Intro To Statistics

The Cardsharps by Caravaggio, c 1594

Trigger Warning

This course will draw upon examples which some people may find repulsive, shocking, or offensive. Areas of discussion may include, but not limited to:

1. Death
2. Violence
3. Illness
4. Abuse
5. Addiction
6. Politics
7. Religion
8. Sex/Gender
9. The Unspeakable Eldritch Horror Of Cthulhu

By: Melanie DeFazio @ stocksy.com

Please notify your instructor of any potential concerns.

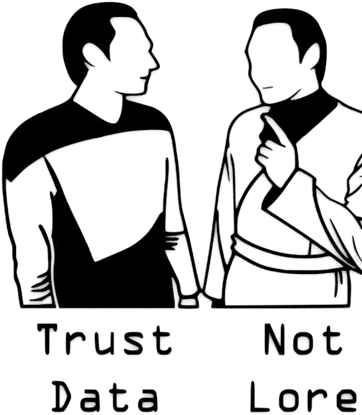
Anonymous Survey

For this activity, you will find yourself in a group of 2 or 3 of your peers. Introduce yourself and ask your partner the following questions:

1. What is your major?
2. How many classes are you taking this semester?
3. How long can you hold your breath?
   (OK to time with your cellphone)
4. How confident do you feel about succeeding in this Statistics class? Choose one of the following responses:
   - Piece of cake
   - A fighting chance
   - Spinning the wheel
   - Need a miracle
   - I am doomed

Enter the responses into a spreadsheet provided by the instructor.

Definitions And Key Terms



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.

Samples And Populations

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.

Definition

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

Examples Of Parameters/Statistics

A sample of college students' ages:

17, 19, 20, 20, 22, 23, 25, 34, 54

size: $n = 9$
mean: $\bar{x} = 26$
median: 22

minumum: 17
maximum: 54
range: 37

A population of beverage containers in my kitchen:

mug	glass	mug	mug	glass	jar
jar	jar	jar	glass	jar	bottle

size: $N = 12$
proportion of mugs: $p = 0.25$

number of categories: 4
maximum frequency: 5 (jars)

Statistical Variables

A **statistical variable** is an attribute or a property of an individual that we are interested in measuring. The value of the variable can "vary" from one entity to another.

A *variable* is distinct from either a *parameter* or a *statistic*. For example, we may look at a population of cars, and measure the exterior color in the hope of estimating the proportion of pink cars in the population.

Given an individual car, we can measure and record its color, so it's a variable in this example. An individual car cannot provide us with the proportion value, so the proportion is not a variable.

On the other hand, it would make no sense to ask: "What is the color of the sample?" But given the color data for a sample of cars, we can compute the proportion of pink cars in the sample, so the proportion is a statistic.

Using Terminology

For each study, describe individual, population, parameter/statistic, variable(s), and data:

Example

US Forest Service is interested in finding the average age of a tree growing in Eldorado National Forest. A ranger takes a sample of 60 trees and measures the age of each tree in years by counting the growth rings.

Example

A coffee shop owner wants to estimate the proportion of customers who purchase pastry with their order of coffee. A sample of customers is taken over the period of one week, and for each customer who buys coffee, the clerk writes down whether they also get pastry or not.

Variable Types

- **Qualitative (categorical)**
  classify individuals based on some attribute
  - **Nominal**
    a variable is used to name, label, or categorize; its values cannot be meaningfully ordered or ranked
  - **Ordinal**
    like nominal, but values can be ordered
- **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

Quantitative variables have additional levels of measurement:

- **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

Variable Type Examples

Determine the type and the level of measurement for the following variables:

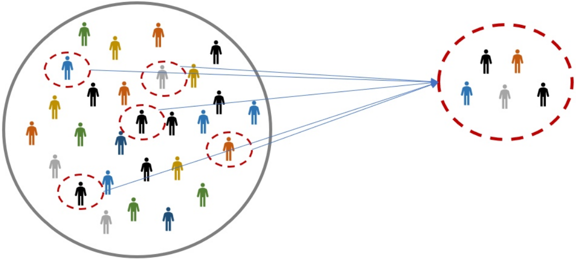
1. Last name
2. Mass of a car
3. Military rank
4. Bank account balance
5. Marital status
6. Temperature in °F
7. Star rating of a cab ride
8. The year an artwork was created
9. Speed limit on a CA road
10. Social Security Number

Data Input

One way to enter data into R is by defining a new variable:

```
x = c(1, 6, 7, 7, 7, 10)
x
[1] 1 6 7 7 7 10

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

Sampling



Random Sampling

Definition

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

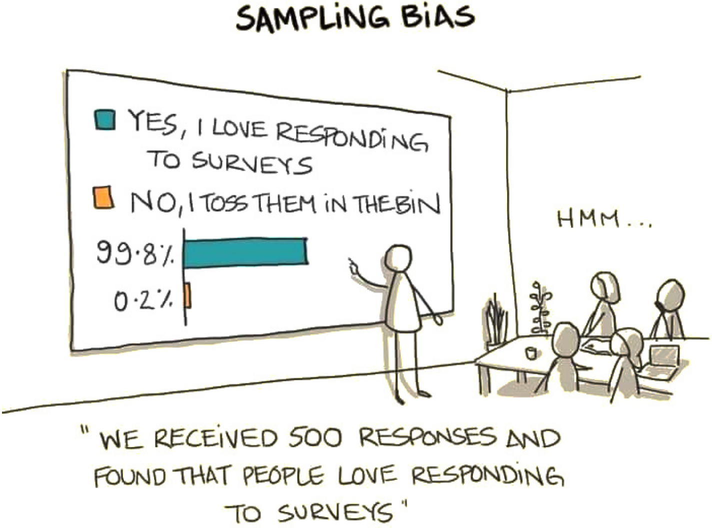
Definition

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.

Definition

Sampling bias is a property of the sampling method or its practical application which leads to some individuals being inherently more likely to be selected for the sampe than others. There are many different types of bias, and they all should be avoided as much as possible.

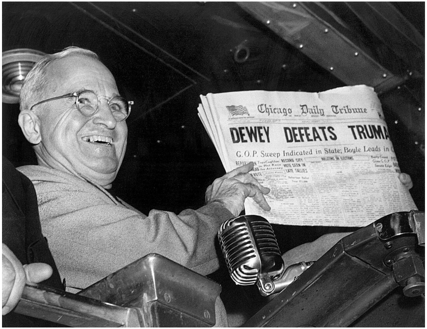
Everyone Loves Surveys


SAMPLING BIAS
🟦 YES, I LOVE RESPONDING TO SURVEYS
🟧 NO, I TOSS THEM IN THE BIN
99.8%
0.2%
HMM...
" WE RECEIVED 500 RESPONSES AND FOUND THAT PEOPLE LOVE RESPONDING TO SURVEYS "

sketchplanations.com

“Dewey Defeats Truman”

In 1948 *The Chicago Tribune* relied in part on a telephone survey to predict the outcome of the US presidential election. Because telephones were not widespread and owners tended to be wealthier, the sample largely excluded working-class voters who heavily supported Truman.



The Associated Press

The “Caveman Effect”

The sample of prehistoric paintings preserved through the millenia and now available to anthropologists seems to suggest that early humans primarily lived in caves, because that’s where most of them are found. But one has to consider that paintings are far more likely to remain preserved in caves, whereas most paintings exposed to the weather have perished a long time ago.

Lubang Jeriji Saléh cave, Indonesia,
over 40,000 years ago

Luc-Henri Fage

High-rise syndrome

Dirk Ingo Franke

In a study performed in 1987 it was reported that cats that survive a fall from less than six stories have greater injuries than cats who fall from higher than six stories. It has been proposed that this might happen because cats reach terminal velocity after righting themselves at about five stories, and after this point they are no longer accelerating, which causes them to relax, leading to less severe injuries than in cats who have fallen from less than six stories. The study was based on a sample of injured cats brought to a vet.

A more straighforward explantion may be the extreme "survivorship bias", whereas dead cats were never brought to a vet in the first place, and the surprising survival rate of cats falling from greater heights can be explained by favorable ground conditions and the small sample size.

Simple Random Sampling Procedure

Ideally, to draw a simple random sample of size n out of a population of size N, one can

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 between 1 and N (sampling *without replacement*),
3. administer the study to the individuals whose numbers were selected.

Sampling Example

Example

A researcher wants to estimate the proportion of incorrect Wikipedia citations for the article about Vladimr Putin: things like dead links, or links to Web pages that changed and no longer provide the correct reference material, or never have. The researcher wants a simple random sample without replacement of size $n = 12$.

```
sample(1:N, n)
```

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 adults currently living in US, then following frames can be used:

- List of Social Security numbers
- List of phone numbers
- List of names obtained from a census
- List of mailing addresses
- Several researchers walking busy streets in several large US cities, and talking to passers by.

Sampling Difficulties

In practice, listing all the individuals in the population may be impractical, and creating a good sampling frame may be very challenging.

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

Example

What sampling difficulties are anticipated during the following experiments?

1. Measuring the age of household cats in California
2. Measuring whether homeless people in Sacramento approve a specific ballot measure

Generating Random Numbers

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

Vietnam war draft lottery, 1969

Sampling Words With R

Sampling 2 different colors out of the 4 colors given:

```
sample(c("red", "green", "blue", "pink"), 2)

[1] "pink" "blue"
```

Sampling 6 outcomes, with possible repetitions,
out of Heads and Tails:

```
sample(c("Heads", "Tails"), 6, replace=T)

[1] "Tails" "Tails" "Heads" "Tails" "Heads" "Heads"
```

Sampling Numbers With R

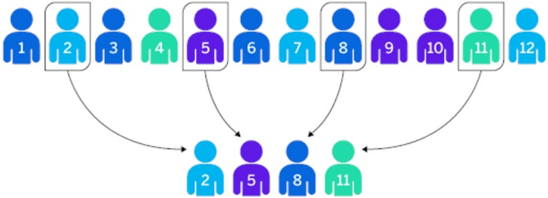
Sampling 5 distinct integers between 1 and 100:

```
sample(1:100, 5)

[1] 12 54 15 52 8
```

Sampling 3 random rational numbers between -1 and 1:

```
runif(3, min=-1, max=1)

[1] -0.1615992 -0.1630165 0.6994203
```

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.

Remark

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

Obtaining a Systematic Sample

To obtain a sample of size approximately n:

1. Estimate the population size N.
2. Let k, the period, be approximately N/n.
3. Let p, the index of the first individual to be chosen, be a random number between 1 and k.
4. Given a sampling frame, the sample will include individuals indexed by

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

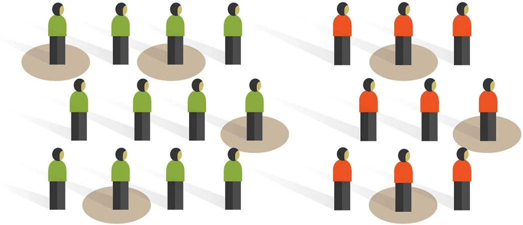
Example

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

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.



There may be two or more strata, and sampling out of them may be proportional or not.

Stratified Sampling Examples

Example

In the Census Bureau's 2022 findings, the percentage of Americans with a bachelor's degree or higher remained stable from the previous year at around 37.7%. An analyst wants to examine people's political preference, so she samples 377 Americans with a BS degree as well as 623 Americans without a BS degree.

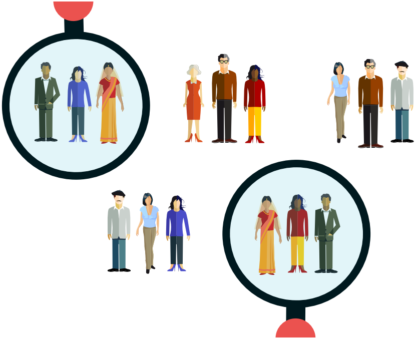
Example

While measuring the mass of an adult Black Widow spider, it makes sense to split the sampling frame by into males and females, since each group is homogeneous, but the respective sub-population parameters and the sampling procedures may be very 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.



Cluster Sample Of Galaxies

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. A more recent, and deeper image is JWST's first deep field.

Vietnam Draft

Example

On December 1, 1969, the Selective Service System of the United States conducted two lotteries to determine the order of call to military service in the Vietnam War for men born 1944-1950.

The days of the year (including February 29) were represented by the numbers 1 through 366 written on slips of paper. The slips were placed in separate plastic capsules that were mixed in a shoebox and then dumped into a deep glass jar. Capsules were drawn from the jar one at a time.

The first number drawn was 258 (September 14), so all registrants with that birthday were assigned lottery number 1. The second number drawn corresponded to April 24, and so forth. All men of draft age (born 1944 to 1950) who shared a birthdate would be called to serve at once. The first 195 birthdates drawn were later called to serve in the order they were drawn.

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.

Experimental Design and Ethics

37% of all stastics are made up on the spot

Variable Types in Relation to Studies

- **Explanatory variables** (also called **independent** or **predictor** variables) are controlled or manipulated by statisticians
- **Response variables** (also called **dependent** or **outcome** variables) are the measured outcomes of the studies

Example

A researcher wants to know whether ethnic background has an effect on academic performance. A sample of students is taken with white/European, Hispanic, African, and Native American ethnicities represented about equally, and each student has the GPA recorded.

Types of Studies

Definition

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

Definition

In a **designed** (or **randomized**) **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

Identify the type of study and categorize the measured variables.

- A survey of 500 male and 500 female Californians asks each participant whether they habitually eat while behind the wheel, and how many passengers they typically transport.
- A researcher places 100 phone calls to random phone numbers in the area and asks each respondent how much they use the cell phone while driving, and how many car crashes they've had in the last 3 years.
- As a part of an FDA study into the efficacy of a new anti-anxiety drug, 30 patients reporting anxiety symptoms are administered the drug, while in a separate sample of size 30 each patient is given an identially looking and tasting sugar pill without any drug. In each case, a participant reports whether they feel that medication was effective.

Pitfalls

- It may be unethical to measure a variable via a designed experiment.
- The sampling procedure may be **biased**, that is, systematically over/underestimating population parameters.

Example

- Measuring the median lethal dose of a recreational drug.
- Measuring the political leaning of Web comments.

Confounding

Definition

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.

Lurking Variable

Example

As ice cream sales increase, the rate of drowning deaths increases sharply. Therefore, ice cream consumption causes drowning.

This example fails to recognize the importance of time of year and temperature to ice cream sales. Ice cream is sold during the hot summer months at a much greater rate than during colder times, and it is during these hot summer months that people are more likely to engage in activities involving water, such as swimming. The increased drowning deaths are simply caused by more exposure to water-based activities, not ice cream. The stated conclusion is false.

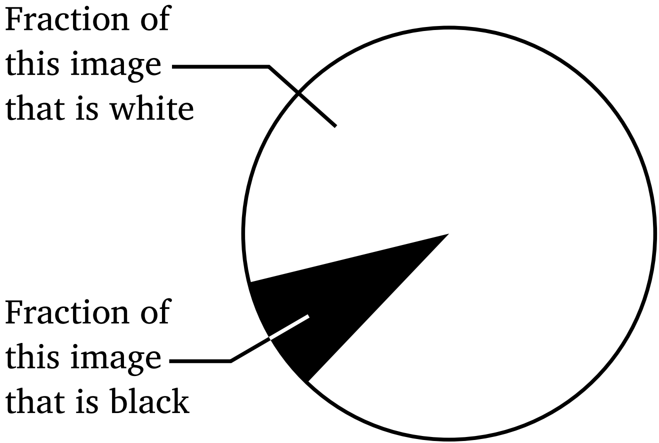
Correlation versus Causation

Example

Young children who sleep with the light on are much more likely to develop myopia in later life. Therefore, sleeping with the light on causes myopia.

This example resulted from a study at the University of Pennsylvania Medical Center. Published in the May 13, 1999 issue of *Nature*, the study received much coverage at the time in the popular press. However, a later study at Ohio State University did not find that infants sleeping with the light on caused the development of myopia. It did find a strong link between parental myopia and the development of child myopia, also noting that myopic parents were more likely to leave a light on in their children's bedroom. In this case, the cause of both conditions is parental myopia, and the above-stated conclusion is false.

Organizing Qualitative Data



xkcd.com

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.

```
colors = c("red", "green", "blue")
x = sample(colors, 50, replace=T)
table(x)
```

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 cannot be adjacent, 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.

Using R for Bar Graphs

Simulate and display a random sample of car body types:

```
cars = c("sedan", "coupe", "suv", "van")
x = sample(cars, 42, replace=T)
barplot(table(x))
```

Pareto Charts

Definition

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

Example

Drug harms in the UK: a multi-criteria decision analysis, by David Nutt, Leslie King and Lawrence Phillips, on behalf of the Independent Scientific Committee on Drugs. The Lancet. 2010.

Using R for Pareto Charts

If `x` is a qualitative data set, then 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.

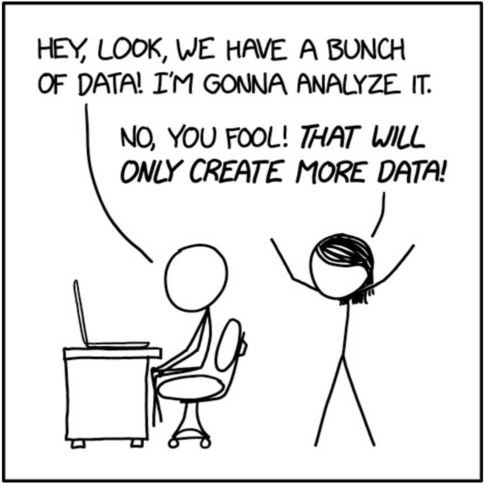
Pie charts are used for the same purpose as bar charts, but are considered inferior by many statisticians. They get overcrowded easily, and some studies have shown that comparing angles of sectors is harder than comparing lengths of bars.

Remark

In R, a pie chart of a qualitative data set x can be created with

```
pie(x)
```

Organizing Quantitative Data


HEY, LOOK, WE HAVE A BUNCH OF DATA! I'M GONNA ANALYZE IT.
NO, YOU FOOL! *THAT WILL ONLY CREATE MORE DATA!*

xkcd.com

Organizing Quantitative 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 left. In other words, a class A–B corresponds to the interval $(A, B]$. We call such classes left-open (or right-closed), and they are conventional.

Class Frequency Table

Example

We can use a built-in dataset to construct the class frequency table by hand, and then compare our work with how R does it.

```
?rivers
rivers
sort(rivers)
hist(rivers, nclass=4)
```

Histograms of Quantitative 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.

```
x = rnorm(100) # simulate a normal sample
hist(x, c="yellow")
```

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

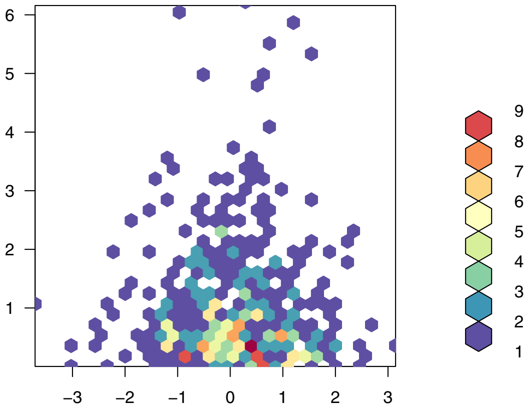
This approach works equally well for discrete and continuous data.

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.
- A **multimodal** distribution has more than one “peak”, indicating that there’s more than one “typical” measurement within the population.

Additional Displays of Quantitative 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 separate the leaves from the stem in order to get the total of 5-20 classes. If the stems are too few or too many, they may be **split** or **joined**.
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. In R, these displays are created with `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.

```
# simulate a random sample of integers:
d = rpois(100, 5)
stripchart(d, method="stack", pch=20, at=0)
```

Note that `stripchart` requires quantitative data.

Time-series Graphs

Definition

A **time-series graph** or **plot** consists of points connected by segments, where 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

Lynx and Hare populations

```
plot(x, y, type="l")
```

Cumulative Frequency Displays

Frequency Polygons

Definition

A **frequency polygon** is a line graph with a segment connecting each consecutive pair of points from the top sides of a frequency histogram. Depending on whether the frequency is cumulative or not, we can use midpoints or right endpoints of each histogram bar.

```
# simulate a random sample of reals:
x = rexp(100)
h = hist(x)

lines(h$mids, h$counts, type="o", pch=20) #overlay
plot(h$mids, h$counts, type="o", pch=20) #polygon
```

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.

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

Frequency Ogives

Definition

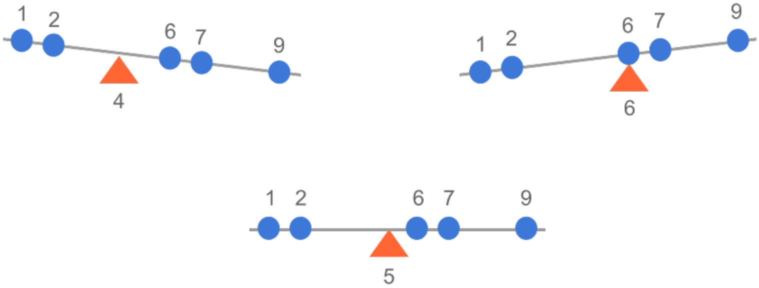
A **frequency ogive** is a frequency polygon for a histogram representing a cumulative (absolute or relative) frequency distribution. To construct the ogive, we use line segments to connect top right corners of cumulative frequency histogram bars.

```
x = rnorm(100, 50, 10)
h = hist(x, ylim=c(0,100))
lines(h$breaks[-1], cumsum(h$counts), type="o", pch=20)
```

Relative frequencies with R:

```
h = hist(rivers)
h$breaks         # breaks between classes
h$counts         # frequencies
cumsum(h$counts) # cumulative frequencies
```

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

$$\bar{x} = \frac{10 + 30 + 45 - 15 + 20}{5} = 18$$

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.

```
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: median $= 20$.

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

$$\text{median} = \frac{10 + 20}{2} = 15$$

```
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 discrete data set with only a few possible values is the most frequent observation that occurs in that data set. For a large qunatitative dataset without frequent observations, a mode is a measurement at the base of the tallest histogram bar.

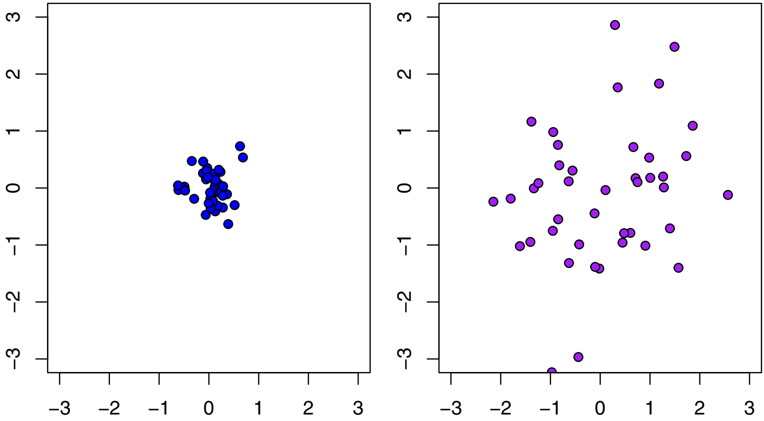
Remark

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. For histograms, one can sometimes identify several peaks, not just the tallest one.

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.

Remark

Measures of dispersion are also known as **measures of variation** and **measures of spread**.

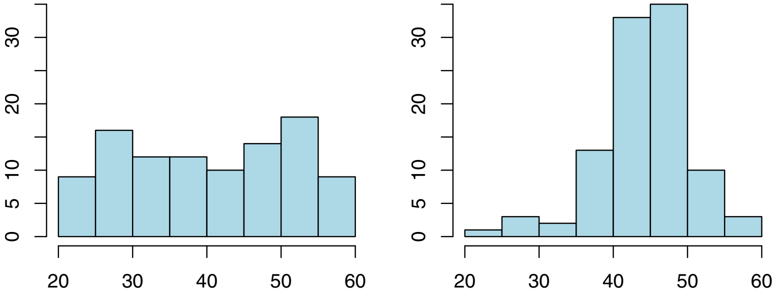
Range

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 $x_M - x_m$.

Range Examples



These two data sets have the same range, but something about the spread seems different.

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 $|\text{median} - x|$, then the **median absolute deviation** of a data set is the median of the set of all absolute deviations.

Example

Compare MAD for these data sets:

$$\begin{array}{ccccc} 0, & 48, & 50, & 55, & 100 \\ 0, & 10, & 50, & 93, & 100 \end{array}$$

```
mad(x, constant=1)
```

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, … , x_N are the observations, N is the population size, and μ is the population mean.

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, ... , 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*.

Computing Standard Deviation

Example

Consider a sample of size $n = 5$ with measurements

$$9,\ 13,\ 1,\ 2,\ 5$$

The sample mean is

$$\bar{x} = (9 + 13 + 1 + 2 + 5)/5 = 6$$

and so the sample standard deviation is

$$\begin{aligned} s &= \sqrt{\big((9-6)^2 + (13-6)^2 + (1-6)^2 + (2-6)^2 + (5-6)^2\big)/(5-1)} \\ &= \sqrt{\big(3^2 + 7^2 + (-5)^2 + (-4)^2 + (-1)^2\big)/4} \\ &= \sqrt{\big(9 + 49 + 25 + 16 + 1\big)/4} \\ &= \sqrt{100/4} = \sqrt{25} = 5 \end{aligned}$$

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.

population variance $\sigma^2 = \frac{1}{N}\sum_{k=1}^{N}(x_k - \mu)^2$

sample variance $s^2 = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \overline{x})^2$

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 for Normal Distributions

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% (approximately all) of data lie within 3 standard deviations of the mean.

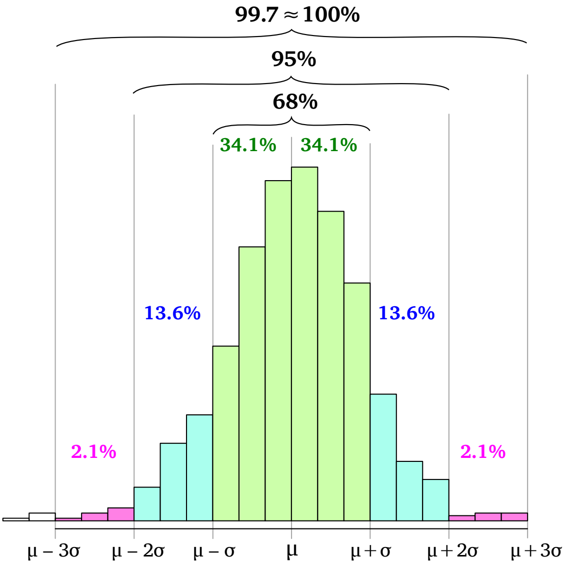
Remark

It may be useful to remember a different set of numbers: 2.5%, 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?

Empirical Rule Example



z-score

Definition

population z-score of a measurement x is $$z = \frac{x - \mu}{\sigma}$$

sample z-score of a measurement x is $$z = \frac{x - \bar{x}}{s}$$

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. z-score is unitless. If the distribution is rougly bell-shaped, then 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 larger z-scores (positive or negative) are less typical.

Chebyshev's Inequality

Theorem (Chebyshev's Inequality)

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

Chebyshev's Inequality Example

Example

Without making any assumption about the distribution shape, estimate the proportion of individuals in the population with measurements within 4 standard deviations away from the population mean.
$k = 4$, so $1 - 1/k^2 = 1 - 1/4^2 = 15/16 = 93.75\%$
Therefore **at least** 93.75% of the population has measurements within 4 σ of the mean.

Example

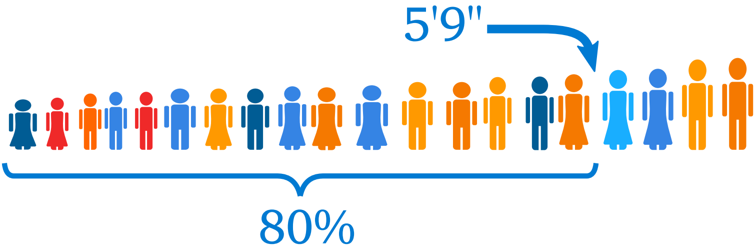
The mean student age is 20.5 years, and the standard deviation is 3.4 years. The shape of the distribution is unknown. Estimate the proportion of students aged between 15.4 and 25.6 years.

ER versus CI

Remark

The Empirical Rule may seem more powerful because it gives an approximation (not merely a lower bound) of 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 a correct bound in every possible situation.

Measures of Position



5'9"

80%

Measures of Position

Measures of position show the location of data values relative to the other data within the same data set.

Remark

Measures of position are also known as **measures of relative standing**.

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 25th percentile, the **second quartile** Q_2 is the 50th percentile, and the **third quartile** Q_3 is the 75th percentile.

Remark

There is no universal agreement on how to compute either percentiles or quartiles.

Computing Quartiles

1. For a data set of size one, $Q_1 = Q_3 = x$.
2. For a large data set, arrange the data in ascending order.
3. Determine Q_2, which is the same as the median.
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

$$\begin{array}{ccccccc} 10 & 12 & 13 & 13 & 17 & 18 & 29 \\ & Q_1 & & Q_2 & & Q_3 & \end{array}$$

Computing Percentiles

Consider a dataset of size n.
To find the percentile of a particular measurement x, sort the data and use the following formula:

$$\text{percentile of } x = \left\lceil \frac{\text{number of values} \leq x}{n} \cdot 100 \right\rceil$$

To find the data value corresponding to kth percentile, compute the index

$$i = \left\lceil \frac{kn}{100} \right\rceil$$

and then the ith percentile P_k is Lth number from the data sorted in the increasing order.

Computing Percentiles

Example

Find the 81st percentile from the sample data, as well as the percentile of the data value 20

$$3,\ 12,\ 20,\ 31,\ 34,\ 42,\ 44,\ 45$$

The 81st percentile has index $\lceil 81 \cdot 8/100 \rceil = \lceil 6.48 \rceil = 7$, so it is the 7th data value $P_{81} = 44$

The percentile of 20 is $\left\lceil \frac{3}{8} \cdot 100 \right\rceil\% = 38\%$

```
quantile(x, 0.81)
ecdf(x)(20)
```

Interquartile Range

Definition

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

$$\mathrm{IQR} = Q_3 - Q_1$$

Example

Compute the IQR for the sample data

$$3,\ 12,\ 20,\ 31,\ 34,\ 42,\ 44,\ 45$$

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

Example

$$\begin{array}{ccccccc} 10 & 12 & 13 & 13 & 17 & 18 & 29 \\ & Q_1 & & Q_2 & & Q_3 & \end{array}$$

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

so 29 is the only outlier.

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_1			Q_2			Q_3		max

```
summary(x)
```

Constructing Box Plot

1. Above a labeled coordinate axis, draw a box starting 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 $*$ or circles $\circ$ or stars $\star$.

Remark

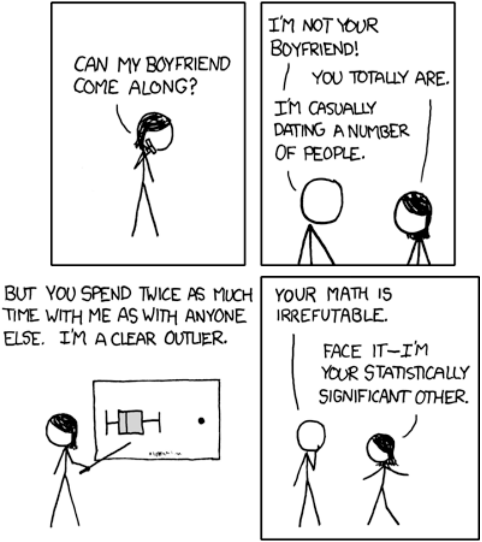
The horizontal lines are sometimes called **whiskers**, and the plot a **box-and-whisker plot**.

Box Plot Example

Box plots are particularly useful for comparing datasets which contain the same type of measurement applied to different poulations.

```
boxplot(Orange$circumference ~ Orange$Tree)
boxplot(Orange$circumference ~ Orange$age)
```

Significant Outlier


CAN MY BOYFRIEND COME ALONG?
I'M NOT YOUR BOYFRIEND!
YOU TOTALLY ARE.
I'M CASUALLY DATING A NUMBER OF PEOPLE.
BUT YOU SPEND TWICE AS MUCH TIME WITH ME AS WITH ANYONE ELSE. I'M A CLEAR OUTLIER.
YOUR MATH IS IRREFUTABLE.
FACE IT—I'M YOUR STATISTICALLY SIGNIFICANT OTHER.

xkcd.com

Scatter Diagrams And Correlation

Scatter Diagrams

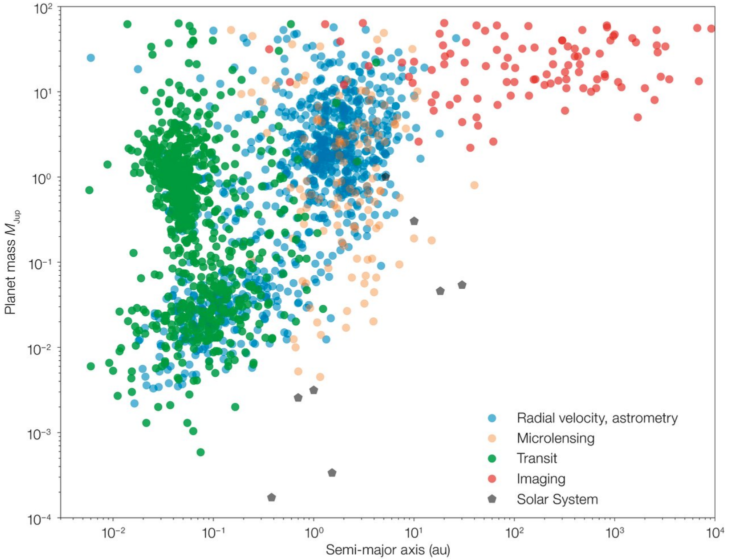
Definition

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.

Scatter Plot Example



Source: `https://www.eso.org/public/images/exoplanets_elt_large/`

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.

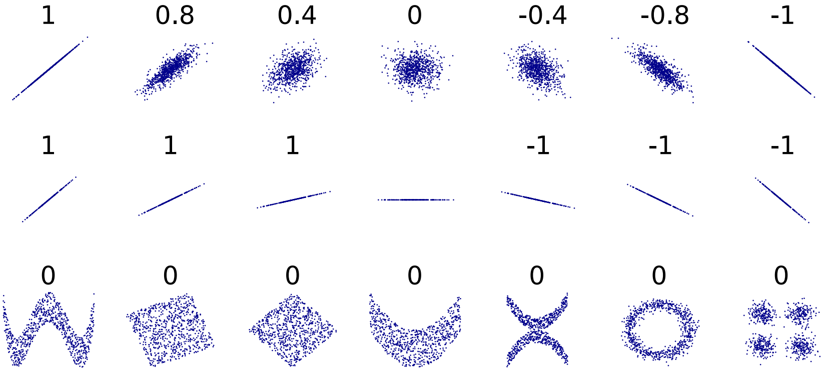
$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E((X-\mu_X)(Y-\mu_Y))}{\sigma_X \sigma_Y}$$

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

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 Examples



Simulating Linear Correlation

We will simulate a sample where the Y variable is a linear function of X plus some random noise $R \sim N(0, \sigma_R)$:

$$Y = mX + b + R$$

The code below simulates a sample of size 100 with X uniform on $(-10, 10)$, $Y = X + R$, and standard deviation of the noise $\sigma_R = 3$. When $\sigma_R = 0$, the linear relation is perfect and $r = 1$. Higher values of σ_R produce samples with r closer to zero.

```
n = 100
sd = 3
x = runif(n, -10, 10)
y = x + rnorm(n, 0, sd)
cor(x, y) # compute correlation coefficient
```

Linear Correlation Versus Causation

Recall that 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.

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

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

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

Using R For Fitting Lines

If we have matched data stored in vectors `x` and `y`, then we can find the correlation coefficient with `cor(x, y)` and the parameters of the line of best fit with `lm(y ~ x)` as the following example illustrates:

```
x = sort(rnorm(10)) # simulate two somewhat
y = sort(rnorm(10)) # related data sets
lm(y ~ x) # find the linear regression line

plot(x, y) # make a scatter plot
abline(lm(y ~ x)) # show the regression line
```

Correlation Example

A matched sample of temperatures X and precipitation levels Y is given in the following table:

X	60,	77,	72,	48,	51,	82,	41,	56,	75,	73
Y	0.1,	0.7,	0.7,	0.4,	0.5,	0.8,	0.4,	0.0,	0.2,	0.7

1. Find the correlation coefficient.
2. Find the equation for the line of best fit in the form $\hat{y} = b_1 x + b_0$

Probability Theory

Probability Experiment

Definition

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

Example

- Shooting with an arrow a target which is 1 meter in diameter and 20 meters away.
- Going to your very first job interview.
- Asking a random US citizen whether they identify as libreal or conservative.
- Conducting a presidential election in the USA.

Sample Space

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).

Example

$S = \{\text{Heads, Tails}\}$

S is the infinite set of points

Compound Events

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:

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

A **complement** of the event E, written as E', is the event which occurs precisely whenever E does not occur.

Probability Measure

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

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

Example

The empty set $\varnothing$ is always a null event.

Disjoint Events

Definition

Events E and F are **mutually exclusive** or **disjoint** if they have no outcomes in common:

$$E \cap F = \varnothing$$

Remark

The definition implies that the event *E and F* can never happen:

$$P(E \cap F) = 0$$

The converse is not true in general. We will consider plenty of continuous experiments where individual outcomes are null events. In situations like that it is possible to have

$$P(E \cap F) = 0 \quad \text{and} \quad E \cap F \neq \varnothing$$

Experiment Example

Roll a six-sided die, let it come to a stop, and record the number X on the top face.

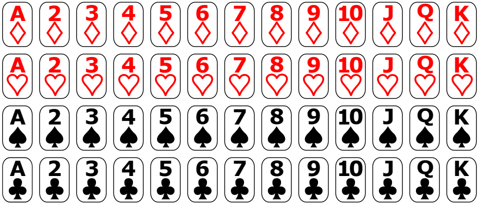
$$S = \{1, 2, 3, 4, 5, 6\}$$

Let E_1 be the event "$X > 4$" and E_2 be the event "X is odd". Describe the following events in English, and as sets:

1. E_1
2. E_2
3. ${E_1}'$
4. $E_1 \cup E_2$
5. $E_1 \cap E_2$

Another Experiment Example

Deal a random card out of the standard 52-card deck.


A♢ 2♢ 3♢ 4♢ 5♢ 6♢ 7♢ 8♢ 9♢ 10♢ J♢ Q♢ K♢
A♡ 2♡ 3♡ 4♡ 5♡ 6♡ 7♡ 8♡ 9♡ 10♡ J♡ Q♡ K♡
A♠ 2♠ 3♠ 4♠ 5♠ 6♠ 7♠ 8♠ 9♠ 10♠ J♠ Q♠ K♠
A♣ 2♣ 3♣ 4♣ 5♣ 6♣ 7♣ 8♣ 9♣ 10♣ J♣ Q♣ K♣

Describe the following events as sets:

1. The card is an Ace
2. The card is a heart suit.
3. The card is a Queen and is red.
4. The card is a diamond suit or a 7.
5. The card is not a number.

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, ...

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

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

$$\text{if} \quad E_1 \cap E_2 = \varnothing \quad \text{then} \quad P(E_1 \cup E_2) = P(E_1) + P(E_2)$$

Probability Model

Definition

A **probability model** is a mathematical object that describes all possible outcomes of an experiment, and allows to compute probabilities of all the possible events. A probability model must be consistent with the probability axioms.

A model can be a list, a table, a formula, or a more general procedure.

Discrete Model Example

Example

Suppose there is a raffle where each participant has a small chance to win a regular prize and an even smaller chance to win the grand prize. A corresponding sample space consists of three outcomes:

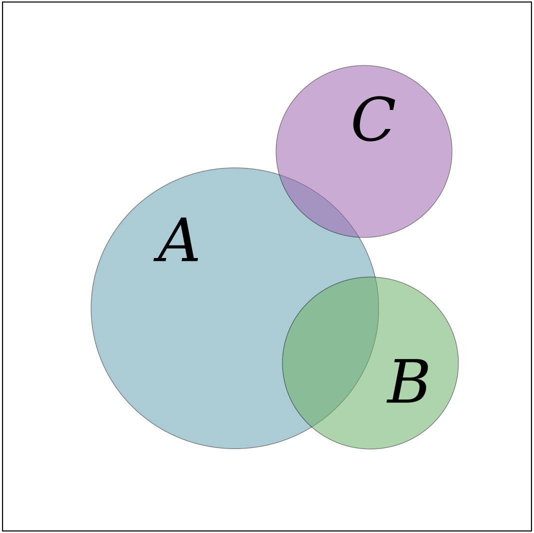
$$S = \{r, g, n\}$$

A model may assign probabilities as follows:

outcome	r	g	n
probability	0.10	0.02	0.88

Note that probabilities of the three disjoint outcomes have to add up to $P(S) = 1$ (axioms 2 and 3).

Geometrical Model Example



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 1751 new-born children is selected out of the population of children born in 2005. 901 of the children are male, and 850 are female. A corresponding empirical probability model for sex ratio at birth will have the sample space $S = \{M, F\}$, with

$$P(M) = \frac{901}{1751} \approx 0.5145 \quad \text{and} \quad P(F) = \frac{850}{1751} \approx 0.4854$$

Contingency Table

Definition

A **contingency table** (**cross tabulation**, **cross tab**, **two-way table**) is a type of table that displays the (multivariate) frequency distribution. 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

90 people were asked the following questions: what is your dominant hand (left, right, both), and are you married or single?

	R	L	B	Total
Married (M)	33	7	2	42
Single (N)	44	3	1	48
Total	77	10	3	90

More Sample Spaces

Example

Toss a fair coin 2 times and record the observed sequence of Heads and Tails.

$$S = \{HH, HT, TH, TT\}$$

Example

Toss a fair coin 3 times and record the observed sequence of Heads and Tails.

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

Here are some events in this space:

- N: no Tails observed
- A: Heads are observed exactly twice

Infinite Sample Spaces

Example

Toss a fair coin as many times as it takes for it to show Tails, and count the number of tosses.

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

Here are some events in this sample space:

- E_1: the first toss resulted in Tails.
- E_2: fewer than 5 tosses were needed.
- E_3: an odd number of tosses was needed.
- E_4: Tails never shows up.

Sample Space Practice

Describe sample spaces for the following experiments:

1. Toss two four-sided dice, one blue and the other one red.
2. Choose randomly two distinct herbs to throw in a stew out of the four available hebrs: basil, dill, mint, oregano.
3. Choose randomly the time for a meeting that has to start at either 2 or 3 pm on a weekday (Monday through Friday).
4. Select a random PIN that consists of 3 Arabic digits.

Classical Probability

Definition

In the classical interpretation of probability, the event space consists of k 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 k.

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 $k = 6$ 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 tend 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.

```
n = 10 # number of tosses
s = 2 # number of sides
barplot(table(floor(runif(n)*s)))
```

Basic Probability Rules

Andrey Kolmogorov, 1903 – 1987

The Complement Rule

Definition

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

Remark

Other notations for the complement of E are E^c and $\overline{E}$.

Theorem

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

Proof.

Since E and E' are mutually exclusive, axioms 2 and 3 imply

$$P(E) + P(E') = P(E \cup E') = P(S) = 1$$

□

The General Addition Rule

Theorem

For any two events E and F,

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

Proof.

Let E_1 be the event when E occurs and F doesn't, and let F_1 be the event when F occurs and E doesn't. Using the axiom 3 we can write

$$P(E) = P(E_1) + P(E \cap F)$$

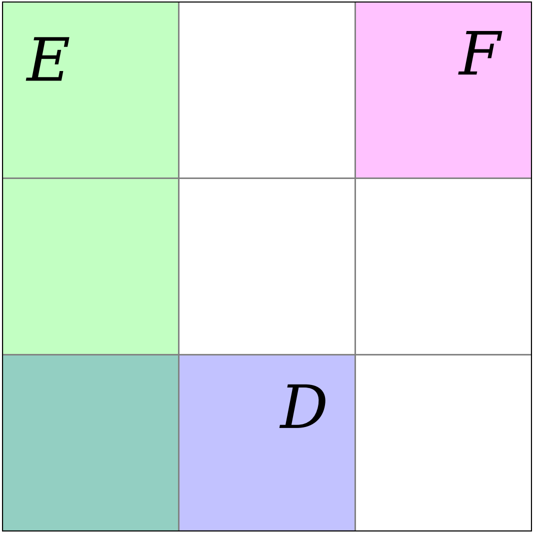
$$P(F) = P(F_1) + P(E \cap F)$$

and hence

$$P(E \cup F) = P(E_1) + P(F_1) + P(E \cap F) = P(E) + P(F) - P(E \cap F)$$

□

Addition Rule Example



Elaborate Example

Example

Our experiment is to draw one card at random out of the standard 52-card deck. How likely is each of the following draws?

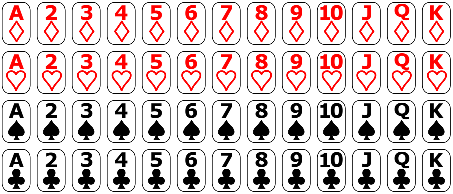
E_1 Any spade

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



A♢	2♢	3♢	4♢	5♢	6♢	7♢	8♢	9♢	10♢	J♢	Q♢	K♢
A♡	2♡	3♡	4♡	5♡	6♡	7♡	8♡	9♡	10♡	J♡	Q♡	K♡
A♠	2♠	3♠	4♠	5♠	6♠	7♠	8♠	9♠	10♠	J♠	Q♠	K♠
A♣	2♣	3♣	4♣	5♣	6♣	7♣	8♣	9♣	10♣	J♣	Q♣	K♣

Independent Events

Definition

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

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

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.

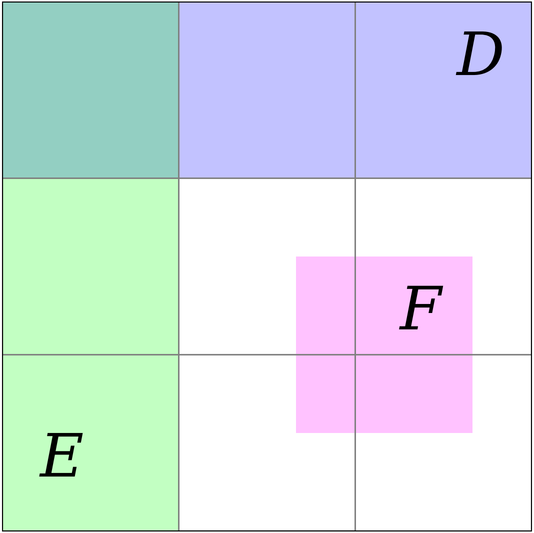
Example

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?

Independence Example



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?

Conditional Probability And 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

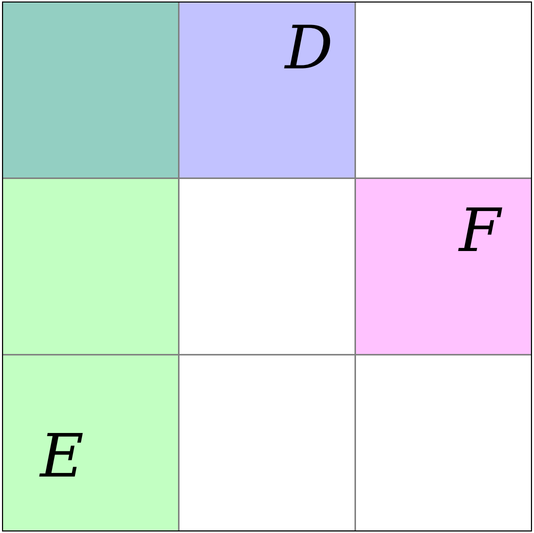
$P(F|E)$ is not defined if $P(E) = 0$.

Theorem

Possible (not null) events E and F are independent iff either

$$P(E) = P(E|F) \quad or \quad P(F) = P(F|E)$$

Conditional Example



Conditional Probability from Table

Example

1. Find $P(\text{Blond} \mid \text{Part-time})$
2. How likely is an employee to be part-time, given that they are a blond?
3. What are the chances that a randomly chosen redhead is employed full-time?
4. What are the chances that an employee is a brunette, given that they are not a part-time blond?

	Part-time	Full-time	Total
Blond	14	16	30
Brunette	12	6	18
Red	2	1	3
Total	28	23	51

Another Example

Example

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

	Windows	OS X	GNU/Linux	Total
Freshman	10	10	0	20
Sophomore	8	8	3	19
Junior	5	10	4	19
Senior	12	17	13	42
Total	35	45	20	100

General Multiplication Rule

Theorem

For any two non-null events E and F,

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

Example

A bag contains 13 apples: 3 of them are Macoun and 10 are Fuji.

1. Pull two apples out of the bag without looking. What is the probability that both are Macoun?
2. If you take 3 apples, what are the chances that all three are Fuji?
3. If you take 2 apples, what are the chances that they are different varieties?

Limitations Of GMR

Remark

The General Multiplication Rule only works if we can break up the compound event into stages in such a way that the chances of succeeding at each stage are unaffected by the way we succeeded at previous stages.

In the example with the apples, the chances of mis-matching the second apple depended on what variety of apple we pulled out first, so the GMR did not apply.

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 (Five Card Draw)

In the game of Five Card Draw each player gets five 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 hands are there?
2. How many will make a full house (three of a kind and a pair)?
3. How many will make four of a kind?
4. How many will make a royal flush? (Suited TJQKA.)
5. How many hands will have at least one Ace?
6. What are the probabilities of the events listed above?

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, **regardless of how the first part of the task was completed**, 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?

Counting with/without Replacement

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.

Definition

To obtain a random sample of size n **without replacement**, simply pick n distinct individuals out of the population. This is the same as Simple Random Sampling.

Remark

We are interested in calculating the probability of obtaining certain samples. All possible samples are equally likely, so we are dealing with classical models and we need a way to compute the total number of possible samples.

Factorial

Definition

For any non-negative integer k, the **factorial** of k, or “k factorial” is

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

while $0! = 1$.

Example

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

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

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 n 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*

$$ {}_N P_n = \frac{N!}{(N-n)!} = (N-n+1)\ldots(N-2)(N-1)N $$

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 10 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 n 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*

$$_N C_n = \frac{N!}{n!(N-n)!}$$

Combinations and Rule of Product

Example (Sack of Apples)

A burlap sack contains 17 apples: 3 of them are Macoun, 4 are Gala, and 10 are Fuji. Pull 9 apples out of the sack without looking. What is the probability that

1. all 9 are Fuji?
2. exactly 3 are Macoun?
3. there are exactly 3 of each variety?

Counting with Replacement

Example (Combination Lock)

A combination lock on a leather briefcase has 6 wheels with digits 0 through 9 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 8 characters long. If a password string is only allowed to have English letters (both cases) and Arabic digits, then how many different 8-character passwords are there?

Proposition

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

Counting Techniques Summary

	with replacement	without replacement
with order	N^n	${}_NP_n = \dfrac{N!}{(N-n)!}$
without order	$\dfrac{(N+n-1)!}{(N-1)!n!}$	${}_NC_n = \dfrac{N!}{n!(N-n)!}$

Table 1: The number of ways to sample n objects out of N objects.

More Counting Techniques †

- The number of ways to put N distinct things into b distinct baskets.
- The number of ways to put N distinct things into b 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 kind, N_2 indistinct objects of the second kind, and so on to N_k indistinct objects of the k-th type.

Permutations of Indistinct Items †

Proposition

The number of permutations (order matters) of N objects of k distinct kinds, such that N_1 of them are of the 1st kind, N_2 of them are of the 2nd kind, and so on, with $N_1 + N_2 + \ldots + N_k = N$, is

$$\frac{N!}{N_1!N_2!\ldots N_k!}$$

Example (Constrained Password)

How many ways are there to make a password that has to consist of 3 letters “a”, 4 letters “b”, and 5 letters “c”?

Example (Anagrams)

How many ways are there to rearrange the letters in the word “rearrange”?

Counting Examples

Example

There are 25 people at a business meeting, and each person shakes hands with everyone else. How many handshakes are there in total?

Example

Matt and Mary want to visit 17 different countries, and this summer they can afford to visit 4 of them. How many trips can they plan if the order in which they visit these 4 countries makes a difference?

Example

A serial number for a laptop consists of a country code (**U** or **G**), followed by 5 digits, followed by 2 capital English letters. How many different serial numbers are there?

More Counting Examples

Example

Bob is taking a quiz with 10 true/false questions. If he answers all questions randomly, what is the probability of him

1. getting all of them right?
2. getting exactly 5 right?
3. getting at least 2 of them wrong?

Example

Alice wants to read the 6 books of *The Lord of the Rings* in a random order. What are the chances she

1. reads them in the correct order?
2. reads *Book I* first and *Book VI* last?

Counting Examples ++

Example

4 couples want to sit in a row of 8 seats in a home movie theater. Assume that all seat assignments are completely random and equally likely.

1. How many sitting arrangements are there?
2. How many ways are there to sit them down so that everyone is sitting next to his/her partner? And what are the chances of that happening?
3. What are the chances that two specific people, Alice and Bob, end up in adjacent seats?

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.

Remark

In other words, a random variable quantifies the outcome of an experiment.

Discrete Versus Continuous

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.

Variable Examples

Example

The random experiment consists in looking at the clock and writing down the time of day.

1. Let the variable X be defined as the number of seconds elapsed since the last midnight.
2. Let the variable M be defined as 0 if the time is at or before noon, and 1 otherwise.

More Variable Examples

Example

The random experiment consists in taking 30 coins at random out of a jar containing 10000 dimes and 10000 quarters.

1. Let the variable C be defined as the number of cents taken out of the jar.
2. Let the variable D be defined as the number of dollars taken out of the jar.
3. Let the variable M be defined as the combined weight in milligrams of the coins taken out of the jar.
4. Let the variable T be defined as the time in seconds it took to get the coins out.

Even More Variable Examples

Example

The random experiment consists of Aisha tossing a fair coin up in the air and letting it fall and rest on the ground, as many times as needed until Tails are observed.

1. Let the variable C be defined as the number of tosses.
2. Let the variable H be defined as the maximum distance between the coin and the floor, measured in meters.
3. Let the variable R be defined as the reciprocal of C.

Discrete Random Variables

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 profit 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 profit 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 pmf.

```
n = 12 # number of coin tosses
p = 0.5 # probability of heads
barplot(dbinom(0:n, n, p), space=0, names.arg=0:n)
```

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

$$\mu_X = EX = x_1 P(X = x_1) + x_2 P(X = x_2) + \ldots = \sum_k x_k P(X = x_k)$$

Computing Mean

Example

Toss a fair six-sided die. Player wins $50 if the die shows six, and loses $10 if the die shows anything else.

x	$P(X = x)$
-10	$5/6$
50	$1/6$

$$\mu_X = EX = -10 \cdot \frac{5}{6} + 50 \cdot \frac{1}{6} = 0$$

Another Mean Example

Example

Toss a fair six-sided die. Player pays $50 for one dot, receives $100 for six dots, and breaks even otherwise.

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

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

Full Example

Example

In a game against the casino, a player gets three random cards from the standard 52-card deck. If the cards are all of same suit, then the player wins \$100. If the cards are all of the same kind, then the player wins \$500. And in all other situations the player loses \$1. Find the expected value for the winnings and decide whether the game is worth playing.

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 **variance** of X is

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

and the **standard deviation** of X is

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

Binomial Distribution

Jakob Bernoulli, 1655–1705

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 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”
- the probability of *success* is the same in each trial.

Binomial Examples

Example

Are these binomial experiments?

1. Toss a fair coin 100 times.
2. Toss a fair six-sided die 100 times.
3. Take a random sample of Bostonians (with replacement) and write down their genders: male or female.
4. Look out of the window every 5 seconds for 10 minutes and check whether it's raining.

Binomial pmf

When we use a random variable 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

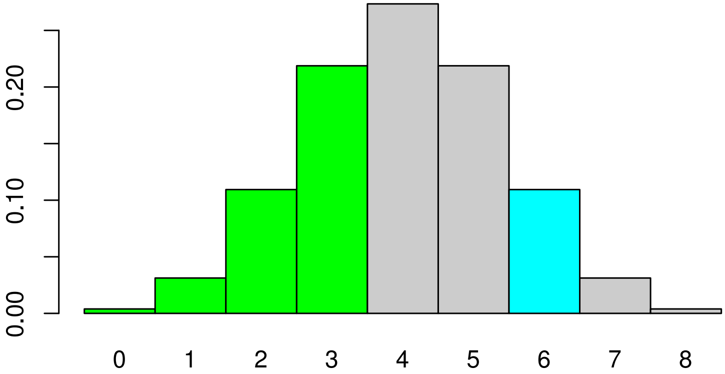
X is a **binomial random variable** if its pmf is

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

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

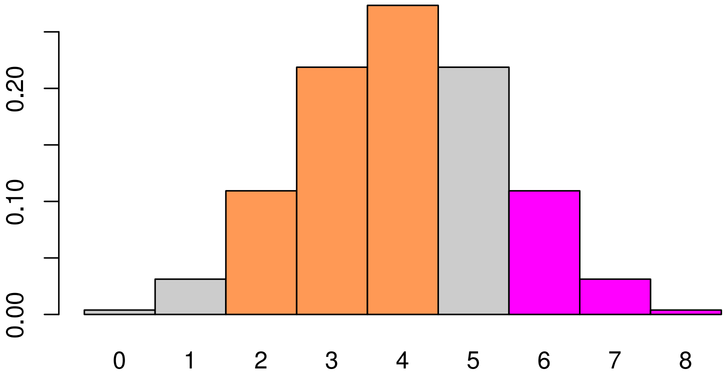
$$X \sim \text{Binom}(n, p)$$

Binomial pmf and cdf with R



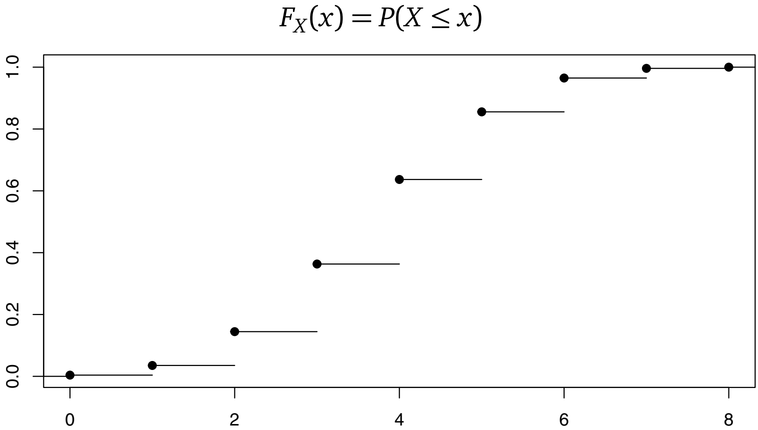
```
pbinom(3, 8, 0.5)
dbinom(6, 8, 0.5)
```

More on Binomial cdf with R



```
pbinom(4, 8, 0.5) - pbinom(1, 8, 0.5)
1 - pbinom(5, 8, 0.5)
```

Binomial cdf



```
n = 8, p = 0.5 # number of trials, probability of success
plot(0:n, pbinom(0:n, n, p), type="p", pch=19)
segments(0:(n-1), pbinom(0:(n-1), n, p), 1:n, pbinom(0:(n-1), n, p))
```

Computing Binomial Probabilities

Example (Disk Drives)

In order to back up sensitive data, a journalist purchased 6 identical hard drives, uploaded an encrypted copy of the data onto each of them, and stored them in safe deposit boxes for one year. Suppose that a drive survives one year in a box with probability 0.9, and that different drives do so independently. After one year, the drives will be taken out of the boxes and inspected. What is the probability that

1. all of the drives survive
2. exactly 3 drives survive
3. at most 2 drives survive
4. 4 or more drives survive

What is the 80th percentile of the number of working drives?

More Binomial Probabilities

Example (Soft Drinks)

Suppose we take a simple random sample of 700 soft drinks out of the population where 34% of all soft drinks are zero-sugar. What are the chances that

1. more than 200 drinks in the sample are zero-sugar?
2. between 210 and 240 drinks in the sample are zero-sugar?
3. fewer than 50 drinks in the sample are zero-sugar?

How many zero-sugar drinks should a sample have so that only 1% of all samples have more zero-sugar drinks than that?

Mean and Standard Deviation

Theorem

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

$$\mu_X = EX = np$$

and standard deviation

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

Example (Disk Drives)

For $X \sim \mathrm{Binom}(n = 6, p = 0.9)$, $EX = 6 \times 0.9 = 5.4$, and

$$\sigma_X = \sqrt{6 \cdot 0.9 \cdot 0.1} \approx 0.7348469$$

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.

```
# Binomial pmf for 1000 fair coin tosses
barplot(dbinom(440:560,1000,0.5),
  names.arg=440:560, space=0)
```

Continuous Distributions

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.

Definition of pdf †

Definition

$f_X(x)$ is a pdf associated with the random variable X if

- $f(x) \geq 0$
- $\displaystyle\int_{\mathbb{R}} f_X(x)dx = 1$
- $P(X \in [a,b]) = \displaystyle\int_a^b f_X(x)dx$

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

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

if

$$f_X(x) = \begin{cases} \dfrac{1}{b-a} & \text{if } a \leq x \leq b \\[2ex] 0 & \text{otherwise} \end{cases}$$

Uniform Properties

Theorem

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

$$\begin{aligned} EX &= \frac{a+b}{2} \\ \sigma_X &= \sqrt{\frac{(b-a)^2}{12}} \end{aligned}$$

Uniform Example

Example

Suppose that your Internet service provider schedules a technician visit to your house on a Thursday between 1 pm and 6 pm. If we don't know anything else about the expected time of arrival, then we can model the arrival time X of the technician by a uniform distribution.

1. Describe the distribution of X formally.
2. Find the mean and the standard deviation of X.

   What is the probability that the technician shows up:

3. after 5 pm?
4. between 1:30 pm and 3 pm?
5. at exactly 4 pm?

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

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

Mean, Variance, Median †

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:

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

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

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

Normal Distribution

Carl Gauss, 1777–1855

Motivation

Approximately normal distributions tend to 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).

Normal Distribution

Definition

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

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

if 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** if

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

Graph Of Normal pdf

Using `curve` function, it is easy to overplot a few members of the normal family with mean zero:

```
curve(dnorm(x, 0, 1), -8, 8, col="red", lwd=2, n=300,
      xlab="Normal pdf with sigma = 1, 2, 3, 4, 5",
      ylab="y")
curve(dnorm(x, 0, 2), -8, 8, col="orange",
      lwd=2, n=300, add=T)
curve(dnorm(x, 0, 3), -8, 8, col="green",
      lwd=2, n=300, add=T)
curve(dnorm(x, 0, 4), -8, 8, col="blue",
      lwd=2, n=300, add=T)
curve(dnorm(x, 0, 5), -8, 8, col="purple",
      lwd=2, n=300, add=T)
```

Or one can look at examples of Normal pdf at Wikimedia.

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 $\frac{1}{2}$
6. the tails get thin fast, but never touch the x-axis
7. X obeys (in fact, determines) the empirical rule
8. the cdf is insensitive to the inequality strictness:

$$P(X \leq x) = P(X < x)$$

Standard Normal cdf With R

If $Z \sim N(0,1)$ is standard normal, then the cdf of Z, which is the probability of $Z \leq z$, can be computed by

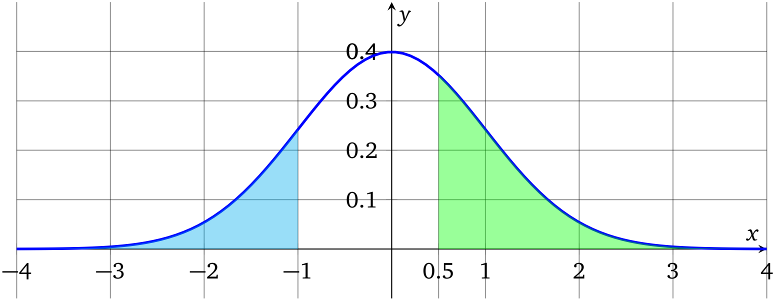
```
pnorm(z, 0, 1)
```

and the q-th quantile of Z can be computed with

```
qnorm(q, 0, 1)
```

Some examples of Normal cdf can be seen at Wikimedia.

Visualizing Normal pdf And cdf



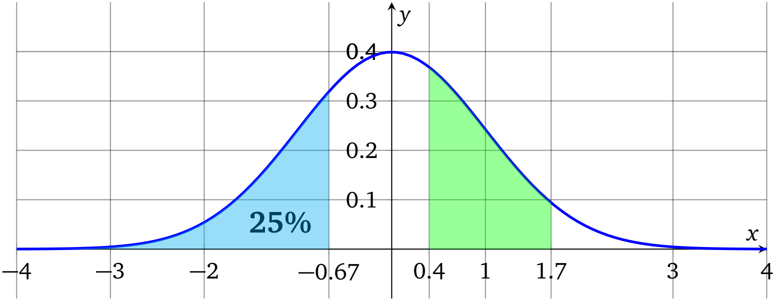
Left tail area $P(Z < -1)$

```
pnorm(-1, 0, 1)
```

Right tail area $P(Z > 0.5)$

```
1 - pnorm(0.5, 0, 1)
```

Visualizing Normal cdf And Quantiles



25th percentile of Z

Interval area $P(0.4 < Z < 1.7)$

```
qnorm(0.25, 0, 1)
```

```
pnorm(1.7, 0, 1)
 - pnorm(0.4, 0, 1)
```

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

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

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.

Normal cdf With R

Before the advent of computers, standardizing was an essential tool for computing normal probabilities. The standard normal cdf was computed once (very slowly) by hand and printed as a large table, and all normal questions had to be rephrased in terms of the corresponding z-scores. Not anymore: modern technology allows us to compute normal probabilities directly.

If X is normal with mean μ and standard deviation σ, then the probability of $X \leq x$ can be computed by

```
pnorm(x, μ, σ)
```

and the q-th quantile of X can be computed with

```
qnorm(q, μ, σ)
```

Application of Normal Probabilities

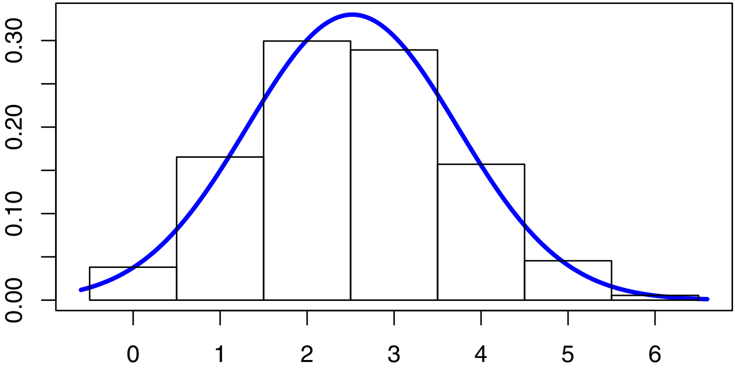
Example

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.

Normal Approximation



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 by de Moivre, 1738, though he's rarely credited for it, since he did not work out the notion of pdf. The proof of the effectiveness of this approximation 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?

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?

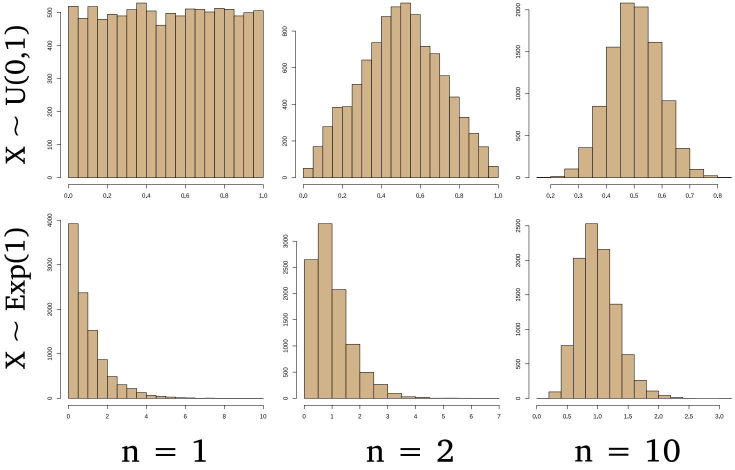
Example (2)

Suppose that a colony of ants has two castes: workers and drones, with one drone per 174 workers. A random sample of 10000 ants is drawn from the colony. What are the chances that there are 50 or more drones in the sample?

Remark

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

Distribution of the Sample Mean



X ~ U(0,1)

X ~ Exp(1)

n = 1

n = 2

n = 10

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 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 than 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 $X_1, \dots, X_n$ be random variables, each with the same mean μ, and a finite variance, not necessarily the same. We can think of $X_1, \dots, X_n$ 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

Given a large enough sample of size n out of a population with mean μ_X and standard deviation σ_X, the distribution of the sample mean $\overline{X}$ is approximately normal:

$$\overline{X} \sim N\left(\mu_X, \frac{\sigma_X}{\sqrt{n}}\right)$$

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; if X is uniform, then $n = 12$ will probably do it; but if X is suspected to be very asymmetric and heavy-tailed, then samples of size 100 or larger may be required to produce the approximate normality of $\overline{X}$.

CLT Application

Example

Suppose that the population of adult black widow spiders has the mean weight $\mu = 1.1$ g and the standard deviation $\sigma = 0.09$ g. Take a simple random sample of 50 spiders. What is the probability that the mean weight for the sample is

1. less than 1.12 g
2. greater than 1.09 g
3. between 1.09 and 1.11 g

Is there a way to estimate the probability of a single spider having weight below 1.12 g?

Another CLT Application

Example

Based on the historical precedent, the average amount of a donation for the office party is \$26, with standard deviation of \$43. Here we will assume that even for small sample sizes, the sample mean is approximately normal.

1. If 7 people donate, what are the chances that the party budget will exceed \$200?
2. What if 8 people donate?

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.

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

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 size is much larger than the sample size, and the population proportion is not too close to 0 or 1.
2. The shape of the sampling distribution is approximately normal when
$$np(1-p) \geq 10$$

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.

Which Distribution to Use?

Example (1)

Suppose that the mass of stars in a star cluster is approximately normally distributed with mean $\mu = 1.26$ solar masses and standard deviation $\sigma = 0.23$ solar masses. What are the chances that a randomly chosen star has mass above 2 solar masses?

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 are at most two hoaxes?

Which Distribution to Use?

Example (3)

Suppose that a sample of 42 cats is taken out of a population of feral cats in California, where the mean age in the population is 1.7 years with standard deviation of 1.3 years. How likely is the mean age in the sample to be between 1 and 2 years?

Example (4)

Suppose that 57% of all land-line calls are spam. What are the chances that in a randomly drawn sample of 1023 land-line phone calls, more than 60% are spam?

Estimating 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.
4. The sample median estimates the population median.
5. The sample maximum estimates the population maximum, even though it systematically falls short of the mark.

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 $(1-\alpha)\cdot 100\%$.

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 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 $\alpha = 0.32$.

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.

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

$z_{\alpha/2}$ using R:

```
qnorm(1 - α/2)
```

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**.

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.

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.*

Clopper-Pearson Interval

The Clopper-Pearson interval for the population proportion is constructed based on exact binomial probabilities, rather than normal approximations, and is commonly known as the “exact” method, since it uses a more appropriate model, applies correctly to small samples, and never produces intervals outside of $[0, 1]$. Unlike the normal approximation CI, the Clopper-Pearson interval is not symmetric about the point estimate, especially in cases when $\hat{p}$ is extreme.

```
binom.test(x, n, conf.level)
```

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 point estimate for the population proportion of stars with planets.
2. Find the 95% confidence interval for the proportion.
3. Find the 80% confidence interval for the proportion.
4. Find the minimal sample size sufficient for constructing the 90% confidence interval with margin of error 3%.
5. Find the minimal sample size sufficient for constructing the 99.9% confidence interval with margin of error 1%.

Relevant News

Current estimates of the number of exoplanets per star suggest at least one planet per Milky Way star on average, which is certainly an underestimate caused by our observation bias.

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 millions of Goldilocks planets, where the temperature and the pressure are neither too high nor too low for bacterial life to exist on the surface.

Estimating Population Mean

William Gosset, 1876–1937

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 normally 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.

t-distribution Shape

Compare the shapes of the *t* distributions with different number of degrees of freedom:

```
r = 6; w = 2; res = 300;
curve(dt(x, 1), -r, r, col="red", lwd=w, n=res,
 ylab="dt(x,df)", xlim=c(-r, r), ylim=c(0, 0.42),
 xlab="pdf of t with 1-6 degrees of freedom, and Z")
grid()
curve(dt(x, 2), -r, r, col="orange", lwd=w, n=res, add=T)
curve(dt(x, 3), -r, r, col="yellow2", lwd=w, n=res, add=T)
curve(dt(x, 4), -r, r, col="green", lwd=w, n=res, add=T)
curve(dt(x, 5), -r, r, col="blue", lwd=w, n=res, add=T)
curve(dt(x, 6), -r, r, col="purple", lwd=w, n=res, add=T)
curve(dnorm(x), -r, r, col="black", lwd=2, n=300, add=T)
```

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 ($n \leq 0.05N$),
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

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

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

```
t.test(x, conf.level)
```

Working with *t* Distribution in R

For a *t*-distribution with `df` degrees of freedom, the cdf $P(t \leq x)$ and the quantile function are implemented as

```
pt(x, df)
qt(x, df)
```

Given a sample of size n and the corresponding t distribution with $n-1$ degrees of freedom, the critical value $t_{\alpha/2,n-1}$ can be computed with

```
qt(1 - α/2, n-1)
```

An Application of *t*-distribution

Example

Suppose that a random sample of size 4 is taken out of an approximately normally distributed population of college class sizes:

$$32, \quad 18, \quad 15, \quad 24$$

1. Find the point estimate for the population mean.
2. Find the 99% confidence interval for the population mean.
3. Find the 99.9% confidence interval for the population mean.

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 approximate 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 $\lceil x \rceil$ is the ceiling function, which rounds x up to the nearest integer.

Sample Size Example

Example

Suppose that a random sample of size 4 is taken out of an approximately normally distributed population of college class sizes:

$$32, \quad 18, \quad 15, \quad 24$$

Find the minimal sample size required to produce

1. a 99% confidence interval for the population mean, with the margin of error $E = 0.5$.
2. a 95% confidence interval for the population mean, with the margin of error $E = 0.1$.

Which Procedure to Use?

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 standard deviation of the weight of a polar bear.

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 versus greater 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. Testing the efficacy of a symptom-relieving drug with a life-threatening side-effect. (H_0: ineffective, H_1: effective)
2. Testing the efficacy of a life-saving drug with annoying but survivable side-effects. (H_0: ineffective, H_1: effective)

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 kg, $\alpha = 0.05$”.
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 kg, $\alpha = 0.05$”.

Interpretation Remark

Remark

Unlike Bayesian statisticians, traditionalist frequentists can never "prove" the null hypothesis. No amount of evidence allows them 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, while making both α and β as small as desired.

But this philosophical interpretation has come under fire during the recent years, and the change was precipitated, ironically, by *conclusive* statistical inquiries into the publication bias of statistical results.

Frequentist Interpretation Bias

The traditional way of preparing a frequentist result for publication is to describe it as *conclusive* or *statistically significant* when H_0 is rejected, and *inconclusive* or *statistically insignificant* when H_0 is not rejected. Recent surveys of publications have shown that *conclusive* studies are 3 times more likely to be published. This phenomenon is now known as the *publication bias*, and is regarded as a serious problem with the traditional frequentist interpretation.

Interpretation Counterexample

Example

There are two drugs on the shelf, X and Y. Both drugs are treating the same condition: acne. Both drugs went through similar clinical efficacy trials, and were found about equal in effect. However, the drug X, unlike the drug Y, also went through another comprehensive study, which controlled for dozens of variables, and yielded a confidence interval for an increase in blood pressure. Specifically, this latter study produced a 99.9% confidence interval covering the zero (meaning no change in blood pressure), and ended with a frequentist conclusion "the data does not provide sufficient evidence to conclude that X increases blood pressure", which in turn precluded publication.

Interpretation Counterexample Part 2

Example

But now imagine yourself having both acne and high blood pressure problems. You are standing in a drug store, choosing between X and Y. Of course if one drug doesn't work, you will try another. But are you going to choose X or Y first? And if you say X, like any normal person who understands statistics would, you just tacitly admitted that the frequentist interpretation is utter nonsense. You cannot keep a straight face and keep calling a result *statistically insignificant* or *inconclusive* if it essentially tied your hands with respect to the choice of medication.

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.

Test For Population Proportion

Tech-Assisted Approach

1. State the null and the alternative hypotheses. The null always takes the form of $p = p_0$. The alternative should be:
   - $p \neq p_0$ for a two-tailed test,
   - $p < p_0$ for a left-tailed test, or
   - $p > p_0$ for a right-tailed test
2. Compute the p-value.
3. State the conclusion.

Note that we will perform two-tailed tests for proportion even when we are aiming to prove one-tailed statement, and this has to be accounted for in your conclusion.

p-value

p-value of a test is the probability of observing a sample of the same size which is more extreme than the one we have observed, given that H_0 is true. We reject or fail to reject H_0 by comparing the p-value with our acceptable significance level α. We reject H_0 if and only if the p-value of the test is less than α.

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 5% significance level.

More Examples

Example (Feral Cats)

A random sample of 130 feral cats has 106 tabbies. With $\alpha = 0.01$, is there enough evidence to conclude that more than 70% of all feral cats are tabbies?

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 19 of them are found to be corked. Run an appropriate test with $\alpha = 0.05$.

Example (Donuts)

A donut shop management believes that about 30% of the donuts sold are with sprinkles. They take a random sample of purchases and observe that out of 197 donuts they sold, 42 are with sprinkles. Is there enough evidence to conclude that the proportion of donuts with sprinkles is significantly different from 30%? Use $\alpha = 0.1$

Test For Population Mean

Classical Approach

1. Determine the null and the alternative hypotheses. The null always takes the form of $\mu = \mu_0$. The alternative should be:
    - $\mu \neq \mu_0$ for a two-tailed test,
    - $\mu < \mu_0$ for a left-tailed test, or
    - $\mu > \mu_0$ for a right-tailed test
2. Determine the distribution of the test statistic: $t_0 \sim t_{n-1}$.
3. Based on α, determine the critical values and the rejection region.
4. Compute the test statistic $t_0 = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$
5. Compute the p-value of the test.
6. State the conclusion.

Finding Critical Value(s)

For the two-tailed t-test, the critical values are $t_{1-\alpha/2}$ and $t_{\alpha/2}$

Example

Find the critical values for a two-tailed t-test with $\alpha = 0.02$ and sample size 17.

Example

Find the critical values for a two-tailed t-test with $\alpha = 0.04$ and sample size 400.

Analyze the *p*-value

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.

For a two-tailed test, the p-value is $2 \cdot P(t > |t_0|)$

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

Example

Example (Hold'em Stats)

Suppose that a Texas Hold'Em player records his profit/loss in 6 different tournaments, in US dollars:

$$-80, \ -160, \ 3200, \ -40, \ 560, \ -200$$

The player would like to know whether his mean earnings per tournament are significantly higher than zero (that is, whether the player makes money on the long run).

1. What is the point estimate for average earnings per tournament?
2. Run a t-test for detecting whether $\mu > 0$ with $\alpha = 0.1$.
3. What could be said for or against applying a t-test to this situation?

Example

Example (Job Search)

Suppose that a random sample of 234 Americans is surveyed to find out how much time they spent on searching for a job to get their current employent. The mean time spent on looking for a job in this sample is 26 days, and the standard deviation is 34 days.

Test the claim that the mean time Americans spend looking for a job is less than 30 days, using significance level $\alpha = 0.05$.

Extreme Event Discussion

Example (Bar Drinks)

Suppose that a bar owner takes a random sample of receipts and records the number of drinks purchased by each patron during their visit:

$$3, \quad 5, \quad 2, \quad 3, \quad 4, \quad 3, \quad 0, \quad 6$$

1. Run the appropriate test at 5% significance level to test the claim that patrons order more than 2 drinks on average.
2. Compare this with the right-tailed test, same significance level.

How Things Get Done

Randall Munroe's succinct explanation of how to make sure that your hypothesis testing results get published: https://xkcd.com/882/

CHI-SQUARED DISTRIBUTION

Chi-Squared Distribution

Definition

If $Z_1, \ldots, Z_k$ are independent standard normal random variables, then the sum of their squares,

$$Q = \sum_{i=1}^{k} Z_i^2$$

is distributed according to the *chi-squared* distribution with k degrees of freedom. This is usually denoted as

$$Q \sim \chi_k^2$$

The chi-squared distribution has one parameter: a positive integer k that specifies the number of degrees of freedom (the number of random variables being summed).

Chi-Squared Application

Theorem

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

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

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

Chi-Squared pdf in R

Using `curve` function, it is easy to overplot a few members of the χ^2 family. Note that the shape of the curve is quite unique for χ_1^2 and χ_2^2. The first one, in particular, is asymptotic to the vertical axis at $x = 0$.

```
curve(dchisq(x, 1), 0, 8, col="red", lwd=2, n=300,
 xlab="pdf of chi-squared with 1-5 degrees of freedom")
curve(dchisq(x, 2), 0, 8, col="orange", lwd=2, n=300, add=T)
curve(dchisq(x, 3), 0, 8, col="green", lwd=2, n=300, add=T)
curve(dchisq(x, 4), 0, 8, col="blue", lwd=2, n=300, add=T)
curve(dchisq(x, 5), 0, 8, col="purple", lwd=2, n=300, add=T)
grid()
```

Properties Of Chi-Squared Distribution

1. The pdf of χ^2_{n-1} depends on the number of degrees of freedom.
2. The pdf is not symmetric.
3. The mean of χ^2_{n-1} is $n-1$.
4. The pdf is zero for all negative numbers.

Working With Chi-Squared Distribution

For a χ^2 distribution with `df` degrees of freedom, the cdf and the quantile function are implemented as

```
pchisq(x, df)
qchisq(x, df)
```

When working out a confidence interval or a two-tailed testing procedure, given a χ^2 distribution with $n-1$ degrees of freedom, the critical values $\chi^2_{\alpha/2,n-1}$ and $\chi^2_{1-\alpha/2,n-1}$ can be computed with

```
qchisq(1−α/2, n−1)
qchisq(α/2, n−1)
```

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 variance.

CI for Population Standard Deviation

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 $(1-\alpha)\cdot 100\%$ confidence interval about the population standard deviation σ is given by

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

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 an approximately normally distributed population: $\{2, 8, 14, 12\}$.

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

Test For Population Standard Deviation

Classical Approach

1. State the null and the alternative hypotheses. The null always takes the form of $\sigma = \sigma_0$. The alternative should be:
   - $\sigma \neq \sigma_0$ for a two-tailed test,
   - $\sigma < \sigma_0$ for a left-tailed test, or
   - $\sigma > \sigma_0$ for a right-tailed test
2. Determine the distribution of the test statistic: $\chi_0^2 \sim \chi_{n-1}^2$
3. Based on α and the type of the test, determine the critical value(s) and the rejection region.
4. Compute the test statistic $\chi_0^2 = \dfrac{(n-1)s^2}{\sigma_0^2}$
5. Compare critical value(s) with the test statistic and state the conclusion.

Example

Example (Hold'em Stats)

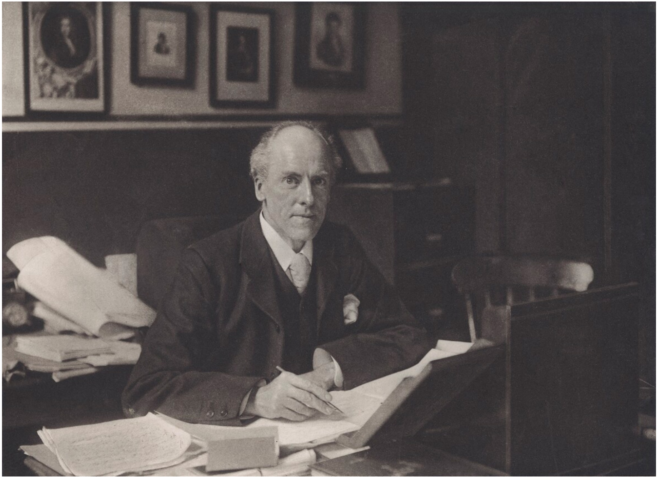
Suppose that a Texas Hold'Em player's earnings in a sample of 41 sessions are approximately normally distributed with mean $\bar{x} = 21.4$ big blinds and standard deviation $s = 91.23$ big blinds. The player would like to know whether the standard deviation of his earnings per session is significantly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

p-value approach

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. We will follow the procedure which is both traditional and popular for continuous unimodal distributions, and double the smaller one-tailed p-value.

Test For Independence



Karl Pearson, 1910

Test For Independence

1. State the null and the alternative hypotheses. The null is that the variables are independent. The alternative is that they are dependent. This is a true right-tailed test, with both the rejection region and the *p*-value being strictly right-tailed.
2. Determine the distribution of the test statistic: $\chi_0^2 \sim \chi_{\text{df}}^2$ where $\text{df} = (r-1)(c-1)$
3. Based on α and the type of the test, determine the critical value(s) and the rejection region.
4. Compute the test statistic $\chi_0^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$
5. Compare critical value(s) with the test statistic and state the conclusion.

Independence Test Example

Example

Consider the sample data whereas people were asked to state their preference for paper books versus digital readers, as well as to name their favorite literary genre.

	Sci-Fi	**Romance**	**Classics**	Totals
Digital	50	125	90	265
Paper	75	175	30	280
Totals	125	300	120	545

Are these preferences independent? Run an appropriate test with significance level $\alpha = 0.05$.

Independence Test With R

Given a table of data like

	A	A'
B	17	28
B'	42	13

To input the data:

```
m = matrix(c(17, 42, 28, 13), ncol=2)
```

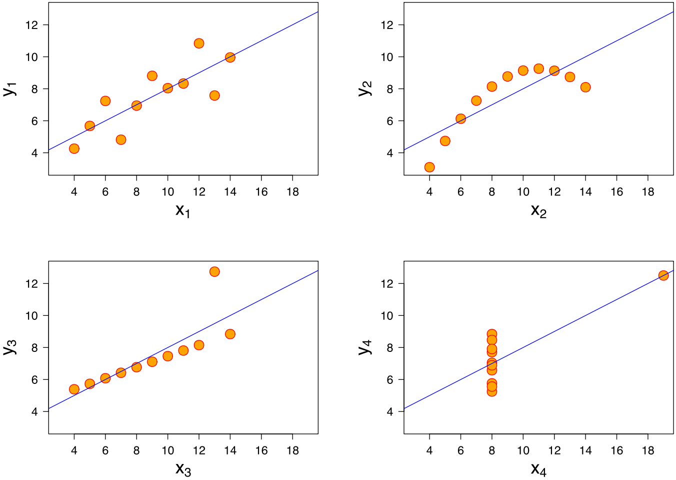
To run a test with continuity correction:

```
chisq.test(m)
```

To run a test without the correction, which corresponds to the manual procedure:

```
chisq.test(m, correct=F)
```

Testing Correlation



Anscombe’s quartet

t-test for Correlation

For pairs from an uncorrelated bivariate normal distribution, the sampling distribution of a function of Pearson's correlation coefficient follows Student's t-distribution with $n-2$ degrees of freedom. Specifically, if the underlying variables have a bivariate normal distribution, the variable

$$t_0 = r\frac{\sqrt{n-2}}{\sqrt{1-r^2}}$$

has a t-distribution in the null case (zero correlation), and so a significant departure from t can be interpreted as a strong correlation. This holds approximately in case of non-normal observed values if sample sizes are large enough.

Remark

The traditional table of critical values for this type of test lists the values of r, not the test statistic.

Classical Approach

1. State the null and the alternative hypotheses. The null is always $\rho = 0$. The alternative should be:
   - $\rho \neq 0$ for a two-tailed test,
   - $\rho < 0$ for a left-tailed test (negative correlation),
   - $\rho > 0$ for a right-tailed test (positive correlation).
2. Determine the distribution of the test statistic: $t_0 \sim t_{n-2}$
3. Based on α, determine the critical value(s) and the rejection region.
4. Compute the test statistic $t_0 = r\dfrac{\sqrt{n-2}}{\sqrt{1-r^2}}$
5. Compare critical value(s) with the test statistic and state the conclusion.

Correlation Test Example

Seven students are sampled randomly from an algebra class, and have their scores for the first two tests recorded:

T_1	59,	39,	51,	15,	60,	67,	54
T_2	63,	62,	49,	41,	56,	71,	68

1. Are the test scores linearly correlated? Use $\alpha = 0.1$
2. If a student gets 70 points on first test, what score can she expect on the second test?

Comparing Populations

Poverty and Wealth by William Powell Frith, 1888

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.

Comparing Proportions

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 statistics 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 $p_1 = p_2$. The alternative should be:
   - $p_1 \neq p_2$ for a two-tailed test,
   - $p_1 < p_2$ for a left-tailed test, or
   - $p_1 > p_2$ for a right-tailed test.
2. Determine the distribution of the test statistic: $z_0 \sim Z$.
3. Based on α and the type of the test, determine the critical value(s) and the rejection region.
4. Compute the test statistic $z_0 = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$
5. Compare critical value(s) with the test statistic and state the conclusion.

Comparing Proportions With R

R implements a different version of a proportion comparison test based on χ^2 distribution:

```
prop.test(c(x1, x2), c(n1, n2), conf.level=)
```

where the first argument is a vector of binomial success counts, the second argument is a vector of sample size, and the third (optional) argument is the confidence level for an interval estimate of $p_1 - p_2$.

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 statistics 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 (320).

Comparing Means: Paired Samples

Matched-pairs Data

For matched-pairs data, the testing procedure is identical to that for working with a single sample (283) of differences.

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.

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

Run an appropriate test to detect whether babies gained more than 2 kg in weight on average, with $\alpha = 0.01$. Also, construct a 90% confidence interval for the difference between the population means.

Comparing 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

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

can be approximated by a t distribution with this many degrees of freedom:

$$\frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{s_2^2}{n_2}\right)^2}$$

Classical Approach

1. Determine the null and the alternative hypotheses. The null always takes the form of $\mu_1 - \mu_2 = \Delta_0$, where Δ_0 is the assumed difference between population means. The alternative should be:
   - $\mu_1 - \mu_2 \neq \Delta_0$ for a two-tailed test,
   - $\mu_1 - \mu_2 < \Delta_0$ for a left-tailed test, or
   - $\mu_1 - \mu_2 > \Delta_0$ for a right-tailed test
2. Determine the distribution of the test statistic.
3. Based on α and the type of the test, determine the critical value(s) and the rejection region.
4. Compute the test statistic.
5. Compare critical value(s) with the test statistic and state the conclusion.

Example

Example (Manatees)

We are trying to detect a difference, if any, between the average weights of female and male manatees. The data for two independent samples of manatees follows.

Male	450,	485,	428,	474,	479,	494,	460,	506,	481,	474,	471
Female	472,	455,	523,	475,	467,	493,	484,	503,	450		

Run an appropriate test with $\alpha = 0.005$.

Example

Example (Air Temperatures)

The data from two independent samples of air temperatures is presented below. Use 0.1 significance level to test the claim that Dubai is 10 or more degrees hotter on average than Sacramento.

Dubai	101,	105,	115,	97,	86,	100,	94
Sacramento	65,	78,	59,	87,	90,	71	

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

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

Testing Review

Hypothesis Testing Review

Example

A company studied two programs for compensating its sales staff, with 9 people participating in the study. In program A, salespeople were paid a higher salary, plus a small commission for each item they sold. In program B they were paid a lower salary with a larger commission. The amounts sold, in thousands of dollars, are summarized below.

Subject	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
A	55	22	34	22	31	61	55	30	68
B	53	24	37	28	25	61	58	38	72

Can we conclude that the mean sales are higher for program B? Use $\alpha = 0.05$.

More Examples

Example

In a simple random sample of 95 families, 70 had one or more pets at home. Can we conclude that the proportion of families with pets is greater than 60% with $\alpha = 0.02$?

Example

A 2012 survey reported that in a sample of 72 women aged 18–25, the mean number of hours of television watched per day was 2.88, with standard deviation of 2.43. Can we conclude that women watch less than 3 hours of TV per day on average? Use $\alpha = 0.01$.

Even More Examples

Example

Two suppliers of machine parts delivered large shipments. A simple random sample of 150 parts was chosen from each shipment. For supplier A, 12 out of 150 parts were defective. For supplier B, 28 out of 150 parts were defective. Is the proportion of defective parts greater for the supplier B? Use $\alpha = 0.05$.

Example

Speeds for a sample of 9 cars were measured by radar along a stretch of highway. The results, in mph, are summarized below.

$$56 \quad 60 \quad 53 \quad 55 \quad 54 \quad 51 \quad 54 \quad 51 \quad 56$$

Can we conclude that the population standard deviation is greater than 2? Use $\alpha = 0.05$.

And More

Example (Volunteers)

In a volunteer group, adults 21 and older volunteer from one to nine hours each week to spend time with a disabled senior citizen. The program recruits among community college students, four-year college students, and nonstudents. The table below is a sample of the adult volunteers and the number of hours they volunteer per week.

	1–3 hrs	4–6 hrs	7–9 hrs
Community College Students	111	96	48
Four-Year College Students	96	133	61
Nonstudents	91	150	53

Is the number of hours independent from the type of volunteer? Use $\alpha = 0.02$.

...And More...

Example (Black Cherry Trees)

Take a look at the `trees` data set, which describes girth, height, and volume of 31 felled black cherry trees. Which pair of the variables is most strongly correlated? How significant are these correlations? Run three tests, each with $\alpha = 0.01$.

The `trees` data set is a *data frame*: it is a table consisting of several columns with matched measurements. One can refer to a specific column (for example, Girth) like so:

```
trees$Girth
```

ANOVA

Ronald Fisher with his sons, 1955
Stained glass window (now removed),
Caius College, Cambridge

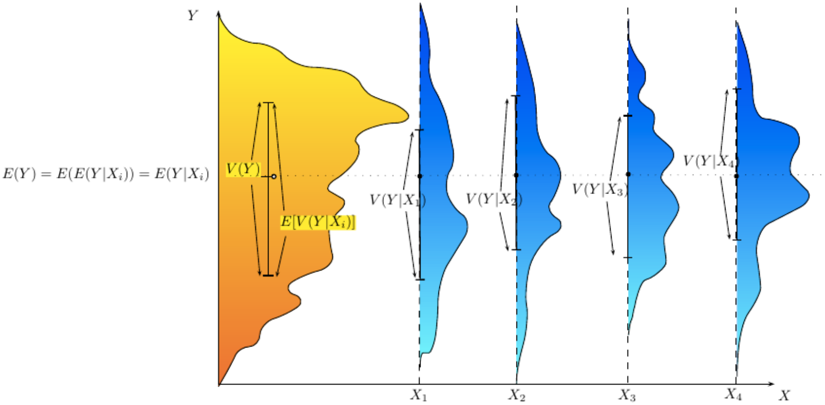
ANOVA

The **an**alysis **o**f **va**riance can be used to describe otherwise complex relations among variables.

For example, we may want to predict the weight of a dog participating in a dog show based on some other characteristics. The question becomes, which other characteristics should we use?

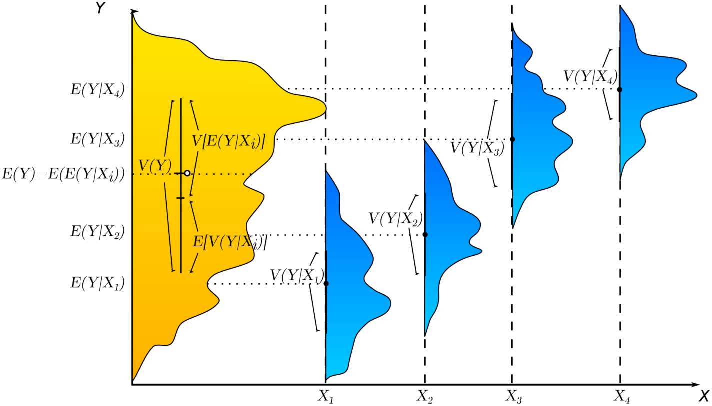
A dog show is not a random sampling of dogs: it is typically limited to individuals that are adult, pure-bred, and exemplary. A histogram of dog weights from a show might plausibly be rather complex.

No Fit



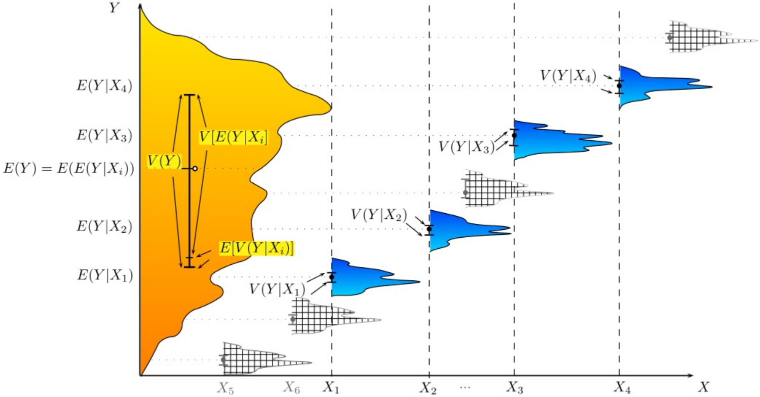
No fit: Young vs old, and short-haired vs long-haired.

Better Fit



Better fit: Pet vs Working breed and less athletic vs more athletic.

Good Fit



Very good fit: weight by breed.

F-distribution

Definition

The F-distribution with d_1 and d_2 degrees of freedom is the distribution of

$$X = \frac{S_1/d_1}{S_2/d_2}$$

where S_1 and S_2 are independent random variables with chi-square distributions with respective degrees of freedom d_1 and d_2.

Also, If $X \sim t_k$ has t-distribution with k degrees of freedom, then

$$X^2 \sim F(1, k)$$

Plot the pdf of F-distribution with d_1, d_2 degrees of freedom:

```
curve(df(x, d1, d2), 0, 5, col="orange", lwd=2, n=300)
```