based on α, determine the critical value(s), and compare them with the test statistic.
5. State the conclusion.

Finding Critical Value(s)

1. For a two-tailed test, the critical values are $-t_{\alpha/2}$ and $t_{\alpha/2}$.
2. For a left-tailed test, the critical value is $-t_{\alpha}$.
3. For a right-tailed test, the critical value is t_{α}.

Example

Find the critical value(s) for a two-sided test with $\alpha = 0.02$ and sample size 17.

Example

Find the critical value(s) for a right-tailed test with $\alpha = 0.04$ and sample size 400.

Example

Example (Hold'em Stats)

Suppose that a Texas Hold'Em player's earnings in a sample of 41 sessions are approximately normally distributed with mean $\bar{x} = 21.4$ big blinds and standard deviation $s = 91.23$ big blinds. The player would like to know whether his mean earnings per session are significantly different from zero (that is, breaking even in the long run). Run an appropriate test with $\alpha = 0.01$.

Step 4: Analyze the P-value

Remark

Only the step (4) in the procedure (219) is different.

Definition

Given the assumed value of the population parameter and a test statistic t_0, the P-**value** is the probability that a more "extreme" value is observed in a random sample of the same size.

1. For a two-tailed test, the P-value is $2 \cdot P(t < -|t_0|)$.
2. For a left-tailed test, the P-value is $P(t < t_0)$.
3. For a right-tailed test, the P-value is $P(t > t_0)$.

Once the P-value is obtained, compare it with α. Reject H_0 if and only if the P-value is less than α.

Confidence Interval Approach

Remark

While the confidence interval approach could be used for one-tailed tests, we will avoid doing that. If the test we wish to run is one-tailed, then we will use either the classical or the P-value approach.

CI Testing Procedure

1. Determine the null and the alternative hypotheses. The null always takes the form of $\mu = \mu_0$. Determine if the test is one-tailed or two-tailed. If one-tailed, use a different approach.
2. Select the significance level α depending on the perceived seriousness of Type I error.
3. Based on the sample data, construct a $(1 - \alpha) \cdot 100\%$ confidence interval for the population mean $\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$.
4. If the confidence interval contains the value of μ assumed in H_0, then the evidence is lacking to reject H_0. Otherwise, we have sufficient evidence to reject H_0.
5. State the conclusion.

10.4 Tests for Population Standard Deviation

Classical Approach

1. Determine the null and the alternative hypotheses. The null always takes the form of $\sigma = \sigma_0$. Determine if the test one-tailed or two-tailed.
2. Select the significance level α depending on the perceived seriousness of Type I error.
3. Compute the test statistic $\chi_0^2 = \dfrac{(n-1)s^2}{\sigma_0^2}$
4. Based on α, determine the critical value(s), and compare them with the test statistic.
5. State the conclusion.

Finding Critical Value(s)

1. For a two-tailed test, the critical values are $\chi^2_{1-\alpha/2}$ and $\chi^2_{\alpha/2}$, and we can reject when the test statistic is outside of this interval.
2. For a left-tailed test, the critical value is $\chi^2_{1-\alpha}$, and we can reject when the test statistic is below it.
3. For a right-tailed test, the critical value is χ^2_{α}, and we can reject when the test statistic is above it.

Example

Find the critical value(s) for a two-sided test with $\alpha = 0.02$ and sample size 17.

Example

Find the critical value(s) for a left-tailed test with $\alpha = 0.05$ and sample size 101.

Example

Example (Hold'em Stats)

Suppose that a Texas Hold'Em player's earnings in a sample of 41 sessions are approximately normally distributed with mean $\bar{x} = 21.4$ big blinds and standard deviation $s = 91.23$ big blinds. The player would like to know whether the standard deviation of his mean earnings per session is significantly lower than 120 big blinds. Run an appropriate test with $\alpha = 0.05$.

P-value approach

Remark

As a rule, the statistical tables are not designed to look up the P-values. With the aid of a computer, though, the P-value approach provides the most informative result.

Remark

Only the step four in the procedure (226) is different. Instead of looking up the critical values we compute the probability of a sample producing a more extreme estimate than the one we observed.

1. For a left-tailed test, the P-value is $P(\chi^2 < \chi_0^2)$.
2. For a right-tailed test, the P-value is $P(\chi^2 > \chi_0^2)$.

Once the P-value is obtained, compare it with α. Reject H_0 if and only if the P-value is less than α.

P-value and Two-tailed Tests

Remark

There is no universal consensus on how the P-value should be defined for two-tailed tests, when the distribution of the test statistic is asymmetric, as is the case with the χ^2 distribution. Because of that, we do not recommend using the P-value approach for these kinds of tests.

10.5 Single Sample Testing Summary

Determining the Population Parameter

1. If the test concerns the proportion of the individuals in the population with a certain characteristic, then the test statistic has the standard normal distribution:

$$z_0 = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}, \ \hat{p} = \frac{x}{n}$$

2. If the test concerns the population mean, then the test statistic has the t distribution with $n-1$ degrees of freedom:

$$t_0 = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

3. If the test concerns the population variance, then the test statistic has the χ^2 distribution with $n-1$ degrees of freedom:

$$\chi_0^2 = \frac{(n-1)s^2}{\sigma^2}$$

Suggested Approach

Note that the classical approach is the only one that works more or less the same way in all three cases, and it can be carried out accurately without the aid of statistical software, using only the tables.

1. Population proportion: classical or *P*-value
2. Population mean: classical, *P*-value, or CI
3. Population standard deviation: classical (and *P*-value for one-tailed tests only)

11.1 Comparing Population Proportions

Independent Versus Dependent Sampling

One way to compare the proportions of individuals with a certain characteristic among two distinct populations is by taking one sample from each population and comparing the corresponding sample proportions. Depending on the context of the statistical study, different sampling techniques have to be used.

Definition

A sampling method is **independent** when the individuals in one sample a chosen independently from the individuals in the other sample. A sampling method is **dependent** when the individuals chosen for one sample determine or influence the choice of individuals for the other sample. In a special case when two samples consist of the same individuals, they are referred to as **matched-pairs** samples.

Examples

Which sampling method can we use? Which one should we use?

Example

Suppose we would like to know whether regular Cannabis smokers are more likely to be diagnosed with lung cancer during their lifetimes, when compared to people who do not consume Cannabis or its active ingredient in any form.

Example

Suppose we would like to know whether a certain kind of family counseling is effective in reducing the amount of reported spousal abuse.

Example

Suppose we would like to know whether flu-sick people of age 18 and older are more likely to report a symptom relief when administered a particular non-toxic cough-suppressant.

The Test Statistic

Having obtained two independent samples and the corresponding sample proportions $\hat{p}_1$ and $\hat{p}_2$, we may want to detect a significant difference between the population proportions, or whether one is greater than the other. In order to do that, we look at the distribution of $\hat{p}_1 - \hat{p}_2$. If both statistics are normally distributed, then so is the difference.

Lemma

If X_1 and X_2 are independent random variables, then the mean of $X_1 + X_2$ is the sum of the corresponding means, and the variance of $X_1 + X_2$ is the sum of the corresponding variances.

Distribution of the Test Statistic

Corollary

If $\hat{p}_1$ and $\hat{p}_2$ are normally distributed with mean p and standard deviation σ, and n_1 and n_2 are the corresponding sample sizes, then

$$z_0 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2}}}$$

has the standard normal distribution.

Pooled Estimate for Proportion

Definition

In practice, the population proportion p is usually unknown. Since we assume that corresponding population proportions are equal (H_0: $p_1 = p_2$), we can use both samples to estimate the common population proportion. So if two independent samples are drawn, with sample sizes n_1, n_2 and sample proportions x_1, x_2 respectively, then the **pooled estimate for population proportion** is

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

Classical Approach

1. Determine the null and the alternative hypotheses. The null always takes the form of $p_1 = p_2$. Determine if the test is two-tailed ($H_1 : p_1 \neq p_2$), left-tailed ($H_1 : p_1 < p_2$), or right-tailed ($H_1 : p_1 > p_2$).
2. Select the significance level α depending on the perceived seriousness of Type I error.
3. Compute the test statistic $z_0 = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$
4. Based on α, determine the critical value(s), and compare them with the test statistic. This step is identical to (212) and (213).
5. State the conclusion.

Example

Example (Astronaut Food)

Suppose that we want to show that plastic containers are better than metal containers at preserving the astronaut food for one year. Two independent samples of astronaut food are drawn, with 90 plastic containers and 70 metal containers. After one year in storage, 14 out of 90 plastic containers are found to contain spoiled food, and 20 out of 70 metal containers are found to be spoiled. Run an appropriate test with the significance level $\alpha = 0.05$.

Confidence Intervals for Proportion Difference

Definition

If two independent samples are drawn, with sample sizes n_1, n_2 and sample proportions x_1, x_2 respectively, then the **confidence interval for the difference between two population proportions** is given by

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}},$$

where $\hat{p}_i = x_i/n_i$ and α is the desired significance level.

Example

Construct a 95% confidence interval for the difference in proportions from the data in example (241).

11.2 Inference About Two Means: Dependent Samples

Matched-pairs Data

Example (Baby Weight Gain)

Suppose we want to show that an average baby gains weight between 1 and 2 years of age, or, more generally, to estimate how much weight is gained. In order to do so, we can take a random sample of babies and measure their weights twice: at 1 year, and then again at 2 years. The result may looks something like this:

1 year	6.1	7.5	8.6	8.2	7.2	8.6	7.7	9.1
2 years	11.2	12.6	13.7	12.8	10.3	14.7	10.3	12.5
difference	5.1	5.1	5.1	4.6	3.1	6.1	2.6	3.4

The rest of the testing procedure is identical to that for working with a single sample (219), only this time the sample is given by the last row in the table, with mean $\overline{d}$ and standard deviation s_d.

Example

Example

Use the data from the example (205) to run an appropriate test to detect a significant weight gain with $\alpha = 0.01$. Also, construct a 95% confidence interval for the difference between the population means.

11.3 Inference About Two Means: Independent Samples

Welch's t test

Remark

In general, comparing means among the populations with possibly unequal variances is a very difficult problem. We will consider an approximate solution known as **Welch's t test**. B. L. Welch have shown that the distribution of

$$t_0 = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

can be approximated by a t distribution. The degrees of freedom in Welch's test are hard to compute, so we will estimate them conservatively by taking the smaller of $(n_1 - 1)$ and $(n_2 - 1)$.

Classical Approach

1. Determine the null and the alternative hypotheses. The null always takes the form of $\mu_1 = \mu_2$. Determine if the test is two-tailed ($H_1 : \mu_1 \neq \mu_2$), left-tailed ($H_1 : \mu_1 < \mu_2$), or right-tailed ($H_1 : \mu_1 > \mu_2$).
2. Select the significance level α depending on the perceived seriousness of Type I error.
3. Compute the test statistic $t_0 = (\bar{x}_1 - \bar{x}_2)/\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

   Here $\bar{x}_1, \bar{x}_2$ are the sample means, s_1^2, s_2^2 are the corresponding sample variances, and n_1, n_2 are the corresponding sample sizes.
4. Based on α, determine the critical value(s), and compare them with the test statistic. This step is identical to (220), except that the number of degrees of freedom is $\min(n_1 - 1, n_2 - 1)$.
5. State the conclusion.

Example

Example (Manatees)

Suppose that we are trying to prove that female manatees weigh more than male manatees. A random sample of $n_1 = 14$ female manatees results in $\bar{x}_1 = 485$ kg and $s_1 = 21$ kg, while a random sample of $n_2 = 23$ male manatees results in $\bar{x}_2 = 470$ kg and $s_2 = 18$ kg. Run anappropriate test with $\alpha = 0.005$

Confidence Interval

Definition

Suppose that a random sample of size n_1 yields the sample mean $\bar{x}_1$ and sample standard deviation s_1. Suppose further that an independent sample of size n_2 is drawn from a different population, and it yields the sample mean $\bar{x}_2$ and sample standard deviation s_2. If the corresponding populations are approximately normally distributed and the sample sizes are sufficiently large, then the confidence interval for the difference of population means $\mu_1 - \mu_2$ is given by

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},$$

where $t_{\alpha/2}$ is the critical point of a t distribution with $\min(n_1 - 1, n_2 - 1)$ degrees of freedom.

Example

Example

Use the data in the example (249) to construct a 95% confidence interval for the difference between the population means.

11.5 Which Procedure to Use?

Statistical Tests

What parameter is being addressed in the problem?

1. Population proportions p_1 and p_2: run the procedure for comparing the proportions from two independent samples.
2. Population means μ_1 and μ_2.
   - 2.1 If the sampling is dependent, then run the one-sample procedure (211) on the sample of differences between matched data points.
   - 2.2 If the sampling is independent, then run the procedure for comparing the means from two independent samples.