Statistics, Probability, and Statistical Models

Pramudita Satria Palar, Ph.D.

Last edited: 28-September-2022

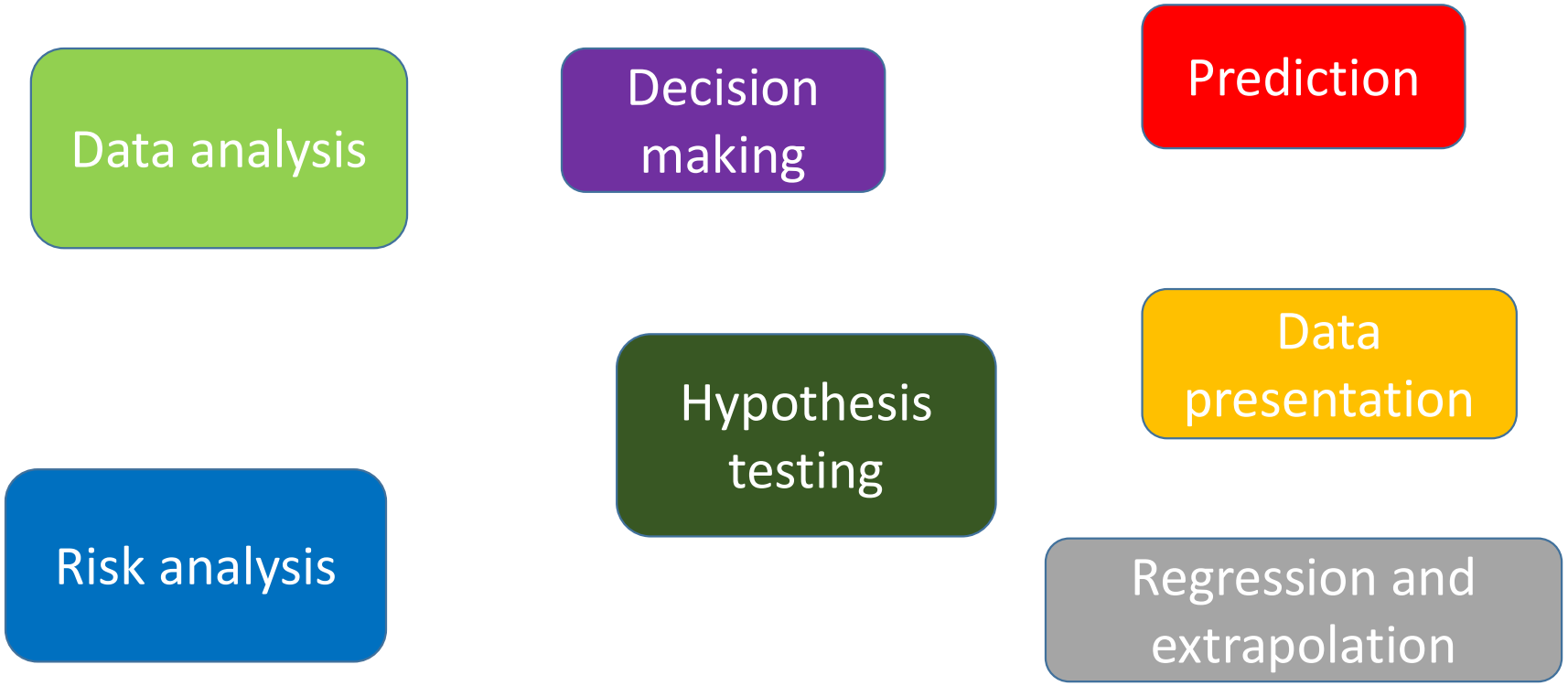
The goals of statistics

- **Statistics: The area of applied mathematics concerned with the data collection, analysis, interpretation, and presentation.**
- Used in practically every scientific discipline.
- Discover patterns and relationships from data and infer useful insight and knowledge from data.
- **You always work with data, right?**

This course will cover:

- Descriptive statistics
- Statistical hypothesis test.
- Basic statistical regression techniques.
- Data visualization and presentation.

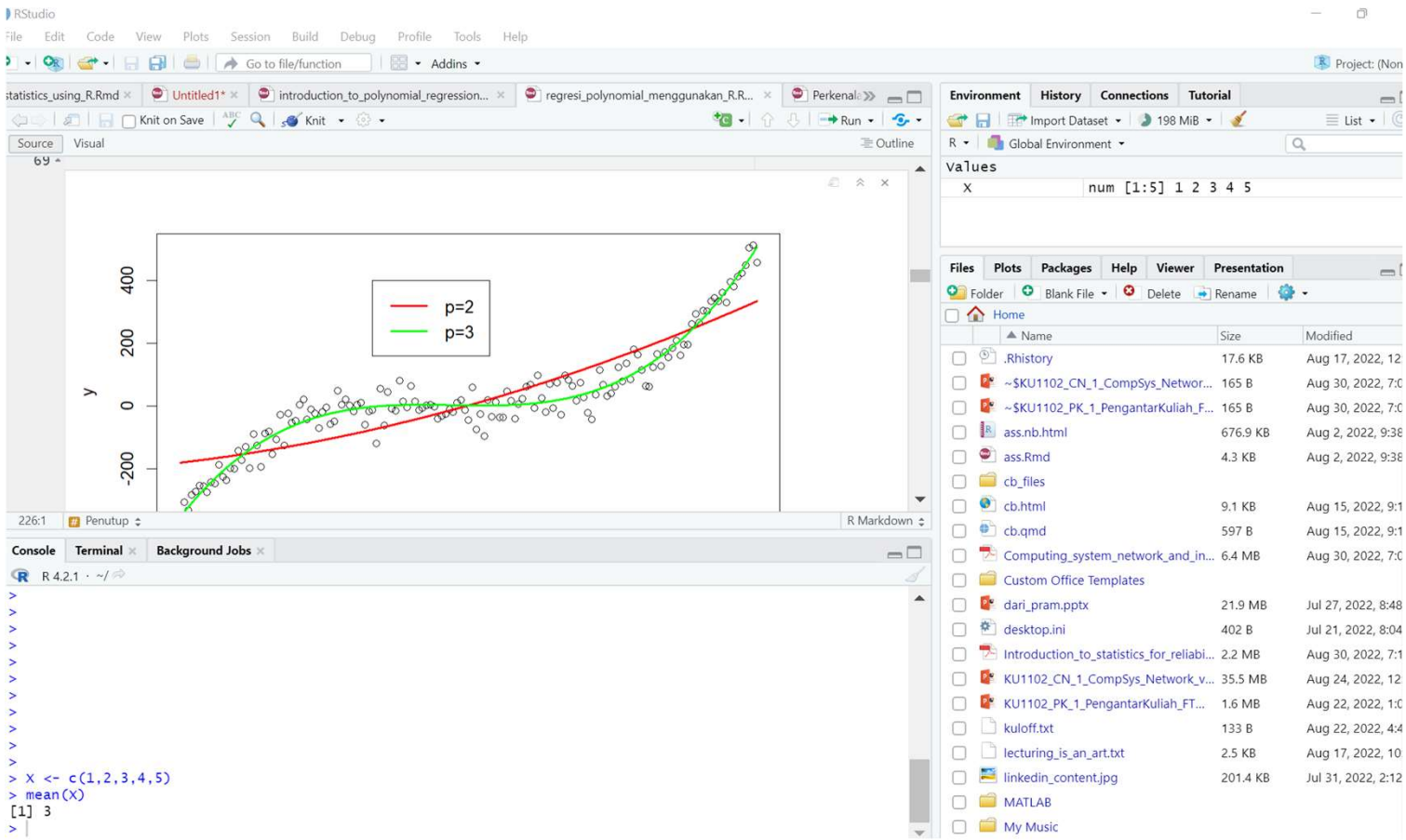
Statistics and probability in research



Tools for statistics

R and R studio

High-level programming



RStudio

Integrated Development Environment for R

A simple example using R

Console | Terminal | Background Jobs

R 4.2.1 · ~/

```
> X <- c(13,12.5,16,16.33,12.3)
> mean(X)
[1] 14.026
> median(X)
[1] 13
> sd(X)
[1] 1.972658
>
```

Save data into a vector

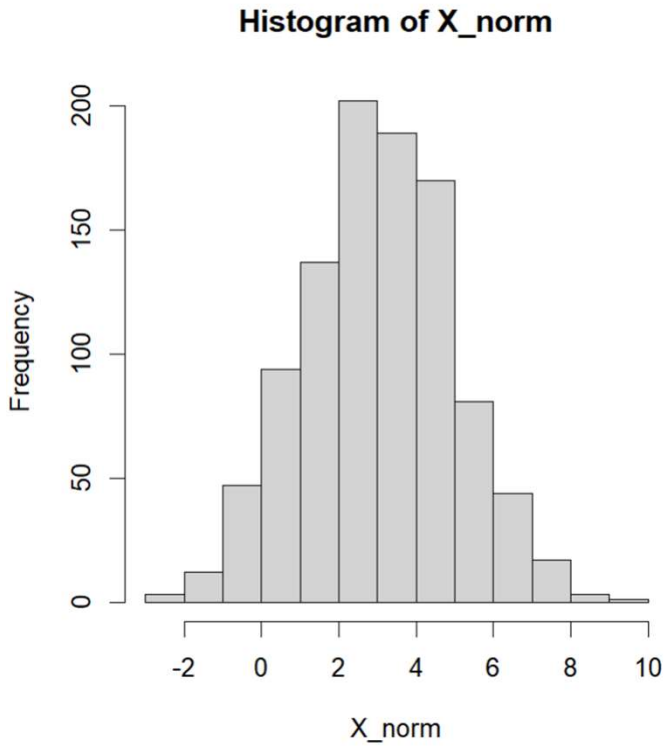
Calculate statistics

Console | Terminal | Background Jobs

R 4.2.1 · ~/

```
> X_norm <- rnorm(1000,mean=3,sd=2)
> hist(X_norm)
>
```

Draw a histogram



Why I switch from Python to R?

- Python might be too technical, especially for non-programmers.
- R is designed by statisticians, from statisticians, for everyone.
- Easy installation (for R).
- R is good for teaching, especially for statistical concepts.
- R is easier to use for non-programmers.
- R has massive packages for statistics and data visualization.

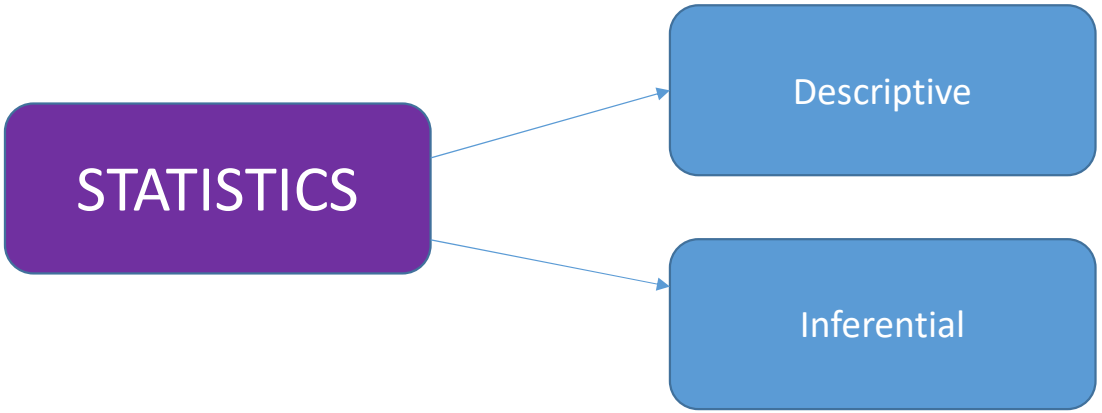
What this course will teach you

Theory

- Descriptive statistics
- Statistical hypothesis test.
- Basic statistical regression techniques.
- Data visualization and presentation.

Practice

BASICS and DESCRIPTIVE STATISTICS

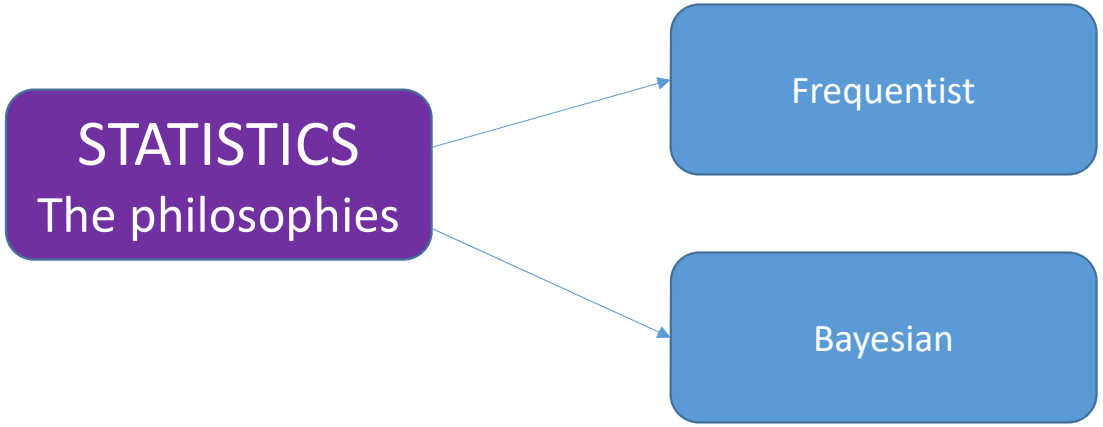


Summarize and describe your data

Example: what is the central tendency of my data?

Drawing conclusion of population from limited sample

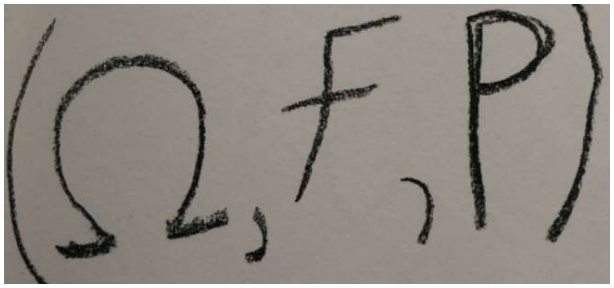
Example: From my samples, what does it tells us about the proportion of people with COVID-19 in Indonesia?



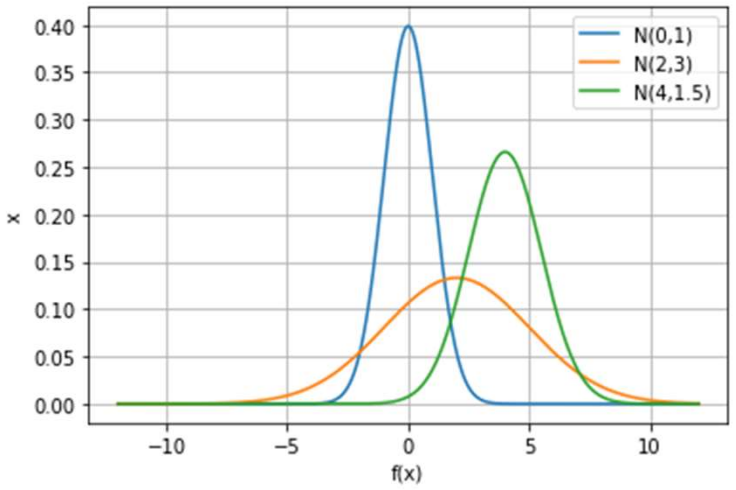
Interpret probabilities as a "probability of an event occurring"

View probabilities as a more general concept. Probabilities can be attached to any event.

Basic terminologies



- **Random variable:** a quantity having a numerical value for each member of a group, especially one whose values occur according to a frequency distribution
- **Probability:** The chance than an event will or will not occur.
- A **probability space** or a **probability triple** (Ω , F , P) is a mathematical construct that models a real-world process (or "experiment") consisting of states that occur randomly

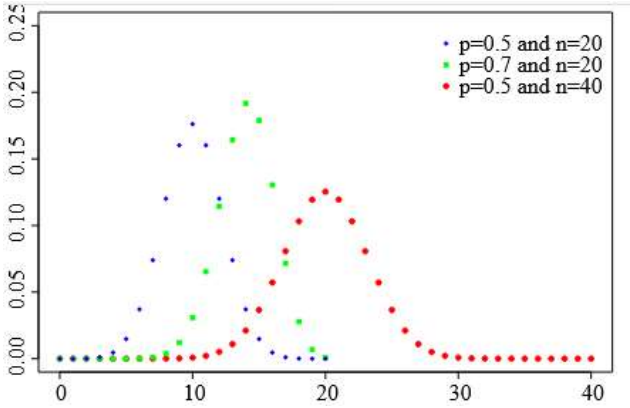


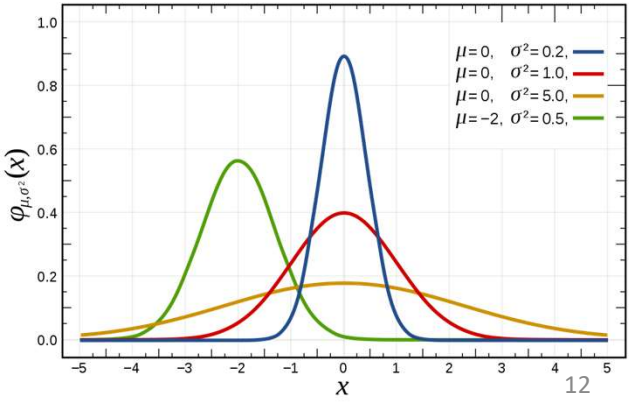
Ω = Sample space, a set of all possible outcomes.
F= Event, each event is a set containing zero or more outcomes(i.e., the subset of the sample space)
P= Probabilities of the event.

Probability distribution (1)

- Is a **mathematical function** that provides the **probabilities of occurrence** of different possible outcomes.
- Description of a **random phenomenon** in terms of the probabilities of event.
- Knowing the shape of the distribution is crucial to the use of **statistical methods** in research analysis!..
- .. since most methods make specific assumption about the distribution curve.

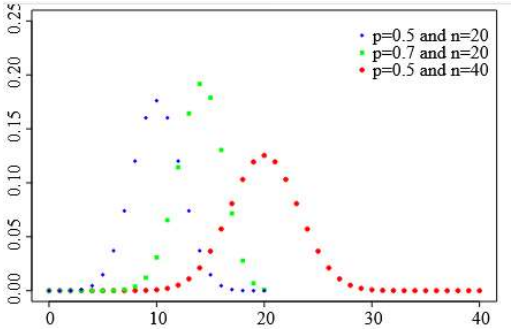
Example:
Binomial distribution
Normal distribution





Probability distribution (2)

- **Probability mass function** *P(X = x) = f(x)* of a discrete random variable *X* satisfies the following properties:



$$P(X = x) = f(x) > 0$$

The probability is non-zero if *x* belongs to the support S

$$\sum_{x \in S} f(x) = 1$$

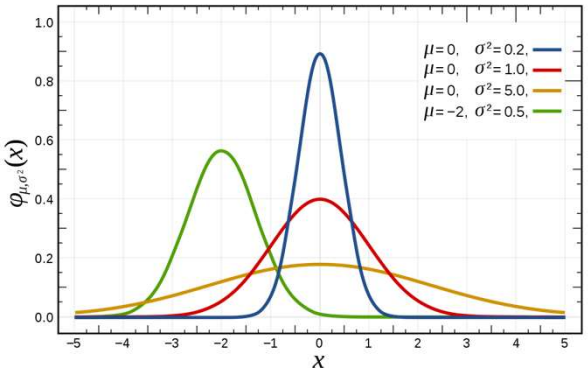
The sum of probabilities of all events equals 1

$$P(x \in A) = \sum_{x \in A} f(x)$$

The probability associated with several possible values is the sum of the probability of the independent events

Probability distribution (2)

- **Probability density function** of a continuous random variable *X* with support *S* is an integrable function *f*(x) satisfying the following:



$$f(x) > 0$$

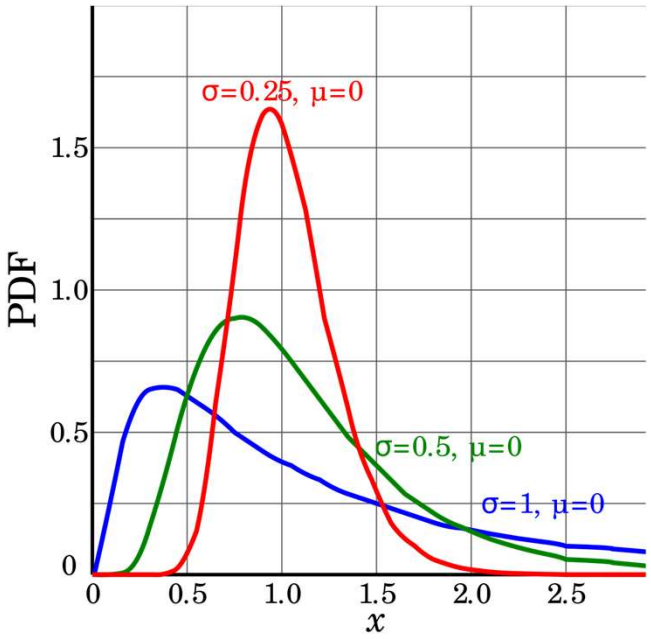
The probability is non-zero if *x* belongs to the support S

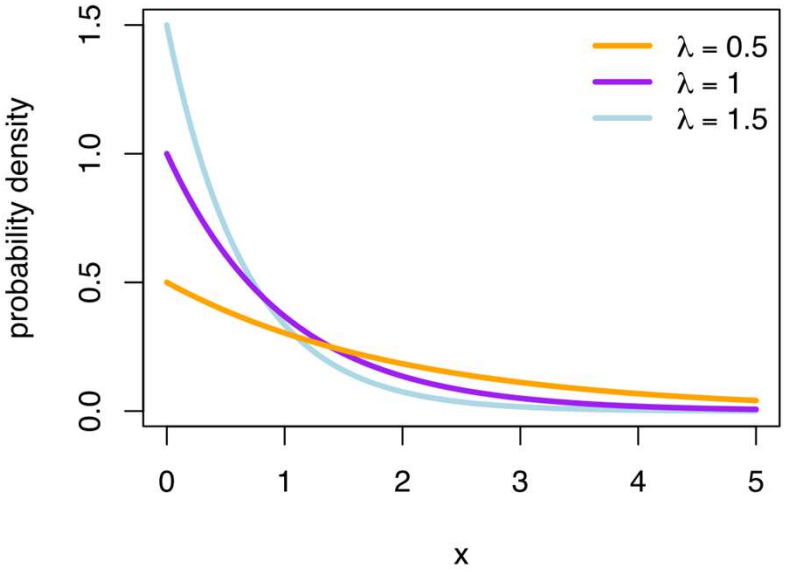
$$\int_S f(x)dx = 1$$

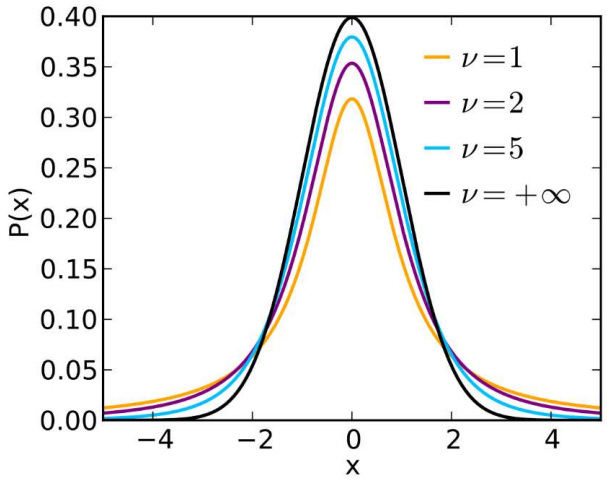
The sum (integral) of probabilities of all events equals 1

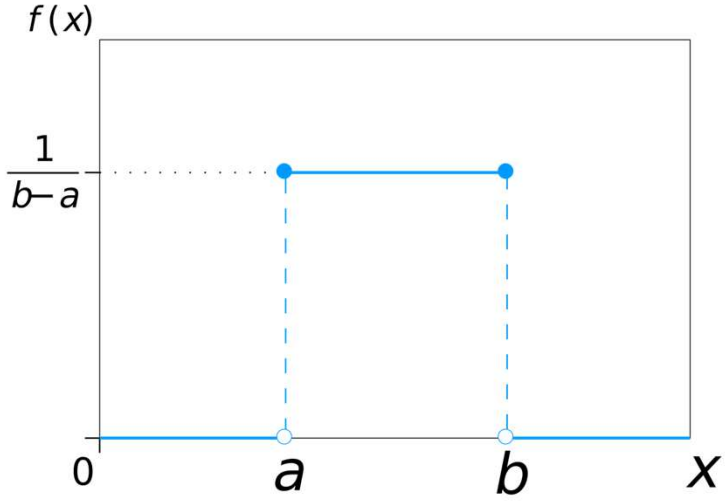
$$p(x \in A) = \int_A f(x)dx$$

The probability that *x* belongs to *A* (an interval) is given by the integral of the PDF over the interval


σ=0.25, μ=0
σ=0.5, μ=0
σ=1, μ=0
PDF
x


λ = 0.5
λ = 1
λ = 1.5
probability density
x


ν=1
ν=2
ν=5
ν=+∞
P(x)
x

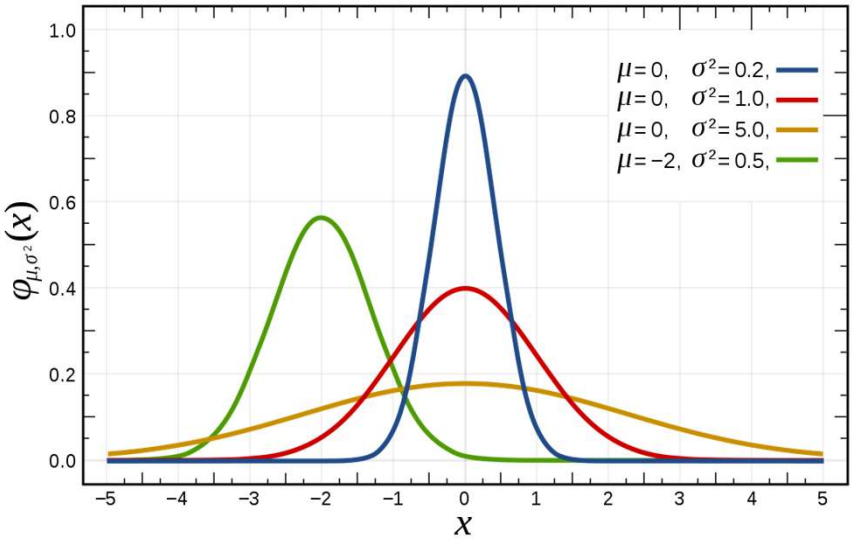

f(x)
1
b−a
0
a
b
x

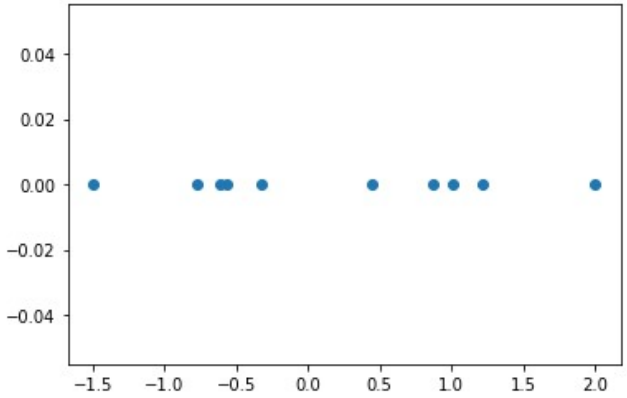
Normal distribution

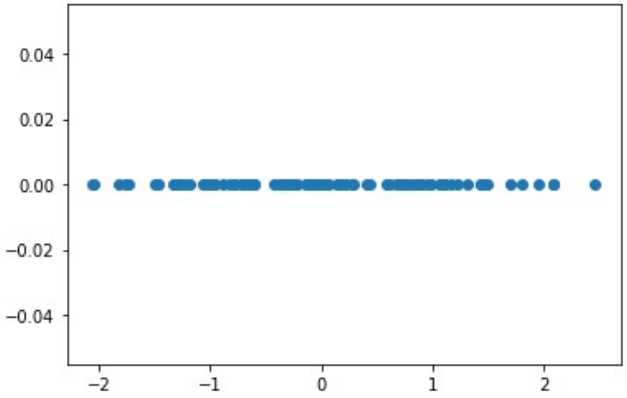
- **Normal distribution** is probably the most important continuous probability distribution.
- It has two parameters, i.e., **the mean** and **the standard deviation** of the distribution.
- It is widely used in statistical hypothesis tests, modeling of physical variables, etc.

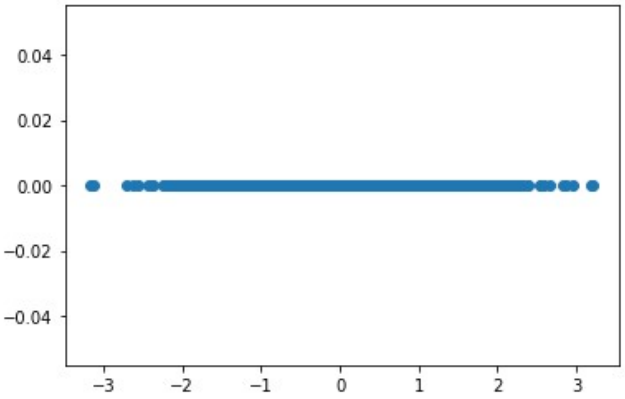
$$f(x) = \frac{1}{\sigma_n\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu_n}{\sigma_n}\right)^2}$$

$$\mathcal{N}(\mu_n, \sigma_n)$$





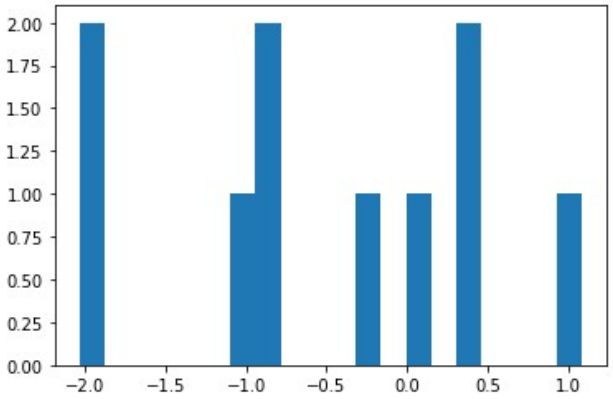


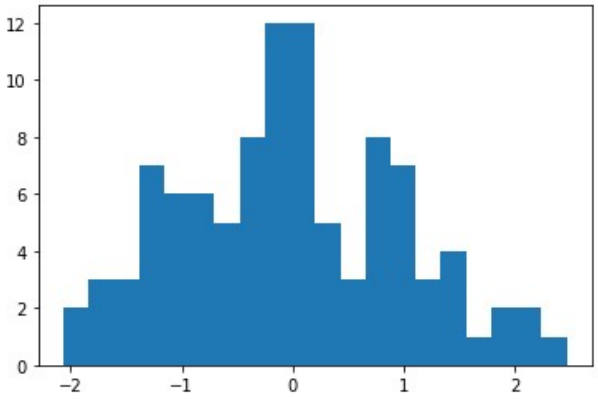


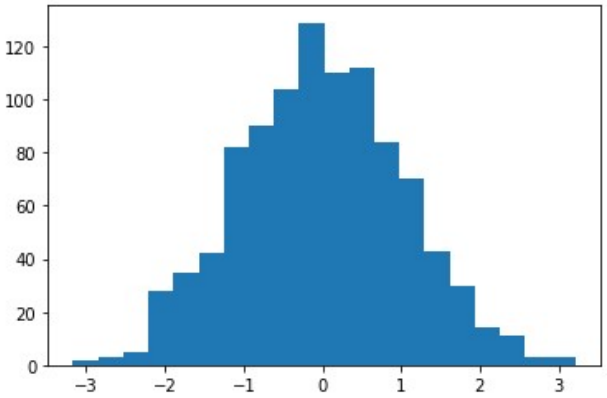
10 samples

100 samples

1000 samples







Descriptive statistics

The aim of **descriptive statistics** is to **describe and summarize your data with several coefficients**.

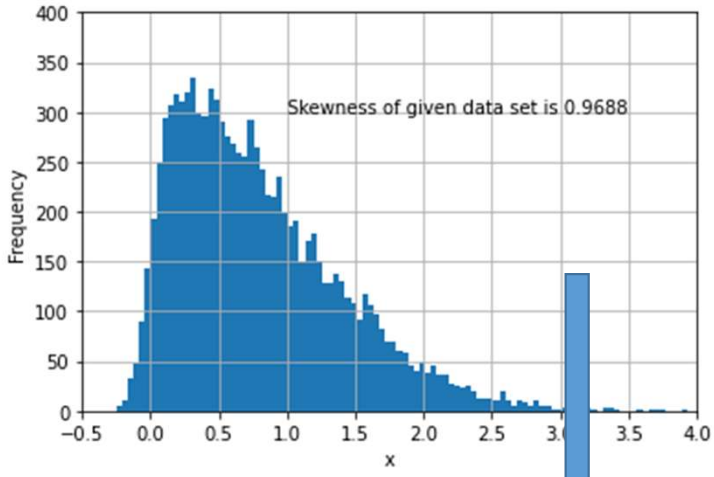
You want to answer questions such as:

- Where is the "center" of my data?
- How far is my data deviates from this center?
- What is the value that occurs most often?

Measures of central tendency

Measures of variability

Measures of correlation



Mean? Std. Dev? Skew? Mode?

Insight!

BASICS and DESCRIPTIVE STATISTICS

Measures of central tendency

Measures of central tendency

- **Measures of central tendency** describe the tendency of where the "center" of your data is located.
- Knowing your measures of central tendency helps you in knowing where the "average" of your data is.

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

(Arithmetic) Mean

$$\text{Value of}\left(\frac{n+1}{2}\right)\text{th item}$$

Median

Value that occurs most often

Mode

Be careful on using these measures, because they might mislead you from the true behavior of the data; thus, use them wisely.

Measures of central tendency

- **Measures of central tendency** describe the tendency of where the "center" of your data is located.
- Knowing your measures of central tendency helps you in knowing where the "average" of your data is.

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

(Arithmetic) Mean

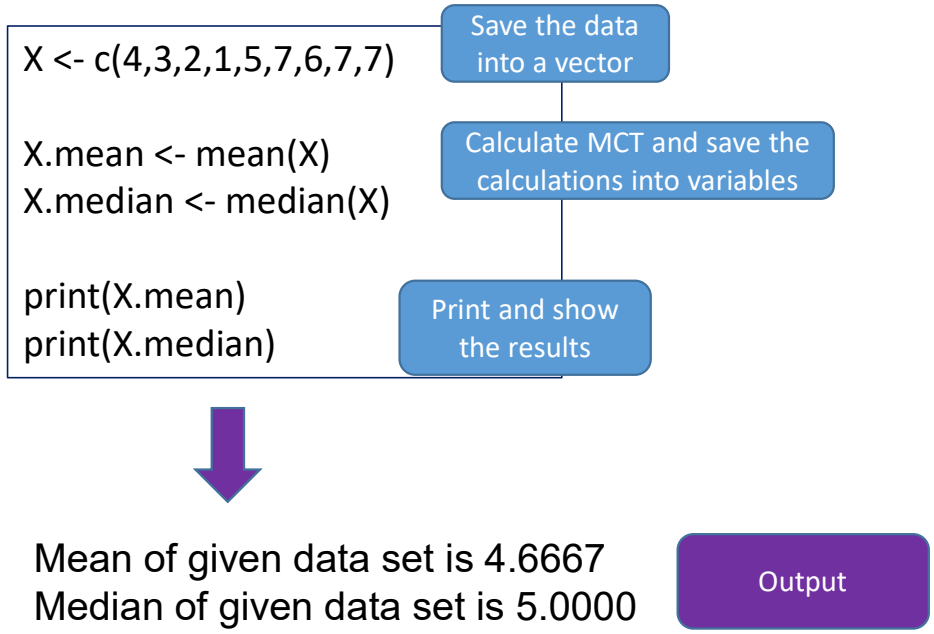
$$\text{Value of}\left(\frac{n+1}{2}\right)\text{th item}$$

Median

Value that occurs most often

Mode

Measures of central tendency: R implementation



Case of mean and median

$$10, 20, 30, 10, 50, 10, 70, 50, 80, 60, 100, 100, 5000$$

$$10, 10, 10, 20, 30, 50, \boxed{50}, 60, 70, 80, 100, 100, 5000$$

Mean and median can give you different impression of the data.

Such differences might be due to:

- The presence of outliers.
- Incorrect / wrong sampling / other errors.
- Highly skewed data

In practice, it is better to observe both the mean and median value.

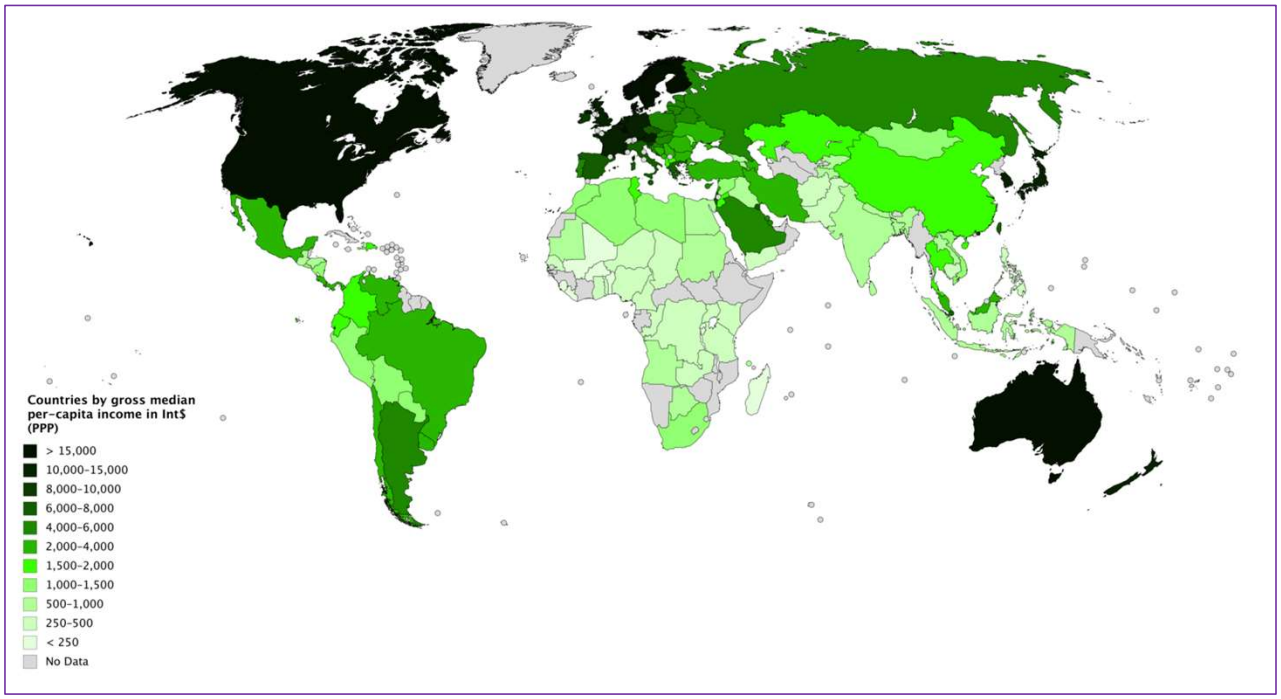
Mean = 430

Median = 50

An example of when median is more useful than mean

- In 2013, Gallup published a list of countries with median annual household income, based on a self-reported survey of approximately 2000 adults from each country.

- Using median, rather than mean income, results in a much more accurate picture of the typical income of the middle class since the data will not be skewed by gains and abnormalities in the extreme ends.



Histogram

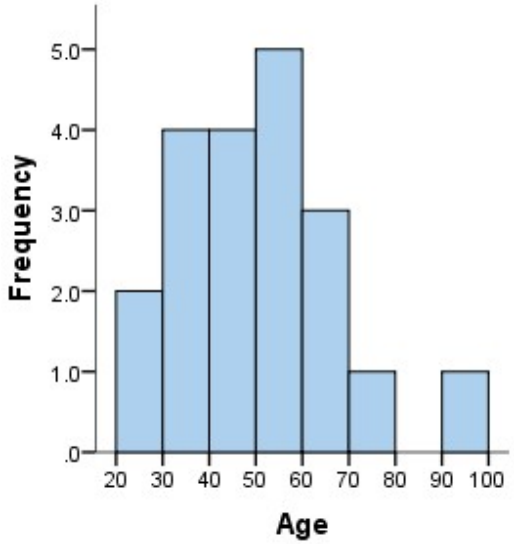
- A histogram is a plot that lets you discover, and show, the **underlying frequency distribution** (shape) of a set of data. This allows the inspection of the data for its underlying distribution (e.g., normal distributions), outliers, skewness, etc.
- **Histogram is particularly useful if you want to visualize the general distribution of your data.**

Source: https://statistics.laerd.com/statistical-guides/understanding-histograms.php

Data

36	25	38	46	55	68	72	55	36	38
67	45	22	48	91	46	52	61	58	55

Histogram

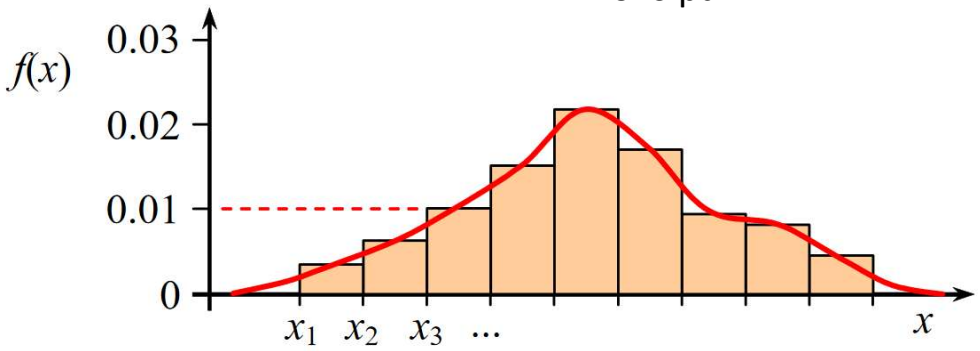


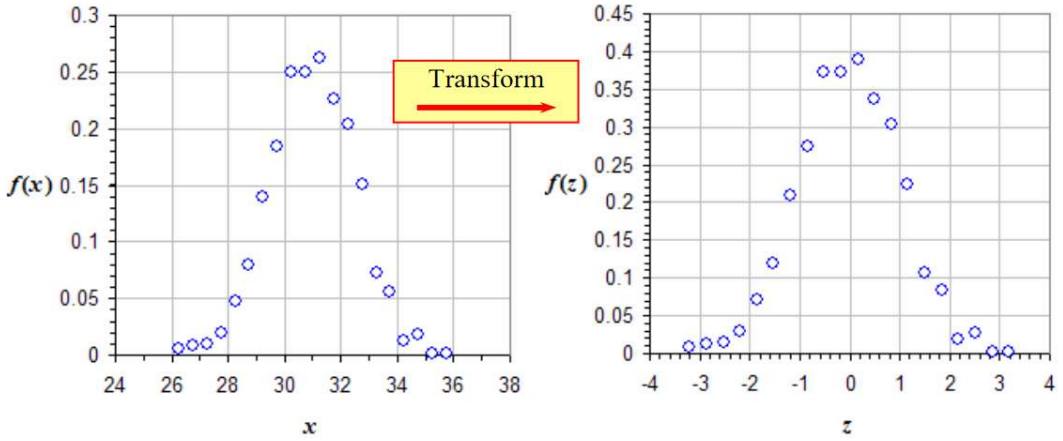
Bin	Frequency	Scores Included in Bin
20-30	2	25,22
30-40	4	36,38,36,38
40-50	4	46,45,48,46
50-60	5	55,55,52,58,55
60-70	3	68,67,61
70-80	1	72
80-90	0	-
90-100	1	91

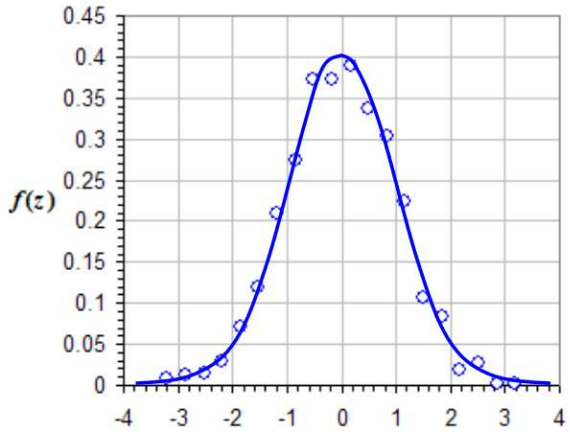
Histogram vs probability distribution

https://www.me.psu.edu/cimbala/me345/Lectures/Probability_density_functions.pdf

- The *y*-axis in PDF is **PDF**, the x-axis in histogram is **frequency**.
- Histogram is **discrete**, PDF is **continuous**
- PDF is the **smooth limit** of a normalized histogram.



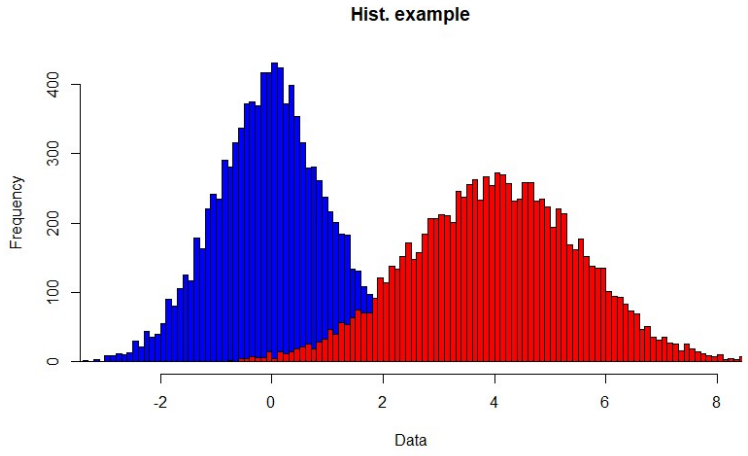




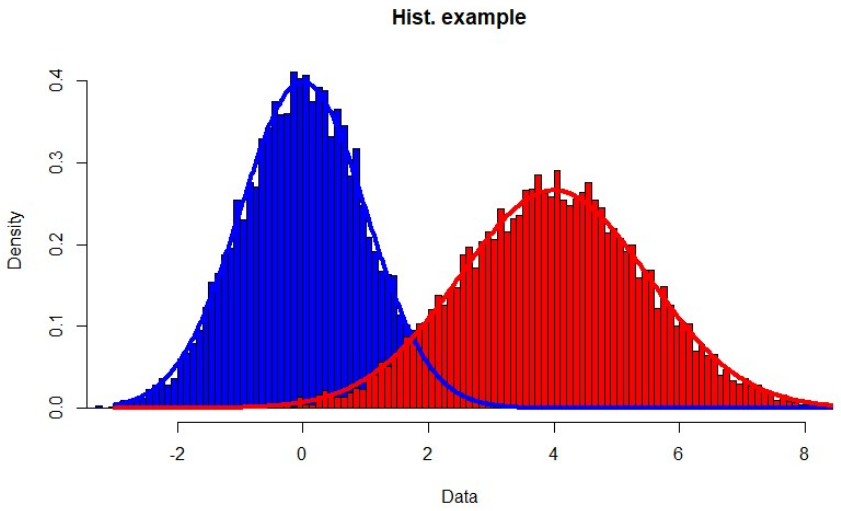
Histogram approximates your PDF when you have large samples (and normalized)

Histogram: R implementation

```r
X_1 <- rnorm(10000,mean=0,sd=1)
X_2 <- rnorm(10000,mean=4,sd=1.5)

hist(X_1,breaks=100,xlim=c(-3,8),col="blue",
     xlab="Data",main="Hist. example")
hist(X_2,breaks=100,add=TRUE,col="red")
```




```r
X_1 <- rnorm(10000,mean=0,sd=1)
X_2 <- rnorm(10000,mean=4,sd=1.5)

hist(X_1,breaks=100,xlim=c(-3,8),col="blue",
     xlab="Data",main="Hist. example",prob=TRUE)
hist(X_2,breaks=100,add=TRUE,col="red",prob=TRUE)

X_p <- seq(-3,9,by=0.02)
y_p1 <- dnorm(X_p,mean=0,sd=1)
y_p2 <- dnorm(X_p,mean=4,sd=1.5)

lines(X_p,y_p1,type = "l",col="blue",lwd=4)
lines(X_p,y_p2,type = "l",col="red",lwd=4)
```



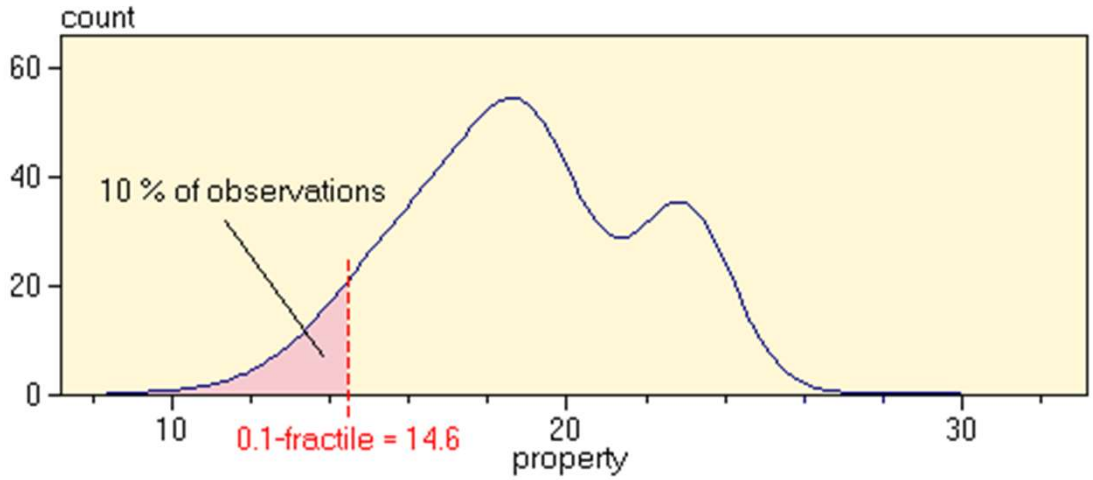
BASICS and DESCRIPTIVE STATISTICS

quantile

Quantile

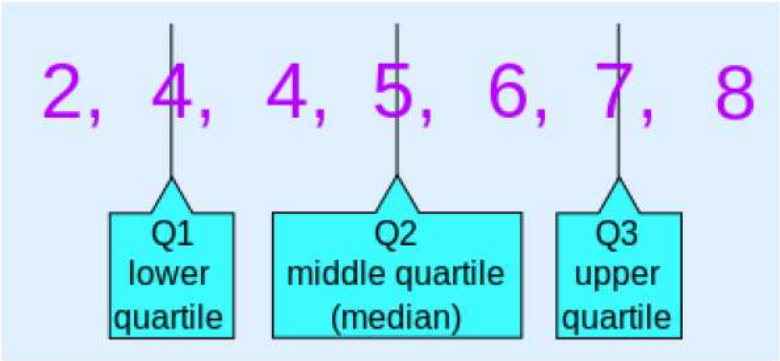
- **By a quantile**, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.
- For example, if the 0.8 quantile of the examination score of subject A is 40, this means that 80% of the students obtain scores lower than 40.

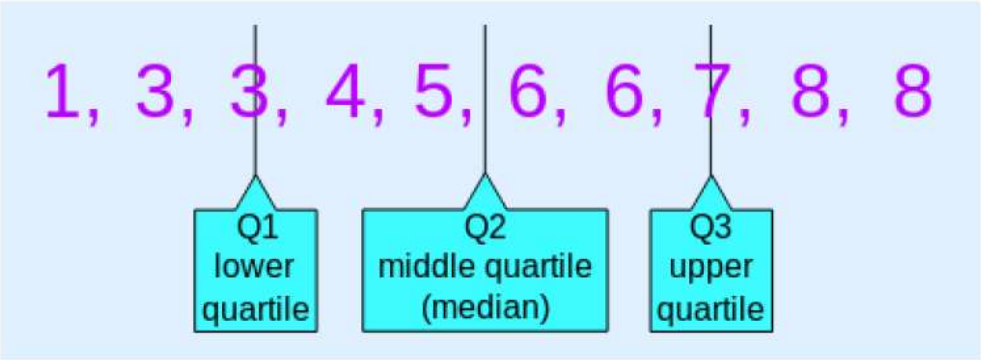
- The 1000-quantiles are called permillages.
- **The 100-quantiles are called percentiles.**
- The 20-quantiles are called vingtiles.
- The 12-quantiles are called duo-deciles.
- The 10-quantiles are called deciles
- The 9-quantiles are called nonile.
- The 5-quantiles are called quintiles.
- **The 4-quantiles are called quartiles.**



Quartile: the 4-quantiles

Quartile is one form of quantile. **Quartiles** are the values that divide a list of numbers into quarters:





BASICS and DESCRIPTIVE STATISTICS

Measures of variability

Measures of variability

- **Measures of variability** gives you the information regarding how far your data deviates from the center:
- Knowing your measures of variability helps you in knowing the spread of your data.

$$\sigma(X) = \sqrt{\frac{1}{n}\sum(X_i - \bar{X})^2}$$

Standard deviation

$$\text{Range}(X) = \max(X) - \min(X)$$

Range

$$\text{IQR}(X) = Q3 - Q1$$

Interquartile range

Be careful on using these measures too, since they might mislead you from the true behavior of the data; thus use them wisely.

Measures of variability: range

Range is defined as the **difference between the values of the extreme items of a series**.

- Pros: It gives an idea of the variability very quickly.
- Cons: The range is affected very greatly by fluctuation of sampling. Its value is never stable.

It is typically only used as a rough measure of variability.

$$\text{Range}(X) = \max(X) - \min(X)$$

$$\text{Range} = \begin{pmatrix}\text{Highest value of an}\\ \text{item in a series}\end{pmatrix} - \begin{pmatrix}\text{Lowest value of an}\\ \text{item in a series}\end{pmatrix}$$

Measures of variability: range

$$\text{Range}(X) = \max(X) - \min(X)$$

Your data

42 20 79 95 13 3 84 70 67 70 74 39 65
17 70

Sorted data

Median

Min 3 13 17 20 39 42 65 67 70 70 70 74
79 84 95 Max

Range = 95-3 = 92

Measures of variability: range

$$\text{Range}(X) = \max(X) - \min(X)$$

$$10, 20, 30, 10, 50, 10, 70, 50, 80, 60, 100, 100, 5000$$

$$10, 10, 10, 20, 30, 50, 50, 60, 70, 80, 100, 100, 5000$$

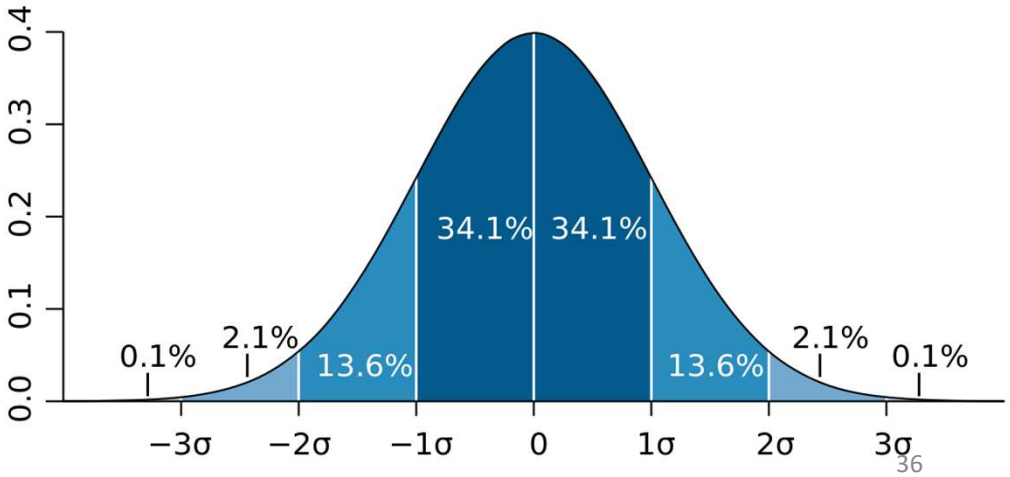
Range = 5000-10 = 4990

This is an example where range might give you wrong impression. However, range can also be used to detect the 'sanity' of your data.

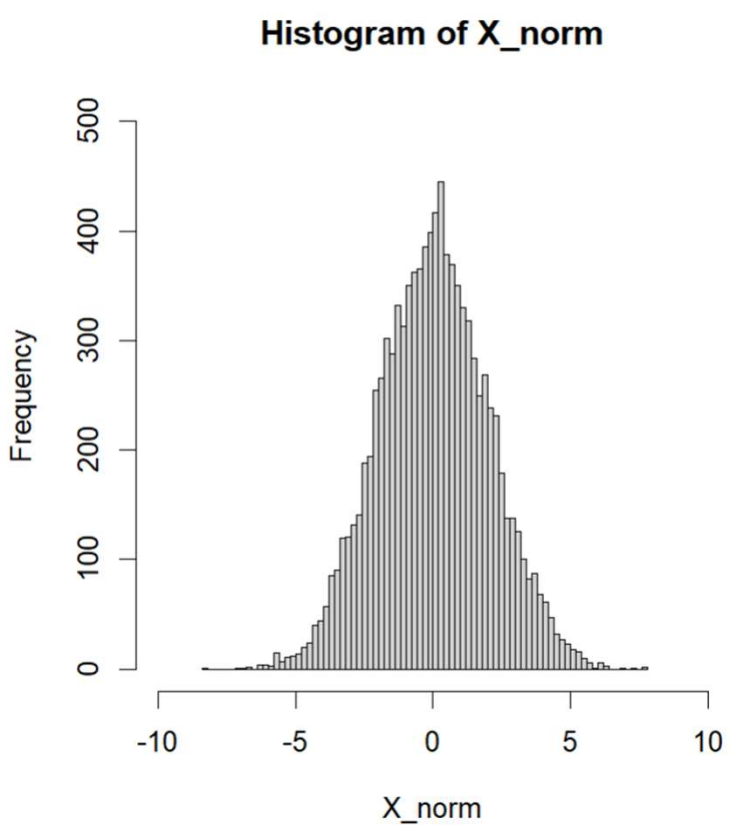
Measures of variability: standard deviation

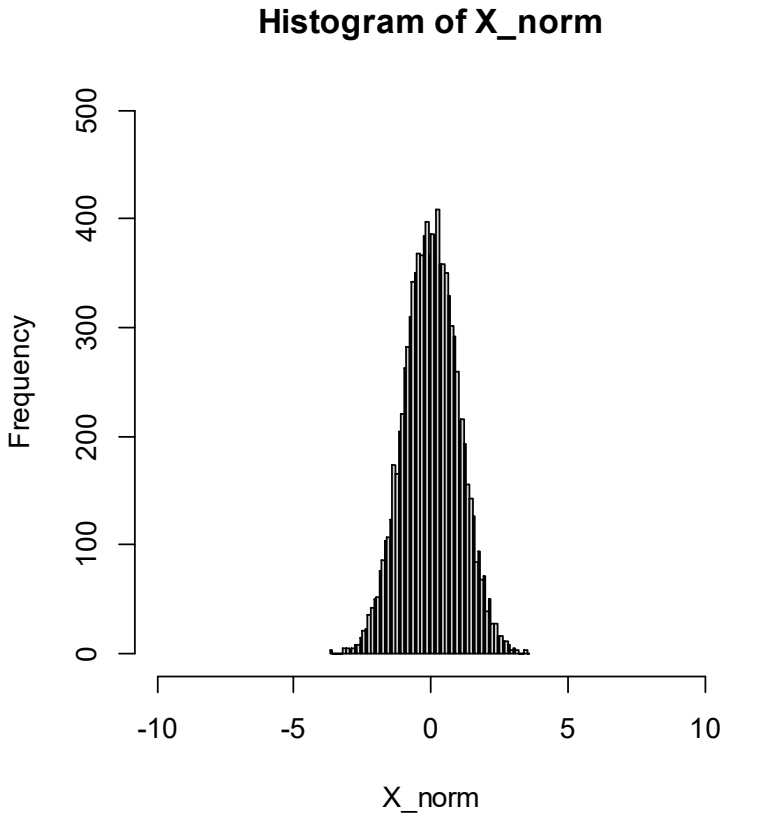
- Standard deviation (σ) is the most widely used measure of dispersion of a series.
- We can also use variance (σ^2), i.e., , mathematical dispersion of the data relative to the mean.
- Standard deviation is more useful to describe the data according to our common sense, while variance is more useful in mathematical derivation.

$$\sigma(X) = \sqrt{\frac{1}{n}\sum(X_i - \bar{X})^2}$$



Measures of variability: standard deviation





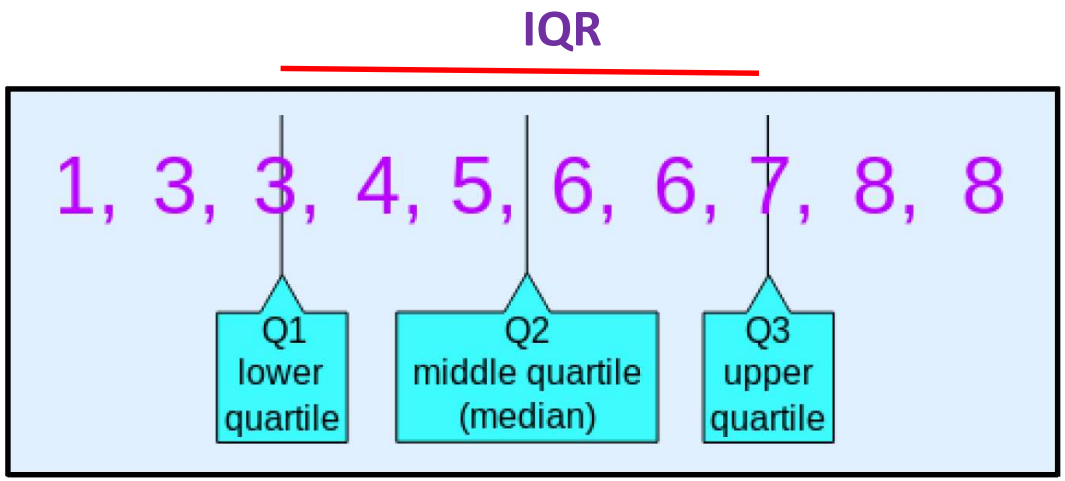

```
> X_norm <- rnorm(10000,mean=0,sd=2)
> hist(X_norm,breaks=100,xlim=c(-10,10),ylim=c(0,500))
```

```
R 4.2.1 · ~/
> X_norm <- rnorm(10000,mean=0,sd=1)
> hist(X_norm,breaks=100,xlim=c(-10,10),ylim=c(0,500))
>
```

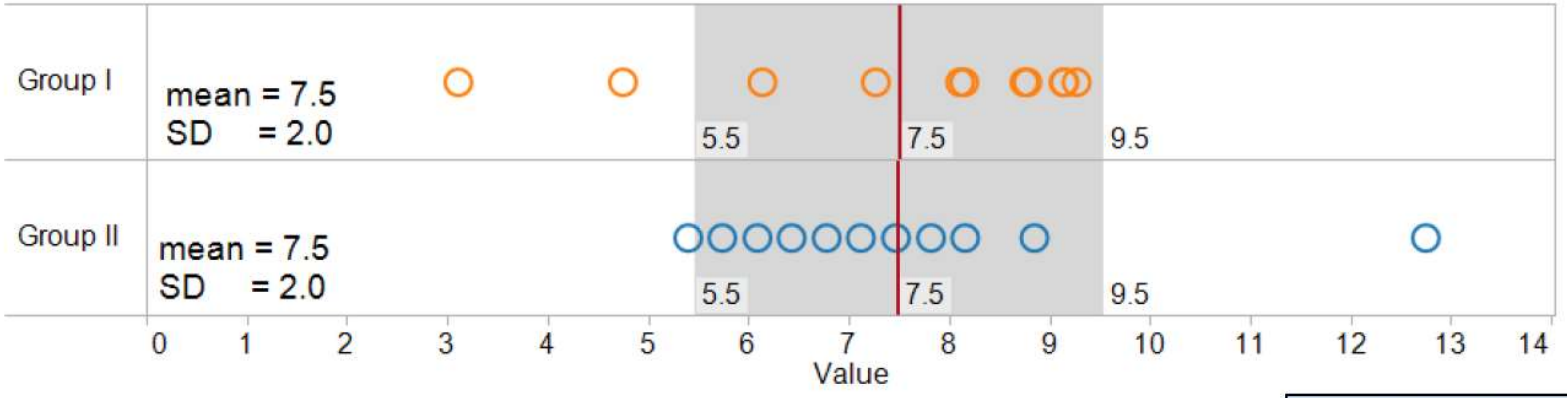
Measures of variability: IQR

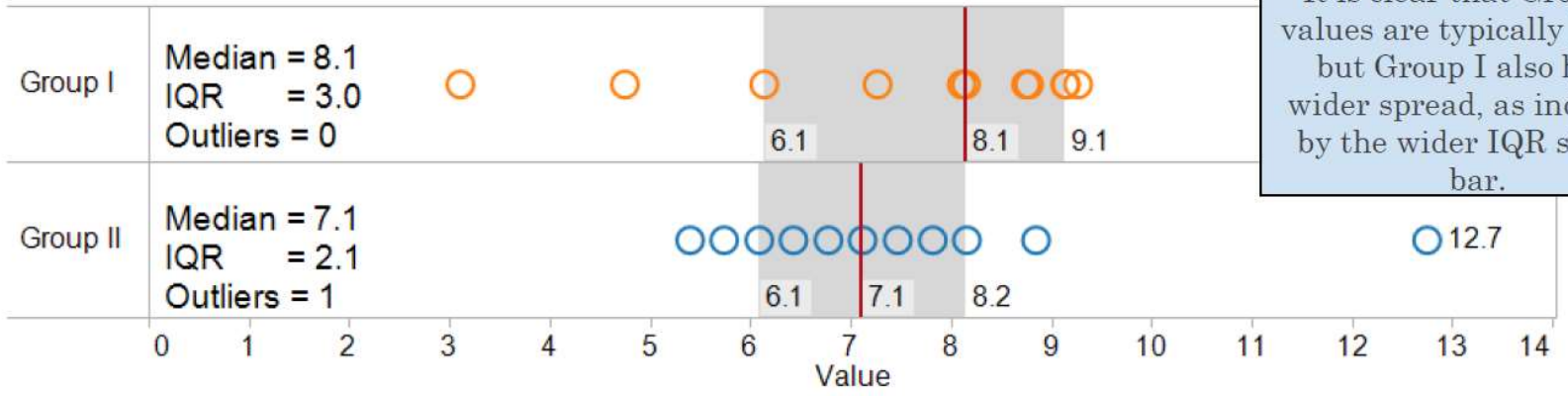
- The **Interquartile Range(IQR)** may be used to characterize the data when there may be extremities that skew the data.
- The interquartile range is a relatively **robust statistic** (also sometimes called "resistance") compared to the rangeandstandard deviation.

$$IQR = Q3\text{-}Q1$$



IQR vs standard deviation

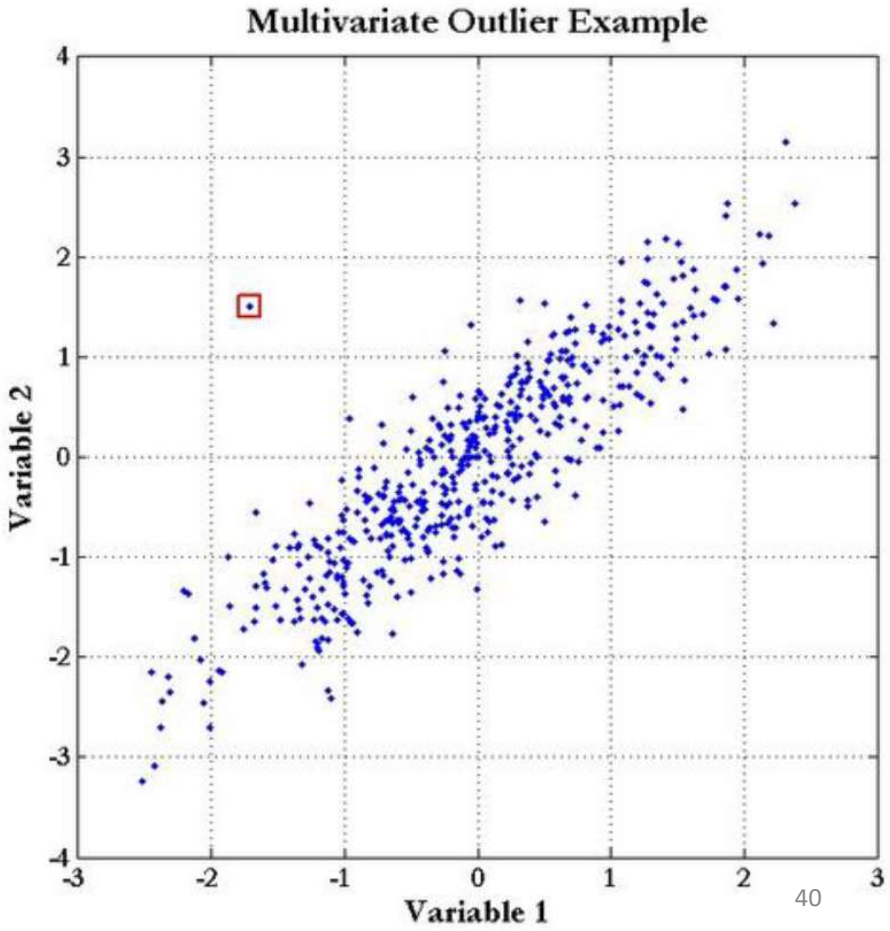




It is clear that Group I's values are typically higher, but Group I also has a wider spread, as indicated by the wider IQR shaded bar.

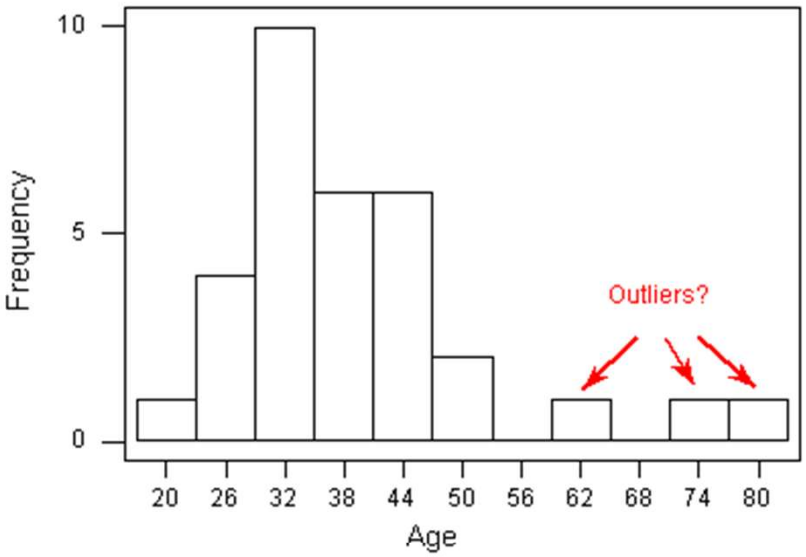
Application of IQR : outlier detection

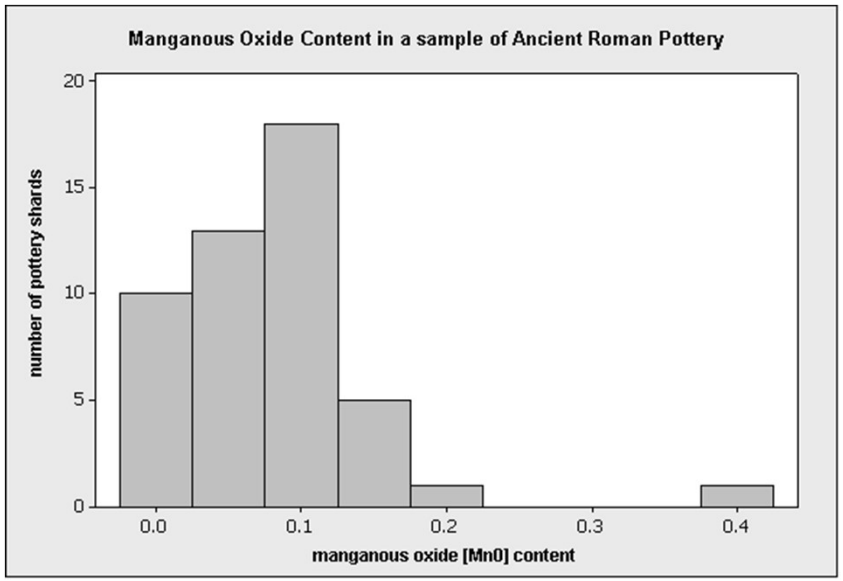
- A data point on a graph or in a set of results that is very much bigger or smaller than the next nearest data point.
- Outlier can cause serious problems in statistical analyses (depending on the context)
- Outliers could mean two things: (1) Measurement error (2) Heavy tailed distribution.
- **Now the question is: how to detect outliers?**

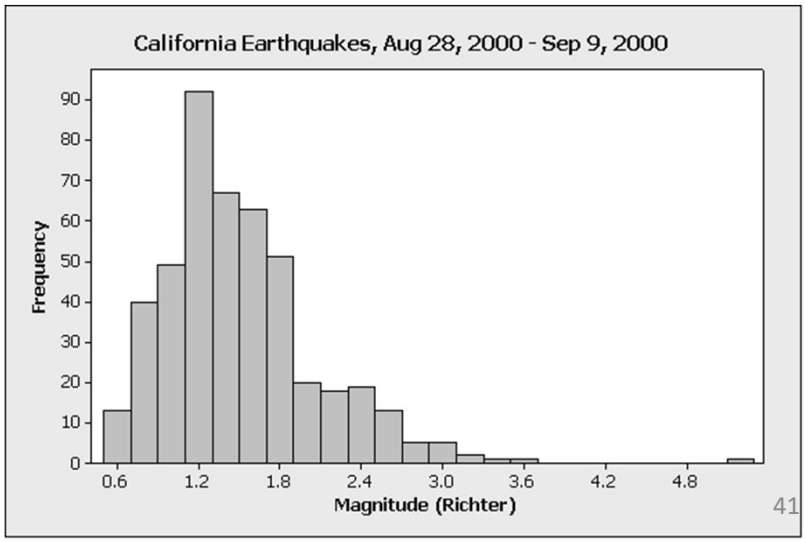


Why outliers occur?

- Experimental error
- Wrong sampling
- Heavy-tailed distribution
- Errors were made on data entry







Application of IQR : outlier detection

- There is no one universally accepted outlier detection method.
- There are some useful methods, one of them is **Tukey's fences.**
- Tukey's fences treat data as outliers when they are outside this range

$$[Q_1 - k(Q_3 - Q_1), Q_3 + k(Q_3 - Q_1)]$$

- If k = 1.5, data outside this range is treated as "outlier".
- If *k* = 3, data outside this range is indicated as "far-out".
- We'll also illustrate this later by using a boxplot.

Application of IQR : outlier detection

$$[Q_1 - k(Q_3 - Q_1), Q_3 + k(Q_3 - Q_1)]$$

- Q_1 = 10
- Q_3 = 30
- **IQR = (Q_3-Q_1) = 20**

Data set

21	22	33	10	8	9	10	4
30	20	30	10	20	17	8	9
10	22	11	2	30	20	22	
33	34	37	21	20	21	23	
11	21	13	8	6	9	100	90
30	70						

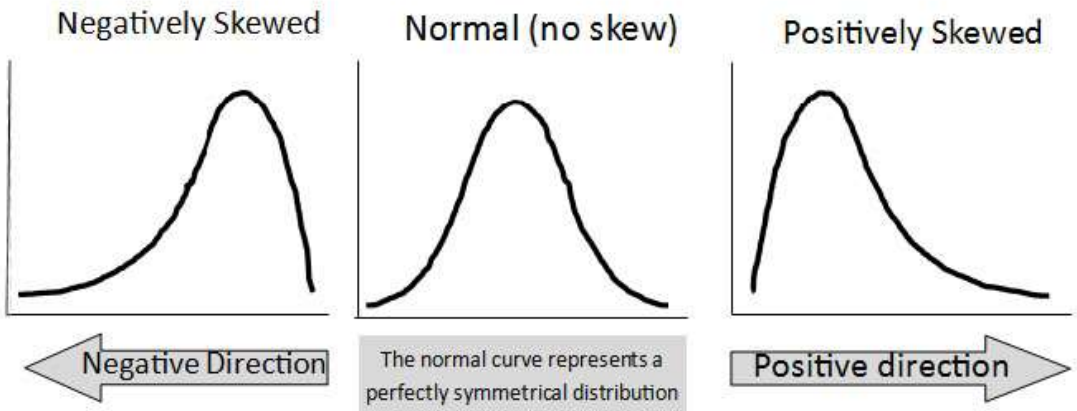
- Lower bound for outlier detection with k= 1.5 is -20
- Upper bound for outlier detection with k= 1.5 60.
- **Thus, the people who score 100, 90 and 70 are outliers.**

BASICS and DESCRIPTIVE STATISTICS

Measures of asymmetry: Skewness

Measures of asymmetry: skewness

- Skewness indicates distribution that is asymmetric.
- If the curve is distorted on the right side, we have positive skewness.
- When the curve is distorted toward left, we have negative skewness.
- If the distribution is perfectly symmetric, the skewness is zero.



$$\text{Skewness} = \frac{3 * (\text{Mean-Median})}{\text{Standard deviation}}$$

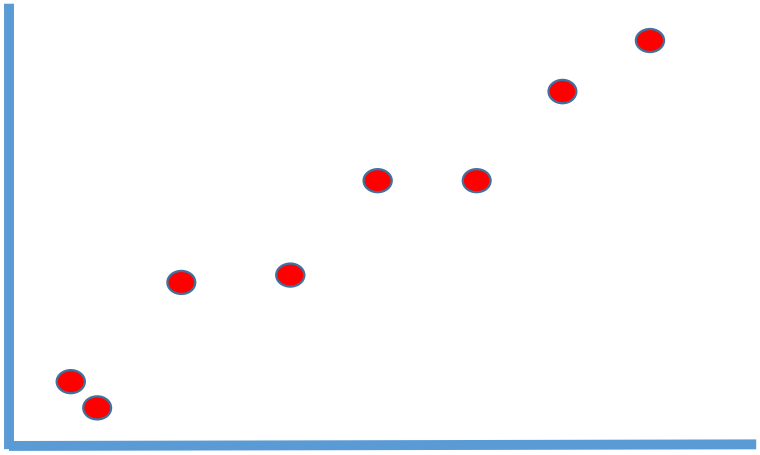
Negative | Zero (or almost) | Positive

BASICS and DESCRIPTIVE STATISTICS

Measures of relationship

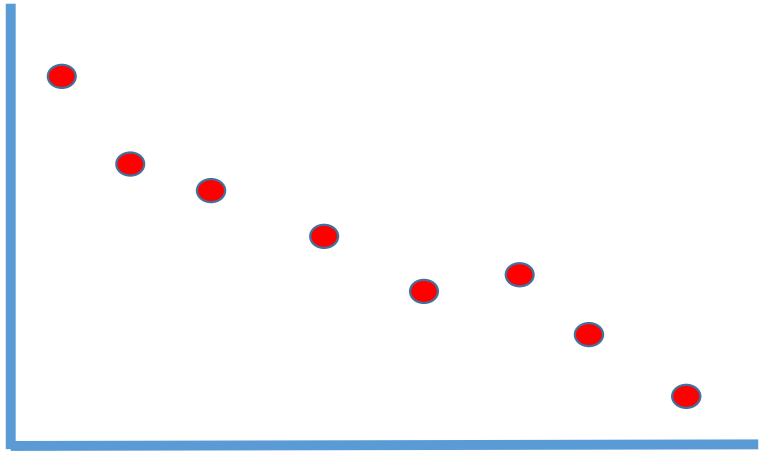
Quiz: How will you describe the following relationship

Stock market returns



Economic Growth

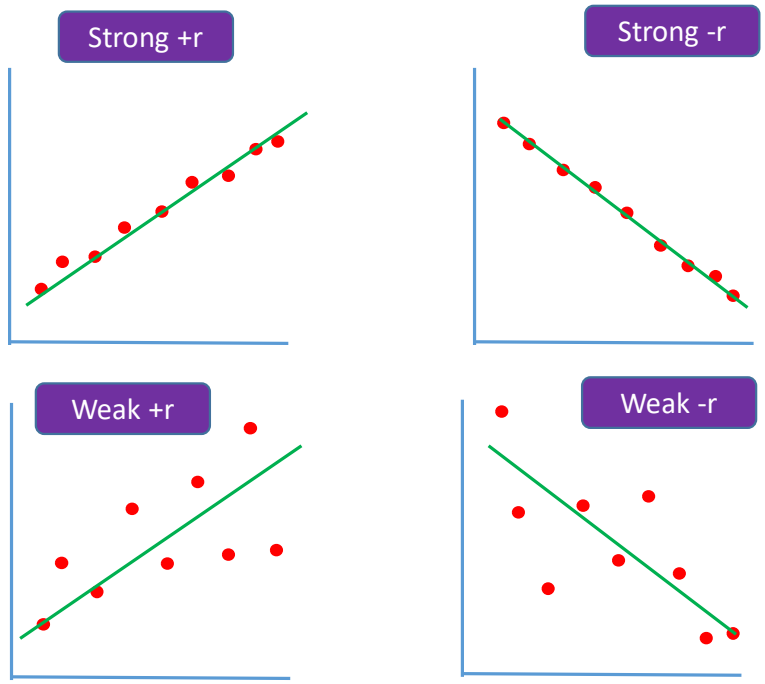
Gasoline prices



World oil production

Correlation and dependence

- **Statistical dependence or association** describes statistical relationship between two random variables.
- The relationship does not indicate causality, **but describes correlation**
- It is a powerful relationship when you want to analyze the relationship between two variables.



Pearson correlation coefficient

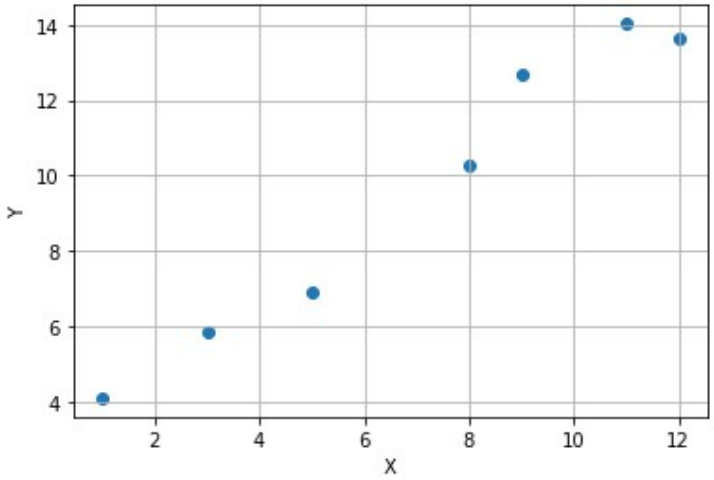
Definition

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

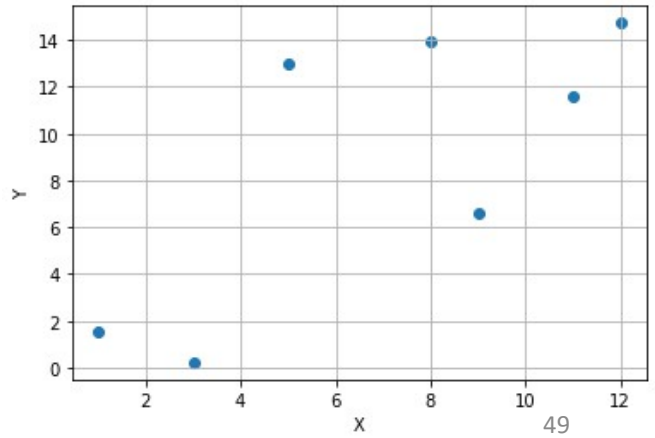
Sample implementation

$$\rho_{X,Y} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

Pearson = 0.9846

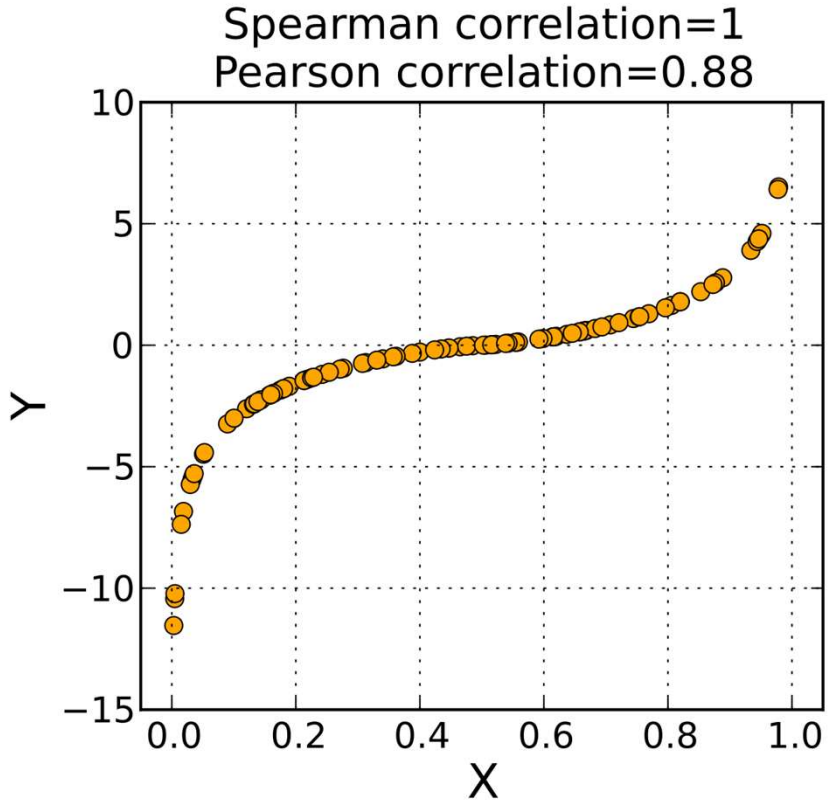


Pearson = 0.7447



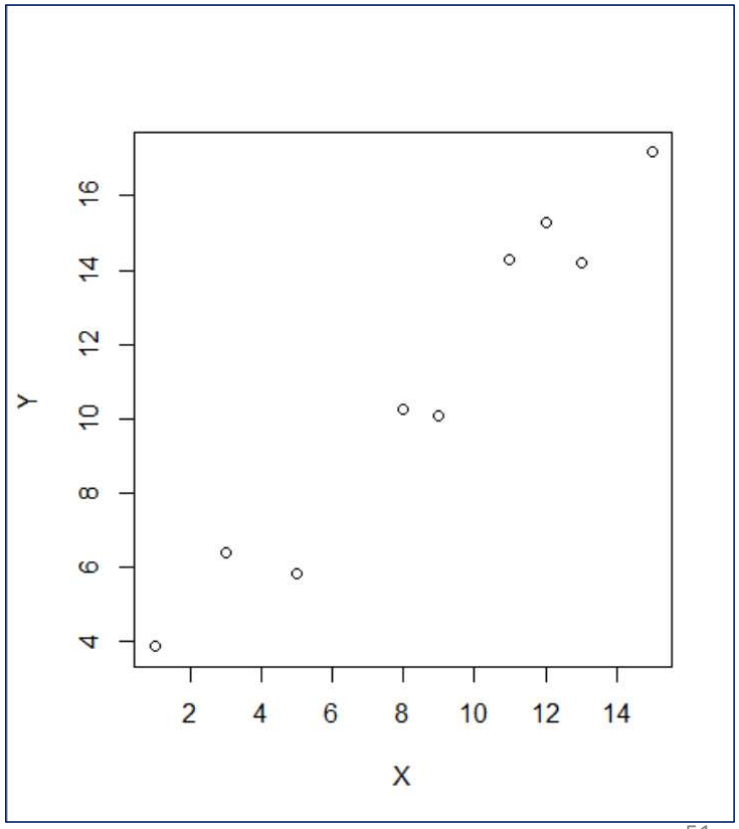
Spearman correlation coefficient

- In contrast to Pearson, Spearman identifies **monotonicity** instead of nonlinearity.
- The Spearman correlation between two variables is equal to the Pearson correlation between the rank values of those two variables



Correlation coefficient in R (example 1)

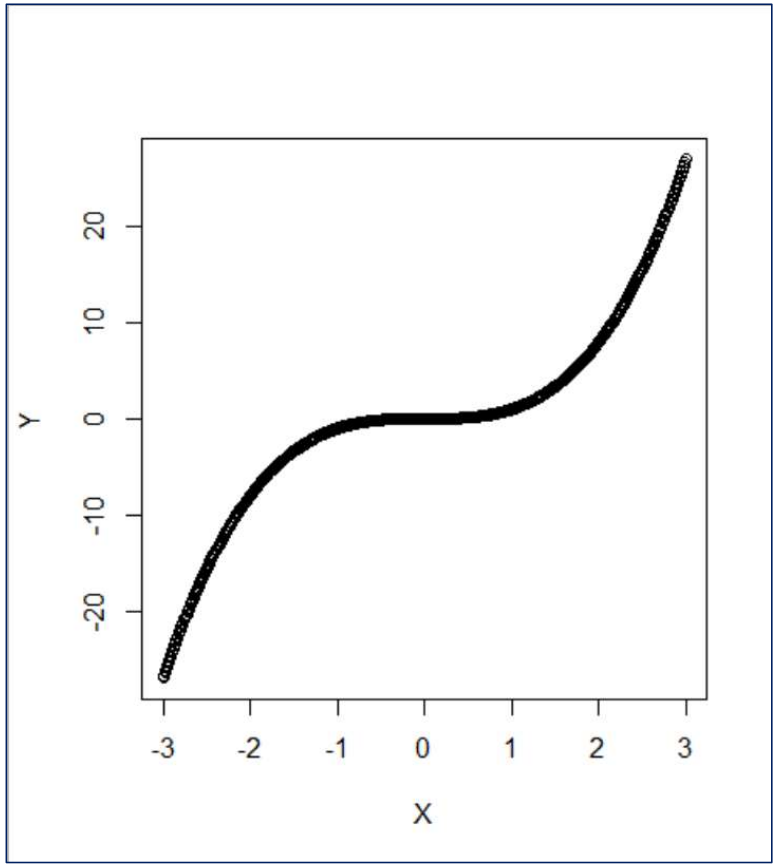
```r
X <- c(1,3,5,8,9,11,12,13,15)
Y <- X + 2 + rnorm(length(X),mean=0,sd=1) # Add random noise

plot(X,Y)
cor(X,Y) # Calculate Pearson corr. coef.
```



Correlation coefficient in R (example 2)

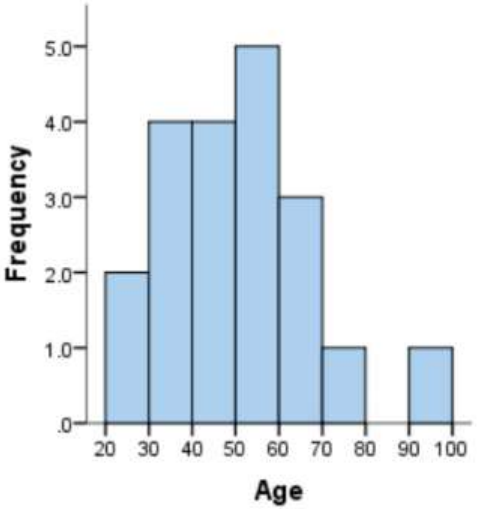
```r
X <- seq(-3,3,by=0.01)
Y <- X^3

plot(X,Y)

cor(X,Y,method="pearson")
cor(X,Y,method="spearman")
cor(X,Y,method="kendall")
```

Pearson = 0.916
Spearman = 1
Kendall = 1



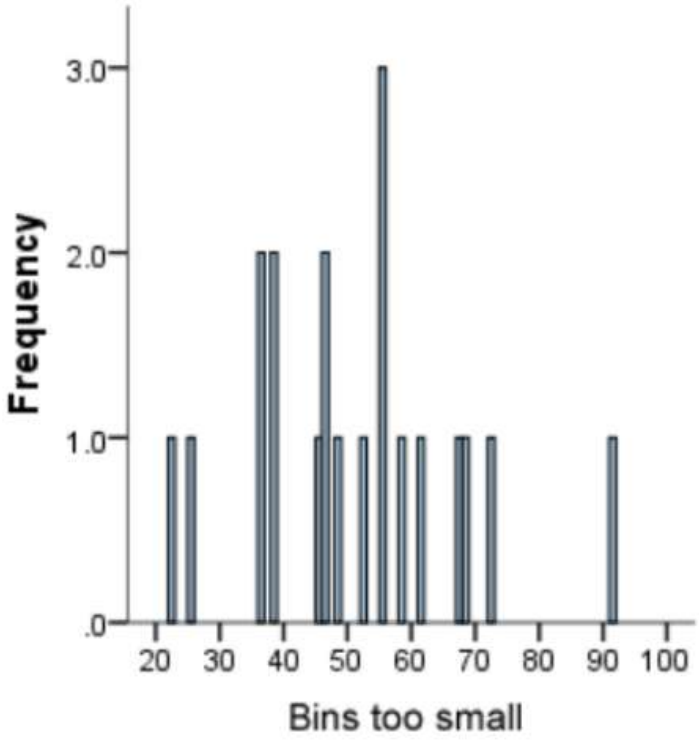
Presenting your results

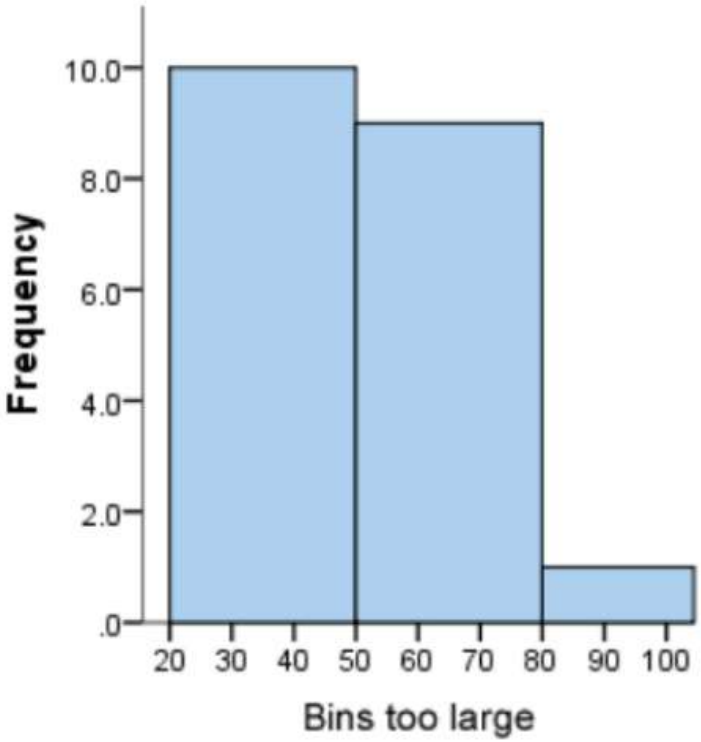
- A histogram is a plot that lets you discover, and show, the underlying frequency distribution (shape) of a set of data. This allows the inspection of the data for its underlying distribution (e.g., normal distribution), outliers, skewness, etc.
- Histogram is very useful if you want to visualize the general distribution of your data



Bin	Frequency	Scores Included in Bin
20-30	2	25,22
30-40	4	36,38,36,38
40-50	4	46,45,48,46
50-60	5	55,55,52,58,55
60-70	3	68,67,61
70-80	1	72
80-90	0	-
90-100	1	91

36	25	38	46	55	68	72	55	36	38
67	45	22	48	91	46	52	61	58	55





Too much:

- The general pattern is difficult to see.
- Shows too much individual data.

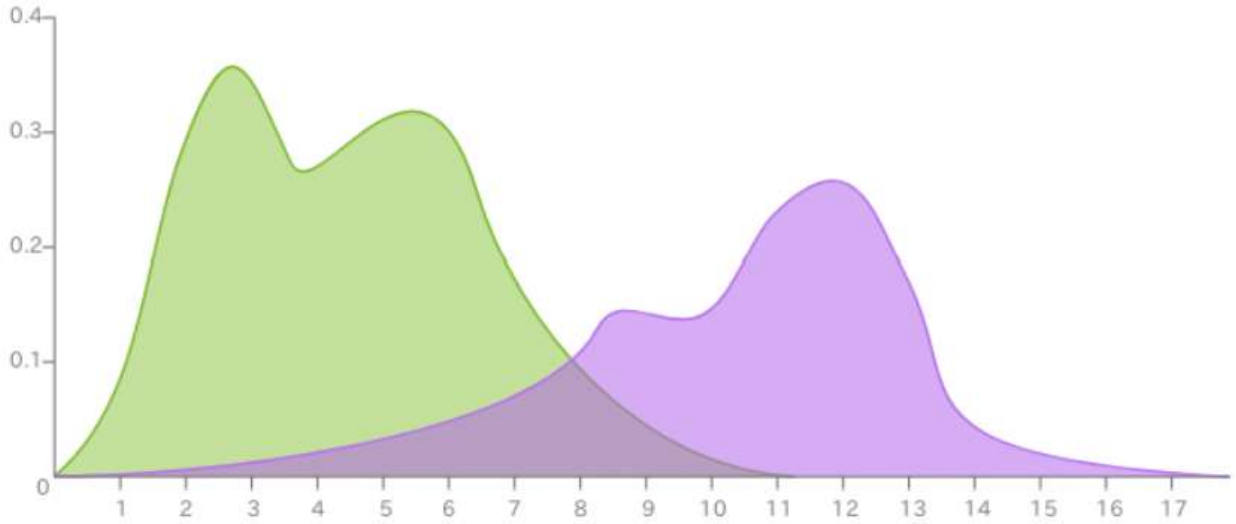
Too few:

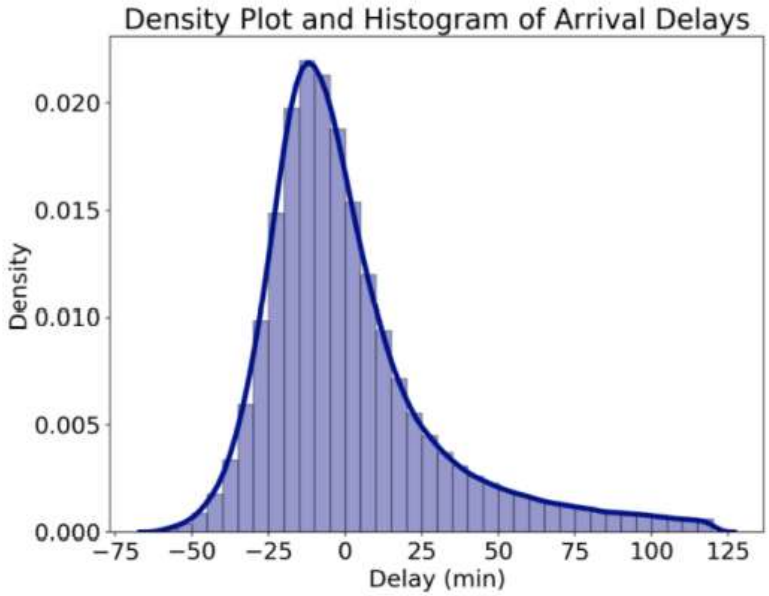
- Might give us incomplete impression of the data

.

Density plots as alternative to histogram

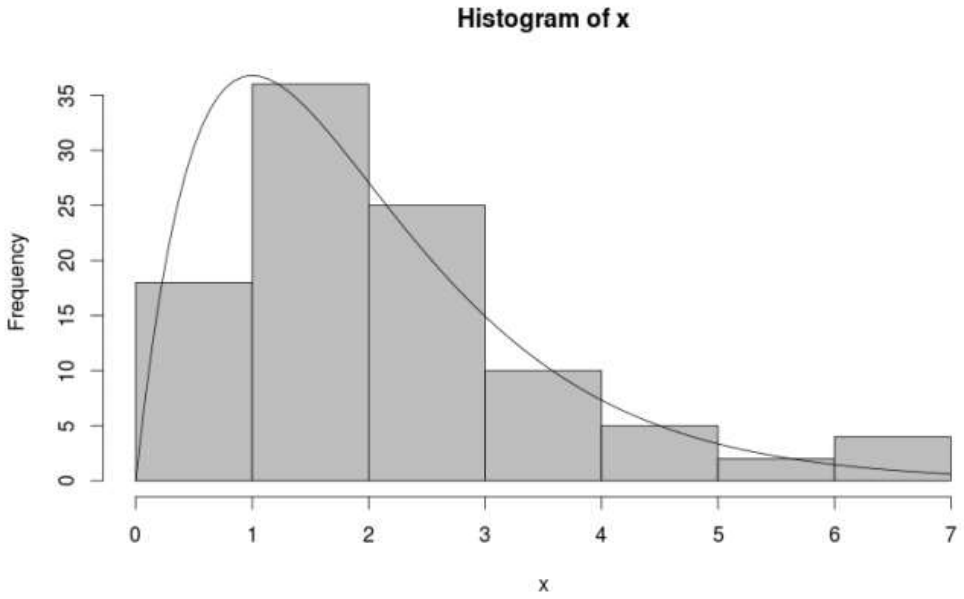
- A Density Plot visualises the distribution of data over a continuous interval or time period.
- Basically, it is a smoothed histogram via kernel smoothing.
- It gives us a clearer picture regarding the shape of the distribution.

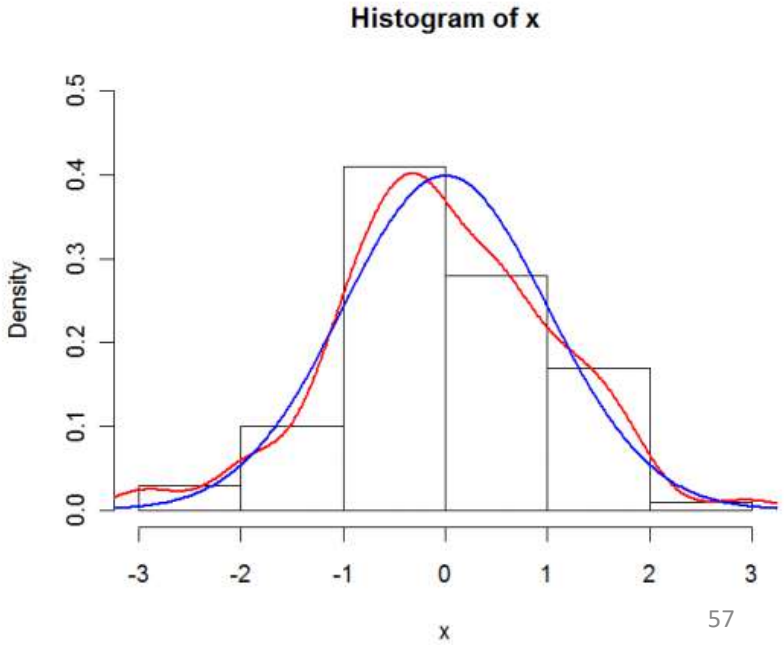




Histogram vs PDF plot

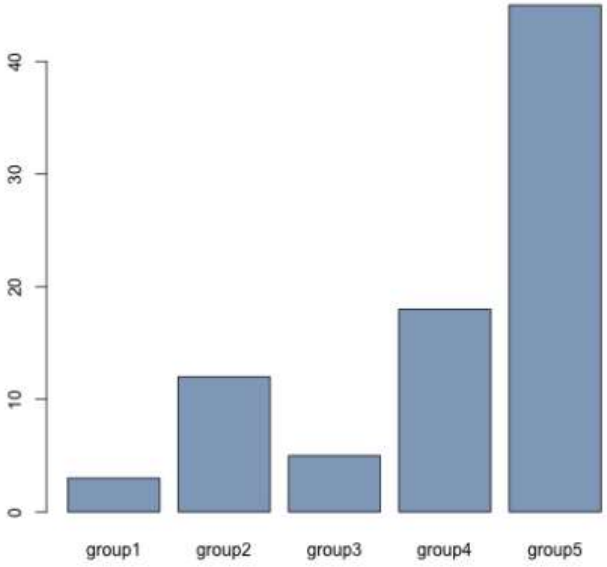
- Histogram is the empirical distribution of your sample.
- PDF is used if you want to describe the hypothesized underlying distribution (normal? exponential?)
- Histogram is usually shown with frequencies as the *y*-axis, while for PDF density is shown (note that the right figure is only for illustrative purpose).

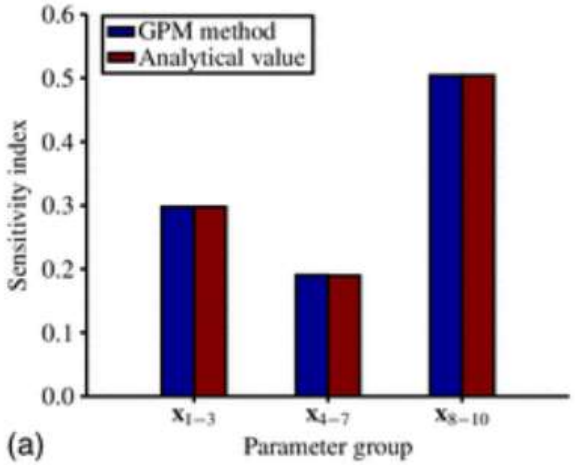


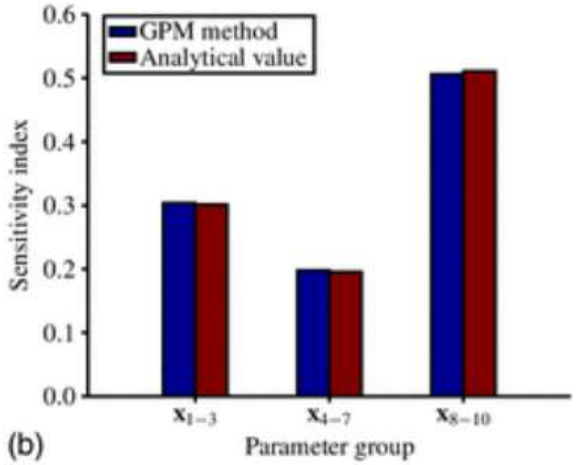


Presenting your results: Barplot

- The major difference is that a histogram is only used to plot the **frequency of score occurrences** in a continuous data set that has been divided into classes, called bins. Bar charts, on the other hand, can be used for a great deal of other types of variables.
- Barplot is very useful is you visualize such data as sales, sensitivity indices, to name a few.

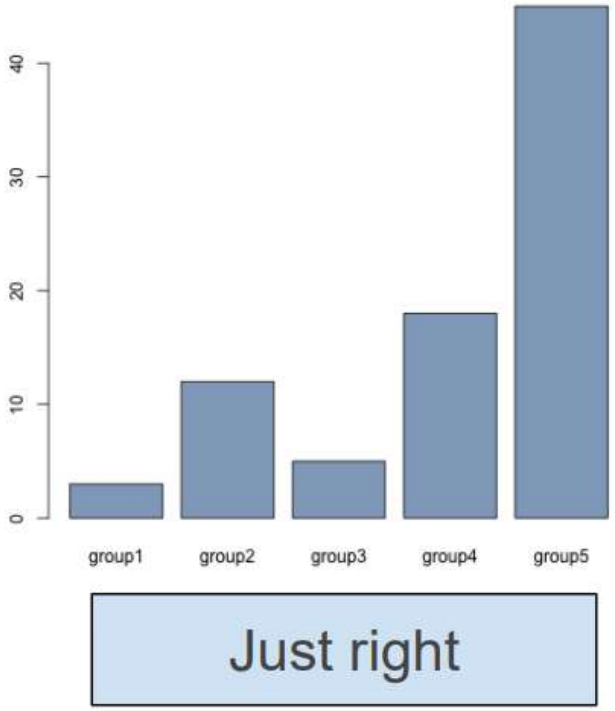


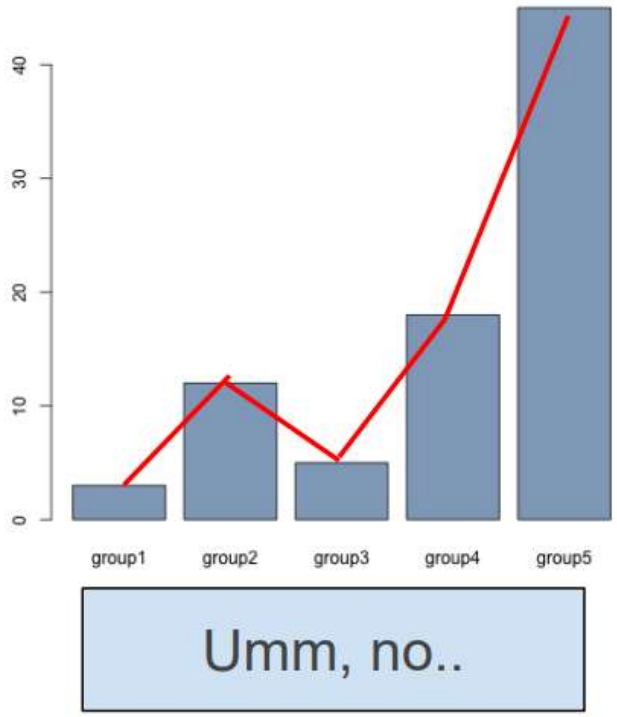




Presenting your results: Barplot

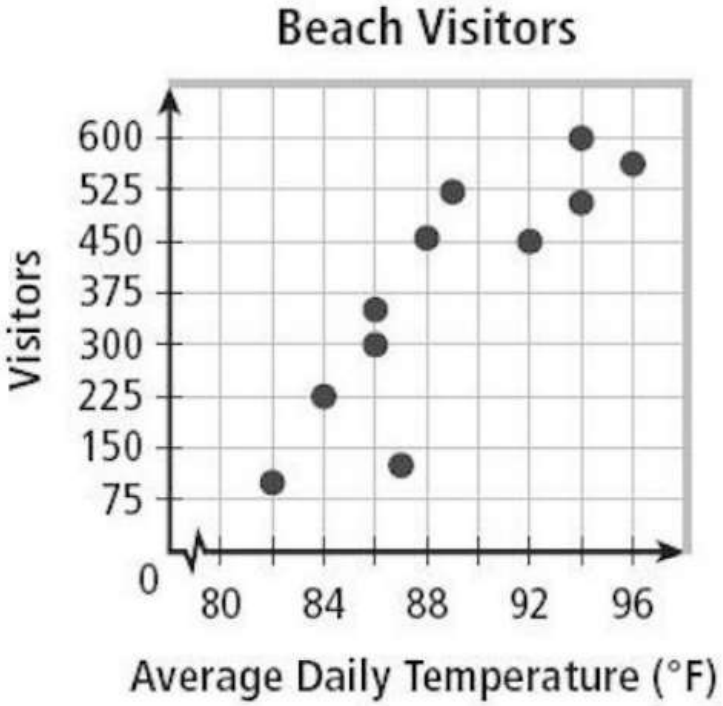
- One common mistake of using continuous plot: using continuous plot when the data should actually visualized with barplot (i.e., when the x-axis is categorical variable).

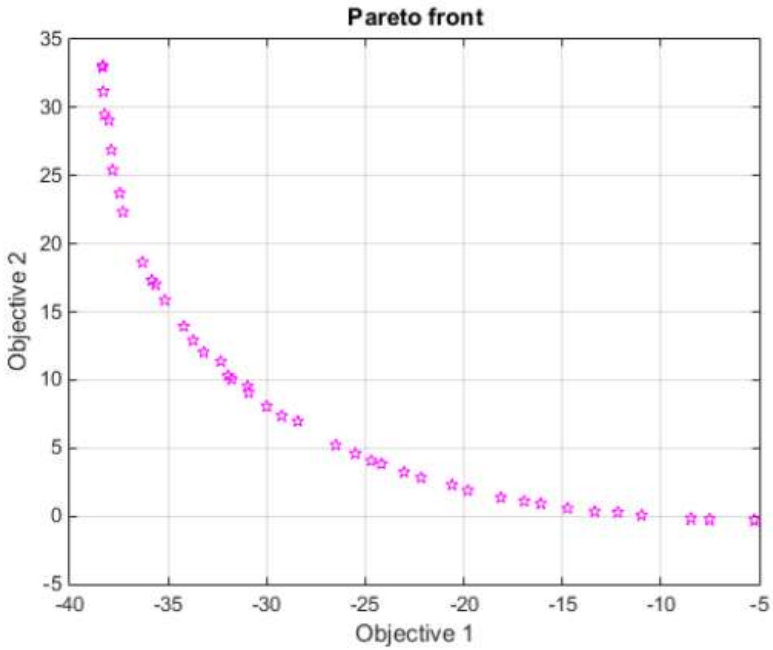




Presenting your results: Scatterplot

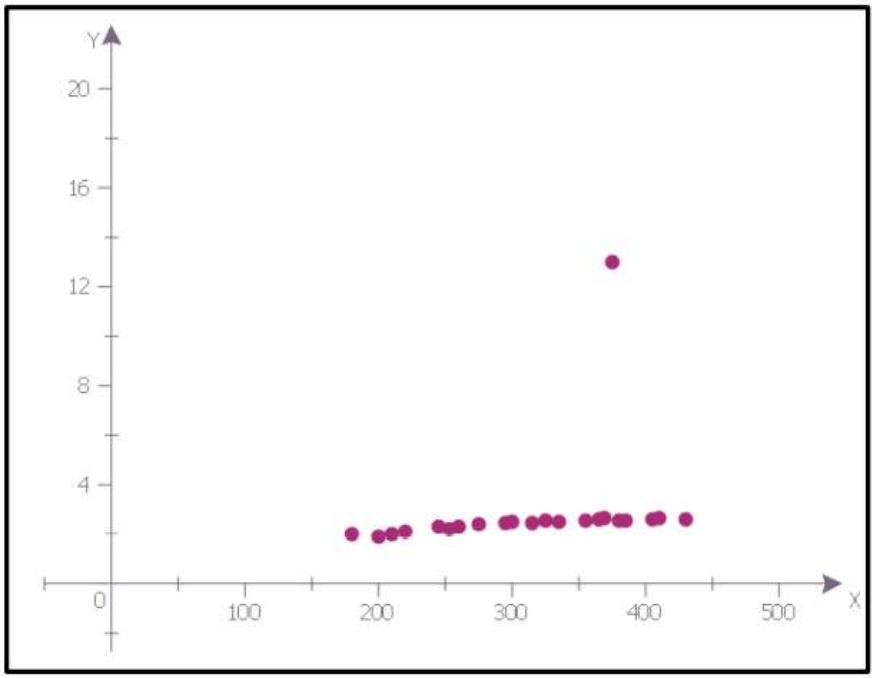
- Scatter plot shows relationship between two sets of data by using points.
- Scatter plot is also typically used in multi-objective decision making/optimization to visualize the Pareto front.

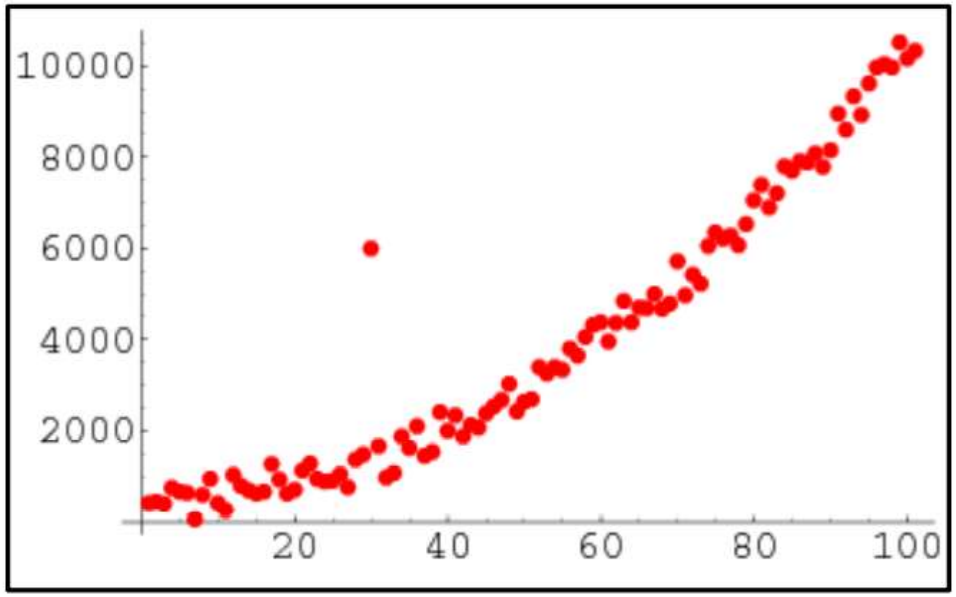




Presenting your results: Scatterplot (2)

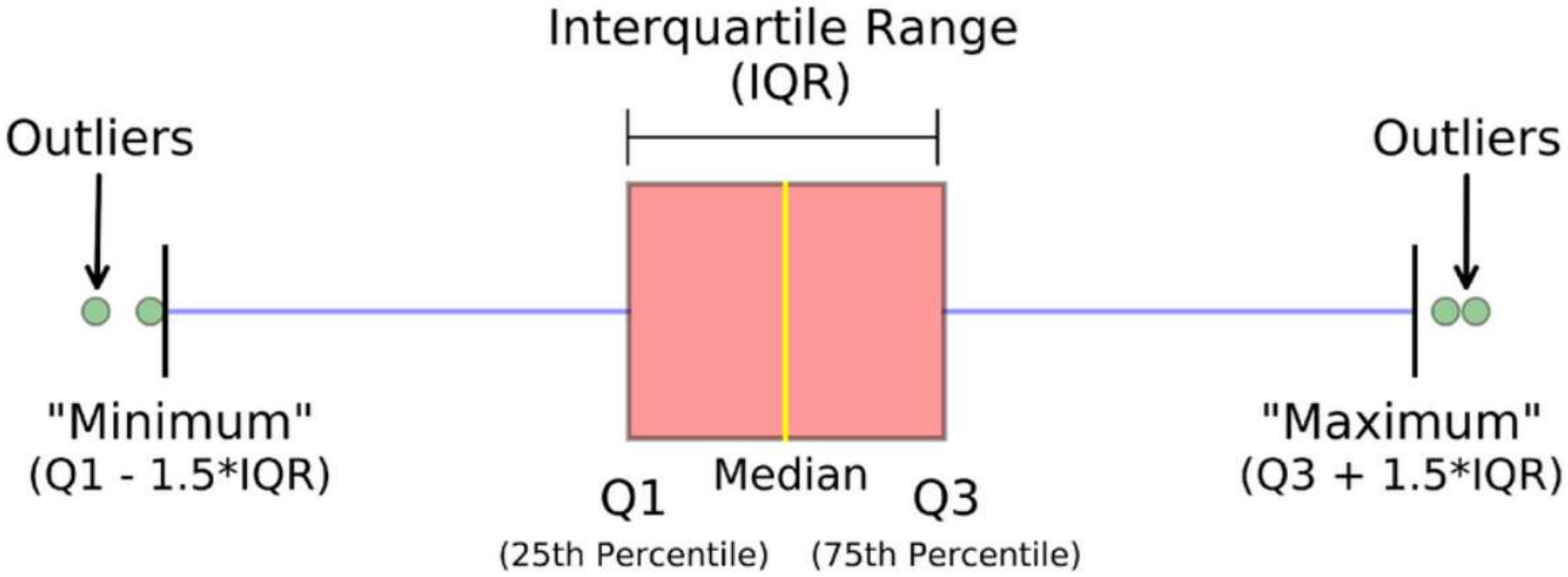
- Scatter plot can also be used for outlier detection.





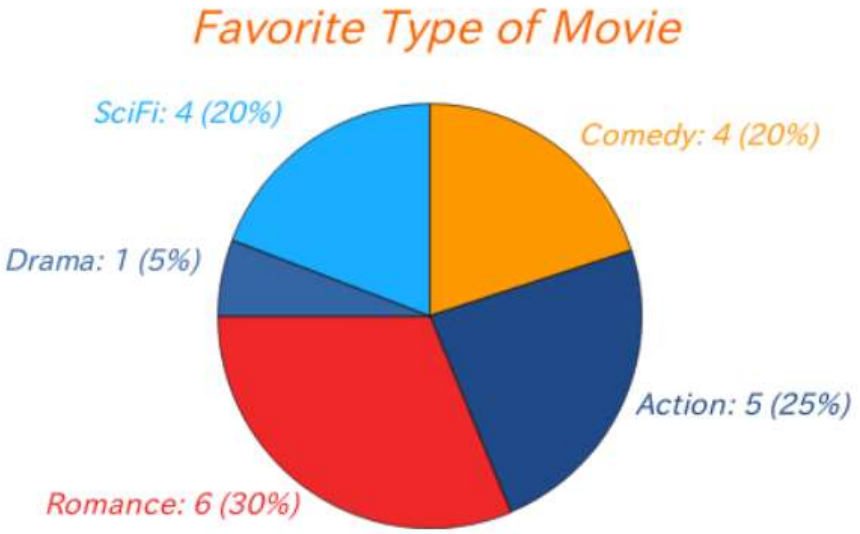
Presenting your results: Boxplot

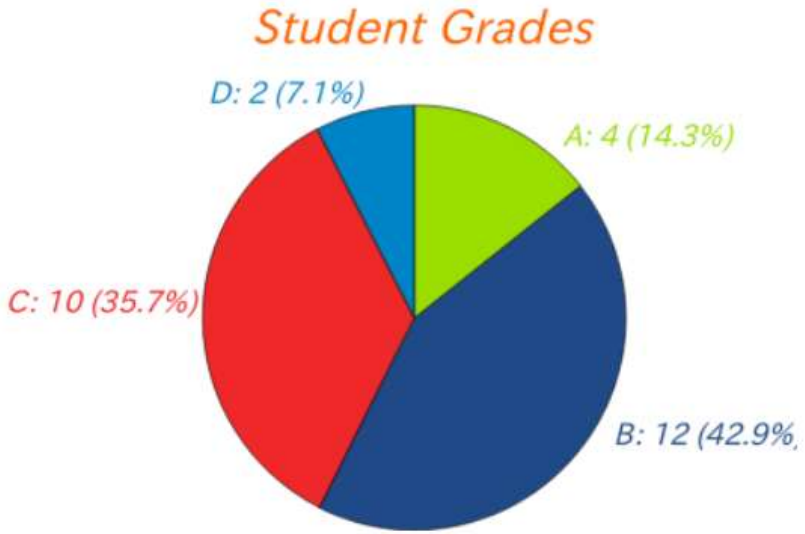
- Displaying the distribution of data based on a five number summary (remember quartiles): Minimum, first quartile, median, third quartile, and maximum.
- Boxplot gives you good indication of how your datas are clustered and spread out.
- Other name of boxplot is “box and whiskers” plot.
- The whiskers can indicate: +/- 1.5 IQR, minimum and maximum of the data, one standard deviation above and below the data, 9th and 91st percentile.



Presenting your results: Pie chart

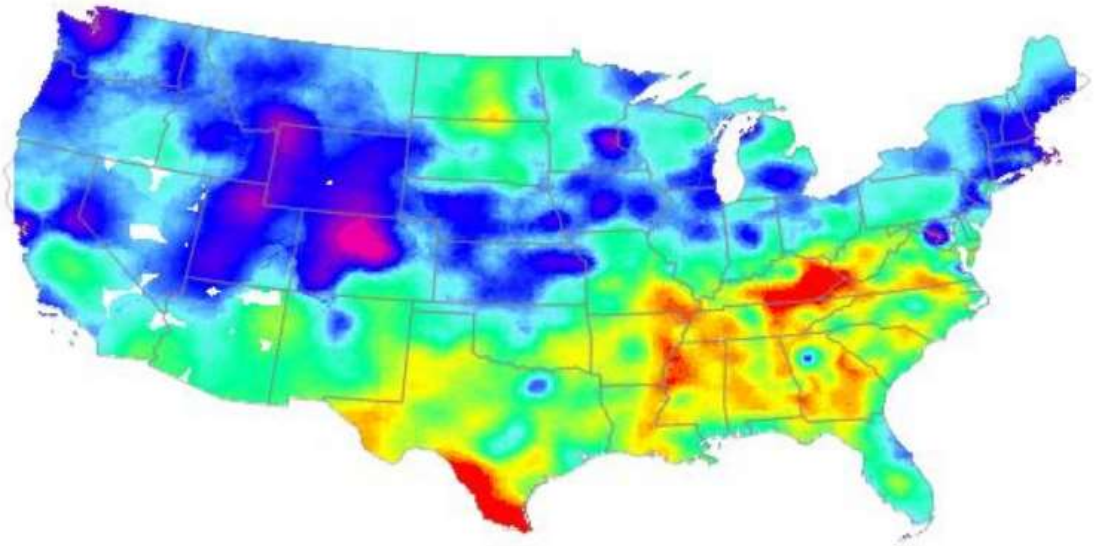
- Pie car is useful when you have a data set such as preferences, sensitivity index, grades, to name a few.

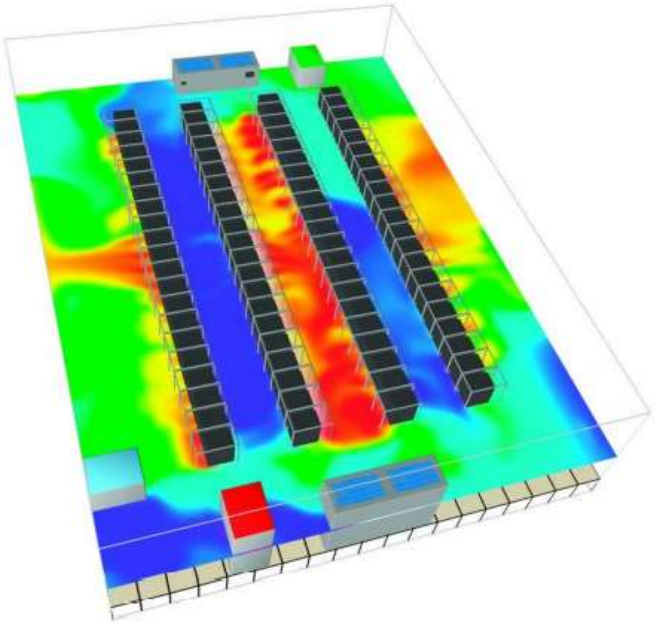




Presenting your results: Heat map

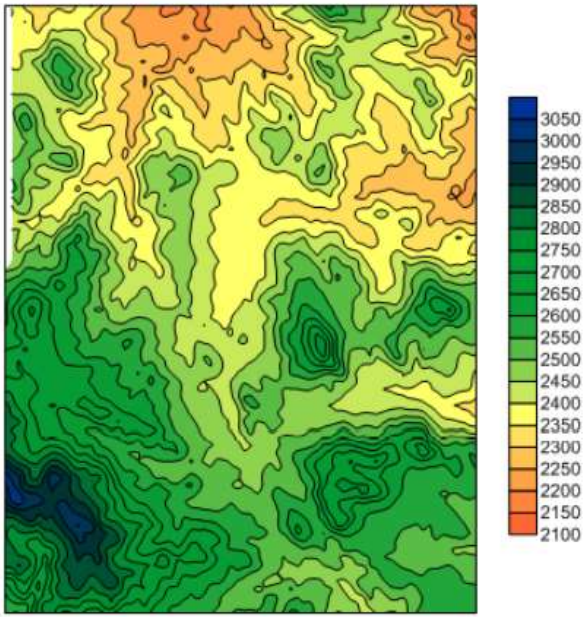
- Heat map is useful to better visualize data with arbitrary borders (e.g., geospatial, CFD/FEM results)

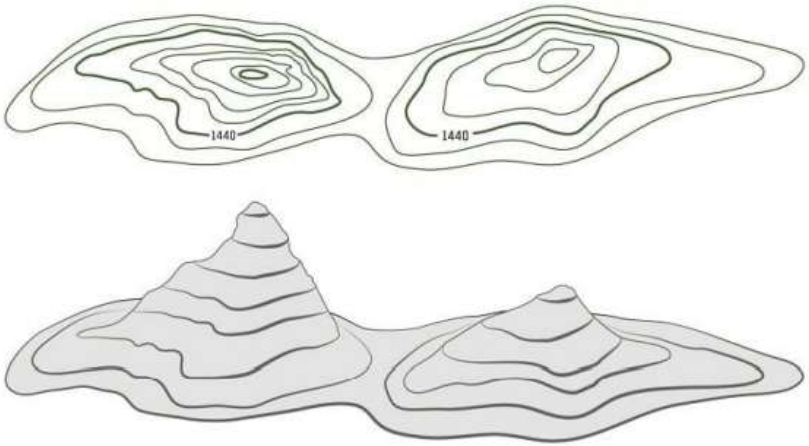


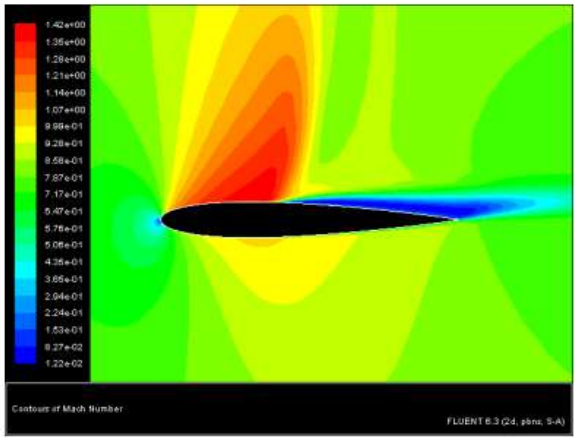


Presenting your results: Contour map.

- Contour map performs visualization by creating multiple lines that share the same values.
- Contour map can be shown with or without colours.
- In some occasions, contour map can be combined with heat map.

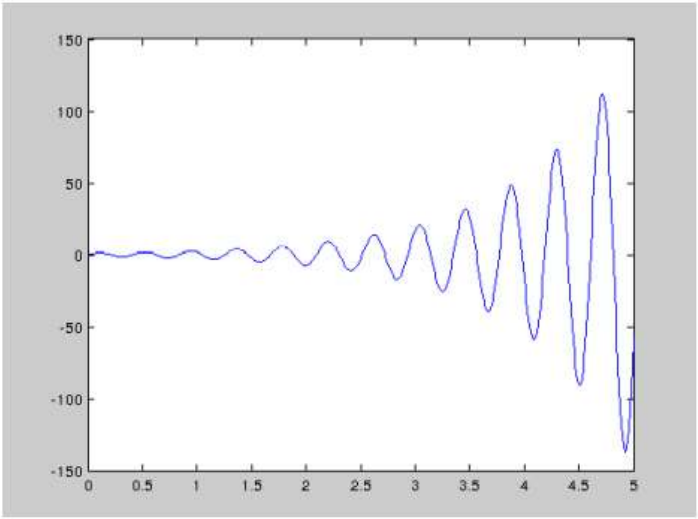




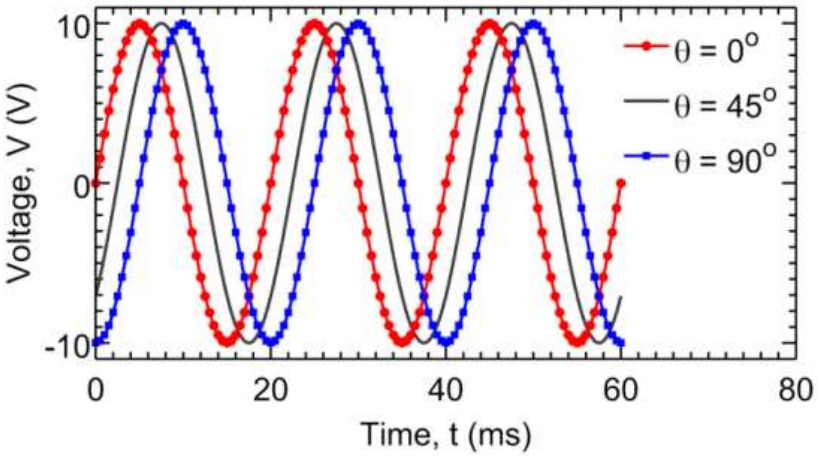


Tips: Use publication-quality figures, please..

- Many journals will comment on the quality of your figures.
- Some reviewers even will 'look down' on your papers when they see poor-quality figures.
- Don't forget to give details! Labels, legend, etc.
- Learn how do do this, properly, <u>properly</u>, **<u>properly</u>**.



Not-acceptable



Acceptable

Sampling distribution

Sampling distribution

- We compute a statistic from a sample selected from the population, and from this statistic we made various statements concerning the values of population parameters that may or may not be true.

Boss: The average content should be 240 ml!

$\bar{x} = 236\ ml$

Is this tolerable?

Should we take action?

Our statistic depends on the random sample and is a random variable

It has a probability distribution

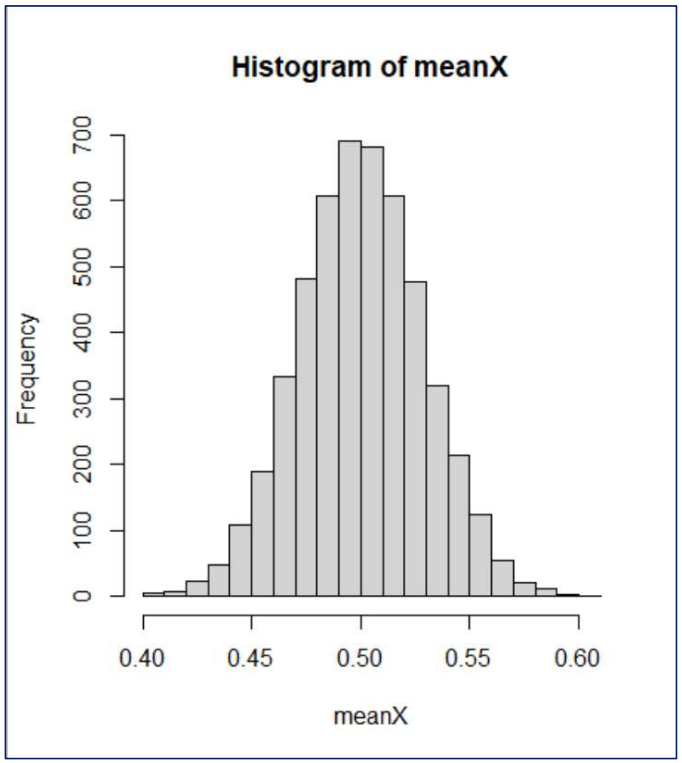
The probability distribution of a statistic is called **sampling distribution**.

Example: Sampling distribution of mean

- Consider a variable *X* with uniform distribution *f(x)* as A = 0, B = 1 (**μ = 0.5)**
- Let's take 100 random samples from this distribution, and we got $\bar{x}$ = 0.53
- Do this again, we got $\bar{x}$ = 0.5104; again, we got $\bar{x}$ = 0.5253; and again.. And again..
- $\bar{x}$ varies for different experiments, **it is random!** It then has a **probability distribution!**

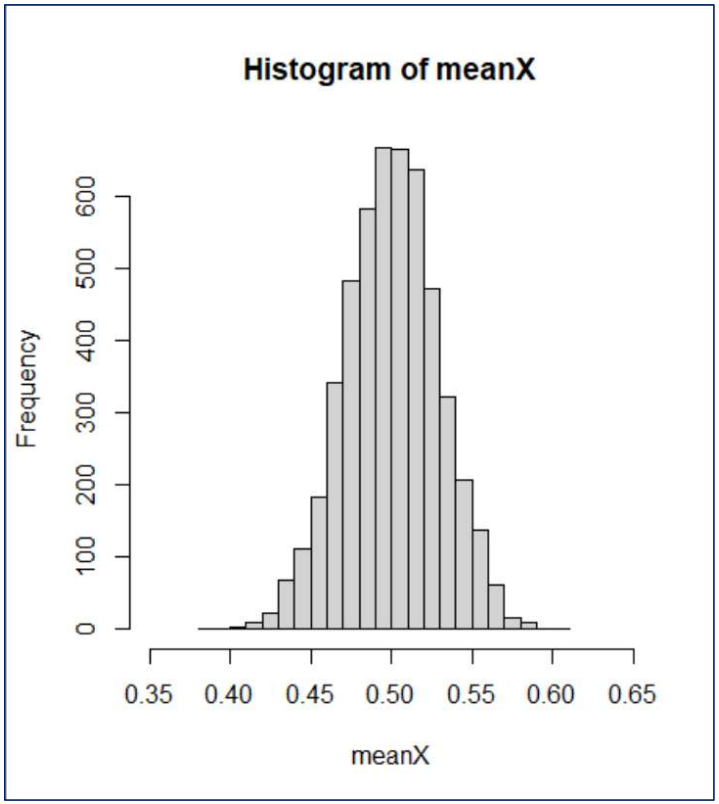
```
nsamp <- 5000 # Number of samples for histogram
meanX <- rep(0,nsamp) # Initialize

# Start the loop
for (i in seq(nsamp)){
  X_unif <- runif(100,min=0,max=1)
  meanX[i] <- mean(X_unif)
}

hist(meanX,breaks=20)
```

Histogram of $\bar{x}$

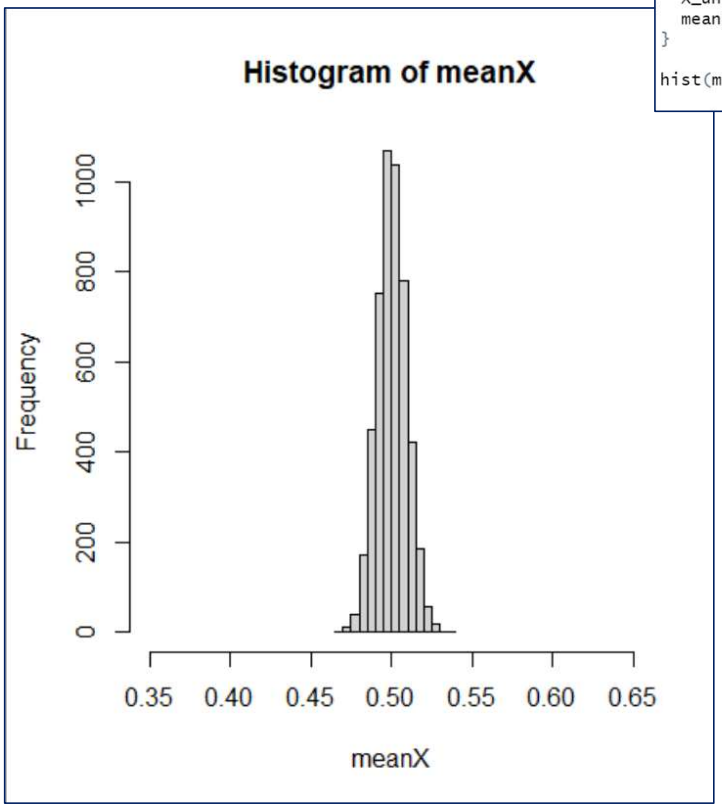


Example: Sampling distribution of mean

Histogram of $\bar{x}$ with n =100



Histogram of $\bar{x}$ with n =1000




```r
nsd <- 5000 # Number of samples for sampling distribution
nsamp <- 1000 # Number of samples for mean calculation

meanX <- rep(0,nsd) # Initialize

# Start the loop
for (i in seq(nsd)){
  X_unif <- runif(nsamp,min=0,max=1)
  meanX[i] <- mean(X_unif)
}

hist(meanX,breaks=20,xlim=c(0.35,0.65))
```

Sampling distribution of mean from a normal distribution

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

X_i is random sample from a normaly distributed variable

$$\mu_{\bar{X}} = \frac{1}{n}(\underbrace{\mu + \mu + \cdots + \mu}_{n \text{ terms}}) = \mu$$

How if the distribution of X is not normal?

$$\sigma^2_{\bar{X}} = \frac{1}{n^2}(\underbrace{\sigma^2 + \sigma^2 + \cdots + \sigma^2}_{n \text{ terms}}) = \frac{\sigma^2}{n}.$$

Central limit theorem

Central Limit Theorem: If $\bar{X}$ is the mean of a random sample of size n taken from a population with mean μ and finite variance σ^2, then the limiting form of the distribution of

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

as $n \to \infty$, is the standard normal distribution $n(z; 0, 1)$.

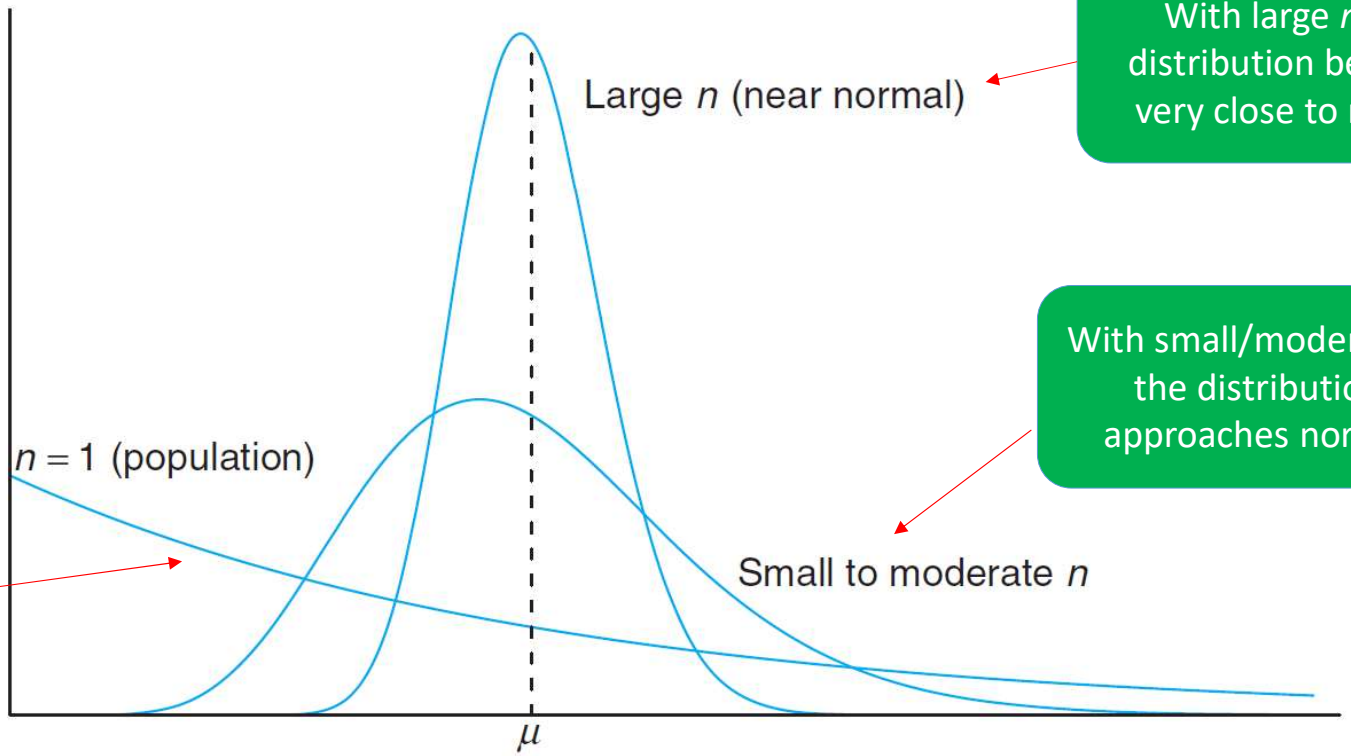
Besides Bayes theorem, CLT is one of the most important theorems in statistics!

This theorem is true for all types of distributions!

Many theories / methods of, hypothesis testing, confidence interval, error analysis, turbulent flow, data science, AI, are built based on CLT.

- This approximation is typically good for *n* >= 30, as long as it is not extremely skewed.
- For *n* <= 30, the approximation is still good if the distribution is close to normal.

Central limit theorem: example, sampling from exponential



With large n, the distribution becomes very close to normal

With small/moderate n, the distribution approaches normal

With n = 1, you sampled from the original exponential distribution

Central limit theorem: example of application

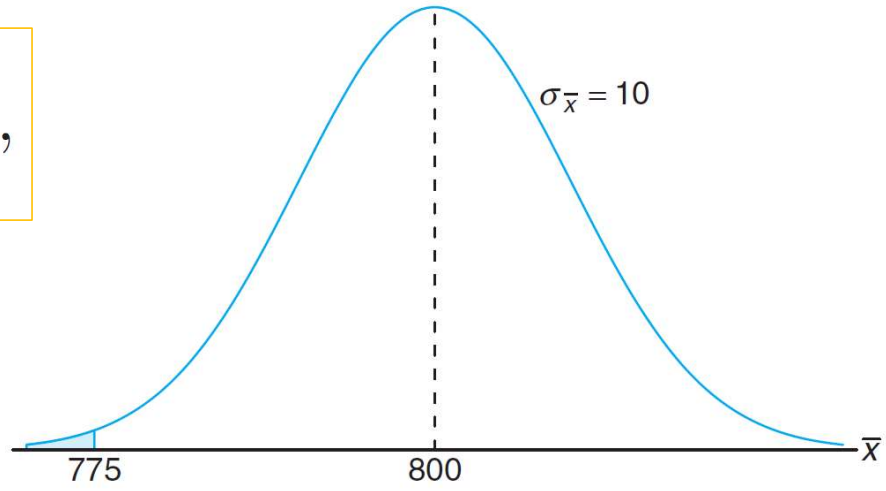
An electrical firm manufactures light bulbs that have a length of life that is approximately normally distributed, with mean equal to 800 hours and a standard deviation of 40 hours. Find the probability that a random sample of 16 bulbs will have an average life of less than 775 hours

Solution

Mean $\mu_{\bar{x}} = 800$

Std error $\sigma_{\bar{x}} = 40/\sqrt{16} = 10$

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



$$z = \frac{775 - 800}{10} = -2.5,$$

$$P(\bar{X} < 775) = P(Z < -2.5) = 0.0062.$$

Statistical hypothesis test

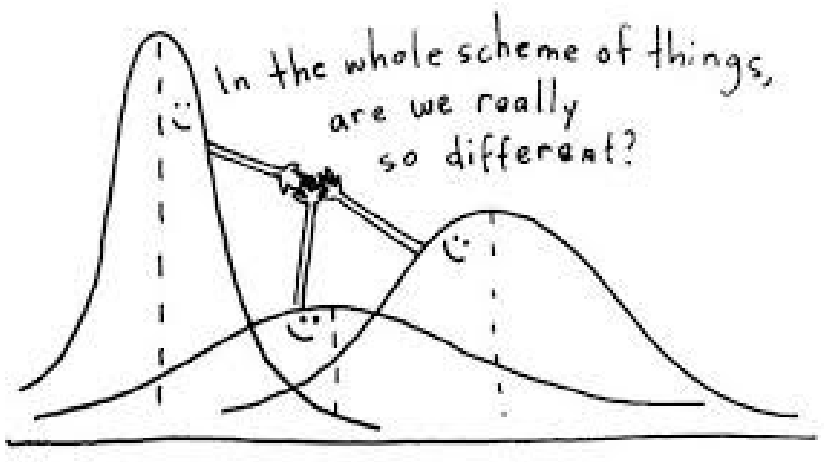
Statistical hypothesis

Stat

- Let's talk about hypothesis: ***"Hypothesis is a formal question that the research intends to resolve"***
- Quite often a research hypothesis is a predictive statement, capable of being tested by scientific methods, that relates an independent to some dependent variable.

The following are good hypotheses:

- Car A consume less fuel than car B
- Students who receive tutorial score higher than other students who do not take tutorial

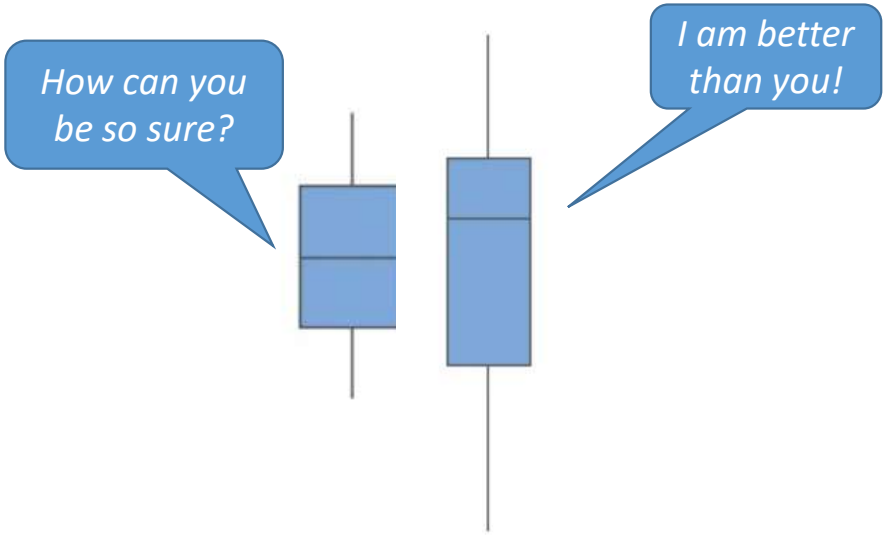


The use of statistical hypothesis

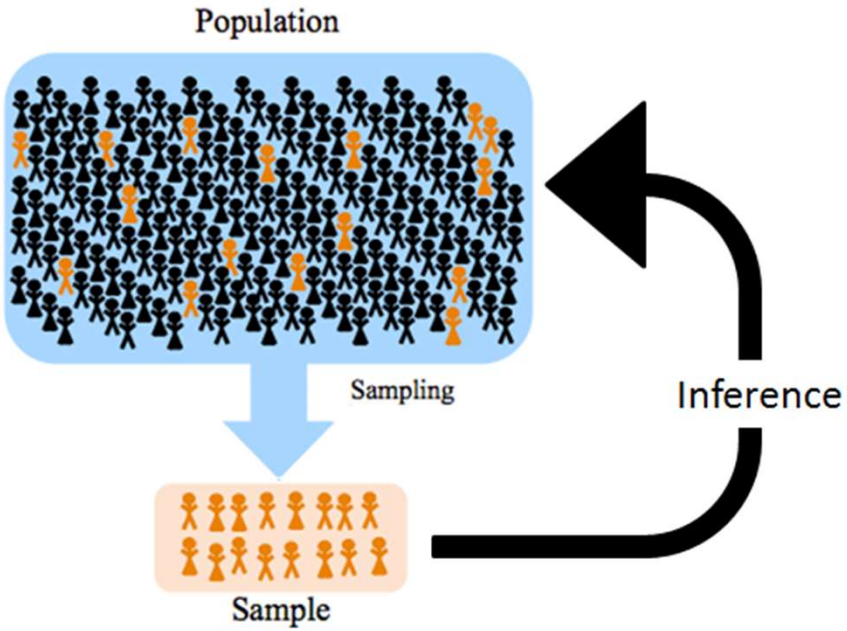
- Does method A significantly performs better than algorithm B? (According to statistical tests)
- Do the students from major X receive significantly higher score than the students from major Y in the calculus class?
- Do male and female undergraduates differ in height on average?
- Do my samples represent the hypothetical values?

Statistical hypothesis test tries to answer these questions

Remember that uncertainties always exist; thus, the need for statistical hypothesis test



Statistical hypothesis

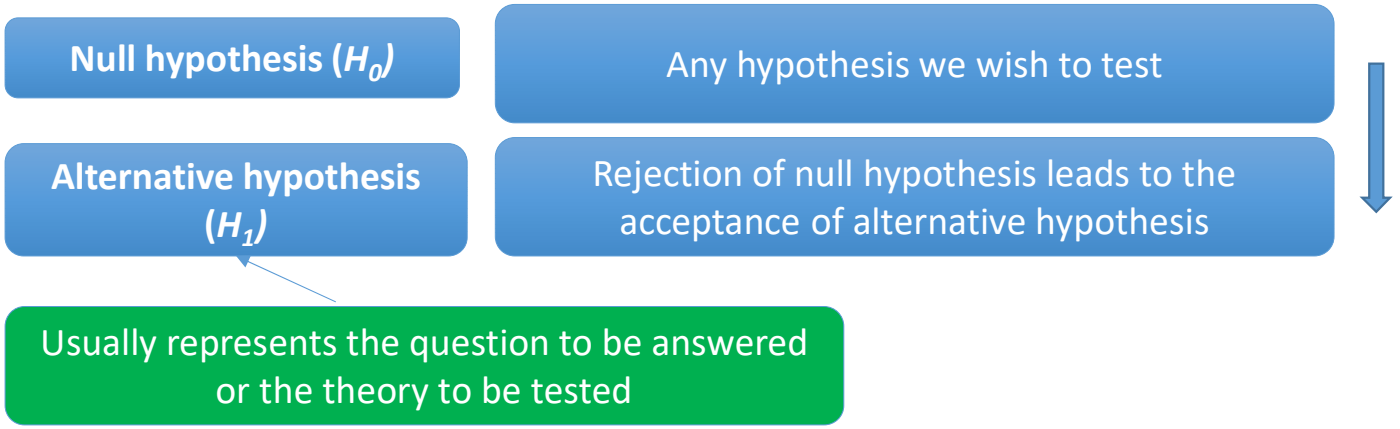


From the sample, we want to find the answer to our hypothesis concerning the population

- **Statistical hypothesis** is an assertion or conjecture concerning one or more populations.
- The truth or falsify of a statistical hypothesis is never known with absolute certainty unless we examine the entire population.
- Thus, we take a random sample and use the data to provide evidence that either support or does not support the hypothesis

Formulating the hypothesis

- If the scientist is interested in *strongly supporting a contention*, he or she hopes to arrive at the contention in the form of rejection of a hypothesis.
- If the medical researches wishes to show strong evidence in favor of the contention that coffee increases the risk of cancer, the hypothesis tested should **"there is no increase in cancer risk produced by drinking coffee"**



Formulating the hypothesis

Null hypothesis (H_0)	Any hypothesis we wish to test
Alternative hypothesis (H_1)	Rejection of null hypothesis leads to the acceptance of alternative hypothesis

reject $\boldsymbol{H_0}$ in favor of H_1 because of sufficient evidence in the data or

fail to reject $\boldsymbol{H_0}$ because of insufficient evidence in the data.

H_0: $p = 0.10$,

H_1: $p > 0.10$.

Reject H_0 if the data produce 20 out of 100 defective items

Fail to reject H_0 if the data produce 12 out of 100 defective items

- Notice that we never say "accept the null hypothesis!", we are just lacking evidence

H_0: defendant is innocent,

H_1: defendant is guilty.

The null hypothesis is the status quo stands in opposition to the alternative and is maintained.
Failure to reject the null hypothesis does not imply innocence! But the evidence is not sufficient

Testing a statistical hypothesis

- A certain type of cold vaccine is known to be only 25% effective after a period of 2 years.
- To determine the superiority of the new samples, 20 people are chosen at random.
- If more than 8 surpass the 2-year period without contracting the virus, the new vaccine is considered superior (notice that the choice of 8 here is quite arbitrary, which is a modest gain over the 5 people)

H_0	The new vaccine is equally effective after a period of 2 years	$H_0: \ p = 0.25,$
H_1	The new vaccine is superior	$H_1: \ p > 0.25.$

Test statistic

- The **test statistic** is X, the number of individuals who receive protection from the new vaccine, in range of 0 to 20. X are divided into two groups

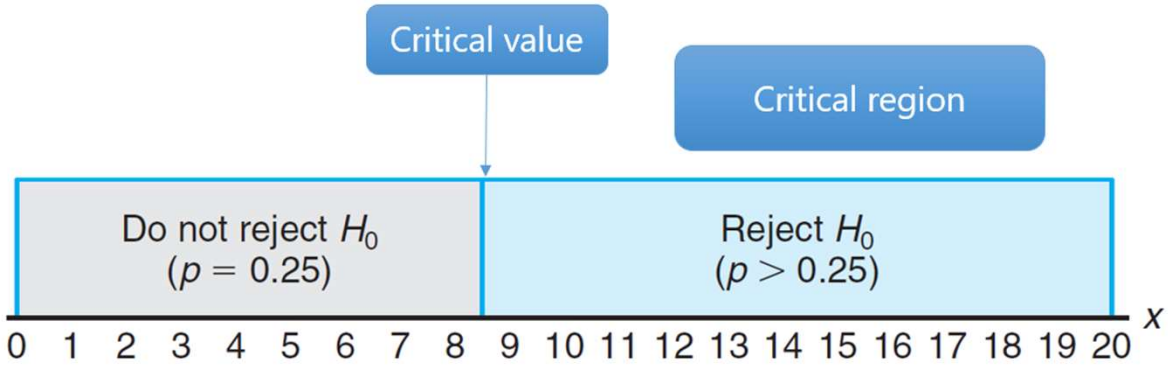


Figure 10.1: Decision criterion for testing $p = 0.25$ versus $p > 0.25$.

Type I and Type II error

	H_0 is true	H_0 is false
Do not reject H_0	Correct decision	Type II error
Reject H_0	Type I error	Correct decision

- The decision procedure could lead to **either of two wrong conclusions**.
- The new vaccine may be no better than the one now in use (null hypothesis true), but we reject H_0
- We would be committing an error by rejecting H_0 in favor of H_1 when in fact the former is true, this is **Type I error**

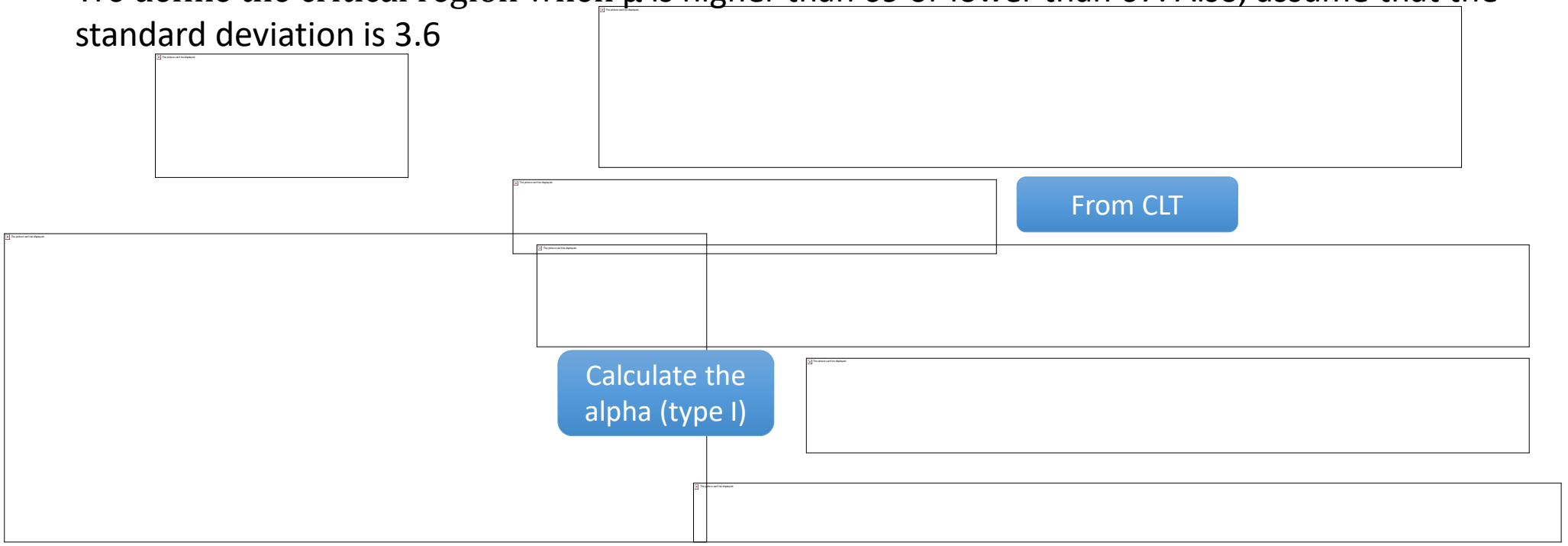
Rejection of the null hypothesis when it is true is called a type I error

- A second kind of error is committed if 8 or fewer of the group surpass the 2-year period and we are unable to conclude that the vaccine is better when it actually is better (H_1 is true). We then fail to reject H_0 when in fact H_0 is false. **This is type II error.**

Nonrejection of the null hypothesis when it is false is called a type II error

Illustration: Continuous random variable

- Consider the null hypothesis that the average weight of male students in a certain college is 68 kilograms against the alternative hypothesis that it is unequal to 68.
- We define the critical region when μ is higher than 69 or lower than 67. Also, assume that the standard deviation is 3.6



P-value

- The most common way in statistical hypothesis is to fix the α to 0.05 or 0.01, and select the critical region accordingly.
- If the test is two tailed and α is set at the 0.05 level, then a z-value is observed from the data and the critical region is:

$$z > 1.96 \qquad \text{or} \qquad z < -1.96,$$

- A value of z in the critical region prompts the statement "The value of the test statistic is significant.
- For example, if

$$H_0: \ \mu = 10,$$
$$H_1: \ \mu \neq 10,$$

If we fail to reject null hypothesis, then we can say "The mean differs significantly from the value 10"

P-value vs classic hypothesis testing

Approach to Hypothesis Testing with Fixed Probability of Type I Error	**1.** State the null and alternative hypotheses. **2.** Choose a fixed significance level α. **3.** Choose an appropriate test statistic and establish the critical region based on α. **4.** Reject H_0 if the computed test statistic is in the critical region. Otherwise, do not reject. **5.** Draw scientific or engineering conclusions.

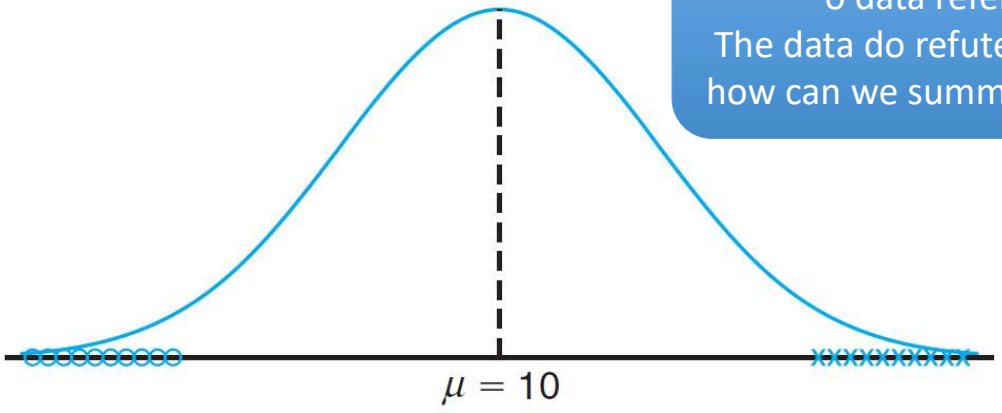
Significance Testing (*P*-Value Approach)	**1.** State null and alternative hypotheses. **2.** Choose an appropriate test statistic. **3.** Compute the *P*-value based on the computed value of the test statistic. **4.** Use judgment based on the *P*-value and knowledge of the scientific system.

We will use P-value approach from now on

P-value: graphical demonstration

- Suppose that two materials are being considered for coating to inhibit corrosion.
- The sample sizes are $n_1 = n_2 = 10$, and corrosion is measured in percent of surface are affected.
- The hypothesis is that the samples came from common distribution with mean $\mu = 10$ (assume that we know the population variance, $\sigma^2 = 10$)

$$H_0: \ \mu_1 = \mu_2 = 10.$$



x data refer to material 1.
o data refer to material 2
The data do refute the null hypothesis, how can we summarize this in number?

The *P*-value can be viewed as simply the probability of obtaining these data given that both samples come from the same distribution

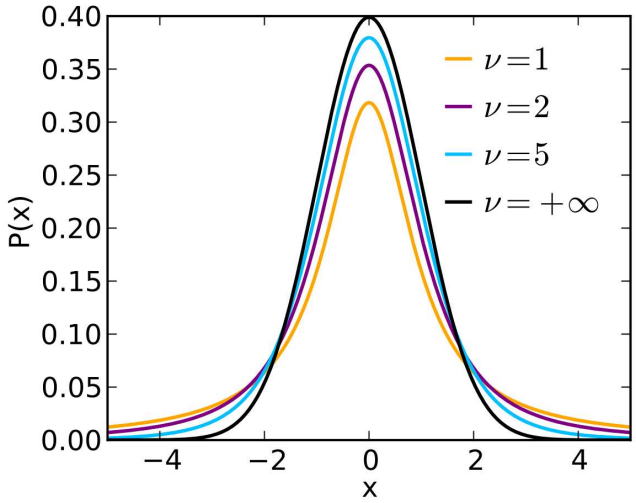
For this problem $P = 0.00000001$!
The small *p*-value refutes null hypothesis, and we conclude that the population means are significantly different

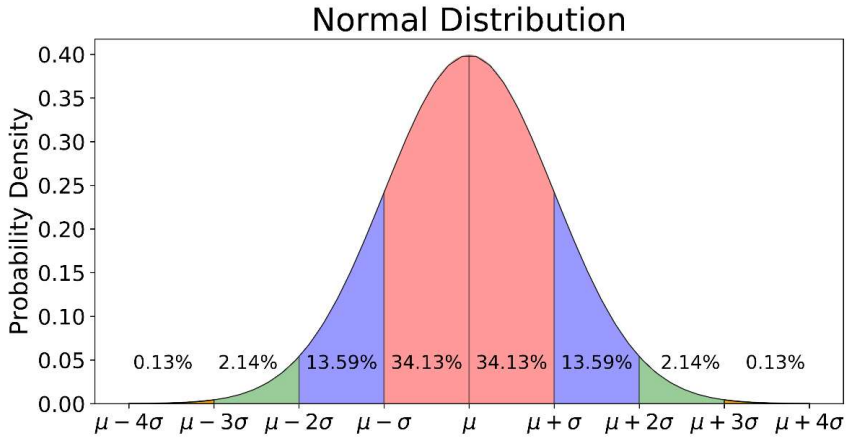
Statistical hypothesis test

Parametric hypothesis test

Parametric test

- **Z-test.** Based on the normal distribution and is used for judging several statistical measures (particularly the mean).
- **t-test.** Based on t-distribution and is used for judging the significance of a sample mean or the difference between two means of two samples in case of small samples.
- **Chi-square test.** Based on chi-square distribution and is used for comparing a sample variance to a theoretical population variance.
- **F-test.** Based on F-distribution and is used to compare the variance of two independent samples





Tests on a single mean (variance known)

- Consider the hypothesis

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

- The appropriate test statistic should be based on the random variable $\bar{X}$, which is normally distributed with mean μ and variance σ^2/n.
- As usual, it is convenient to standardize $\bar{X}$

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

- Under H_0, that is if $\mu = \mu_0$ $\sqrt{n}(\bar{X} - \mu_0)/\sigma$ follows a normal distribution, hence:

$$P\left(-z_{\alpha/2} < \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1 - \alpha$$

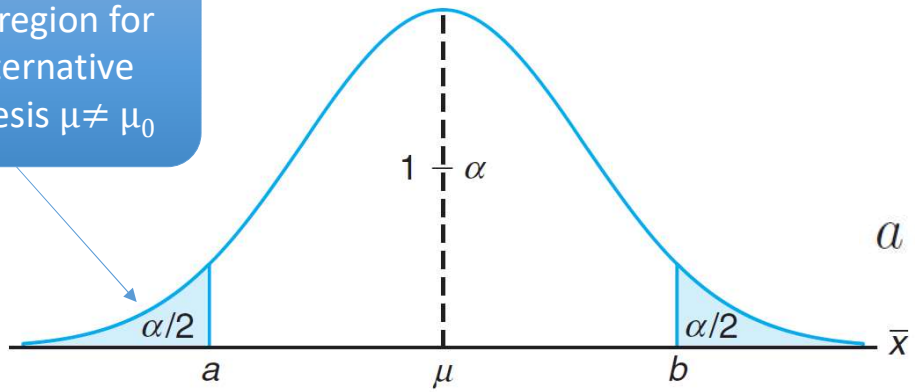
Tests on a single mean (variance known)

Test procedure on a single mean (variance known)

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} > z_{\alpha/2} \quad \text{or} \quad z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} < -z_{\alpha/2}$$

If $-z_{\alpha/2} < z < z_{\alpha/2}$, do not reject H_0. Rejection of H_0, of course, implies acceptance of the alternative hypothesis $\mu \neq \mu_0$. With this definition of the critical region, it should be clear that there will be probability α of rejecting H_0 (falling into the critical region) when, indeed, $\mu = \mu_0$.

Critical region for the alternative hypothesis $\mu \neq \mu_0$



Or in terms of the original variables..

$$\textit{reject } H_0 \text{ if } \bar{x} < a \text{ or } \bar{x} > b,$$

$$a = \mu_0 - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \qquad b = \mu_0 + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

The signal that favors H_1 comes from large or small values of *z!*

Tests on a single mean (variance known)

For one (upper) sided

$$H_0: \ \mu = \mu_0,$$
$$H_1: \ \mu > \mu_0.$$

The signal that favors H_1 comes from large values of *z!*

Reject H_0 if $z > z_\alpha$

- For one (lower) sided, surely you can just change the alternative hypothesis to

$$H_1: \mu < \mu_0$$

The signal that favors H_1 comes from small values of *z!*

Reject H_0 if $z < -z_\alpha$

Statistical hypothesis testing, illustration 1

Population normal, population infinite, sample size may be large or small but variance of the population is known, H_a may be one-sided or two-sided:

A sample of 400 male students is found to have a mean height 67.47 inches. Can it be reasonably regarded as a sample from a large population with mean height 67.39 inches and standard deviation 1.30 inches? Test at 5% level of significance.

- In this problem, we would like to know whether these 400 male students represent the general population.
- The first step is to formulate the hypothesis. The null hypothesis is that the mean height of the population is equal to 67.39 inches:

$$H_0 : \mu_{H_0} = 67.39''$$
$$H_a : \mu_{H_0} \neq 67.39''$$

Given information

$$\bar{X} = 67.47'', \ \sigma_p = 1.30'', \ n = 400$$

We assume normal distribution and we use *z*-test

Accept

$$z = \frac{\bar{X} - \mu_{H_0}}{\sigma_p / \sqrt{n}} = \frac{67.47 - 67.39}{1.30/\sqrt{400}} = \frac{0.08}{0.065} = 1.231$$

Std. error of the mean

$$\text{Acceptance Region } A : |Z| \leq 1.96$$
$$\text{Rejection Region } R : |Z| > 1.96$$

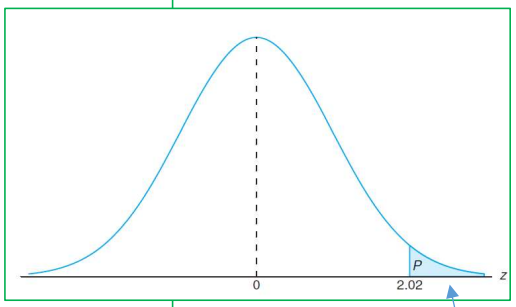
Given information

T
e

Example 10.3: A random sample of 100 recorded deaths in the United States during the past year showed an average life span of 71.8 years. Assuming a population standard deviation of 8.9 years, does this seem to indicate that the mean life span today is greater than 70 years? Use a 0.05 level of significance.

Solution:

1. H_0: $\mu = 70$ years.
2. H_1: $\mu > 70$ years.
3. $\alpha = 0.05$.
4. Critical region: $z > 1.645$, where $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$.
5. Computations: $\bar{x} = 71.8$ years, $\sigma = 8.9$ years, and hence $z = \frac{71.8 - 70}{8.9/\sqrt{100}} = 2.02$.
6. Decision: Reject H_0 and conclude that the mean life span today is greater than 70 years.



The P-value corresponding to $z = 2.02$ is given by the area of the shaded region in Figure 10.10.

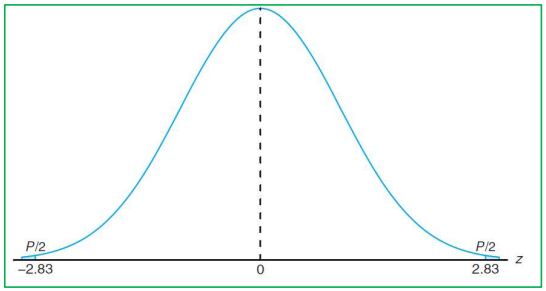
Using Table A.3, we have

$$P = P(Z > 2.02) = 0.0217.$$

As a result, the evidence in favor of H_1 is even stronger than that suggested by a 0.05 level of significance.

A manufacturer of sports equipment has developed a new synthetic fishing line that the company claims has a mean breaking strength of 8 kilograms with a standard deviation of 0.5 kilogram. Test the hypothesis that $\mu = 8$ kilograms against the alternative that $\mu \neq 8$ kilograms if a random sample of 50 lines is tested and found to have a mean breaking strength of 7.8 kilograms. Use a 0.01 level of significance.

1. H_0: $\mu = 8$ kilograms.
2. H_1: $\mu \neq 8$ kilograms.
3. $\alpha = 0.01$.
4. Critical region: $z < -2.575$ and $z > 2.575$, where $z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$.
5. Computations: $\bar{x} = 7.8$ kilograms, $n = 50$, and hence $z = \frac{7.8-8}{0.5/\sqrt{50}} = -2.83$.
6. Decision: Reject H_0 and conclude that the average breaking strength is not equal to 8 but is, in fact, less than 8 kilograms.



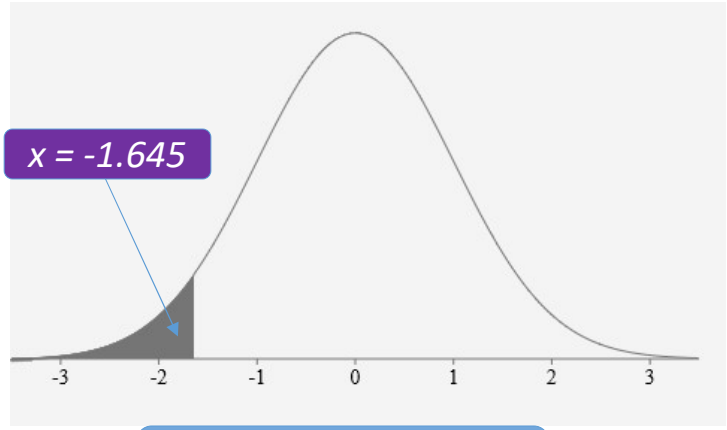
Since the test in this example is two tailed, the desired P-value is twice the area of the shaded region in Figure 10.11 to the left of $z = -2.83$. Therefore, using Table A.3, we have

$$P = P(|Z| > 2.83) = 2P(Z < -2.83) = 0.0046,$$

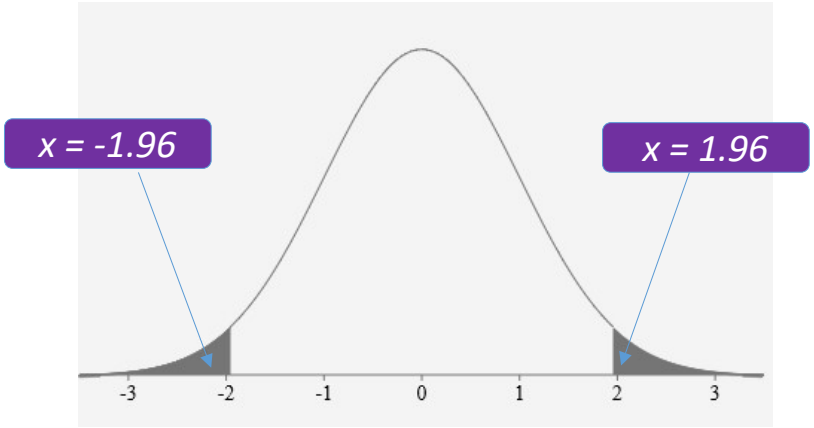
which allows us to reject the null hypothesis that $\mu = 8$ kilograms at a level of significance smaller than 0.01.

Important values from the normal distribution table

http://onlinestatbook.com/2/calculators/normal_dist.html



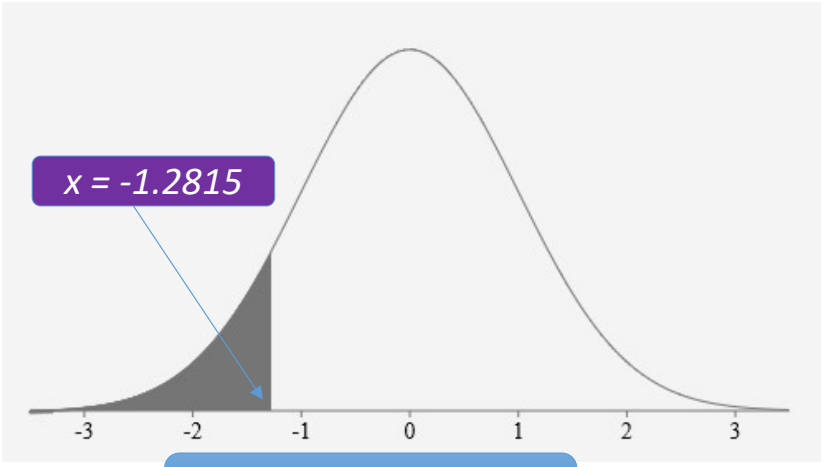
α = 0.05, one-sided



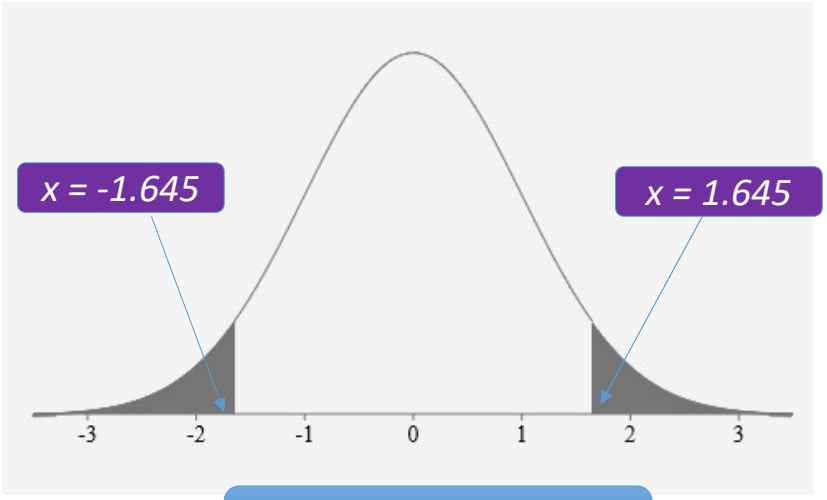
α = 0.05, two-sided

Important values from normal distribution table

http://onlinestatbook.com/2/calculators/normal_dist.html



α = 0.1, one-sided



α = 0.1, two-sided

Tests on a single mean (variance unknown)

For the two-sided hypothesis

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

we reject H_0 at significance level α when the computed t-statistic

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

exceeds $t_{\alpha/2,n-1}$ or is less than $-t_{\alpha/2,n-1}$.

Degree of freedom = *n*-1

For one-sided, we use the same principle as in the case of known variance

The Edison Electric Institute has published figures on the number of kilowatt hours used annually by various home appliances. It is claimed that a vacuum cleaner uses an average of 46 kilowatt hours per year. If a random sample of 12 homes included in a planned study indicates that vacuum cleaners use an average of 42 kilowatt hours per year with a standard deviation of 11.9 kilowatt hours, does this suggest at the 0.05 level of significance that vacuum cleaners use, on average, less than 46 kilowatt hours annually? Assume the population of kilowatt hours to be normal.

1. H_0: $\mu = 46$ kilowatt hours.
2. H_1: $\mu < 46$ kilowatt hours.
3. $\alpha = 0.05$.
4. Critical region: $t < -1.796$, where $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$ with 11 degrees of freedom.
5. Computations: $\bar{x} = 42$ kilowatt hours, $s = 11.9$ kilowatt hours, and $n = 12$. Hence,

$$t = \frac{42 - 46}{11.9/\sqrt{12}} = -1.16, \qquad P = P(T < -1.16) \approx 0.135.$$

6. Decision: Do not reject H_0 and conclude that the average number of kilowatt hours used annually by home vacuum cleaners is not significantly less than 46.

Two samples: Tests on Two Means (known variances)

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

Z has a standard normal distribution

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{1/n_1 + 1/n_2}}.$$

Z if we assume that $\sigma_1 = \sigma_2 = \sigma$

Null hypothesis

$$H_0\colon \mu_1 - \mu_2 = d_0.$$

Test statistic

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}},$$

Alternative hypothesis

$$H_1\colon \mu_1 - \mu_2 \neq d_0$$

Condition to reject H_0

$$z > z_{\alpha/2} \text{ or } z < -z_{\alpha/2}$$

For upper-sided, one tail

$$H_1\colon \mu_1 - \mu_2 > d_0$$

Two samples: Tests on Two Means (unknown but equal variance) = The pooled t-test

$$H_0: \ \mu_1 = \mu_2,$$
$$H_1: \ \mu_1 \neq \mu_2,$$

Two sided null and alternative hypothesis

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{s_p\sqrt{1/n_1 + 1/n_2}},$$

Test statistic

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

Conditions to reject null hypothesis

exceeds $t_{\alpha/2,n_1+n_2-2}$ or is less than $-t_{\alpha/2,n_1+n_2-2}$.

For one-sided alternative hypothesis

for H_1: $\mu_1 - \mu_2 > d_0$, reject H_1: $\mu_1 - \mu_2 = d_0$ when $t > t_{\alpha,n_1+n_2-2}$.

An experiment was performed to compare the abrasive wear of two different laminated materials. Twelve pieces of material 1 were tested by exposing each piece to a machine measuring wear. Ten pieces of material 2 were similarly tested. In each case, the depth of wear was observed. The samples of material 1 gave an average (coded) wear of 85 units with a sample standard deviation of 4, while the samples of material 2 gave an average of 81 with a sample standard deviation of 5. Can we conclude at the 0.05 level of significance that the abrasive wear of material 1 exceeds that of material 2 by more than 2 units? Assume the populations to be approximately normal with equal variances.
Let μ_1 and μ_2 represent the population means of the abrasive wear for material 1 and material 2, respectively.

1. H_0: $\mu_1 - \mu_2 = 2$.
2. H_1: $\mu_1 - \mu_2 > 2$.
3. $\alpha = 0.05$.
4. Critical region: $t > 1.725$, where $t = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{s_p\sqrt{1/n_1 + 1/n_2}}$ with $v = 20$ degrees of freedom.
5. Computations:

$$\begin{aligned} \bar{x}_1 &= 85, \quad & s_1 &= 4, \quad & n_1 &= 12, \\ \bar{x}_2 &= 81, \quad & s_2 &= 5, \quad & n_2 &= 10. \end{aligned}$$

Hence

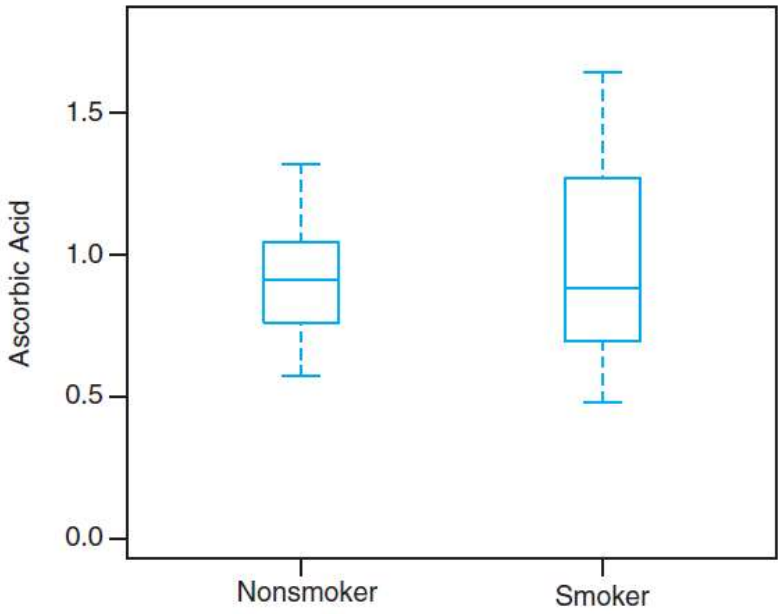
$$s_p = \sqrt{\frac{(11)(16) + (9)(25)}{12 + 10 - 2}} = 4.478,$$

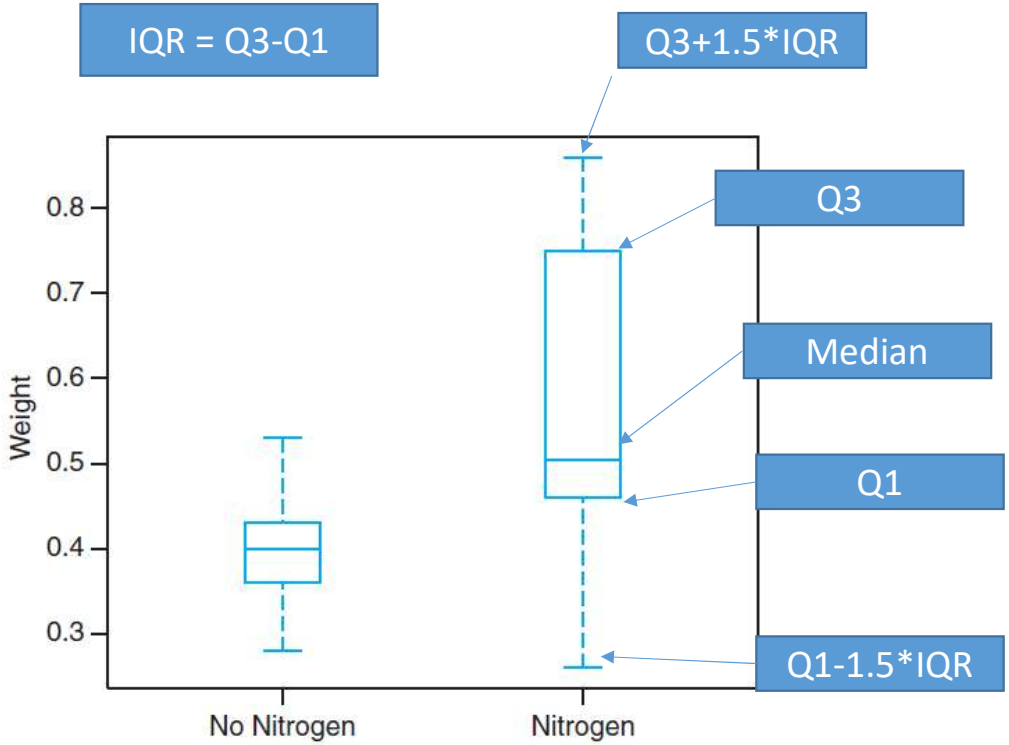
$$t = \frac{(85 - 81) - 2}{4.478\sqrt{1/12 + 1/10}} = 1.04,$$

$$P = P(T > 1.04) \approx 0.16. \qquad \text{(See Table A.4.)}$$

6. Decision: Do not reject H_0. We are unable to conclude that the abrasive wear of material 1 exceeds that of material 2 by more than 2 units.

The use of box-plot





Other cases

- Hypothesis testing for comparing two related samples.
- Hypothesis testing for difference between proportions.
- Hypothesis testing for comparing a variance to some hypothesized population variance.
- Hypothesis testing of correlation coefficients.
- ….

Hypothesis testing: the debate

"p-value < 0.05 is not strict, we should set p-value to < 0.005!"

Why?

- By setting a lower threshold, science will get benefit by having a lower probability of accepting false observations.
- Increasing reproducibility of research. Studies that yielded highly significant results (< 0.01) are more likely to reproduce than those that are just barely significant at the 0.05 level

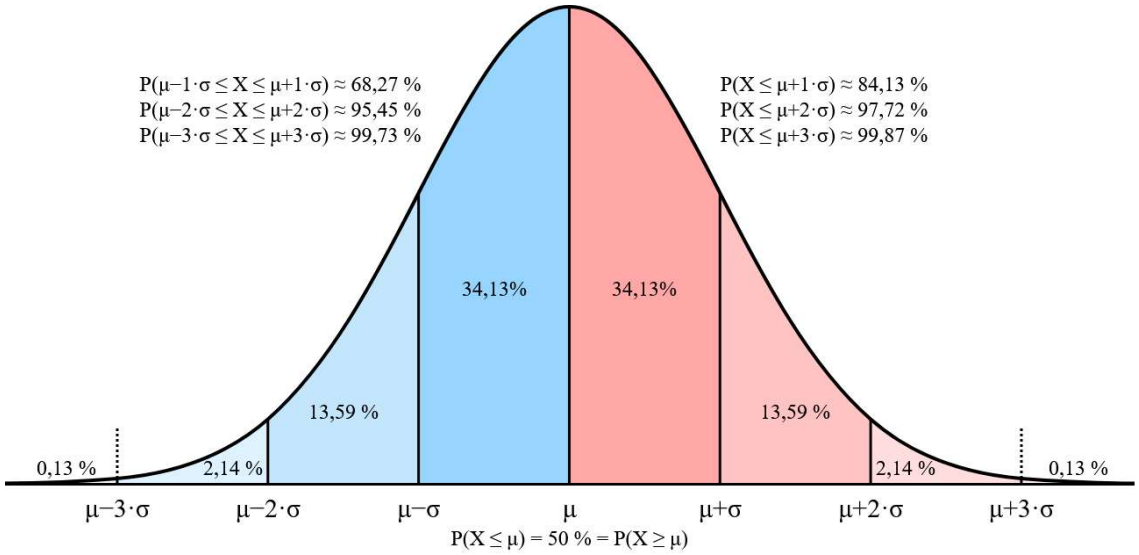
"Statistical significance erodes the culture of science, they should be banned"

Why?

- Many researchers give too much focus on answering the questions of where there is an effect or not, but giving less importance on why a particular phenomenon happen.
- However, both things should actually go hand in hand, it is the culture that should change.

Hypothesis testing: the debate

- Although it is a common practice to use $p < 0.05$, there are some discplines that already use very low p-values.
- Genetics use $p < 0.00000005$, astrophysics uses $p < 0.0000003$ (5-sigma, this is how they can be so sure that Higgs-Boson exists).



Notes on hypothesis testing

- **Testing is not decision making,** the tests are only useful aids for decision making.
- **Test do not explain the reasons as to why does the difference exists**. They tell us whether the difference is due to chances or due to specific reasons.
- Test do not tell us "true" or "false". They give us the probability of accepting or rejecting evidences (e.g., very strong evidence, weak evidence, etc).

The take home point is that statistical testing should be combined with adequate knowledge of the subject

ANOVA

Design of experiment

Statistically based experimental design techniques are particularly useful in the engineering world for solving many important problems.

- **Discovery** of new basic phenomena that can lead to new products and commercialization of new.
- **Improvement** of existing products and processes.

Some applications:

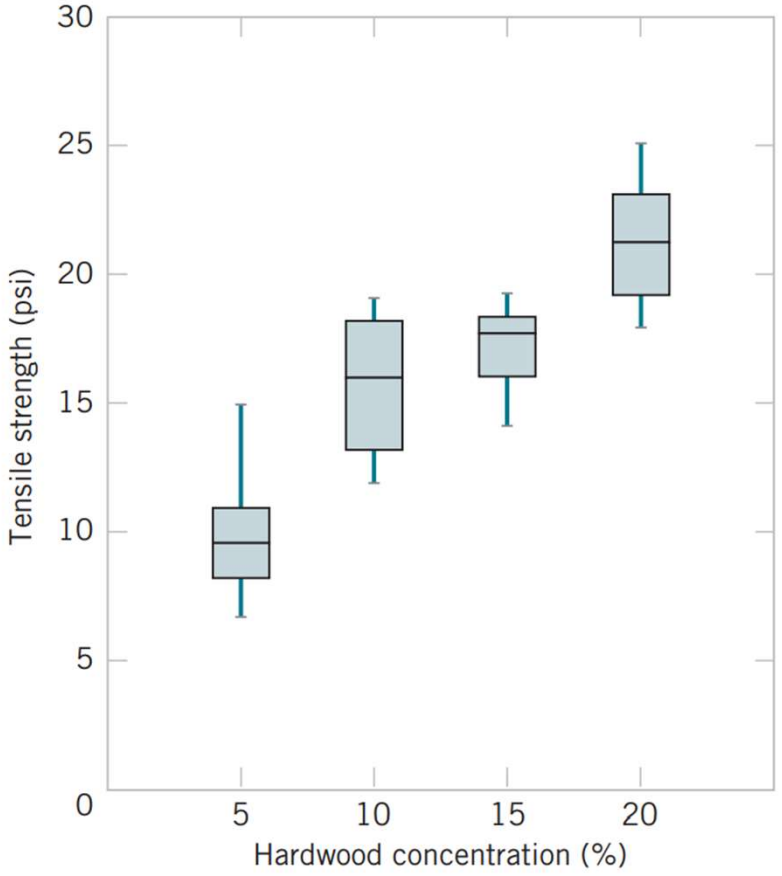
- Evaluation and comparison of basic design configurations
- Evaluation of different materials
- Selection of design parameters so that the product will work well under a wide variety of field conditions (or so that the design will be robust)
- Determination of key product design parameters that affect product performance

Single-factor experiment

Example: Tensile strength experiment. “It is believed that tensile strength is a function of the hardwood concentration in the pulp”

Hardwood Concentration (%)	Observations						Totals	Averages
	1	2	3	4	5	6		
5	7	8	15	11	9	10	60	10.00
10	12	17	13	18	19	15	94	15.67
15	14	18	19	17	16	18	102	17.00
20	19	25	22	23	18	20	127	21.17
							383	15.96

- An experiment with four different levels of **factor** and six **replicates.**
- **Replication is important**.



Hardwood Concentration (%)	Observations 1	2	3	4	5	6	Totals	Averages
5	7	8	15	11	9	10	60	10.00
10	12	17	13	18	19	15	94	15.67
15	14	18	19	17	16	18	102	17.00
20	19	25	22	23	18	20	127	21.17
							383	15.96

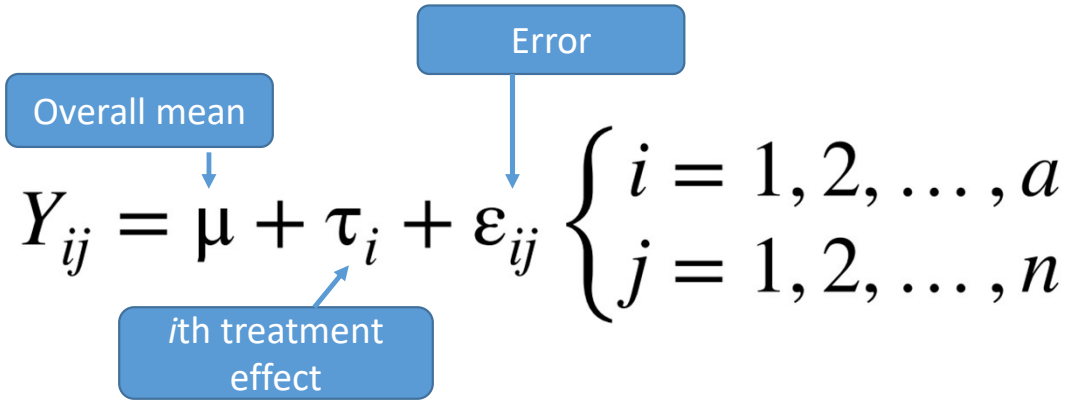
What can you infer from the boxplot?

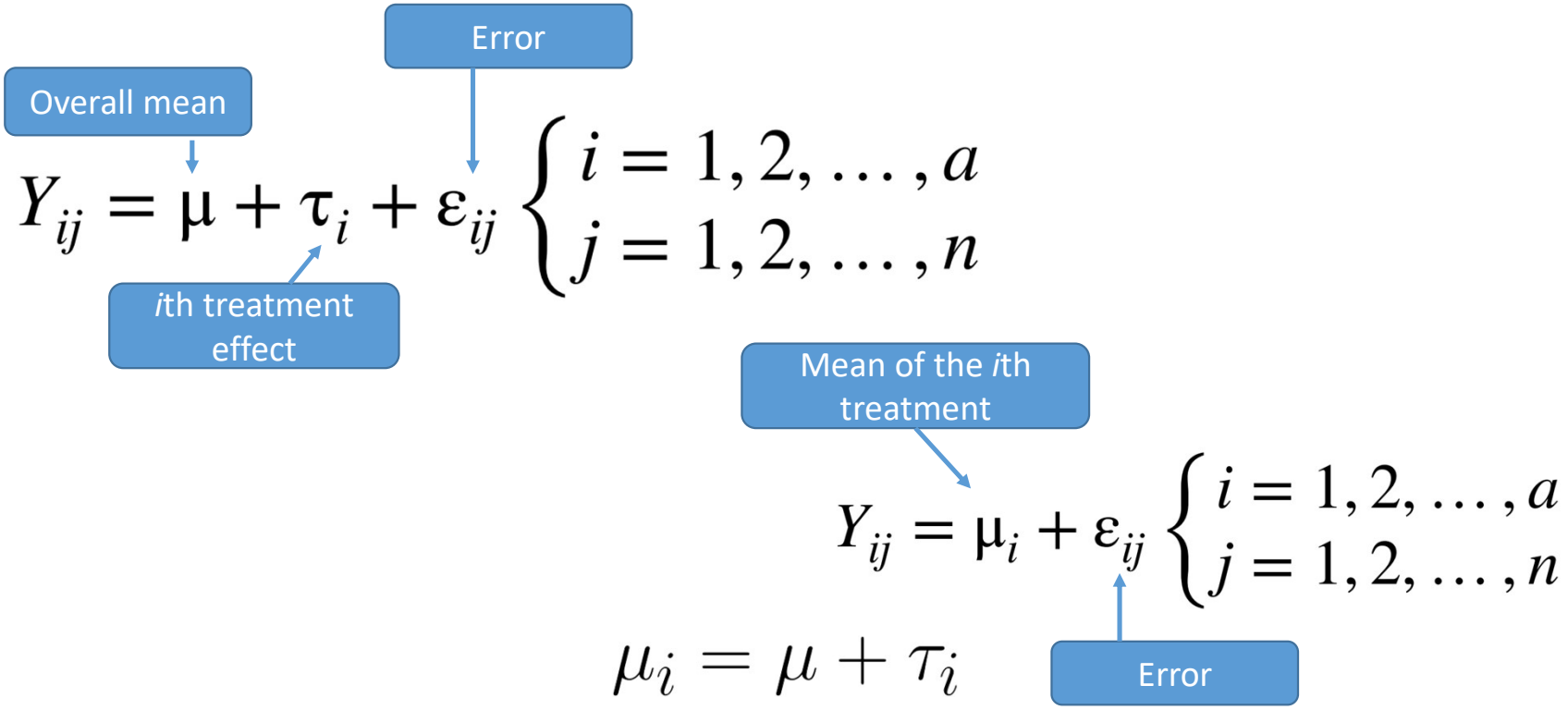
- Hardwood concentration seems to affect tensile strength.
- The distribution is near symmetric.
- The variability does not change.
- Box plots show the variability of the observations within a treatment (factor level) and the variability between treatments

Typical data for a single factor experiment

Treatment	Observations				Totals	Averages
1	y_{11}	y_{12}	$\cdots$	y_{1n}	$y_{1\cdot}$	$\bar{y}_{1\cdot}$
2	y_{21}	y_{22}	$\cdots$	y_{2n}	$y_{2\cdot}$	$\bar{y}_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots\vdots\vdots$	$\vdots$	$\vdots$	$\vdots$
a	y_{a1}	y_{a2}	$\cdots$	y_{an}	$y_{a\cdot}$	$\bar{y}_{a\cdot}$
					$y_{\cdot\cdot}$	$\bar{y}_{\cdot\cdot}$

- Each factor level is called a treatment.
- The response for each of the *a* treatments is a random variable.
- Each entry, y_{ij}, represents the *j*th observation taken under treatment *i*





- The error is assumed to be normally distributed, independent, with zero mean and variance σ^2
- Each treatment can be thought of as a normal population with mean μ_i and variance σ^2.

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij} \begin{cases} i = 1, 2, \ldots, a \\ j = 1, 2, \ldots, n \end{cases}$$

$$Y_{ij} = \mu_i + \varepsilon_{ij} \begin{cases} i = 1, 2, \ldots, a \\ j = 1, 2, \ldots, n \end{cases}$$

The treatment effects τ_i are usually defined as deviations from the overall mean μ,

$$\Rightarrow \sum_{i=1}^{a} \tau_i = 0$$

$$H_0: \tau_1 = \tau_2 = \cdots = \tau_a = 0 \qquad H_1: \tau_i \neq 0 \text{ for at least one } i$$

- If the null hypothesis is true, each observation consists of the overall mean μ plus a realization of the random error component ε_{ij}.
- This is equivalent to saying that all N observations are taken from a normal distribution with mean μ and variance σ^2.
- Therefore, **if the null hypothesis is true, changing the levels of the factor has no effect on the mean response**

ANOVA identity

Total sum of squares

Treatment sum of squares

Error sum of squares

$$\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{..})^2 = n\sum_{i=1}^{a}(\bar{y}_{i.}-\bar{y}_{..})^2 + \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^2,$$

$$SS_T = SS_{\text{treatments}} + SS_{\text{error}}$$

- $y_{i.} = \sum_{j=1}^{n} y_{ij}$ (total of the observations under the i-th treatment)
- $y_{..} = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij})$ (Grand total of all observations)
- $\bar{y}_{i.} = y_{i.}/n$ (average of the observations under the i-th treatment)
- $\bar{y}_{ii} = y_{..}/N$ (grand mean of all observations)

$$\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{..})^2 = n\sum_{i=1}^{a}(\bar{y}_{i.}-\bar{y}_{..})^2 + \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^2,$$

ANOVA identity

$$SS_T = SS_{\text{treatments}} + SS_{\text{error}}$$

$$an-1 = a-1+a(n-1)$$

$$df_{\text{total}} = df_{\text{treatments}} + df_{\text{error}}$$

Expected value for the mean square for treatments and errors

$$MS_{\text{treatments}} = \frac{SS_{\text{treatments}}}{a-1}$$

Mean square for treatments

$$MS_{\text{error}} = \frac{SS_{\text{error}}}{a(n-1)}$$

Mean square for errors

$$\mathbb{E}(MS_{\text{treatments}}) = \sigma^2 + \frac{n\sum_{i=1}^{a}\tau_i^2}{a-1}$$

$$\mathbb{E}(MS_{\text{error}}) = \sigma^2$$

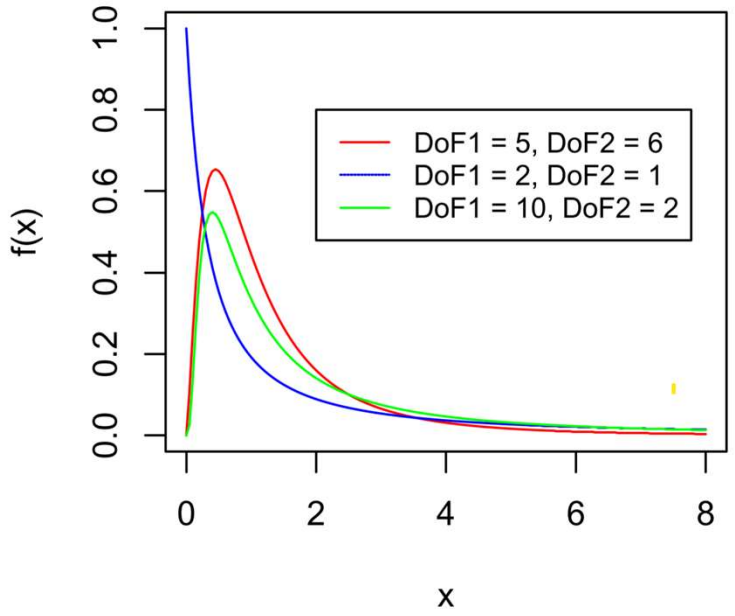
To check the null hypothesis:

$$F_0 = \frac{SS_{\text{treatments}}/(a-1)}{SS_{\text{error}}/[a(n-1)]} = \frac{MS_{\text{treatments}}}{MS_{\text{error}}}$$

What does this test actually tells us?

$$\mathbb{E}(MS_{\text{treatments}}) = \sigma^2 + \frac{n\sum_{i=1}^{a}\tau_i^2}{a-1}$$

$$\mathbb{E}(MS_{\text{error}}) = \sigma^2$$



H_0: $\tau_1 = \tau_2 = \cdots = \tau_a = 0$

H_1: $\tau_i \neq 0$ for at least one i

ANOVA table

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Treatments	$SS_{\text{Treatments}}$	$a-1$	$MS_{\text{Treatments}} = \dfrac{SS_{\text{Treatments}}}{a-1}$	$\dfrac{MS_{\text{Treatments}}}{MS_E}$
Error	SS_E	$a(n-1)$	$MS_{\text{Error}} = \dfrac{SS_{\text{Error}}}{a(n-1)}$	
Total	SS_T	$an-1$		

EXAMPLE 13.1 | Tensile Strength ANOVA

Consider the paper tensile strength experiment described in Section 13.2.1. This experiment is a completely randomized design. We can use the analysis of variance to test the hypothesis that different hardwood concentrations do not affect the mean tensile strength of the paper. The hypotheses are

$$H_0: \tau_1 = \tau_2 = \tau_3 = \tau_4 = 0 \qquad H_1: \tau_i \neq 0 \text{ for at least one } i$$

We use $\alpha = 0.01$. The sums of squares for the analysis of variance are computed as follows:

$$SS_T = 512.96$$
$$SS_{\text{Treatments}} = 382.79$$
$$SS_E = SS_T - SS_{\text{Treatments}} = 512.96 - 382.79 = 130.17$$

The ANOVA is summarized in Table 13.4. Because $f_{0.01,3,20} = 4.94$, we reject H_0 and conclude that hardwood concentration in the pulp significantly affects the mean strength of the paper. We can also find a P-value for this test statistic from computer software to be:

$$P(F_{3,20} > 19.60) \simeq 3.59 \times 10^{-6}$$

Computer software is used here to obtain the probability. Because the P-value is considerably smaller than $\alpha = 0.01$, we have strong evidence to conclude that H_0 is not true.

Practical Interpretation: There is strong evidence to conclude that hardwood concentration has an effect on tensile strength. However, the ANOVA does not tell us which levels of hardwood concentration result in different tensile strength means. We see how to answer this question in Section 13.2.3.

Hardwood Concentration (%)	Observations						Totals	Averages
	1	2	3	4	5	6		
5	7	8	15	11	9	10	60	10.00
10	12	17	13	18	19	15	94	15.67
15	14	18	19	17	16	18	102	17.00
20	19	25	22	23	18	20	127	21.17
							383	15.96

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	f_0	P-value
Hardwood concentration	382.79	3	127.60	19.60	3.59 E-6
Error	130.17	20	6.51		
Total	512.96	23			

HOMEWORK

Homework (deadline: 28 October 2022, 23:59)

1. Collect the height data of all participants in the Research Methodology course (approximate is fine). Let us call this data **"data set A".**
2. Collect the typing speed data of all participants in the Research Methodology course. Let us call this data **"data set B"**
3. For data set A and B, use descriptive statistics tools to summarize the data in a few important numbers.
4. For data set A and B, visualize the distribution of the data. What insight that you get from the distribution?
5. For data set A and B, calculate the quartiles, and 0.1, 0.2, 0.8 and 0.9-quantiles of your data
6. For data set A, compare the distribution of the height data with the mean height of the Indonesian people. Use statistical hypothesis test and answer: what inference that you obtain from the test?
7. For data set B, compare the distribution of the obtained typing speed with the mean typing speed of all the people who took the test. Use statistical hypothesis test.
8. Based on the data, is there any correlation between typing speed and body height?
9. Is there any difference between the typing speed of Male and Female students?

Fill out the data here: https://forms.gle/JKnVfoSHLGQXQkPs9

I will share the data here: https://github.com/optimuspram/Research_methodology_ITB_R

This is an **individual** *homework!*

Do not just give the answer in numbers of simple yes/no. Give the reason.

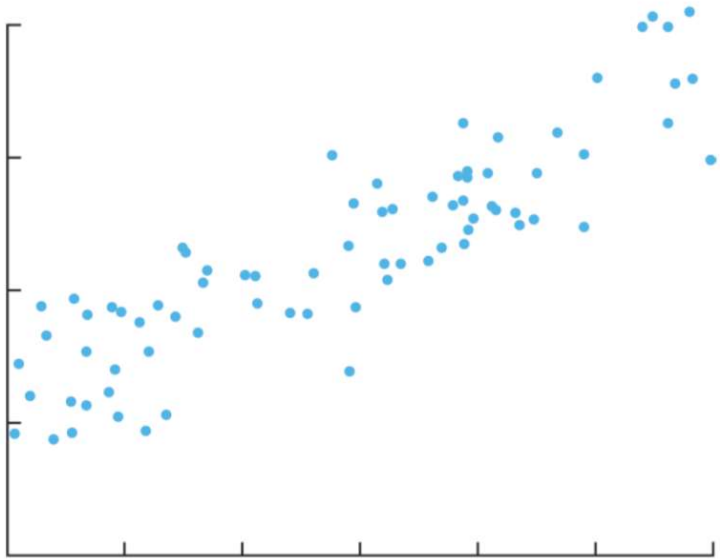
REGRESSION TECHNIQUES

Regression analysis

- **Regression analysis** is used to model the relationship between **a response variable** and **one or more predictor variables.**
- The goal is understand the **general trend** that relates the independent and dependent variable.
- We can also use regression analysis to **identify the impact of the input variables to the output.**

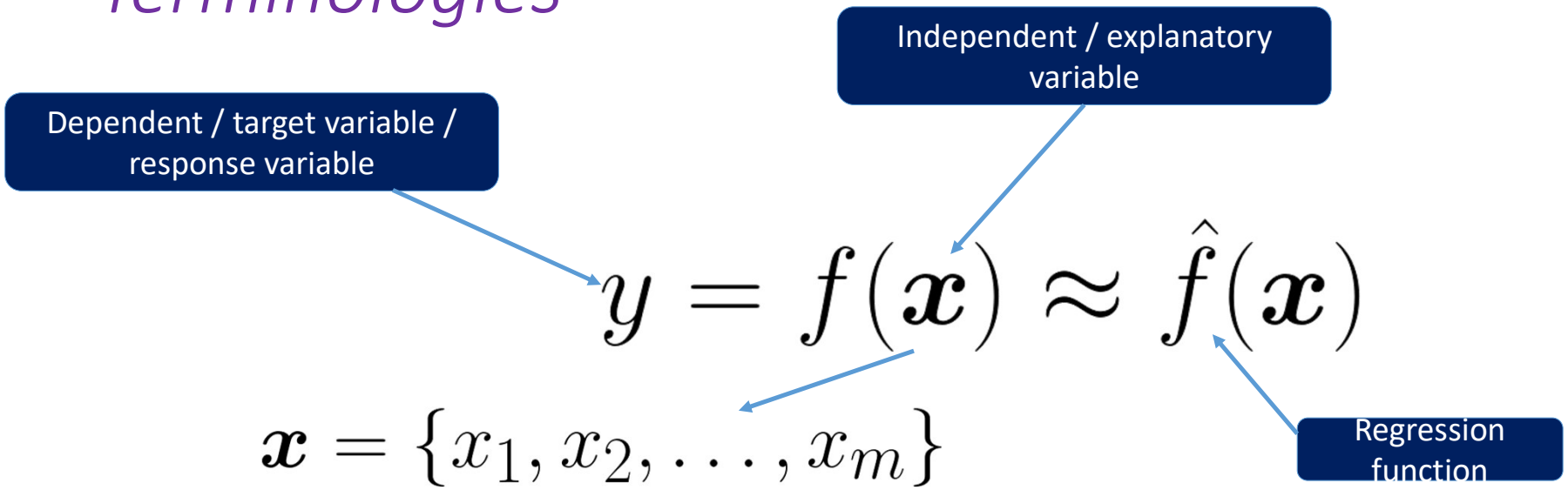
Regression is one of the fundamental technologies in statistics, AI, and science

Is there any relationship between these two variables?



SOURCE HBR.ORG

Terminologies



- ***f(x)*** is the true mathematical relationship that, in practice, we don't know what it's really like.
- $\hat{\boldsymbol{f}}$**(x)** is the approximation of the true mathematical relationship by a regression function.

Why do regression?

- ***Regression analysis*** eases the process of **identifying relationship by a quantitative measure.**
- Regression analysis can also give use the confidence of how accurate our regression function approximates the relationship.
- Regression allows us to make prediction (the goal of machine learning); but in statistics, **the goal is to uncover relationship between variables.**

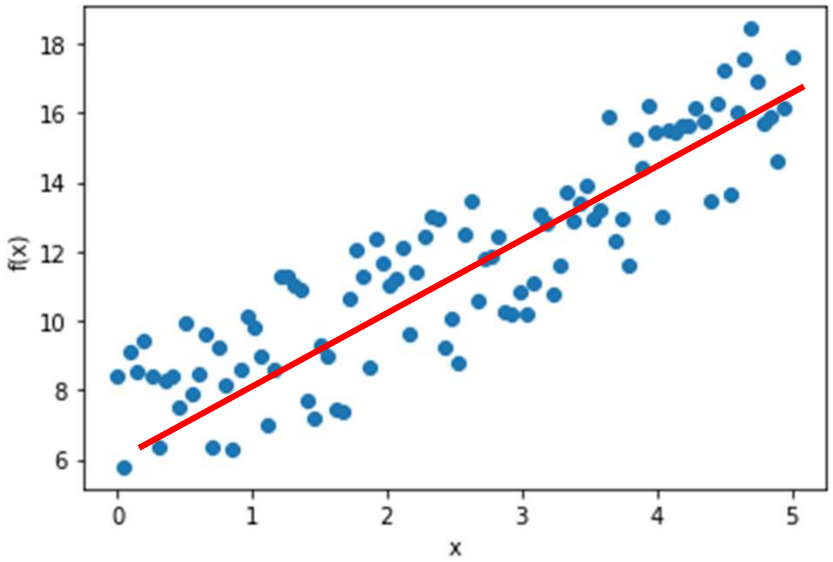
Thus, the following two aspects must be considered:

- **Predictive power:** Accuracy of the regression model in capturing the true function.
- **Interpretability**: The mathematical model of the regression function gives us insight regarding the trend/behavior the problem.

Correlation vs regression

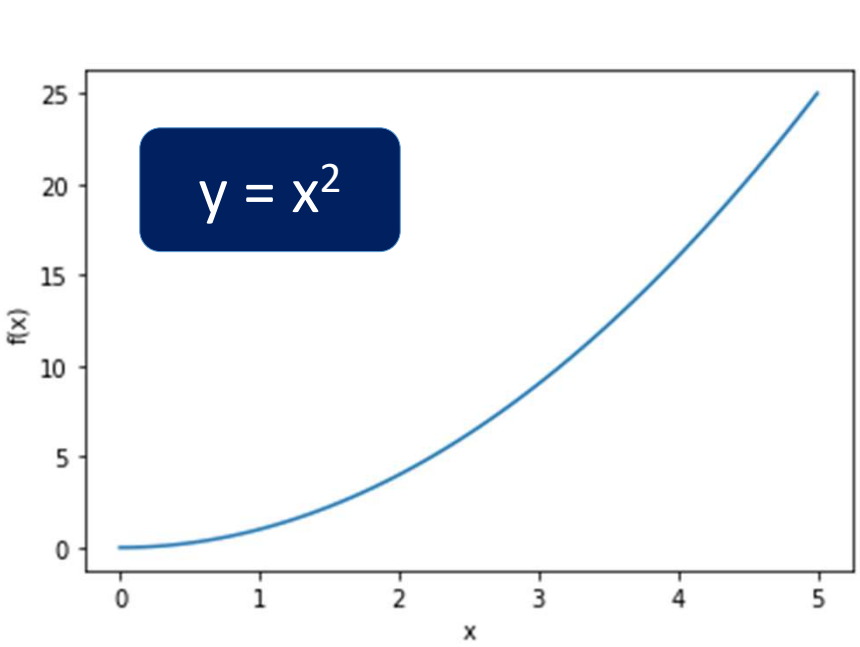
- ***Correlation*** is described as the analysis which lets us know the association of the relationship between two variables.
- ***Regression*** aims to predict the value of the dependent variable based on the known value of the independent variable.
- More precisely, regression describes how an independent variable is numerically related to the dependent variable

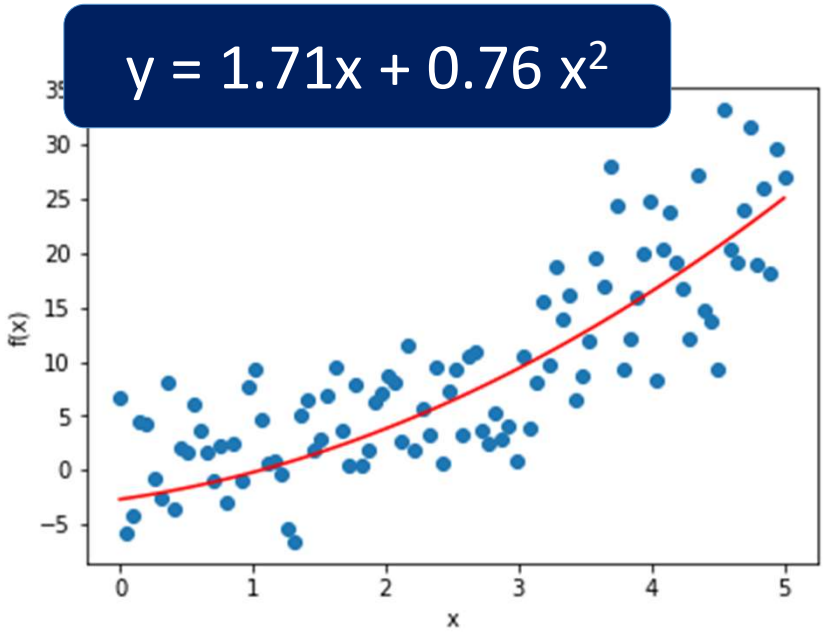
Correlation: If we increase *x,* will *f(x)* increase or decrease?



Regression: What is the best model that describes this data?

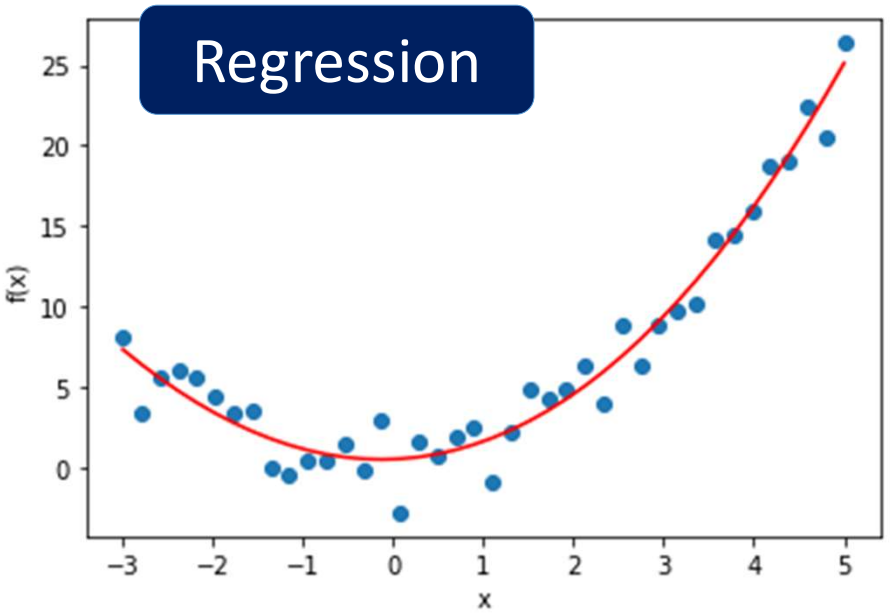
Deterministic vs statistical relationship

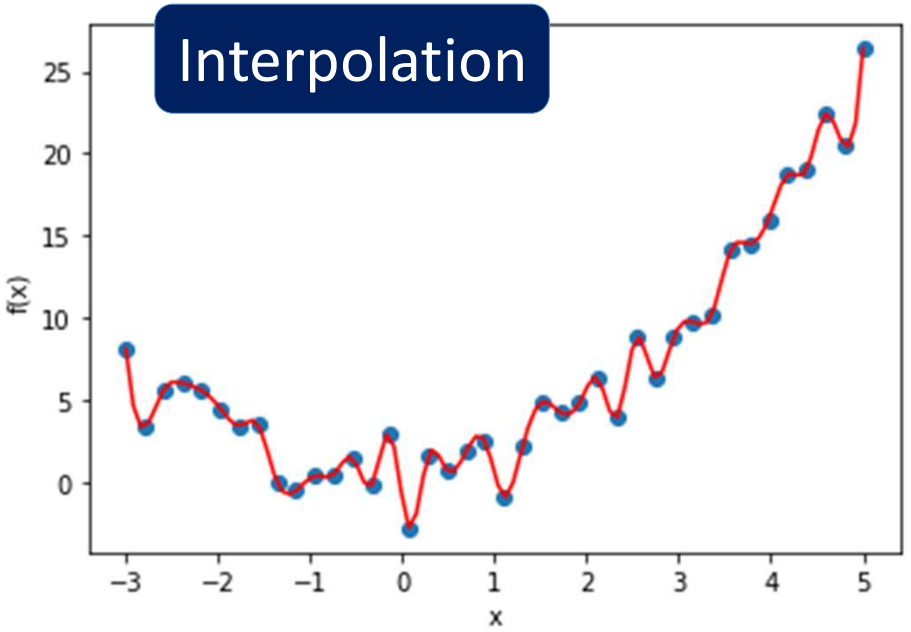




- ***Deterministic relationship*** is when the true relationship is EXACTLY known.
- We use ***Statistical relationship*** to model/approximate the deterministic relationship that is not known.

Interpolation vs regression

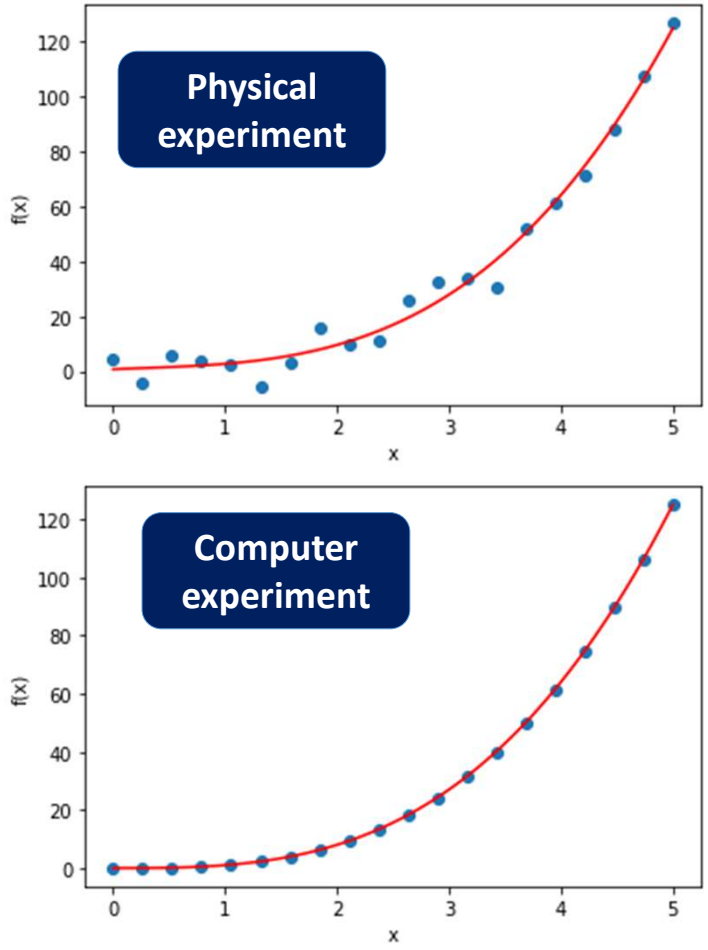




- **Interpolation** passes through sampling points.
- **Regression** does not passes through sampling points

Physical vs computer experiment

- **Physical experiments** observe the input/output at select samples using experimental apparatus.
- On the other hand, **computer experiments** use computer simulations.
- The outputs from physical experiment is typically corrupted by **noise**, while for experiment it is typically **smooth**.



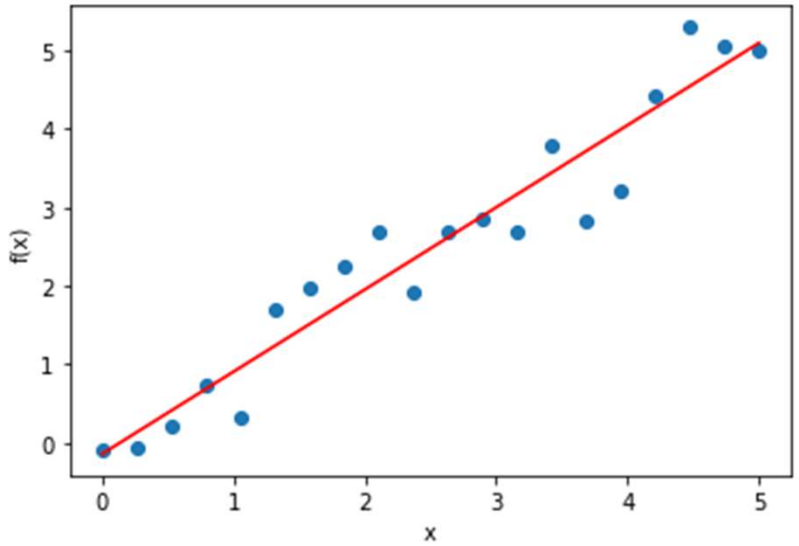
Linear regression

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Intercept

Slope

Error/noise



- ***Linear regression*** is used to find a linear relationship between a target and one or more predictor
- Linear regression is the simplest regression and particularly useful when the underlying relationship is linear.
- You can do it even with your calculator

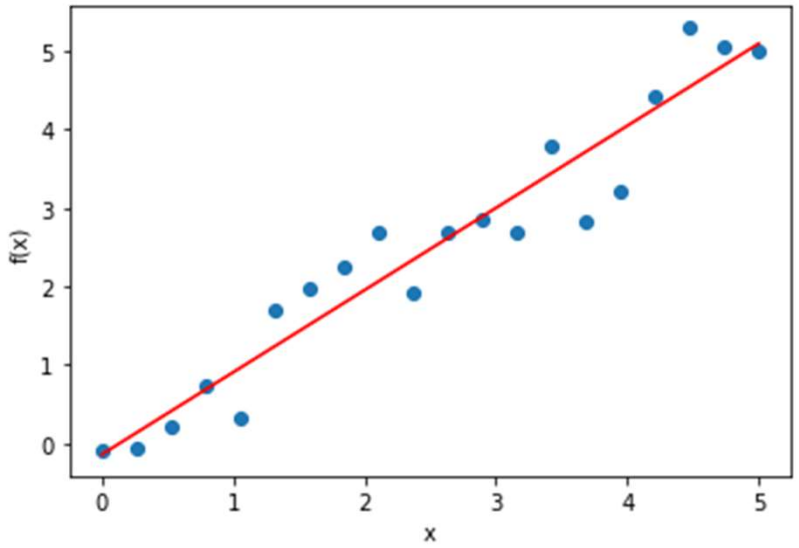
Linear regression: calculating the coefficients

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Intercept

Slope

Error/noise



- ***How to find the coefficients?***

Find the line that minimizes the error between that line and the data

Linear regression: calculating the coefficients

Error

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x)^2$$

$$y = \beta_0 + \beta_1 x + \varepsilon$$

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x) = 0$$

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial \beta_1} = -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x) = 0$$

A function reaches on optimum when the gradients are zero

$$\beta_1 = \frac{(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \qquad \beta_0 = \bar{y} - \beta_1 \bar{x}$$

Assessing the accuracy of the model

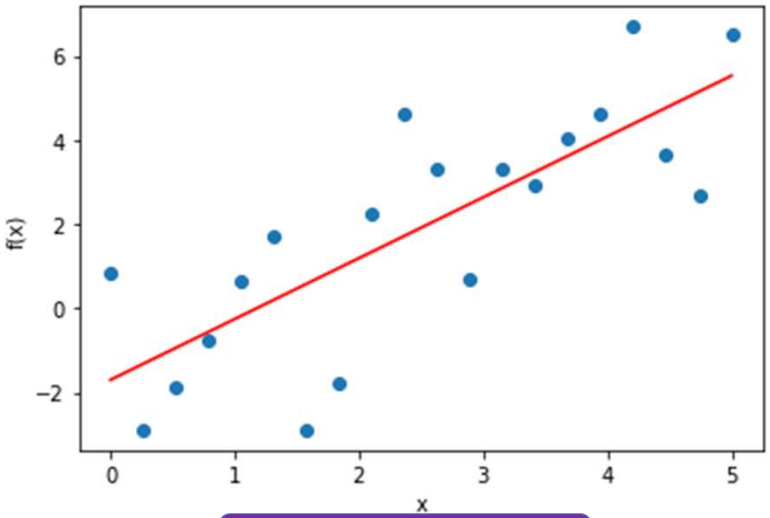
$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

$$\text{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

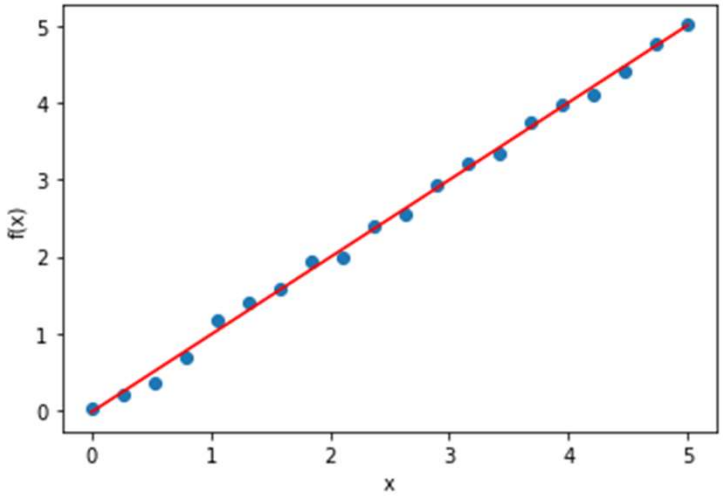
Residual sum of squares

$$\text{TSS} = \sum(y_i - \bar{y})^2$$

Total sum of squares



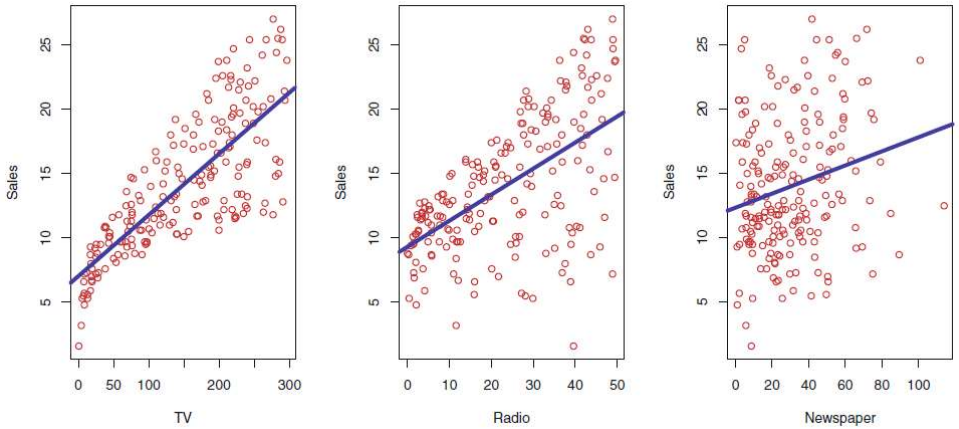
R-squared = 0.61



R-squared = 0.99

Multiple linear regression

- Often, most problems have two or more input variables.
- You can do a simple linear regression for each variable but that is **not really a proper method**.
- We need **multiple linear regression** that can handle **multiple variables**.
- However, the principle is still the same except now you have more terms and more coefficients.



Simple linear regression models (three different models, one for each variable)

The problems are..

- It is unclear how to make a single prediction given levels of three variables.
- Each of the three regression equations ignores the other two variables.
- If the variables correlate with each other, this simple approach can lead to misleading estimates.

Multiple linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

Average effect on *Y* of a one unit increase in X_j, *holding all other predictors fixed*

$$\texttt{sales} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{newspaper} + \epsilon$$

To find the coefficients

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

$$\begin{aligned} \text{RSS} &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2. \end{aligned}$$

A matrix algebra approach to obtain the coefficients for multiple linear regression

Linear regression matrix (n x p+1)

$$\boldsymbol{F} = (\mathbf{1}\ \boldsymbol{X})$$

$$\boldsymbol{F} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \\ \end{bmatrix}$$

Vector of outputs (n x 1)

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}.$$

Obtain the coefficients by solving the following system of linear equations

$$\boldsymbol{F}\boldsymbol{\beta} = \boldsymbol{y}$$

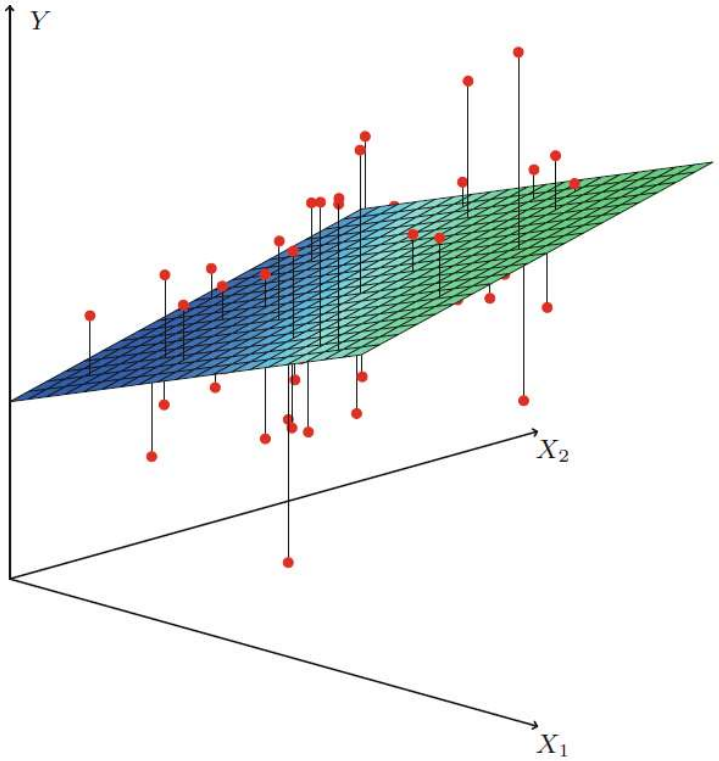
$$\hat{\boldsymbol{\beta}} = (\boldsymbol{F}^T\boldsymbol{F})^{-1}\boldsymbol{F}^T\boldsymbol{y}$$

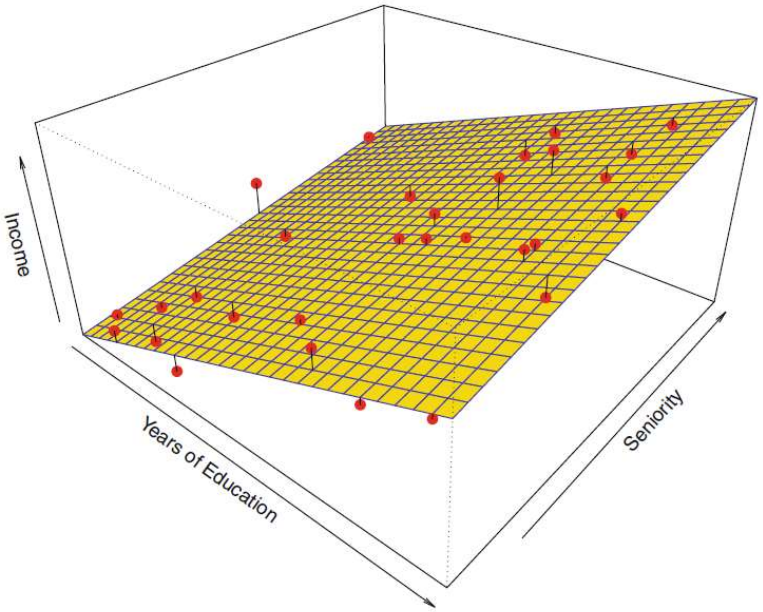
Linear regression coefficients (p+1 x 1)

$$\hat{\boldsymbol{\beta}} = \{\beta_0, \beta_1, \beta_2, \dots, \beta_p\}^T$$

- This method is called ordinary least squares. This basically finds the coefficients that minimize RSS.

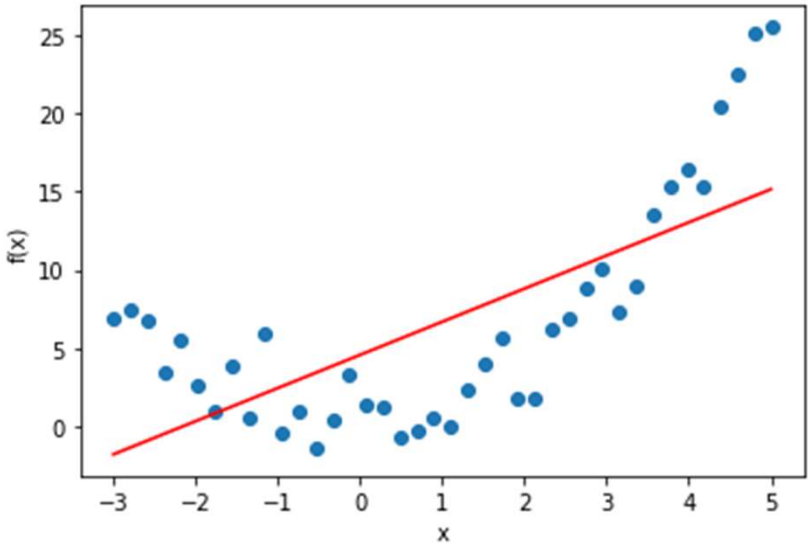
Multiple linear regression model





- Dengan dua variable input, multiple linear regression model berbentuk bidang dua dimensi

Limitation of linear regression



- Linear regression assumes that the true relationship in linear.
- Therefore, linear regression cannot capture nonlinear relationship!
- Solution: **We need more sophisticated methods.**

Polynomial regression

Polynomial regression

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \ldots + \beta_n x^n + \varepsilon$$

- **PR** models the relationship between independent and dependent variables by using *n*-th order polynomial
- PR is useful when the relationship does not exhibits a clear linear relationship

Linear model

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

Quadratic model

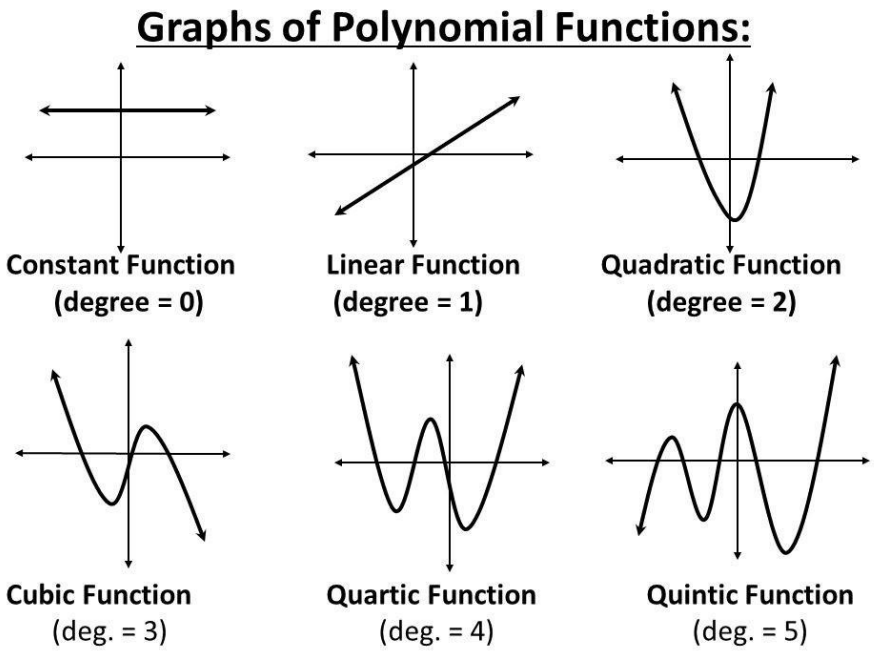
$$y = \beta_0 + \beta_1 x + \varepsilon$$

Polynomial regression

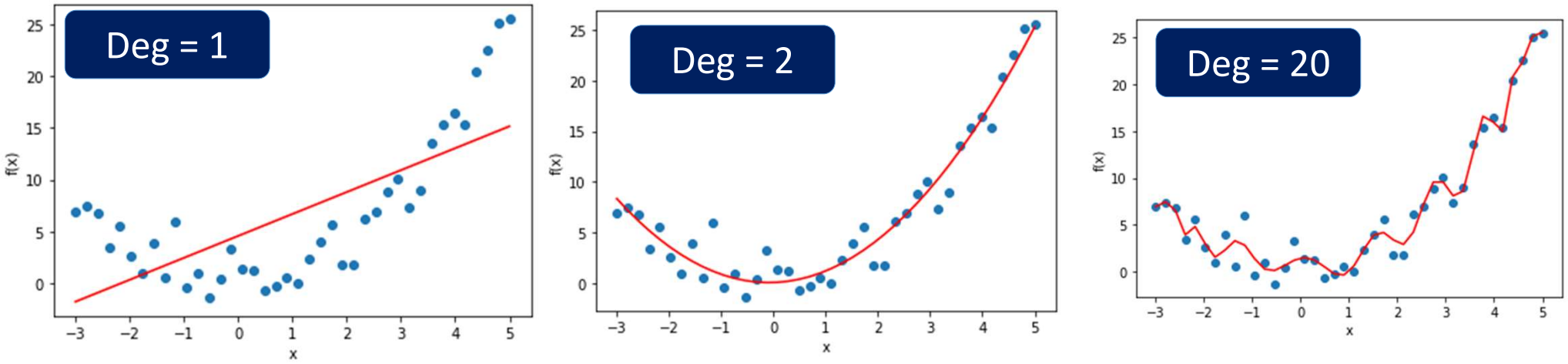
Polynomial regression

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \ldots + \beta_n x^n + \varepsilon$$

- The higher the order, the more complex the behavior of polynomial functions.
- The best order totally depends on the true behavior of the function and also noises



Polynomial regression

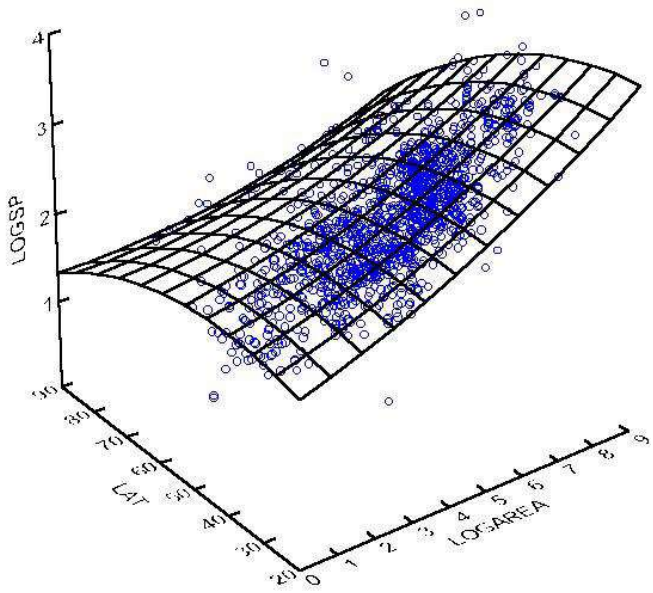


When using PR, be careful of two cases that might happen:

- **Underfitting:** The PR misses the true behavior of the unknown function.
- **Overfitting:** The PR captures the "trend" that should be there (e.g. noises)

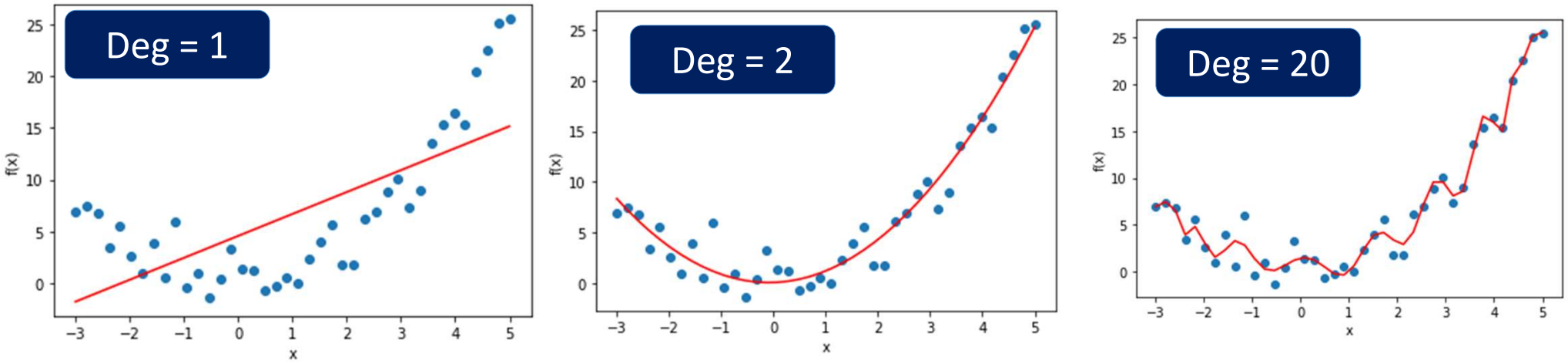
Multivariate linear regression

- The extension of multivariable linear regression.
- Capable of capturing non-linearity in high-dimensional space (high number of dependent variables).
- Might also involve interaction between independent variables



$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_1^2 + \beta_5 x_1 x_2 + \varepsilon$$

Polynomial regression

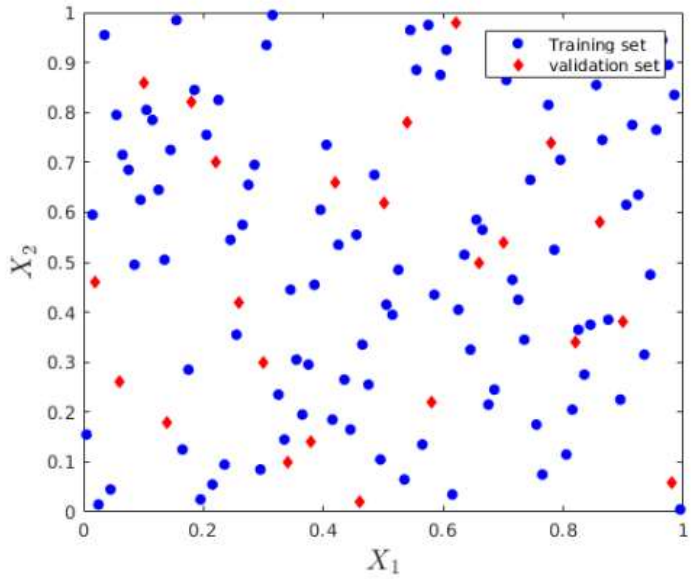


When using PR, be careful of two cases that might happen:

- **Underfitting:** The PR misses the true behavior of the unknown function.
- **Overfitting:** The PR captures the "trend" that should be there (e.g. noises)

Validation – Testing the accuracy of your model

- The goal of validation is to check the accuracy of our model.
- Validation is performed by testing the accuracy of the model on the validation set.
- The ideal condition is to have a separate training set (the one that is used to build the model) and validation set.



$$y = f(\boldsymbol{x}) \approx \hat{f}(\boldsymbol{x})$$

Use this regression model

$$e = \hat{y} - y$$

Use this on the validation set

Error metrics

- There are several available error metrics such as mean absolute error (MAE), root mean squared absolute error (RMSE), mean absolute relative error (MARE), and root mean squared relative error (RMSRE).
- RMSE gives more penalty to large error. This means the RMSE is most useful when large errors are particularly undesirable.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2} \qquad \text{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

- Which one is better? That depends on your objective.
- RMSE and MAE might not be easy to interpret when multiple cases are considered. In some cases, people use the relative value (RMSRE, MARE, normalized MAE, normalized RMSE)

Is your model accurate or not? Do cross-validation!

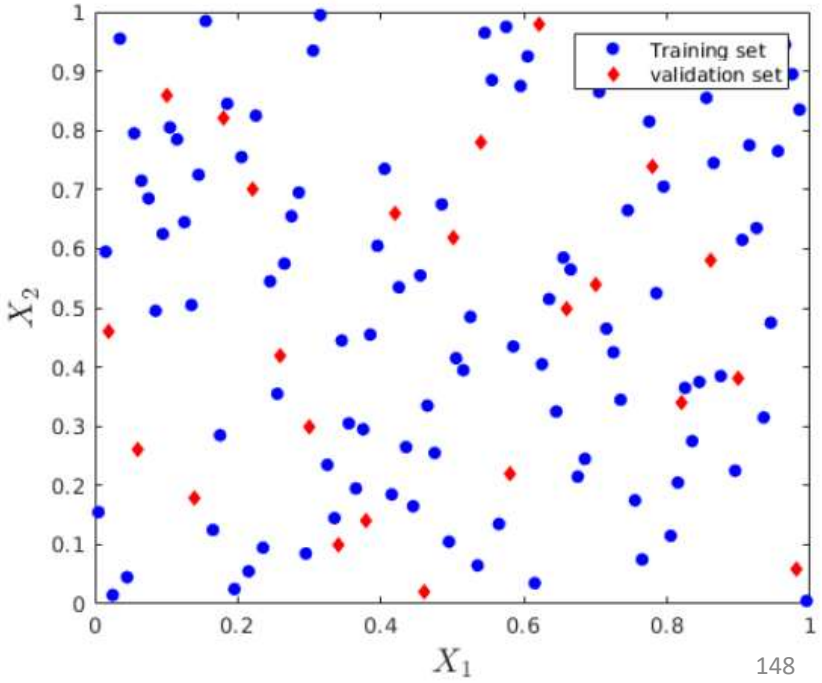
- Cross-validation (CV) is the way to assess the accuracy of surrogate models without external validation set.
- Why? Because in many cases, we don't have the external validation set.
- CV separates a training set (sampling points to create Kriging) into a smaller training set and a validation set.

CV techniques

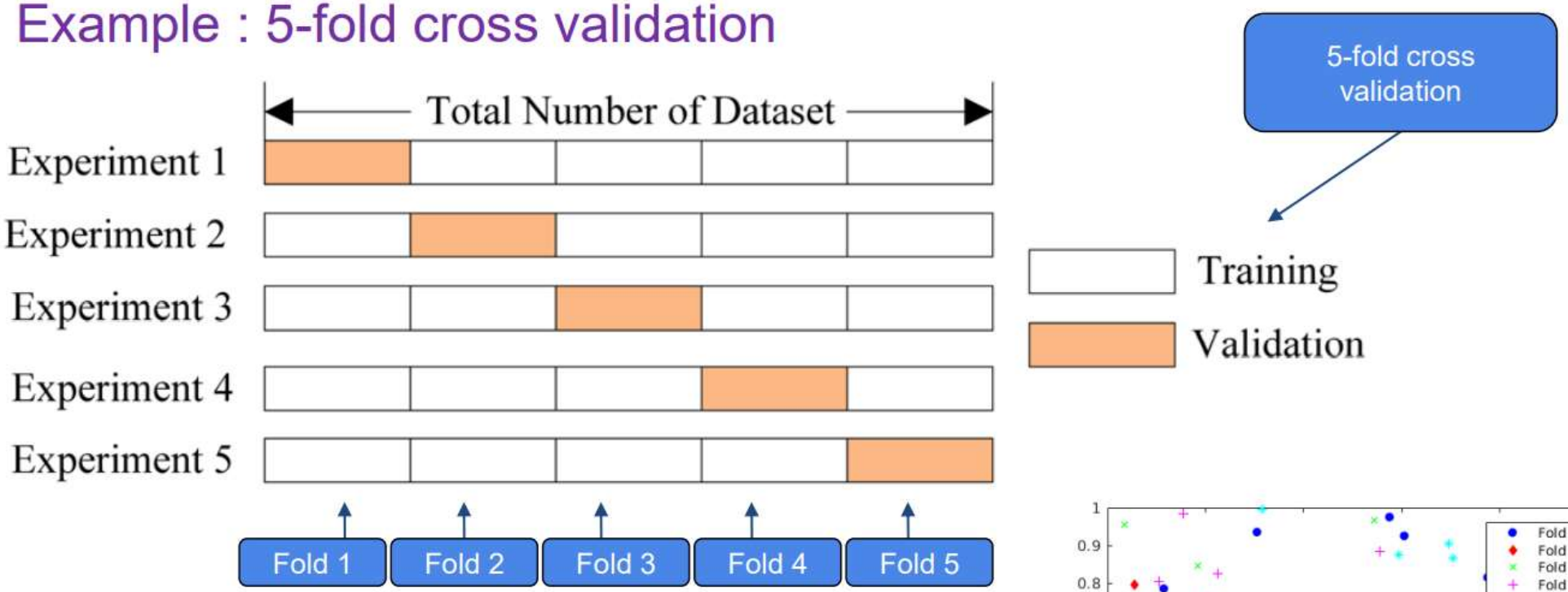
Two popular CV approaches:
- Leave-p-out CV.
- K-fold CV

Error metrics

- RMSE $\text{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$
- MAE
- R2 (coefficient of determination)
- Mean absolute relative error.
- Root-mean-squared-relative error.
- etc..

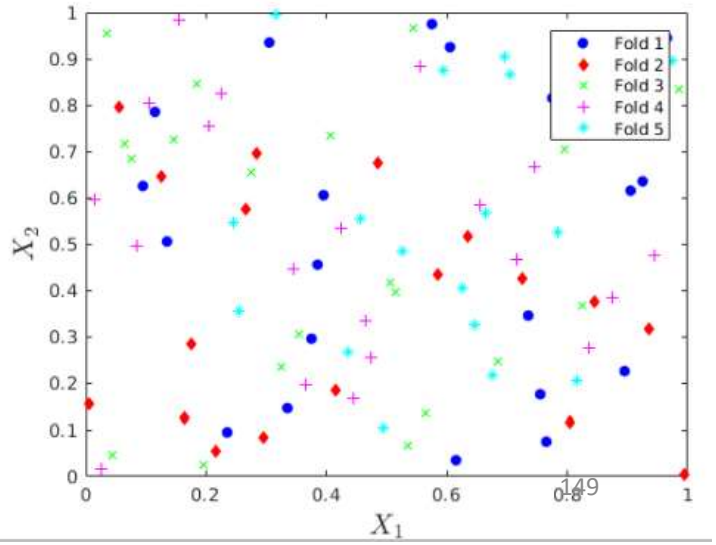


Example : 5-fold cross validation



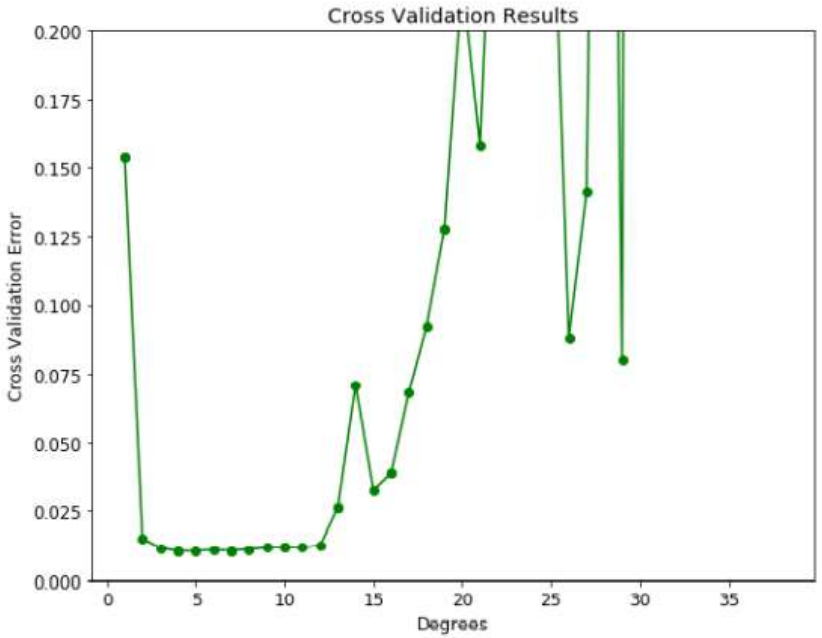
Steps:

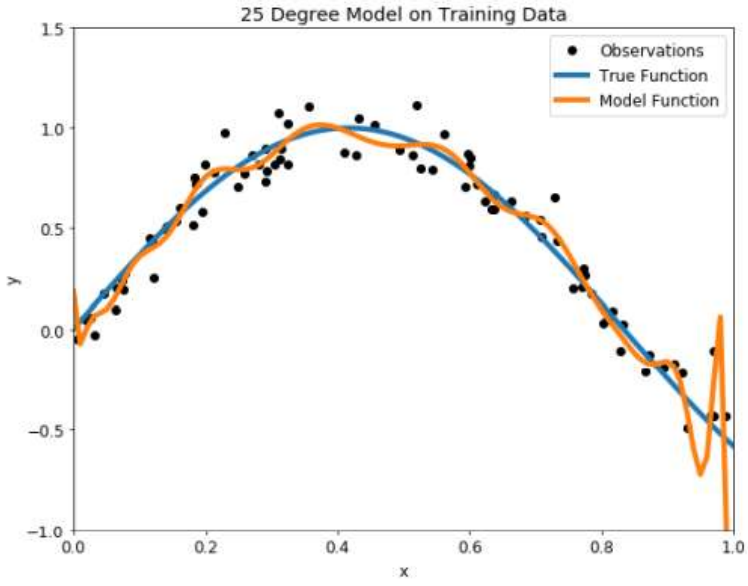
1. Use fold 2-5 to create a surrogate model, use fold 1 as validation samples. Compute error values.
2. Use fold 1,3,4,5 to create a surrogate model, use fold 2 as validation samples. Compute error values.
3. Continue until all folds have been used.
4. Average the error values.



How to avoid underfit/overfit? Do cross validation

	degrees	cross_valid
0	4	0.010549
1	5	0.010637
2	7	0.010665
3	6	0.010887
4	8	0.011182
5	3	0.011695
6	9	0.011757
7	11	0.011769
8	10	0.011902
9	12	0.012642



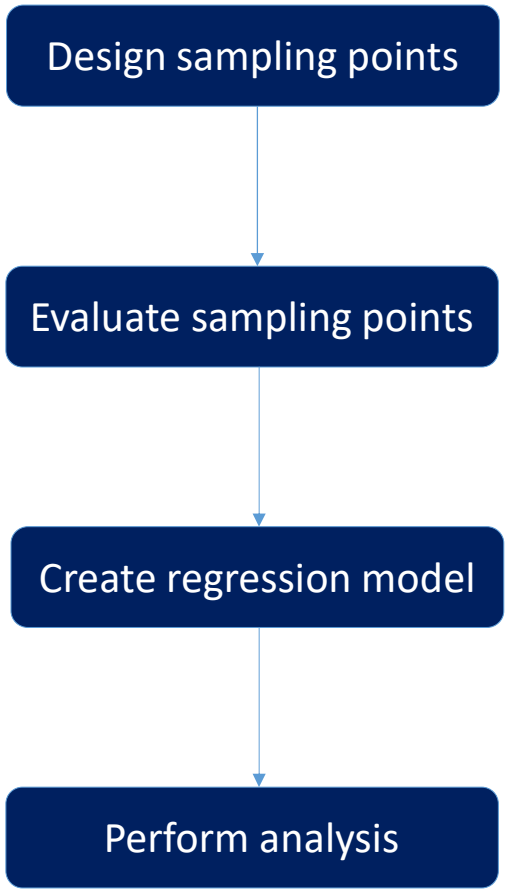


Designing your experiment

- Let's say that we want to perform experiments to find the relationship between ***x*** and ***y***
- A single experiment is costly!
- With limited budget, how can we maximize the information.

Typical objectives are:

- To find the relationship between ***x*** and ***y.***
- To identify important and less important independent variables

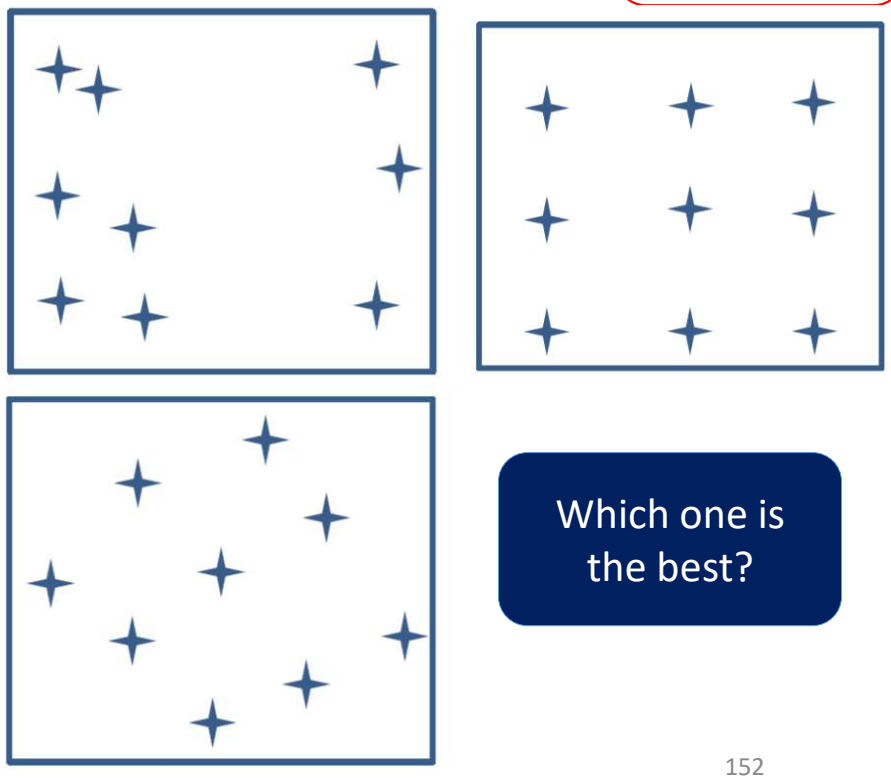


Designing your experiment

- The samples can be obtained via two means: (1) Sampling, (2) Observation, (3) data collection from different sources.
- Data from (1), in many cases, can be adjusted.
- Data from (3) are often highly scattered.
- Data from (2) can be of two types, adjusted or scatter.
- In statistics, it is often that we can do sampling by ourselves.

We need samples to construct this

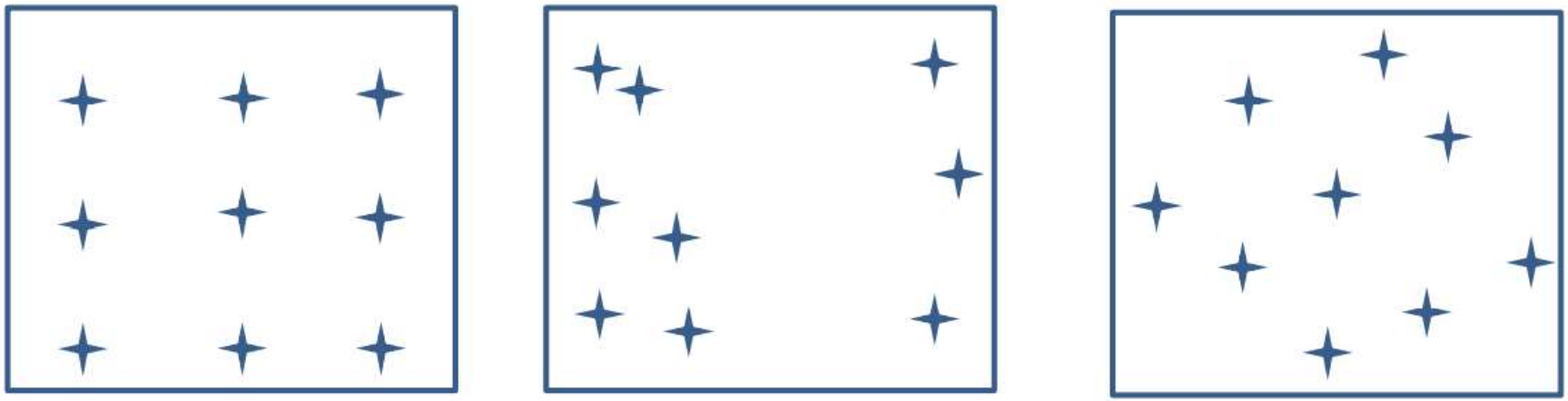
$$y = f(\boldsymbol{x}) \approx \hat{f}(\boldsymbol{x})$$



Which one is the best?

Sampling points

- It is important to decide the sampling points, i.e., the points where you will evaluate $\boldsymbol{y}$ (e.g. by CFD, FEM, or physical experiments).
- Basically, we want to maximize the information from limited available samples.
- To that end, the sampling points need to be carefully placed.

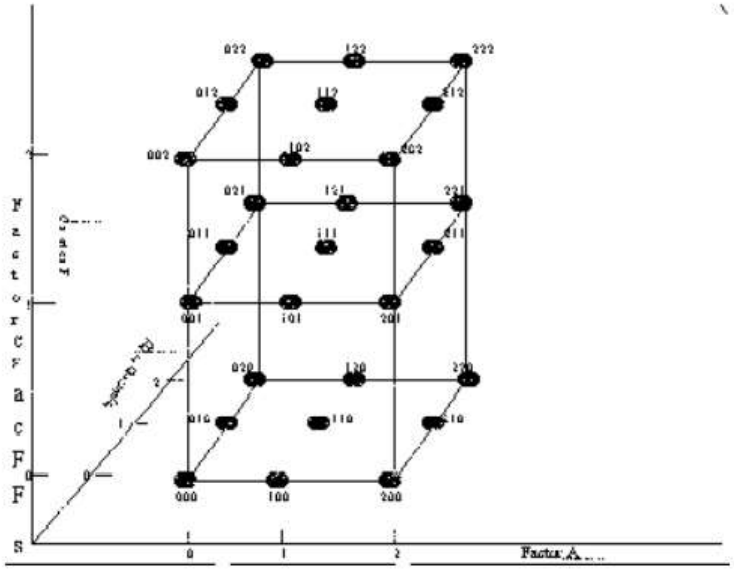


Which one that you think is the best?

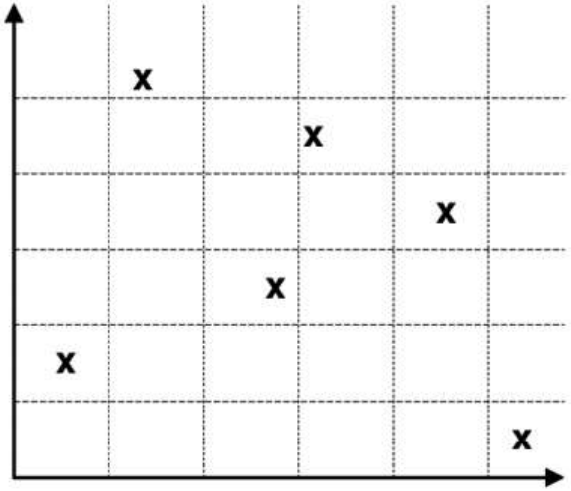
Sampling points

- There are several available techniques for generating sampling points.
- Among them are full factorial design, latin hypercube sampling, and low-discrepancy sequences.
- Good sampling points should 'fill' the space.

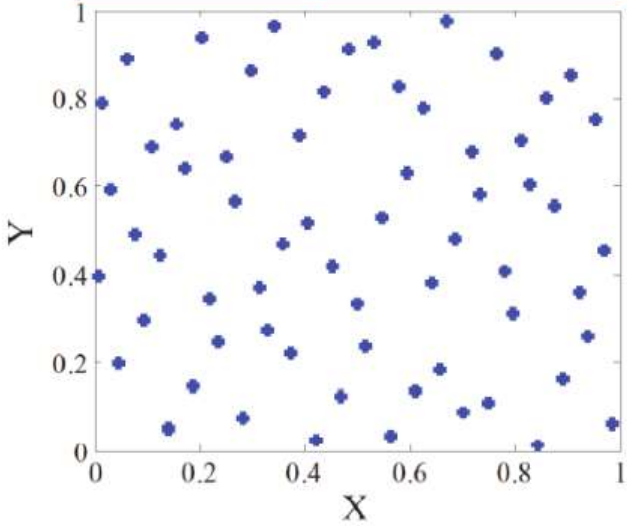
Full factorial design



Latin hypercube sampling

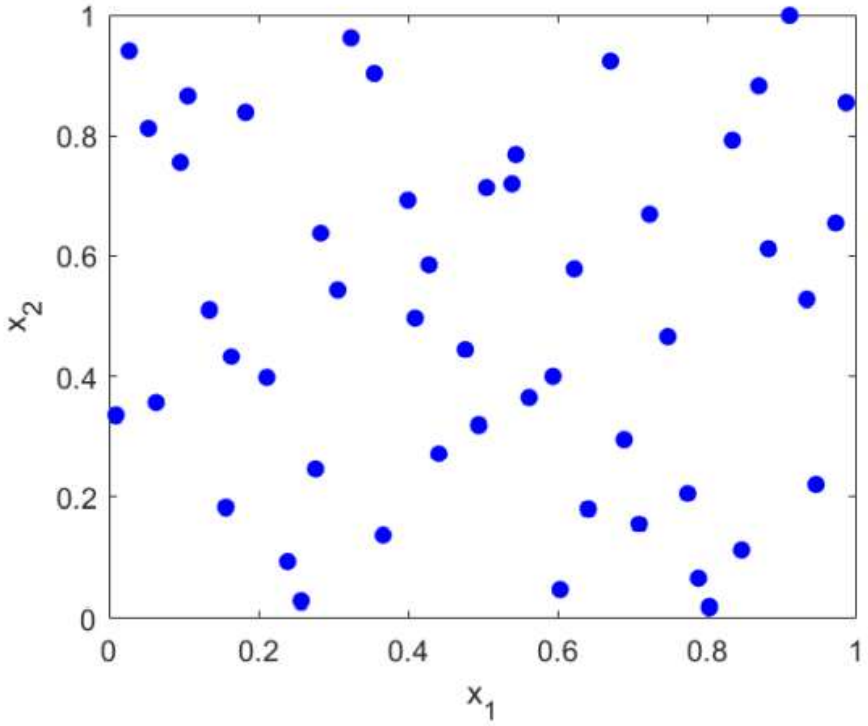


Low discrepancy sequence

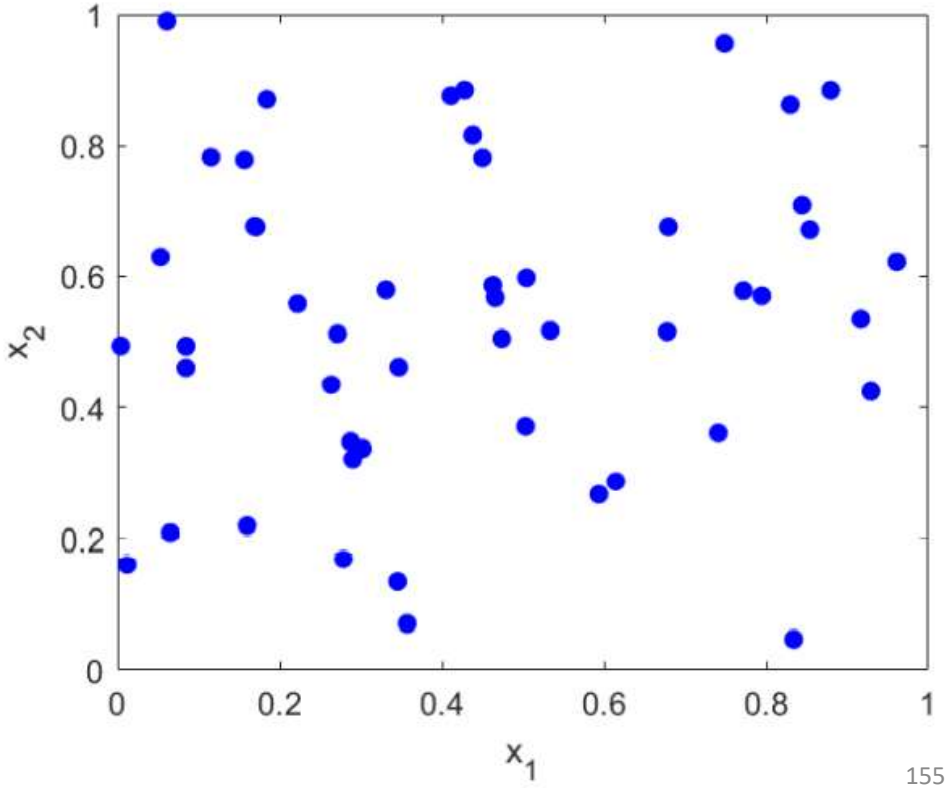


Sampling points: Example using MATLAB

LHS



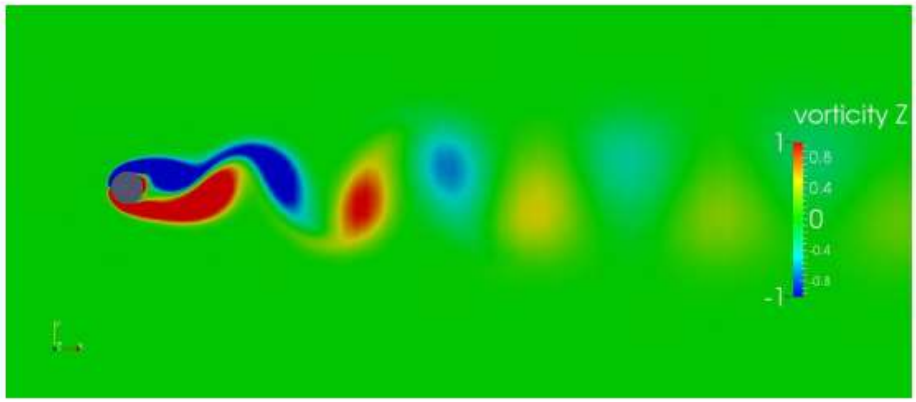
Random

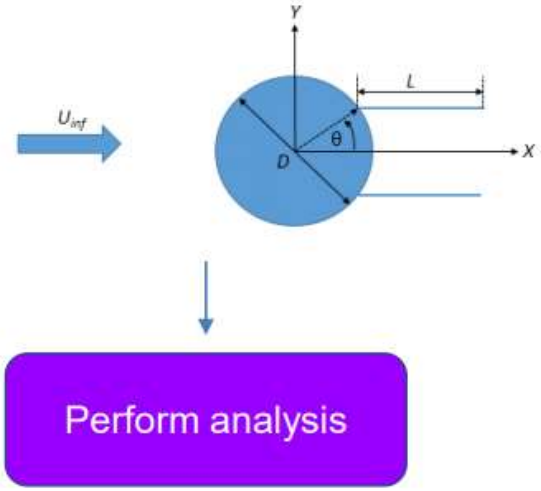


An example of computer experiment

- One of my former student designed a computer experiment to study the impact of adding a plate to reduce vortex induced vibration.
- He used GPR to create a regression model.
- His independent variables was the length and angle of the plate attached to the cylinder.
- His dependent variables (output of interests) were mean drag coefficient and vibration

SURROGATE ASSISTED GENETIC ALGORITHM FOR COMPUTATIONALLY EXPENSIVE FLUID FLOW OPTIMIZATION
Master Thesis, Yohanes Bimo Dwianto, Bandung Institute of Technology, Dept. of Aerospace Engineering



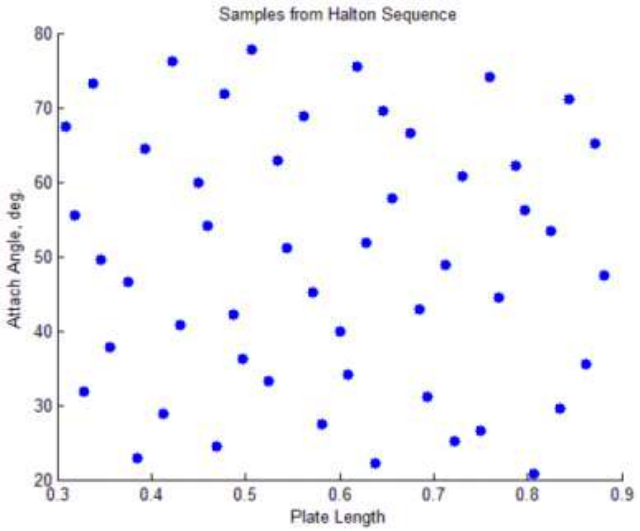


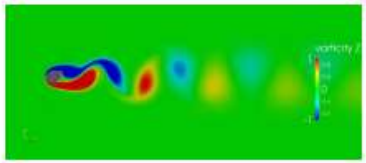
An example of computer experiment

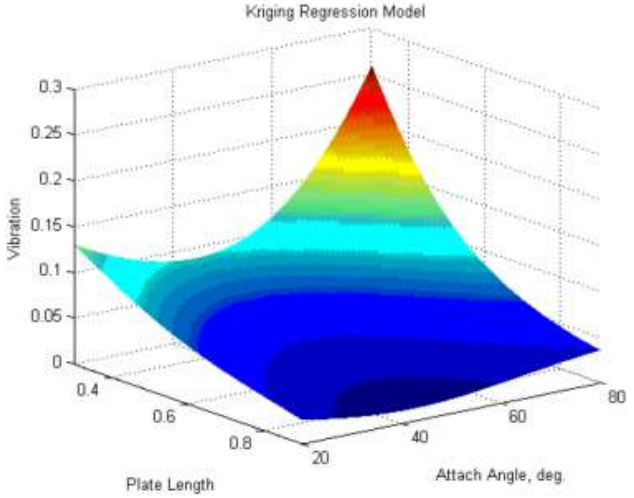
Design sampling points

Perform experiment

Create regression model







Analyze

Why you need to know all of these?