Basic Statistics for Data Science with Python

Nur Andi Setiabudi

July 9, 2021

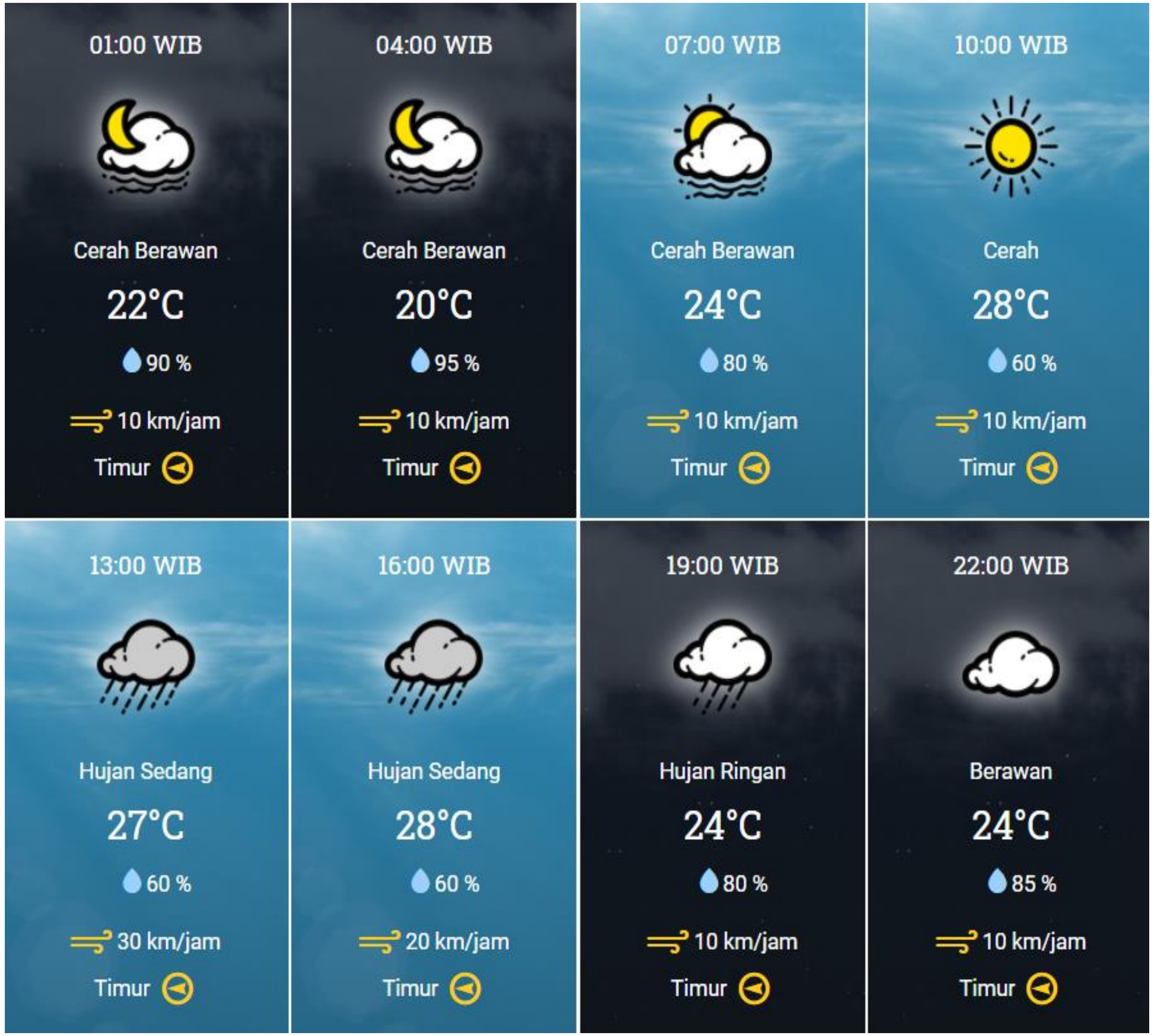
Food Tasting

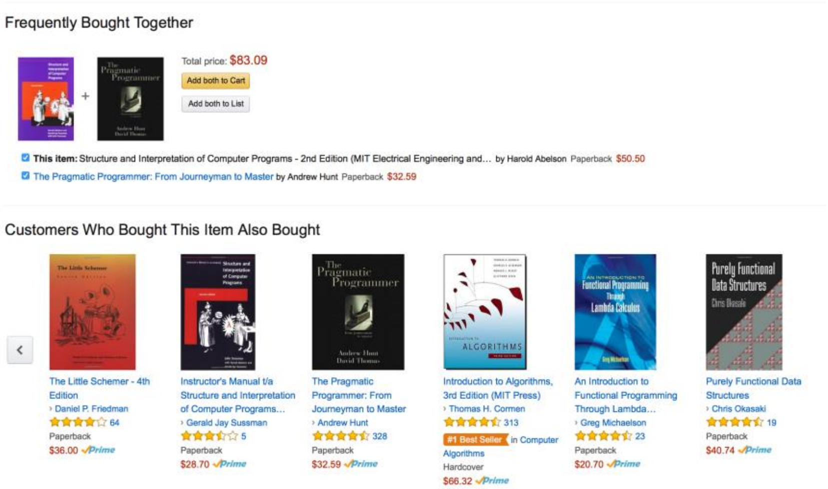
- Stir it well
- Pick only small portion of it
- Taste
- Conclude

That was

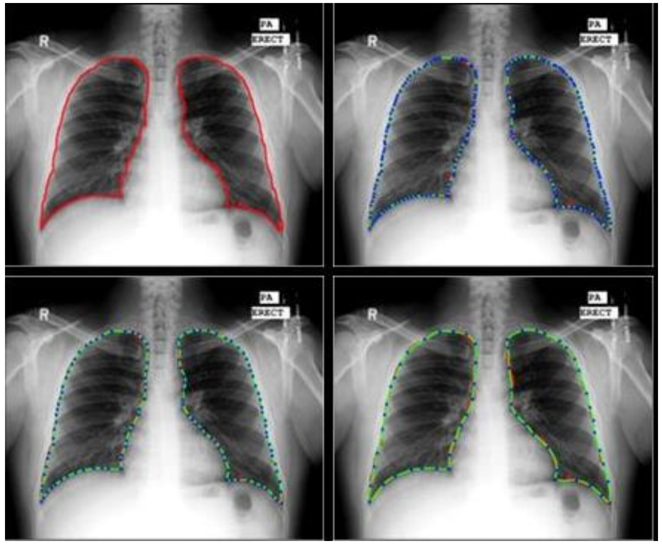
Statistics


MATCH STATISTICS
QUARTER-FINALS - FIRST LEG
LIVERPOOL
PORTO
FULL TIME
2 - 0
58% BALL POSSESSION 42%
3 ATTEMPTS ON TARGET 3
14 TOTAL ATTEMPTS 9
1 SAVES 3
4 CORNERS 5
4 OFFSIDES 2
107.45 km DISTANCE COVERED 109.26 km
539 (82%) PASSES COMPLETED 253 (68%)
9 FOULS COMMITTED 7
0 / 0 YELLOW/RED CARDS 2 / 0


01:00 WIB
Cerah Berawan
22°C
90 %
10 km/jam
Timur
04:00 WIB
Cerah Berawan
20°C
95 %
10 km/jam
Timur
07:00 WIB
Cerah Berawan
24°C
80 %
10 km/jam
Timur
10:00 WIB
Cerah
28°C
60 %
10 km/jam
Timur
13:00 WIB
Hujan Sedang
27°C
60 %
30 km/jam
Timur
16:00 WIB
Hujan Sedang
28°C
60 %
20 km/jam
Timur
19:00 WIB
Hujan Ringan
24°C
80 %
10 km/jam
Timur
22:00 WIB
Berawan
24°C
85 %
10 km/jam
Timur

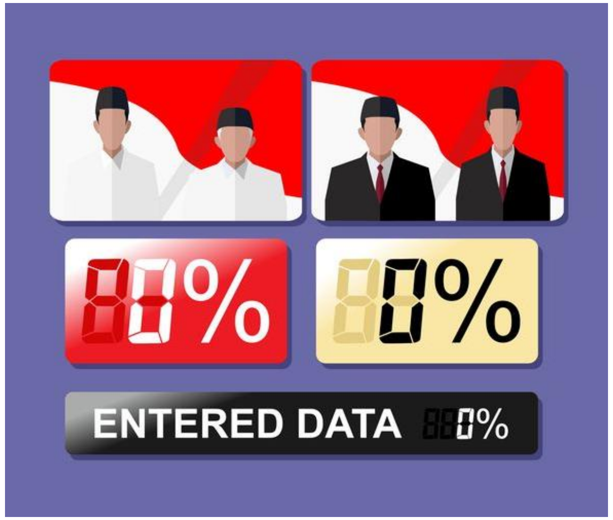

Frequently Bought Together
Total price: $83.09
Add both to Cart
Add both to List
This item: Structure and Interpretation of Computer Programs - 2nd Edition (MIT Electrical Engineering and… by Harold Abelson Paperback $50.50
The Pragmatic Programmer: From Journeyman to Master by Andrew Hunt Paperback $32.59
Customers Who Bought This Item Also Bought
The Little Schemer - 4th Edition
Daniel P. Friedman
64
Paperback
$36.00
Instructor's Manual t/a Structure and Interpretation of Computer Programs…
Gerald Jay Sussman
5
Paperback
$28.70
The Pragmatic Programmer: From Journeyman to Master
Andrew Hunt
328
Paperback
$32.59
Introduction to Algorithms, 3rd Edition (MIT Press)
Thomas H. Cormen
313
#1 Best Seller in Computer Algorithms
Hardcover
$66.32
An Introduction to Functional Programming Through Lambda…
Greg Michaelson
23
Paperback
$20.70
Purely Functional Data Structures
Chris Okasaki
19
Paperback
$40.74


A
B




HENDERSON LAND (00012)
AASTOCKS.com Copyright ©
3 months (daily)
07/10/2020 O:29.0 H:29.25 L:28.75 C:29.2
SMA(10):28.59 SMA(20):28.84 SMA(50):29.421
31.275
30.575
29.875
29.175
28.475
Aug
Sep
Oct
Turnover = 45.594M SMA (20) = 85.274M Close Up Close Down
183.003M


CREDIT SCORE
720


ENTERED DATA

But, statistics (and also data science, of course) ***is not a magic***

Definition

What is statistics?

- The practice or **science** of **collecting** and **analyzing** numerical **data** in large quantities, especially for the purpose of **inferring proportions** in a whole from those in a **representative sample** – Oxford
- The **science** of learning from **data**, and of **measuring**, **controlling** and **communicating uncertainty** – American Statistical Association

Basic steps in statistics

Studying a problem through the use of statistical data analysis usually involves four basic steps.

- Defining the problem
- Collecting/preparing the data
- Analyzing the data
- Reporting the results

Population and Sample

Statistics

Sampling

Inference

Sample

Descriptive

Population
(Unkown)

Technology make it possible to gather and analyze whole population

Population and sample

- Data consists of information coming from observations, counts, measurements, or responses
- A population is the entire group that you want to draw conclusions about.
- A sample is the specific group that you will collect data from. The size of the sample is always less than the total size of the population.

Sample → **Population**

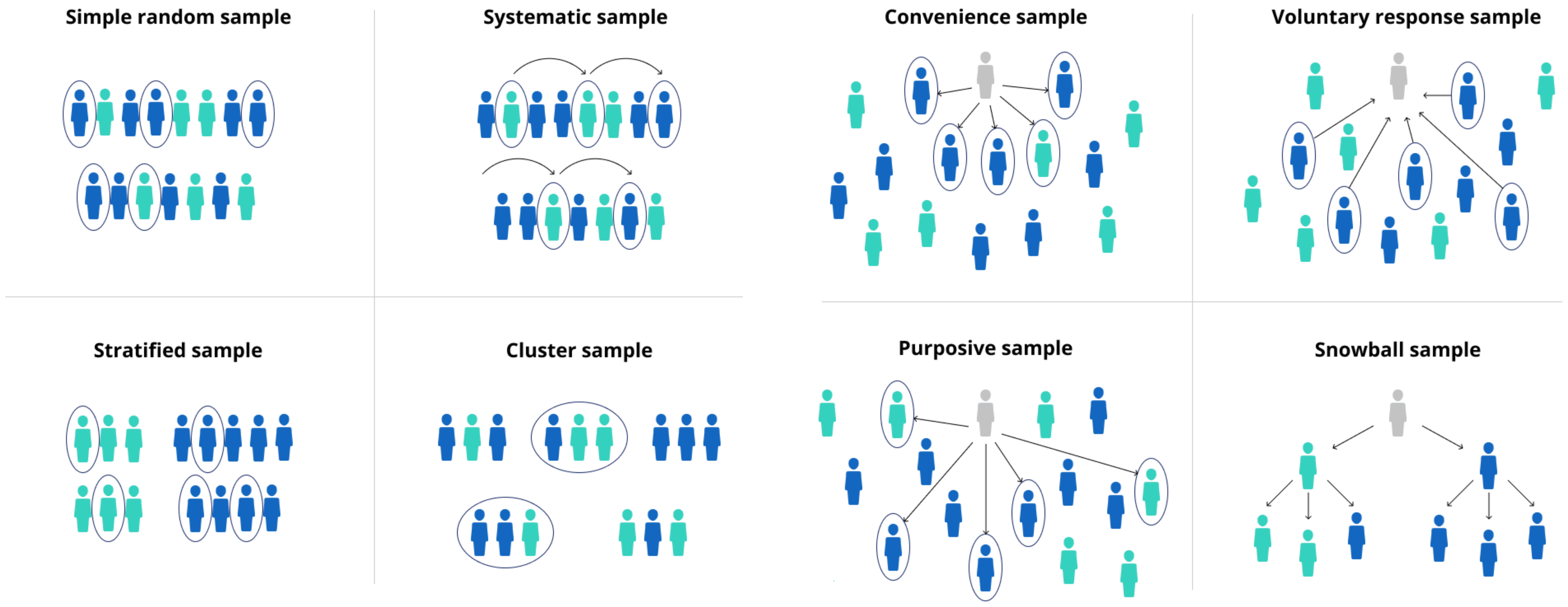
Statistics → Parameter

- A parameter is a numerical description of a *population* characteristic.
- A statistic is a numerical description of a *sample* characteristic.

Characteristics of good sample

- **Representative** of population: should be an accurate representative of the universe from which it is taken
- **Random selection**: should be selected at random. This means that any item in the group has a full and equal chance of being selected and included in the sample. This makes the selected sample truly representative
- **Sampling error** can be quantified. **Non sampling error** is corrected for as much as possible
- **Economical**: should be achieved with minimum cost and effort
- **Practical**: should be capable of being understood and followed in the fieldwork

Sampling method



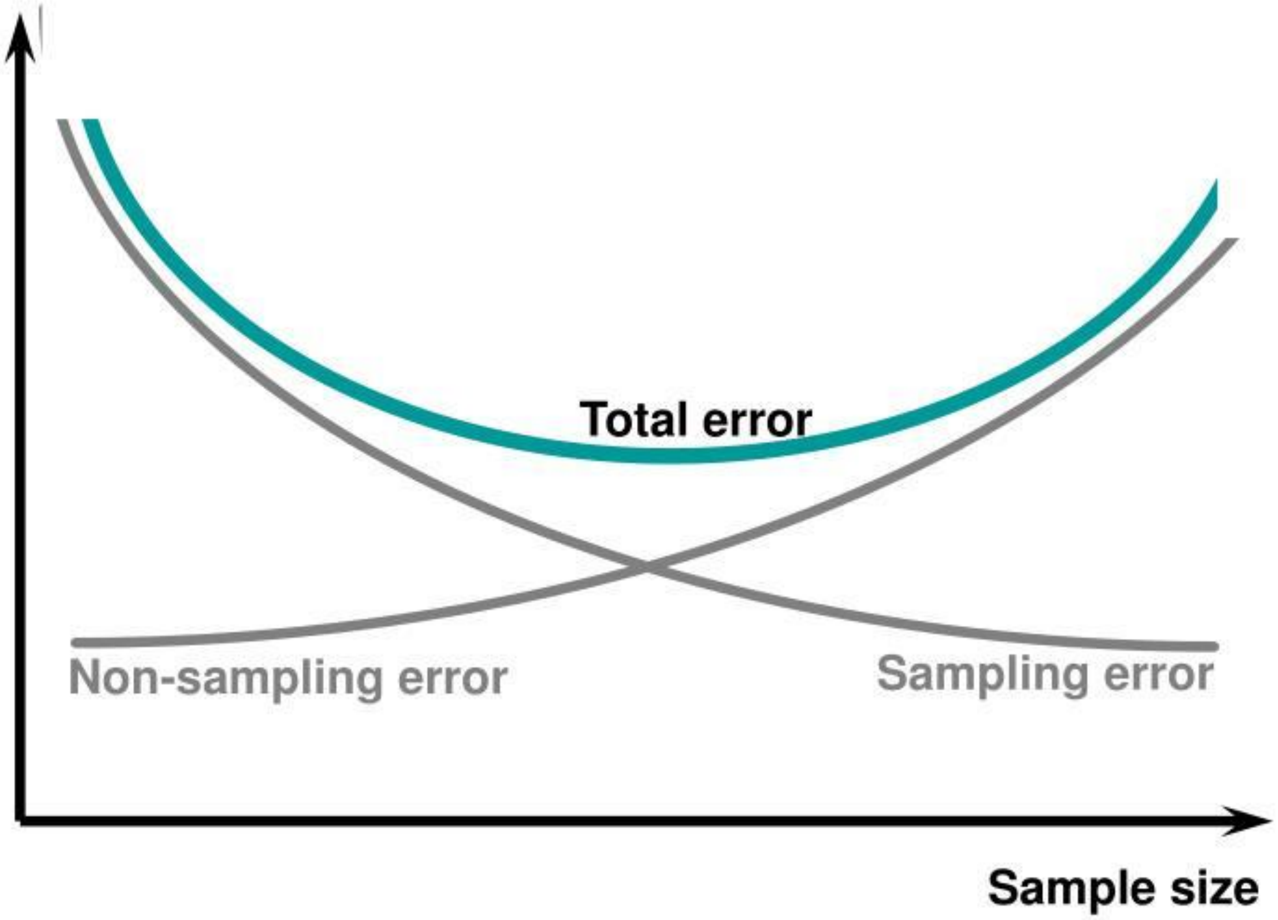
https://www.scribbr.com/methodology/sampling-methods/

Error and bias

Potential source of error
in estimating population parameter using sample

Sampling error	Non-sampling error		
Sample is not the whole population	Behavioral effect	Questionnaire and tools error	Poor sampling method

Error vs sample size



Two Branches of Statistics

Statistics

Descriptive statistics

Involves the organization, summarization, and display of data

- **Frequency, count**
- **Basic measurements**
- **Charts**

Inferential statistics

Involves using a sample to draw conclusions about a population

- **Estimation**
- **Hypothesis testing**
- **Confidence Interval**
- **Significance testing**
- **Modeling**

Descriptive Statistics

Measurement level

Qualitative		Quantitative	
Nominal	Ordinal	Interval	Ratio
Named	Named	Named	Named
Color **Gender**	Natural order	Natural order	Natural order
	Education **Satisfaction level**	Equal interval between variables	Equal interval between variables
		Temperature **Exam score**	Has a "true zero" value, thus ratio between values can be calculated
			Age **Monthly income**

Descriptive statistics

- Descriptive statistics is the term given to the analysis of data that helps describe, show or summarize data in a meaningful way such that, for example, patterns might emerge from the data.
- Descriptive statistics **do not**, however, allow us to make conclusions beyond the data we have analyzed or reach conclusions regarding any hypotheses we might have made.
- They are simply a way to describe our data.
- **Measures of central tendency:** these are ways of describing the central position of a frequency distribution for a group of data
- **Measures of spread (dispersion):** these are ways of summarizing a group of data by describing how spread out the scores are

Measures of central tendency

ways of describing the central position of a frequency distribution for a group of data

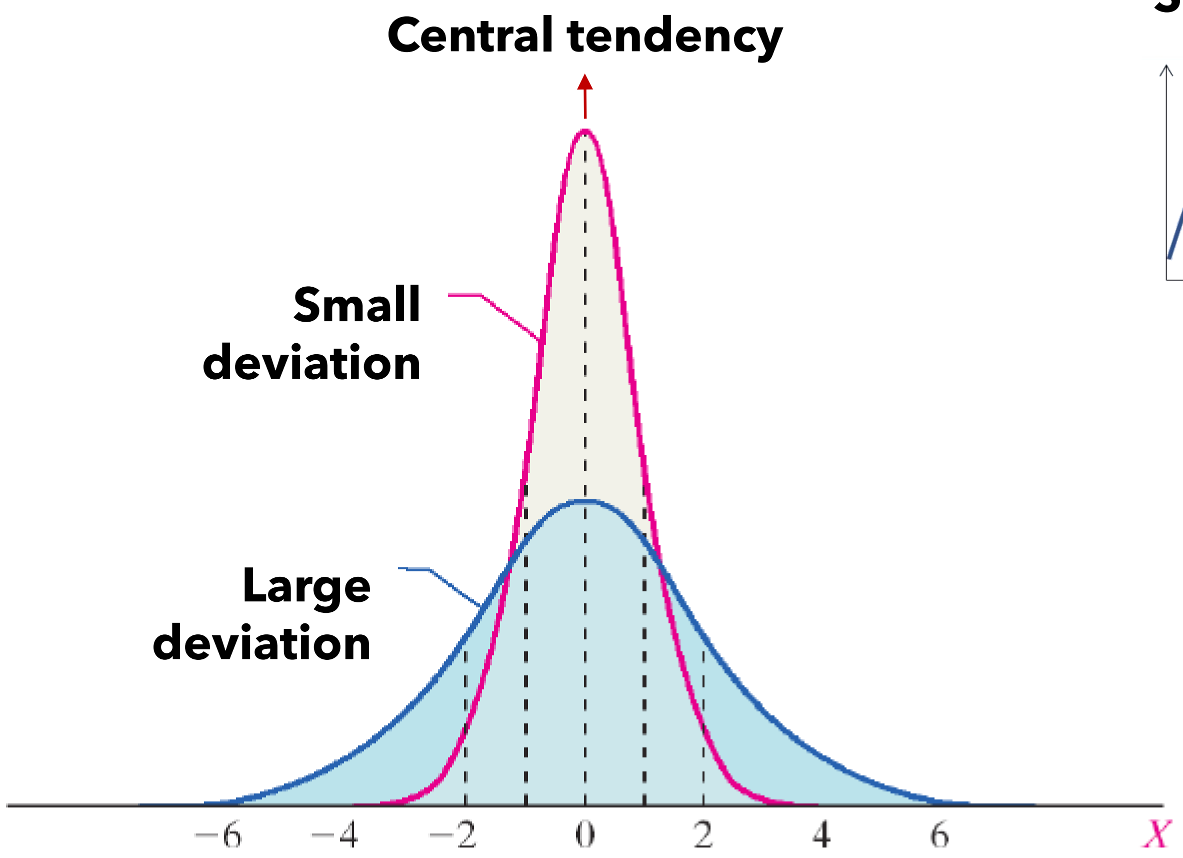
Measure of Central Tendency	Appropriate to choose when …	Should not be used when…
Mean the balance point of a data distribution	No situation precludes it	• Extreme scores • Skewed distribution • Ordinal scale • Nominal scale
Median the midpoint of a data distribution	• Extreme scores • Skewed distribution • Ordinal scale	• Nominal scale
Mode score or category that has the greatest frequency	• Nominal scales • Discrete variables • Describing shape	• Interval or ratio data, except to accompany mean or median

Measures of dispersion

Summarizing a group of data by describing how spread out the scores are

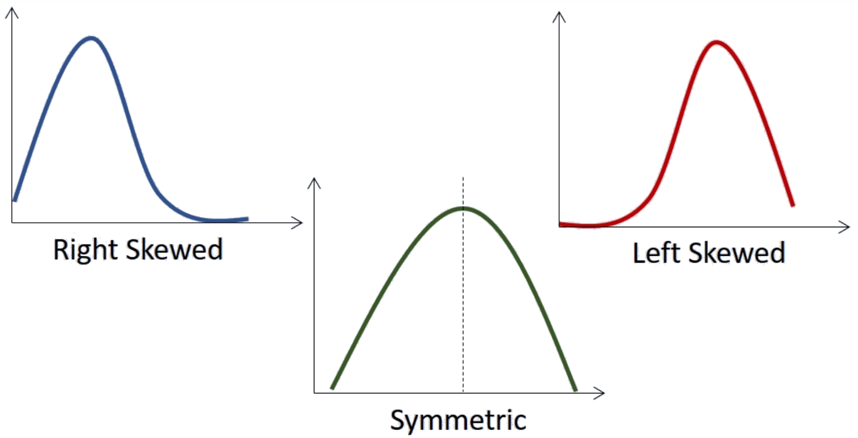
- **Range**: The interval between the highest and lowest measures
- **Percentile**: The value below / above which a particular percentage of values fall
- The **standard deviation** is a statistic that measures the dispersion of a dataset relative to its mean and is calculated as the square root of the variance

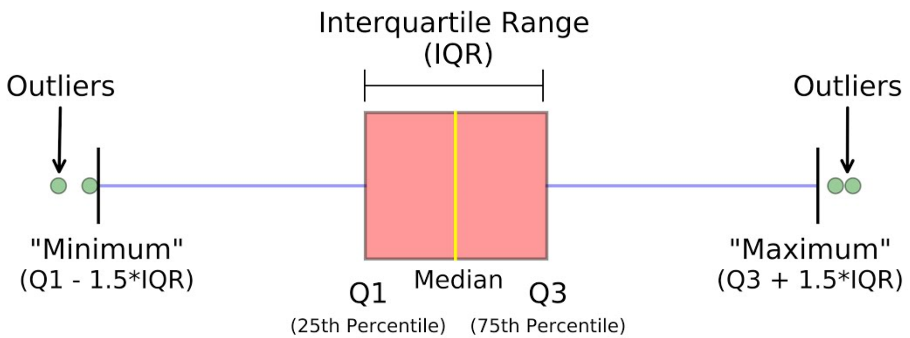
Data Distribution



Histogram

Skewness

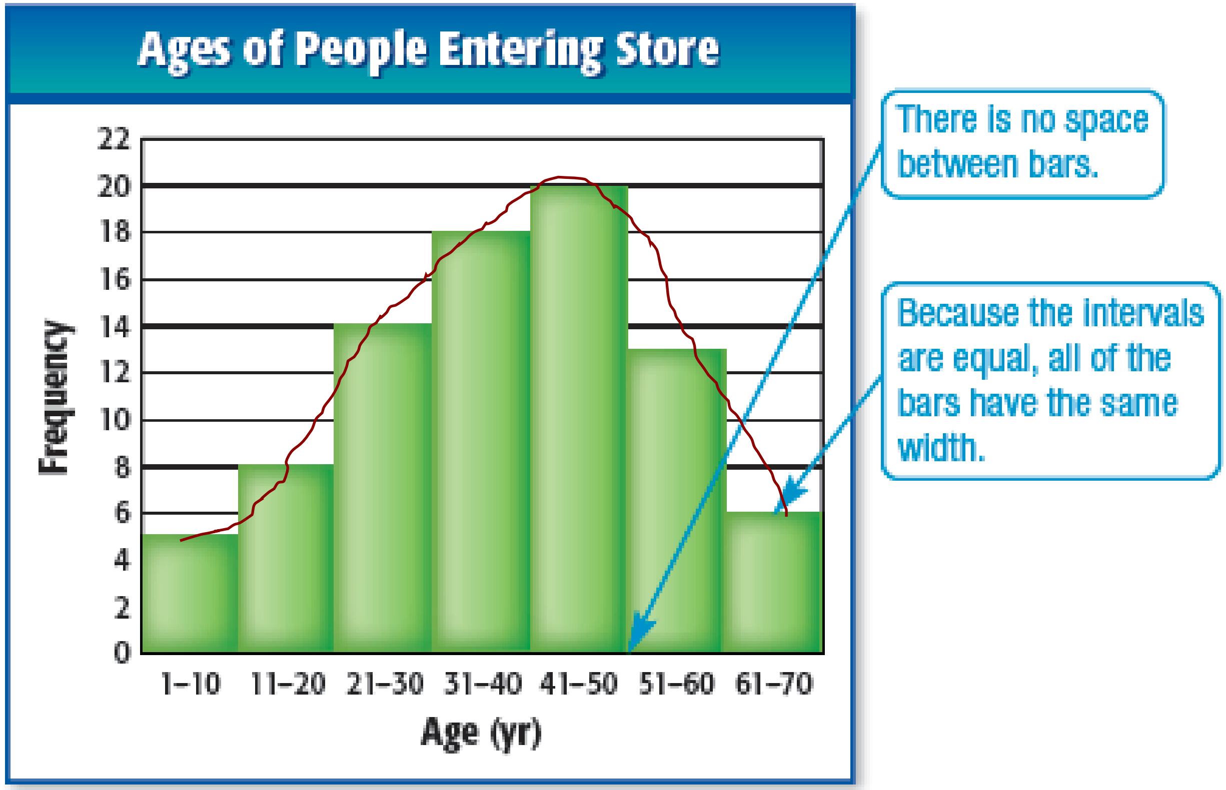




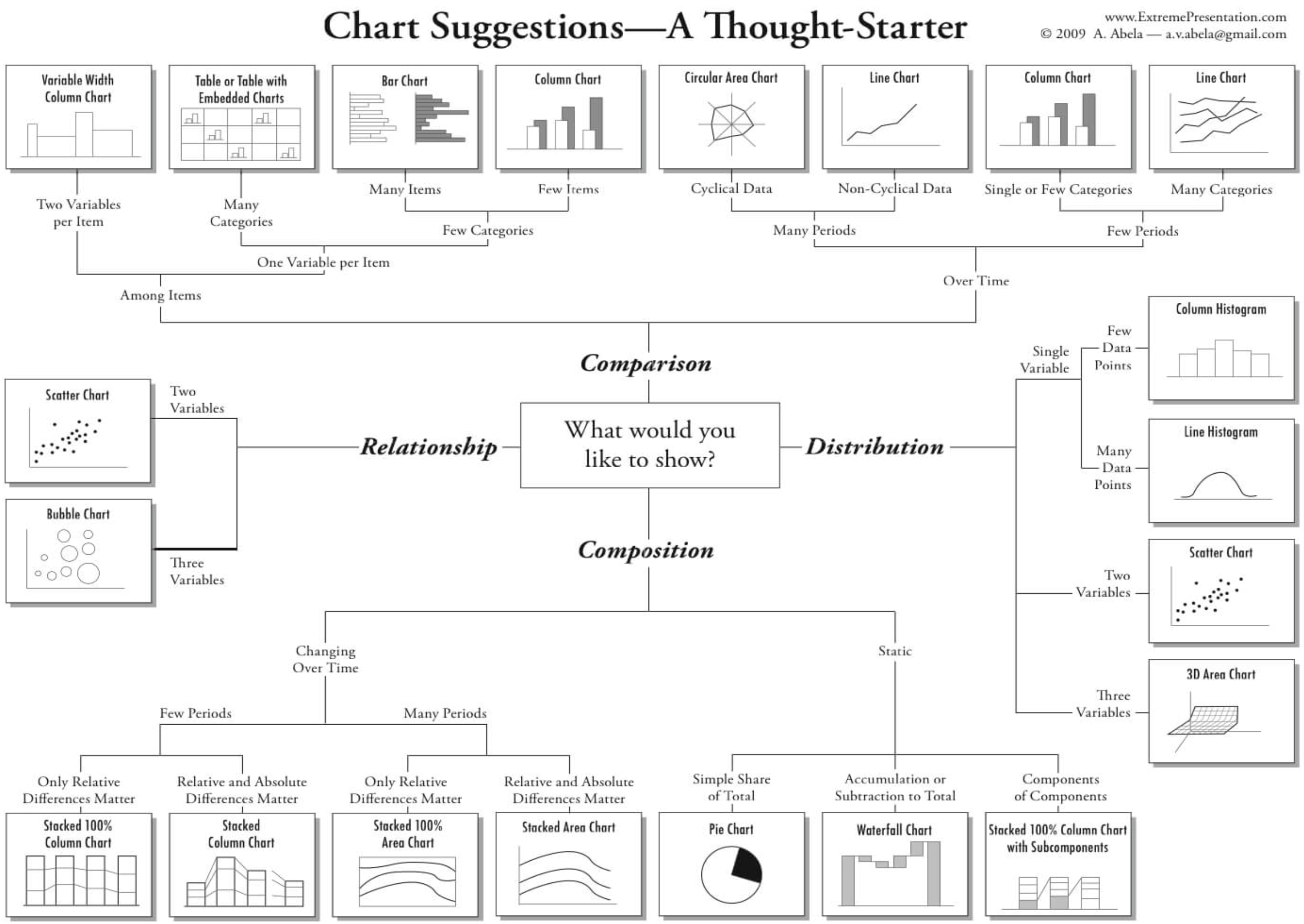
Box Plot

Presenting data: Table and chart

Age	Tally	Frequency
1–10	𝍸	5
11–20	𝍸 III	8
21–30	𝍸 𝍸 IIII	14
31–40	𝍸 𝍸 𝍸 III	18
41–50	𝍸 𝍸 𝍸 𝍸	20
51–60	𝍸 𝍸 III	13
61–70	𝍸 I	6



Presenting data: chart



Probability Distribution

Probability distribution terminology

- Probability is the measure of the likelihood that an event will occur in a random experiment. It is quantified as a number between 0 and 1
- A random experiment is a physical situation whose outcome cannot be predicted until it is observed
- A sample space, is a set of all possible outcomes of a random experiment

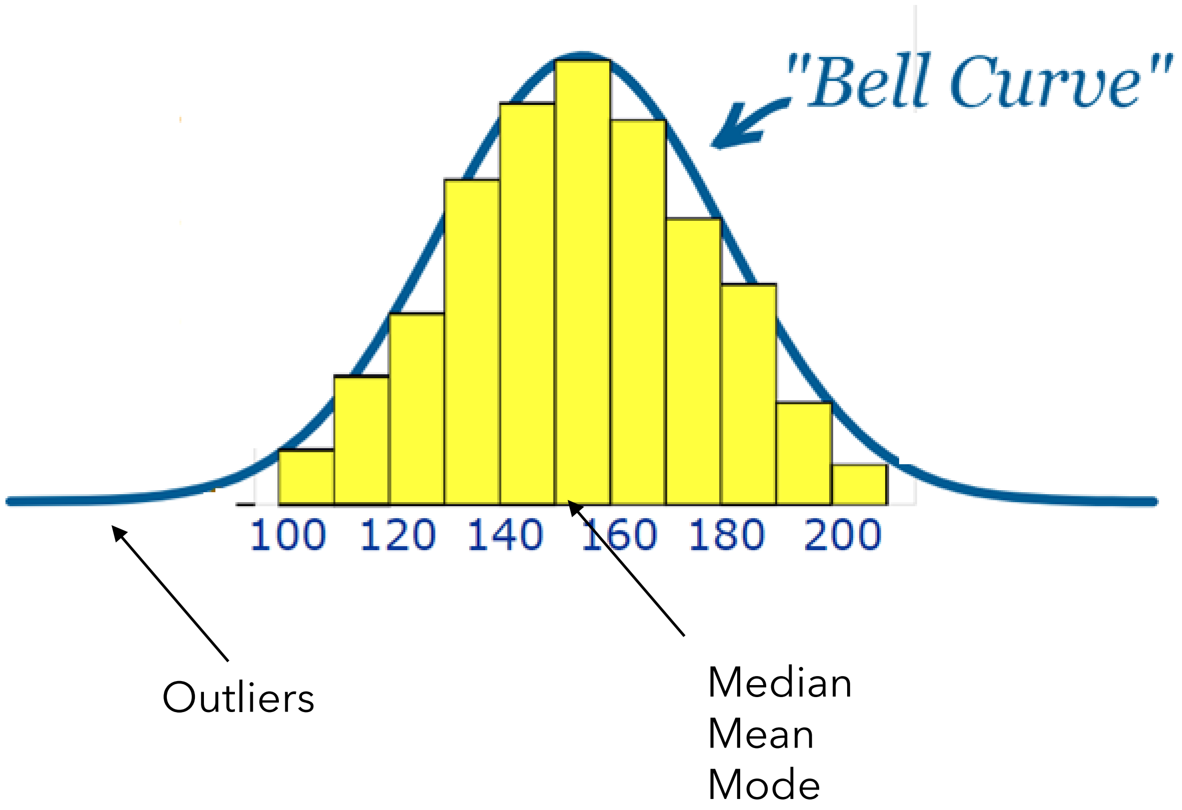
Probability distribution terminology

- A sample will form a **distribution**
- The distribution provides a **parameterized mathematical function** that can be used to calculate the probability for any individual observation from the sample space
- Data type:
  - Discrete: take only specified values
  - Continuous: take any value within a given range

Common data distribution

- Many data conform to well-known and well-understood mathematical functions
  - **Bernoulli**: two possible values, eg. 0/1, success/fail, etc
  - **Uniform**: finite number of values are equally likely to be observed, eg. rolling dice
  - **Binomial**: *n* times random experiment of Bernoulli
  - **Normal**: represents the behavior of most of the situations in the universe
  - Chi-square, Poisson, Gamma, Exponential, etc

Normal (**z**) distribution



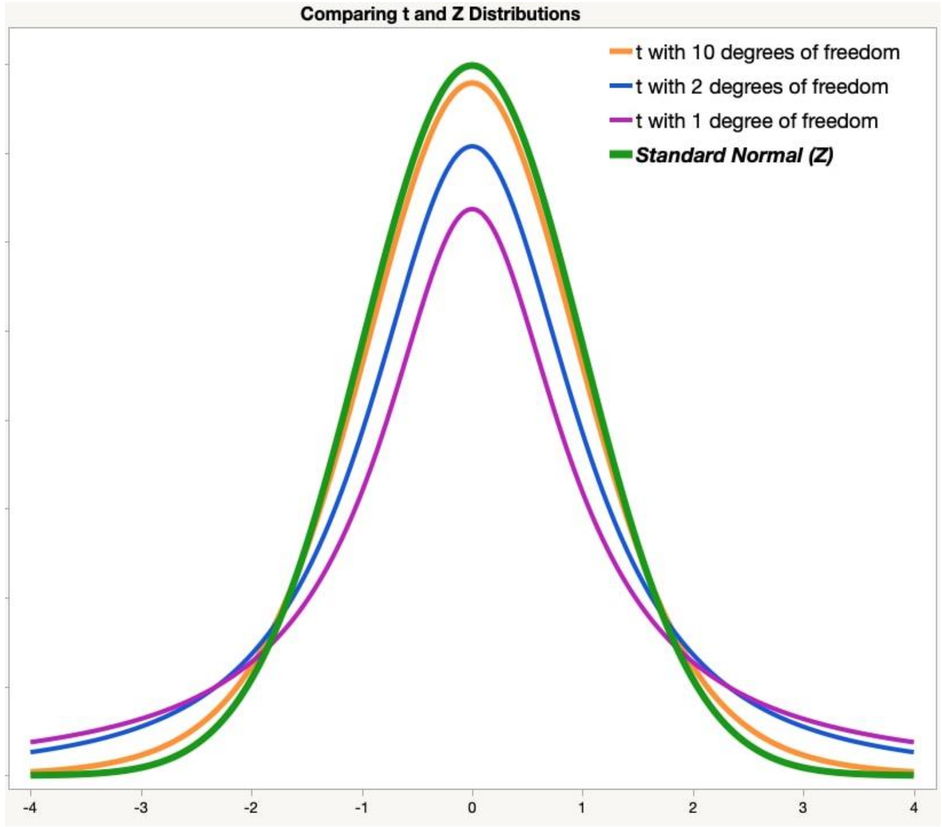
The Normal Distribution has:

- mean = median = mode
- symmetry about the center
- 50% of values less than the mean
- and 50% greater than the mean

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

Student's t distribution

- Like a standard normal distribution (or z-distribution), the *t*-distribution has a mean of zero.
- The normal distribution assumes that the population standard deviation is known. The *t*-distribution does not make this assumption.
- The *t*-distribution is defined by the *degrees of freedom*. These are related to the sample size.
- The *t*-distribution is most useful for small sample sizes, when the population standard deviation is not known, or both.
- As the sample size increases, the *t*-distribution becomes more similar to a normal distribution.



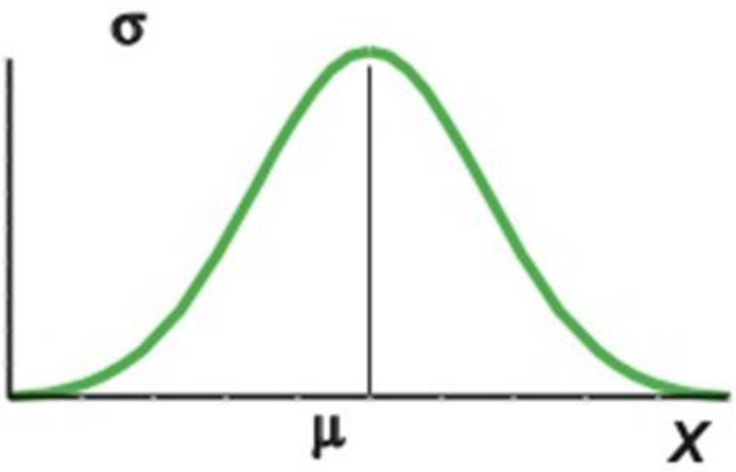
jmp.com

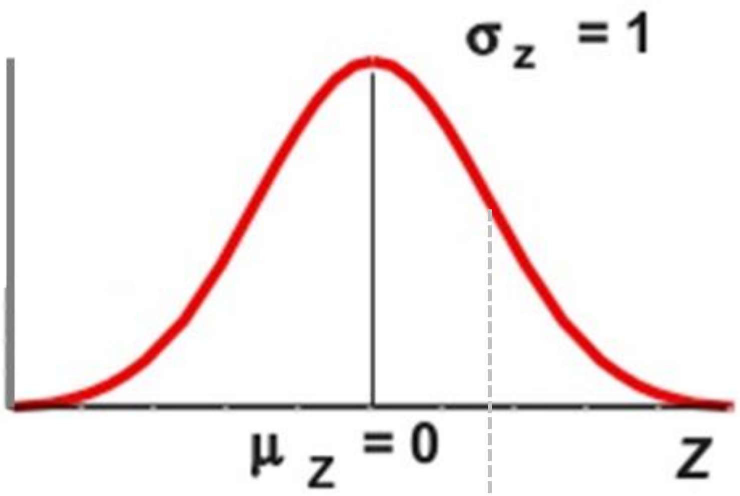
Standardized Normal distribution

Normal distribution

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

Standardized Normal distribution

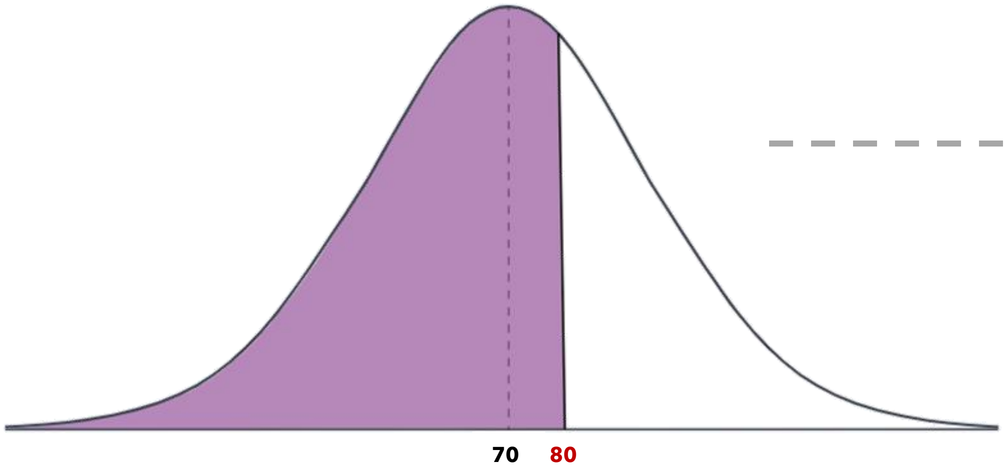


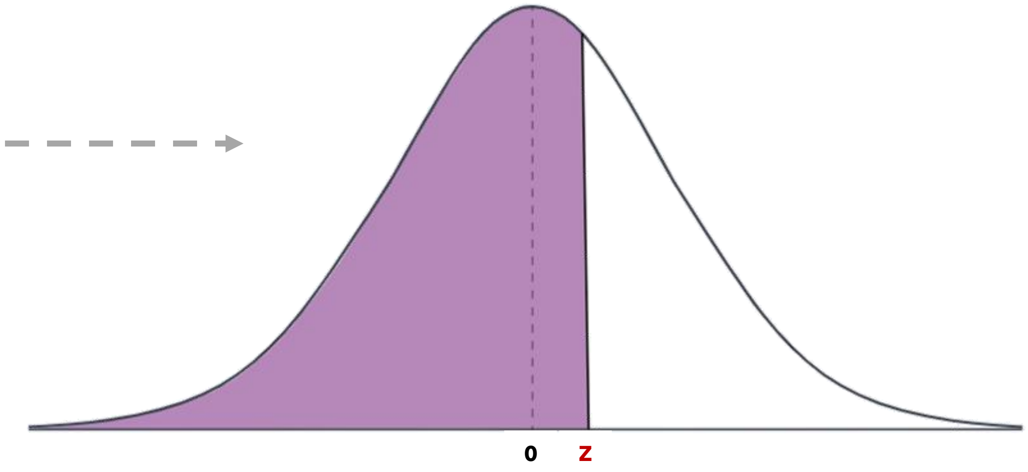


Area under curve = 1
(probability of all-possible events)

Calculate probability: **example**

Problem: The weights of adult-males are known to be normally distributed with a mean of 70 kgs and a standard deviation of 13 kgs. Find the percentage of adult-males with weights less than 80 kgs

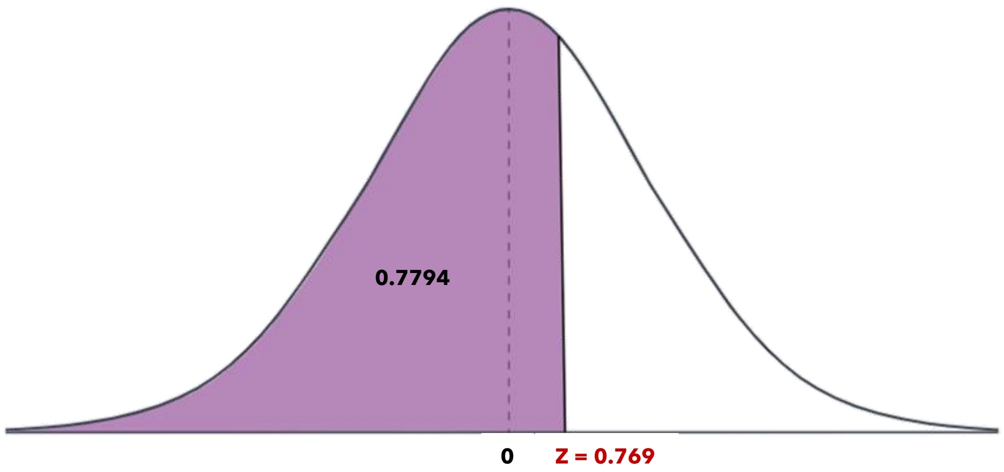




For $\boldsymbol{X} = 80 \rightarrow \boldsymbol{Z} = (80 - 70)/13 = 0.769$

So, $P(\boldsymbol{X} < 80) = P(\boldsymbol{Z} < 0.769)$

Calculate probability: **example**



P(***X*** < 80) = P(***Z*** < 0.769) = 0.7794

The percentage of adult-males with weights less than 80 kgs is around 78%

Standard Normal Probabilities

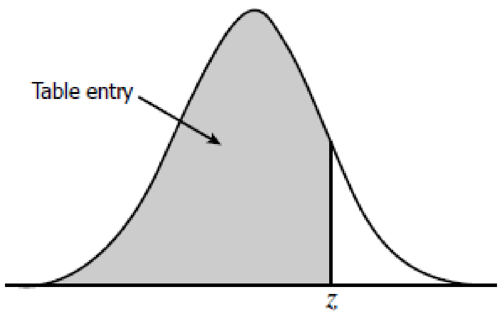


Table entry for z is the area under the standard normal curve to the left of z.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177

Inferential statistics

Inferential statistics

Eight **out of 10 owners said their cat** prefers it → 80% ?

How confident can we be about such statistic?

8 out of 10?
80 out of 100?
800 out of 1000?
80000 out of 100000?

Inferential statistics

- Process to draw conclusions about some unknown aspect of a population based on a random sample from that population

Statistic of Sample

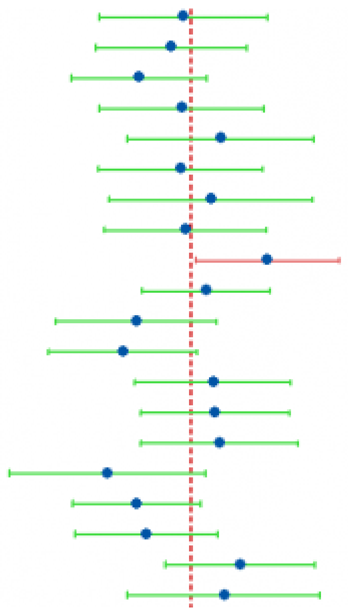
Or **parameter estimator**

- **Hypothesis testing**
- **Confidence Interval**
- **Significance testing**

Parameter of Population (Unknown)

Confidence interval

- A confidence interval calculates the probability that a population parameter will fall between two set values.
- Confidence intervals measure the degree of uncertainty or certainty in a sampling method.
- Most often, confidence intervals reflect confidence levels of 95% or 99%.
- If we say that we are 95% confident that the unknown population <u>mean</u> is contained in this interval, then we are really saying that we found the interval using a method that is successful in giving correct results 95% of the time



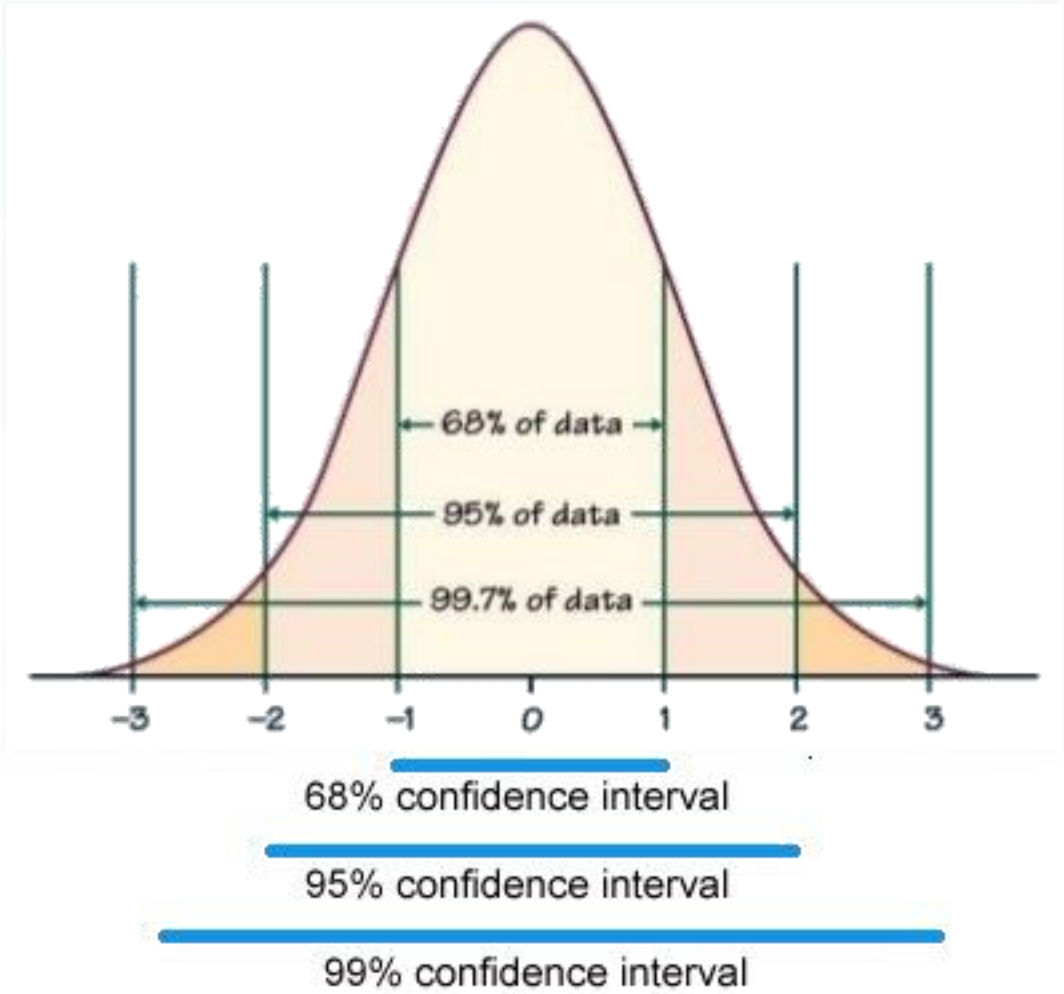
Interpreting confidence interval

"In a sample of 659 parents with toddlers, about 85%, stated they use a car seat for all travel with their toddler. From these results, a 95% confidence interval was provided, going from about 82.3% up to 87.7%."

- we are 95% certain that the population proportion who use a car seat for all travel with their toddler will fall between 82.3% and 87.7%.
- if we take a different sample from these 659 people, 95% of the time, the percentage of the population who use a car seat in all travel with their toddlers will be in between 82.3% and 87.7%.
- **95% confidence interval does not mean 95% probability**

Confidence level *vs* confidence interval

The greater the confidence level, the wider the confidence interval



Confidence interval formula

- Confidence interval for mean

$$\bar{X} - Z_\alpha\left({}^{\sigma}/_{\sqrt{n}}\right) \le \mu \le \bar{X} + Z_\alpha\left({}^{\sigma}/_{\sqrt{n}}\right)$$

(Population variance known)

$$\bar{x} - t_{v,\alpha}\left(\frac{s}{\sqrt{n}}\right) \le \mu \le \bar{x} + t_{v,\alpha}\left(\frac{s}{\sqrt{n}}\right)$$

(Population variance unknown)

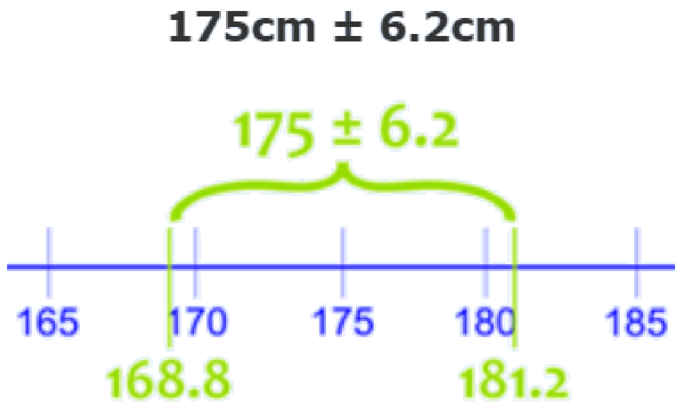
- Confidence interval for proportion

$$\text{p'} - Z_\alpha\sqrt{\frac{\text{p'q'}}{n}} \le p \le \text{p'} + Z_\alpha\sqrt{\frac{\text{p'q'}}{n}}$$

Confidence interval: **example**

Problem: We measure the heights of 40 randomly chosen men, and get a mean height of 175cm. We have already known that standard deviation is 20cm. Calculate the 95% confidence interval for mean height!

$$Z_\alpha\left({}^{\sigma}\!/_{\sqrt{n}}\right) = 1.96 * (20/\sqrt{40}) = 6.2$$



This says the true mean of ALL men height is likely to be between 168.8cm and 181.2cm.

Hypothesis testing

Hypothesis testing

- Hypothesis testing is a statistical method that is used in making statistical decisions using sample
- Hypothesis Testing is basically an assumption that we make about the population parameter
- Hypothesis is made before running the test

Null Hypothesis (H0)

Assumes that the observation is due to a chance factor

"Two website designs generate equal revenue"

VS

Alternative Hypothesis (H1)

Shows that observations are the result of a real effect
This is hypothesis that we propose

"New website design generate more revenue than the existing one"

Hypothesis testing …

Null Hypothesis (H0)

"Two website designs generate equal revenue"

VS

Alternative Hypothesis (H1)

"New website design generate more revenue than the existing one"

- Collect **sufficient** evidence to reject H0 → accept H1
- How sufficient is sufficient? → Confidence level (90%, 95%, 99%, ect.)
- When reject H0, we conclude that H1 is true at certain level of confidence
- We never accept H0. Instead, we say: No sufficient evidence to reject H0

Two types of error

	Reject H0	Fail to Reject H0
Reality: H0 is True	**Type I error** (α)	**Correct decision** Probability (1 - α)
Reality: H0 is False	**Correct decision** Power (1 - β)	**Type II error** (β)

Hypothesis testing: the court

Man
About 40s
180 cm tall
Black sedan

Souce: Rahul Patwari @ Youtube

Hypothesis testing: the court

Souce: Rahul Patwari @ Youtube

Courtroom hypothesis

Innocent

Our job is to disapprove this

until proven

If innocent, very unlikely to find this evidence

EVIDENCE

Guilty

Alternative hypothesis

So we can accept this

HOW UNLIKELY?

50%: chance he's innocent and we found that evidence

20%

5%

1%

Significant level (alpha)

Hypothesis testing: the court

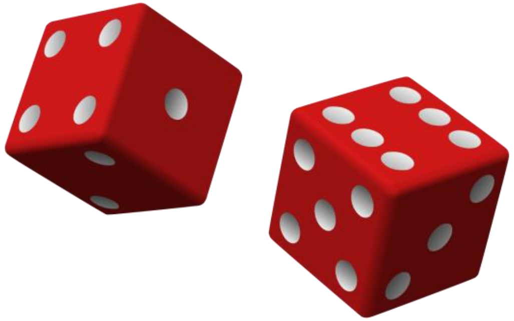
Judge doesn't reject his innocent
CORRECT DECISION

Judge doesn't reject his innocent
ERROR (Type II) : β

Judge reject his innocent
ERROR (Type I) : α

Judge reject his innocent
CORRECT DECISION

Souce: Rahul Patwari @ Youtube



- There are un-avoidable errors in statistics
- That's why statistics is also defined as **science of uncertainty**
- The role of statisticians is to **quantify** and **minimize** the error

Statistical test

Correlational: test an association or relationship between variables	
Pearson Correlation	Tests for the strength of the association between two continuous variables
Spearman Correlation	Tests for the strength of the association between two ordinal variables
Comparison of Means: test the difference between the means of variables	
T-Test	Tests for the difference between two group
ANOVA	Tests for the difference between group means (more than 2 groups)
Regression: test if change in one variable predicts change in another variable	
Linear regression	Tests how change in the predictor variable predicts the level of change in the outcome variable
Logistics regression	Tests how change in the predictor variable predicts the level of change in the **categorical** outcome variable

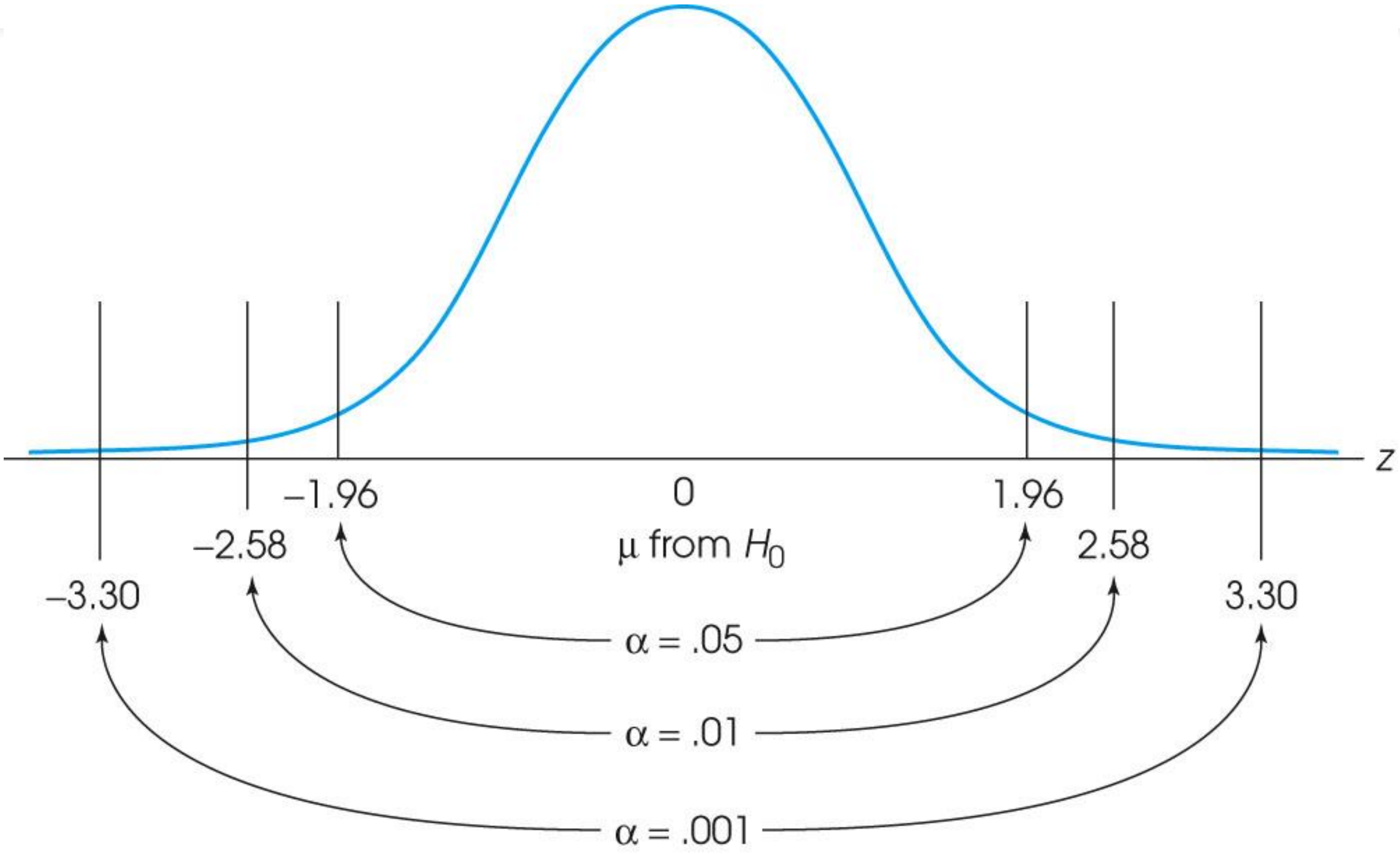
Hypothesis testing steps

1. State the null hypothesis, $\mathbf{H_0}$ and the alternative hypothesis, $\mathbf{H_1}$
2. Choose an alpha α, our significance level
3. Select a **statistical test**, and calculate the observed **test statistic**
4. Find the **critical value** of the test statistic (and/or ***p*-value**)
5. Compare the observed test statistic with the critical value, (or compare the **p-value** with α), and decide to accept or reject H_0

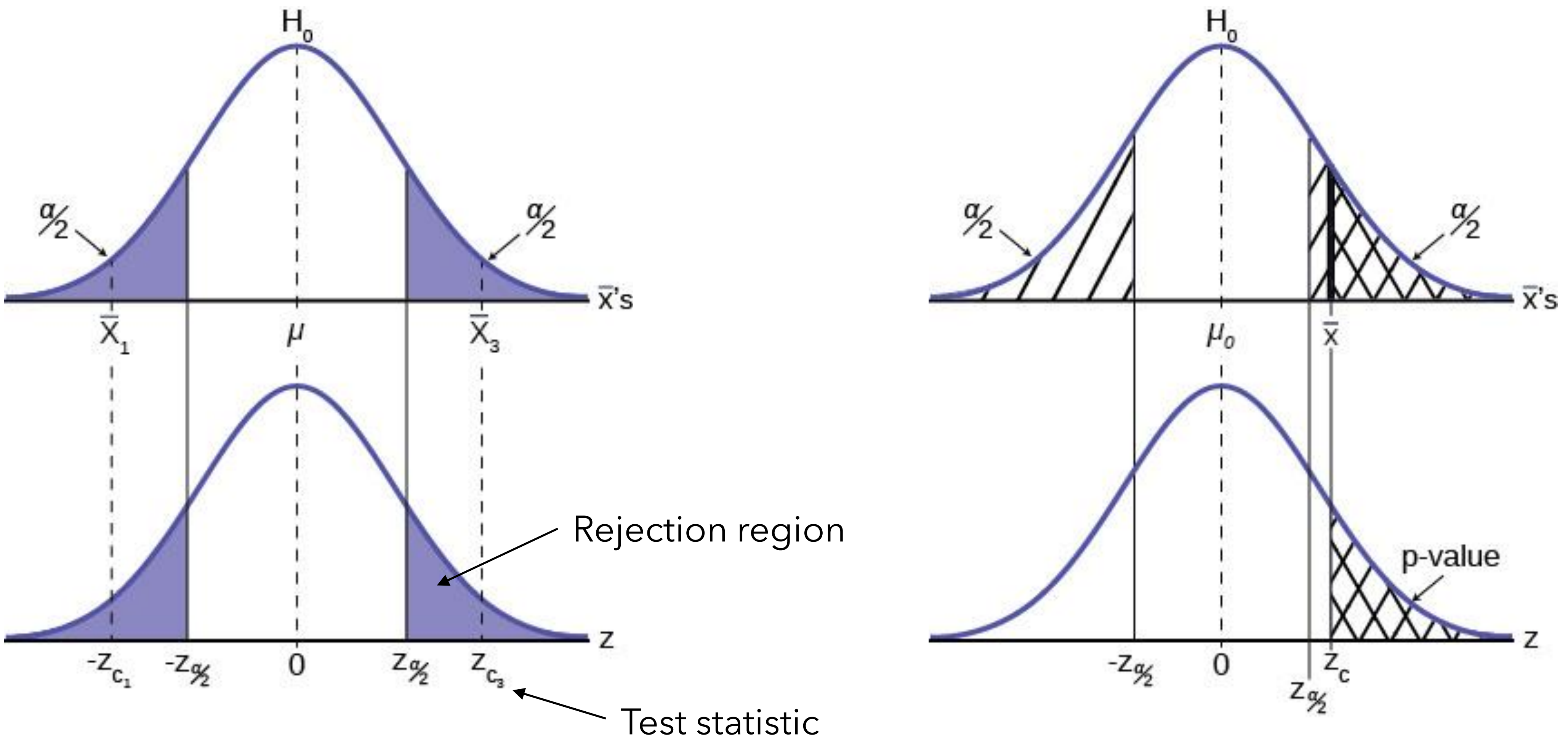
Two and one-tailed hypothesis testing

$H_0: \mu = 15.00$	$H_0: \mu \geq 15$	$H_0: \mu \leq 15$
$H_1: \mu \neq 15.00$	$H_1: \mu < 15$	$H_1: \mu > 15$
Two-tailed	One-tailed	

Alpha/significance level



Alpha, critical value, test statistic and p-value



Making conclusion

If test statistic is large enough to be in the critical region, we conclude that the difference is significant or that the treatment has a significant effect**. In this case we reject the null hypothesis**.

If the mean difference is relatively small, then the test statistic will have a low value. In this case, we conclude that the **evidence from the sample is not sufficient**, and the decision is **fail to reject the null hypothesis**

Hypothesis testing about mean

- **One sample**:
  - Z test if population variance is known $$z_h = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$
  - Z test if population variance is unknown $$t_h = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

Hypothesis testing about mean

- **Two samples**
  - Both variances are unknown, but assumed equal)

$$t_h = \frac{(\bar{x}_1 - \bar{x}_2) - \delta_0}{s_{(\bar{x}_1 - \bar{x}_2)}} \qquad s_{\bar{x}_1 - \bar{x}_2} = \sqrt{s_{gab}^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

$$s_{gab}^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \ \text{dan} \ \ v = n_1 + n_2 - 2$$

Hypothesis testing about mean

- **Two paired samples**

$$t = \frac{m}{s/\sqrt{n}}$$

Since ***m*** is mean of pairwise difference between two-sample, we can simply apply standard one-sample T-Test

Hypothesis testing about proportion

- **One sample proportion**

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{\left(p_0(1 - p_0)\right)}{n}}}$$

- **Comparison of two-proportions**

$$z = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}(1 - \hat{p})\left(1/n_1 + 1/n_2\right)}}$$

Levene's test

- Inferential statistic used to assess the **equality of variances** for a variable calculated for two or more groups
- Often used before a comparison of means

$$W = \frac{(N-k)}{(k-1)} \cdot \frac{\sum_{i=1}^{k} N_i (Z_{i\cdot} - Z_{\cdot\cdot})^2}{\sum_{i=1}^{k} \sum_{j=1}^{N_i} (Z_{ij} - Z_{i\cdot})^2},$$

where

- k is the number of different groups to which the sampled cases belong,
- N_i is the number of cases in the ith group,
- N is the total number of cases in all groups,
- Y_{ij} is the value of the measured variable for the jth case from the ith group,
- $Z_{ij} = \begin{cases} |Y_{ij} - \bar{Y}_{i\cdot}|, & \bar{Y}_{i\cdot} \text{ is a mean of the } i\text{-th group,} \\ |Y_{ij} - \tilde{Y}_{i\cdot}|, & \tilde{Y}_{i\cdot} \text{ is a median of the } i\text{-th group.} \end{cases}$

Sample size to conduct hypothesis testing

- One sample

$$n = \left(\frac{Z_{1-\alpha/2} + Z_{1-\beta}}{ES} \right)^2$$

$$ES = \frac{|\mu_1 = \mu_2|}{\sigma}$$

- Two independent sample

$$n_i = 2\left(\frac{Z_{1-\alpha/2} + Z_{1-\beta}}{ES} \right)^2$$

$$ES = \frac{|\mu_1 = \mu_2|}{\sigma}$$

- Two paired sample

$$n = \left(\frac{Z_{1-\alpha/2} + Z_{1-\beta}}{ES} \right)^2$$

$$ES = \frac{\mu_d}{\sigma_d}$$

- One sample proportion

$$n = \left(\frac{Z_{1-\alpha/2} + Z_{1-\beta}}{ES} \right)^2$$

$$ES = \frac{p_1 - p_0}{\sqrt{p_1(1-p_1)}}$$

- Two ind. Sample – proportion

$$n_i = 2\left(\frac{Z_{1-\alpha/2} + Z_{1-\beta}}{ES} \right)^2$$

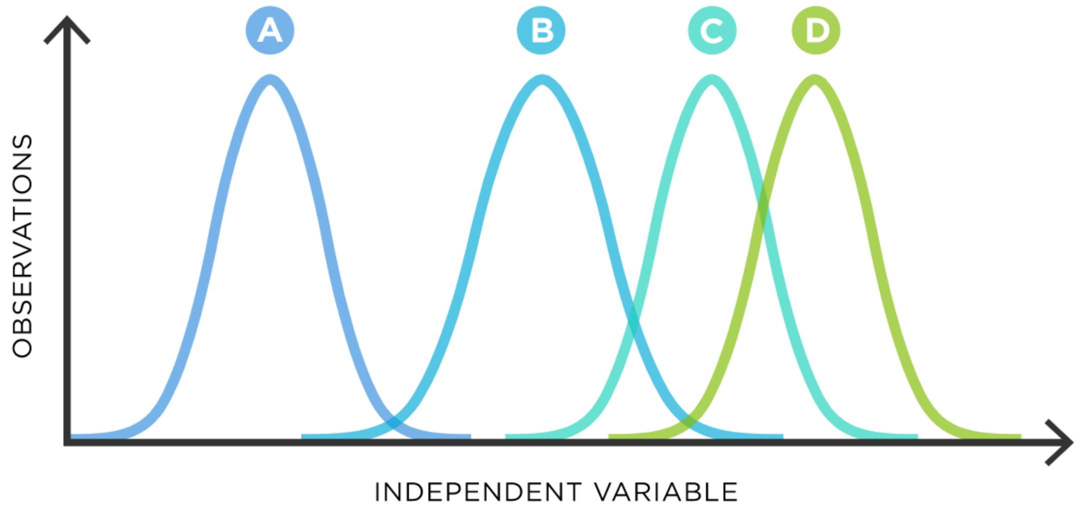
$$ES = \frac{|p_1 = p_2|}{\sqrt{p(1-p)}}$$

ANOVA – Analysis of Variance

- ANOVA is a statistical technique that is used to check if the means of two or more groups are significantly different from each other.
- When we have only two samples, t-test and ANOVA give the same results
- If we conduct multiple t-tests for comparing more than two samples, it will have a compounded effect on the error rate of the result
- ANOVA uses F-tests to statistically test the equality of means.

H_0: $\mu_0 = \mu_1 = \ldots = \mu_m$

H_1: at least one μ is unequal



Hypothesis testing: **example**

Problem: We measure the heights of 40 randomly chosen men, and get a mean height of 175cm. We have already known that standard deviation is 20cm. With alpha 5%, can we conclude that mean height of ALL men is greater than 170cm?

H_0: $\mu = 170$

H_1: $\mu > 170$

Critical value: $\mathbf{z} > 1.645$

Test statistic: $(175–170)/(20/\sqrt{40}) = 1.581$

Test statistic is less than critical value, so we failed to reject $H_{0,}$ and conclude that there are not sufficient evidences that population mean of men's height is more than 170cm

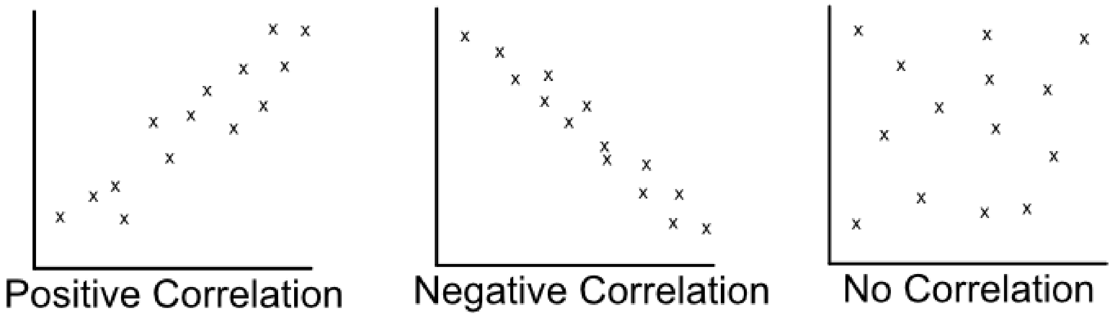
Correlation

Correlation

- Correlation is a statistical technique that can show whether and how strongly pairs of numerical variables are related
- Denoted by a **correlation coefficient** (or "r"). It ranges from -1.0 to +1.0.
- The closer r is to +1 or -1, the more closely the two variables are related

Pearson's correlation:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

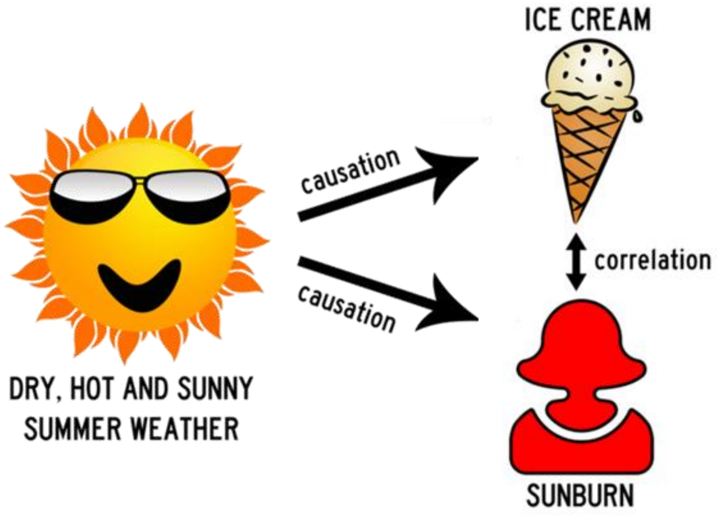


Association

- The **Chi-Square Test** of Independence determines whether there is an association between categorical variables
- This test utilizes a contingency table (also known as *cross-tabulation*, *crosstab*, or *two-way table*) to analyze the data

	Smartphones	Laptops	Total Sales (in $M)
Region 2	132	14	146
Region 3	92	16	108
Total Sales (in $M)	224	30	254

Correlation does not imply causation



- Seeing two variables moving together does not necessarily mean we know whether one variable causes the other to occur.
- It may be the result of random chance, where the variables appear to be related, but there is no true underlying relationship
- There may be a third, lurking variable that that makes the relationship appear stronger (or weaker) than it actually is
- ...but with well-designed empirical research, we can establish causation!
- **So, how do we explore causation? With the right kind of investigation!**

Regression

Statistical modeling

The majority of the problems in statistical inference can be considered to be problems related to statistical modeling

-- Konishi & Kitagawa (2008)

Supervised

- Linear Regression
- Logistics Regression
- Naïve Bayes classification
- Decision tree
- Discriminant analysis
- …

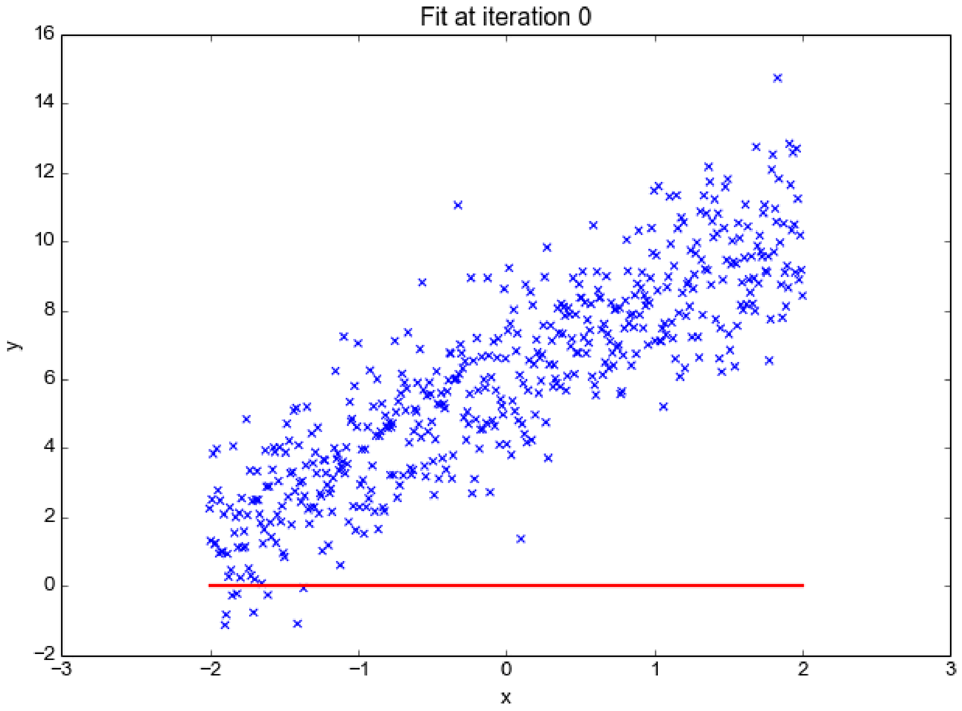
Unsupervised

- K-Mean clustering
- K-Median
- Hierarchical
- Hidden Markov model
- …

Time series

- Auto Regression
- Moving Average
- ARIMA
- **ARCH model**
- …

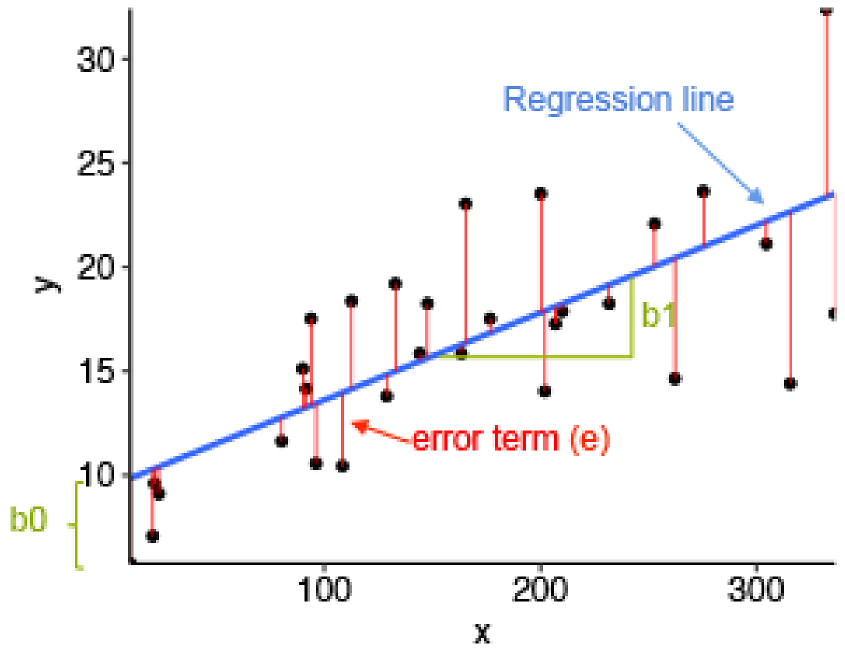
Regression



- Linear regression is a basic and commonly used type of predictive analysis.
- In regression, we examine:
  - does a set of predictor variables do a good job in predicting an outcome (dependent) variable?
  - Which variables in particular are significant predictors of the outcome variable, and in what way do they impact the outcome variable?

Regression

- Regression is one types of statistical modeling
- The **goal** of is to adjust the values of the model's parameters to find the line or curve that comes closest to your data



Simple:

$$\boldsymbol{y} = \mathrm{b}_0 + \mathrm{b}_1\, \boldsymbol{x} + \boldsymbol{error}$$

Multiple:

$$\boldsymbol{y} = \mathrm{b}_0 + \mathrm{b}_1\, \boldsymbol{x_1} + \mathrm{b}_2\, \boldsymbol{x_2} + \ldots + \boldsymbol{error}$$

- We then use statistical inference to test whether regression coefficients (b0, b1, b2, …) are significant

Regression

- Regression only accepts numerical variables for both **predictors** and **response**
- Categorical variable can be use as predictors after be transformed into **dummy variables**
- Significance of whole model → **F** test (ANOVA), Significance of each predictors → **t**-test
- **R-Squared (or Adjusted R-Squared)** is commonly used to quantify the goodness of the model

Regression

Assumptions for linear regression and ANOVA

- Each group sample is drawn from a normally distributed population
- All populations have a common variance (homoscedasticity)
- All samples are drawn independently of each other
- Within each sample, the observations are sampled randomly and independently of each other (no auto-correlation)
- No correlation between predictors (no multicollinearity)

Statistics *vs.* Machine Learning

- The two are highly related and share some underlying machinery, but they are different:

	Statistics	**Machine Learning**
Focus	**Building models**	**Creating system** that learn from data
Purpose	**Inference**, **relationship** between variables	**Prediction** accuracy, optimization
Prior assumption about data	Some knowledge (**assumption**) about population usually required	**Without assumption**
Dimensionality of data	Usually applied to low-dimensional data	Usually applied to high-dimensional data
Knowledge overlap	No machine learning knowledge required	Some stats knowledge usually needed. Stats is basis for algorithm

Q/A