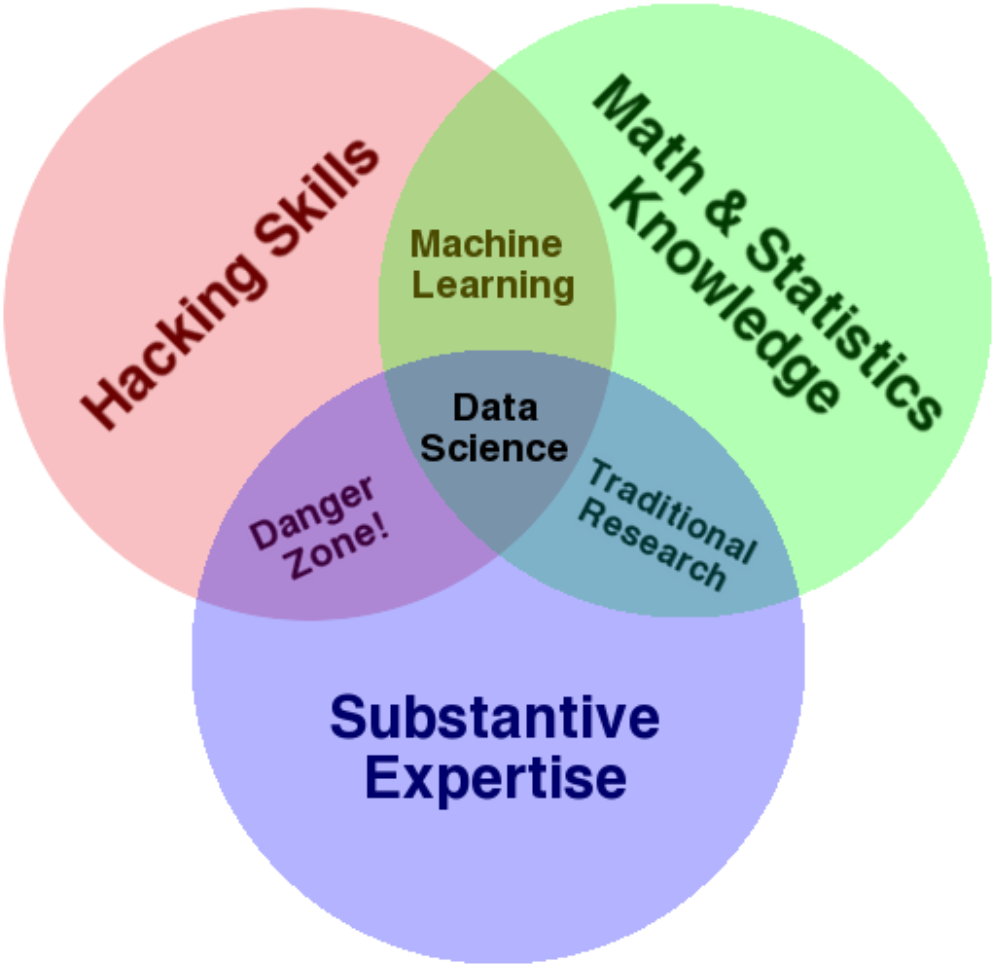
Topic 1. Introduction to Data Science Exploratory Data Analysis vs Statistical Inference

Dr. Lídia Montero - UPC
including notes from John Canny, Michael Franklin, Dan Bruckner, Evan Sparks, Shivaram Venkataraman and Eric Xing

Skills for Data Science



http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

The first war: Terminology

- Analyzing data has a long history!
- There have been many terms that have been used to describe such endeavors:
  - Statistics
  - Artificial Intelligence
  - Machine learning
  - Data analytics
- Since I happen to work in “Data Science” projects perhaps I may be allowed the indulgence of using that terminology…

The Good

Experiments, observations, and numerical simulations in many areas of science and business are currently generating terabytes of data, and in some cases are on the verge of generating petabytes and beyond. Analyses of the information contained in these data sets have already led to major breakthroughs in fields ranging from genomics to astronomy and high-energy physics and to the development of new information-based industries.

- Frontiers in Massive Data Analysis, National Research Council of the National Academies

The Bad

Given a large mass of data, we can by judicious selection construct perfectly plausible unassailable theories—all of which, some of which, or none of which may be right.

- Paul Arnold Srere

The Hopeful

The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that's going to be a hugely important skill in the next decades, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids. Because now we really do have essentially free and ubiquitous data. So the complimentary scarce factor is **the ability to understand that data and extract value from it.**

– Hal Varian, Google's Chief Economist, http://www.mckinsey.com/insights/innovation/hal_varian_on_how_the_web_challenges_managers

My personal goal: Getting students to be able to think critically about data.

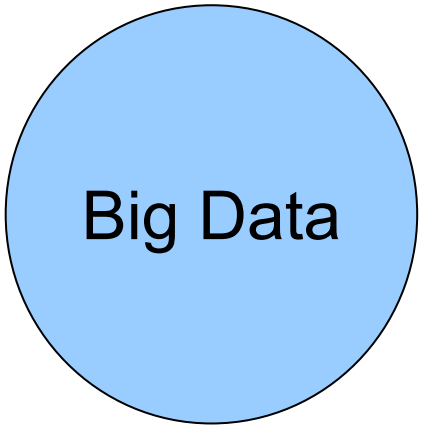
What is Big Data?

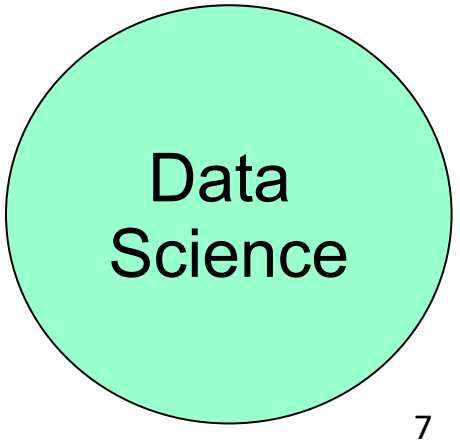
- The are many examples of "data", but what makes some of it “big”? The classic definition revolves around the **three Vs**.
- **Volume, velocity, and variety.**
  - **Volume:** There is a just a lot of it being generated all the time. Things get interesting and “big”, when you can’t fit it all on one computer anymore. Why? There are many ideas here such as MapReduce, Hadoop, etc. that all revolve around being able to process data that goes from Terabytes, to Petabytes, to Exabytes.
  - **Velocity:** Data is being generated very quickly. Can you even store it all? If not, then what do you get rid of and what do you keep?
  - **Variety:** The data types you mention all take different shapes. What does it mean to store them so that you can play with or compare them?
  - **Veracity –** the truthfulness/reliability/quality of data and data sources

http://pl.wikipedia.org/wiki/Green_Giant#mediaviewer/Plik:Jolly_green_giant.jpg

Is Big Data the same as Data Science?

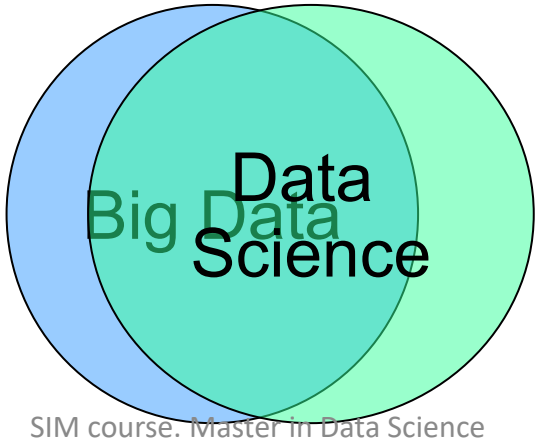
- Are Big Data and Data Science the same thing?
  - I wouldn't say so...
  - Data Science can be done on small data sets.
  - And not everything done using Big Data would necessarily be called Data Science.



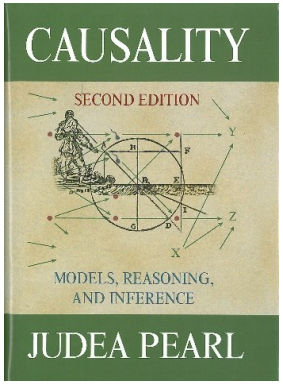


Is Big Data the same as Data Science?

- Are Big Data and Data Science the same thing?
  - I wouldn't say so...
  - Data Science can be done on small data sets.
  - And not everything done using Big Data would necessarily be called Data Science.
  - But there certainly is a substantial overlap!



But... what does getting "knowledge" from data really mean? Are we searching for causality?



•"Causation: The relation between mosquitoes and mosquito bites. Easily understood by both parties but never satisfactorily defined by philosophers and scientists."

•- http://freshspectrum.com/causation/ Michael Scriven, Evaluation Thesaurus, 1991

Most strikingly, society will need to shed some of its obsession for causality in exchange for simple correlations: not knowing why but only what.

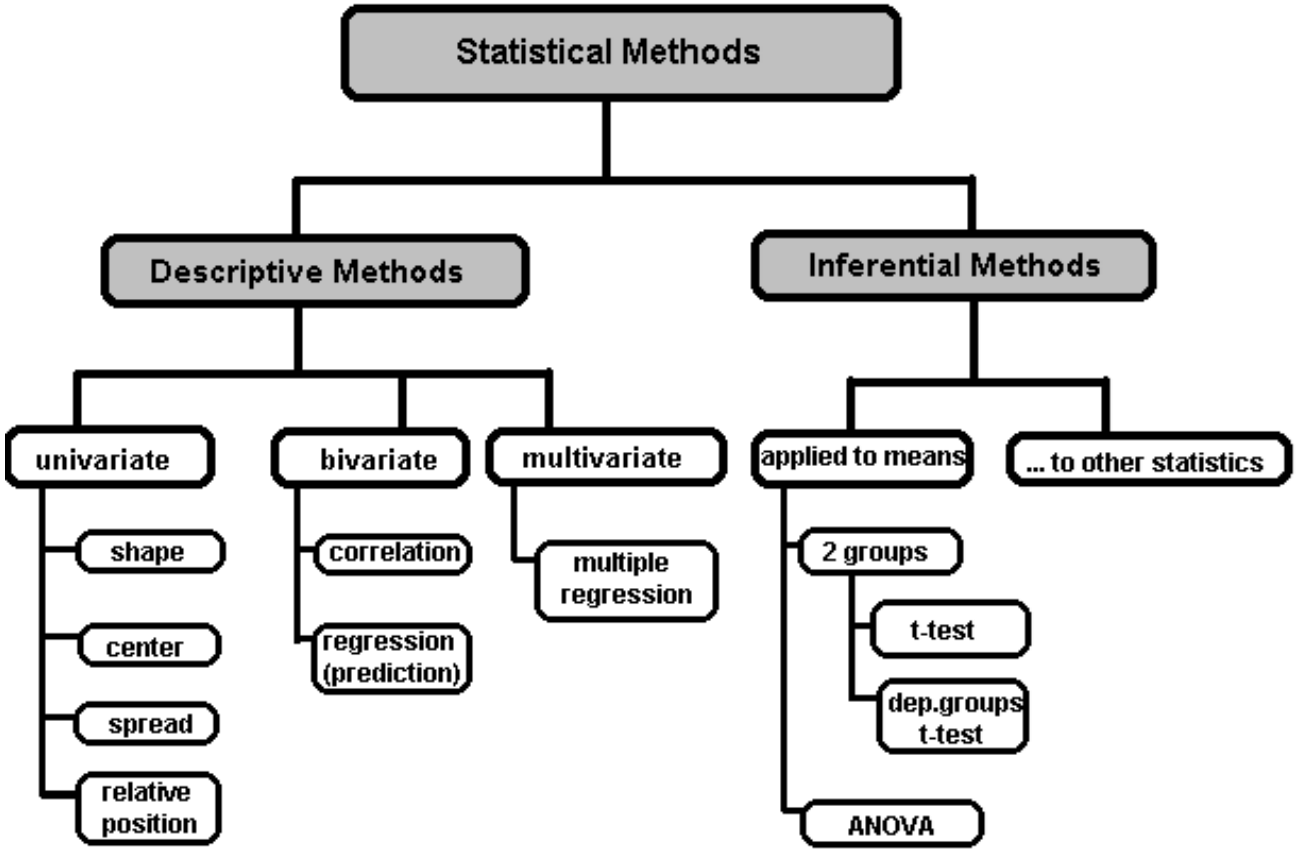
- Big Data: A Revolution that Transform How We Live, Work, and Think, Viktor Mayer-Schönberger and Kenneth Cukier.

Data Science Tools for Students: Free!

- Software:
- R: https://www.r-project.org/
  - Many library contributions
- Rstudio: https://www.rstudio.com/
  - Integrated development environment (IDE) for R
- Python
- http://www.python.org/
  - iPython: http://ipython.org/
  - Numpy: http://www.numpy.org/
  - Pandas: http://pandas.pydata.org/
  - Matplotlib: http://matplotlib.org/
  - Mayavi: http://mayavi.sourceforge.net/
  - Scikit-learn: http://scikit-learn.org/stable/

- Data:
- UCI Machine learning repository
  - http://archive.ics.uci.edu/ml/
- Kaggle
  - https://www.kaggle.com/
- U.S. Government
  - https://www.data.gov/

A Taxonomy of Statistics



Statistical Description of Data

- Statistics describes a numeric set of data by its
  - Center – Location (mean or robust median)
  - Variability – Scale (variance or quartiles/percentiles)
  - Shape (not always present in RV) (skewness, kurtosis)

  - Moments: raw and centered

- Statistics describes a categorical set of data by
  - Frequency, percentage or proportion of each category

Some Definitions

Variable - any characteristic of an individual or entity. A variable can take different values for different individuals. Variables can be *categorical* or *quantitative*. Per S. S. Stevens…

- **Nominal** - Categorical variables with no inherent order or ranking sequence such as names or classes (e.g., gender). Value may be a numerical, but without numerical value (e.g., I, II, III). The only operation that can be applied to Nominal variables is enumeration.
- **Ordinal** - Variables with an inherent rank or order, e.g. mild, moderate, severe. Can be compared for equality, or greater or less, but not *how much* greater or less.
- **Interval** - Values of the variable are ordered as in Ordinal, and additionally, differences between values are meaningful, however, the scale is not absolutely anchored. Calendar dates for example. Addition and subtraction, but not multiplication and division are meaningful operations.
- **Ratio** - Variables with all properties of Interval plus an absolute, non-arbitrary zero point, e.g. age, weight, temperature (Kelvin). Addition, subtraction, multiplication, and division are all meaningful operations. Just, numerical and continuous values.

Some Definitions

Distribution - (of a variable) tells us what values the variable takes and how often it takes these values.

- Unimodal - having a single peak
- Bimodal - having two distinct peaks
- Symmetric - left and right half are mirror images.
  - Skewness: asymmetry (0 for normally distributed data)
  - Kurtosis: Peak/flat shape (3 for normally distributed data)

Outline

- Exploratory Data Analysis
  - Numerical Summary
  - Chart types
  - Some important distributions
  - Hypothesis Testing

Numerical Presentation

A fundamental concept in summary statistics is that of a *central value* for a set of observations and the extent to which the central value characterizes the whole set of data.

Measures of central value such as the mean or median must be coupled with measures of data dispersion (e.g., average distance from the mean) to indicate how well the central value characterizes the data as a whole.

To understand how well a central value characterizes a set of observations, let us consider the following two sets of data:

A: 30, 50, 70
B: 40, 50, 60

The mean of both two data sets is 50. But, the distance of the observations from the mean in data set A is larger than in the data set B. Thus, the mean of data set B is a better representation of the data set than is the case for set A.

Methods of Center Measurement

Center measurement is a summary measure of the overall level of a dataset

Commonly used methods are mean, median, mode, geometric mean etc.

Mean: Summing up all the observation and dividing by number of observations. Mean of 20, 30, 40 is (20+30+40)/3 = 30.

Notation : Let $x_1, x_2, \ldots x_n$ are n observations of a variable x. Then the mean of this variable,

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

Methods of Center Measurement

Median: The middle value in an ordered sequence of observations. That is, to find the median we need to order the data set and then find the middle value. In case of an even number of observations the average of the two middle most values is the median. For example, to find the median of {9, 3, 6, 7, 5}, we first sort the data giving {3, 5, 6, 7, 9}, then choose the middle value 6. If the number of observations is even, e.g., {9, 3, 6, 7, 5, 2}, then the median is the average of the two middle values from the sorted sequence, in this case, (5 + 6) / 2 = 5.5.

Mode: The value that is observed most frequently. The mode is undefined for sequences in which no observation is repeated.

Mean or Median

The median is less sensitive to outliers (extreme scores) than the mean and thus a better measure than the mean for highly skewed distributions, e.g. family income.

For example mean of 20, 30, 40, and 990 is (20+30+40+990)/4 =270.

The median of these four observations is (30+40)/2 =35.

Here 3 observations out of 4 lie between 20-40. So, the mean 270 really fails to give a realistic picture of the major part of the data. It is influenced by extreme value 990.

Methods of Variability Measurement

Variability (or dispersion) measures the amount of scatter in a dataset.

Commonly used methods: *range*, *variance*, *standard deviation*, *interquartile range*, *coefficient of variation etc*.

Range: The difference between the largest and the smallest observations. The range of 10, 5, 2, 100 is (100-2)=98. It's a crude measure of variability.

Methods of Variability Measurement

Variance: The variance of a set of observations is the average of the squares of the deviations of the observations from their mean. In symbols, the variance of the n observations x_1, x_2,…x_n is

$$S^2 = \frac{(x_1 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1}$$

Variance of 5, 7, 3? Mean is (5+7+3)/3 = 5 and the variance is

$$\frac{(5-5)^2 + (3-5)^2 + (7-5)^2}{3-1} = 4$$

Standard Deviation: Square root of the variance. The standard deviation of the above example is 2.

Methods of Variability Measurement

Quartiles: Data can be divided into four regions that cover the total range of observed values. Cut points for these regions are known as quartiles.

In notations, quartiles of a data is the $((n+1)/4)q^{th}$ observation of the data, where q is the desired quartile and n is the number of observations of data.

The first quartile (Q1) is the first 25% of the data. The second quartile (Q2) is between the 25^{th} and 50^{th} percentage points in the data. The upper bound of Q2 is the median. The third quartile (Q3) is the 25% of the data lying between the median and the 75% cut point in the data.

Q1 is the median of the first half of the ordered observations and Q3 is the median of the second half of the ordered observations.

Methods of Variability Measurement

In the following example Q1= ((15+1)/4)1 =4th observation of the data. The 4th observation is 11. So Q1 is of this data is 11.

An example with 15 numbers

3 6 7 11 13 22 30 40 44 50 52 61 68 80 94

Q1 Q2 Q3

The first quartile is Q1=11. The second quartile is Q2=40 (This is also the Median.) The third quartile is Q3=61.

Inter-quartile Range (IQR): Difference between Q3 and Q1. Inter-quartile range of the previous example is 61- 11=50. The middle half of the ordered data lie between 11 and 61.

Deciles and Percentiles

Deciles: If data is ordered and divided into 10 parts, then cut points are called Deciles

Percentiles: If data is ordered and divided into 100 parts, then cut points are called Percentiles. 25^{th} percentile is the Q1, 50^{th} percentile is the Median (Q2) and the 75^{th} percentile of the data is Q3.

In notations, percentiles of a data is the ((n+1)/100)p th observation of the data, where p is the desired percentile and n is the number of observations of data.

Coefficient of Variation: The standard deviation of data divided by it's mean. It is usually expressed in percent.

Coefficient of Variation = $\frac{\sigma}{\bar{x}} \times 100$

Descriptive vs. Inferential Statistics

- **Descriptive:** e.g., Median; describes data you have but can't be generalized beyond that
  - We'll talk about Exploratory Data Analysis
- **Inferential:** e.g., t-test, that enable inferences about the population beyond our data
  - These are the techniques we'll leverage for Machine Learning and Prediction

Examples of Business Questions

- **Simple (descriptive) Stats**
  - "Who are the most profitable customers?"
- **Hypothesis Testing**
  - "Is there a difference in value to the company of these customers?"
- **Segmentation/Classification**
  - What are the common characteristics of these customers?
- **Prediction**
  - Will this new customer become a profitable customer? If so, how profitable?

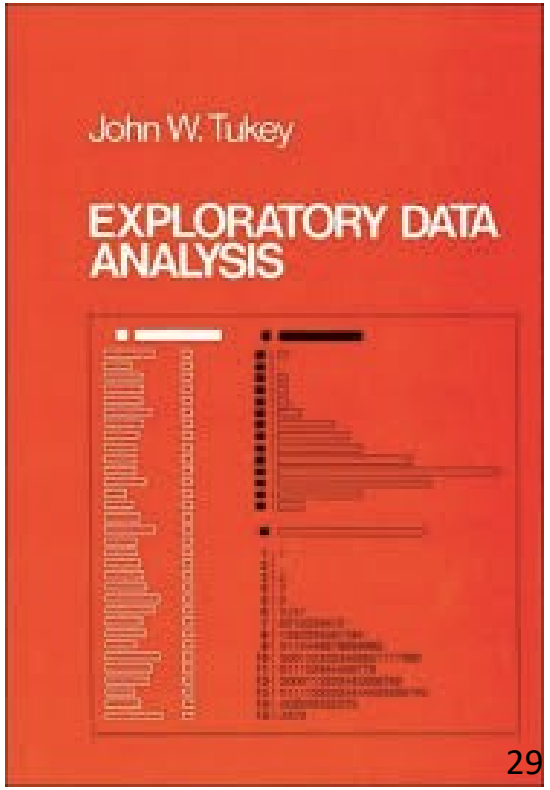
adapted from Provost and Fawcett, "Data Science for Business"

Applying techniques

- Most business questions are causal: what would happen if? (e.g. I show this ad)
- But its easier to ask correlational questions, (what happened in this past when I showed this ad).
- **Supervised Learning:**
  - Classification and Regression
- **Unsupervised Learning:**
  - Clustering and Dimension reduction
- Note: Unsupervised Learning is often used inside a larger Supervised learning problem.

Exploratory Data Analysis 1977

- Based on insights developed at Bell Labs in the 60's
- Techniques for visualizing and summarizing data
- What can the data tell us? (in contrast to "confirmatory" data analysis)
- Introduced many basic techniques:
  - 5-number summary, box plots, stem and leaf diagrams,…
- **5 Number summary**:
  - extremes (min and max)
  - median & quartiles
  - More robust to skewed & longtailed distributions



The Trouble with Summary Stats

Set A		Set B		Set C		Set D	
X	Y	X	Y	X	Y	X	Y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.11	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

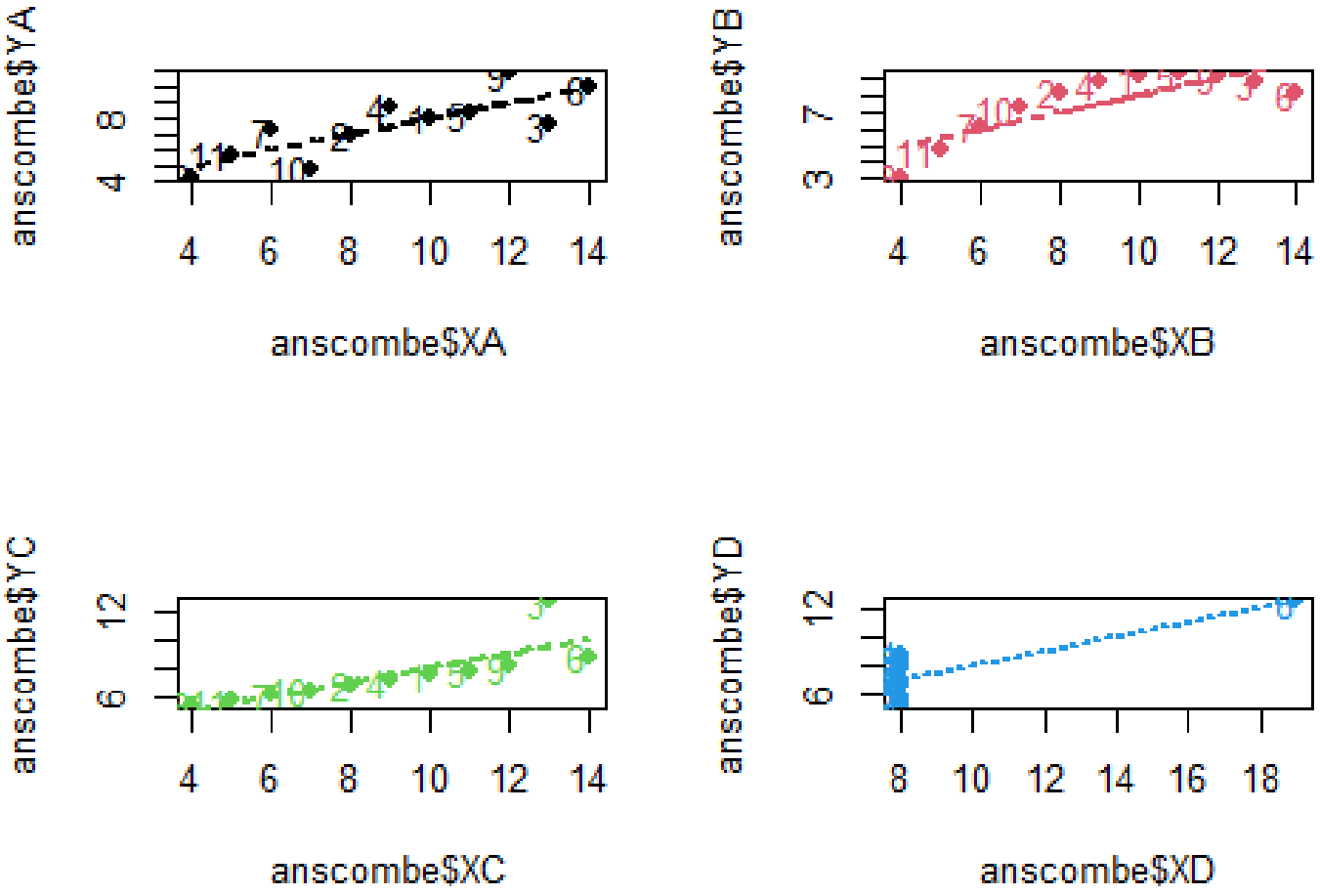
Summary Statistics Linear Regression

$u_X = 9.0$ $\sigma_X = 3.317$ $Y = 3 + 0.5\ X$

$u_Y = 7.5$ $\sigma_Y = 2.03$ $R^2 = 0.67$

[Anscombe 73]

Looking at Data (this is the aim of one lab session)



Data Presentation

- Data Art

The "R" Language

- An evolution of the "S" language developed at Bell labs for EDA.
- Idea was to allow interactive exploration and visualization of data.
- The preferred language for statisticians, used by many other data scientists.
- Features:
  - Probably the most comprehensive collection of statistical models and distributions.
  - CRAN: a very large resource of open source statistical models.

Chart examples from Jeff Hammerbacher's 2012 CS194 class

Chart types

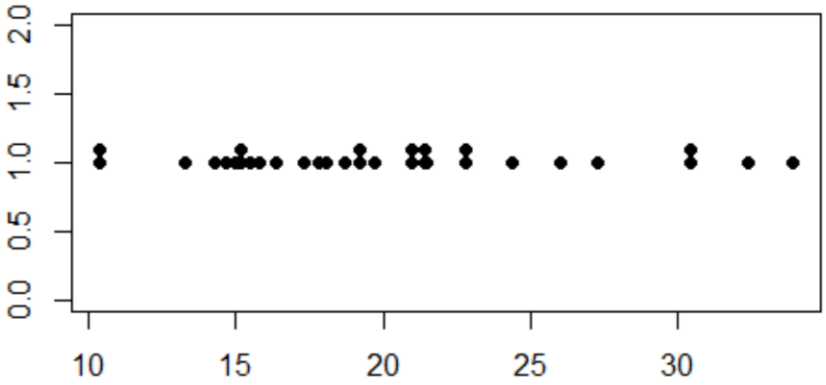
- Single variable
  - Dot plot
  - Jitter plot
  - Error bar plot
  - Box-and-whisker plot
  - Histogram
  - Kernel density estimate
  - Cumulative distribution function

(note: examples using base and ggplot2 library from R)

Chart types

- **Dot plot**

```
#create stacked dot plot
stripchart(mtcars$mpg, method = "stack")
axis(2)
```




```
ggplot(mtcars, aes(x = mpg)) +
geom_dotplot(binwidth=1)
```

#> Bin width defaults to 1/30 of the range of the data. Pick better value with `binwidth`

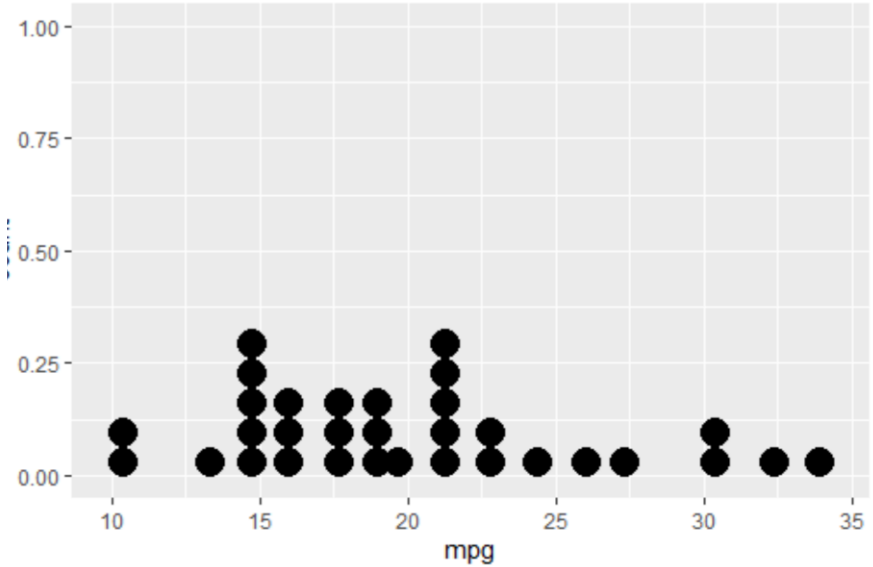
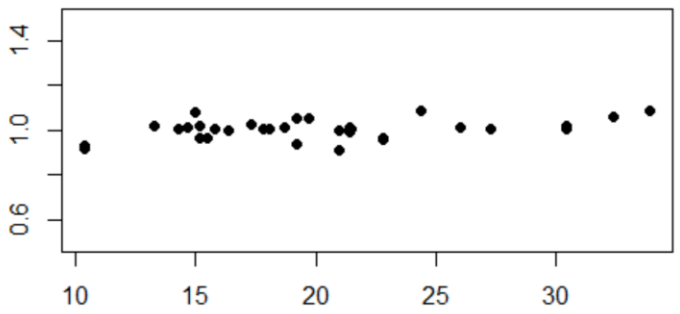


Chart types

- **Jitter plot**
- Noise added to the y-axis to spread the points

stripchart(mtcars$mpg, method = "jitter", pch=19)
axis(2)



ggplot(mtcars, aes(x = mpg, y=1)) +
geom_jitter()

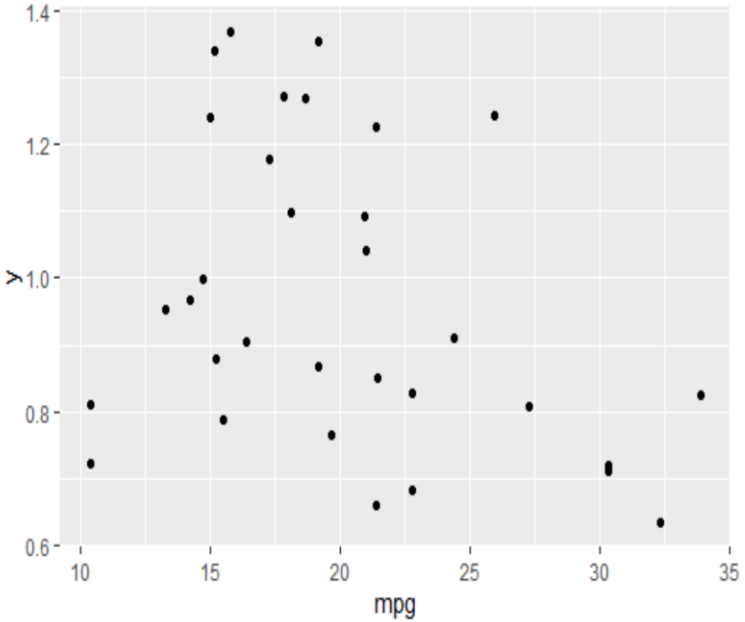


Chart types

- **Box-and-whisker plot** : a graphical form of 5-number summary (Tukey)

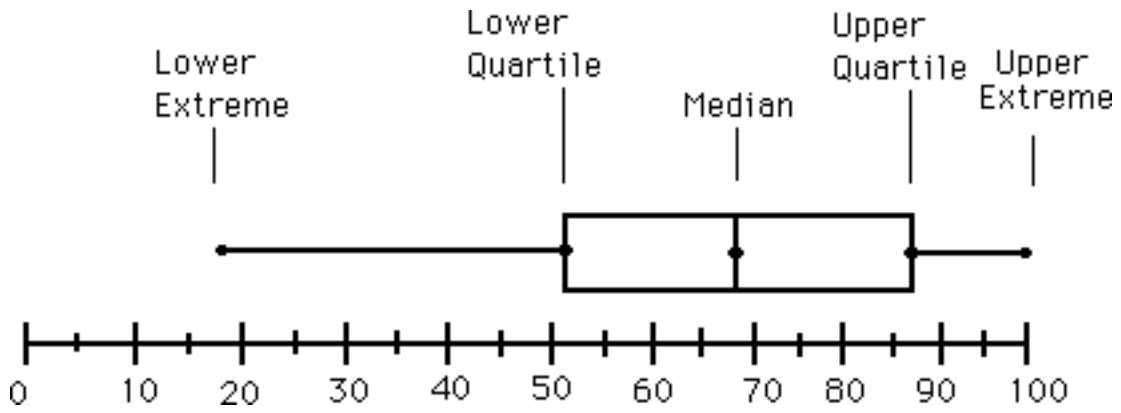
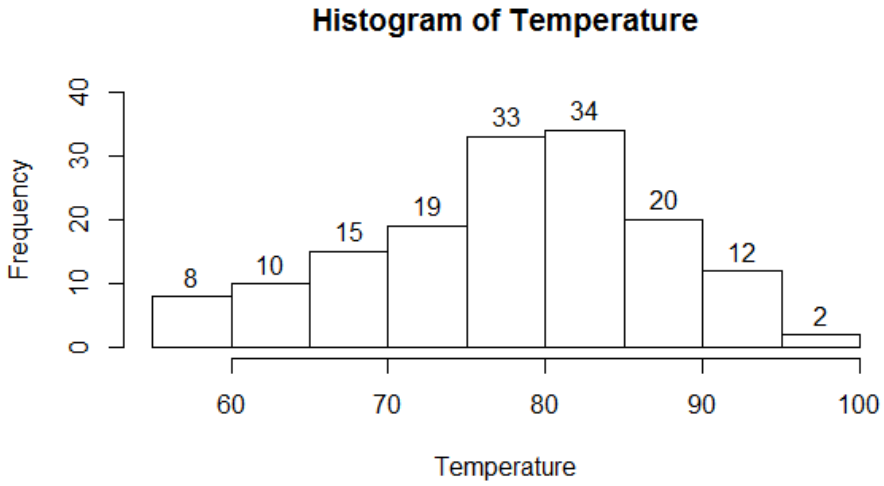
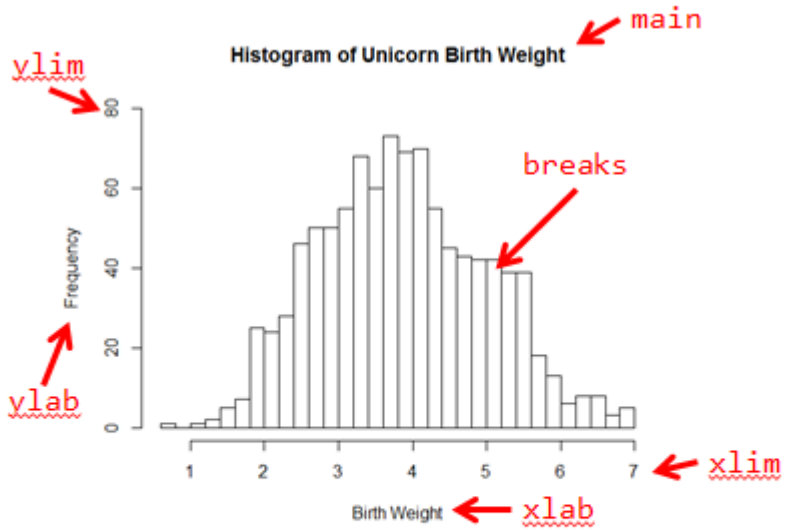


Chart types

- **Histogram:** hist(var, freq=T, breaks="Sturges")




```
h <- hist(Temperature,ylim=c(0,40))
# From  airquality  data set
text(h$mids,h$counts,labels=h$counts)
```

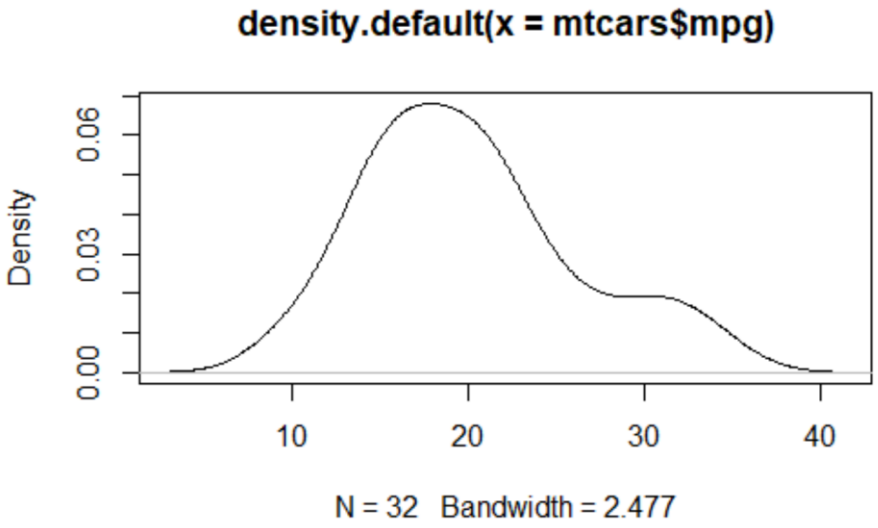



```
#~~ FINAL PLOT:

hist(unicorns$birthweight,                          # x value
     breaks = 40,                                   # number of cells
     xlab = "Birth Weight",                         # x-axis label
     main = "Histogram of Unicorn Birth Weight",    # plot title
     ylim = c(0,80))                                # limits of the y axis (min,max)

```

Chart types

- **Kernel density estimate**

plot(density(mtcars$mpg))



ggplot(mtcars, aes(x = mpg))
+ geom_density()

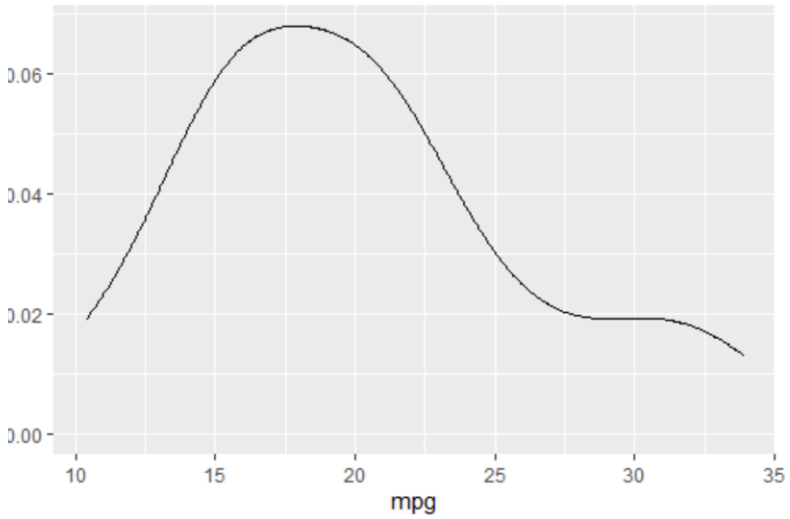


Chart types

- Histogram and Kernel Density Estimates
  - Histogram
    - Proper selection of bin width is important
    - Outliers should be discarded
  - KDE (like a smooth histogram)
    - Kernel function
      - Box, Epanechnikov, Gaussian
    - Kernel bandwidth

Chart types

- **Cumulative distribution function**

```
> plot(ecdf(f500.ca$revenues))
```

- Integral of the histogram – simpler to build than KDE (don't need smoothing)

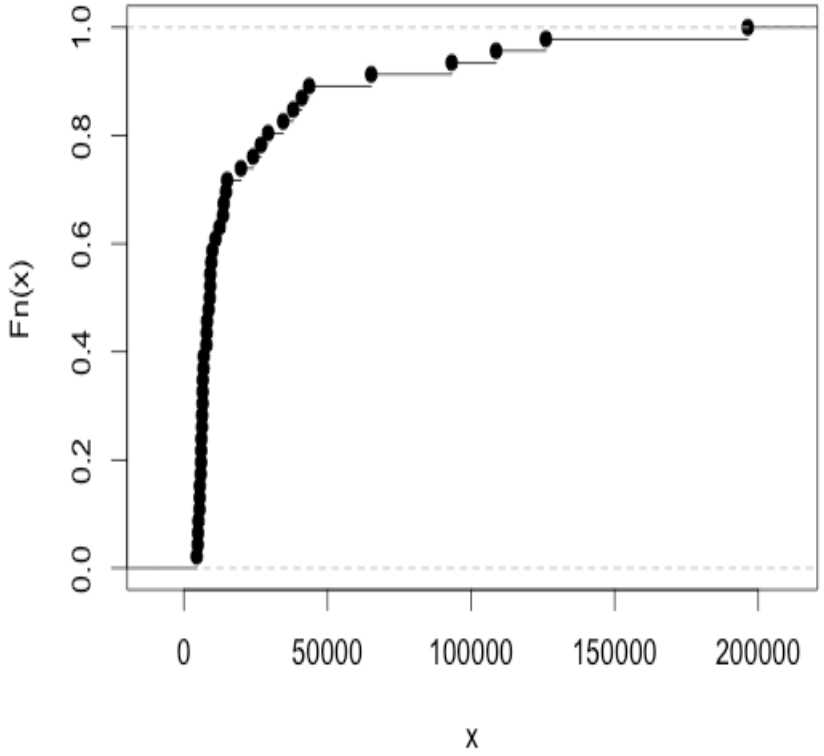
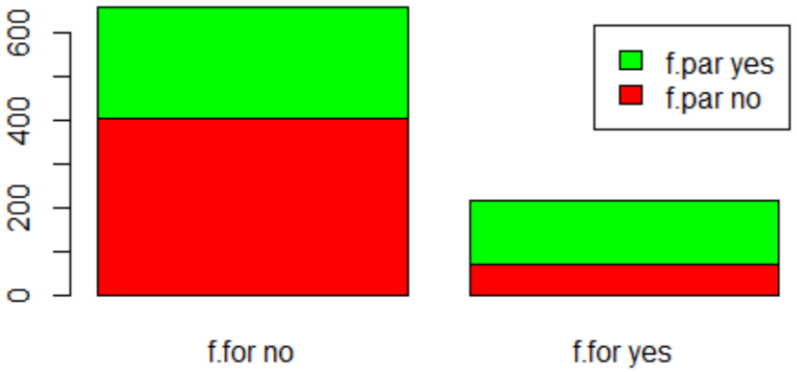


Chart types

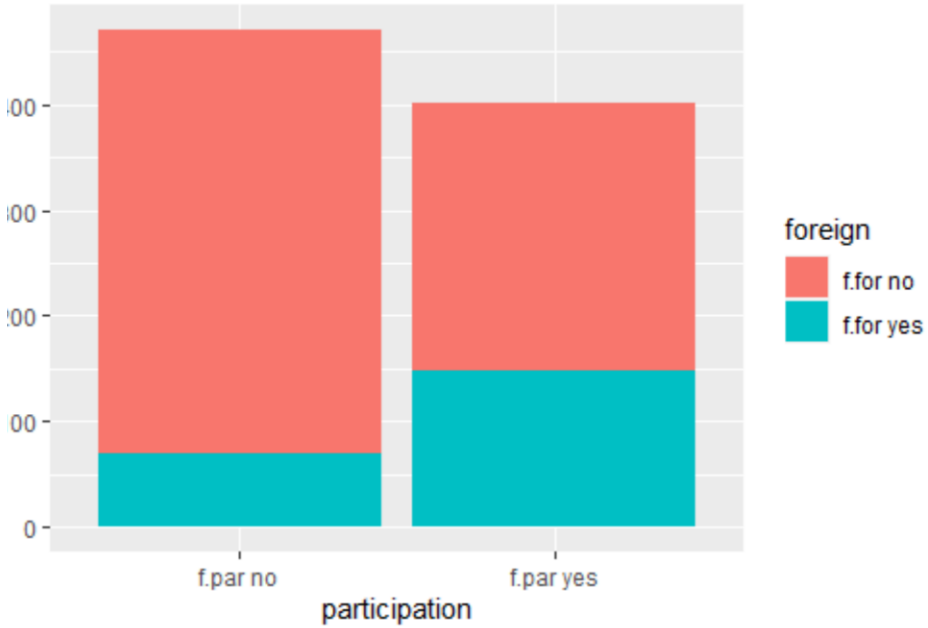
- Two variables
  - Bar chart
  - Scatter plot
  - Line plot
  - Log-log plot

Chart types

- **Bar plot:** at least one variable is discrete



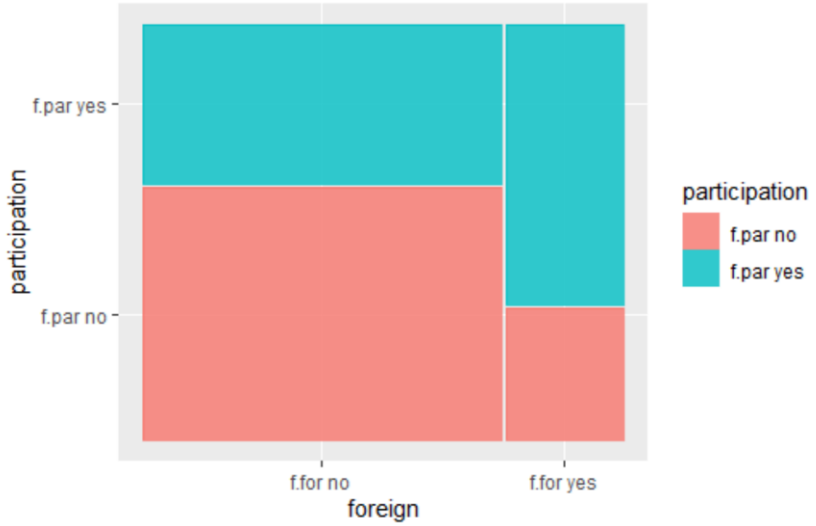

```
tt<-
table(df$participation,df$foreign)

barplot(tt, col = c("red","green"),
legend.text = rownames(tt), beside
= FALSE)
```




```
ggplot(data = df, aes( x = participation )) +
  geom_bar(aes(fill = foreign))
```

Chart types

- **Mosaic plot:** two variables are discrete

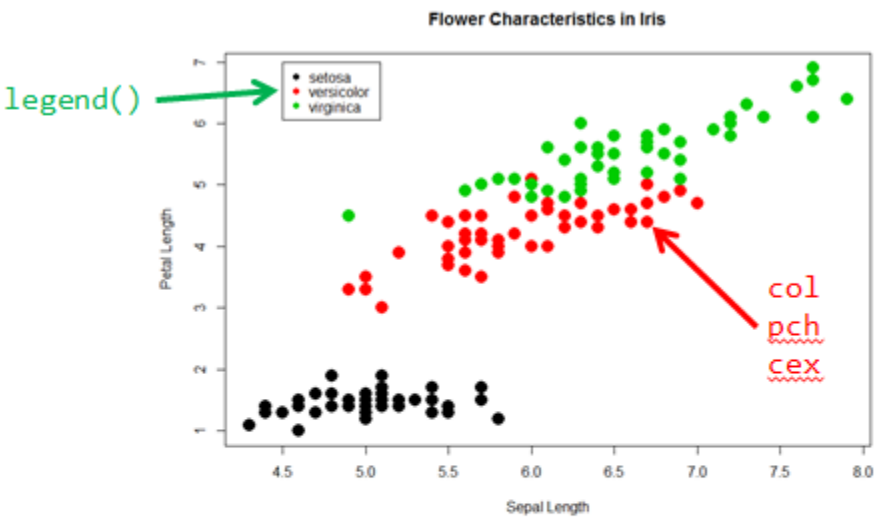



```
plot(participation~foreign, data = df , col =
c("green","red"))
```

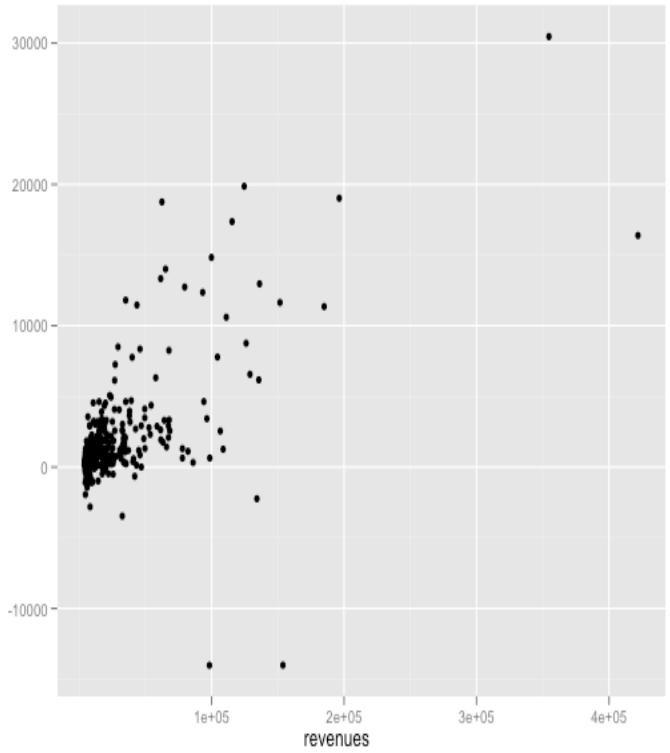
```
ggplot(data = df) +
  geom_mosaic(aes(x = product(foreign),
fill = participation))
```

Chart types

- **Scatter plot: plot(x=revenues, y=profits, main="xxx")**




```
261 # FINAL PLOT
262
263 plot(iris$Sepal.Length, iris$Petal.Length,        # x variable, y variable
264      col = iris$Species,                          # colour by species
265      pch = 16,                                    # type of point to use
266      cex = 2,                                     # size of point to use
267      xlab = "Sepal Length",                       # x axis label
268      ylab = "Petal Length",                       # y axis label
269      main = "Flower Characteristics in Iris")     # plot title
270
271 legend (x = 4.5, y = 7, legend = levels(iris$Species), col = c(1:3), pch = 16)
272 # legend with titles of iris$Species and colours 1 to 3, point type pch at coords (x,y)
273
```



ggplot(stores, aes(revenues, profits))
+ geom_point(size=4)

Chart types

- **Line plot:** plot(Days, Stock, type = "l“)

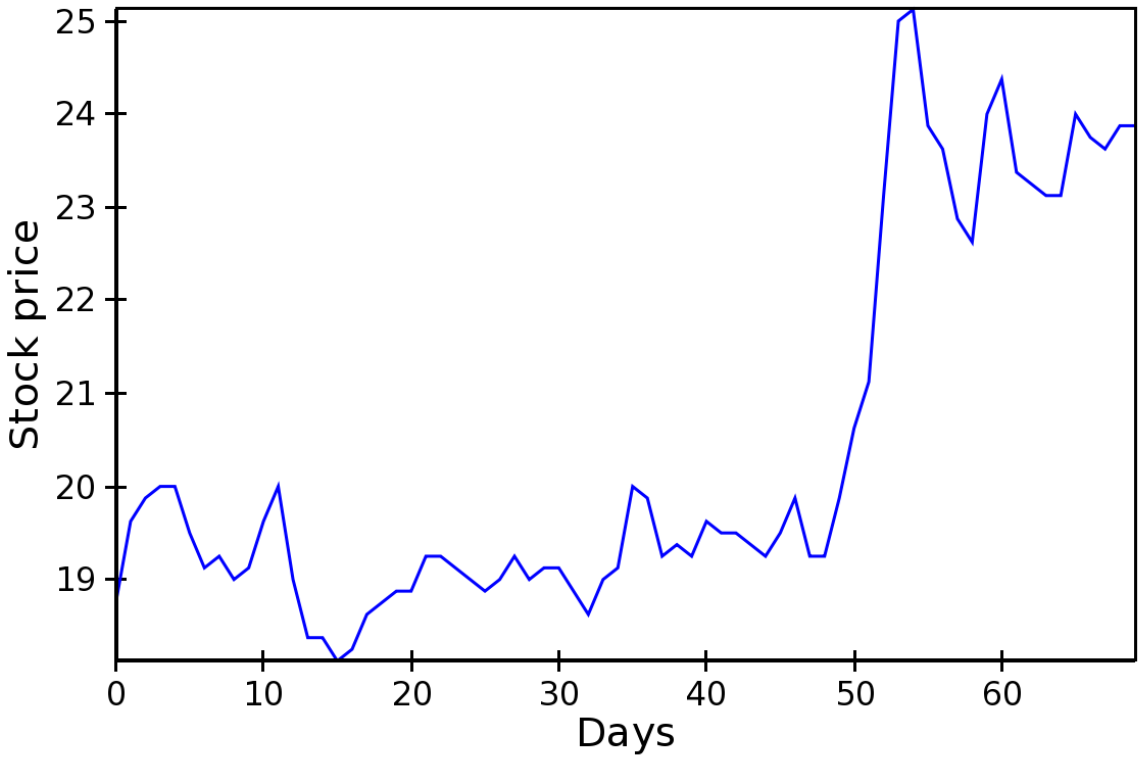
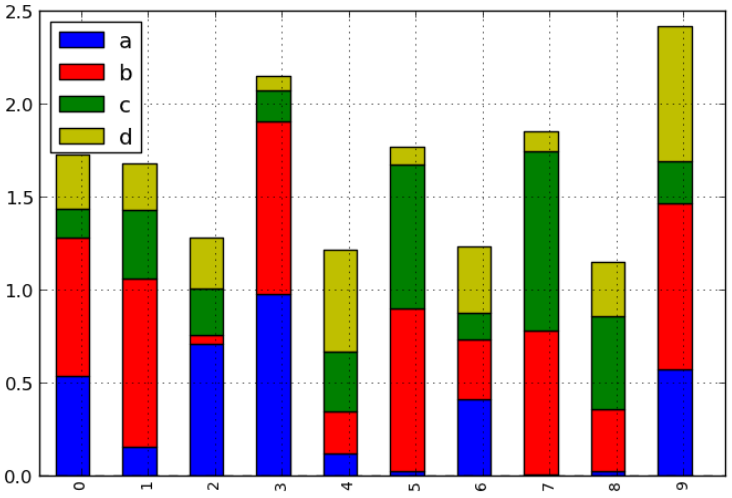


Chart types

- More than two variables
  - Stacked plots
  - Parallel coordinate plot
  - Radar plots

Chart types

- **Stacked plot:** stack variable is discrete:



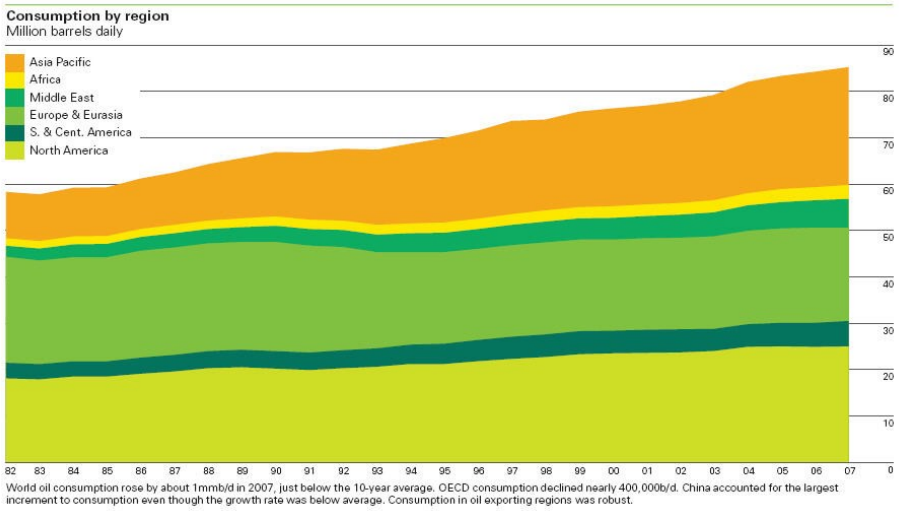
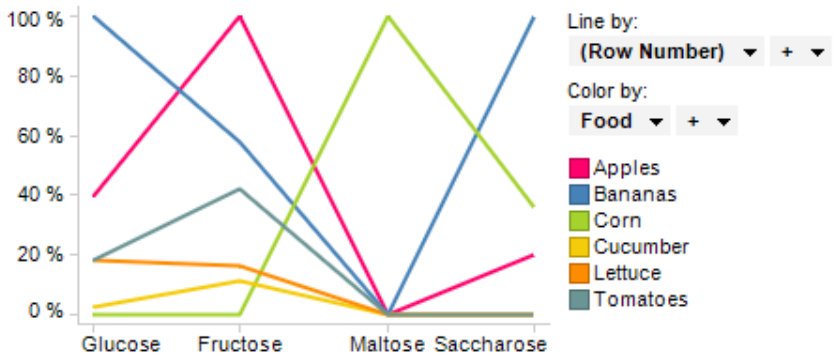


Chart types

- **Parallel coordinate plot:** one discrete variable, an arbitrary number of other variables:



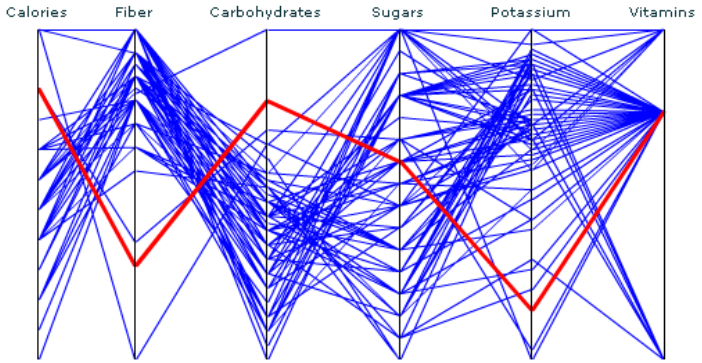
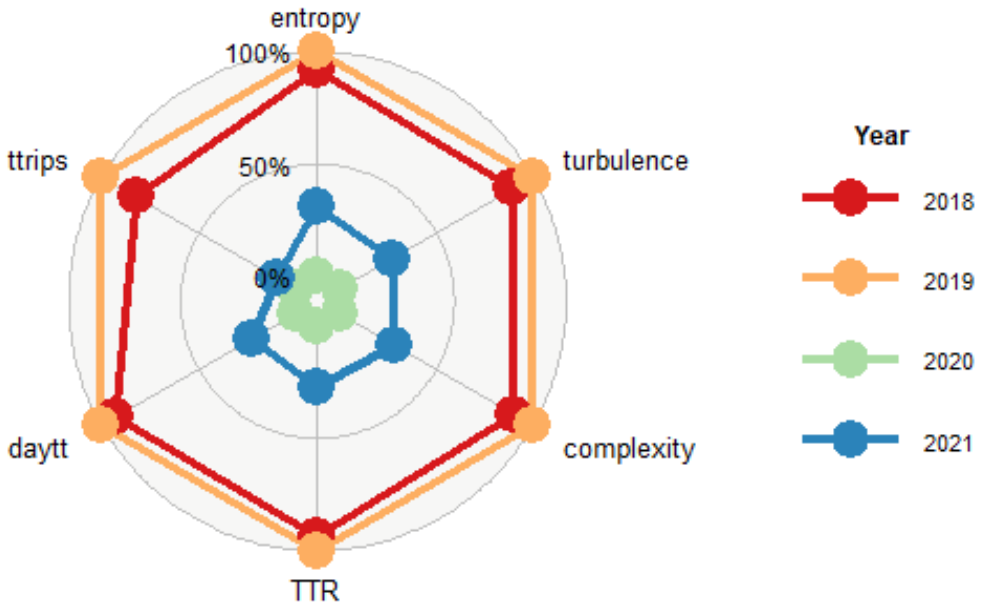


Chart types

- **Radar plot:** one discrete variable, an arbitrary number of other variables



Normal Distributions, Mean, Variance

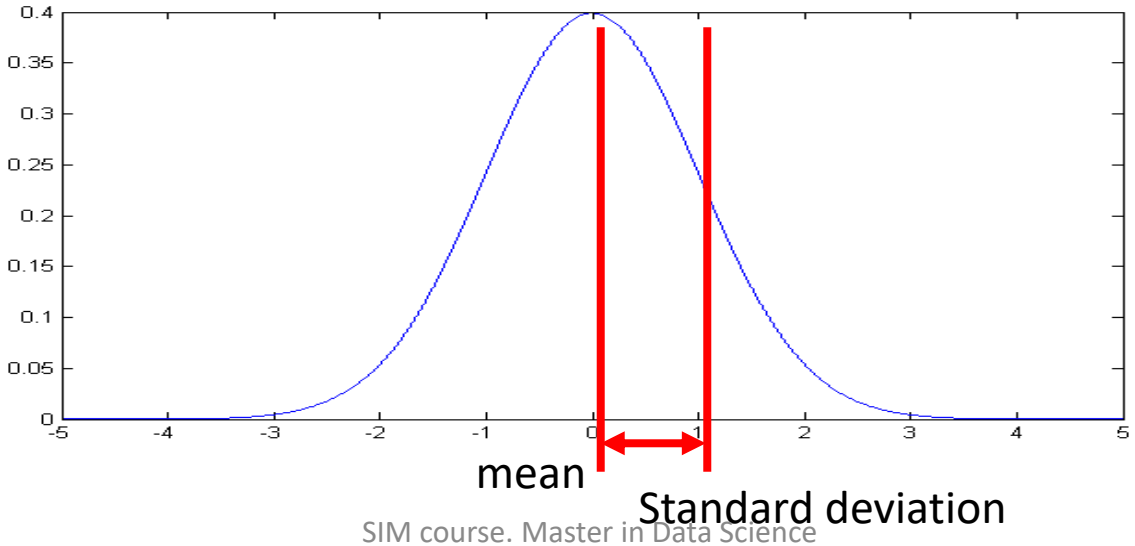
The mean of a set of values is just the average of the values.

Variance a measure of the width of a distribution. Specifically, the variance is the mean squared deviation of samples from the sample mean:

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

The standard deviation is the square root of variance.

The normal distribution is completed characterized by mean and variance.



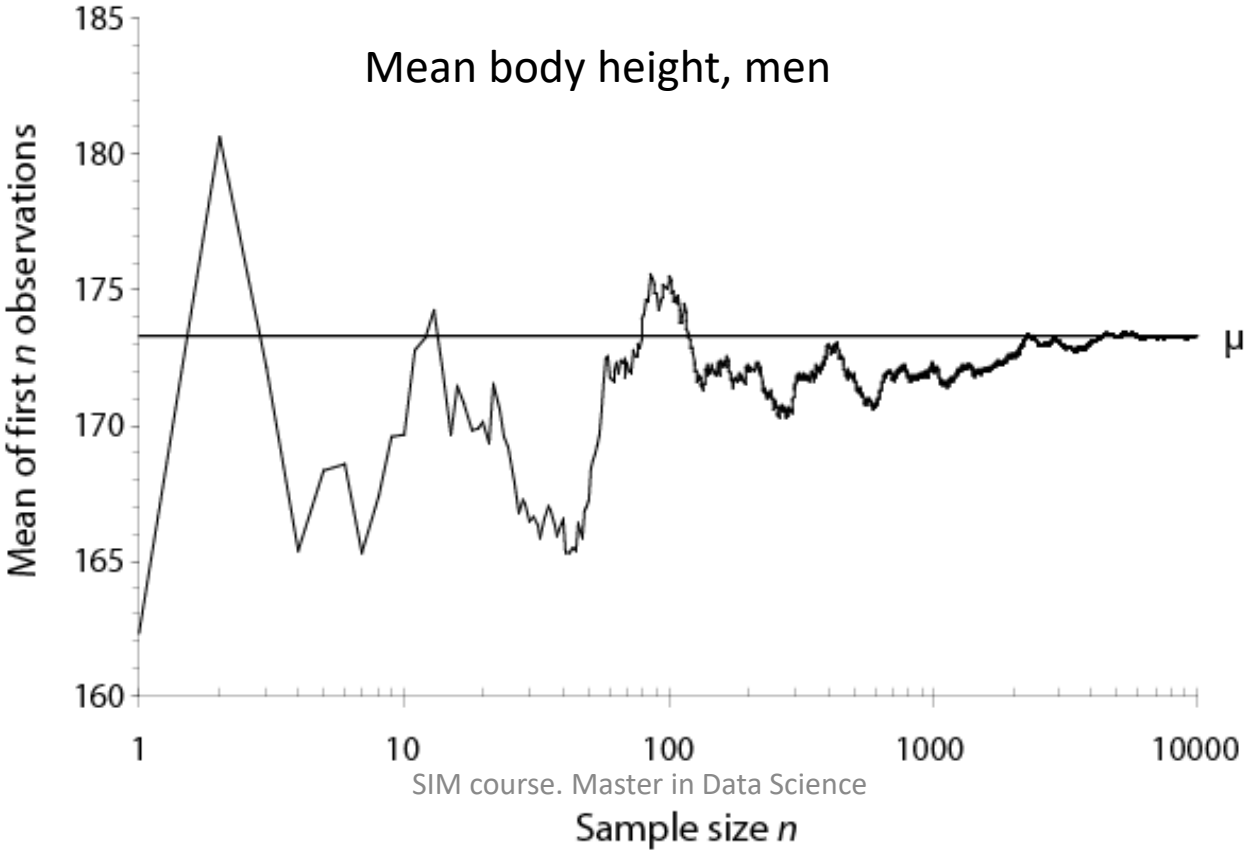
Distributions

Some other important distributions:

- **Poisson:** the distribution of counts that occur at a certain "rate".
  - Observed frequency of a given term in a corpus.
  - Number of web site clicks in an hour.
- **Exponential:** the interval between two such events.
- **Zipf/Pareto/Yule distributions:** govern the frequencies of different terms in a document, or web site visits.
- **Binomial/Multinomial:** The number of counts of events (e.g. die tosses = 6) out of n trials.
- **Chi-squared, Student-t, Fisher distributions**: useful for hypothesis testing.
  - You should understand the distribution of your data before applying any model.

Law of Large Numbers

As a sample gets larger and larger, the x-bar approaches μ. Figure demonstrates results from an experiment done in a population with μ = 173.3

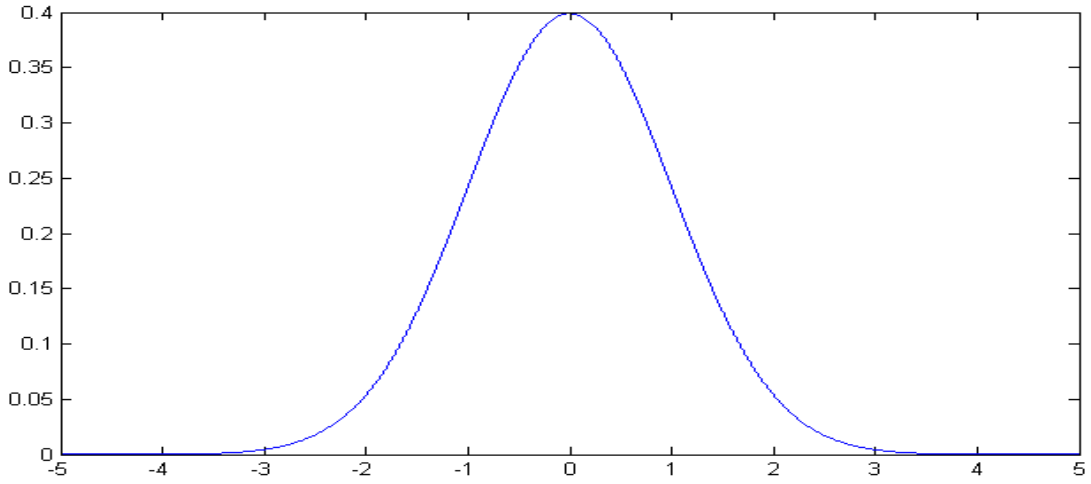


Central Limit Theorem

The distribution of the sum (or mean) of a set of n independent identically-distributed random variables Xi approaches a normal distribution as $n \rightarrow \infty$.

The common parametric statistical tests, like t-test and ANOVA assume normally-distributed data, but depend on sample mean and variance measures of the data.

They typically work reasonably well for data that are not normally distributed as long as the samples are not too small.



Central Limit Theorem

- The CLT states that regardless of the distribution of the original data, the average of the data is Normally distributed

- Why such a big deal?

- Allows for hypothesis testing (p-values) and CI's to be estimated

Central Limit Theorem

- If a random sample is drawn from a population, a statistic (like the sample average) follows a distribution called a "sampling distribution".

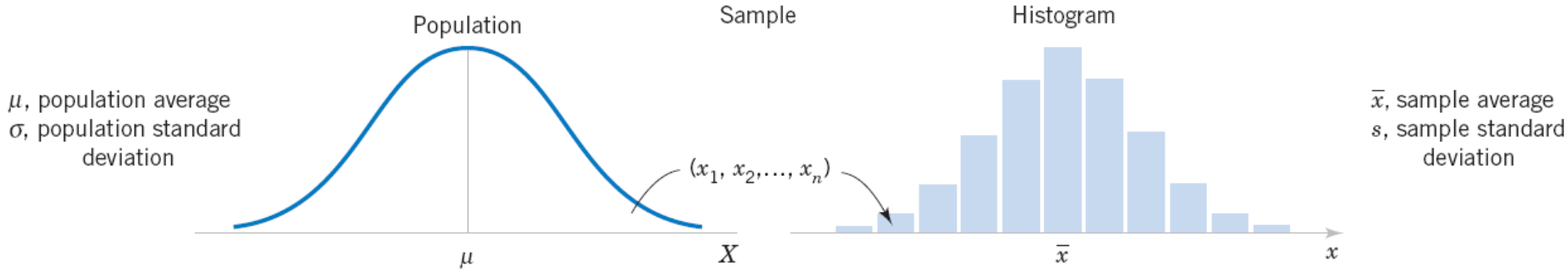
- CLT tells us the sampling distribution of the average is a Normal distribution, regardless of the distribution of the original observations, as the sample size increases.

1-1 Statistical Inference

- The field of statistical inference consists of those methods used to make decisions or draw conclusions about a **population.**

- These methods utilize the information contained in a **sample** from the population in drawing conclusions.

1-1 Statistical Inference



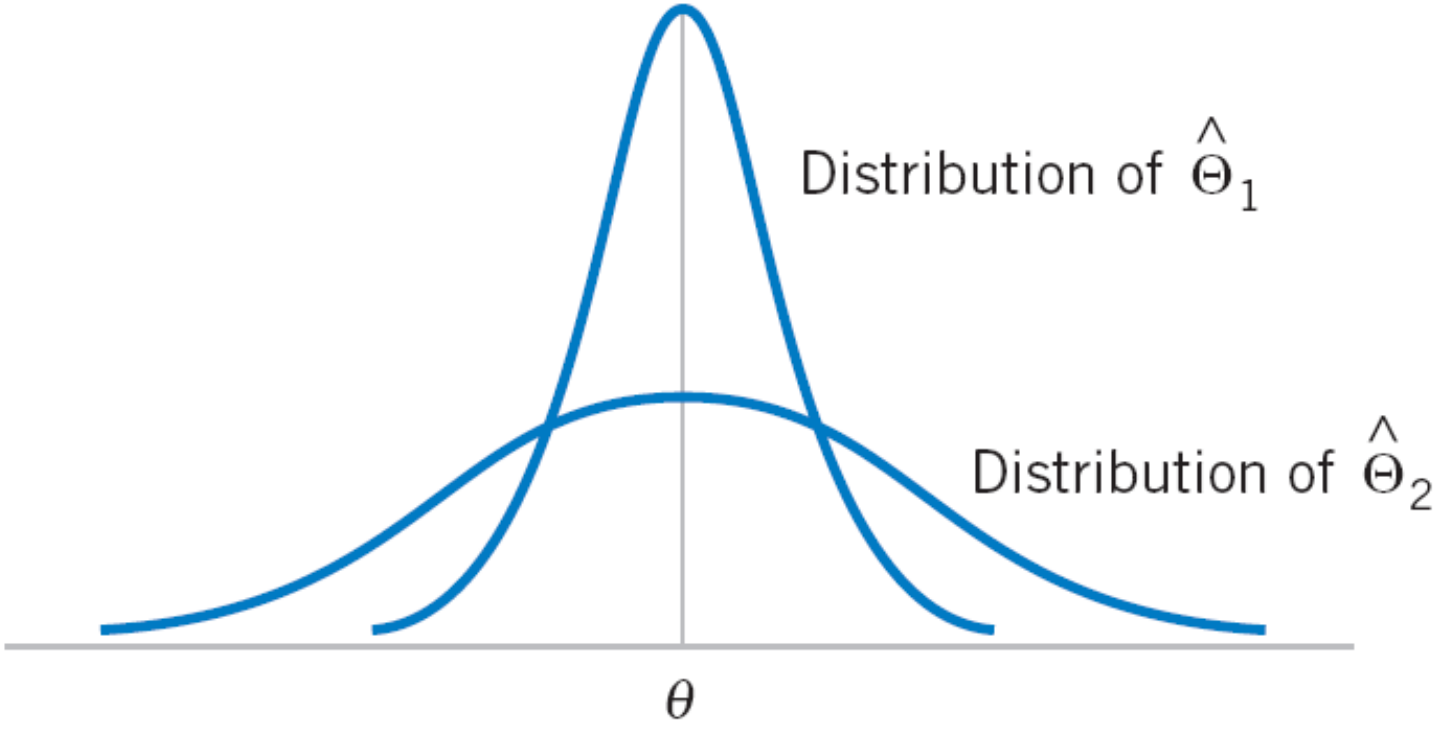
Relationship between a population and a sample.

1-2 Point Estimation

A **point estimate** of some population parameter θ is a single numerical value $\hat{\theta}$ of a statistic $\hat{\Theta}$.

Unknown Parameter θ	Statistic $\hat{\Theta}$	Point Estimate $\hat{\theta}$
μ	$\bar{X} = \frac{\Sigma X_i}{n}$	$\bar{x}$
σ^2	$S^2 = \frac{\Sigma(X_i - \bar{X})^2}{n-1}$	s^2
p	$\hat{P} = \frac{X}{n}$	$\hat{p}$
$\mu_1 - \mu_2$	$\bar{X}_1 - \bar{X}_2 = \frac{\Sigma X_{1i}}{n_1} - \frac{\Sigma X_{2i}}{n_2}$	$\bar{x}_1 - \bar{x}_2$
$p_1 - p_2$	$\hat{P}_1 - \hat{P}_2 = \frac{X_1}{n_1} - \frac{X_2}{n_2}$	$\hat{p}_1 - \hat{p}_2$

1-2 Point Estimation



The sampling distributions of two unbiased estimators $\hat{\Theta}_1$ and $\hat{\Theta}_2$.

1-2 Point Estimation

If we consider all unbiased estimators of θ, the one with the smallest variance is called the **minimum variance unbiased estimator** (MVUE).

The **mean square error** of an estimator $\hat{\Theta}$ of the parameter θ is defined as

$$\text{MSE}(\hat{\Theta}) = E(\hat{\Theta} - \theta)^2 = VAR[\hat{\Theta}] + Bias[\hat{\Theta} - \theta]^2$$

The **standard error** of a statistic is the standard deviation of its sampling distribution. If the standard error involves unknown parameters whose values can be estimated, substitution of these estimates into the standard error results in an **estimated standard error.**

1-3 Hypothesis Testing

1-3.1 Statistical Hypotheses

We like to think of statistical hypothesis testing as the data analysis stage of a **comparative experiment**, in which the engineer is interested, for example, in comparing the mean of a population to a specified value (e.g. mean height in Catalan population).

A **statistical hypothesis** is a statement about the parameters of one or more populations.

1-3 Hypothesis Testing

1-3.1 Statistical Hypotheses

For example, suppose that we are interested in the burning rate of a solid propellant used to power aircrew escape systems.

- Now burning rate is a random variable that can be described by a probability distribution.
- Suppose that our interest focuses on the **mean** burning rate (a parameter of this distribution).
- Specifically, we are interested in deciding whether or not the mean burning rate is 50 centimeters per second.

1-3 Hypothesis Testing

1-3.1 Statistical Hypotheses

Two-sided Alternative Hypothesis

$$H_0: \mu = 50 \text{ cm/s}$$
$$H_1: \mu \neq 50 \text{ cm/s}$$

One-sided Alternative Hypotheses

$H_0: \mu = 50$ cm/s $\quad H_1: \mu < 50$ cm/s $\quad$ or $\quad H_0: \mu = 50$ cm/s $\quad H_1: \mu > 50$ cm/s

1-3 Hypothesis Testing

1-3.1 Statistical Hypotheses

Test of a Hypothesis

• A procedure leading to a decision about a particular hypothesis

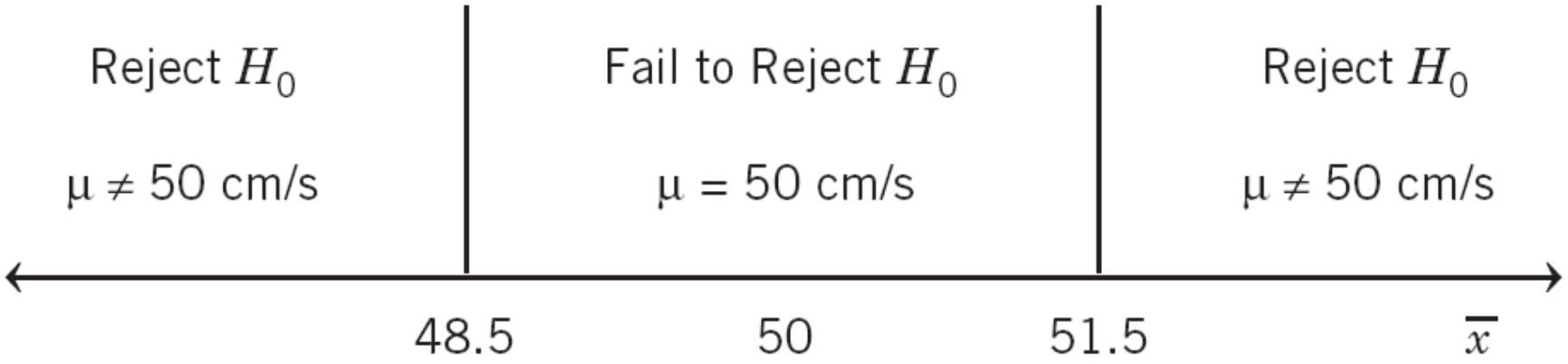
• Hypothesis-testing procedures rely on using the information in a **random sample from the population of interest**.

• If this information is *consistent* with the hypothesis, then we will conclude that the hypothesis can't be rejected (**true)**; if this information is *inconsistent* with the hypothesis, we will conclude that the hypothesis is rejected (**false)**.

1-3 Hypothesis Testing

1-3.2 Testing Statistical Hypotheses

$$H_0: \mu = 50 \text{ cm/s}$$
$$H_1: \mu \neq 50 \text{ cm/s}$$



Decision criteria for testing $H_0: \mu = 50$ cm/s versus $H_1: \mu \neq 50$ cm/s.

1-3 Hypothesis Testing

1-3.2 Testing Statistical Hypotheses

Rejecting the null hypothesis H_0 when it is true is defined as a **type I error.**

Failing to reject the null hypothesis when it is false is defined as a **type II error.**

1-3 Hypothesis Testing

1-3.2 Testing Statistical Hypotheses

Decisions in Hypothesis Testing

Decision	H_0 Is True	H_0 Is False
Fail to reject H_0	No error	Type II error
Reject H_0	Type I error	No error

$$\alpha = P(\text{type I error}) = P(\text{reject } H_0 \text{ when } H_0 \text{ is true})$$

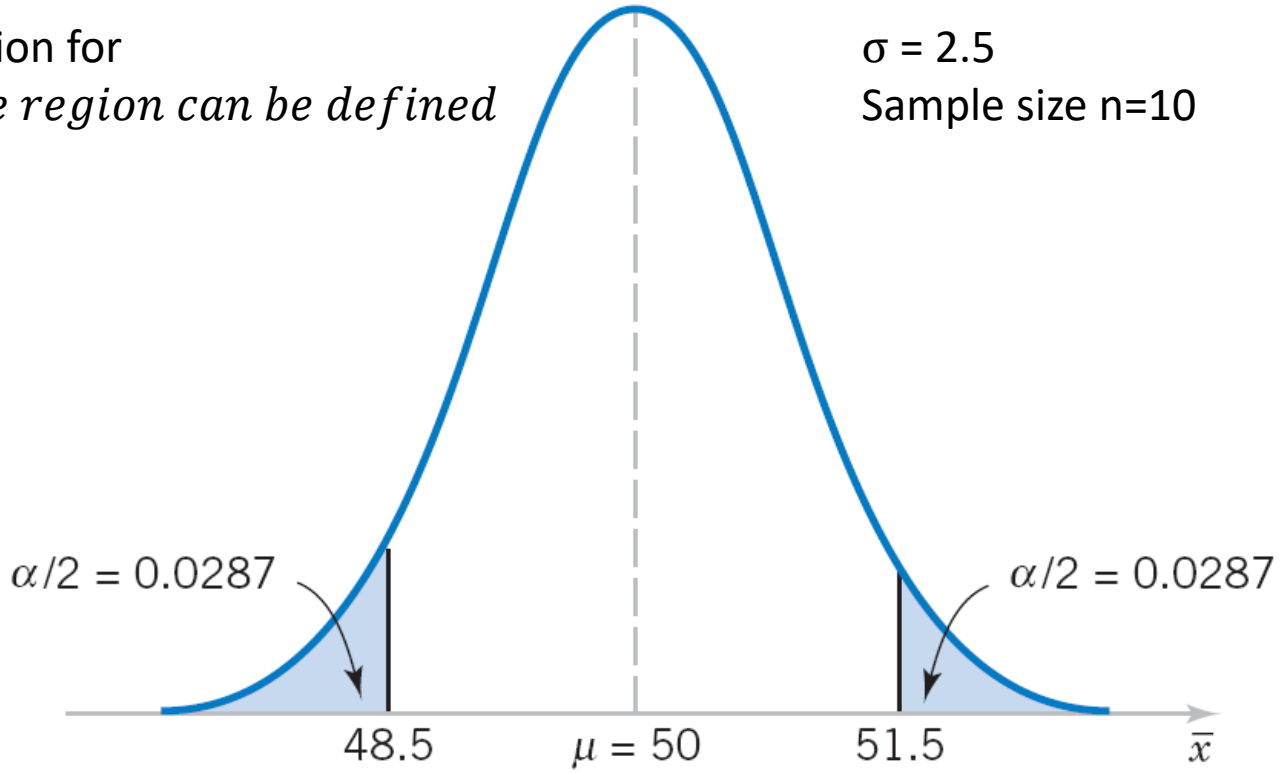
Sometimes the type I error probability is called the **significance level**, or the **α-error**, or the **size** of the test.

1-3 Hypothesis Testing

1-3.2 Testing Statistical Hypotheses

Sampling distribution for
$\bar{X} \rightarrow Confidence\ region\ can\ be\ defined$

σ = 2.5
Sample size n=10

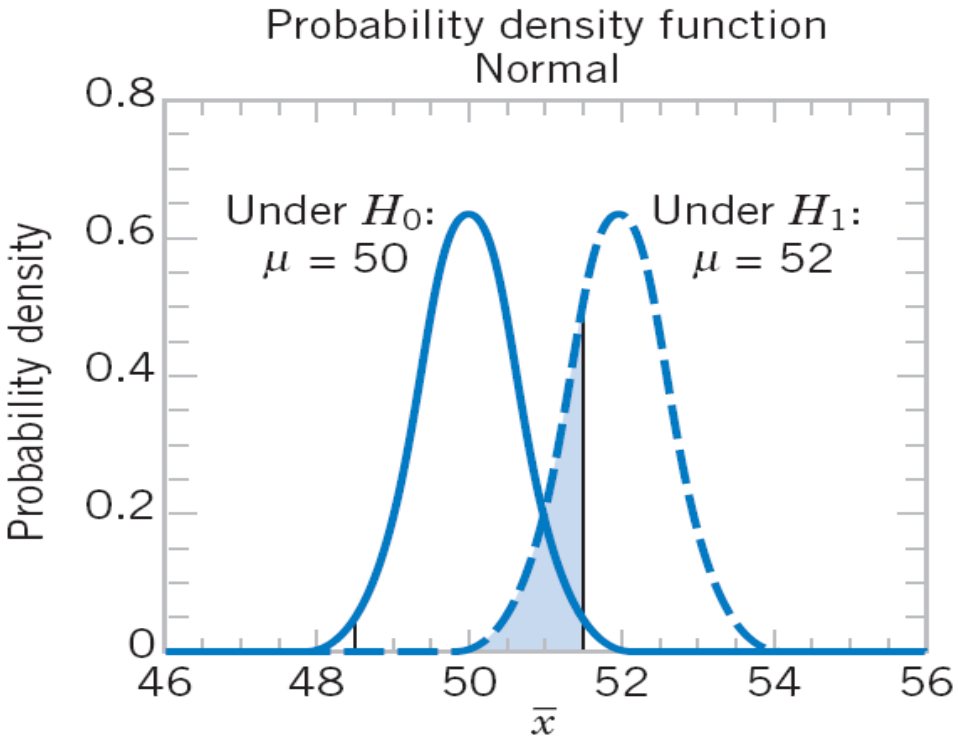


The critical region for H_0: $\mu = 50$ versus H_1: $\mu \neq 50$ and $n = 10$.

1-3 Hypothesis Testing

1-3.2 Testing Statistical Hypotheses

$$\beta = P(\text{type II error}) = P(\text{fail to reject } H_0 \text{ when } H_0 \text{ is false})$$



The probability of type II error when $\mu = 52$ and $n = 10$.

1-3 Hypothesis Testing

1-3.2 Testing Statistical Hypotheses

$\sigma = 2.5$

Fail to Reject H_0 When	Sample Size	α	β at $\mu = 52$	β at $\mu = 50.5$
$48.5 < \bar{x} < 51.5$	10	0.0574	0.2643	0.8923
$48 < \bar{x} < 52$	10	0.0114	0.5000	0.9705
$48.5 < \bar{x} < 51.5$	16	0.0164	0.2119	0.9445
$48 < \bar{x} < 52$	16	0.0014	0.5000	0.9918

1-3 Hypothesis Testing

1-3.2 Testing Statistical Hypotheses

The **power** of a statistical test is the probability of rejecting the null hypothesis H_0 when the alternative hypothesis is true.

- The power is computed as 1 - β, and power can be interpreted as *the probability of correctly rejecting a false null hypothesis*. We often compare statistical tests by comparing their **power** properties.

- For example, consider the propellant burning rate problem when we are testing H_0 : μ = 50 centimeters per second against H_1 : μ not equal 50 cm per second . Suppose that the true value of the mean is μ = 52. When n = 10, we found that β = 0.2643, so the power of this test is 1 - β = 1 - 0.2643 = 0.7357 when μ = 52.

1-3 Hypothesis Testing

1-3.3 One-Sided and Two-Sided Hypotheses

Two-Sided Test:

$$H_0: \mu = \mu_0$$
$$H_1: \mu \neq \mu_0$$

One-Sided Tests:

$$H_0: \mu = \mu_0 \quad H_1: \mu > \mu_0 \qquad \text{or} \qquad H_0: \mu = \mu_0 \quad H_1: \mu < \mu_0$$

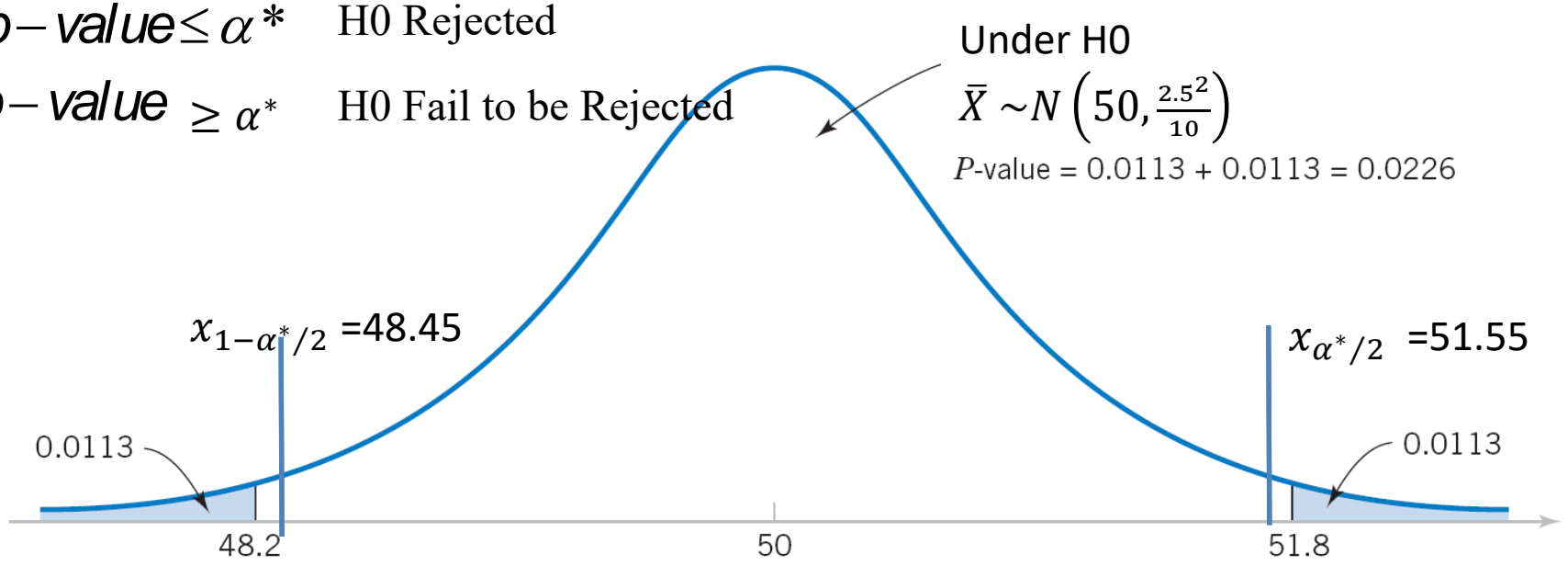
1-3 Hypothesis Testing

1-3.3 P-Values in Hypothesis Testing

α^* Significance level - Commonly α^*=0.05

$p-value \leq \alpha^*$ H0 Rejected

$p-value \geq \alpha^*$ H0 Fail to be Rejected



Calculating the *P*-value for the propellant burning rate problem.

1-3 Hypothesis Testing

1-3.5 General Procedure for Hypothesis Testing

1. **Parameter of interest:** From the problem context, identify the parameter of interest.
2. **Null hypothesis, H_0:** State the null hypothesis, H_0.
3. **Alternative hypothesis, H_1:** Specify an appropriate alternative hypothesis, H_1.
4. **Test statistic:** State an appropriate test statistic.
5. **Reject H_0 if:** Define the criteria that will lead to rejection of H_0.
6. **Computations:** Compute any necessary sample quantities, substitute these into the equation for the test statistic, and compute that value.
7. **Conclusions:** Decide whether or not H_0 should be rejected and report that in the problem context. This could involve computing a P-value or comparing the test statistic to a set of critical values.

Steps 1–4 should be completed prior to examination of the sample data.

1-4 Inference on the Mean of a Population, Variance Known

Assumptions

1. $X_1, X_2, \ldots, X_n$ is a random sample of size n from a population.
2. The population is normally distributed, or if it is not, the conditions of the central limit theorem apply.

Under the previous assumptions, the quantity

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

has a standard normal distribution, $N(0, 1)$.

1-4 Inference on the Mean of a Population, Variance Known

1-4.1 Hypothesis Testing on the Mean

We wish to test:

$$H_0: \mu = \mu_0$$
$$H_1: \mu \neq \mu_0$$

The **test statistic** is:

$$Z_0 = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

1-4 Inference on the Mean of a Population, Variance Known

1-4.1 Hypothesis Testing on the Mean (two-sided)

$Fixed\ \alpha\ (significance\ level)$: $z_{\alpha/2}$ s.t. P(Z > $z_{\alpha/2}$) = $\alpha/2$ Z ~$N(0,1)$

Reject H_0 if the observed value of the test statistic z_0 is either:

$$z_0 > z_{\alpha/2} \qquad z_0 < -z_{\alpha/2}$$

Fail to reject H_0 if

$$-z_{\alpha/2} \leq z_0 \leq z_{\alpha/2}$$

1-4 Inference on the Mean of a Population, Variance Known

1-4.1 Hypothesis Testing on the Mean

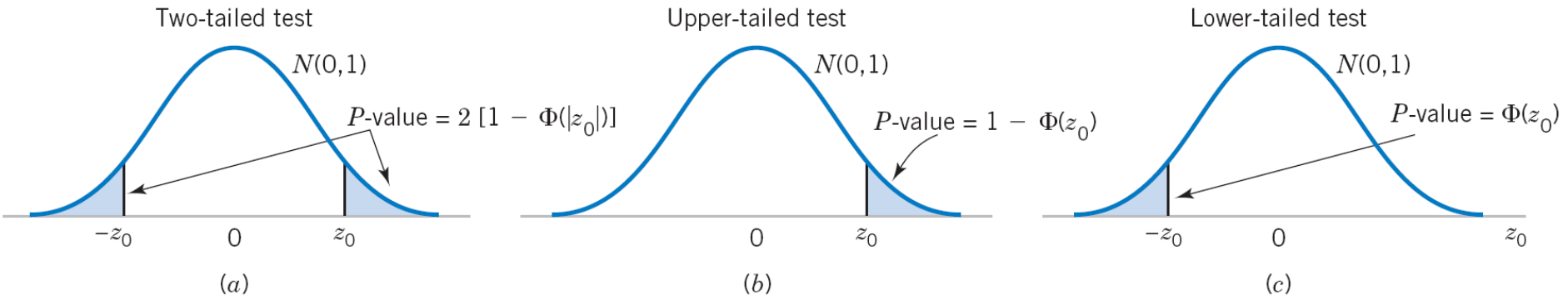
Two-sided

$P - value = 2\mathrm{P}(\mathrm{Z} > z_0)$

One-sided

$P - value = \mathrm{P}(\mathrm{Z} > z_0)$

$P - value = \mathrm{P}(\mathrm{Z} < -z_0)$



The P-value for a z-test. (a) The two-sided alternative H_1: $\mu \neq \mu_0$. (b) The one-sided alternative H_1: $\mu > \mu_0$. (c) The one-sided alternative H_1: $\mu < \mu_0$.

$z_{\alpha/2}$ s.t. $\mathrm{P}(\mathrm{Z} > z_{\alpha/2}) = \alpha/2$
$\mathrm{Z} \sim N(0,1)$

z_{α} s.t. $\mathrm{P}(\mathrm{Z} > z_{\alpha}) = \alpha$
$\mathrm{Z} \sim N(0,1)$

z_{α} s.t. $\mathrm{P}(\mathrm{Z} < -z_{\alpha}) = \alpha$
$\mathrm{Z} \sim N(0,1)$

1-4 Inference on the Mean of a Population, Variance Known

1-4.1 Hypothesis Testing on the Mean

Testing Hypotheses on the Mean, Variance Known

Null hypothesis: $H_0: \mu = \mu_0$

Test statistic: $$Z_0 = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

Alternative Hypotheses	*P*-Value	Rejection Criterion for Fixed-Level Tests
$H_1: \mu \neq \mu_0$	Probability above z_0 and probability below $-z_0$, $P = 2[1-\Phi(\lvert z_0 \rvert)]$	$z_0 > z_{\alpha/2}$ or $z_0 < -z_{\alpha/2}$
$H_1: \mu > \mu_0$	Probability above z_0, $P = 1-\Phi(z_0)$	$z_0 > z_\alpha$
$H_1: \mu < \mu_0$	Probability below z_0, $P = \Phi(z_0)$	$z_0 < -z_\alpha$

1-4 Inference on the Mean of a Population, Variance Known

1-4.4 Some Practical Comments on Hypothesis Testing

Statistical versus Practical Significance

H0: $\mu=50$
H1: $\mu \neq 50$

$\sigma = 2$

Sample Size n	P-Value When $\bar{x} = 50.5$	Power (at $\alpha = 0.05$) When $\mu = 50.5$
10	0.4295	0.1241
25	0.2113	0.2396
50	0.0767	0.4239
100	0.0124	0.7054
400	5.73×10^{-7}	0.9988
1000	2.57×10^{-15}	1.0000

1-4 Inference on the Mean of a Population, Variance Known

1-4.4 Some Practical Comments on Hypothesis Testing

Statistical versus Practical Significance

Be careful when interpreting the results from hypothesis testing when the sample size is large because any small departure from the hypothesized value μ_0 will probably be detected, even when the difference is of little or no practical significance.

1-4 Inference on the Mean of a Population, Variance Known

1-4.5 Confidence Interval on the Mean

Two-sided confidence interval:

$$P(L \leq \mu \leq U) = 1 - \alpha$$

One-sided confidence intervals:

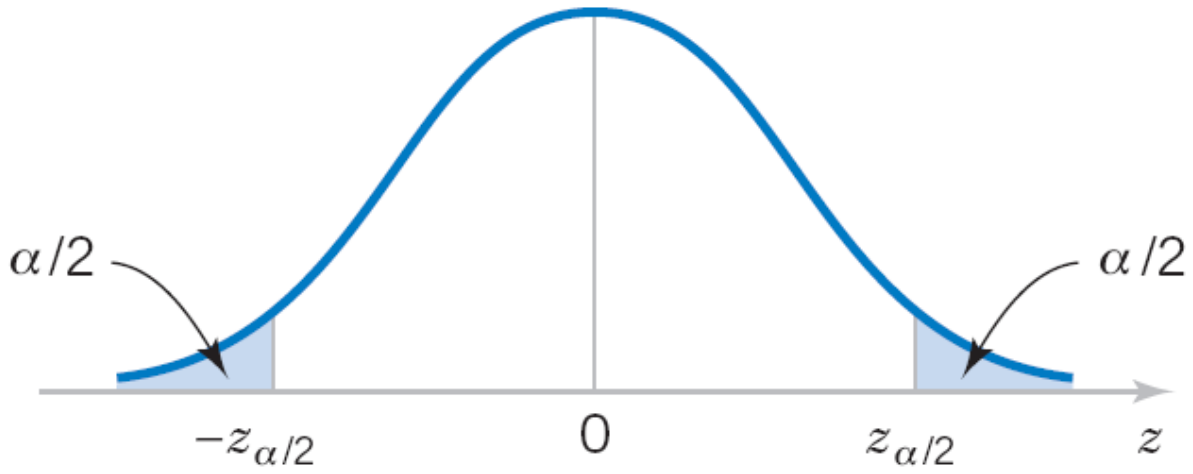
$$P(L \leq \mu) = 1 - \alpha \qquad P(\mu \leq U) = 1 - \alpha$$

Confidence coefficient: $1 - \alpha$

1-4 Inference on the Mean of a Population, Variance Known

1-4.6 Confidence Interval on the Mean

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \qquad P\{-z_{\alpha/2} \leq Z \leq z_{\alpha/2}\} = 1 - \alpha$$



The distribution of Z.

1-4 Inference on the Mean of a Population, Variance Known

1-4.5 Confidence Interval on the Mean

Confidence Interval on the Mean, Variance Known

If $\bar{x}$ is the sample mean of a random sample of size n from a population with known variance σ^2, a $100(1-\alpha)\%$ **confidence interval on μ** is given by

$$\bar{x} - \frac{z_{\alpha/2}\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + \frac{z_{\alpha/2}\sigma}{\sqrt{n}}$$

where $z_{\alpha/2}$ is the upper $100\alpha/2$ percentage point and $-z_{\alpha/2}$ is the lower $100\alpha/2$ percentage point of the standard normal distribution in Appendix A Table I.

1-4 Inference on the Mean of a Population, Variance Known

1-4.5 Confidence Interval on the Mean

Relationship between Tests of Hypotheses and Confidence Intervals

If $[l,u]$ is a $100(1 - \alpha)$ percent confidence interval for the parameter, then the test of significance level α of the hypothesis

$$H_0: \theta = \theta_0$$

$$H_1: \theta \neq \theta_0$$

will lead to rejection of H_0 if and only if the hypothesized value is not in the $100(1 - \alpha)$ percent confidence interval $[l, u]$.

1-4 Inference on the Mean of a Population, Variance Known

1-4.5 Confidence Interval on the Mean

Confidence Level and Precision of Estimation

The length of the two-sided 95% confidence interval is

$$2(1.96\,\sigma/\sqrt{n}) = 3.92\,\sigma/\sqrt{n}$$

whereas the length of the two-sided 99% confidence interval is

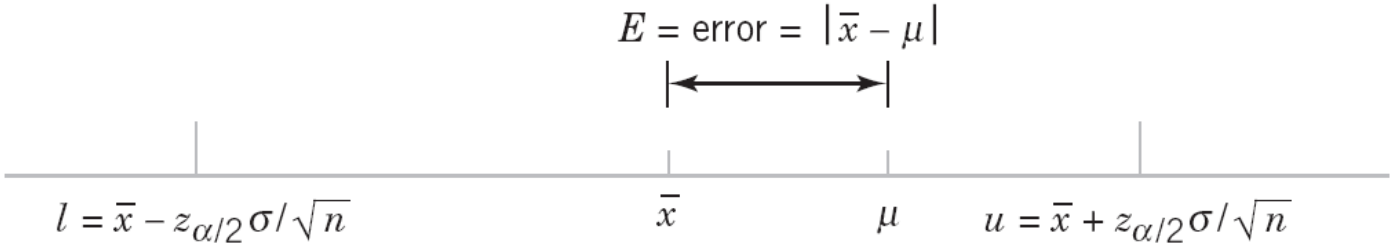
$$2(2.58\,\sigma/\sqrt{n}) = 5.16\,\sigma/\sqrt{n}$$

1-4 Inference on the Mean of a Population, Variance Known

1-4.5 Confidence Interval on the Mean

Choice of Sample Size

Error in estimating μ with $\bar{x}$.



Sample Size for a Specified E on the Mean, Variance Known

If $\bar{x}$ is used as an estimate of μ, we can be $100(1-\alpha)\%$ confident that the error $|\bar{x}-\mu|$ will not exceed a specified amount E when the sample size is

$$n = \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2$$

1-4 Inference on the Mean of a Population, Variance Known

1-4.5 Confidence Interval on the Mean

Choice of Sample Size

Note the general relationship between sample size, desired length of the confidence interval $2E$, confidence level $100(1-\alpha)\%$, and standard deviation σ:

- As the desired length of the interval $2E$ decreases, the required sample size n increases for a fixed value of σ and specified confidence.
- As σ increases, the required sample size n increases for a fixed desired length $2E$ and specified confidence.
- As the level of confidence increases, the required sample size n increases for fixed desired length $2E$ and standard deviation σ.

$$n = \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2$$

1-4 Inference on the Mean of a Population, Variance Known

1-4.5 Confidence Interval on the Mean

One-Sided Confidence Bounds

One-Sided Confidence Bounds on the Mean, Variance Known

The $100(1-\alpha)\%$ **upper-confidence bound** for μ is

$$\mu \leq u = \bar{x} + z_\alpha \sigma/\sqrt{n}$$

and the $100(1-\alpha)\%$ **lower-confidence bound** for μ is

$$\bar{x} - z_\alpha \sigma/\sqrt{n} = l \leq \mu$$

1-5 Inference on the Mean of a Population, Variance Unknown

1-5.1 Hypothesis Testing on the Mean

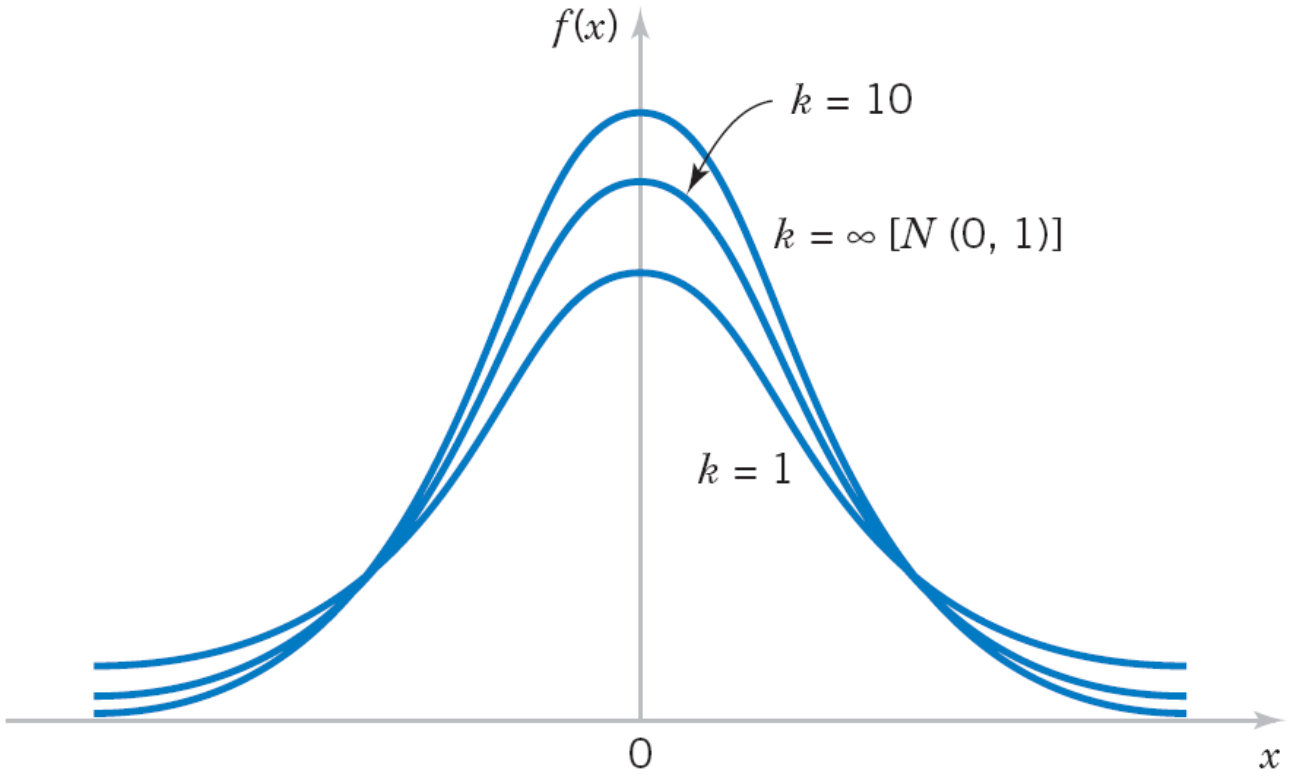
Let $X_1, X_2, \ldots, X_n$ be a random sample for a normal distribution with unknown mean μ and unknown variance σ^2. The quantity

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

has a t distribution with $n - 1$ degrees of freedom.

1-5 Inference on the Mean of a Population, Variance Unknown

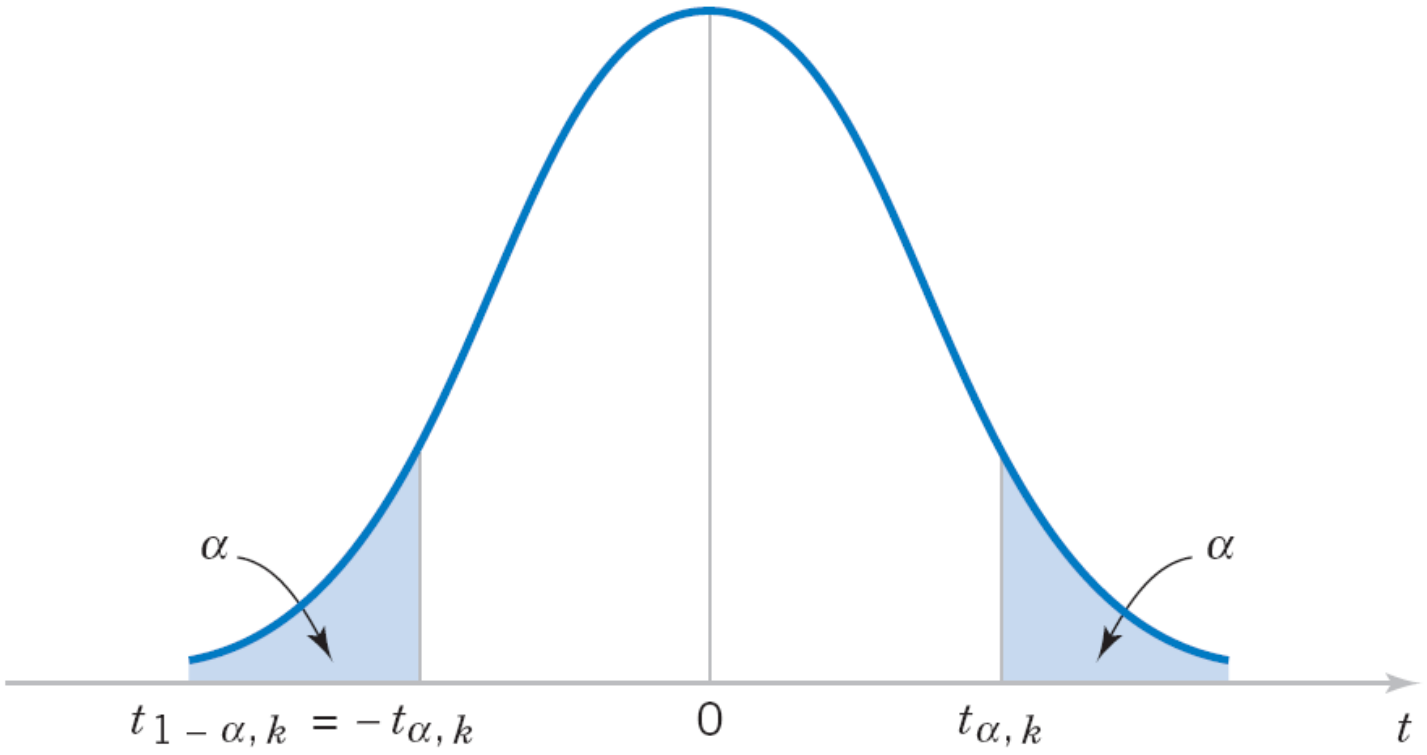
1-5.1 Hypothesis Testing on the Mean



Probability density functions of several t distributions.

1-5 Inference on the Mean of a Population, Variance Unknown

1-5.1 Hypothesis Testing on the Mean

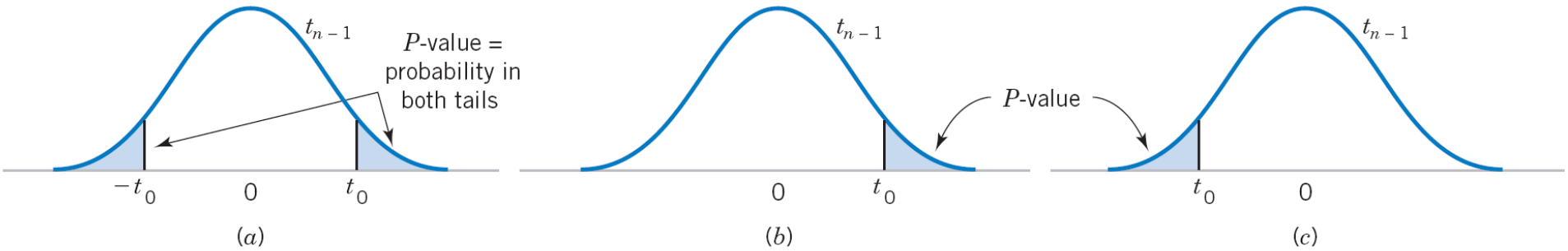


Percentage points of the t distribution.

1-5 Inference on the Mean of a Population, Variance Unknown

1-5.1 Hypothesis Testing on the Mean

Calculating the P-value



Calculating the P-value for a t-test: (a) H_1: $\mu \neq \mu_0$; (b) H_1: $\mu > \mu_0$; (c) H_1: $\mu < \mu_0$.

1-5 Inference on the Mean of a Population, Variance Unknown

1-5.1 Hypothesis Testing on the Mean

Testing Hypotheses on the Mean of a Normal Distribution, Variance Unknown

Null hypothesis: $H_0: \mu = \mu_0$

Test statistic: $T_0 = \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}}$

Alternative Hypotheses	*P*-Value	Rejection Criterion for Fixed-Level Tests
$H_1: \mu \neq \mu_0$	Sum of the probability above t_0 and the probability below $-t_0$	$t_0 > t_{\alpha/2,n-1}$ or $t_0 < -t_{\alpha/2,n-1}$
$H_1: \mu > \mu_0$	Probability above t_0	$t_0 > t_{\alpha,n-1}$
$H_1: \mu < \mu_0$	Probability below t_0	$t_0 < -t_{\alpha,n-1}$

The locations of the critical regions for these situations are shown in Fig. 4-19*a*, *b*, and *c*, respectively.

1-5 Inference on the Mean of a Population, Variance Unknown

1-5.3 Confidence Interval on the Mean (two-sided)

Confidence Interval on the Mean of a Normal Distribution, Variance Unknown

If $\bar{x}$ and s are the mean and standard deviation of a random sample from a normal distribution with unknown variance σ^2, a $100(1-\alpha)\%$ CI on μ is given by

$$\bar{x} - t_{\alpha/2,n-1}s/\sqrt{n} \leq \mu \leq \bar{x} + t_{\alpha/2,n-1}s/\sqrt{n}$$

where $t_{\alpha/2,n-1}$ is the upper $100\alpha/2$ percentage point of the t distribution with $n-1$ degrees of freedom.

1-5 Inference on the Mean of a Population, Variance Unknown

1-5.3 Confidence Interval on the Mean

Golf Clubs

Reconsider the golf club coefficient of restitution problem in Example 4-7. We know that $n = 15$, $\bar{x} = 0.83725$, and $s = 0.02456$. Find a 95% CI on μ.

Solution. From equation 4-50 we find ($t_{\alpha/2,n-1} = t_{0.025,14} = 2.145$):

$$\bar{x} - t_{\alpha/2,n-1}s/\sqrt{n} \leq \mu \leq \bar{x} + t_{\alpha/2,n-1}s/\sqrt{n}$$

$$0.83725 - 2.145(0.02456)/\sqrt{15} \leq \mu \leq 0.83725 + 2.145(0.02456)/\sqrt{15}$$

$$0.83725 - 0.01360 \leq \mu \leq 0.83725 + 0.01360$$

$$0.82365 \leq \mu \leq 0.85085$$

In R

- If I have the data:

```
> x = c(2100, 2180, 2200, 2030, 2186, 2203, 2080, 2111)
> t.test(x)

        One Sample t-test data:  x
t = 93.371, df = 7, p-value = 4.258e-12
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 2082.149 2190.351
sample estimates:
mean of x
  2136.25
```

CI for Mean of Paired Differences

- Recall the notation: $\mu_d = \mu_1 - \mu_2$ $\quad \bar{d} = \bar{x}_1 - \bar{x}_2$ $\quad se(\bar{d}) = s_d / \sqrt{n}$
- If the standard deviation σ_d **is known**

$$\text{CI} = \bar{d} \pm z_{\alpha/2} \times \frac{\sigma_d}{\sqrt{n}}$$

- If the standard deviation σ_d **is unknown**

$$\text{CI} = \bar{d} \pm t_{\alpha/2,n-1} \times \frac{s_d}{\sqrt{n}}$$

Conditions: $n \geq 30$ or bell-shaped distribution

Example

- A researcher measures the weight of 5 students, and after a week of diet, he measures again. The goal is to see how much weight they lost by being in the diet. He finds

student	Day 1	Day 7	Difference(d)
a	75	74	1
b	63	64	-1
c	88	83	5
d	79	77	2
e	67	64	3

$$s_d = \sqrt{\frac{\sum_{i=1}^{n}(d_i - \bar{d})^2}{n-1}} = \sqrt{\frac{(1-2)^2}{5-1} + \frac{(-1-2)^2}{5-1} + \ldots + \frac{(3-2)^2}{5-1}} = 2.23$$

Fine or wrong ?
CI? 98%

$$CI = \bar{d} \pm z_{\alpha/2} \times \frac{\sigma_d}{\sqrt{n}} = 2 \pm 2.36 \times \frac{2.23}{\sqrt{5}} = (-0.35, 4.35)$$

Example

- A researcher measures the weight of 5 students, and after a week of diet, he measures again. The goal is to see how much weight they lost by being in the diet. He finds

student	Day 1	Day 7	Difference(d)
a	75	74	1
b	63	64	-1
c	88	83	5
d	79	77	2
e	67	64	3

$$CI = \bar{d} \pm t_{\alpha/2,n-1} \times \frac{s_d}{\sqrt{n}} = 2 \pm 2.78 \times \frac{2.23}{\sqrt{5}} = (-0.78, 4.77)$$

OK, 95% CI

In R:

t.test(x = c(75, 63, 88, 79, 67), y = c(74, 64, 83, 77, 64), paired = TRUE)

CI for Difference in Two Sample Means

- CI for $\mu_1 - \mu_2$

$$CI = \bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2,df} \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

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

> t.test(x = c(75, 63, 88, 79, 67), y = c(74, 64, 83, 77, 64, 64)) # paired and var.equal FALSE as default

CI for Difference in Two Sample Means – Pooled Std.Deviation

- If we assume that the standard deviation of the two populations is the same, then we can use the pooled standard error to estimate it:

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

- And then the CI becomes:

$$CI = \bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2,df} \times s_p \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

- Note that $df = n_1 + n_2 - 2$

CI for Difference in Two Sample Means

> t.test(x = c(75, 63, 88, 79, 67), y = c(74, 64, 83, 77, 64, 64)) # paired and var.equal FALSE as default

```
        Welch Two Sample t-test
data:  c(75, 63, 88, 79, 67) and c(74, 64, 83, 77, 64, 64)
t = 0.61304, df = 7.8367, p-value = 0.5572
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -9.436022 16.236022
sample estimates:
mean of x mean of y
    74.4      71.0
```

> t.test(x = c(75, 63, 88, 79, 67), y = c(74, 64, 83, 77, 64, 64), var.equal=TRUE) # paired FALSE as default, now a pooled variance – Classical Two Sample t-test

```
        Two Sample t-test
data:  c(75, 63, 88, 79, 67) and c(74, 64, 83, 77, 64, 64)
t = 0.62465, df = 9, p-value = 0.5477
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -8.913008 15.713008
sample estimates:
mean of x mean of y
    74.4      71.0
```

CI for a Population Mean

- If the standard deviation σ **is known**

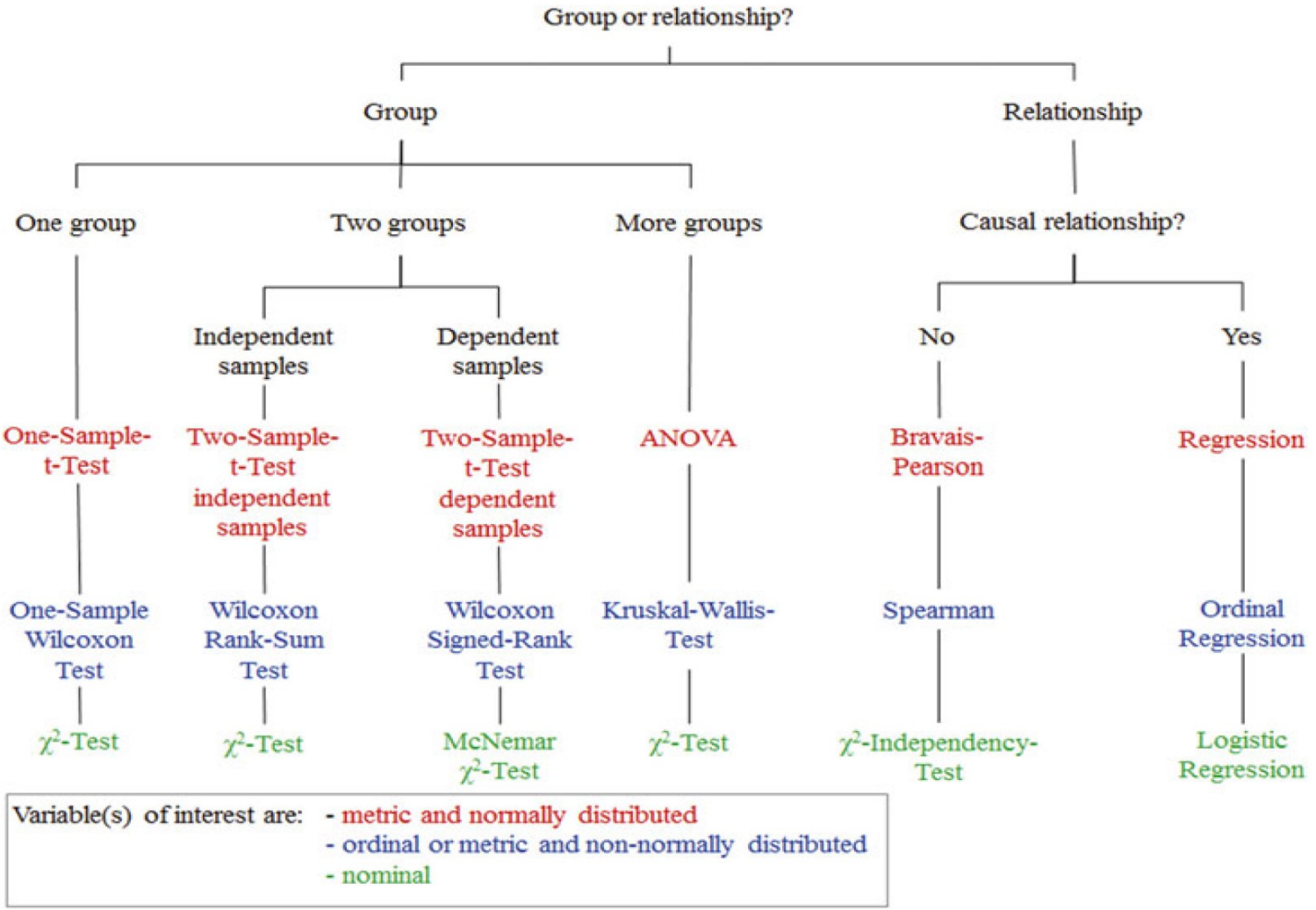
$$\text{CI for } \mu \text{ is } \bar{x} \pm z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$$

- If the standard deviation σ **is unknown**

$$\text{CI for } \mu \text{ is } \bar{x} \pm t_{\alpha/2,n-1} \times \frac{s}{\sqrt{n}} \quad \text{where} \quad s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

Conditions: $n \geq 30$ or bell-shaped distribution

Guide to test selection



F. Kronthaler and S. Zöllner, Data Analysis with RStudio, https://doi.org/10.1007/978-3-662-62518-7_6

1-6 Inference on the Variance of a Normal Population

1-6.1 Hypothesis Testing on the Variance of a Normal Population

Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal distribution with unknown mean μ and unknown variance σ^2. The quantity

$$X^2 = \frac{(n-1)S^2}{\sigma^2}$$

has a chi-square distribution with $n-1$ degrees of freedom, abbreviated as χ^2_{n-1}. In general, the probability density function of a chi-square random variable is

$$f(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{(k/2)-1} e^{-x/2} \qquad x > 0$$

where k is the number of degrees of freedom and $\Gamma(k/2)$ is generalized gamma function

1-6 Inference on the Variance of a Normal Population

1-6.1 Hypothesis Testing on the Variance of a Normal Population

$$H_0: \sigma^2 = \sigma_0^2$$
$$H_1: \sigma^2 \neq \sigma_0^2$$

$$X_0^2 = \frac{(n-1)S^2}{\sigma_0^2}$$

1-6 Inference on the Variance of a Normal Population

1-6.1 Hypothesis Testing on the Variance of a Normal Population in R

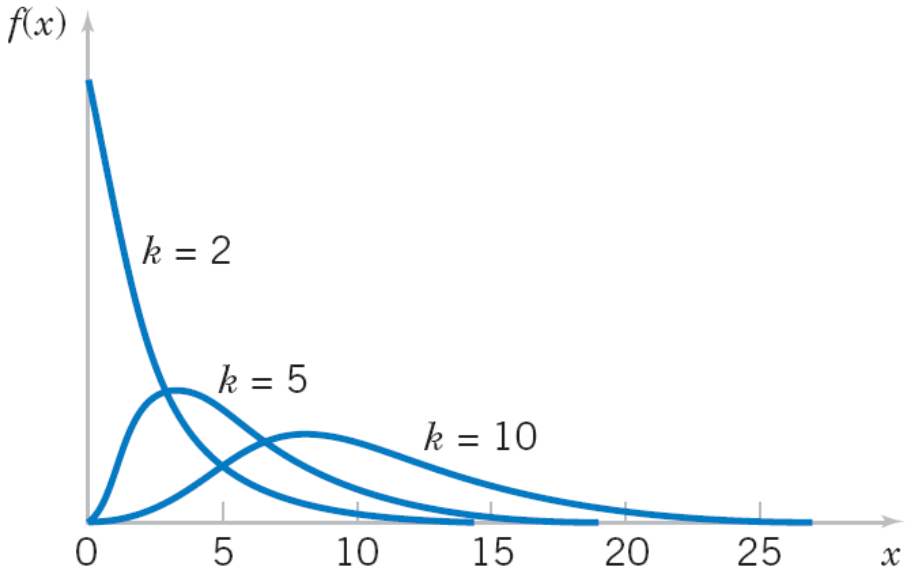
```
> x = c(2100, 2180, 2200, 2030, 2186, 2203, 2080, 2111)
> library(EnvStats)
> varTest(x)

        Chi-Squared Test on Variance

data:  x
Chi-Squared = 29314, df = 7, p-value < 2.2e-16
alternative hypothesis: true variance is not equal to 1
95 percent confidence interval:
  1830.633 17346.609
sample estimates:
variance
4187.643
```

1-6 Inference on the Variance of a Normal Population

1-6.1 Hypothesis Testing on the Variance of a Normal Population

The mean and variance of the χ^2 distribution are

$$\mu = k \quad \text{and} \quad \sigma^2 = 2k$$

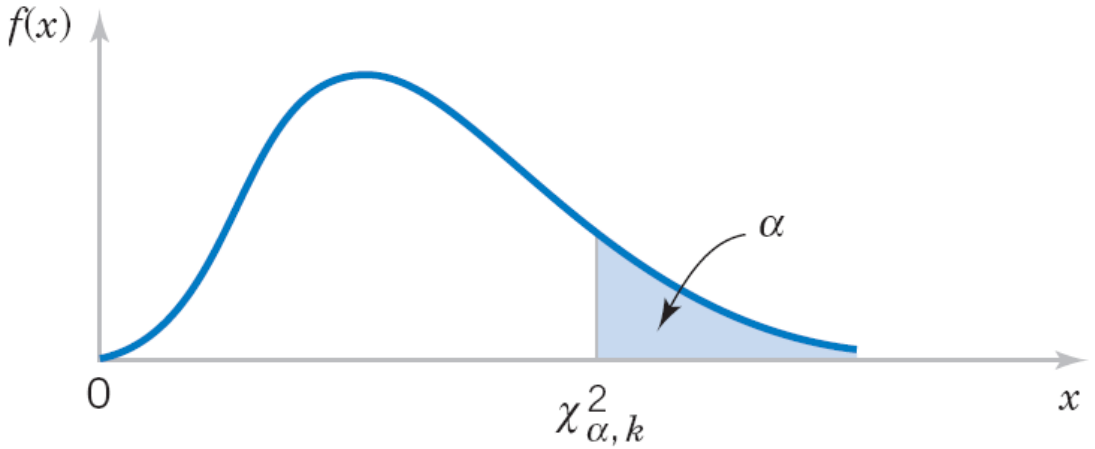


Probability density functions of several χ^2 distributions.

1-6 Inference on the Variance of a Normal Population

1-6.1 Hypothesis Testing on the Variance of a Normal Population

$$P(X^2 > \chi^2_{\alpha,k}) = \int_{\chi^2_{\alpha,k}}^{\infty} f(u)\,du = \alpha$$



Percentage point $\chi^2_{\alpha,k}$ of the χ^2 distribution.

1-6 Inference on the Variance of a Normal Population

1-6.1 Hypothesis Testing on the Variance of a Normal Population

Testing Hypotheses on the Variance of a Normal Distribution

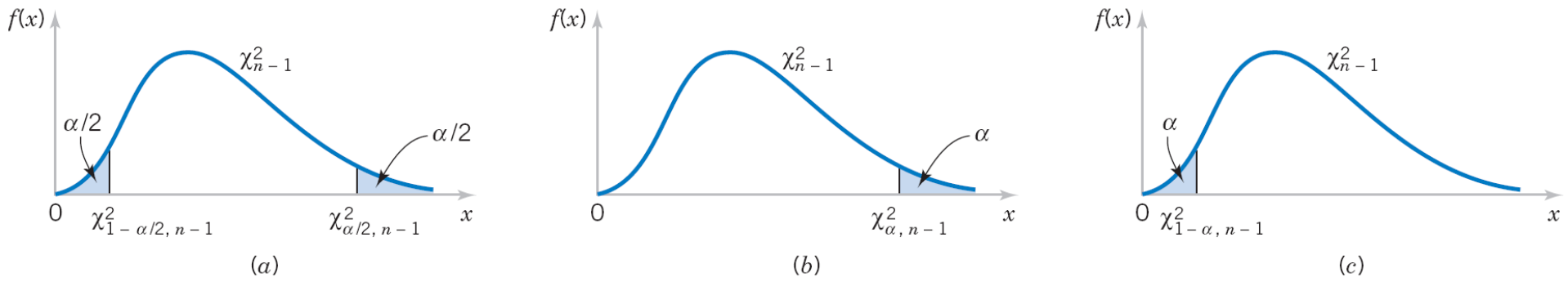
Null hypothesis: $H_0: \sigma^2 = \sigma_0^2$

Test statistic: $\chi_0^2 = \dfrac{(n-1)S^2}{\sigma_0^2}$

Alternative Hypotheses	Rejection Criterion
$H_1: \sigma^2 \neq \sigma_0^2$	$\chi_0^2 > \chi_{\alpha/2,n-1}^2$ or $\chi_0^2 < \chi_{1-\alpha/2,n-1}^2$
$H_1: \sigma^2 > \sigma_0^2$	$\chi_0^2 > \chi_{\alpha,n-1}^2$
$H_1: \sigma^2 < \sigma_0^2$	$\chi_0^2 < \chi_{1-\alpha,n-1}^2$

1-6 Inference on the Variance of a Normal Population

1-6.1 Hypothesis Testing on the Variance of a Normal Population



Distribution of the test statistic for H_0: $\sigma^2 = \sigma_0^2$ with critical region values for (a) H_1: $\sigma^2 \neq \sigma_0^2$, (b) H_0: $\sigma^2 > \sigma_0^2$, and (c) H_0: $\sigma^2 < \sigma_0^2$.

1-6 Inference on the Variance of a Normal Population

1-6.2 Confidence Interval on the Variance of a Normal Population

Confidence Interval on the Variance of a Normal Distribution

If s^2 is the sample variance from a random sample of n observations from a normal distribution with unknown variance σ^2, a $100(1-\alpha)\%$ CI on σ^2 is

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}$$

where $\chi^2_{\alpha/2,n-1}$ and $\chi^2_{1-\alpha/2,n-1}$ are the upper and lower $100\alpha/2$ percentage points of the chi-square distribution with $n-1$ degrees of freedom, respectively.

1-6 Inference on the Variance of Normal Populations

1-6.3 Confidence Interval on the Variance of two Normal Populations

$$H0\colon \sigma_X^2 = \sigma_Y^2 \qquad H1\colon \sigma_X^2 \neq \sigma_Y^2$$

- Two independent samples of size n_X and n_Y with sample variances s_X^2 and s_Y^2 are available.
- Then $\frac{s_X^2/\sigma_X^2}{s_Y^2/\sigma_Y^2}$ is F distributed df1=$(n_X - 1)$ and df2=$(n_Y - 1)$

1-6 Inference on the Variance of Normal Populations

1-6.3 Confidence Interval on the Variance of two Normal Populations

$$H0\colon \sigma_X^2 = \sigma_Y^2 \qquad H1\colon \sigma_X^2 \neq \sigma_Y^2$$

- $f_0 = \frac{s_X^2}{s_Y^2}$ is F distributed $\mathrm{df1} = (n_X - 1)$ and $\mathrm{df2} = (n_Y - 1)$.
- Two-sided. Rejection criteria:

$f_0 > F^{-1}(1 - \alpha/2; n_X - 1; n_Y - 1)$ or $f_0 < F^{-1}(\alpha/2; n_X - 1; n_Y - 1)$ where $F^{-1}()$ is the inverse of the distribution function (so percentiles)

- One-sided. Rejection criteria: $H1\colon \sigma_X^2 > \sigma_Y^2$

$$f_0 > F^{-1}(1 - \alpha; n_X - 1; n_Y - 1)$$

1-7 Inference on Population Proportion

1-7.1 Hypothesis Testing on a Binomial Proportion

We will consider testing:

$$H_0: p = p_0$$
$$H_1: p \neq p_0$$

Let X be the number of observations in a random sample of size n that belongs to the class associated with p. Then the quantity

$$Z = \frac{X - np}{\sqrt{np(1-p)}}$$

has approximately a standard normal distribution, $N(0, 1)$.

1-7 Inference on Population Proportion

1-7.1 Hypothesis Testing on a Binomial Proportion

Testing Hypotheses on a Binomial Proportion

Null hypotheses: $H_0: p = p_0$

Test statistic: $$Z_0 = \frac{X - np_0}{\sqrt{np_0(1 - p_0)}}$$

$P(Z \leq z) = \Phi(z)$ is the distribution function of the standard normal variable

Alternative Hypotheses	*P*-Value	Rejection Criterion for Fixed-Level Tests
$H_1: p \neq p_0$	Probability above z_0 and probability below $-z_0$, $P = 2[1 - \Phi(\lvert z_0 \rvert)]$	$z_0 > z_{\alpha/2}$ or $z_0 < -z_{\alpha/2}$
$H_1: p > p_0$	Probability above z_0, $P = 1 - \Phi(z_0)$	$z_0 > z_\alpha$
$H_1: p < p_0$	Probability below z_0, $P = \Phi(z_0)$	$z_0 < -z_\alpha$

1-7 Inference on Population Proportion

1-7.1 Hypothesis Testing on a Binomial Proportion

Engine Controllers

A semiconductor manufacturer produces controllers used in automobile engine applications. The customer requires that the process fallout or fraction defective at a critical manufacturing step not exceed 0.05 and that the manufacturer demonstrate process capability at this level of quality using $\alpha = 0.05$. The semiconductor manufacturer takes a random sample of 200 devices and finds that 4 of them are defective. Can the manufacturer demonstrate process capability for the customer?

Solution. We may solve this problem using the seven-step hypothesis testing procedure as follows:

1. **Parameter of interest:** The parameter of interest is the process fraction defective p.
2. **Null hypothesis, H_0:** $p = 0.05$
3. **Alternative hypothesis, H_1:** $p < 0.05$
   This formulation of the problem will allow the manufacturer to make a strong claim about process capability if the null hypothesis H_0: $p = 0.05$ is rejected.

1-7 Inference on Population Proportion

1-7.1 Hypothesis Testing on a Binomial Proportion

4. **Test statistic:** The test statistic is (from equation 4-64)

$$z_0 = \frac{x - np_0}{\sqrt{np_0(1 - p_0)}}$$

where $x = 4$, $n = 200$, and $p_0 = 0.05$.

5. **Reject H_0 if:** Reject H_0: $p = 0.05$ if the P-value is less than 0.05.

6. **Computations:** The test statistic is

$$z_0 = \frac{4 - 200(0.05)}{\sqrt{200(0.05)(0.95)}} = -1.95$$

7. **Conclusions:** Because $z_0 = -1.95$, the P-value is $\Phi(-1.95) = 0.0256$; since this is less than 0.05, we reject H_0 and conclude that the process fraction defective p is less than 0.05. We conclude that the process is capable.

1-7 Inference on Population Proportion

1-7.1 Hypothesis Testing on a Binomial Proportion – Exact Test

```
> #perform two-tailed Binomial test
> binom.test(x=4, n=200, p=0.05, alternative = "two.sided")
```

```
        Exact binomial test

data:  4 and 200
number of successes = 4, number of trials = 200, p-value = 0.05025
alternative hypothesis: true probability of success is not equal to 0.05
95 percent confidence interval:
 0.005475566 0.050413609
sample estimates:
probability of success
                  0.02
```

1-7 Inference on Population Proportion

1-7.1 Hypothesis Testing on a Binomial Proportion– Exact Test

```
> #perform one-tailed Binomial test
> binom.test(x=4, n=200, p=0.05, alternative = "less")

        Exact binomial test

data:  4 and 200
number of successes = 4, number of trials = 200, p-value = 0.02645
alternative hypothesis: true probability of success is less than 0.05
95 percent confidence interval:
 0.00000000 0.04518041
sample estimates:
probability of success
                  0.02
```

1-7 Inference on Population Proportion

1-7.1 Hypothesis Testing on a Binomial Proportion – Exact Test

```
> #perform one-tailed Binomial test
> binom.test(x=4, n=200, p=0.05, alternative = "greater")
```

```
        Exact binomial test

data:  4 and 200
number of successes = 4, number of trials = 200, p-value = 0.991
alternative hypothesis: true probability of success is greater than 0.05
95 percent confidence interval:
 0.006859719 1.000000000
sample estimates:
probability of success
                  0.02
```

1-7 Inference on Population Proportion

1-7.1 Hypothesis Testing on a Binomial Proportion

```
> #perform one-tailed Binomial test
> prop.test(x=4, n=200, p=0.05, alternative = “two.sided")
```

```
        1-sample proportions test with continuity correction

data:  4 out of 200, null probability 0.05
X-squared = 3.1842, df = 1, p-value = 0.07435
alternative hypothesis: true p is not equal to 0.05
95 percent confidence interval:
 0.006426013 0.053757461
sample estimates:
   p
0.02
```

1-7 Inference on Population Proportion

1-7.1 Hypothesis Testing on a Binomial Proportion

```
> #perform one-tailed Binomial test
> prop.test(x=4, n=200, p=0.05, alternative = “less")
```

```
        1-sample proportions test with continuity correction

data:  4 out of 200, null probability 0.05
X-squared = 3.1842, df = 1, p-value = 0.03718
alternative hypothesis: true p is less than 0.05
95 percent confidence interval:
 0.00000000 0.04715366
sample estimates:
   p
0.02
```

1-7 Inference on Population Proportion

1-7.1 Hypothesis Testing on a Binomial Proportion

```
> #perform one-tailed Binomial test
> prop.test(x=4, n=200, p=0.05, alternative = “greater")
```

```
        1-sample proportions test with continuity correction

data:  4 out of 200, null probability 0.05
X-squared = 3.1842, df = 1, p-value = 0.9628
alternative hypothesis: true p is greater than 0.05
95 percent confidence interval:
 0.0074791 1.0000000
sample estimates:
   p
0.02
```

1-7 Inference on Population Proportion

1-7.3 Confidence Interval on a Binomial Proportion

Confidence Interval on a Binomial Proportion

If $\hat{p}$ is the proportion of observations in a random sample of size n that belong to a class of interest, an approximate $100(1-\alpha)\%$ CI on the proportion p of the population that belongs to this class is

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

where $z_{\alpha/2}$ is the upper $100\,\alpha/2$ percentage point of the standard normal distribution.

1-7 Inference on Population Proportion

1-7.3 Confidence Interval on a Binomial Proportion

Crankshaft Bearings

In a random sample of 85 automobile engine crankshaft bearings, 10 have a surface finish that is rougher than the specifications allow. Find a 95% confidence interval on the proportion of defective bearings.

Solution. A point estimate of the proportion of bearings in the population that exceeds the roughness specification is $\hat{p} = x/n = 10/85 = 0.12$. A 95% two-sided CI for p is computed from equation 4-73 as

$$\hat{p} - z_{0.025}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le p \le \hat{p} + z_{0.025}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

or

$$0.12 - 1.96\sqrt{\frac{0.12(0.88)}{85}} \le p \le 0.12 + 1.96\sqrt{\frac{0.12(0.88)}{85}}$$

which simplifies to

$$0.05 \le p \le 0.19$$

1-7 Inference on Population Proportion

1-7.3 Confidence Interval on a Binomial Proportion

Choice of Sample Size

Sample Size for a Specified E on a Binomial Proportion

If $\hat{P}$ is used as an estimate of p, we can be $100(1-\alpha)\%$ confident that the error $|\hat{P}-p|$ will not exceed a specified amount E when the sample size is

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

For a specified error E, an upper bound on the sample size for estimating p is

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

Example

- In a sample of 230 soccer players, 34 were left footed.
- What can you say about the proportion of left footed players of the entire league with 95% confidence?

n = 230

$$\hat{p} = \frac{34}{230} = 0.15$$

- CI(95%) = $$\left[0.15 - 1.96\sqrt{\frac{0.15(1-0.15)}{230}}, 0.15 + 1.96\sqrt{\frac{0.15(1-0.15)}{230}}\right]$$

$$= (0.10, 0.19)$$

In R: prop.test(x = 34, n = 230, p = 0.5, correct = FALSE)

CI for Difference in two Population Proportions

$$\hat{p}_1 - \hat{p}_2 \pm z^*_{\alpha} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Example: Snoring and Heart Disease

- 1105 snorers: 86 had heart disease
- 1379 non-snorers: 24 had heart disease

$$n_1 = 1105 \qquad \hat{p}_1 = 86/1105 = .0778$$
$$n_2 = 1379 \qquad \hat{p}_2 = 24/1379 = .0174$$
$$\hat{p}_1 - \hat{p}_2 = .0604$$

- **Conclusion:** With 99% confidence, the difference in the proportion of people with heart disease between the snores and non snorers is between 0.038 and 0.083.

$$= (.038, .083)$$

$$CI(99\%) = \left[.0778 - .0174 \pm 2.58 \sqrt{\frac{.0778(1-.0778)}{1105} + \frac{.0174(1-.0174)}{1379}} \right]$$

Example: Snoring and Heart Disease (cont.)

- 1105 snorers: 86 had heart disease
- 1379 non-snorers: 24 had heart disease

```
> prop.test( c(86, 24), c(1105, 1359), correct = F, conf.level = 0.99)

    2-sample test for equality of proportions without continuity correction

data:  c(86, 24) out of c(1105, 1359)
X-squared = 51.731, df = 1, p-value = 6.364e-13
alternative hypothesis: two.sided
99 percent confidence interval:
 0.03746030 0.08287572
sample estimates:
    prop 1     prop 2
0.07782805 0.01766004
```

Example: Snoring and Heart Disease (cont.)

- 1105 snorers: 86 had heart disease
- 1379 non-snorers: 24 had heart disease

```
> prop.test( c(86, 24), c(1105, 1359), correct = F, conf.level = 0.99,alternative ="less")

    2-sample test for equality of proportions without continuity correction

data:  c(86, 24) out of c(1105, 1359)
X-squared = 51.731, df = 1, p-value = 1
alternative hypothesis: less
99 percent confidence interval:
 -1.00000000  0.08067637
sample estimates:
    prop 1     prop 2
0.07782805 0.01766004
```

Example: Snoring and Heart Disease (cont.)

- 1105 snorers: 86 had heart disease
- 1379 non-snorers: 24 had heart disease

```
> prop.test( c(86, 24), c(1105, 1359), correct = F, conf.level =
0.99,alternative ="greater")

    2-sample test for equality of proportions without continuity
correction

data:  c(86, 24) out of c(1105, 1359)
X-squared = 51.731, df = 1, p-value = 3.182e-13
alternative hypothesis: greater
99 percent confidence interval:
 0.03965965 1.00000000
sample estimates:
    prop 1     prop 2
0.07782805 0.01766004
```

1-8 Three important tests

- **T-test:** compare two groups, or two interventions on one group.

- **CHI-squared and Fisher's test.** Compare the counts in a "contingency table".

- **ANOVA:** compare outcomes under several discrete interventions.

Chi-squared test

Often you will be faced with discrete (count) data. Given a table like this:

	Prob(X)	Count(X)
X=0	0.5	10
X=1	0.5	50

Where Prob(X) is part of a null hypothesis about the data (e.g. that a coin is fair).

The CHI-squared statistic lets you test whether an observation is consistent with the data:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

$$\chi_1^2 = \frac{(10-30)^2}{30} + \frac{(50-30)^2}{30} = \frac{800}{30} = 26.6$$

pvalue=0

O_i is an observed count, and E_i is the expected value of that count. It has a chi-squared distribution, whose p-values you compute to do the test.

1-9 Testing for Goodness of Fit - IDA

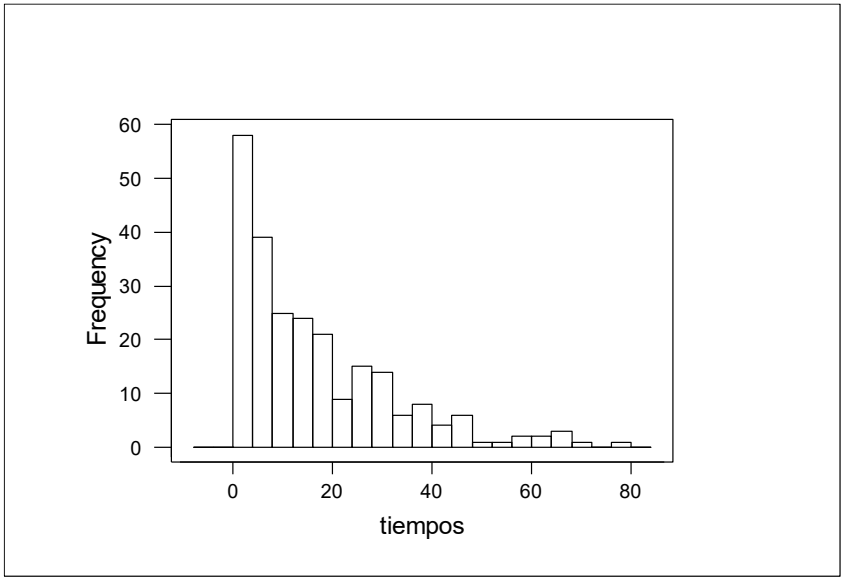
- So far, we have assumed the population or probability distribution for a particular problem is known.
- There are many instances where the underlying distribution is not known, and we wish to test a particular distribution.
- Use a goodness-of-fit test procedure based on the chi-square distribution.

$$X_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Interarrival times to a queueing system (I)

0.0056	0.0967	0.1807	0.2262	0.2389	0.2628	0.4112	0.5298	0.5525
0.5861	0.6353	0.6510	0.6731	0.7571	0.7587	1.0471	1.0651	1.1033
1.1745	1.1891	1.2248	1.2468	1.3303	1.3524	1.3836	1.4123	1.4896
1.5370	1.5647	1.6214	1.6926	1.7026	1.8342	1.9083	1.9799	2.0623
2.1642	2.1831	2.2502	2.3509	2.3802	2.4070	2.4423	2.5339	2.7035
2.7615	2.7816	2.8736	2.9131	2.9503	2.9543	3.3028	3.4130	3.4334
3.7070	3.7101	3.7975	3.9954	4.0307	4.2456	4.3496	4.3519	4.4318
4.4611	4.5225	4.6649	4.8233	4.8930	4.9570	5.0024	5.0737	5.1245
5.4059	5.4331	5.4764	5.5202	5.5676	5.7514	5.9360	5.9792	6.1983
6.3889	6.4247	6.4959	6.6339	6.6549	6.7329	6.9757	7.1417	7.1698
7.1863	7.3967	7.6357	7.6863	7.7195	7.8727	7.9692	8.1695	8.3292
8.5473	8.5664	8.5788	8.7657	8.7732	8.7898	8.9490	9.0254	9.2393
9.5786	9.6571	9.7378	9.9193	10.0864	10.4750	10.5871	10.8425	10.8702
11.0024	11.1306	11.1442	11.5322	11.5343	12.1626	12.2130	12.5622	13.0473
13.0725	13.2578	13.6691	13.8986	14.0160	14.0924	14.1088	14.2914	14.3412
14.3476	14.3792	14.3882	14.5239	14.5886	15.1880	15.2705	15.4756	15.5005
15.6594	15.9621	16.0468	16.0755	16.3010	16.3375	16.5762	16.7091	16.7427
16.8580	16.9638	17.2893	17.3069	17.3494	17.6276	18.0241	18.1331	18.5128
19.3180	19.3221	19.7837	19.8183	19.9215	20.4094	21.2502	21.3287	21.3893
21.9020	22.5750	22.6202	23.0887	23.4108	24.2013	24.2341	24.3140	24.8287
25.5066	26.0540	26.1539	26.3157	26.4727	26.8948	26.9533	26.9969	27.0315
27.7501	27.9995	28.8724	29.1316	29.2133	29.5204	29.8822	30.1434	30.2184
30.2375	30.3884	30.4502	31.3845	31.6547	31.7689	31.9442	32.2406	33.2581
33.4450	34.4730	35.3383	35.5309	36.2030	37.0102	37.3794	37.3898	38.2246
38.5539	38.6309	38.6955	40.2452	40.7250	41.0193	41.6327	44.2501	45.0125
45.0766	46.2803	46.6233	47.8474	48.5374	53.3607	56.1566	58.3313	60.6484
61.9682	66.3411	66.7154	67.0041	70.4577	79.3826			

Interarrival times to a queueing system (II): sample moments



Histogram 240 observed inter arrival times

Mean and standard deviation of the sample :

$$\hat{\mu} = 16.398179$$

$$\hat{\sigma} = 15.855114$$

Coefficient of Variation $cv = 0.966882 \approx 1.0 \rightarrow$

Hipothesis : exponential inter arrival times : $f(x) = 0.060982e^{-0.60982x}$

Interarrival times to a queueing system (III): chisquared test of the theoretical distribution

- Group observations into intervals: plot them in a histogram
- N_j is the number of observations in the sample belonging to j-th class, whose lower and upper values are: $[a_{j-1}, a_j)$
- According to the theoretical distribution considered f(x) (depending on the hypothesis), then the expected number of observation in j-th class Np_j, being N the total sample size, and p_j the expected probability for j-th interval, that according to f(x) is:

$$p_j = \int_{a_{j-1}}^{a_j} f(x)dx$$

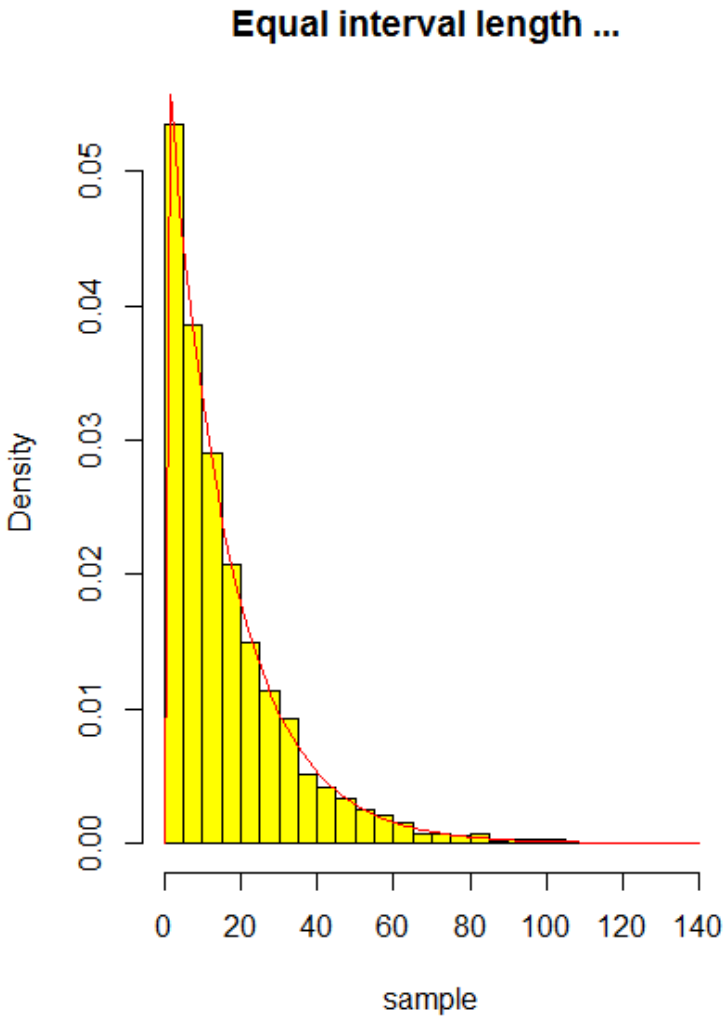
- Compute the χ^2 statistic for distribution matching, distributed as χ^2_{k-1}:

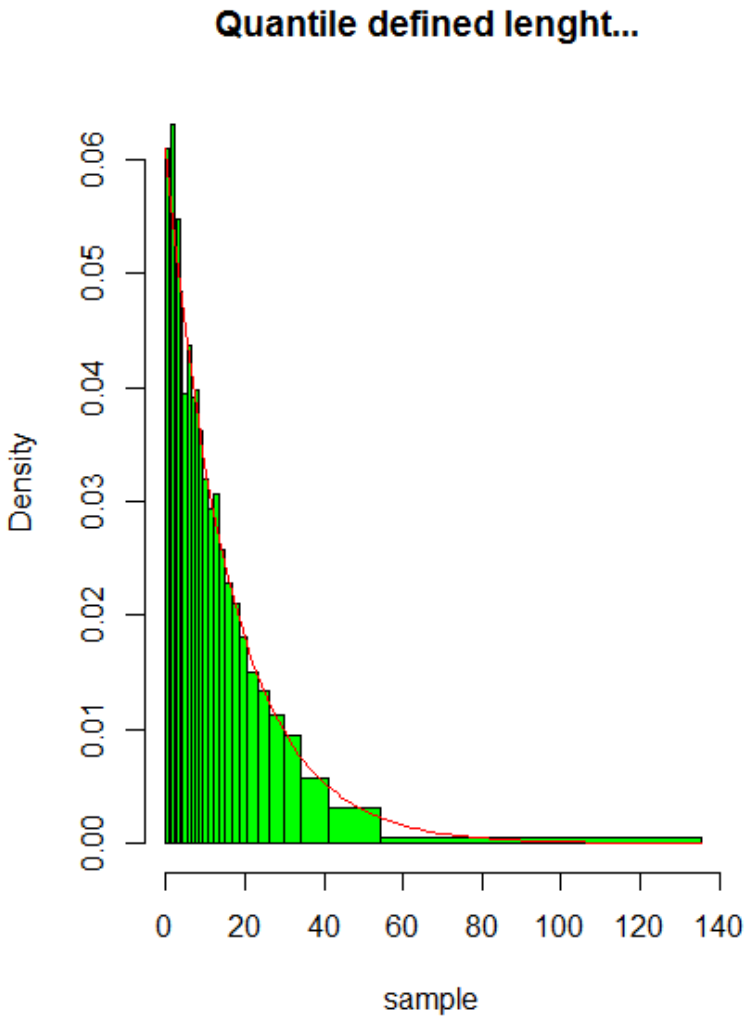
$$\chi^2 = \sum_{j=1}^{k} \frac{(N_j - Np_j)^2}{Np_j}$$

If the fit is good enough then χ^2 value should be less than the theoretical distribution of the statistic $\chi^2_{k-1,\alpha}$, for a confidence level of α.

- In the example taking equal probability for each interval p_j=0.04, fiven k = 25 classes, then the theoretical number of observations should be Np_j = 240 x 0.04 = 9.6. Lower and upper interval values are given by $a_j = -16.398179 \ln(1 - \frac{j}{25})$, leading to the following table .

Ensure equal number of expected obs. according to hypothetical distribution





Ensure equal number of expected obs. according to hypothetical distribution

```
par(mfrow=c(1,2))
hist(sample,freq=F,breaks=25,col="yellow",main="Equal
   interval length ... ")
curve(dgamma(x,shape=1,scale=16.398179),col=2,add=T)

sequence<-seq(0,1,by=0.04)
# qualist<-quantile(sample,sequence) # Restricted to
   when non-candidate parametric distribution is stated
qualist<-qgamma(sequence,shape=1,scale=16.398179)
qualist[1]<-0
sequence;qualist
hist(sample,freq=F,breaks=qualist,col="green",main="Qua
   ntile defined lenght... ")
curve(dgamma(x,shape=1,scale=16.398179),col=2,add=T)
```

Interarrival times to a queueing system (IV): validation of theoretical distribution

k	$[a_{j-1},a_j)$	N_j	Np_j	$(N_j - Np_j)^2/Np_j$
1	[0,0.6694)	12	9.6	0.6
2	[0.6694,1.3673)	12	9.6	0.6
3	[1.3673,2.0962)	12	9.6	0.6
4	[2.0962,2.8591)	11	9.6	0.2041
5	[2.8591,3.6591)	7	9.6	0.7041
6	[3.6591,4.5002)	10	9.6	0.0166
7	[4.5002,5.3868)	8	9.6	0.2666
8	[5.3868,6.3242)	9	9.6	0.0375
9	[6.3242,7.3183)	10	9.6	0.0166
10	[7.3183,8.3766)	8	9.6	0.2666
11	[8.3766,9.5079)	9	9.6	0.0375
12	[9.5079,10.7232)	7	9.6	0.7041
13	[10.7332,12.0357)	7	9.6	0.7041
14	[12.0357,13.4626)	6	9.6	1.35
15	[13.4626,15.0255)	12	9.6	0.6
16	[15.0255,16.7532)	13	9.6	1.2041
17	[16.7532,18.6846)	9	9.6	0.0375
18	[18.6846,20.8743)	6	9.6	1.35
19	[20.8743,23.4021)	7	9.6	0.7041
20	[23.4021,26.3918)	9	9.6	0.0375
21	[26.3918,30.051)	12	9.6	0.6
22	[30.051,34.7865)	13	9.6	1.2041
23	[34.7865,41.4173)	13	9.6	1.2041
24	[41.4173,52.7837)	8	9.6	0.2666
25	[52.7837,∞)	10	9.6	0.0166
				χ^2=13.3324

Threshold for chi-square distribution $\chi^2_{24,0.10} = 33.196 \geq$ **χ^2=13.3324 statistic $\Rightarrow$ H0 accepted**

Closing Words

All the tests so far are parametric tests that assume the data are **normally distributed**, and that the samples are **independent of each other and all have the same distribution** (IID).

They may be arbitrarily inaccurate if those assumptions are not met. Always make sure your data satisfies the assumptions of the test you're using. e.g. watch out for:

- Outliers – will corrupt many tests that use variance estimates.
- Correlated values as samples, e.g. if you repeated measurements on the same subject.
- Skewed distributions – give invalid results.

Non-parametric tests (CSI lab session)

These tests make no assumption about the distribution of the input data, and can be used on very general datasets:

- K-S test
- Bootstrap confidence intervals
- Non – parametric tests for means and variances in group (CSI lab session).

K-S test

The K-S (Kolmogorov-Smirnov) test is a very useful test for checking whether two (continuous or discrete) distributions are the same.

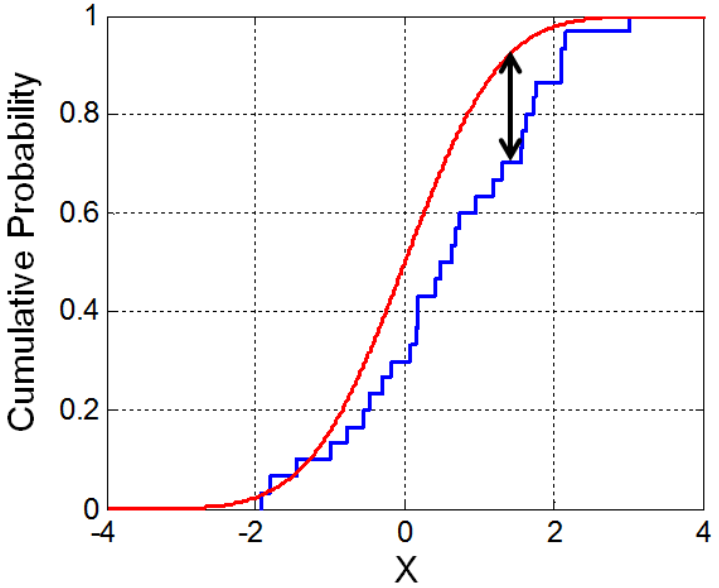
In the **one-sided test**, an observed distribution (e.g. some observed values or a histogram) is compared against a reference distribution.

In the **two-sided test**, two observed distributions are compared.

The K-S statistic is just the **max distance between the CDFs** of the two distributions.

While the statistic is simple, its distribution is not!

But it is available in most stat packages.



K-S test

The K-S test can be used to test **whether a data sample has a normal distribution** or not.

Thus it can be used as a sanity check for any common parametric test (which assumes normally-distributed data).

It can also be used to compare distributions of data values in a large data pipeline: **Most errors will distort the distribution of a data parameter and a K-S test can detect this**.

K-S test

```
> x <- rnorm( 500)
> ks.test(x, "pnorm")
        One-sample Kolmogorov-Smirnov test data:  x
D = 0.039321, p-value = 0.422
alternative hypothesis: two-sided

> y <- rnorm( 500, mean = 0, sd = 5)
> ks.test(y, "pnorm") # Tentative distribution parameter values have to be included
        One-sample Kolmogorov-Smirnov test data:  y
D = 0.34317, p-value < 2.2e-16
alternative hypothesis: two-sided

> ks.test(y, "pnorm", mean = mean(y), sd = sd(y))  # Correct
        One-sample Kolmogorov-Smirnov test data:  y
D = 0.022137, p-value = 0.9671
alternative hypothesis: two-sided

> # Two-sample case
> ks.test( x, y )
        Two-sample Kolmogorov-Smirnov test data:  x and y
D = 0.356, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Non-parametric tests (CSI lab session)

Permutation tests

Bootstrap confidence intervals

- We won't discuss these in detail, but it's important to know that non-parametric tests using one of the above methods exist for many forms of hypothesis.

- They make no assumptions about the distribution of the data, but in many cases are just as sensitive as parametric tests.

- They use computational cycles to simulate sample data, to derive p-value estimates approximately, and accuracy improves with the amount of computational work done.

Non-parametric tests (CSI lab session)

- Parametric T-TEST (dicothomic factor):
  - t.test(*formula, dataframe, var.equal=c(TRUE,FALSE),alternative*)
  - Non parametric version: *wilcox.test(formula, dataframe)*
- ONEWAY – Analysis of Variance for 1 factor:
  - ONEWAY – Parametric Analysis of Variance for 1 factor: aov(*formula, dataframe*) o oneway.test(*formula, dataframe, var.equal=c(TRUE,FALSE)). Ex: oneway.test(*Y ~ A)
  - Non Parametric contrast for the equal mean hypothesis in groups defined by the level of 1 factor: kruskal.test(*formula,dataframe,var.equal=c(TRUE,FALSE)). Ex:* kruskal.*test(*Y ~ A)

Non-parametric tests (CSI lab session)

- Correlation test for 2 numeric variables is given in R by:
  - Parametric version for normal-like variables: cor(var1, var2,method=”Pearson”) (default option in R)
  - Non-parametric version for general variables: cor(var1, var2,method=”Spearman”)
- Parametric contrasts (assuming normal distribution of Y) for equal dispersion (variance) in groups defined by levels of the studied factor (Y ~ A is the formula parameter):
  - Dichotomous Case: var.test(*formula,dataframe)*
  - Polytomous Case: bartlett.test(*formula,dataframe).*
  - Breusch Pagan Test: bptest(prestige~type) # popular in econometrics.

Non-parametric tests (CSI lab session)

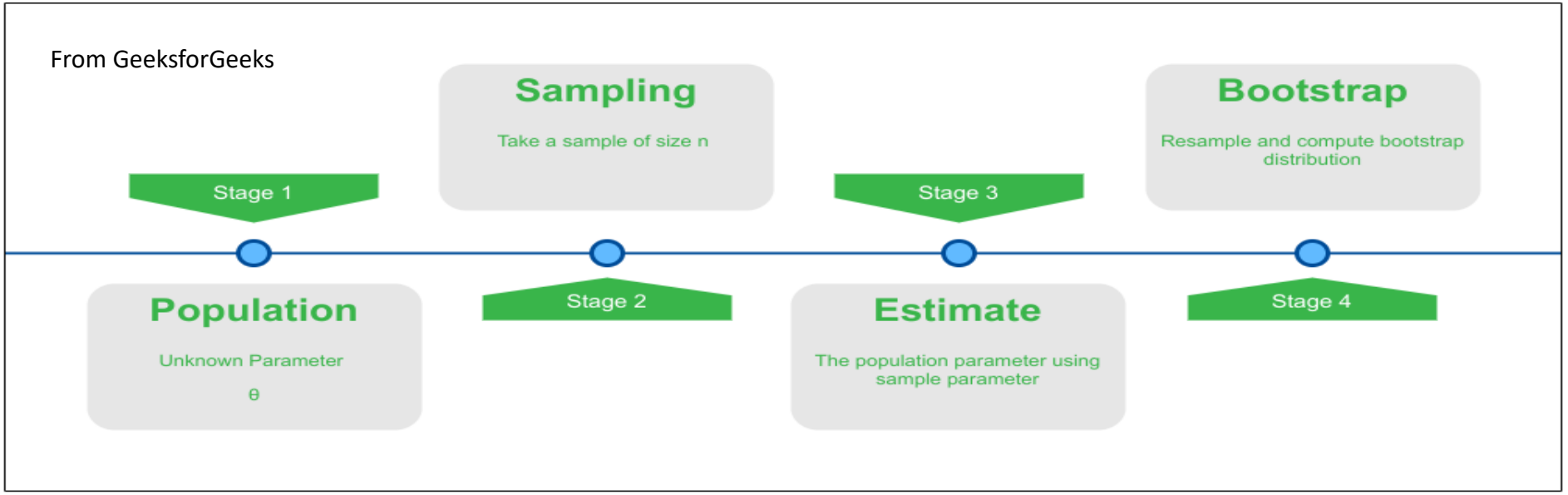
- Non Parametric contrasts (normal distribution of Y not required) for equal dispersion (variance) in groups defined by levels of the studied factor (Y ~ A is the formula parameter):
  - fligner.test(*formula,dataframe).*
- Comparison between individual group means: Provided that F test shows a difference between groups, the question arises of wherein the difference lies.
  - Parametric version: pairwise.t.test(Y, A) .
  - Non-Parametric version: pairwise.wilcox.test(Y, A) .

Non-parametric tests (CSI lab session)

- **Bootstrap confidence intervals** are base on:
- First proposed by Bradley Efron.
- Defining multiple *resamples* (with replacement) from a single set of observations, and computes the effect size of interest on each of these resamples (ex: difference in means).
- The bootstrap resamples of the effect size can then be used to determine the 95% CI.
- Resampling distribution of the difference in means approaches to a normal distribution due to the Central Limit Theorem.
- Asymmetrical resampling for skewed distributions

Non-parametric tests: Bootstrap process

- Sample n elements with replacement from original sample data.
- For every sample calculate the desired statistic eg. mean, median etc.
- Repeat steps 1 and 2 m times and save the calculated stats.
- Determine α and 1- α percentiles for the bootstrap distribution to calculate the 95% CI



Non-parametric tests: Bootstrap confidence interval in R – Example Davison and Hinkley

- Determine a 95% CI for the population of 49 U.S. cities between 1920 and 1930. The measurements are the population (in 1000's) of 49 U.S. cities in 1920 and 1930 (bigcity dataset).

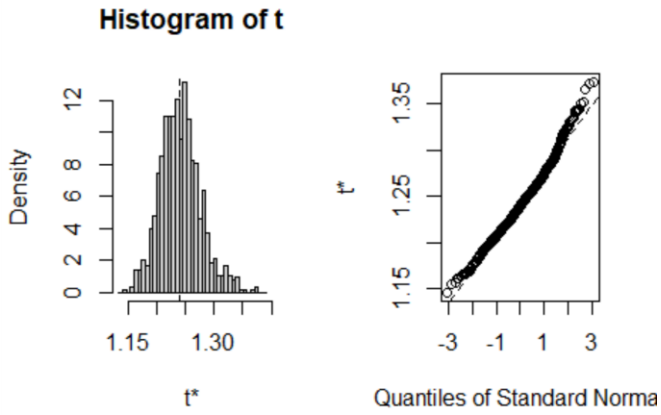
```
> head(bigcity)
> names(bigcity) <- c("Pop.1920", "Pop.1930")
> ratio <- function( df, w) sum(df$Pop.1930*w)/sum(df$Pop.1920*w)
> bootdis<-boot(bigcity, ratio, R = 999, stype = "w")
> plot(bootdis)
> boot.ci( boot.out=bootdis )
```

BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 999 bootstrap replicates
CALL : boot.ci(boot.out = bootdis)

Intervals :

Level	Normal	Basic
95%	(1.168, 1.306)	(1.155, 1.301)

Level	Percentile	BCa
95%	(1.177, 1.323)	(1.179, 1.327)



IMPORTANT TERMS AND CONCEPTS

Alternative hypothesis
Bias in estimation
Chi-squared distribution
Comparative experiment
Confidence bound
Parameter estimation
Point estimation
Power of a test
Practical significance versus statistical significance
Precision of estimation
~~Prediction interval~~
Confidence coefficient
Confidence interval
Confidence level
Confidence limits
Coverage
Critical region
Estimated standard error
Probability of a type I error
Probability of a type II error
Procedure for hypothesis testing
~~Relative efficiency~~ of an estimator
Fixed significance level hypothesis testing
Goodness of fit
Hypothesis testing
Minimum variance unbiased estimator
Null hypothesis
Sample size determination
Significance level
Standard error
Statistical hypothesis
Statistical inference
t-distribution
One-sided alternative hypothesis
One-sided confidence bounds
Op~~erating characteristic~~ curves
P-values
Test statistic
Tolerance interval
Two-sided alternative hypothesis
Type I error
Type II error

Maximum Likelihood

Introduction to Maximum Likelihood Estimation

- There are two common approaches to parameter estimation: maximum likelihood and Bayesian estimation.
- Maximum Likelihood: treat the parameters as quantities whose values are fixed but unknown.
- Bayes: treat the parameters as random variables having some known prior distribution. Observations of samples converts this to a posterior.
- Bayesian Learning: sharpen the *a posteriori* density causing it to peak near the true value.

Likelihood Function

- **Objective : Estimating the unknown parameters θ of a population distribution based on a random sample $\chi_1,\ldots,\chi_n$ from that distribution**

- Previous chapters : Intuitive Estimates
  => Sample Means for Population Mean

- To improve estimation, R. A. Fisher (1890~1962) proposed MLE in 1912~1922.

Joint p.d.f. vs. Likelihood Function

- Identical quantities
- Different interpretation
- Joint p.d.f. of X_1 ,…, X_n :
  - A function of χ_1,…,χ_n for given θ
  - Probability interpretation

$$f\left(x_1,\ldots,x_n|\theta\right)=f\left(x_1|\theta\right)f\left(x_2|\theta\right)\ldots f\left(x_n|\theta\right)=\prod_{i=1}^{n} f\left(x_i|\theta\right)$$

- Likelihood Function of θ :
  - A function of θ for given x_1,…,x_n
  - No probability interpretation

$$L\left(\theta|x_1,\ldots,x_n\right)=f\left(x_1,\ldots,x_n|\theta\right)=f\left(x_1|\theta\right)\ldots f\left(x_n|\theta\right)=\prod_{i=1}^{n} f\left(x_i|\theta\right)$$

Example : Normal Distribution

- Suppose $x_1,\ldots,x_n$ is a random sample from a normal distribution with p.d.f.:

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\{-\frac{(x-\mu)^2}{2\sigma^2}\}$$

parameter (μ, σ^2), Likelihood Function:

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} [\frac{1}{\sigma\sqrt{2\pi}} \exp\{-\frac{(x_i-\mu)^2}{2\sigma^2}\}]$$

$$= (\frac{1}{\sigma\sqrt{2\pi}})^n \exp\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i-\mu)^2\}$$

Calculation of Maximum Likelihood Estimators (MLE)

- MLE of an unknown parameter θ :

  The value $\theta = \theta(x_1, \ldots, x_n)$ which maximizes the likelihood function $L(\theta | x_1, \ldots, x_n)$

- Example of MLE:
  - Distribution fitting: exponential distribution

Example of MLE

- Mean $\theta = \mu > 0, f_\mu(x) = \frac{1}{\mu} e^{-x/\mu}, x \geq 0$
- Likelihood function :

$$L(\mu) = \left(\frac{1}{\mu} e^{-x_1/\mu}\right)\left(\frac{1}{\mu} e^{-x_2/\mu}\right)\ldots\left(\frac{1}{\mu} e^{-x_n/\mu}\right) = \frac{1}{\mu^n} \exp\left(-\frac{1}{\mu}\sum_{i=1}^{n} x_i\right)$$

- Log-likelihood function:

$$l(\mu) = \ln[L(\mu)] = -n\ln\mu - \frac{1}{\mu}\sum_{i=1}^{n} x_i$$

- Determine the value of $\hat{\mu}$ maximizing $L(\mu)$ for $\mu \geq 0$:

$$\max_{\mu} L(\mu) \equiv \max_{\mu} l(\mu)$$

$$\frac{dl(\mu)}{d\mu} = -\frac{n}{\mu} + \frac{1}{\mu^2}\sum_{i=1}^{n} x_i \Rightarrow \hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- According to KKT conditions:

$$\frac{dl^2(\mu)}{d\mu^2} = \frac{n}{\mu^2} - \frac{2}{\mu^3}\sum_{i=1}^{n} x_i < 0 \text{ for } \mu = \hat{\mu}$$

Properties of MLE's

- **Objective**

  optimality properties in large sample

- **Fisher information (continuous case)**

$$I(\theta) = \int_{-\infty}^{\infty} \left[\frac{d \ln f(x \mid \theta)}{d\theta} \right]^2 f(x \mid \theta) dx = E\left\{ \left[\frac{d \ln f(x \mid \theta)}{d\theta} \right]^2 \right\}$$

- **Alternatives of Fisher information**

$$I(\theta) = E\left\{ \left[\frac{d \ln f(x \mid \theta)}{d\theta} \right]^2 \right\} = Var\left\{ \frac{d \ln f(x \mid \theta)}{d\theta} \right\} \quad (1)$$

$$I(\theta) = \int_{-\infty}^{\infty} \left[\frac{d \ln f(x \mid \theta)}{d\theta} \right]^2 f(x \mid \theta) dx = -E\left\{ \left[\frac{d^2 \ln f(x \mid \theta)}{d\theta^2} \right] \right\} \quad (2)$$

Score function and properties

- The first derivative of the log-likelihood function is called Fisher's score function:

$$\mathbf{u}(\boldsymbol{\theta}) = \frac{\partial \log \mathrm{L}(\boldsymbol{\theta}; \mathbf{y})}{\partial \boldsymbol{\theta}}$$

- If the log-likelihood is concave, one can find the maximum likelihood estimator by setting the score to zero, i.e. by solving the system of equations

$$\mathbf{u}(\hat{\boldsymbol{\theta}}) = \mathbf{0}$$

Score function and properties. Proof

$$\int_{-\infty}^{\infty} f(x \mid \theta) dx = 1$$

$$\int_{-\infty}^{\infty} \frac{df(x \mid \theta)}{d\theta} dx = \frac{d}{d\theta} 1 = 0$$

$$\int_{-\infty}^{\infty} \frac{df(x \mid \theta)}{d\theta} dx = \int_{-\infty}^{\infty} \frac{df(x \mid \theta)}{d\theta} \frac{1}{f(x \mid \theta)} f(x \mid \theta) dx$$

$$= \int_{-\infty}^{\infty} \frac{d \ln f(x \mid \theta)}{d\theta} f(x \mid \theta) dx$$

$$= E\left\{ \frac{d \ln f(x \mid \theta)}{d\theta} \right\} = 0$$

The Information matrix

- The score is a random vector with some interesting statistical properties.
- In particular, the score evaluated at the true parameter value has mean zero,

$$\mathrm{E}[\mathbf{u}(\boldsymbol{\theta})] = \mathbf{0}$$

- and variance-covariance matrix given by the information matrix

$$\mathrm{var}[\mathbf{u}(\boldsymbol{\theta})] = \mathrm{E}[\mathbf{u}(\boldsymbol{\theta})\mathbf{u}'(\boldsymbol{\theta})] = \mathbf{I}(\boldsymbol{\theta})$$

The Information matrix

- Under mild regularity conditions, the information matrix can also be obtained as minus the expected value of the second derivatives of the loglikelihood:

$$\mathbf{I}(\boldsymbol{\theta}) = -\mathrm{E}[\frac{\partial^2 \log \mathrm{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}]$$

- The matrix of negative observed second derivatives is sometimes called the observed information matrix

Proof

$$I(\theta)=\int_{-\infty}^{\infty}\left[\frac{d\ln f(x\mid\theta)}{d\theta}\right]^2 f(x\mid\theta)dx=-E\left\{\left[\frac{d^2\ln f(x\mid\theta)}{d\theta^2}\right]\right\}$$

$$\text{diffrentiating}\int_{-\infty}^{\infty}\frac{d\ln f(x\mid\theta)}{d\theta}f(x\mid\theta)dx$$

$$\int_{-\infty}^{\infty}\left[\frac{d^2\ln f(x\mid\theta)}{d\theta^2}f(x\mid\theta)+\frac{d\ln f(x\mid\theta)}{d\theta}\frac{df(x\mid\theta)}{d\theta}\right]dx$$

$$=\int_{-\infty}^{\infty}\left[\frac{d^2\ln f(x\mid\theta)}{d\theta^2}+\frac{d\ln f(x\mid\theta)}{d\theta}\frac{1}{f(x\mid\theta)}\right]f(x\mid\theta)dx$$

$$=\int_{-\infty}^{\infty}\left[\frac{d^2\ln f(x\mid\theta)}{d\theta^2}+\left\{\frac{d\ln f(x\mid\theta)}{d\theta}\right\}^2\right]f(x\mid\theta)dx=0$$

MLE (Continued)

- **Define the Fisher information for an i.i.d. sample**

$$X_1, X_2, \cdots, X_n \text{ i.i.d. sample from p.d.f } f(x \mid \theta)$$

$$I_n(\theta) = -E\left\{\frac{d^2 \ln f(X_1, X_2, \cdots, X_n \mid \theta)}{d\theta^2}\right\}$$

$$= -E\left\{\frac{d^2}{d\theta^2}\left[\ln f(X_1 \mid \theta) + \ln f(X_2 \mid \theta) \cdots + \ln f(X_n \mid \theta)\right]\right\}$$

$$= -E\left\{\frac{d^2 \ln f(X_1 \mid \theta)}{d\theta^2}\right\} - E\left\{\frac{d^2 \ln f(X_2 \mid \theta)}{d\theta^2}\right\} - \cdots - E\left\{\frac{d^2 \ln f(X_n \mid \theta)}{d\theta^2}\right\}$$

$$= I(\theta) + I(\theta) + \cdots I(\theta) = nI(\theta)$$

MLE (Continued)

- **Generalization of the Fisher information for k-dimensional vector parameter**

p.d.f. of an r.v. X is $f(x \mid \theta)$, where $\theta = (\theta_1, \theta_2, \cdots, \theta_k)$

information matrix of θ, $I(\theta)$, is given by

$$I_{ij}(\theta) = E\left\{\left[\frac{\partial \ln f(x \mid \theta)}{\partial \theta_i}\right]\left[\frac{\partial \ln f(x \mid \theta)}{\partial \theta_j}\right]\right\}$$

$$= -E\left\{\frac{\partial^2 \ln f(x \mid \theta)}{\partial \theta_i \partial \theta_j}\right\}$$

MLE (Continued)

- **Newton-Raphson and Fisher Scoring**
- Calculation of the mle often requires iterative procedures. Consider expanding the score function evaluated at the **mle** $\widehat{\theta}$ around a trial value θ_0 using a first order Taylor series, so that

$$\mathbf{u}(\hat{\boldsymbol{\theta}}) \approx \mathbf{u}(\boldsymbol{\theta}_0) + \frac{\partial \mathbf{u}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$$

- Let **H** denote the Hessian or matrix of second derivatives of the log-likelihood function

$$\mathbf{H}(\boldsymbol{\theta}) = \frac{\partial^2 \log \mathrm{L}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = \frac{\partial \mathbf{u}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

MLE (Newton-Raphson and Fisher Scoring)

- Setting the left-hand-size of the equation to zero and solving for $\widehat{\theta}$

$$\mathbf{0} = \mathbf{u}(\hat{\boldsymbol{\theta}}) \approx \mathbf{u}(\boldsymbol{\theta}_0) + \frac{\partial \mathbf{u}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$$

- gives the first-order approximation

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0 - \mathbf{H}^{-1}(\boldsymbol{\theta}_0)\mathbf{u}(\boldsymbol{\theta}_0)$$

- An alternative procedure first suggested by Fisher is to replace minus the Hessian by its expected value, the information matrix. The resulting procedure is known as Fisher Scoring and takes as our improved estimate

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0 + \mathbf{I}^{-1}(\boldsymbol{\theta}_0)\mathbf{u}(\boldsymbol{\theta}_0)$$

MLE (Continued)

- **Cramér-Rao Lower Bound**
- A random sample $X_1, X_2, \ldots, X_n$ from p.d.f $f(x|\theta)$.
- Let $\hat{\theta}$ be any estimator of θ with $E(\hat{\theta}) = \theta + B(\theta)$,

  where $B(\theta)$ is the bias of $\hat{\theta}$.
- If $B(\theta)$ is differentiable in θ and if certain regularity conditions holds, then

$$Var(\hat{\theta}) \geq \frac{[1+B'(\theta)]^2}{nI(\theta)}$$

(Cramér-Rao inequality)

- The ratio of the lower bound to the variance of any estimator of θ is called the **efficiency** of the estimator.
- An estimator has efficiency = 1 is called the **efficient estimator**.
- **ML estimators are efficient**

Large Sample Inference Based on the MLE's

Large sample inference on unknown parameter θ, normally distributed

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

estimate

$$I(\hat{\theta}) = -\frac{1}{n}\sum_{i=1}^{n}\left[\frac{d^2 \ln f(X_i|\theta)}{d\theta^2}\right]_{\theta=\hat{\theta}}$$

100(1-α)% CI for θ

$$\hat{\theta} - z_{\alpha/2}\frac{1}{\sqrt{nI(\hat{\theta})}} \leq \theta \leq \hat{\theta} + z_{\alpha/2}\frac{1}{\sqrt{nI(\hat{\theta})}}$$

Delta Method for Approximating the Variance of an Estimator

- **Delta method**

  estimate a nonlinear function $h(\theta)$

  suppose that $E(\hat{\theta}) = \theta$ and $Var(\hat{\theta})$ is a known function of θ.

expand $h(\hat{\theta})$ around θ using first-order taylor series

$$h(\hat{\theta}) \cong h(\theta) + (\hat{\theta} - \theta)h'(\theta)$$

using $E(\hat{\theta} - \theta) = 0,$ $Var\left[h(\hat{\theta})\right] \cong \left[h'(\theta)\right]^2 Var(\hat{\theta})$

Likelihood Ratio Tests

The last section presented an inference for pointwise estimation based on likelihood theory. In this section, we present a corresponding inference for testing hypotheses.

Let $f(x;\theta)$ be a probability density function where θ is a real valued parameter taking values in an interval Θ that could be the whole real line. We call the parameter space. An alternative hypothesis H_1 will restrict the parameter θ to some subset Θ_1 of the parameter space Θ. The null hypothesis H_0 is then the complement θ of with respect to Θ.

Consider the two-sided hypothesis

$H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$, where θ_0 is a specified value.

We will test versus H_0 on the basis of the random sample H_1 from

$X_1, X_2, \ldots, X_n$ $f(x;\theta)$

If the null hypothesis holds, we would expect the likelihood $L(\theta) = \prod_{i=1}^{n} f(x_i;\theta)$ to be relatively large, when evaluated at the prevailing value θ_0. Consider the ratio of two likelihood functions, namely $\lambda = \frac{L(\theta_0)}{L(\hat{\theta})}$

Note that $\lambda \leq 1$, but if H_0 is true λ should be close to 1; while if H_1 is true, λ should be smaller. For a specified significance level α, we have the decision rule, reject H_0

in favor of H_1 if $\lambda \leq c$, where c is such that $\alpha = P_{\theta_0}[\lambda \leq c]$

This test is called the likelihood ratio test.

Example 1

Let $X_1, X_2, ..., X_n$ be a random sample of size n from a normal distribution with known variance. Obtain the likelihood ratio for testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$

$$L(\mu / X_1, ..., X_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = (2\pi\sigma^2)^{\frac{-n}{2}} e^{\frac{-\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}}$$

$$\ln L(\mu) = \frac{-n}{2} \ln(2\pi\sigma^2) - \frac{\sum (x_i - \mu)2}{2\sigma^2}$$

$$\frac{\partial}{\partial\mu} \ln L(\mu) = \frac{\sum (x_i - \mu)}{\sigma^2} = 0.$$ So $\hat{\mu} = \bar{x}$ is a maximum since

$$\frac{\partial^2}{\partial\mu^2} \ln L(\mu) = \frac{-1}{\sigma^2}$$ < 0 . Thus $\hat{\mu} = \bar{x}$ is the MLE of μ .

Example 1 (continued)

$$\lambda = \frac{L(\mu_0)}{L(\hat{\mu})} = \frac{(2\pi\sigma^2)^{\frac{-n}{2}} e^{\frac{-\sum (x_i-\mu_0)^2}{2\sigma^2}}}{((2\pi\sigma^2)^{\frac{-n}{2}} e^{\frac{-\sum (x_i-\bar{x})^2}{2\sigma^2}}} = e^{\frac{-\sum (x_i-\mu_0)^2-(x_i-\bar{x})^2}{2\sigma^2}} = e^{\frac{-\sum [(x_i-\bar{x})+(\bar{x}-\mu_0)]^2-(x_i-\bar{x})^2}{2\sigma^2}}$$

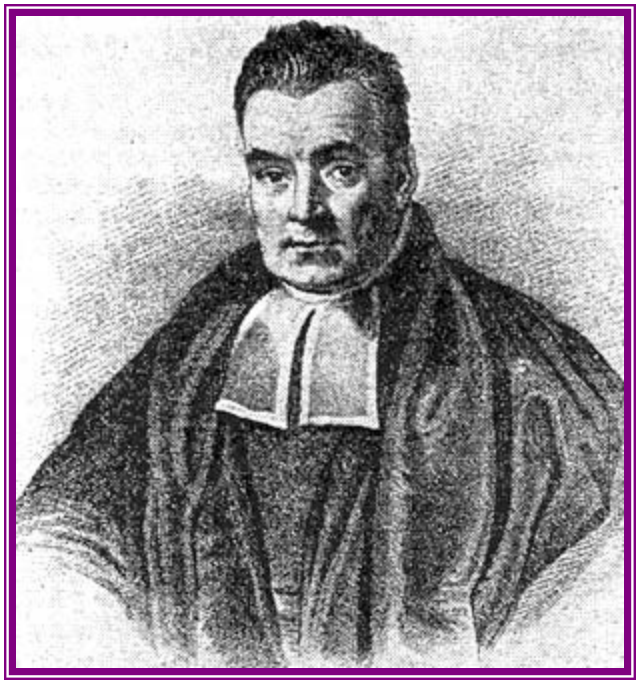
$$= e^{\frac{-\sum (x_i-\bar{x})^2+2(x_i-\bar{x})(\bar{x}-\mu_0)+(\bar{x}-\mu_0)^2-(x_i-\bar{x})^2}{2\sigma^2}} = e^{\frac{-\sum (\bar{x}-\mu_0)^2}{2\sigma^2}} = e^{\frac{-n(\bar{x}-\mu_0)^2}{2\sigma^2}}$$

$$= e^{\frac{-z_0^2}{2}}$$

thus $\lambda \le c$ is equivalent to $e^{\frac{-z_0^2}{2}} \le c$, or $z_0^2 \ge c^*$

So $P\left(|z_0| \ge c^{**}\right) = \alpha$ thus $c^{**} = z_{\alpha/2}$

Bayesian Inference



Thomas Bayes (c. 1702 – April 17, 1761)
Source: www.wikipedia.com

Thomas Bayes (pictured above) was a Presbyterian minister and a mathematician born in London who developed a special case of Bayes' theorem which was published and studied after his death.

Bayes' Theorem (review):

$$f(A|B) = f(A \cap B) \,/\, f(B) = f(B \mid A)\, f(A) \,/\, f(B)$$

since, $f(A \cap B) = f(B \cap A) = f(B \mid A)\, f(A)$

Some Key Terms in Bayesian Inference……in plain English

- ***prior distribution*** – probability tendency of an uncertain quantity, θ, that expresses previous knowledge of θ from, for example, a past experience, with the absence of some proof
- ***posterior distribution*** – this distribution takes proof into account and is then the conditional probability of θ. The posterior probability is computed from the prior and the likelihood function using Bayes' theorem.
- ***posterior mean*** – the mean of the posterior distribution
- ***posterior variance*** – the variance of the posterior distribution
- ***conjugate priors*** - a family of prior probability distributions in which the key property is that the posterior probability distribution also belongs to the family of the prior probability distribution

Bootstrap Resampling method

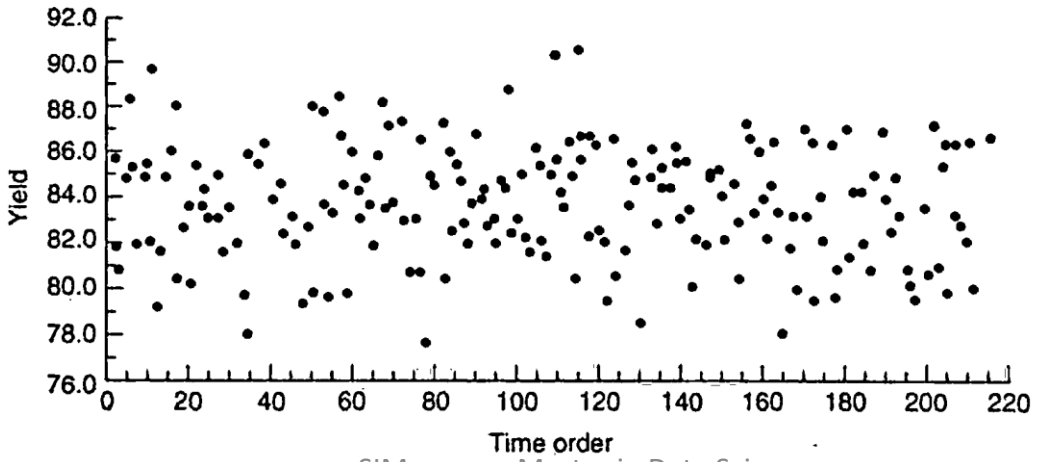
- Bootstrap is a powerful, computer-based method for statistical inference relying on few assumptions.
- The basic idea of bootstrap is make inference about a estimate for a population parameter Θ on sample data.
- It is a resampling method by independently sampling with replacement from an existing sample data with same sample size n and performing inference among these resampled data.
- Let $Z = \begin{pmatrix} z_1 & \dots & z_n \end{pmatrix}^T$ where $z_i = \begin{pmatrix} x_i & , & y_i \end{pmatrix}$.
- The basic idea is to randomly draw datasets with replacement from the training data, each sample the same size as the original training set. This is done B times (B = 100 say), producing B bootstrap datasets.
- Then we refit the model to each of the bootstrap datasets, and examine the behavior of the fits over the B replications.

Hypothesis Testing when no external reference distribution is available

- **First discussion Box-Hunter-Hunter (1978)**
- An experiment was performed on a manufacturing plant by making in sequence 10 batches of a chemical using production method A followed by 10 batches using a modified method B.
- What evidence do the data provide that method B gives higher yields than method A?
- They found that: $\bar{y}_A = 84.24, \qquad \bar{y}_B = 85.54$
- The modified method seems to give an average that is 1.3 units higher.
- However, considering the variability in batch to batch outcomes. Could they reasonably claim whether the new process B is better or whether the observed difference in the average could just be a chance event?

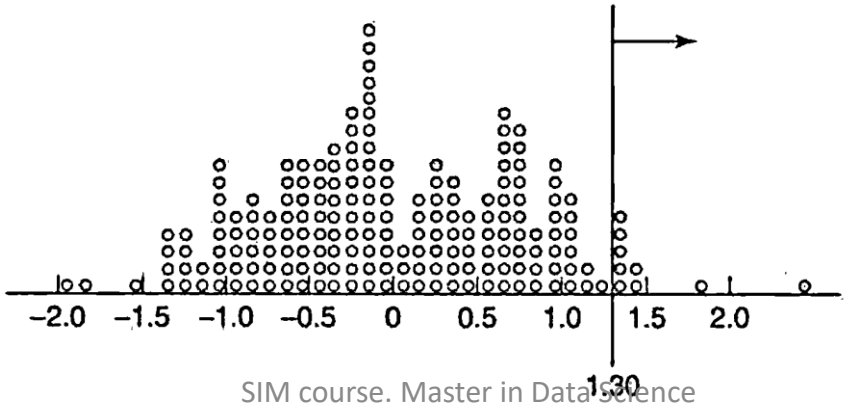
Hypothesis Testing when no external reference distribution is available

- **First discussion Box-Hunter-Hunter (1978)**
- They found that: $\bar{y}_A = 84.24, \qquad \bar{y}_B = 85.54$
- Could they reasonably claim whether the new process B is better or whether the observed difference in the average could just be a chance event?
- To answer this question, yields of the process for the previous 210 batches can be used.



Hypothesis Testing when no external reference distribution is available

- **First discussion Box-Hunter-Hunter (1978)**
- They found that: $\bar{y}_A = 84.24, \qquad \bar{y}_B = 85.54$
- Could they reasonably claim whether the new process B is better or whether the observed difference in the average could just be a chance event?
- **Reference distribution** of 191 differences between averages of adjacent sets of 10 observations

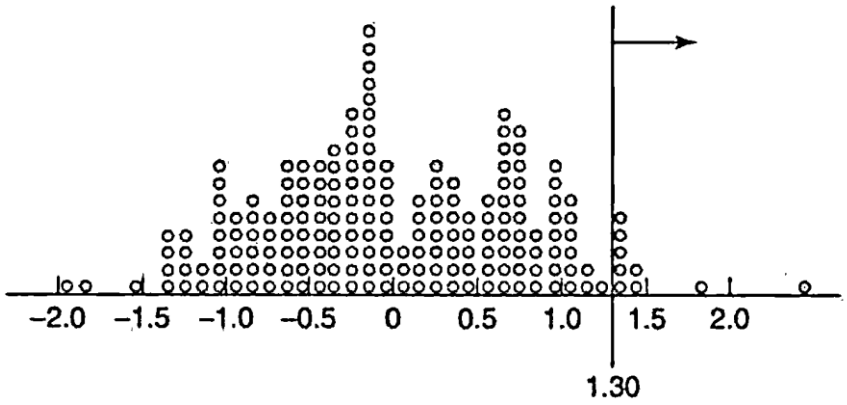


Hypothesis Testing when no external reference distribution is available

- **The null hypothesis is the observed difference is a member of the reference set**

$$\bar{y}_A = 84.24, \qquad \bar{y}_B = 85.54$$

- **Reference distribution** of 191 differences between averages of adjacent sets of 10 observations.



- Only 9/191=0.047, 4.7% significance level. Thus, the observed difference is statistically significant. No reference distribution (normality or independence) has been used, just the observed distribution.