

# BioStatistics

Gaurav Ahuja Ph.D  
Associate Professor  
Department of Computational Biology



INDRAPRASTHA INSTITUTE of  
INFORMATION TECHNOLOGY **DELHI**

# Course Structure

Weekly Lecture Plan	
Week Number	Lecture Topic
Week 1	Sampling methods and Sample size determination, Statistical Epidemiology
Week 2	Longitudinal Data Analysis, Categorical Data Analysis
Week 3	Time to event data analysis, Clinical Trials
Week 4	Quantitative Genetics: Relationship between genotype and phenotype, Types of Quantitative Characteristics,
Week 5	Determining Gene Number for a Polygenic Characteristic. Statistical Methods Required for Analyzing Quantitative
Week 6	Phenotypic and Genotypic Variance, Calculating Heratibility. Locating Genes That Affect Quantitative Characteristics.
Week 7	Genetically Variable Traits Change in Response to Selection
Week 8	Predicting the Response to Selection.
Week 9	Quantitative expect of Horizontal Gene Transfer,
Week 10	Introduction to Population genetics
Week 11	Hardy-Weinberg law, its implications and its extensions.
Week 12	Estimation of Genomic or Allelic Frequencies with the Hardy-Weinberg Law
Week 13	Quantitative aspect of Mendalian and non-mendalian Genetics

# Scoring Strategy

Assignments: 15% (one)  
Quiz: 25% (average of best 2 out of 3/4)  
In-class assessment: 10%  
Mini Project: 10%  
Mid-Sem= 20%  
Final-Sem (viva)= 20%

## Evaluation

### Resource Material

Type	Title
Textbook	B. Pierce, Genetics: A conceptual approach
Textbook	Principles and Practice of Biostatistics - by B Antonisamy

bar inference  
analysis sample inferences  
population precision research  
make dichotomous  
ordinal outcomes biostatistical variables estimate  
classified generalizeable endpoints continuous  
subset histogram prevalence question  
range categorical random  
chart analyses

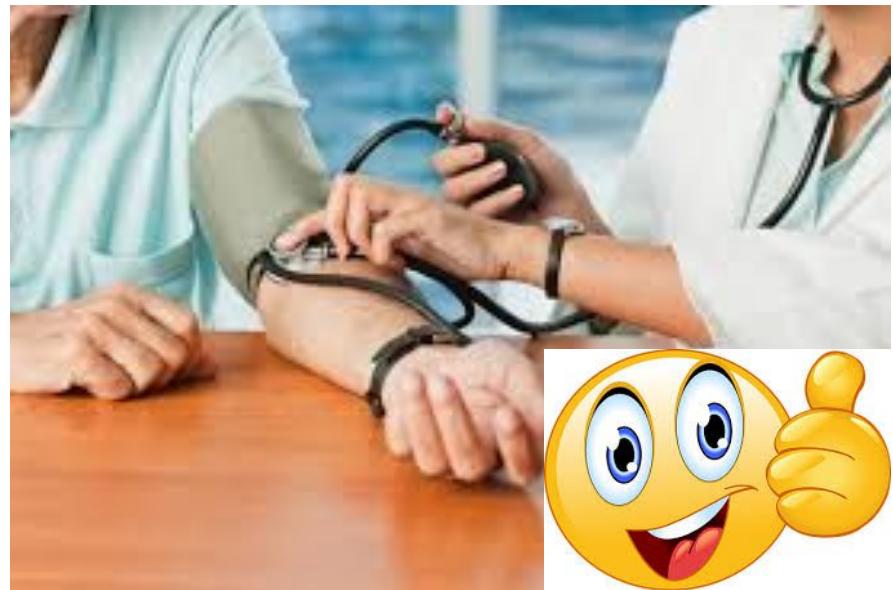
# Why study Statistics?

- Statistical methods are required to ensure that data are interpreted correctly and that apparent relationships are meaningful (or “significant”) and not simply chance occurrences.
- A “statistic” is a numerical value that describes some property of a data set.
- Two important concepts in statistics are the “population” and the “sample”.

# Problem with a Close Friend



May get CVD or Cardiac Stroke in near future



No high blood pressure on Old School Machine

# Classical Example



VS



# Problems to Solve

- How many machines should we test?
- How many participants should we test at each machine?
- In what order should we take the measurements? That is, should the human observer or the machine take the first measurement? Under ideal circumstances we would have taken both the human and machine readings simultaneously, but this was logistically impossible.
- What data should we collect on the questionnaire that might influence the comparison between methods?
- How should we record the data to facilitate computerization later?
- How should we check the accuracy of the computerized data?

# Simple Solution

1) How many machines should we test?

**Let's check all of them !**

(2) How many participants should we test at each machine?

**Using the methods of sample-size estimation.**

# Simple Solution

(3) In what order should we take the measurements? That is, should the human observer or the machine take the first measurement? Under ideal circumstances we would have taken both the human and machine readings simultaneously, but this was logically impossible.

First measurements are always a problem! High blood pressure !

A conventional technique we used here was to randomize the order in which the measurements were taken, so that for any person it was equally likely that the machine or the human observer would take the first measurement

# Simple Solution

(4) What data should we collect on the questionnaire that might influence the comparison between methods?

**We must collect the meta-data of the patient (age, weight, height etc), and check if any of these parameters create bias in the study.**

(5) How should we record the data to facilitate computerization later?

(6) How should we check the accuracy of the computerized data?

**Automate the process!**

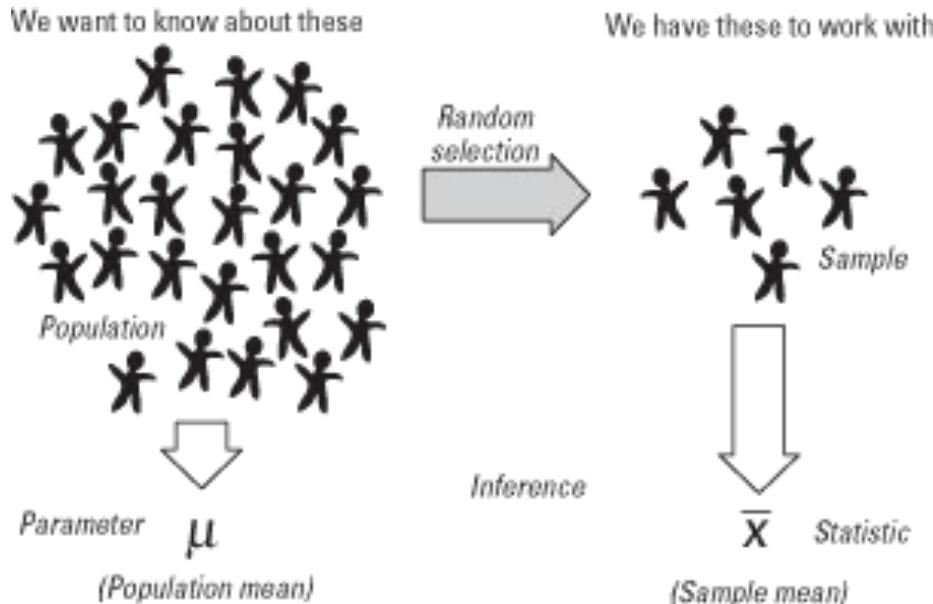
# Real World Example

Mean blood pressures and differences between machine  
and human readings at four locations

Location	Number of people	Systolic blood pressure (mm Hg)					
		Machine		Human		Difference	
		Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
A	98	142.5	21.0	142.0	18.1	0.5	11.2
B	84	134.1	22.5	133.6	23.2	0.5	12.1
C	98	147.9	20.3	133.9	18.3	14.0	11.7
D	62	135.4	16.7	128.5	19.0	6.9	13.6

Difference of 14 mm Hg in mean systolic blood pressure between the two methods for the 98 people we interviewed at location C, this difference might not hold up if we interviewed 98 other people at this location at a different time, and we wanted to have some idea as to the error in the estimate of 14 mm Hg.

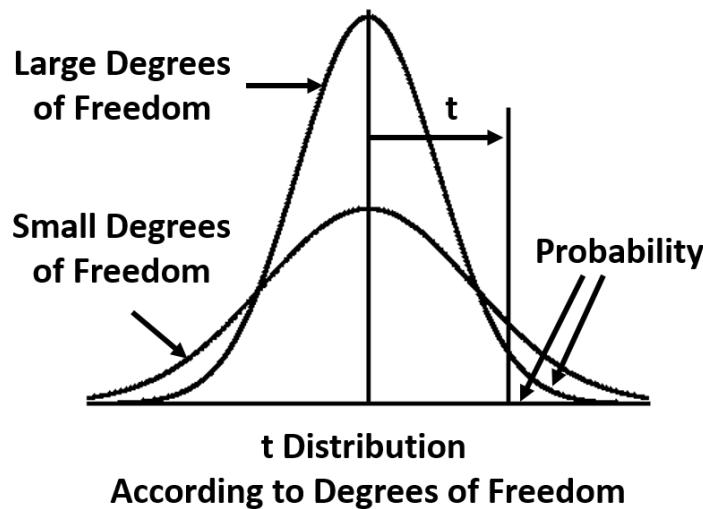
# What's going on ?



We needed to develop a probability model !

# What Can We Do now ?

Probability model will tell us how likely it is that we would obtain a 14-mm Hg difference between the two methods in a sample of 98 people if there were no real difference between the two methods over the entire population of users of the machine e.g. using a probability model based on the t distribution.

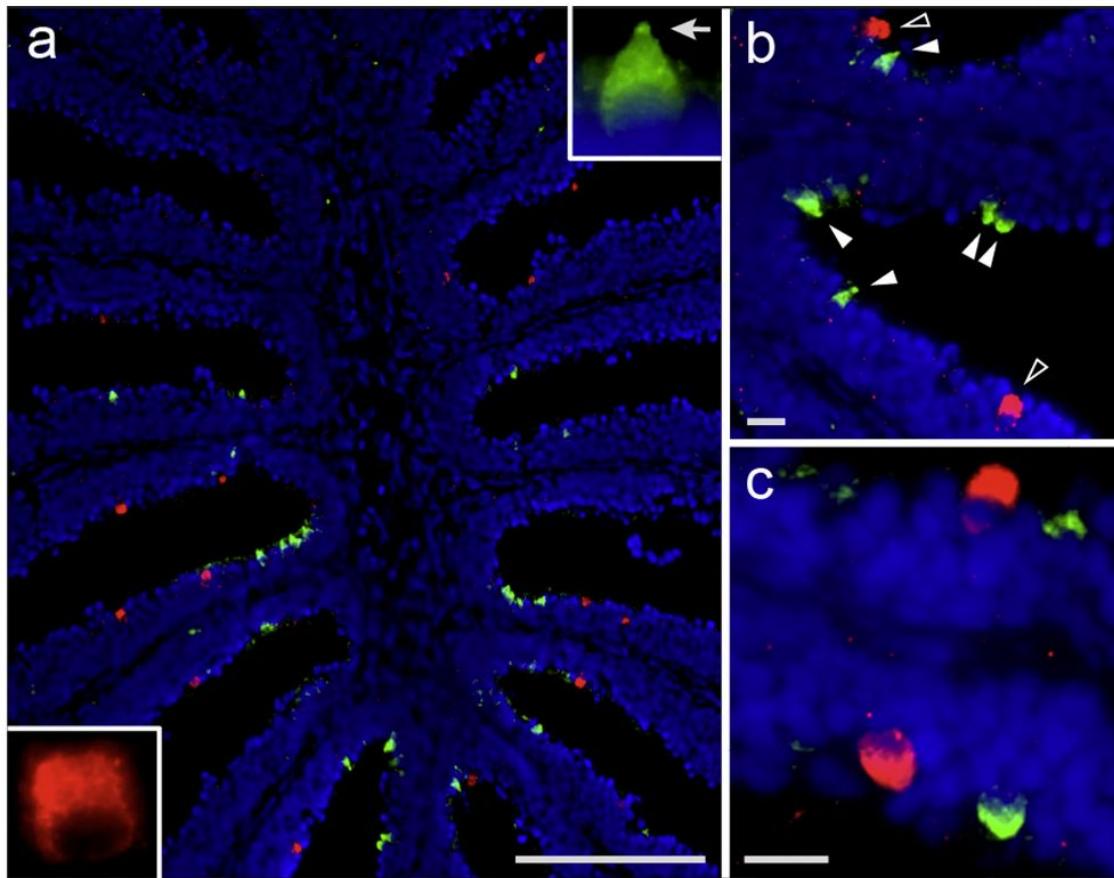


# Importance of Descriptive Statistics

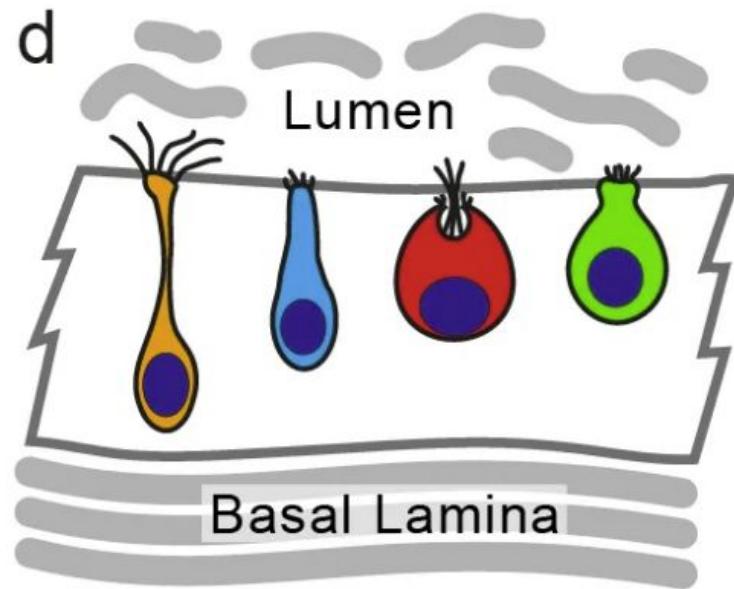
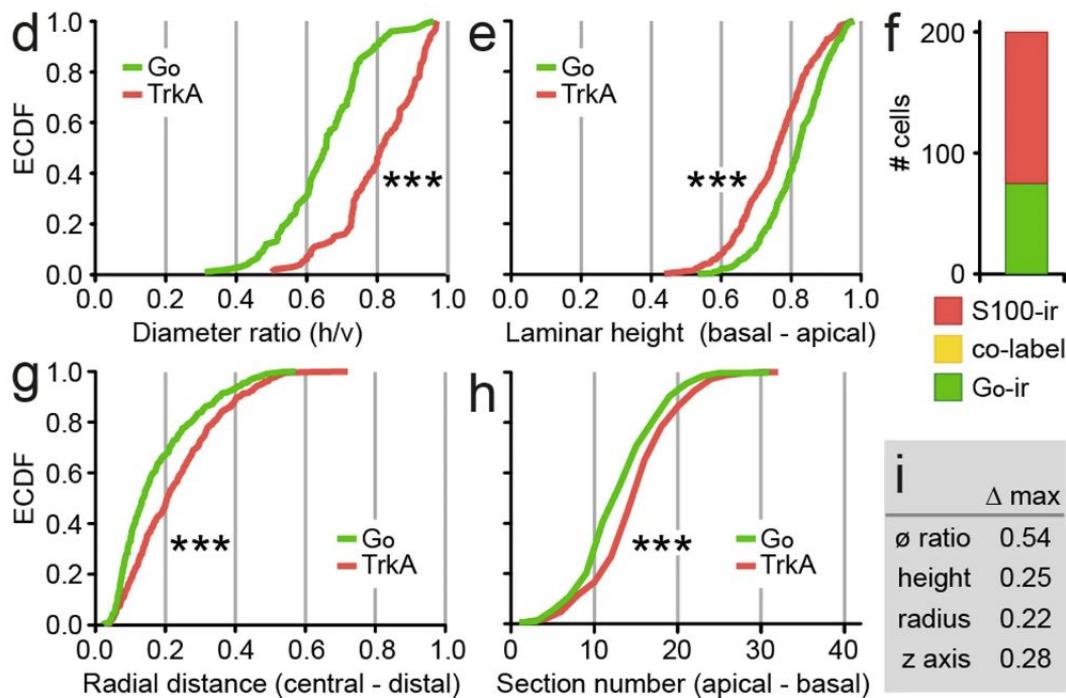
- The first step in looking at data is to describe the data at hand in some concise way.
- In smaller studies this step can be accomplished by listing each data point.
- In general, however, this procedure is tedious or impossible and, even if it were possible, would not give an overall picture of what the data look like.

Some real world  
Examples

# Identification of a Novel cell population



# Identification of a Novel cell population



# Wrong Statistics May Invite Troubles



DOI, PMID, arXiv ID, keyword,

Home / Publications

## Parental olfactory experience influences behavior and neural structure in subsequent generations

Nature Neuroscience (2014) - 6 Comments

pubmed: 24292232 doi: 10.1038/nn.3594 issn: 1546-1726 issn: 1097-6256

Brian G Dias, Kerry J Ressler

1] Department of Psychiatry and Behavioral Sciences, Emory University School of Medicine, Atlanta, G...

Gonzalo Otazu commented 6 years ago



PubMed COMMONS

The statistical tests in the paper, both for the behavioral measurements as well as for the size of the M71 glomeruli , use as n, number of samples, the number of F1 and F2 individuals. This would be fine if the individuals were actually independent samples. However, they arise from a presumably small number of FO males. The numbers of FO males are not given in the paper. This is a major concern given that there is a lot of variability in the levels of expression of olfactory receptors in these mice that might be inheritable. As an example, for Figure 1a, the authors compared 16 F1-Ace-C57 mice with 13 F1-Home-C57 mice and find a p value of 0.043, with 27

Stanley Lazic commented 6 years ago

PubMed COMMONS

An excellent point and discussed in Lazic SE, 2013 and references therein.

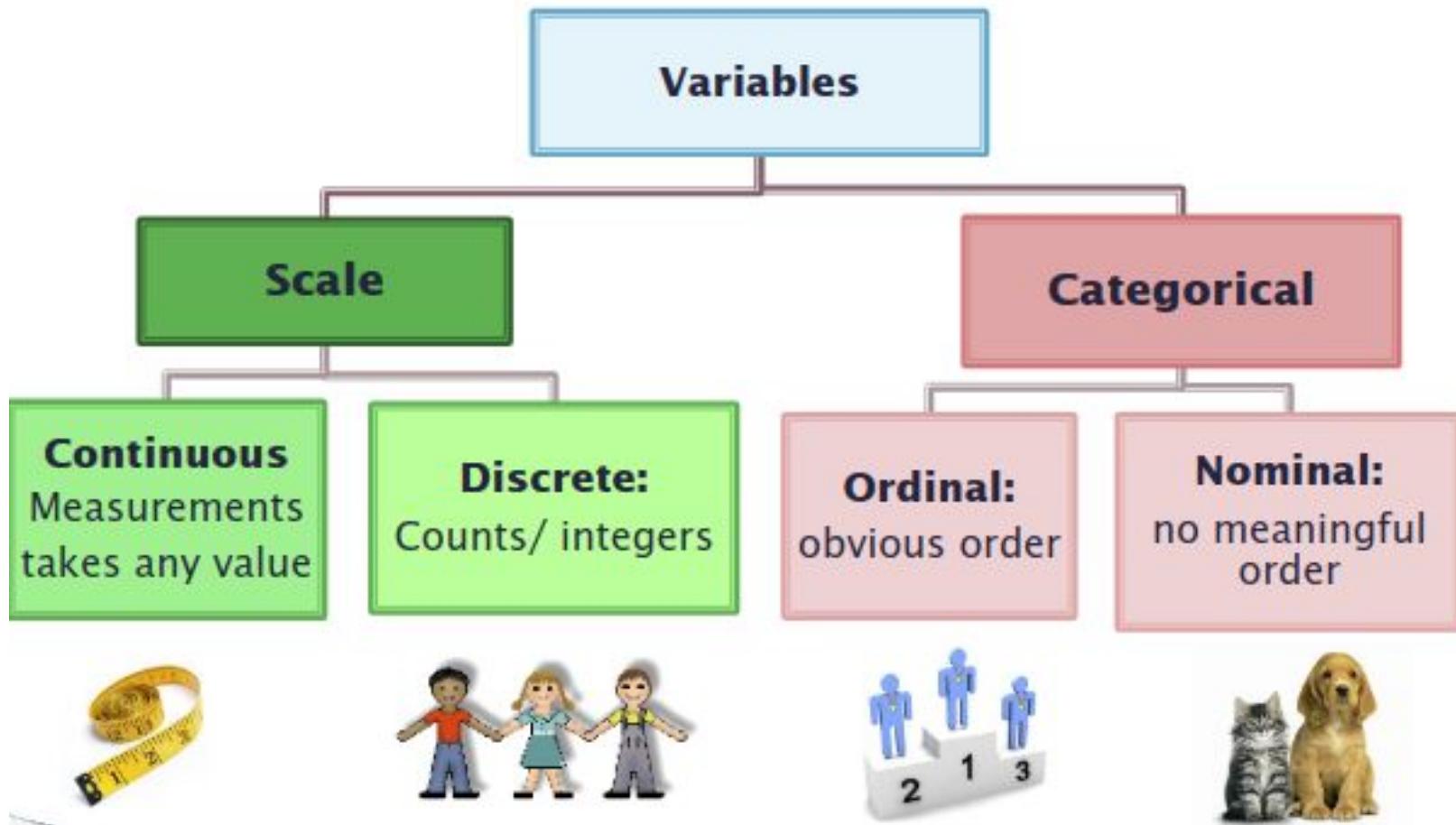
Also, it is not clear why the dorsal and medial glomerulus have different n for what appears to be same animals. These discrepancies occur for all histological data (e.g. Fig4 panel (i) has n=23 and n=16 for dorsal while panel (j) has n=16 and n=19 for ventral).

# Elements of Structured Data

- Data comes from many sources: sensor measurements, events, text, images, and videos.
- The Internet of Things (IoT) is spewing out streams of information. Much of this data is unstructured:
  - images are a collection of pixels with each pixel containing RGB (red, green, blue) color information.
  - Texts are sequences of words and non word characters, often organized by sections, subsections, and so on.

# KEY TERMS FOR DATA TYPES

- Continuous Data that can take on any value in an interval.  
Synonyms: interval, float, numeric
- Discrete Data that can take on only integer values, such as counts.  
Synonyms: integer, count
- Categorical Data that can take on only a specific set of values representing a set of possible categories.  
Synonyms: enums, enumerated, factors, nominal, polychotomous
- Binary: A special case of categorical data with just two categories of values (0/1, true/false).  
Synonyms: dichotomous, logical, indicator, boolean
- Ordinal: Categorical data that has an explicit ordering.  
Synonyms: ordered factor



# Fact!

## METRICS AND ESTIMATES

Statisticians often use the term *estimates* for values calculated from the data at hand, to draw a distinction between what we see from the data, and the theoretical true or exact state of affairs. Data scientists and business analysts are more likely to refer to such values as a *metric*. The difference reflects the approach of statistics versus data science: accounting for uncertainty lies at the heart of the discipline of statistics, whereas concrete business or organizational objectives are the focus of data science. Hence, statisticians estimate, and data scientists measure.

# Descriptive Statistics

# Mean

The most basic estimate of location is the mean, or average value. The mean is the sum of all the values divided by the number of values.

## Representation I

$$\text{Mean} = \bar{x} = \frac{\sum_i^n x_i}{n}$$

## Representation II

If  $a$  and  $b$  are integers, where  $a < b$ , then

$$\sum_{i=a}^b x_i$$

means  $x_a + x_{a+1} + \dots + x_b$ .

The arithmetic mean is, in general, a very natural measure of location. **One of its main limitations, however, is that it is very sensitive to extreme values.**

# Mean

## Sample Data

$i$	$x_i$	$i$	$x_i$	$i$	$x_i$	$i$	$x_i$
1	3265	6	3323	11	2581	16	2759
2	3260	7	3649	12	2841	17	3248
3	3245	8	3200	13	3609	18	3314
4	3484	9	3031	14	2838	19	3101
5	4146	10	2069	15	3541	20	2834

$$\bar{x} = (3265 + 3260 + \dots + 2834)/20 = 3166.9 \text{ g}$$

## NOTE

$N$  (or  $n$ ) refers to the total number of records or observations. In statistics it is capitalized if it is referring to a population, and lowercase if it refers to a sample from a population. In data science, that distinction is not vital so you may see it both ways.

# The Arithmetic Mean

The arithmetic mean is, in general, a very natural measure of location. **One of its main limitations, however, is that it is over sensitive to extreme values.**

Sample of birthweights (g) of live-born infants born at a private hospital in San Diego, California, during a 1-week period

$i$	$x_i$	$i$	$x_i$	$i$	$x_i$	$i$	$x_i$
1	3265	6	3323	11	2581	16	2759
2	3260	7	3649	12	2841	17	3248
3	3245	8	3200	13	3609	18	3314
4	3484	9	3031	14	2838	19	3101
5	4146	10	2069	15	3541	20	2834

Take home: **The arithmetic mean is a poor measure of central location** because it does not reflect the center of the sample. Nevertheless, the arithmetic mean is by far the most widely used measure of central location.

# Trimmed-Mean

- A variation of the mean is a trimmed mean.
- Calculated by dropping a fixed number of sorted values at each end and then taking an average of the remaining values.
- Representing the sorted values by  $x(1), x(2), x(3) \dots x(n)$ , where  $x(1)$  is the smallest value and the  $x(n)$  is the largest, the formula to compute the trimmed mean with  $P$  smallest and largest values omitted is:

$$\text{Trimmed mean} = \bar{x} = \frac{\sum_{i=p+1}^{n-p} x_{(i)}}{n - 2p}$$

**A trimmed mean eliminates the influence of extreme values.**

# Weighted Mean

- Calculate by multiplying each data value  $x_i$  by a weight  $w_i$  and dividing their sum by the sum of the weights.

$$\text{Weighted mean} = \bar{x}_w = \frac{\sum_{i=1}^n w_i x_i}{\sum_i^n w_i}$$

# Weighted Mean

**There are two main motivations for using a weighted mean:**

- Some values are intrinsically more variable than others, and highly variable observations are given a lower weight. For example, if we are taking the average from multiple sensors and one of the sensors is less accurate, then we might downweight the data from that sensor.
- The data collected does not equally represent the different groups that we are interested in measuring. For example, because of the way an online experiment was conducted, we may not have a set of data that accurately reflects all groups in the user base. To correct that, we can give a higher weight to the values from the groups that were underrepresented.



# Hands-on exercises

Compute the mean & trimmed mean

A	B	C	D
State	Population	Murder.Rate	Abbreviation
Alabama	4779736	5.7	AL
Alaska	710231	5.6	AK
Arizona	6392017	4.7	AZ
Arkansas	2915918	5.6	AR
California	37253956	4.4	CA
Colorado	5029196	2.8	CO
Connecticut	3574097	2.4	CT
Delaware	897934	5.8	DE
Florida	18801310	5.8	FL
Georgia	9687653	5.7	GA
Hawaii	1360301	1.8	HI
Idaho	1567582	2	ID
Illinois	12830632	5.3	IL
Indiana	6483802	5	IN
Iowa	3046355	1.9	IA
Kansas	2853118	3.1	KS
Kentucky	4339367	3.6	KY
Louisiana	4532272	10.2	LA

$$\text{Mean} = \bar{x} = \frac{\sum_i^n x_i}{n}$$

$$\text{Trimmed mean} = \bar{x} = \frac{\sum_{i=p+1}^{n-p} x_{(i)}}{n - 2p}$$

# Properties of Arithmetic Mean

Consider a sample  $x_1, \dots, x_n$ , which will be referred to as the original sample. To create a translated sample  $x_1 + c, \dots, x_n + c$ , add a constant  $c$  to each data point.

Let  $y_i = x_i + c, i = 1, \dots, n$ .

$$\begin{aligned} \text{If } \quad & y_i = x_i + c, \quad i = 1, \dots, n \\ \text{then } \quad & \bar{y} = \bar{x} + c \end{aligned}$$

# The Median

An alternative measure of location, perhaps second in popularity to the arithmetic mean, is the median or, more precisely, the sample median.

The **sample median** is

- (1) The  $\left(\frac{n+1}{2}\right)$ th largest observation if  $n$  is odd
- (2) The average of the  $\left(\frac{n}{2}\right)$ th and  $\left(\frac{n}{2}+1\right)$ th largest observations if  $n$  is even

- A. Why we are dividing by 2 in the definition ?
- B. Why median is better measure of central location than mean ?

The rationale for these definitions is to ensure an equal number of sample points on both sides of the sample median.

# The Median

Calculate the **Median** now !

Sample of birthweights (g) of live-born infants born at a private hospital in San Diego, California, during a 1-week period

$i$	$x_i$	$i$	$x_i$	$i$	$x_i$	$i$	$x_i$
1	3265	6	3323	11	2581	16	2759
2	3260	7	3649	12	2841	17	3248
3	3245	8	3200	13	3609	18	3314
4	3484	9	3031	14	2838	19	3101
5	4146	10	2069	15	3541	20	2834

1. First, arrange the sample in ascending order.
2. Because n is even, Sample median = average of the 10th and 11th largest observations =  $(3245 + 3248)/2 = 3246.5$  g

Mean= **3166.9 g**

Median= **3246.5 g**

The main strength of the sample median is that it is insensitive to very large or very small values.

# Outliers

- An outlier is any value that is very distant from the other values in a data set.
- The exact definition of an outlier is somewhat subjective, although certain conventions are used in various data summaries and plots (see “Percentiles and Boxplots”).
- Being an outlier in itself does not make a data value invalid or erroneous.
- Still, outliers are often the result of data errors such as mixing data of different units (kilometers versus meters) or bad readings from a sensor.
- When outliers are the result of bad data, the mean will result in a poor estimate of location, while the median will be still be valid.
- In any case, outliers should be identified and are usually worthy of further investigation

# Comparison of the Arithmetic Mean and the Median

In many samples, the relationship between the arithmetic mean and the sample median can be used to assess the symmetry of a distribution.

- In particular, for symmetric distributions the arithmetic mean is approximately the same as the median.
- For positively skewed distributions, the arithmetic mean tends to be larger than the median.
- For negatively skewed distributions, the arithmetic mean tends to be smaller than the median.

# The Geometric Mean

- In mathematics, the geometric mean is a mean or average, which indicates the central tendency or typical value of a set of numbers by using the product of their values (as opposed to the arithmetic mean which uses their sum).
- The geometric mean is defined as the nth root of the product of n numbers, i.e., for a set of numbers  $x_1, x_2, \dots, x_n$ , the geometric mean is defined as

$$\left( \prod_{i=1}^n x_i \right)^{\frac{1}{n}} = \sqrt[n]{x_1 x_2 \cdots x_n}$$

# The Geometric Mean

The minimum inhibitory concentration (MIC) of penicillin G in the urine for *N. gonorrhoeae* in 74 patients

## Distribution of minimum inhibitory concentration (MIC) of penicillin G for *N. gonorrhoeae*

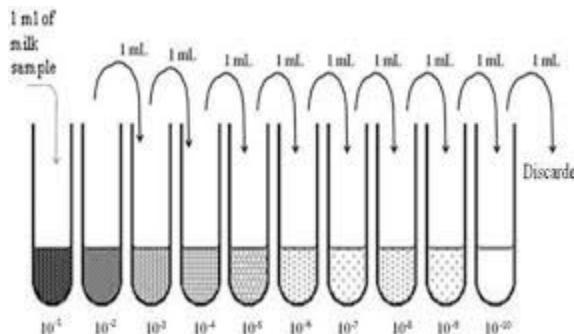
Concentration ( $\mu\text{g/mL}$ )	Frequency	Concentration ( $\mu\text{g/mL}$ )	Frequency
$0.03125 = 2^0(0.03125)$	21	$0.250 = 2^3(0.03125)$	19
$0.0625 = 2^1(0.03125)$	6	$0.50 = 2^4(0.03125)$	17
$0.125 = 2^2(0.03125)$	8	$1.0 = 2^5(0.03125)$	3

Source: Based on JAMA, 220, 205–208, 1972.

The arithmetic mean is not appropriate as a measure of location in this situation because the distribution is very skewed.

# The Geometric Mean

- Many types of laboratory data, specifically data in the form of concentrations of one substance in another, as assessed by serial dilution techniques, can be expressed either as multiples of 2 or as a constant multiplied by a power of 2; that is, outcomes can only be of the form  $2^k c$ ,  $k = 0, 1, \dots$ , for some constant  $c$ .
- For example, the data in Table 2.5 represent the minimum inhibitory concentration (MIC) of penicillin G in the urine for *N. gonorrhoeae* in 74 patients [2]



# The Geometric Mean

The solution is to work with the distribution of the logs of the concentrations.

The log concentrations have the property that successive possible concentrations differ by a constant; that is,  
 $\log(2k+1c) - \log(2k c) = \log(2k+1) + \log c - \log(2k ) - \log c = (k + 1) \log 2 - k \log 2 = \log 2$ .

Thus the log concentrations are equally spaced from each other, and the resulting distribution is now not as skewed as the concentrations themselves.

**The arithmetic mean can then be computed in the log scale; that is,**

$$\overline{\log x} = \frac{1}{n} \sum_{i=1}^n \log x_i$$

However, it is usually preferable to work in the original scale by taking the antilogarithm of  $\log x$  to form the geometric mean

**Home work:** Write a script in R to compute the geometric mean for the sample

**Distribution of minimum inhibitory concentration (MIC)  
of penicillin G for *N. gonorrhoeae***

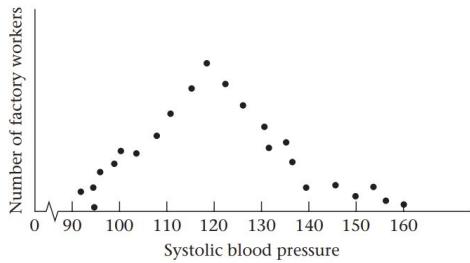
Concentration ( $\mu\text{g/mL}$ )	Frequency	Concentration ( $\mu\text{g/mL}$ )	Frequency
$0.03125 = 2^0(0.03125)$	21	$0.250 = 2^3(0.03125)$	19
$0.0625 = 2^1(0.03125)$	6	$0.50 = 2^4(0.03125)$	17
$0.125 = 2^2(0.03125)$	8	$1.0 = 2^5(0.03125)$	3

Source: Based on *JAMA*, 220, 205–208, 1972.

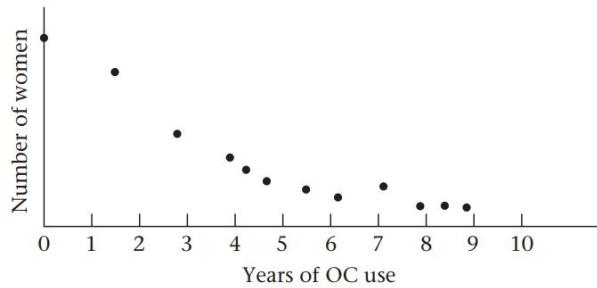
Hint: Use log base=10 for easy computation!

# Comparison of the Arithmetic Mean and the Median

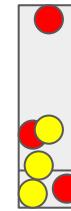
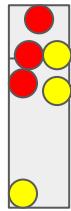
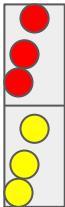
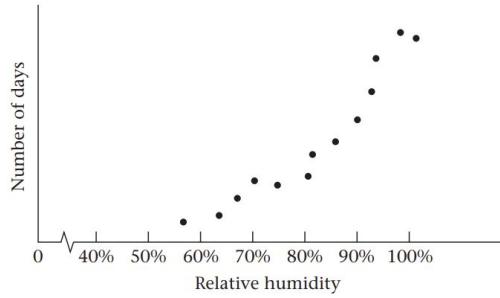
Symmetric Distribution



Positively Skewed Distribution

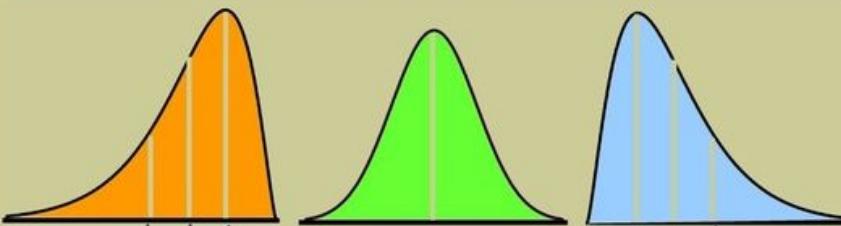


Negatively Skewed Distribution



# Comparison of the Arithmetic Mean and the Median

## Mean, Median, Mode



Mean  
Median  
Mode

Negatively  
Skewed

Mean  
Median  
Mode

Symmetric  
(Not Skewed)

Mode  
Median  
Mean

Positively  
Skewed

# The Mode

The mode is the most frequently occurring value among all the observations in a sample.

**Sample of time intervals between successive menstrual periods (days)  
in college-age women**

Value	Frequency	Value	Frequency	Value	Frequency
24	5	29	96	34	7
25	10	30	63	35	3
26	28	31	24	36	2
27	64	32	9	37	1
28	185	33	2	38	1

Some distributions have more than one mode. In fact, one useful method of classifying distributions is by the number of modes present. A distribution with one mode is called unimodal; two modes, bimodal; three modes, trimodal; and so forth.



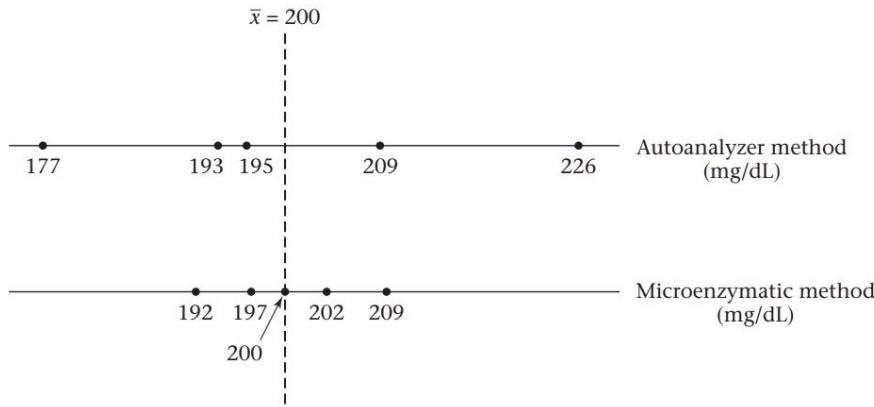
# Hands-on exercises

- Use of R to compute the mean and median for the birth weight data

```
> attach(bwt)
> names(bwt)
[1] "id" "birthwt"
> birthwt
[1] 3265 3260 3245 3484 4146 3323 3649 3200 3031 2069 2581 2841
3609 2838 3541
[16] 2759 3248 3314 3101 2834
> mean(birthwt)
[1] 3166.9
> median(birthwt)
[1] 3246.5
```

# Estimates of Variability

# Measure of Spread



## Measurement methods

1. Autoanalyzer
2. Microenzymatic

Arithmetic means are both 200 mg/dL

greater variability or spread

# The Range

The range is the difference between the largest and smallest observations in a sample.

Sample of birthweights (g) of live-born infants born at a private hospital in San Diego, California, during a 1-week period

$i$	$x_i$	$i$	$x_i$	$i$	$x_i$	$i$	$x_i$
1	3265	6	3323	11	2581	16	2759
2	3260	7	3649	12	2841	17	3248
3	3245	8	3200	13	3609	18	3314
4	3484	9	3031	14	2838	19	3101
5	4146	10	2069	15	3541	20	2834

Calculate Range ?

Advantage of the range is that it is very easy to compute once the sample points are ordered.

Disadvantage is that it is very sensitive to extreme observations. Hence, if the lightest infant in Table above weighed 500 g rather than 2069 g, then the range would increase dramatically to  $4146 - 500 = 3646$  g.

Another disadvantage of the range is that it depends on the sample size ( $n$ ). That is, the larger  $n$  is, the larger the range tends to be. This complication makes it difficult to compare ranges from data sets of differing size.

# Quantiles

Another approach that addresses some of the shortcomings of the range in quantifying the spread in a data set is the use of quantiles or percentiles.

The  $p$ th percentile is defined by

- (1) The  $(k + 1)$ th largest sample point if  $np/100$  is not an integer (where  $k$  is the largest integer less than  $np/100$ ).
- (2) The average of the  $(np/100)$ th and  $(np/100 + 1)$ th largest observations if  $np/100$  is an integer.

Percentiles are also sometimes called **quantiles**.

The spread of a distribution can be characterized by specifying several percentiles. For example, the 10th and 90th percentiles are often used to characterize spread.

Percentiles have the advantage over the range of being less sensitive to outliers and of not being greatly affected by the sample size ( $n$ ).

Compute the 10th and 90th percentiles for the birthweight data (below)

Sample of birthweights (g) of live-born infants born at a private hospital in San Diego, California, during a 1-week period

$i$	$x_i$	$i$	$x_i$	$i$	$x_i$	$i$	$x_i$
1	3265	6	3323	11	2581	16	2759
2	3260	7	3649	12	2841	17	3248
3	3245	8	3200	13	3609	18	3314
4	3484	9	3031	14	2838	19	3101
5	4146	10	2069	15	3541	20	2834

Solution: Because  $20 \times .1 = 2$  and  $20 \times .9 = 18$  are integers, the 10th and 90th percentiles are defined by

10th percentile: average of the second and third largest values =  $(2581 + 2759)/2 = 2670$  g

90th percentile: average of the 18th and 19th largest values =  $(3609 + 3649)/2 = 3629$  g

We would estimate that 80% of birthweights will fall between 2670 g and 3629 g, which gives an overall impression of the spread of the distribution

Compute the 20th percentile for the white-blood-count data

**Sample of admission white-blood counts  
( $\times 1000$ ) for all patients entering a hospital  
in Allentown, Pennsylvania, on a given day**

$i$	$x_i$	$i$	$x_i$
1	7	6	3
2	35	7	10
3	5	8	12
4	9	9	8
5	8		

Because  $np/100 = 9 \times .2 = 1.8$  is not an integer, the 20th percentile is defined by the  $(1 + 1)$ th largest value = second largest value = 5000.

### Use of R to compute sample quantiles

```
> quantile(birthwt,probs = c(0.1, 0.9),na.rm = TRUE, type = 2)
```

10% 90%  
2670 3629

```
> quantile(white.count, probs = 0.2, na.rm = TRUE, type = 2)  
20% 5
```

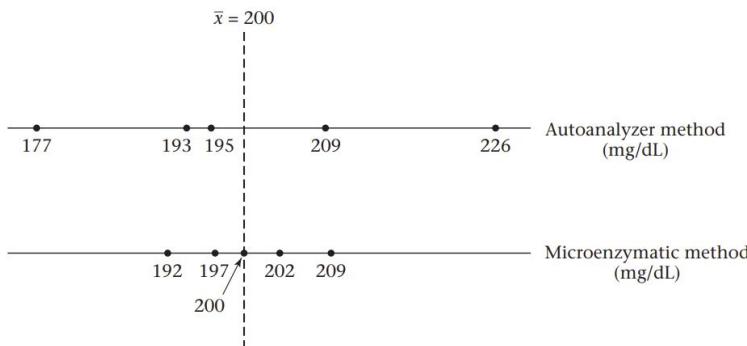
# The Variance and Standard Deviation

If the center of the sample is defined as the arithmetic mean, then a measure that can summarize the difference (or deviations) between the individual sample points and the arithmetic mean is needed; that is

$$x_1 - \bar{x}, x_2 - \bar{x}, \dots, x_n - \bar{x}$$

One simple measure that would seem to accomplish this goal is

$$d = \frac{\sum_{i=1}^n (x_i - \bar{x})}{n}$$



The sum of the deviations of the individual observations of a sample about the sample mean is always zero.

# A little bit Improvement: mean deviation

$$\sum_{i=1}^n |x_i - \bar{x}| / n$$

The mean deviation is a reasonable measure of spread but does not characterize the spread as well as the standard deviation if the underlying distribution is bell-shaped.

# Sample Variance (or variance)

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$$

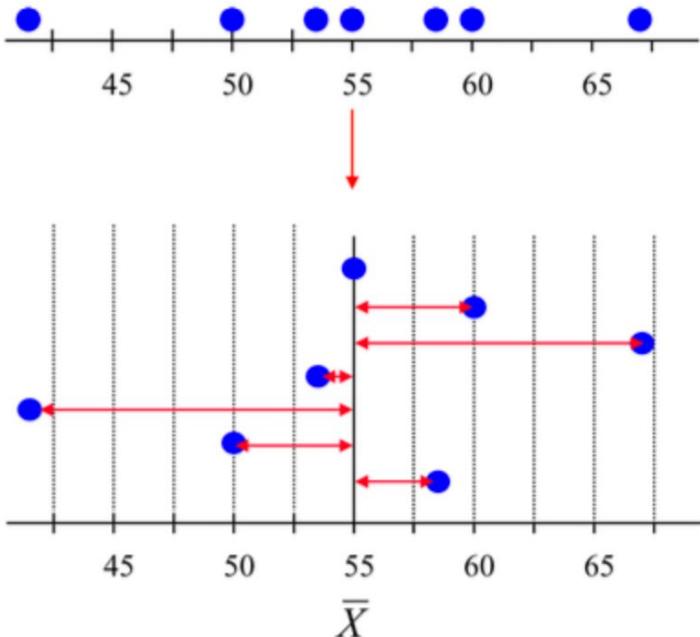
$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

Course activity 1: Why we use “n-1” not n, while calculating Variance ?

# Standard deviation

- The standard deviation is the most common dispersion measure in statistics.
- Like the mean for the location measures, if we have to present one statistics which summarizes the spread of the data, it is usually the standard deviation.
- As its name suggests, the standard deviation tells what is the “normal” deviation of the data.
- It actually computes the average deviation from the mean.
- The larger the standard deviation, the more scattered the data are.
- On the contrary, the smaller the standard deviation, the more the data are centred around the mean.

# Standard deviation



- There are two formulas depending on whether we face a sample or a population.
- The standard deviation for a population is denoted  $\sigma$ .
- The standard deviation for a sample is denoted  $s$ .

# Standard deviation

Population

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

the standard deviation is  
actually the average deviation  
of the data from their mean

$\mu$

Sample

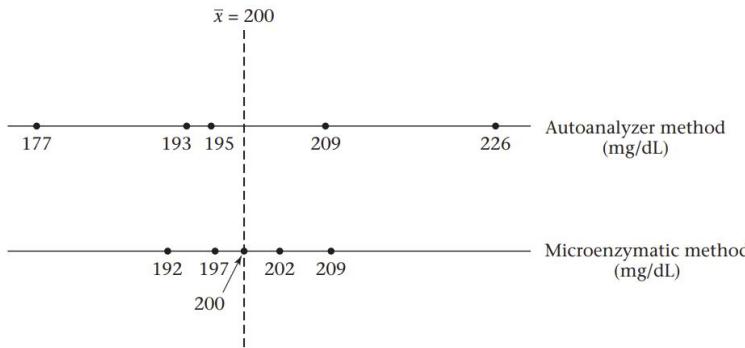
$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

the standard deviation is  
actually the average deviation  
of the data from the sample  
mean

# Standard Deviation

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}} = \sqrt{\text{sample variance}}$$

Compute the variance and standard deviation for the Autoanalyzer- and Microenzymatic-method data



## Solution: Autoanalyzer Method

$$\begin{aligned}s^2 &= \left[ (177-200)^2 + (193-200)^2 + (195-200)^2 + (209-200)^2 + (226-200)^2 \right] / 4 \\&= (529 + 49 + 25 + 81 + 676) / 4 = 1360 / 4 = 340 \\s &= \sqrt{340} = 18.4\end{aligned}$$

## Microenzymatic Method

$$\begin{aligned}s^2 &= \left[ (192-200)^2 + (197-200)^2 + (200-200)^2 + (202-200)^2 + (209-200)^2 \right] / 4 \\&= (64 + 9 + 0 + 4 + 81) / 4 = 158 / 4 = 39.5 \\s &= \sqrt{39.5} = 6.3\end{aligned}$$

# Properties of Variance and SD



shifted by a constant c



Suppose there are two samples

$$x_1, \dots, x_n \quad \text{and} \quad y_1, \dots, y_n$$

$$\text{where } y_i = x_i + c, \quad i = 1, \dots, n$$

If the respective sample variances of the two samples are denoted by

$$s_x^2 \text{ and } s_y^2$$

$$\text{then } s_y^2 = s_x^2$$

# Properties of Variance and SD

Suppose there are two samples

$$x_1, \dots, x_n \quad \text{and} \quad y_1, \dots, y_n$$

where  $y_i = cx_i$ ,  $i = 1, \dots, n$ ,  $c > 0$

Then  $s_y^2 = c^2 s_x^2$   $s_y = cs_x$

$$\begin{aligned}s_y^2 &= \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1} = \frac{\sum_{i=1}^n (cx_i - c\bar{x})^2}{n-1} \\&= \frac{\sum_{i=1}^n [c(x_i - \bar{x})]^2}{n-1} = \frac{\sum_{i=1}^n c^2(x_i - \bar{x})^2}{n-1} \\&= \frac{c^2 \sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = c^2 s_x^2 \\s_y &= \sqrt{c^2 s_x^2} = cs_x\end{aligned}$$

# Standard deviation vs. Variance

- Standard deviation and variance are often used interchangeably and both quantify the spread of a given dataset by measuring how far the observations are from their mean.
- However, the standard deviation can be more easily interpreted because the unit for the standard deviation is the same than the unit of measurement of the data (while it is the unit<sup>2</sup> for the variance).

	Population Sample	
Standard deviation	$\sigma$	$s$
Variance	$\sigma^2$	$s^2$

## Practical Application !

WK	Time (min)( $x_i$ )	WK	Time (min)( $x_i$ )
1	12.80	10	11.57
2	12.20	11	11.73
3	12.25	12	12.67
4	12.18	13	11.92
5	11.53	14	11.67
6	12.47	15	11.80
7	12.30	16	12.33
8	12.08	17	12.55
9	11.72	18	11.83

Q1. What is the mean 1 mile running time over 18 weeks?

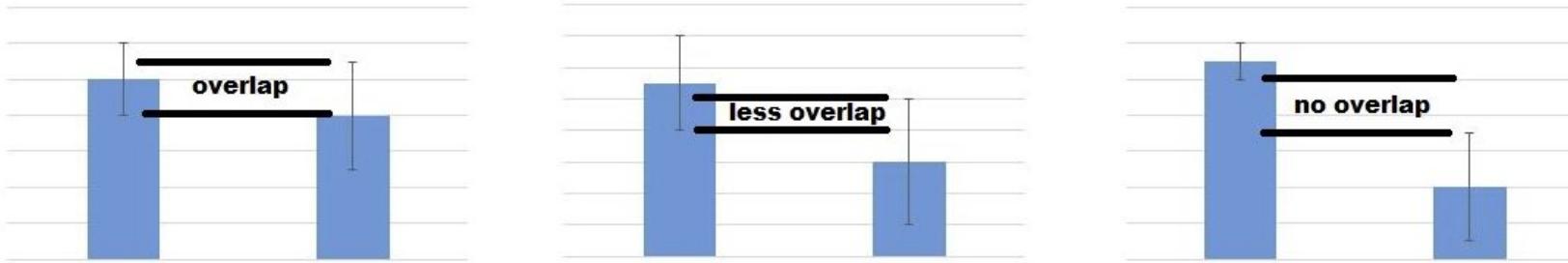
Q2. What is standard deviation of the 1 mile running time over 18 weeks?

Q3. Suppose we construct a new variable called  $\text{time\_100} = 100 \times \text{time}$  (e.g., for week 1,  $\text{time\_100} = 1280$ ). What is the mean and standard deviation of  $\text{time\_100}$ ?

Q4. Suppose the man does not run for 6 months over the winter due to snow on the ground. He resumes running once a week in the spring and records a running time = 12.97 minutes in his first week of running in the spring. Is this an outlying value relative to the distribution of running times recorded the previous year? Why or why not?

# Coefficient of Variation (CV)

- It is useful to relate the arithmetic mean and the standard deviation to each other



- In all the above cases, we need to perform Statistical test to ensure if the means are different.
- A standard deviation of 10 means something different conceptually if the arithmetic mean is 10 versus if it is 1000.

The coefficient of variation (CV) is defined by

$$100\% \times (s/\bar{x})$$

# Coefficient of Variation (CV)

The coefficient of variation (CV) is defined by

$$100\% \times (s/\bar{x})$$

This measure remains the same regardless of what units are used because if the units change by a factor c, then both the mean and standard deviation change by the factor c; while the CV, which is the ratio between them, remains unchanged.

Question: Compute the coefficient of variation for the data in Table below when the birth weights are expressed in either grams or ounces.

Sample of birthweights (g) of live-born infants born at a private hospital in San Diego, California, during a 1-week period

$i$	$x_i$	$i$	$x_i$	$i$	$x_i$	$i$	$x_i$
1	3265	6	3323	11	2581	16	2759
2	3260	7	3649	12	2841	17	3248
3	3245	8	3200	13	3609	18	3314
4	3484	9	3031	14	2838	19	3101
5	4146	10	2069	15	3541	20	2834

$$CV = 100\% \times (s/\bar{x}) = 100\% \times (445.3\text{ g}/3166.9\text{ g}) = 14.1\%$$

If the data were expressed in ounces, then

$$CV = 100\% \times (15.7 \text{ oz}/111.71 \text{ oz}) = 14.1\%$$

# Coefficient of Variation (CV)

- The CV is most useful in comparing the variability of several different samples, each with different arithmetic means.
- This is because a higher variability is usually expected when the mean increases, and the CV is a measure that accounts for this variability.
- Thus, if we are conducting a study in which air pollution is measured at several sites and we wish to compare day-to-day variability at the different sites, we might expect a higher variability for the more highly polluted sites.
- A more accurate comparison could be made by comparing the CVs at different sites than by comparing the standard deviations.
- The CV is also useful for comparing the reproducibility of different variables.

# Grouped Data

Sometimes the sample size is too large to display all the raw data.

Consider the data set in Table below, which represents the birthweights from 100 consecutive deliveries at a Boston hospital. Suppose we wish to display these data for publication purposes.

**Question:** How can we do this?

**Answer:** The simplest way to display the data is to generate a frequency distribution using a statistical package.

Sample of birthweights (oz) from 100 consecutive deliveries at a Boston hospital

58	118	92	108	132	32	140	138	96	161
120	86	115	118	95	83	112	128	127	124
123	134	94	67	124	155	105	100	112	141
104	132	98	146	132	93	85	94	116	113
121	68	107	122	126	88	89	108	115	85
111	121	124	104	125	102	122	137	110	101
91	122	138	99	115	104	98	89	119	109
104	115	138	105	144	87	88	103	108	109
128	106	125	108	98	133	104	122	124	110
133	115	127	135	89	121	112	135	115	64

# Grouped Data

A frequency distribution is an ordered display of each value in a data set together with its frequency, that is, the number of times that value occurs in the data set. In addition, the percentage of sample points that take on a particular value is also typically given.

Frequency distribution of the birthweight data  
on Table 2.9 using the FREQ procedure of SAS

Birthweight	Frequency	Percent	Cumulative Frequency	Cumulative Percent
32	1	1.00	1	1.00
58	1	1.00	2	2.00
64	1	1.00	3	3.00
67	1	1.00	4	4.00
68	1	1.00	5	5.00
83	1	1.00	6	6.00
85	2	2.00	8	8.00
86	1	1.00	9	9.00

For any particular birthweight b, the Cumulative Frequency is the number of birthweights in the sample that are less than or equal to b.

The Percent =  $100 \times \text{Frequency}/n$

Cumulative Percent =  $100 \times \text{Cumulative Frequency}/n$  = the percentage of birthweights less than or equal to b.

If the number of unique sample values is large, then a frequency distribution may still be too detailed a summary for publication purposes.

# Grouped Data

Here are some general instructions for categorizing the data:

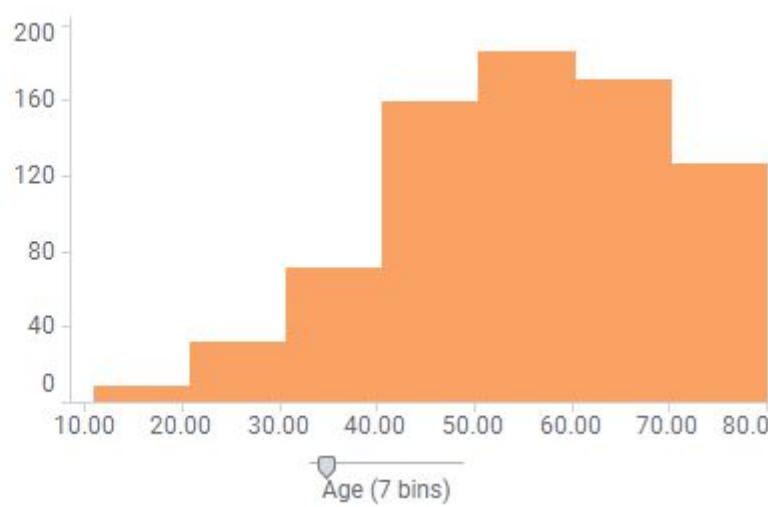
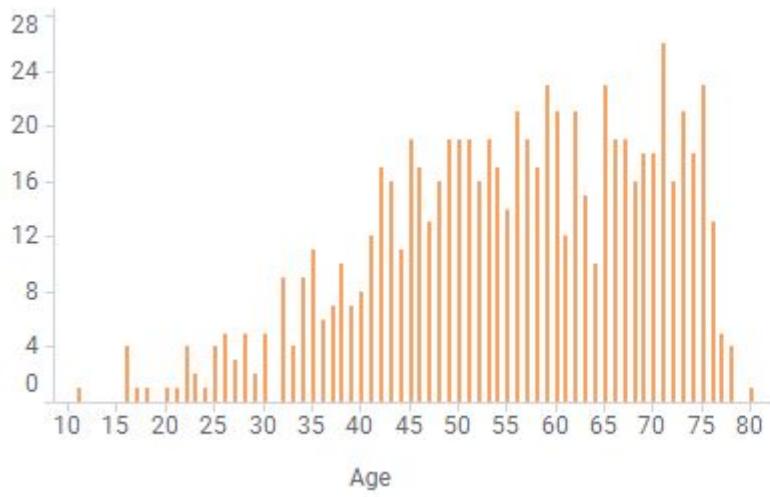
General layout of grouped data

Group interval	Frequency
$y_1 \leq x < y_2$	$f_1$
$y_2 \leq x < y_3$	$f_2$
.	.
.	.
$y_i \leq x < y_{i+1}$	$f_i$
.	.
.	.
$y_k \leq x < y_{k+1}$	$f_k$

Grouped frequency distribution of the birthweight (oz) from 100 consecutive deliveries

The FREQ Procedure				
Group_interval	Frequency	Percent	Cumulative Frequency	Cumulative Percent
29.5 $\leq$ x $<$ 69.5	5	5.00	5	5.00
69.5 $\leq$ x $<$ 89.5	10	10.00	15	15.00
89.5 $\leq$ x $<$ 99.5	11	11.00	26	26.00
99.5 $\leq$ x $<$ 109.5	19	19.00	45	45.00
109.5 $\leq$ x $<$ 119.5	17	17.00	62	62.00
119.5 $\leq$ x $<$ 129.5	20	20.00	82	82.00
129.5 $\leq$ x $<$ 139.5	12	12.00	94	94.00
139.5 $\leq$ x $<$ 169.5	6	6.00	100	100.00

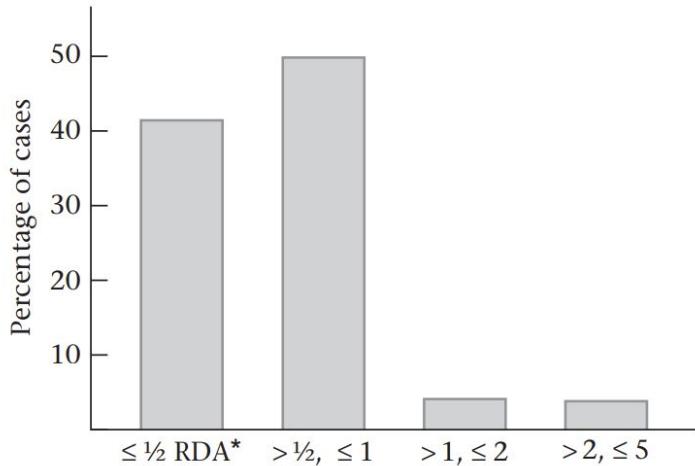
# Binning continuous data is a bad scientific practice (IMHO) !



# **Graphics Methods**

# Graphic Methods

## Bar Graphs



## Stem-and-leaf plot

Stem	Leaf	#
16	1	1
15	5	1
15		
14	6	1
14	014	3
13	557888	6
13	222334	6
12	5567788	7
12	0111222234444	13
11	555556889	10
11	0012223	7
10	5567888899	10
10	012344444	9
9	568889	6
9	12344	5
8	556788999	9
8	3	1
7		
7		
6	78	2
6	4	1
5	8	1
5		
4		
4		
3		
3	2	1

Multiply Stem.Leaf by  $10^{**+1}$

```
> stem.leaf(bwt$birthweight,  
unit=1, trim.outliers=FALSE)
```

## Problems:

- (1) The definition of the groups is somewhat arbitrary
- (2) the sense of what the actual sample points are within the respective groups is lost.

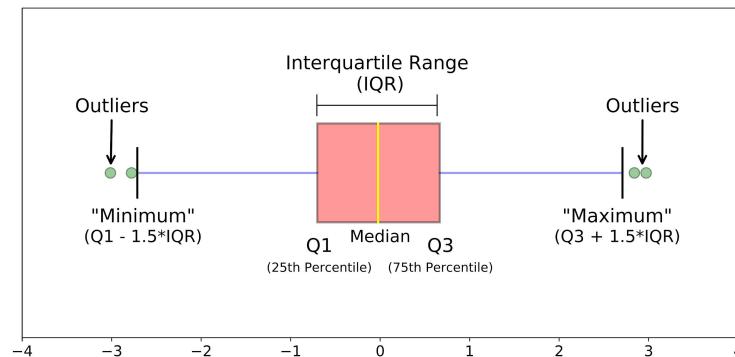
# BOX PLOT

A box plot uses the relationships among the median, upper quartile, and lower quartile to describe the skewness of a distribution

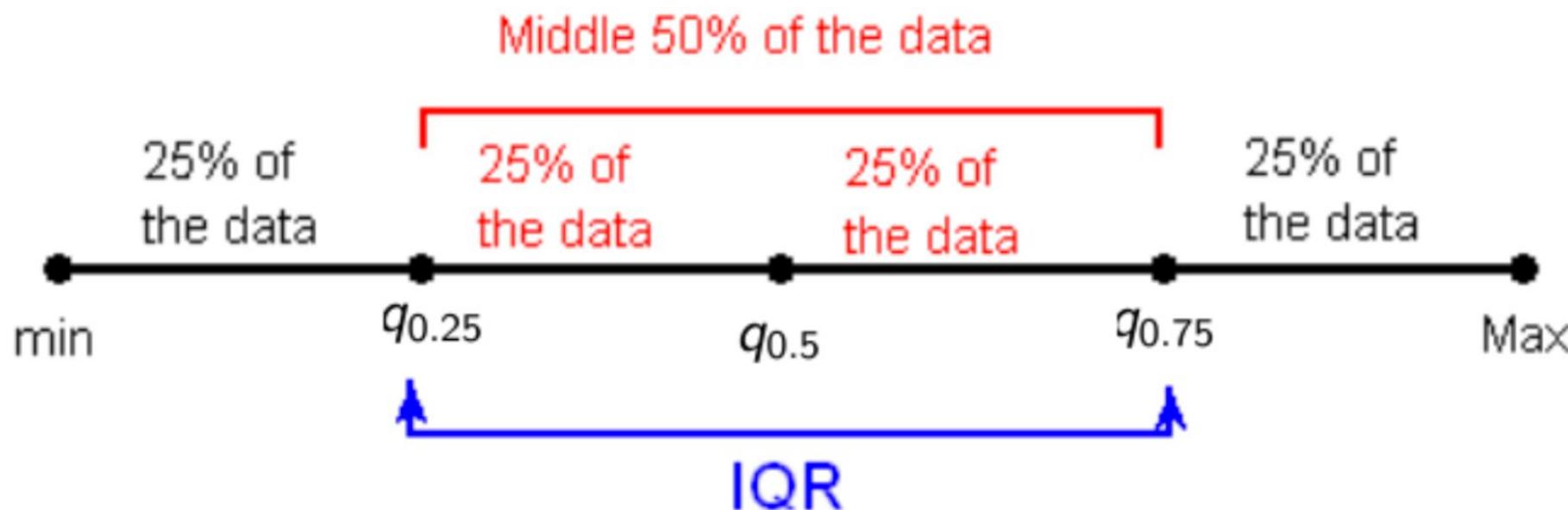
The upper and lower quartiles can be thought of conceptually as the approximate 75th and 25th percentiles of the sample—that is, the points  $3/4$  and  $1/4$  along the way in the ordered sample.

How can the median, upper quartile, and lower quartile be used to judge the symmetry of a distribution?

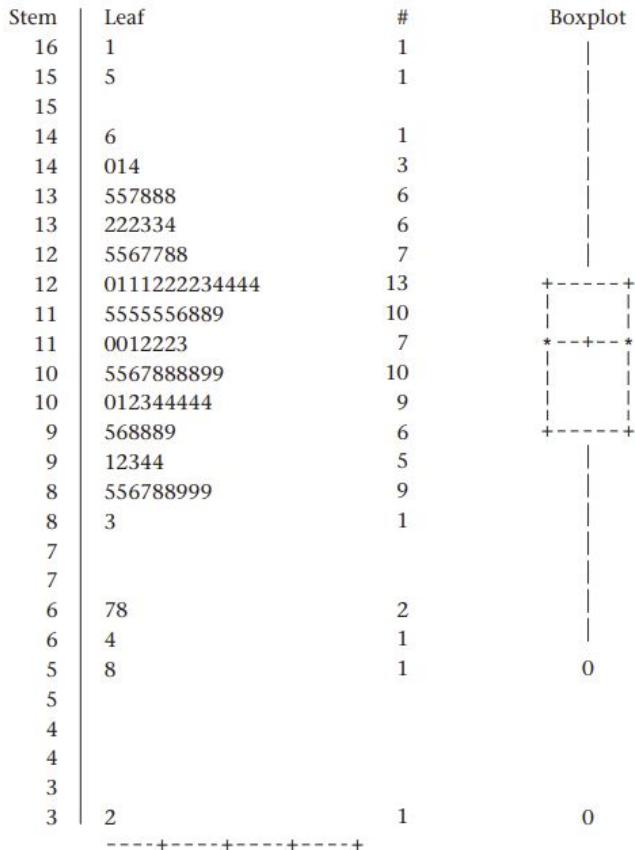
- (1) If the distribution is symmetric, then the upper and lower quartiles should be approximately equally spaced from the median.
- (2) If the upper quartile is farther from the median than the lower quartile, then the distribution is positively skewed.
- (3) If the lower quartile is farther from the median than the upper quartile, then the distribution is negatively skewed.



# BOX PLOT



## BOX PLOT/STEM-LEAF PLOT



An **outlying value** is a value  $x$  such that either

- (1)  $x > \text{upper quartile} + 1.5 \times (\text{upper quartile} - \text{lower quartile})$  or
- (2)  $x < \text{lower quartile} - 1.5 \times (\text{upper quartile} - \text{lower quartile})$

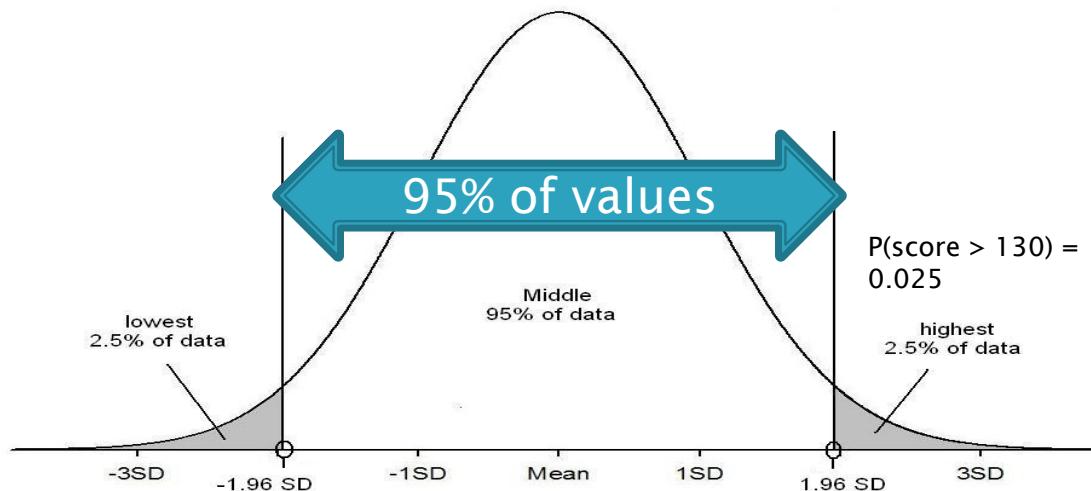
An **extreme outlying value** is a value  $x$  such that either

- (1)  $x > \text{upper quartile} + 3.0 \times (\text{upper quartile} - \text{lower quartile})$  or
- (2)  $x < \text{lower quartile} - 3.0 \times (\text{upper quartile} - \text{lower quartile})$

The box plot is then completed by

- (1) Drawing a vertical bar from the upper quartile to the largest nonoutlying value in the sample
- (2) Drawing a vertical bar from the lower quartile to the smallest nonoutlying value in the sample
- (3) Individually identifying the outlying and extreme outlying values in the sample by zeroes (0) and asterisks (\*), respectively

## 95% 1.96 x SD's from the mean



**70**

$$mean - (1.96 \times SD)$$

$$100 - (1.96 \times 15.3) = 70$$

**100**

**130**

$$mean + (1.96 \times SD)$$

$$100 + (1.96 \times 15.3) = 130$$

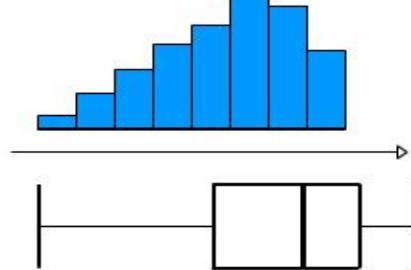
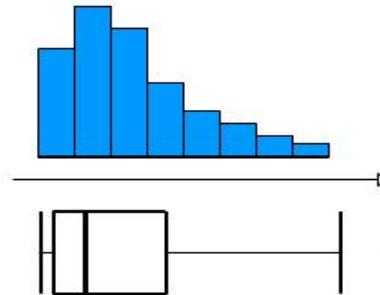
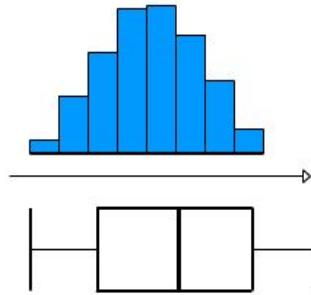
**95% of people have an IQ between 70 and 130**

## Graphical Representation

Charts can be used to **informally** assess whether data is:

Normally  
distributed

Or....Skewed

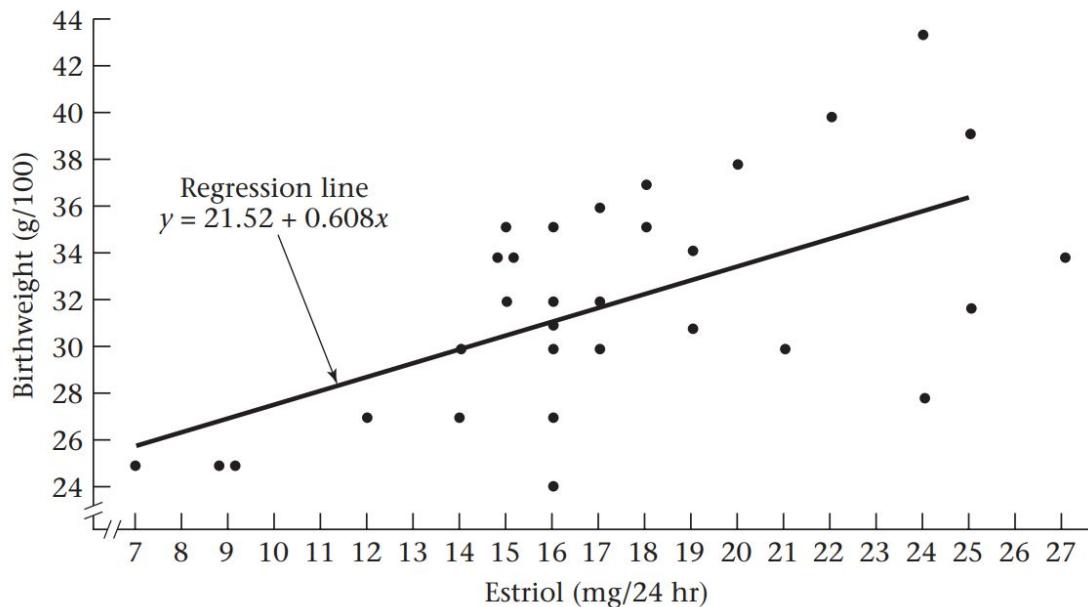


**The mean and median are very different for skewed data.**

# **Regression and Correlation Methods**

# Regression

$$E(y|x) = \alpha + \beta x$$



The line  $y = \alpha + \beta x$  is the **regression line**, where  $\alpha$  is the intercept and  $\beta$  is the slope of the line.

# Regression

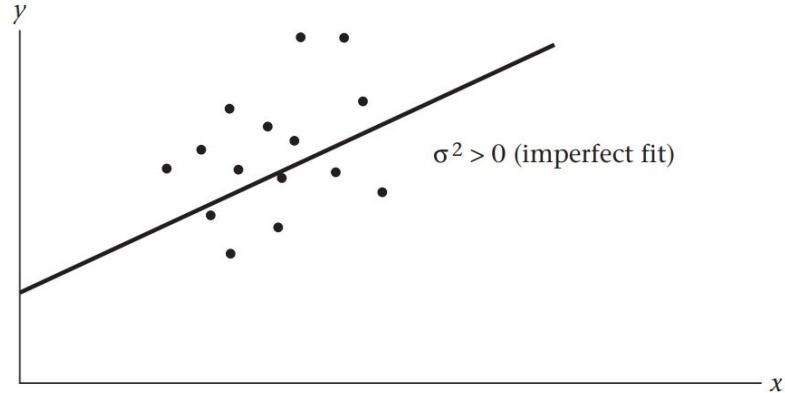
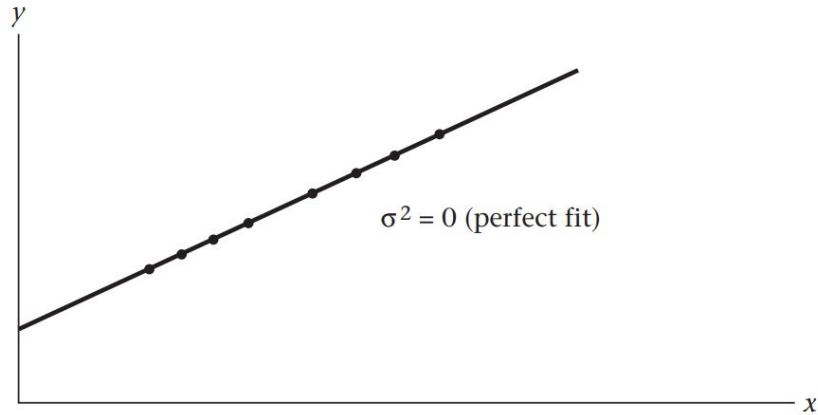
For any linear-regression equation of the form  $y = \alpha + \beta x + e$ ,  $y$  is called the **dependent variable** and  $x$  is called the **independent variable** because we are trying to predict  $y$  as a function of  $x$ .

$$y = \alpha + \beta x + e$$

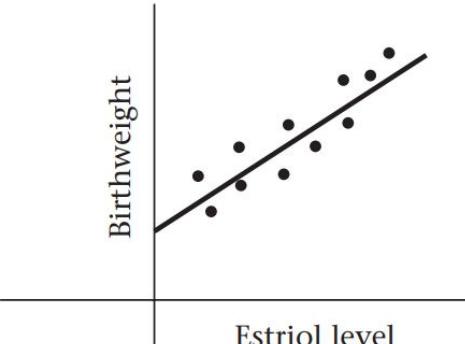
e=Error Term

where  $e$  is normally distributed with mean 0 and variance  $\sigma^2$ .

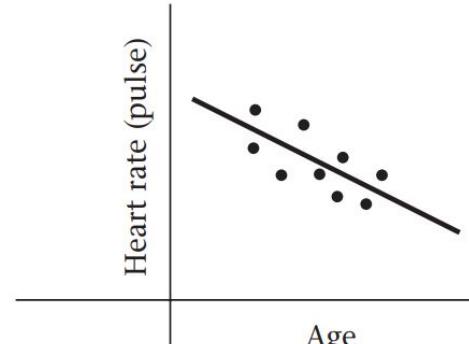
# The effect of $\sigma^2$ on the goodness of fit of a regression line



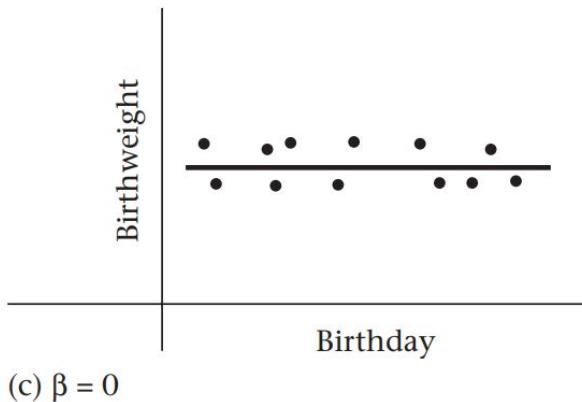
# Interpretation of the regression line for different values of $\beta$



(a)  $\beta > 0$

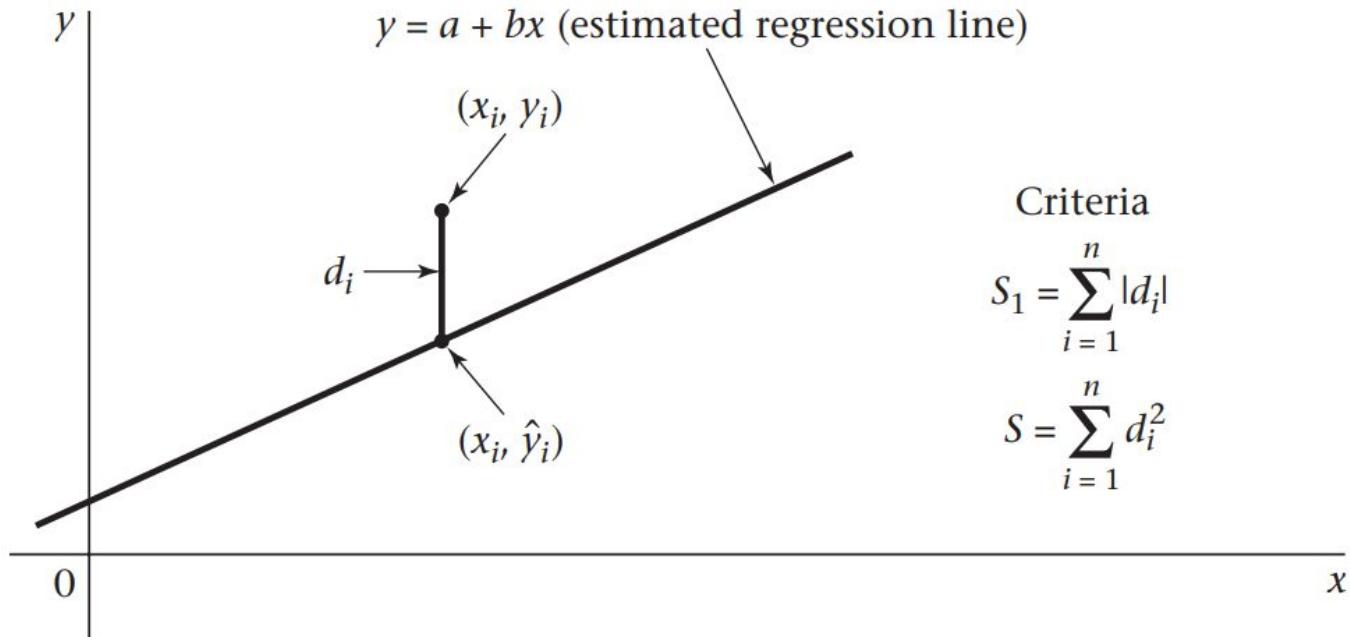


(b)  $\beta < 0$



(c)  $\beta = 0$

# Fitting Regression Lines: The Least Square Method



# Fitting Regression Lines: The Least Square Method

Consider the data in Figure 11.4 and the estimated regression line  $y = a + bx$ . The distance  $d_i$  of a typical sample point  $(x_i, y_i)$  from the line could be measured along a direction parallel to the  $y$ -axis. If we let  $(x_i, \hat{y}_i) = (x_i, a + bx_i)$  be the point on the estimated regression line at  $x_i$ , then this distance is given by  $d_i = y_i - \hat{y}_i = y_i - a - bx_i$ . A good-fitting line would make these distances as small as possible. Because the  $d_i$  cannot all be 0, the criterion  $S_1 = \text{sum of the absolute deviations of the sample points from the line} = \sum_{i=1}^n |d_i|$  can be used and the line that minimizes  $S_1$  can be found. Instead, for both theoretical reasons and ease of derivation, the following least-squares criterion is commonly used:

$S = \text{sum of the squared distances of the points from the line}$

$$= \sum_{i=1}^n d_i^2 = \sum_{i=1}^n (y_i - a - bx_i)^2$$

# Fitting Regression Lines: The Least Square Method

---

The **least-squares line**, or **estimated regression line**, is the line  $y = a + bx$  that minimizes the sum of squared distances of the sample points from the line given by

$$S = \sum_{i=1}^n d_i^2$$

This method of estimating the parameters of a regression line is known as the **method of least squares**.

---

# Correlation Coefficient

**Cardiovascular Disease** Serum cholesterol is an important risk factor in the etiology of cardiovascular disease. Much research has been devoted to understanding the environmental factors that cause elevated cholesterol levels. For this purpose, cholesterol levels were measured on 100 genetically unrelated spouse pairs. We are not interested in predicting the cholesterol level of a husband from that of his wife but instead would like some quantitative measure of the relationship between their levels. We will use the correlation coefficient for this purpose.

First, we discuss the related concept of covariance. The *covariance* is a measure used to quantify the relationship between two random variables.

# Correlation Coefficient

---

The **covariance** between two random variables  $X$  and  $Y$  is denoted by  $\text{Cov}(X, Y)$  and is defined by

$$\text{Cov}(X, Y) = E[(X - \mu_x)(Y - \mu_y)]$$

$$cov_{x,y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N - 1}$$

which can also be written as  $E(XY) - \mu_x\mu_y$ , where  $\mu_x$  is the average value of  $X$ ,  $\mu_y$  is the average value of  $Y$ , and  $E(XY) = \text{average value of the product of } X \text{ and } Y$ .

---

It can be shown that if the random variables  $X$  and  $Y$  are independent, then the covariance between them is 0. If large values of  $X$  and  $Y$  tend to occur among the same subjects (as well as small values of  $X$  and  $Y$ ), then the covariance is positive. If large values of  $X$  and small values of  $Y$  (or conversely, small values of  $X$  and large values of  $Y$ ) tend to occur among the same subjects, then the covariance is negative.

# Correlation Coefficient

---

The **correlation coefficient** between two random variables  $X$  and  $Y$  is denoted by  $\text{Corr}(X, Y)$  or  $\rho$  and is defined by

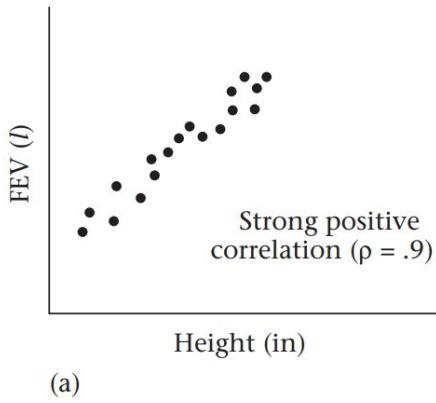
$$\rho = \text{Corr}(X, Y) = \text{Cov}(X, Y) / (\sigma_x \sigma_y)$$

where  $\sigma_x$  and  $\sigma_y$  are the standard deviations of  $X$  and  $Y$ , respectively.

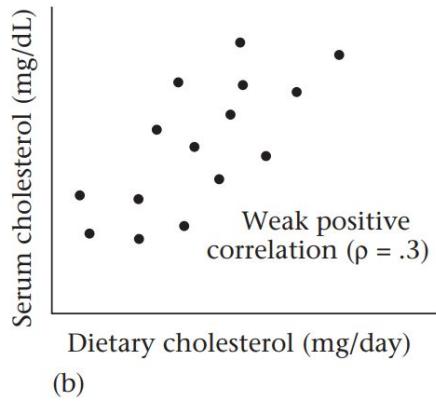
---

- Unlike the covariance, the correlation coefficient is a dimensionless quantity that is independent of the units of  $X$  and  $Y$  and ranges between  $-1$  and  $1$ .
- For random variables that are approximately linearly related, a correlation coefficient of  $0$  implies independence.
- A correlation coefficient close to  $1$  implies nearly perfect positive dependence with large values of  $X$  corresponding to large values of  $Y$  and small values of  $X$  corresponding to small values of  $Y$ .

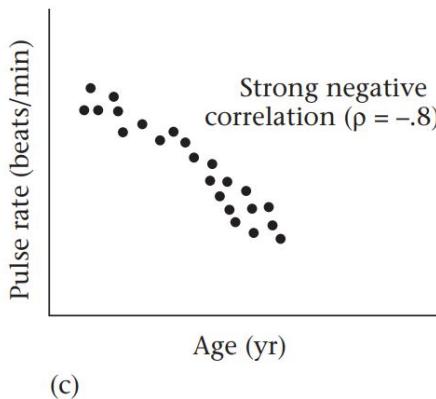
# Interpretation of various degrees of correlation



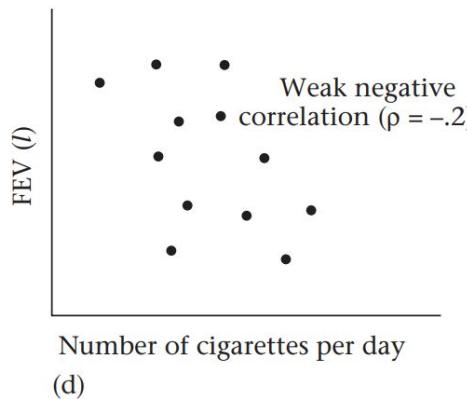
(a)



(b)



(c)



(d)

# Hands-On Session

[https://colab.research.google.com/drive/1KakItqQgloDT7dvM6ekIFHQPbjHjg7yT?usp=share\\_link](https://colab.research.google.com/drive/1KakItqQgloDT7dvM6ekIFHQPbjHjg7yT?usp=share_link)

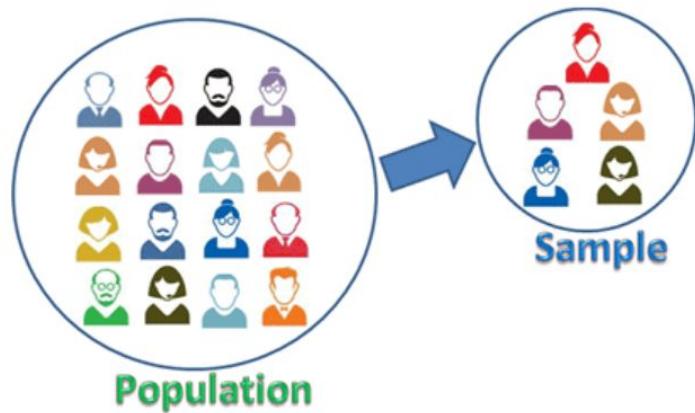


# Sampling

- Data Sampling is the selection of statistical samples from the population to estimate the characteristics of the entire population.
- It is the main technique for data collection when you want to create a statistically sound conclusion from a subset of a population of data.
- Data sampling helps to make statistical inferences about the population.

# Populations and Samples

- **Population:** Population is the group of elements which has common characteristics. It is a collection of observations we would like to make inferences about.
- **Sample:** A sample is the subset of a population
- **Sampling:** A collection of samples from the population is a sampling. In other words, sampling units are an overlapping collection of elements from the population.



# Why Use Data Sampling?

- Sometimes, gathering information on a complete population is too expensive, time-consuming, or nonsensical.
- For example: The process we are measuring would require destructive testing (think taste tests, car crash tests, etc.).
- Getting data from the entire population is too expensive or would take longer than we have.
- Getting total population data is just too hard.



Link For R Script is here:

[SamplingError.R](#)

# Sampling Error

- Sampling error is the deviation between the estimate of an ideal sample and the true population.
- The core assumption of data sampling is that samples are a subset of the population, and the sample mean is equal to the mean of the population.
- To the degree that doesn't happen is the term Sampling Error
- We can reduce sampling error by following sampling best practices, like having a large enough sample size, choosing the right kind of sampling, and avoiding sampling bias.



Link For R Script is here:

[SamplingError.R](#)

# Data Sampling Methods

**There are Two Types of Sampling:**

- **Probability Sampling:** Every element in the sample population has an equal chance of being selected. A sampling method is biased if every member of the population doesn't have an equal likelihood of being in the sample.
- **Non-probability Sampling:** Opposite to Probability Sampling

# Different Types of Probability Sampling

- Simple Random Sampling:
- Stratified Sampling
- Systematic Sampling
- Cluster Sampling

# Simple Random Sampling

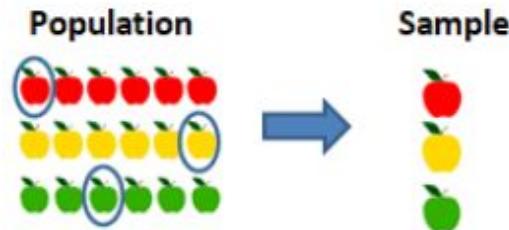
It is a method of sampling in which every element of the universe has an equal probability of being chosen. For example, choose an individual from a lottery. The advantage of this method is that it is free from personal bias, and the universe gets fairly represented by samples.



Link For R Script is here:  
[SimpleRandomSampling.R](#)

# Stratified Sampling

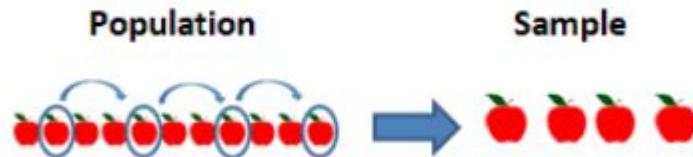
The population is broken down into non-overlapping groups. In other words, strata (elements within the subgroups that are homogenous or heterogeneous). Then random samples are taken from each stratum, representing the entire population. The advantage of this method is it covers all the elements of the population. But there is a possibility of bias at the time of classification of the population.



Link For R Script is here:  
[StatifizedSampling.R](#)

# Systematic Sampling

Samples are selected from the population according to a pre-determined rule. In other words, every nth element is selected from the population as a sample. Arrange all the elements in a sequence and then select the samples from the population at regular intervals.

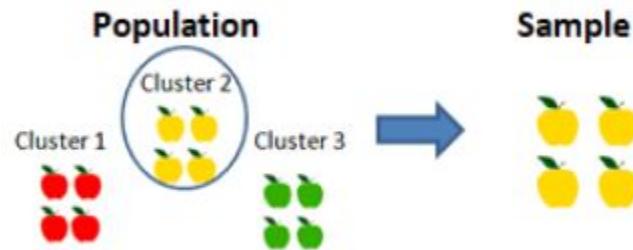


Link For R Script is here:  
[SystematicSampling\\_KthStart.R](#)

# Cluster Sampling

The population is divided into many different clusters, then clusters or subgroups are randomly selected.

For example, clusters are of different ages, sex, locations, etc.



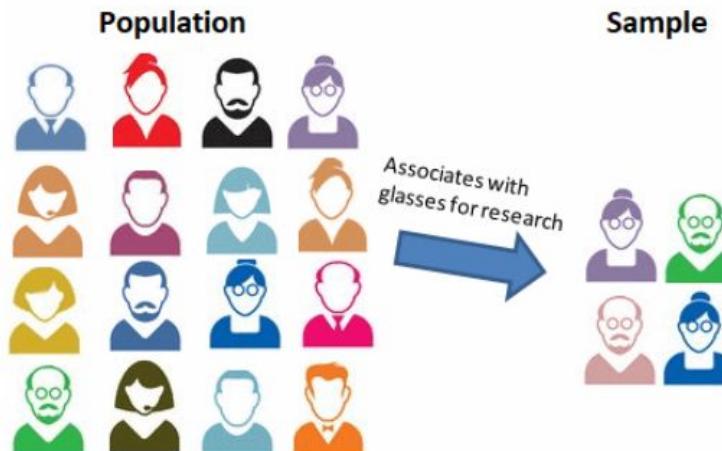
Link For R Script is here:  
[ClusteredSampling.R](#)

# Different Types of Non-Probability Sampling

- Purposive Sampling
- Convenience Sampling
- Quota Sampling
- Snowball/referral Sampling

# Purposive Sampling

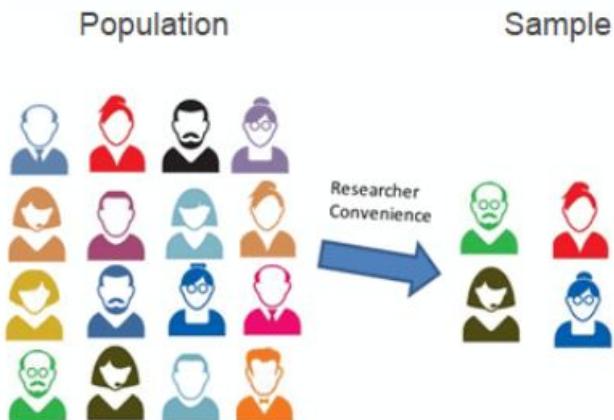
Purposive sampling is also known as judgment sampling. Samples are selected based on the purpose or intention of the research. The method is flexible to include those items in the sample that are of special significance.



Link For R Script is here:  
[PurposiveSampling.R](#)

# Convenience Sampling

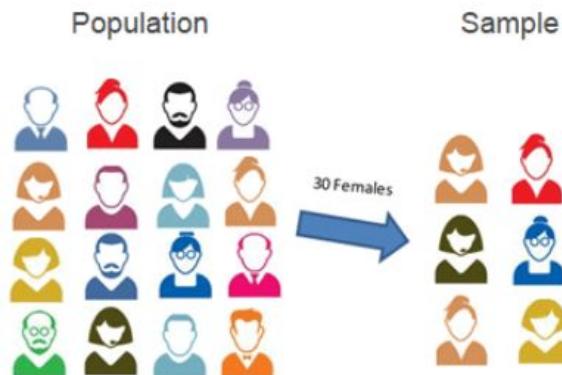
Convenience sampling is one of the easiest sampling methods. Samples selection is based on availability and the selection of convenient samples for the researcher.



Link For R Script is here:  
[convenience\\_sample.R](#)

# Quota Sampling

It is one type of stratified sampling, where samples are collected in each subgroup until the desired quota is met. The proportion of this sample does not match the proportion of the group to the population.



Link For R Script is here:

[Quota\\_Sampling.R](#)

# Snowball/Referral Sampling

Snowball sampling or referral sampling is the method famous in medical and social science surveys where the population is unknown, and difficult to get the sample. Hence researchers will take help from the existing elements to refer the others as samples who can fit in the population. Since it is based on referrals, there is a chance of bias.



Link For R Script is here:  
[SnowBall\\_Sampling.R](#)

# Kinds of Sampling Bias

Following are the different types of sampling bias

- **Response Bias:** A response or data bias is a systematic bias that occurs during data collection that influences the response.
- **Voluntary response Bias:** Occurs when individuals can choose to participate.
- **Non-response Bias:** Non-response bias occurs when units selected as part of the sampling procedure do not respond in whole or part.
- **Convenience Bias:** When a sample is taken from individuals that are conveniently available.

# Representative Sample

- A representative sample refers to a subset selected from a broader statistical population or a collection of factors, ensuring an accurate reflection of the larger group in terms of the specific characteristic or quality being investigated.
- Increasing the sample size reduces the probability of sampling errors and enhances the probability that the sample effectively mirrors the characteristics of the target population.
- Representative samples reduce the risk of selection bias, ensuring that all segments of the population have an equal chance of being included in the sample.



Link For R Script is here:

[RepresentativeSample.R](#)

# Homogeneity

- Homogeneity in data sampling represents the degree of similarity among elements within a sample. When a sample is homogeneous, its elements share common characteristics, contributing to a more focused and targeted analysis.
- While a representative sample aims to encompass the diversity of the entire population, homogeneity becomes particularly important when studying specific traits or characteristics within a subgroup. This focus on similarity enhances the precision and accuracy of results within that particular subset of the population.



Link For R Script is here:  
[HomogenousSample.R](#)

# Sample Size for Data Sampling

- How Large Should a Data Sample Be?

- When you pick a sample size, there will always be a trade-off between precision and cost.

This trade-off depends on:

- The type of data being sampled (continuous or discrete)
  - How precise do you want your statistical inferences to be?
  - The estimate of the standard deviation for the entire population.
  - The confidence level desired.

Sample size is the number of observations collected from a population; it is a subset of the population to make inferences about the population.



Link For R Script is here:

[SampleSize.R](#)

# Sample Size Needed for Hypothesis Testing Depends on:

- Desired Risk (Both alpha and beta)
- The minimum value to be detected between the population means
- The variation in the characteristic being measured ( $S$  or  $\sigma$ )—the population variance.
- Population size does NOT come into the determination of how big a population is.

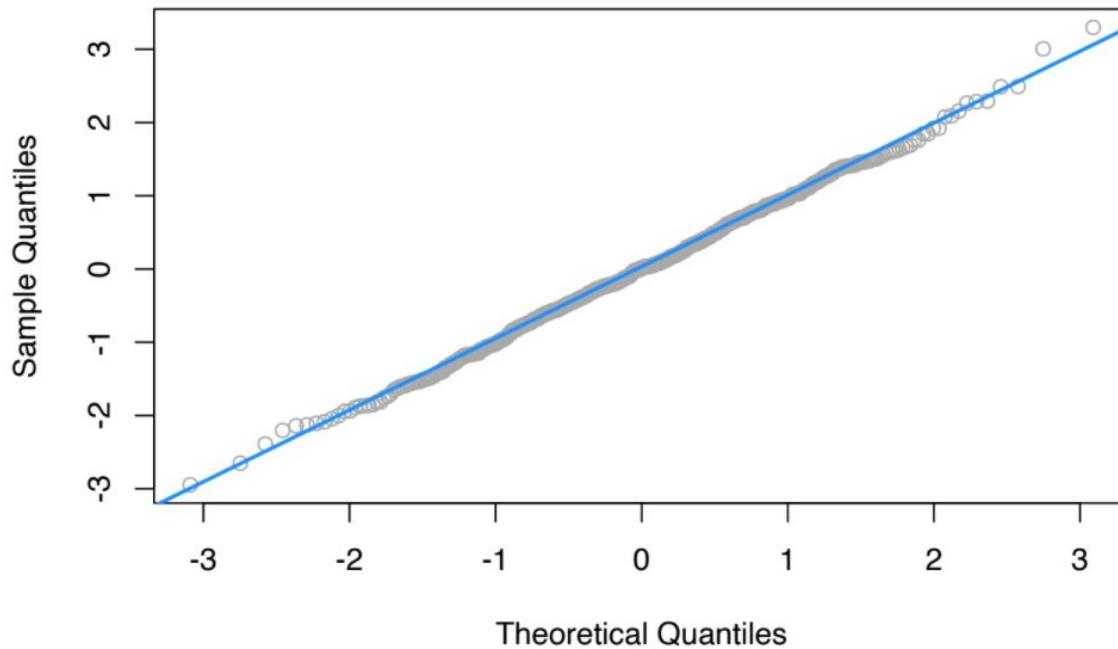
# Test for Normality

# Q-Q plot

Another visual method for assessing the normality of errors, which is more powerful than a histogram, is a normal quantile-quantile plot, or **Q-Q plot** for short.

```
qqnorm(vector, main = "Normal Q-Q Plot, fit_1", col = "darkgrey")  
qqline(vector, col = "dodgerblue", lwd = 2)
```

### Normal Q-Q Plot, fit\_1



In short, if the points of the plot do not closely follow a straight line, this would suggest that the data do not come from a normal distribution.

# Shapiro-Wilk Test

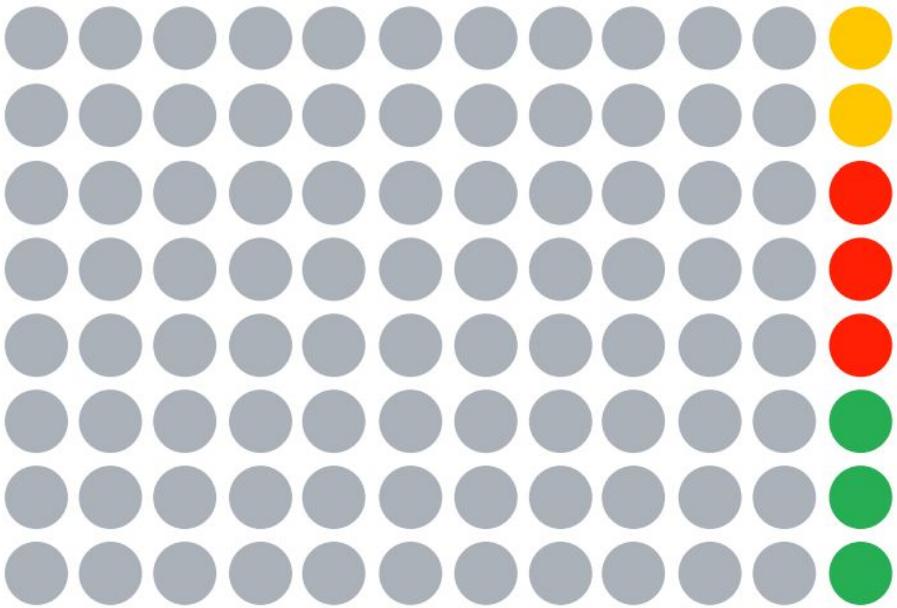
Histograms and Q-Q Plots give a nice visual representation of the residuals distribution, however if we are interested in formal testing, there are a number of options available. A commonly used test is the **Shapiro–Wilk test**, which is implemented in R.

```
shapiro.test(rexp(25))
```

```
##  Shapiro-Wilk normality test
##
## data:  rexp(25)
## W = 0.71164, p-value = 1.05e-05
```

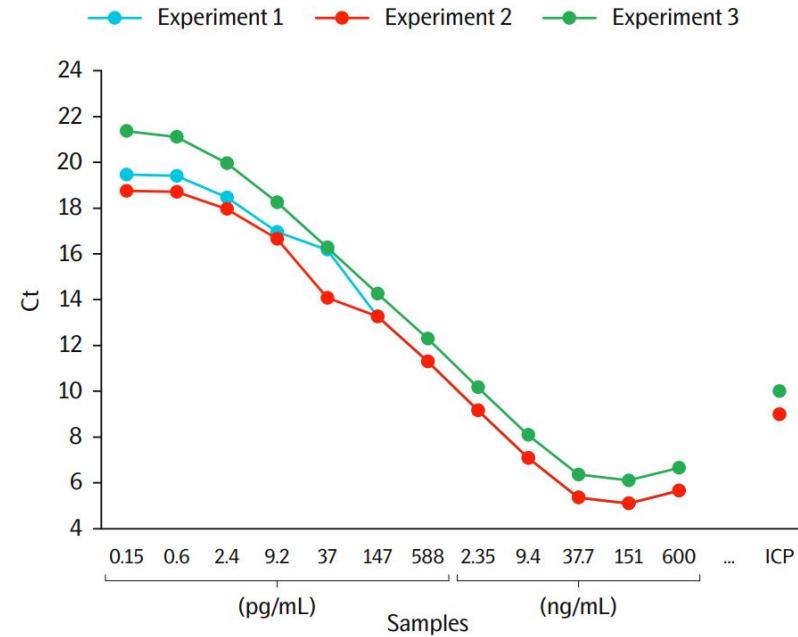
The null hypothesis of this test is that the data follows a normal distribution. A small p-value indicates rejection of the null hypothesis, suggesting non-normality.

# Data Transformation



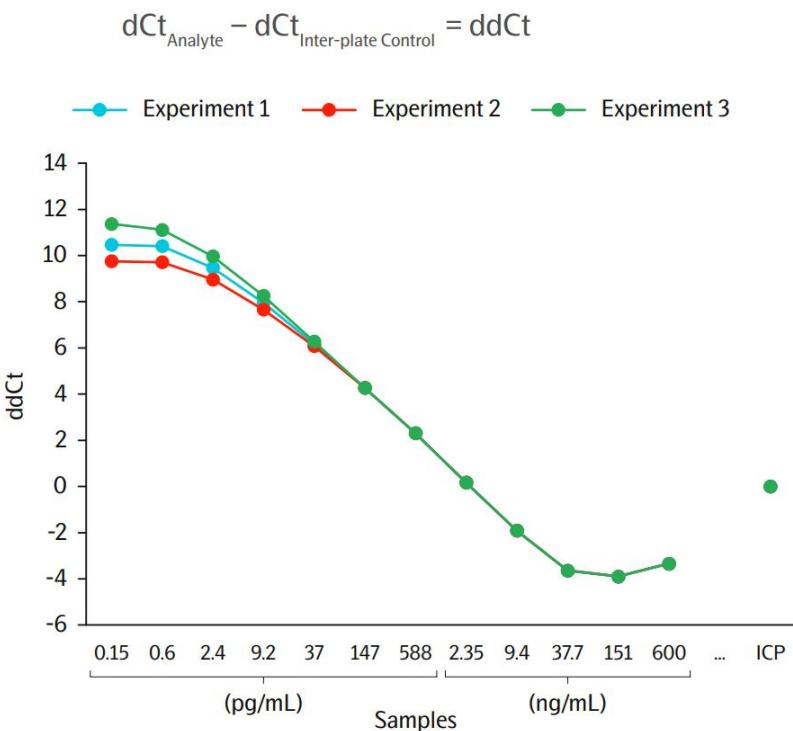
- **Inter-plate controls**
- **Negative controls**
- **Sample controls**

Below is an example of raw data for three runs with a calibrator curve. There are deviating samples in the middle, and one curve has a parallel shift.



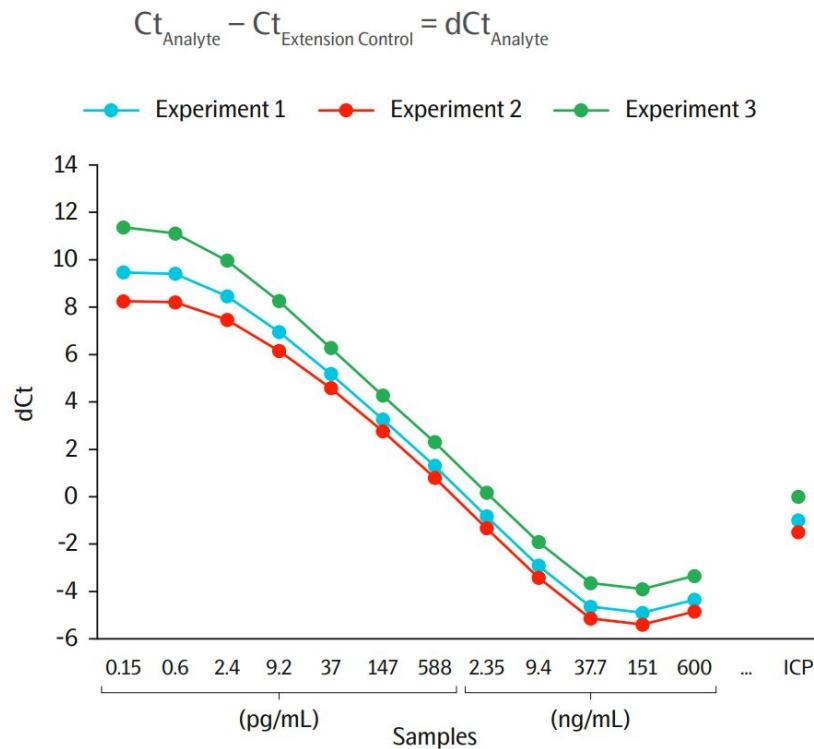
## Normalized against IPCs

To improve inter-assay repeatability, the data is then normalized against the IPCs. This is done per assay and plate and may be followed by intensity or reference sample normalization depending on the study characteristics.



## Adjusted against Extension Control

The raw data is then adjusted against the Extension Control per sample to improve intra-assay repeatability by reducing technical variation introduced in the extension step.



BEFORE

AFTER

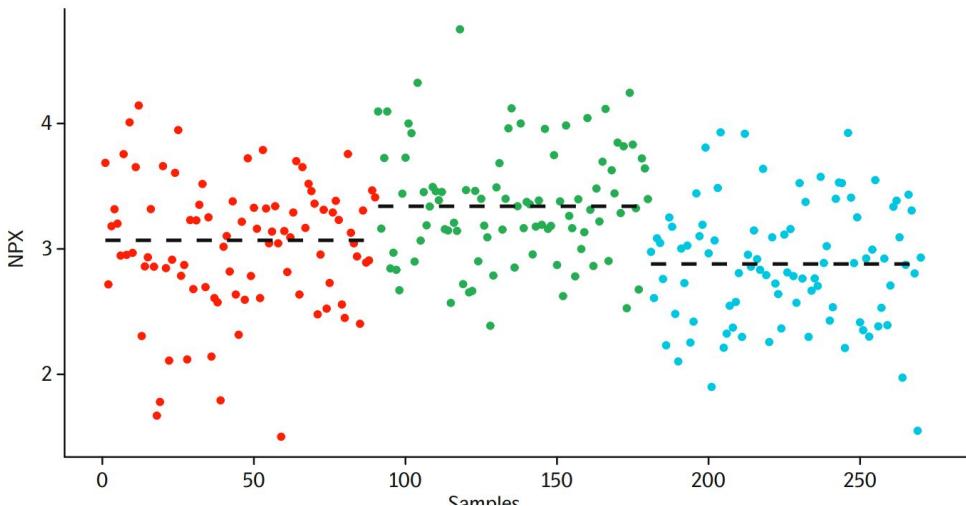


Plate 1  
Plate 2  
Plate 3

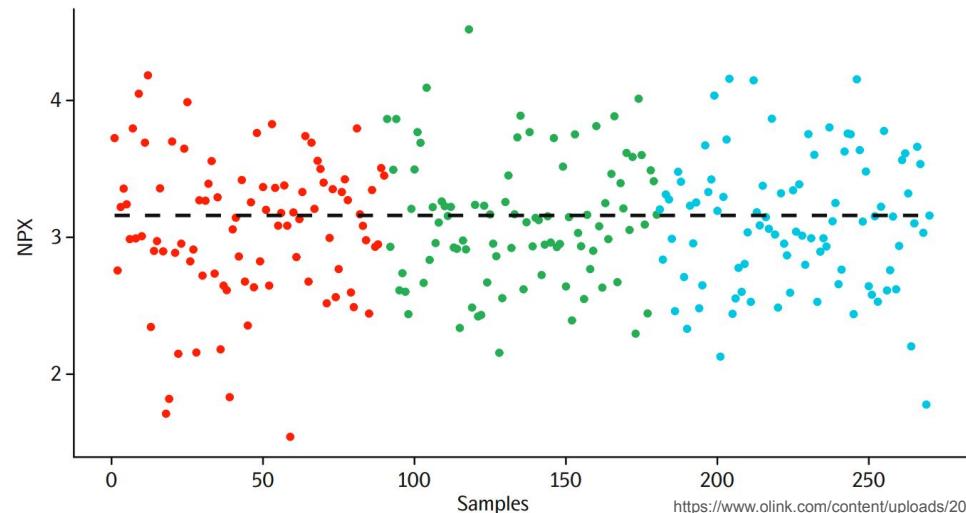
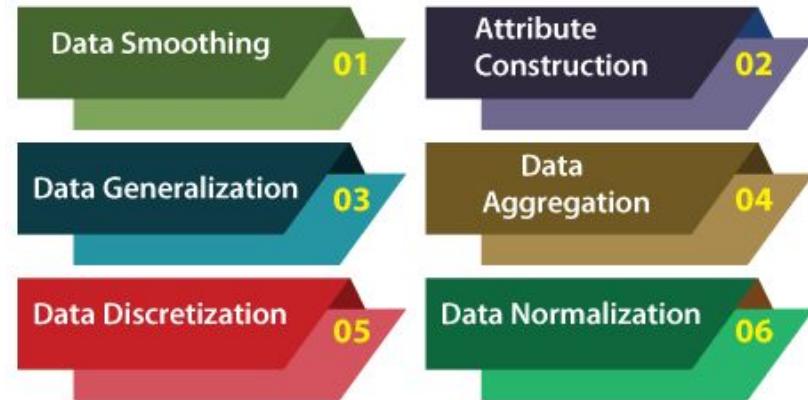


Plate 1  
Plate 2  
Plate 3

# What is Data Transformation

- Data transformation is the mutation of data characteristics to improve access or storage.
- Transformation may occur on the format, structure, or values of data. With regard to data analytics, transformation usually occurs after data is extracted or loaded.
- Data transformation increases the efficiency of analytic processes and enables data-driven decisions. Raw data is often difficult to analyze and too vast in quantity to derive meaningful insight, hence the need for clean, usable data.

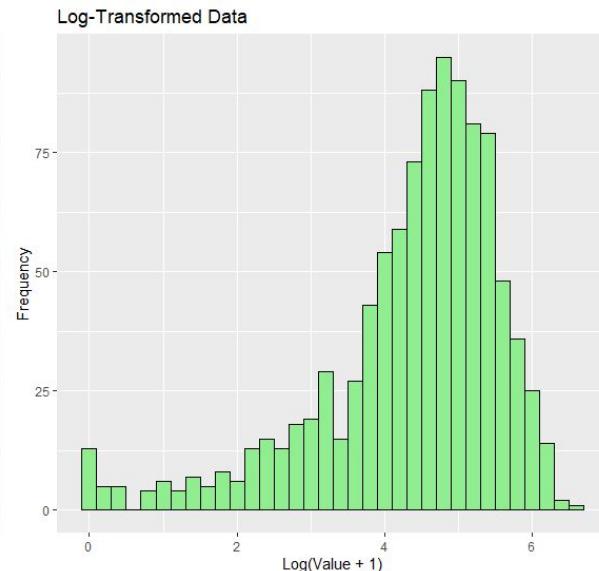
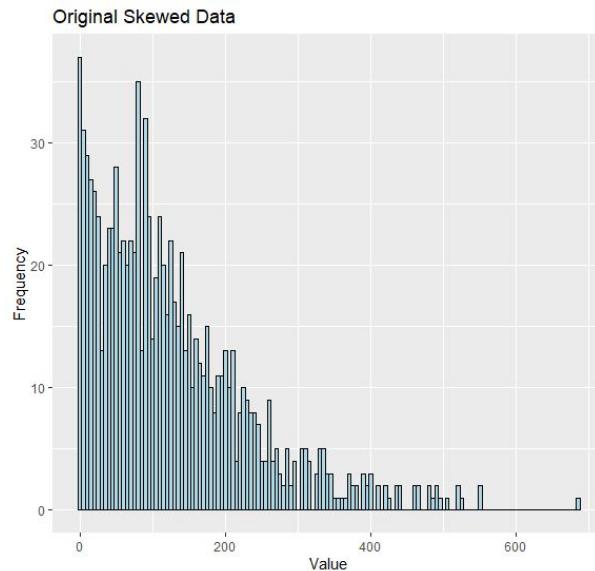


# Types of Data Transformation

- **Constructive:** The data transformation process adds, copies, or replicates data.
- **Destructive:** The system deletes fields or records.
- **Aesthetic:** The transformation standardizes the data to meet requirements or parameters.
- **Structural:** The database is reorganized by renaming, moving, or combining columns.

# Constructive Data Transformation

**Constructive data transformation involves modifying the original dataset to improve the performance of a statistical model or to meet the assumptions of a particular analysis.**



# Constructive Data Transformation

## 1. Log Transformation:

- Parameter: Base of the logarithm (e.g., natural logarithm, base-10 logarithm).
- Purpose: Reduces the impact of extreme values and stabilizes variance.

## 2. Square Root Transformation:

- Parameter: None.
- Purpose: Stabilizes variance and reduces the impact of extreme values.

## 3. Box-Cox Transformation:

- Parameter: Lambda ( $\lambda$ ) parameter.
- Purpose: Generalization of power transformations, stabilizes variance, and handles different types of transformations.

## 4. Reciprocal Transformation:

- Parameter: None.
- Purpose: Inverts the values, useful for data with reciprocal relationships.

# Constructive Data Transformation

## 5. Exponential Transformation:

- Parameter: Power or rate of growth.
- Purpose: Amplifies the differences between small values, useful for data with exponential relationships.

## 6. Square Transformation:

- Parameter: None.
- Purpose: Amplifies the differences between large values, useful for data with quadratic relationships.

## 7. Inverse Hyperbolic Sine (ASinh) Transformation:

- Parameter: None.
- Purpose: Stabilizes variance, particularly for count data or percentages.

## 8. Rank Transformation:

- Parameter: None.
- Purpose: Transforms the data into ranks, useful for non-parametric analyses and handling non-normal distributions.

# Constructive Data Transformation

## 9. Winsorizing:

- Parameter: Trimming percentage or threshold values.
- Purpose: Reduces the impact of outliers by capping extreme values at a specified threshold.

## 10. Centering and Scaling:

- Parameter: Mean, standard deviation, or other scaling factors.
- Purpose: Centers the data around a specific value and scales it to a desired range.

## 11. Winsorized Standardization:

- Parameter: Trimming percentage or threshold values.
- Purpose: Combines winsorizing and standardization to handle outliers.

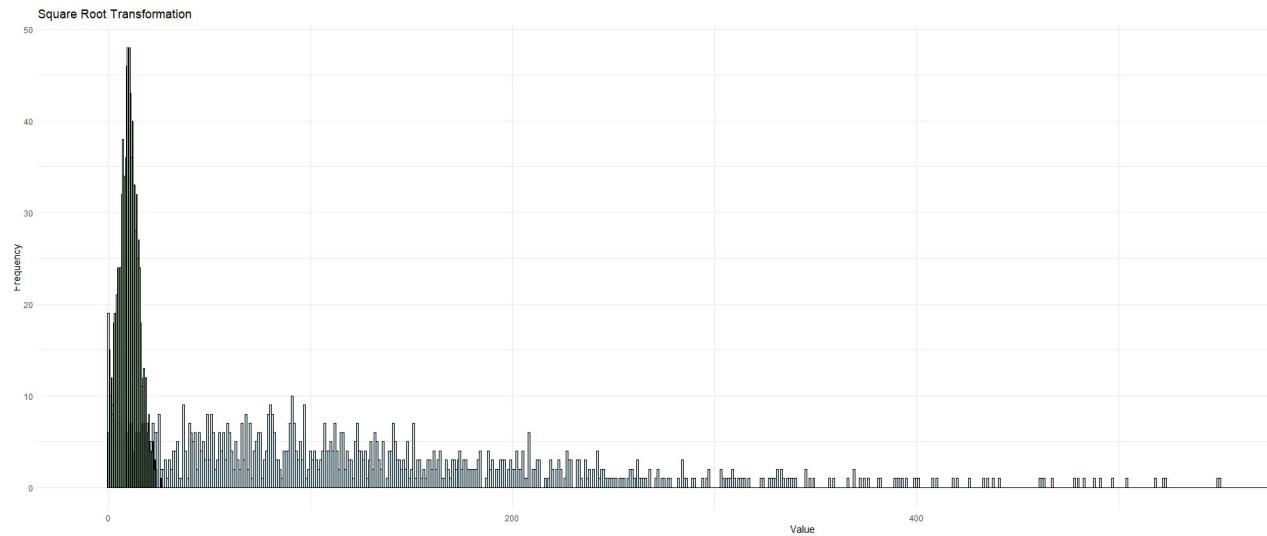
## 12. Truncate Transformation:

- Parameter: Truncation points.
- Purpose: Cuts off extreme values beyond specified thresholds.

# Square Root Transformation

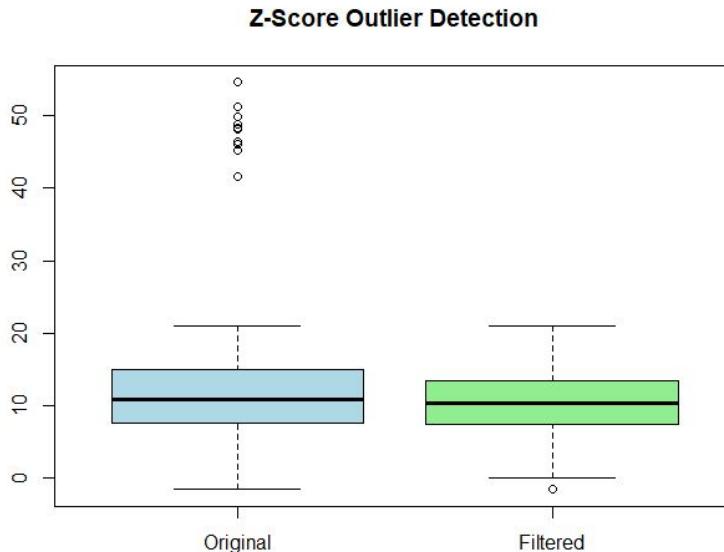
## Square Root Transformation:

- Parameter: None.
- Purpose: Stabilizes variance and reduces the impact of extreme values.



# Destructive Data Transformation

Destructive data transformation involves removing or altering data points, you might be interested in outlier detection and handling techniques.



**1. Z-Score or Standard Score:**

- Identify and remove data points that fall outside a certain threshold of standard deviations from the mean.

**2. Modified Z-Score:**

- Similar to the Z-score but more robust to outliers, calculated using the median and median absolute deviation.

**3. IQR (Interquartile Range) Method:**

- Identify outliers based on the spread of the middle 50% of the data.

**4. Trimmed Mean:**

- Calculate the mean after removing a certain percentage of extreme values from both ends of the dataset.

**5. Winsorizing:**

- Cap extreme values at a certain threshold, effectively limiting their impact on the analysis.

**6. Tukey's Fences:**

- Determine outliers using a combination of the median and the interquartile range.

**7. Mahalanobis Distance:**

- Identify outliers based on the distance of each data point from the center of the data distribution.

**8. DBSCAN (Density-Based Spatial Clustering of Applications with Noise):**

- Clustering method that identifies outliers as points not belonging to any cluster.

**9. Isolation Forest:**

- Tree-based algorithm that isolates outliers in the data.

**10. One-Class SVM (Support Vector Machine):**

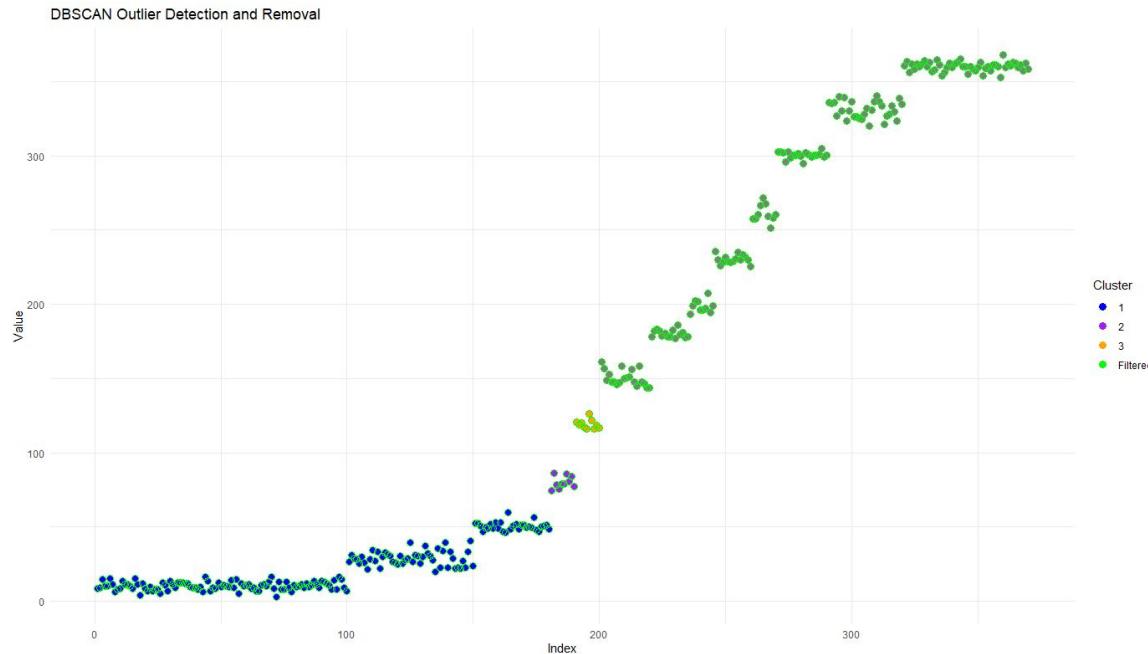
- Anomaly detection algorithm that identifies outliers based on deviations from the majority class.

**11. Quantile-Quantile (Q-Q) Plot:**

- Visual method to identify outliers by comparing the quantiles of the observed data with the quantiles of a theoretical distribution.

# DBSCAN

**DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a clustering algorithm that can be used for outlier detection.**



# Data Smoothing

**Data smoothing is a technique used to reduce noise in a dataset, making underlying patterns or trends more apparent. Several methods can be employed for data smoothing.**

## **Moving Average:**

- Simple technique that calculates the average of a subset of neighboring data points.

## **Exponential Smoothing:**

- Weighted average method where recent data points are given more weight than older ones.

## **Lowess (Locally Weighted Scatterplot Smoothing):**

- Non-parametric method that fits localized linear regressions to the data, giving more weight to nearby points.

## **Savitzky-Golay Filter:**

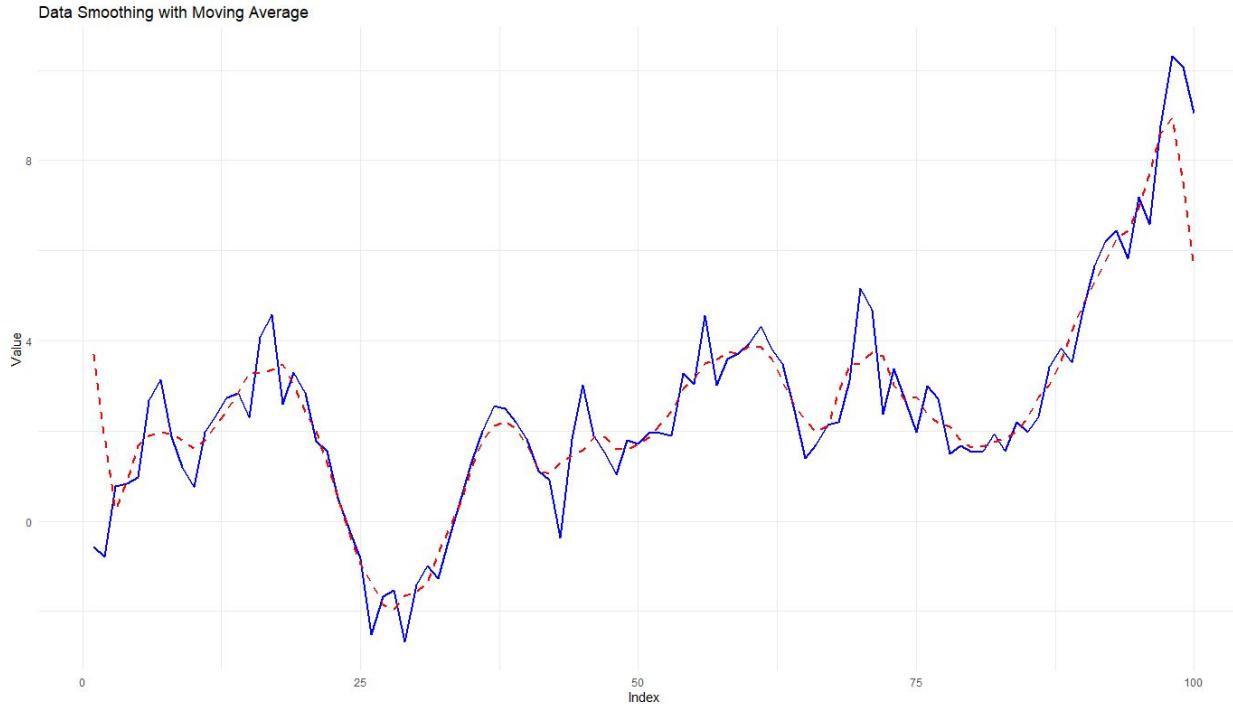
- Polynomial smoothing technique that fits a local polynomial to the data and uses convolution to smooth the signal.

## **Kernel Smoothing:**

- Involves placing a kernel function at each data point and averaging the values based on the kernel weight.

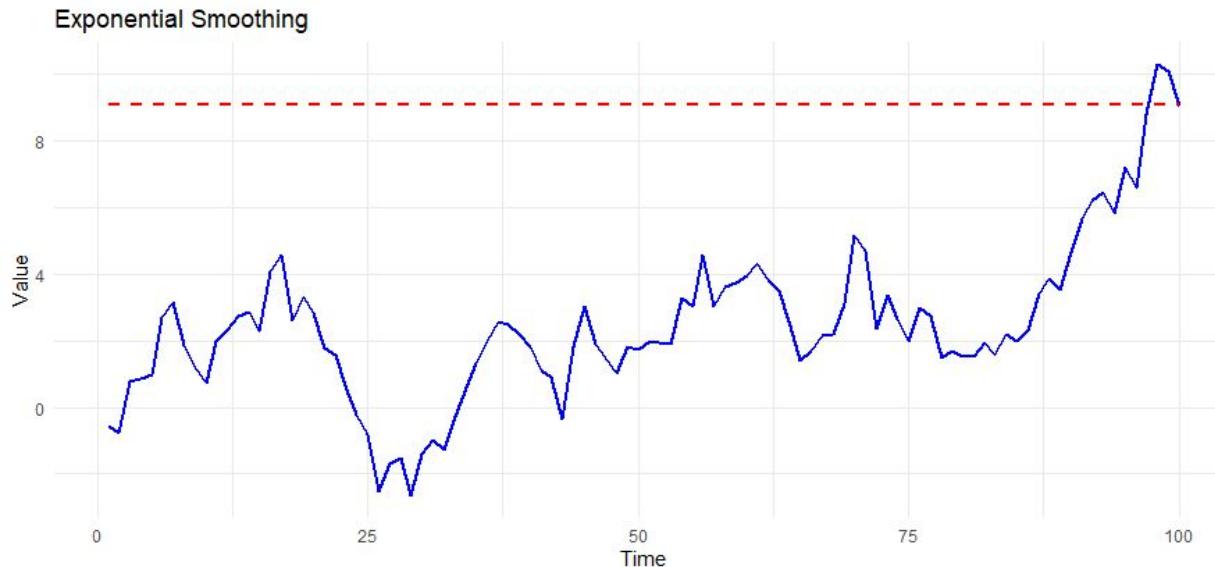
# Moving Average Methods

Simple technique that calculates the average of a subset of neighboring data points.



# Exponential Smoothing

Exponential smoothing is a time series forecasting and data smoothing technique that assigns exponentially decreasing weights to past observations. It is particularly useful for capturing trends and patterns in time-ordered data. The basic idea behind exponential smoothing is to give more importance to recent observations while gradually reducing the influence of older observations.



# Attribute Construction

Attribute construction, also known as feature construction or feature engineering, is a process in which new attributes (features) are created from the existing set of attributes in a dataset. This is done with the goal of enhancing the performance of a machine learning model or improving the interpretability of the data. Attribute construction involves creating new, meaningful features that may capture important relationships, patterns, or information in the data. It can be a crucial step in the data preprocessing phase.

# Attribute Construction

Attribute construction, also known as feature construction or feature engineering, is a process in data transformation where new attributes (features) are created from existing ones or external sources to improve the performance of a machine learning model. It involves creating meaningful, relevant, and informative features that can enhance the representation of the data and the predictive power of the model.

Here are some common techniques used in attribute construction:

## 1. Creating Interaction Terms:

- Combine two or more existing features to represent their interaction. For example, if you have features for "length" and "width," you can create a new feature for "area" by multiplying the two.

## 2. Polynomial Features:

- Introduce polynomial features to capture non-linear relationships. For example, if you have a feature "x," you can create a new feature " $x^2$ " or " $x^3$ " to allow the model to capture quadratic or cubic relationships.

## 3. Binning or Discretization:

- Convert continuous features into discrete bins or categories. This can help capture non-linear patterns and reduce the sensitivity to outliers.

#### **4. Encoding Categorical Variables:**

- Convert categorical variables into numerical representations, such as one-hot encoding or label encoding, so that they can be used in numerical models.

#### **5. Time-based Features:**

- Extract features related to time, such as day of the week, month, or year. These features can be valuable in time-series analysis.

#### **6. Aggregation and Grouping:**

- Create new features by aggregating or grouping existing ones. For instance, calculate the mean, sum, or count of a specific attribute within a certain group.

#### **7. Domain-specific Features:**

- Incorporate knowledge about the domain to create features that may be relevant for the specific problem you are trying to solve.

#### **8. Text Processing:**

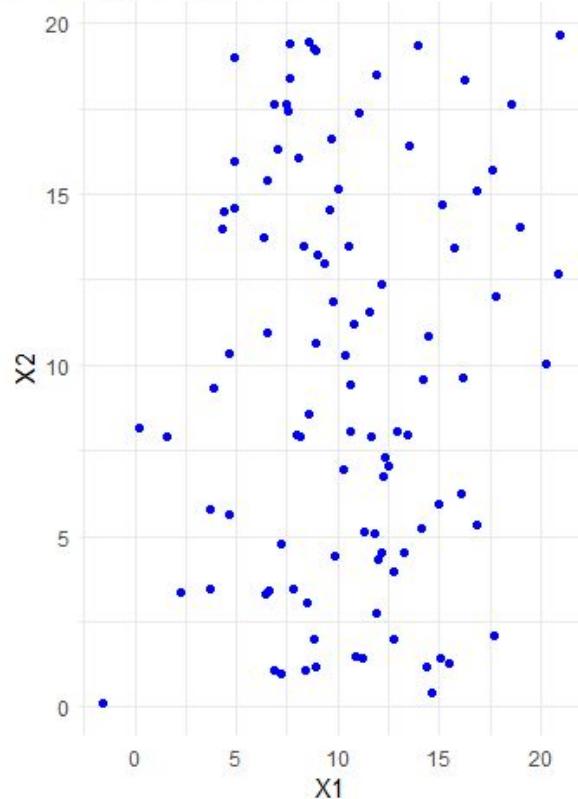
- For natural language processing tasks, convert text data into features, such as word frequencies, n-grams, or TF-IDF (Term Frequency-Inverse Document Frequency).

#### **9. Dimensionality Reduction:**

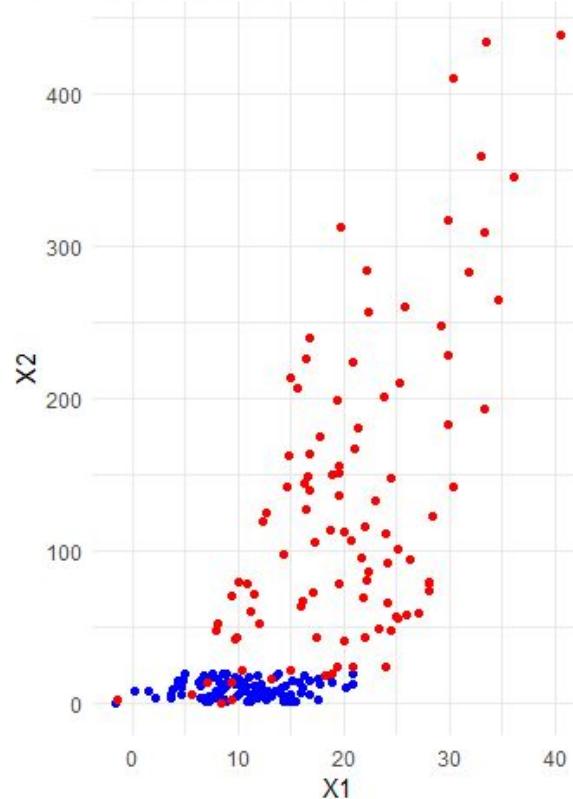
- Techniques like Principal Component Analysis (PCA) or t-Distributed Stochastic Neighbor Embedding (t-SNE) can be used to reduce the dimensionality of the data and create new features.

# Attribute Construction

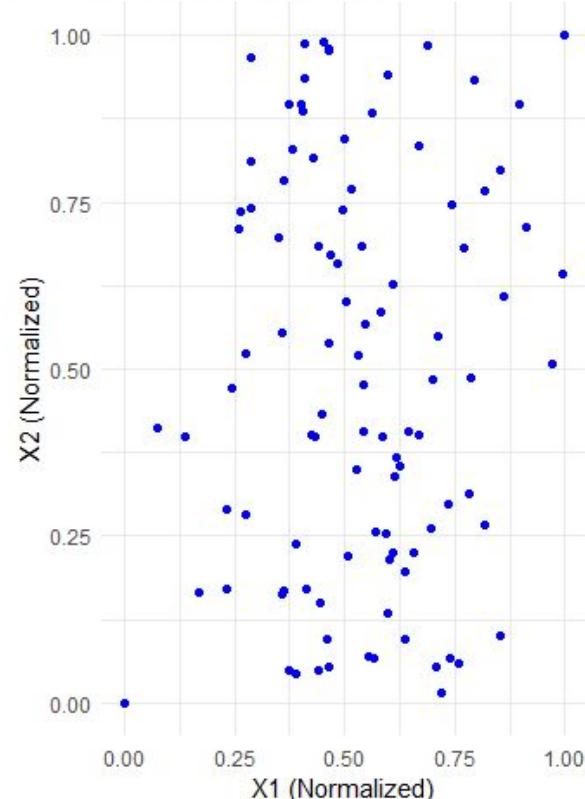
A Original Attributes



B Constructed Attributes



C Normalized Attributes



# Attribute Construction



<https://www.geeksforgeeks.org/feature-engineering-in-r-programming/>

# Data Generalization

- Data generalization is a process in data transformation that involves representing summarized or abstracted information rather than detailed, specific data.
- The goal of data generalization is to reduce the level of detail in the data while preserving its essential characteristics.
- This is often done to protect privacy, reduce noise, simplify analysis, or create higher-level views of the data.

# Data Discretization

- Data discretization is a data transformation technique that involves converting continuous data into discrete categories or bins.
- The main purpose of discretization is to simplify the data, reduce noise, and facilitate the analysis or modeling of the data.
- This process is particularly useful when dealing with continuous variables in machine learning or data mining, where algorithms might perform better with discrete inputs.

# Data Discretization

## Equal Width (or Equal Interval) Discretization:

In this approach, the range of values for a continuous variable is divided into a specified number of equal-width intervals or bins. Each bin represents a range of values, and data points falling within a specific range are assigned to the corresponding bin.

## Equal Frequency (or Equal Depth) Discretization:

In this approach, data points are grouped into bins such that each bin contains approximately the same number of data points. This method helps ensure that each bin captures a similar portion of the data distribution.

# Data Discretization

- **Data :** 0, 4, 12, 16, 16, 18, 24, 26, 28

- **Equal width**

- Bin 1: 0, 4 [ -, 10)
  - Bin 2: 12, 16, 16, 18 [ 10, 20)
  - Bin 3: 24, 26, 28 [ 20, +)

- **Equal frequency**

- Bin 1: 0, 4, 12 [ -, 14)
  - Bin 2: 16, 16, 18 [ 14, 21)
  - Bin 3: 24, 26, 28 [ 21, +)

- Data discretization is commonly used in scenarios where machine learning algorithms or statistical analyses require categorical or ordinal input variables.
- It can be applied to various types of data, such as age, income, temperature, and more.
- However, it's important to note that discretization introduces some level of information loss, and the choice of discretization method and the number of bins should be carefully considered based on the characteristics of the data and the goals of the analysis.

# Data Normalization

**Data normalization is a crucial step in data preprocessing for several reasons, and it is often necessary before applying various machine learning algorithms. Here are some key reasons why data normalization is important:**



<https://www.geeksforgeeks.org/how-to-normalize-data-in-r/>

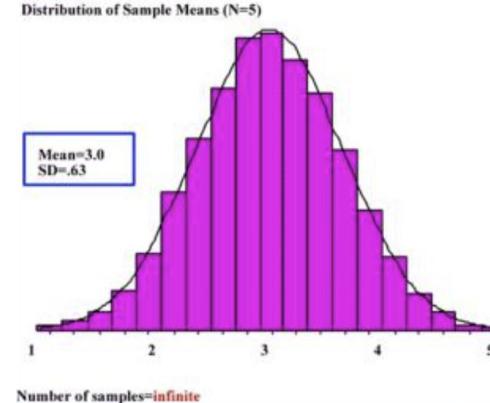
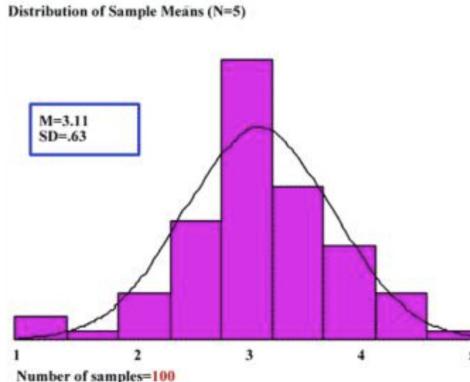
# Case Study of All Techniques



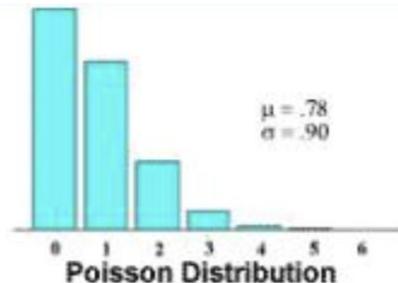
# Central Limit Theorem

# Central Limit Theorem

- The central limit theorem states that if you have a population with mean  $\mu$  and standard deviation  $\sigma$  and take sufficiently large random samples from the population with replacement then the distribution of the sample means will be approximately normally distributed.
- This will hold true regardless of whether the source population is normal or skewed, provided the sample size is sufficiently large (usually  $n > 30$ ).



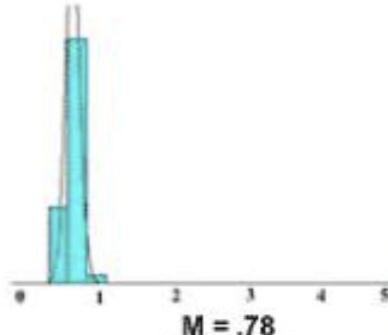
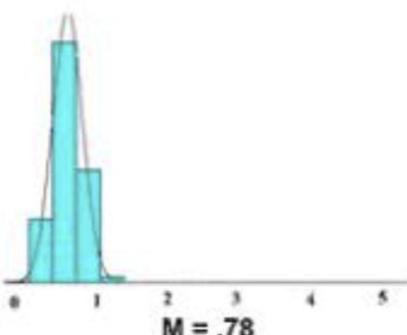
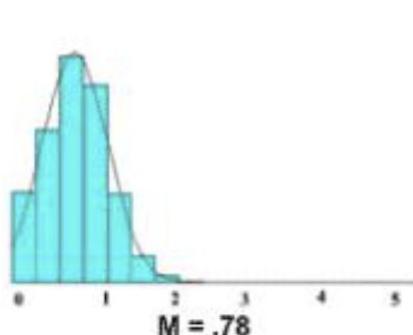
# Central Limit Theorem with a Poisson Distribution



N = 5

N = 25

N = 100



```
# Central Limit Theorem Simulation in R

# Set seed for reproducibility
set.seed(123)

# Number of samples to generate
num_samples <- 1000

# Number of observations in each sample
sample_size <- 30

# Original distribution (e.g., exponential distribution)
original_distribution <- rexp

# Create a matrix to store sample means
sample_means <- matrix(0, nrow = num_samples, ncol = sample_size)

# Generate samples and calculate sample means
for (i in 1:num_samples) {
  sample_means[i, ] <- original_distribution(sample_size)
}

# Calculate means of each sample
means <- rowMeans(sample_means)

# Plot the original distribution
hist(original_distribution(1000), main = "Original Distribution", col =
"lightblue", xlim = c(0, 5))

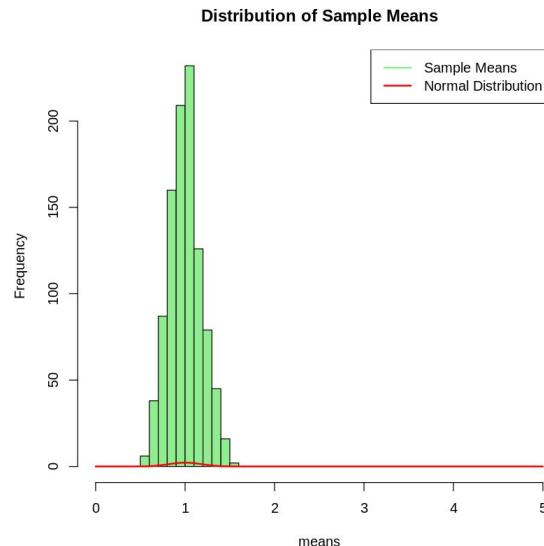
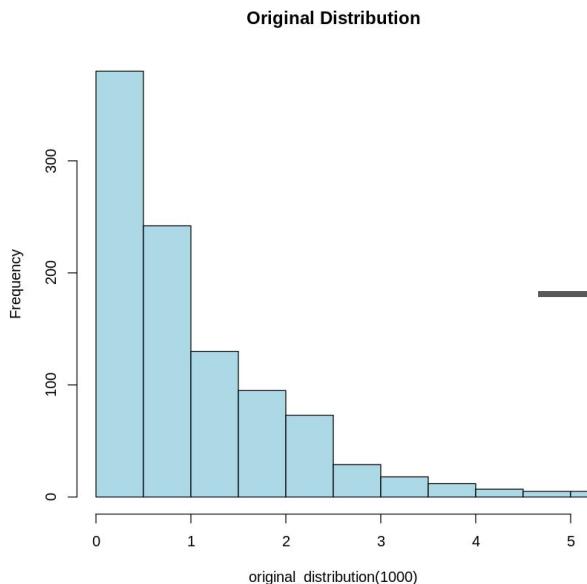
# Plot the distribution of sample means
hist(means, main = "Distribution of Sample Means", col =
"lightgreen", xlim = c(0, 5))

# Add a normal distribution curve for comparison
curve(dnorm(x, mean = mean(means), sd = sd(means)), col =
"red", lwd = 2, add = TRUE)

# Add legend
legend("topright", legend = c("Sample Means", "Normal
Distribution"), col = c("lightgreen", "red"), lwd = 2)

# Print mean and standard deviation of sample means
cat("Mean of sample means:", mean(means), "\n")
cat("Standard deviation of sample means:", sd(means), "\n")
```

The Central Limit Theorem (CLT) states that the distribution of the sum (or average) of a large number of independent, identically distributed random variables approaches a normal distribution, regardless of the original distribution of the variables.



# Probability

# Breast Cancer

- Cancer One theory concerning the etiology of breast cancer states that women in a given age group who give birth to their first child relatively late in life (after age 30) are at greater risk for eventually developing breast cancer over some time period  $t$  than are women who give birth to their first child early in life (before age 20).
- Because women in upper social classes tend to have children later, this theory has been used to explain why these women have a higher risk of developing breast cancer than women in lower social classes.

# Breast Cancer

- To test this hypothesis, we might identify 2000 postmenopausal women from a particular census tract who are currently ages 45–54 and have never had breast cancer, of whom 1000 had their first child before the age of 20 (call this group A) and 1000 after the age of 30 (group B).
- These 2000 women might be followed for 5 years to assess whether they developed breast cancer during this period.
- **Suppose there are four new cases of breast cancer in group A and five new cases in group B.**

*Is this evidence enough to confirm a difference in risk between the two groups?*

**Most people would feel uneasy about concluding that on the basis of such a limited amount of data.**

# Breast Cancer

- Suppose we had a more ambitious plan and sampled 10,000 postmenopausal women each from groups A and B and at follow-up found 40 new cases in group A and 50 new cases in group B and asked the same question.
- Although we might be more comfortable with the conclusion because of the larger sample size, we would still have to admit that this apparent difference in the rates could be due to chance

The problem is that we need a conceptual framework to make these decisions but have not explicitly stated what the framework is. This framework is provided by the underlying concept of **probability**

# Our Work Plan

- How to define probability and some rules for working with probabilities are introduced.
- Understanding probability is essential in calculating and interpreting p-values in the statistical tests .
- It also permits the discussion of sensitivity, specificity, and predictive values of screening tests

# **Definition of Probability**

The **sample space** is the set of all possible outcomes. In referring to probabilities of events, an **event** is any set of outcomes of interest. The **probability of an event** is the relative frequency of this set of outcomes over an indefinitely large (or infinite) number of trials.

**Pulmonary Disease** The tuberculin skin test is a routine screening test used to detect tuberculosis. The results of this test can be categorized as either positive, negative, or uncertain. If the probability of a positive test is .1, it means that if a large number of such tests were performed, about 10% would be positive. The actual percentage of positive tests will be increasingly close to .1 as the number of tests performed increases.

# Equation

- (1) The probability of an event  $E$ , denoted by  $Pr(E)$ , always satisfies  $0 \leq Pr(E) \leq 1$ .
- (2) If outcomes  $A$  and  $B$  are two events that cannot both happen at the same time, then  $Pr(A \text{ or } B \text{ occurs}) = Pr(A) + Pr(B)$ .

# Example

**Hypertension:** Let **A** be the event that a person has normotensive diastolic blood pressure (DBP) readings ( $DBP < 90$ ), and let **B** be the event that a person has borderline DBP readings ( $90 \leq DBP < 95$ ). Suppose that  $Pr(A) = .7$ , and  $Pr(B) = .1$ . Let **Z** be the event that a person has a  $DBP < 95$ . Then

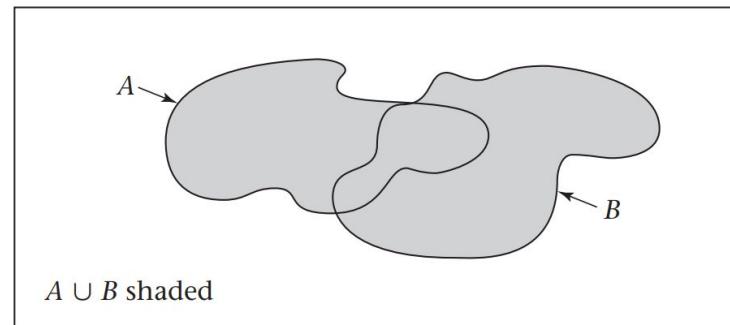
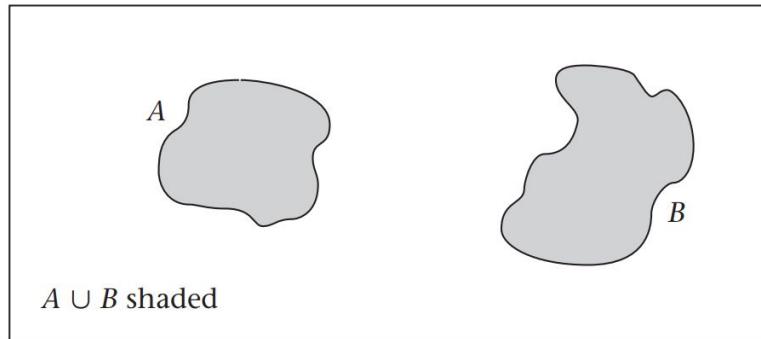
$$Pr(Z) = Pr(A) + Pr(B) = .8$$

**Because the events A and B cannot occur at the same time.**

**Two events A and B are mutually exclusive if they cannot both happen at the same time.**

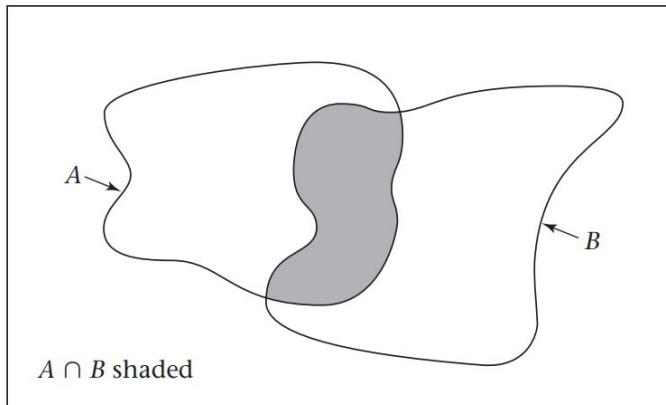
# Useful Probability Notions

- The symbol { } is used as shorthand for the phrase “the event.”
- $A \cup B$  is the event that either A or B occurs, or they both occur

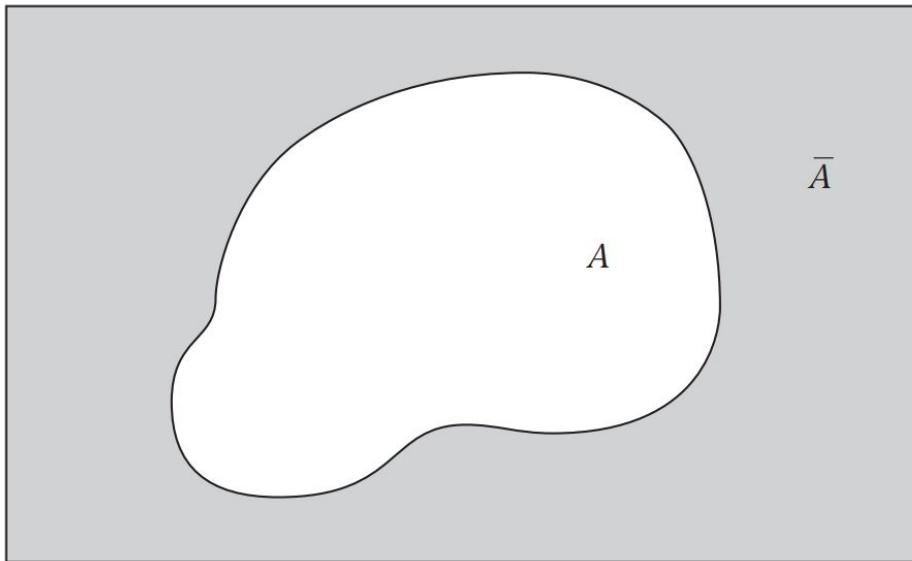


# Useful Probability Notions

- The symbol { } is used as shorthand for the phrase “the event.”
- $A \cap B$  is the event that both A and B occur simultaneously.



# Useful Probability Notions



$\bar{A}$  is the event that  $A$  does not occur. It is called the **complement** of  $A$ . Notice that  $Pr(\bar{A}) = 1 - Pr(A)$ , because  $\bar{A}$  occurs only when  $A$  does not occur. Event  $\bar{A}$  is diagrammed in Figure 3.3.

# Multiplication Law of Probability

**Hypertension, Genetics** Suppose we are conducting a hypertension-screening program in the home. Consider all possible pairs of DBP measurements of the mother and father within a given family, assuming that the mother and father are not genetically related. This sample space consists of all pairs of numbers of the form  $(X, Y)$  where  $X > 0, Y > 0$ . Certain specific events might be of interest in this context. In particular, we might be interested in whether the mother or father is hypertensive, which is described, respectively, by events  $A = \{\text{mother's DBP} \geq 90\}$ ,  $B = \{\text{father's DBP} \geq 90\}$ . These events are diagrammed in Figure 3.4.

Suppose we know that  $Pr(A) = .1$ ,  $Pr(B) = .2$ . What can we say about  $Pr(A \cap B) = Pr(\text{mother's DBP} \geq 90 \text{ and father's DBP} \geq 90) = Pr(\text{both mother and father are hypertensive})$ ? We can say nothing unless we are willing to make certain assumptions.

**Hypertension, Genetics** Compute the probability that both mother and father are hypertensive if the events in Example 3.12 are independent.

# In Class Assessment

**Sexually Transmitted Disease** Suppose two doctors, A and B, test all patients coming into a clinic for syphilis. Let events  $A^+$  = {doctor A makes a positive diagnosis} and  $B^+$  = {doctor B makes a positive diagnosis}. Suppose doctor A diagnoses 10% of all patients as positive, doctor B diagnoses 17% of all patients as positive, and both doctors diagnose 8% of all patients as positive. Are the events  $A^+$ ,  $B^+$  independent?

Two events  $A$ ,  $B$  are **dependent** if

$$Pr(A \cap B) \neq Pr(A) \times Pr(B)$$

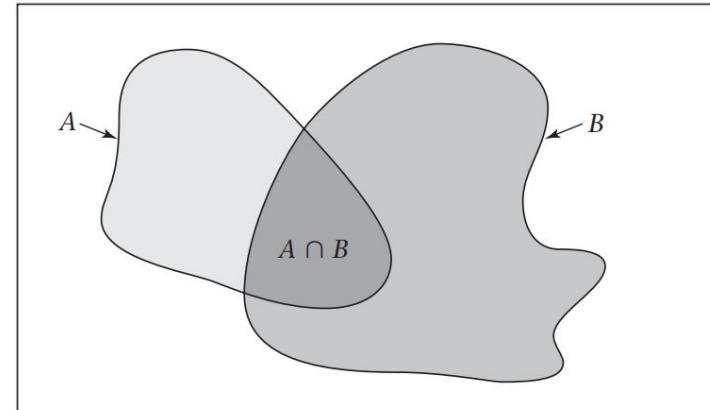
# Addition Law of Probability

We have seen from the definition of probability that if A and B are mutually exclusive events, then  $\Pr(A \cup B) = \Pr(A) + \Pr(B)$ . A more general formula for  $\Pr(A \cup B)$  can be developed when events A and B are not necessarily mutually exclusive.

## Addition Law of Probability

If A and B are any events,

$$\text{then } \Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$$



Thus, to compute  $\Pr(A \cup B)$ , add the probabilities of A and B separately and then subtract the overlap, which is  $\Pr(A \cap B)$ .

# In Class Assessment

$$Pr(A^+) = .1 \quad Pr(B^+) = .17 \quad Pr(A^+ \cap B^+) = .08$$

**Sexually Transmitted Disease** Consider the data in Example 3.15. Suppose a patient is referred for further lab tests if either doctor A or B makes a positive diagnosis. What is the probability that a patient will be referred for further lab tests?

# Conditional Probability

# Conditional Probability

Suppose we want to compute the probability of several events occurring simultaneously. If the events are independent, then we can use the multiplication law of probability to do so. If some of the events are dependent, then a quantitative measure of dependence is needed to extend the multiplication law to the case of dependent events.

# Conditional Probability

**Cancer** Physicians recommend that all women over age 50 be screened for breast cancer. The definitive test for identifying breast tumors is a breast biopsy. However, this procedure is too expensive and invasive to recommend for *all* women over the age of 50. Instead, women in this age group are encouraged to have a mammogram every 1 to 2 years. Women with positive mammograms are then tested further with a biopsy. Ideally, the probability of breast cancer among women who are mammogram positive would be 1 and the probability of breast cancer among women who are mammogram negative would be 0. The two events {mammogram positive} and {breast cancer} would then be completely dependent; the results of the screening test would automatically determine the disease state. The opposite extreme is achieved when the events {mammogram positive} and {breast cancer} are completely independent. In this case, the probability of breast cancer would be the same regardless of whether the mammogram is positive or negative, and the mammogram would not be useful in screening for breast cancer and should not be used.

# Conditional Probability

These concepts can be quantified in the following way. Let  $A = \{\text{mammogram}^+\}$ ,  $B = \{\text{breast cancer}\}$ , and suppose we are interested in the probability of breast cancer ( $B$ ) given that the mammogram is positive ( $A$ ). This probability can be written  $\Pr(A \cap B)/\Pr(A)$ .

---

The quantity  $\Pr(A \cap B)/\Pr(A)$  is defined as the **conditional probability of  $B$  given  $A$** , which is written  $\Pr(B|A)$ .

---

# Conditional Probability

These concepts can be quantified in the following way. Let  $A = \{\text{mammogram}^+\}$ ,  $B = \{\text{breast cancer}\}$ , and suppose we are interested in the probability of breast cancer ( $B$ ) given that the mammogram is positive ( $A$ ). This probability can be written  $\Pr(A \cap B)/\Pr(A)$ .

---

The quantity  $\Pr(A \cap B)/\Pr(A)$  is defined as the **conditional probability of  $B$  given  $A$** , which is written  $\Pr(B|A)$ .

---

- (1) If  $A$  and  $B$  are independent events, then  $\Pr(B|A) = \Pr(B) = \Pr(B|\bar{A})$ .
- (2) If two events  $A$ ,  $B$  are dependent, then  $\Pr(B|A) \neq \Pr(B) \neq \Pr(B|\bar{A})$  and  $\Pr(A \cap B) \neq \Pr(A) \times \Pr(B)$ .

# Bayes Rule and Screening Tests

# Bayes Rule and Screening Tests

**Cancer** Physicians recommend that all women over age 50 be screened for breast cancer. The definitive test for identifying breast tumors is a breast biopsy. However, this procedure is too expensive and invasive to recommend for *all* women over the age of 50. Instead, women in this age group are encouraged to have a mammogram every 1 to 2 years. Women with positive mammograms are then tested further with a biopsy. Ideally, the probability of breast cancer among women who are mammogram positive would be 1 and the probability of breast cancer among women who are mammogram negative would be 0. The two events {mammogram positive} and {breast cancer} would then be completely dependent; the results of the screening test would automatically determine the disease state. The opposite extreme is achieved when the events {mammogram positive} and {breast cancer} are completely independent. In this case, the probability of breast cancer would be the same regardless of whether the mammogram is positive or negative, and the mammogram would not be useful in screening for breast cancer and should not be used.

These concepts can be quantified in the following way. Let  $A = \{\text{mammogram}^+\}$ ,  $B = \{\text{breast cancer}\}$ , and suppose we are interested in the probability of breast cancer ( $B$ ) given that the mammogram is positive ( $A$ ). This probability can be written  $\Pr(A \cap B)/\Pr(A)$ .

# Bayes Rule and Screening Tests

If the risk of developing health problems is known to increase with age, Bayes' theorem allows the risk to an individual of a known age to be assessed more accurately (by conditioning it on their age) than simply assuming that the individual is typical of the population as a whole.

# Bayes Rule and Screening Tests

---

The **predictive value positive (PV<sup>+</sup>)** of a screening test is the probability that a person has a disease given that the test is positive.

$$Pr(\text{disease} \mid \text{test}^+)$$

The **predictive value negative (PV<sup>-</sup>)** of a screening test is the probability that a person does *not* have a disease given that the test is negative.

$$Pr(\text{no disease} \mid \text{test}^-)$$

---

**Cancer** Suppose that among 100,000 women with negative mammograms 20 will be diagnosed with breast cancer within 2 years, or  $Pr(B|\bar{A}) = 20 / 10^5 = .0002$ , whereas 1 woman in 10 with positive mammograms will be diagnosed with breast cancer within 2 years, or  $Pr(B|A) = .1$ . The two events  $A$  and  $B$  would be highly dependent, because

**Cancer** Find  $PV^+$  and  $PV^-$  for mammography given the data in Example 3.19.

**Solution:** We see that  $PV^+ = Pr(\text{breast cancer} \mid \text{mammogram}^+) = .1$

whereas  $PV^- = Pr(\text{breast cancer}^- \mid \text{mammogram}^-)$

$$= 1 - Pr(\text{breast cancer} \mid \text{mammogram}^-) = 1 - .0002 = .9998$$

Thus, if the mammogram is negative, the woman is virtually certain *not* to develop breast cancer over the next 2 years ( $PV^- \approx 1$ ); whereas if the mammogram is positive, the woman has a 10% chance of developing breast cancer ( $PV^+ = .10$ ).

A symptom or a set of symptoms can also be regarded as a screening test for disease. The higher the  $PV$  of the screening test or symptoms, the more valuable the test will be. Ideally, we would like to find a set of symptoms such that both  $PV^+$  and  $PV^-$  are 1. Then we could accurately diagnose disease for each patient. However, this is usually impossible.

Clinicians often cannot directly measure the  $PV$  of a set of symptoms. However, they can measure how often specific symptoms occur in diseased and normal people. These measures are defined as follows:

## These measures are defined as follows:

---

The **sensitivity** of a symptom (or set of symptoms or screening test) is the probability that the symptom is present given that the person has a disease.

---

The **specificity** of a symptom (or set of symptoms or screening test) is the probability that the symptom is *not* present given that the person does *not* have a disease.

---

A **false negative** is defined as a negative test result when the disease or condition being tested for is actually present. A **false positive** is defined as a positive test result when the disease or condition being tested for is not actually present.

---

**Cancer** Suppose the disease is lung cancer and the symptom is cigarette smoking. If we assume that 90% of people with lung cancer and 30% of people without lung cancer (essentially the entire general population) are smokers, then the sensitivity and specificity of smoking as a screening test for lung cancer are .9 and .7, respectively. Obviously, cigarette smoking cannot be used by itself as a screening criterion for predicting lung cancer because there will be too many false positives (people without cancer who are smokers).

**Cancer** Suppose the disease is breast cancer in women and the symptom is having a family history of breast cancer (either a mother or a sister with breast cancer). If we assume 5% of women with breast cancer have a family history of breast cancer but only 2% of women without breast cancer have such a history, then the sensitivity of a family history of breast cancer as a predictor of breast cancer is .05 and the specificity is  $.98 = (1 - .02)$ . A family history of breast cancer cannot be used by itself to diagnose breast cancer because there will be too many false negatives (women with breast cancer who do not have a family history).

# Bayes' Rule

The level of prostate-specific antigen (PSA) in the blood is frequently used as a screening test for prostate cancer. Punglia et al. [5] reported the following data regarding the relationship between a positive PSA test ( $\geq 4.1$  ng/dL) and prostate cancer.

**Association between PSA and prostate cancer**

PSA test result	Prostate cancer	Frequency
+	+	92
+	-	27
-	+	46
-	-	72

- What are the sensitivity and specificity of the test?
- What are the  $PV^+$  and  $PV^-$  of the test?

## Bayes' Rule

Review Question 3C.3 assumes that each PSA<sup>+</sup> and PSA<sup>-</sup> participant (or at least a representative sample of PSA<sup>+</sup> and PSA<sup>-</sup> participants) is evaluated for the presence of prostate cancer. Thus, one can directly evaluate  $PV^+$  and  $PV^-$  from the data provided. Instead, in many screening studies, a random sample of cases and controls is obtained. One can estimate sensitivity and specificity from such a design. However, because cases are usually oversampled relative to the general population (e.g., if there are an equal number of cases and controls), one cannot directly estimate  $PV^+$  and  $PV^-$  from the frequency counts available in a typical screening study. Instead, an indirect method known as Bayes' rule is used for this purpose.

The general question then becomes how can the sensitivity and specificity of a symptom (or set of symptoms or diagnostic test), which are quantities a physician can estimate, be used to compute  $PVs$ , which are quantities a physician needs to make appropriate diagnoses?

Let  $A$  = symptom and  $B$  = disease. From Definitions 3.12, 3.13, and 3.14, we have

$$\text{Predictive value positive} = PV^+ = \Pr(B|A)$$

$$\text{Predictive value negative} = PV^- = \Pr(\bar{B}|\bar{A})$$

$$\text{Sensitivity} = \Pr(A|B)$$

$$\text{Specificity} = \Pr(\bar{A}|\bar{B})$$

Question on  
previous slide

Let  $\Pr(B)$  = probability of disease in the reference population. We wish to compute  $\Pr(B|A)$  and  $\Pr(\bar{B}|\bar{A})$  in terms of the other quantities. This relationship is known as Bayes' rule.

## Bayes' Rule

Let  $A$  = symptom and  $B$  = disease.

$$PV^+ = Pr(B|A) = \frac{Pr(A|B) \times Pr(B)}{Pr(A|B) \times Pr(B) + Pr(A|\bar{B}) \times Pr(\bar{B})}$$

In words, this can be written as

$$PV^+ = \frac{\text{Sensitivity} \times x}{\text{Sensitivity} \times x + (1 - \text{Specificity}) \times (1 - x)}$$

where  $x = Pr(B)$  = prevalence of disease in the reference population. Similarly,

$$PV^- = \frac{\text{Specificity} \times (1 - x)}{\text{Specificity} \times (1 - x) + (1 - \text{Sensitivity}) \times x}$$

# In Class Assessment!

**Hypertension** Suppose 84% of hypertensives and 23% of normotensives are classified as hypertensive by an automated blood-pressure machine. What are the  $PV^+$  and  $PV^-$  of the machine, assuming 20% of the adult population is hypertensive?

# Attention!

Example 3.26 considered only two possible disease states: hypertensive and normotensive. In clinical medicine there are often more than two possible disease states. We would like to be able to predict the most likely disease state given a specific symptom (or set of symptoms). Let's assume that the probability of having these symptoms among people in each disease state (where one of the disease states may be normal) is known from clinical experience, as is the probability of each disease state in the reference population. This leads us to the generalized Bayes' rule:

## Generalized Bayes' Rule

Let  $B_1, B_2, \dots, B_k$  be a set of mutually exclusive and exhaustive disease states; that is, at least one disease state must occur and no two disease states can occur at the same time. Let  $A$  represent the presence of a symptom or set of symptoms. Then,

$$Pr(B_i|A) = Pr(A|B_i) \times Pr(B_i) \Bigg/ \left[ \sum_{j=1}^k Pr(A|B_j) \times Pr(B_j) \right]$$

Question on  
previous slide

**Pulmonary Disease** Suppose a 60-year-old man who has never smoked cigarettes presents to a physician with symptoms of a chronic cough and occasional breathlessness. The physician becomes concerned and orders the patient admitted to the hospital for a lung biopsy. Suppose the results of the lung biopsy are consistent either with lung cancer or with sarcoidosis, a fairly common, usually nonfatal lung disease. In this case

$$A = \{\text{chronic cough, results of lung biopsy}\}$$

Disease state

$$\begin{cases} B_1 = \text{normal} \\ B_2 = \text{lung cancer} \\ B_3 = \text{sarcoidosis} \end{cases}$$

Suppose that  $Pr(A|B_1) = .001$   $Pr(A|B_2) = .9$   $Pr(A|B_3) = .9$

and that in 60-year-old, never-smoking men

$$Pr(B_1) = .99 \quad Pr(B_2) = .001 \quad Pr(B_3) = .009$$

The first set of probabilities  $Pr(A|B_i)$  could be obtained from clinical experience with the previous diseases, whereas the latter set of probabilities  $Pr(B_i)$  would have to be obtained from age-, gender-, and smoking-specific prevalence rates for the diseases in question. The interesting question now becomes what are the probabilities  $Pr(B_i|A)$  of the three disease states given the previous symptoms?

**Solution:** Bayes' rule can be used to answer this question. Specifically,

$$\begin{aligned}Pr(B_1|A) &= Pr(A|B_1) \times Pr(B_1) / \left[ \sum_{j=1}^3 Pr(A|B_j) \times Pr(B_j) \right] \\&= .001(.99) / [.001(.99) + .9(.001) + .9(.009)] \\&= .00099 / .00999 = .099\end{aligned}$$

$$\begin{aligned}Pr(B_2|A) &= .9(.001) / [.001(.99) + .9(.001) + .9(.009)] \\&= .00090 / .00999 = .090\end{aligned}$$

$$\begin{aligned}Pr(B_3|A) &= .9(.009) / [.001(.99) + .9(.001) + .9(.009)] \\&= .00810 / .00999 = .811\end{aligned}$$

Thus, although the unconditional probability of sarcoidosis is very low (.009), the conditional probability of the disease given these symptoms and this age-gender-smoking group is .811. Also, although the symptoms and diagnostic tests are consistent with both lung cancer and sarcoidosis, the latter is much more likely among patients in this age-gender-smoking group (i.e., among never-smoking men).

# In Class Assessment!

**Pulmonary Disease** Now suppose the patient in Example 3.27 smoked two packs of cigarettes per day for 40 years. Then assume  $Pr(B_1) = .98$ ,  $Pr(B_2) = .015$ , and  $Pr(B_3) = .005$  in this type of person. What are the probabilities of the three disease states for this type of patient, given these symptoms?

# Probability Distribution Function

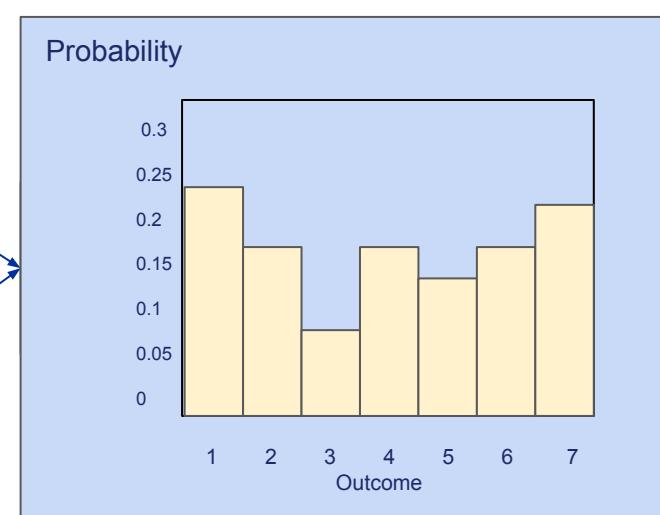
# Probability Distribution Function

Discrete

Probability mass  
function  
(PMF)

Continuous

Probability  
density function  
(PDF)



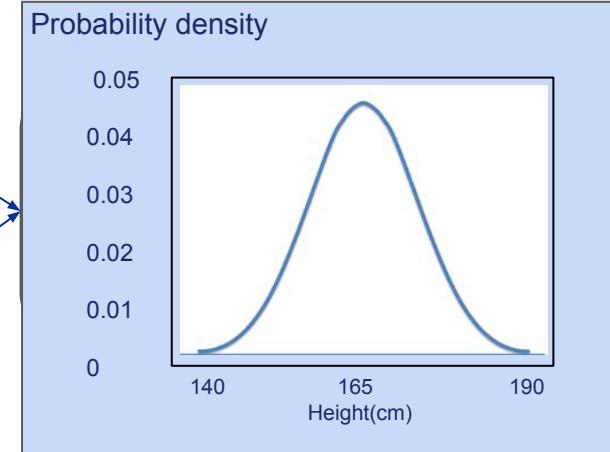
# Probability Distribution Function

Discrete

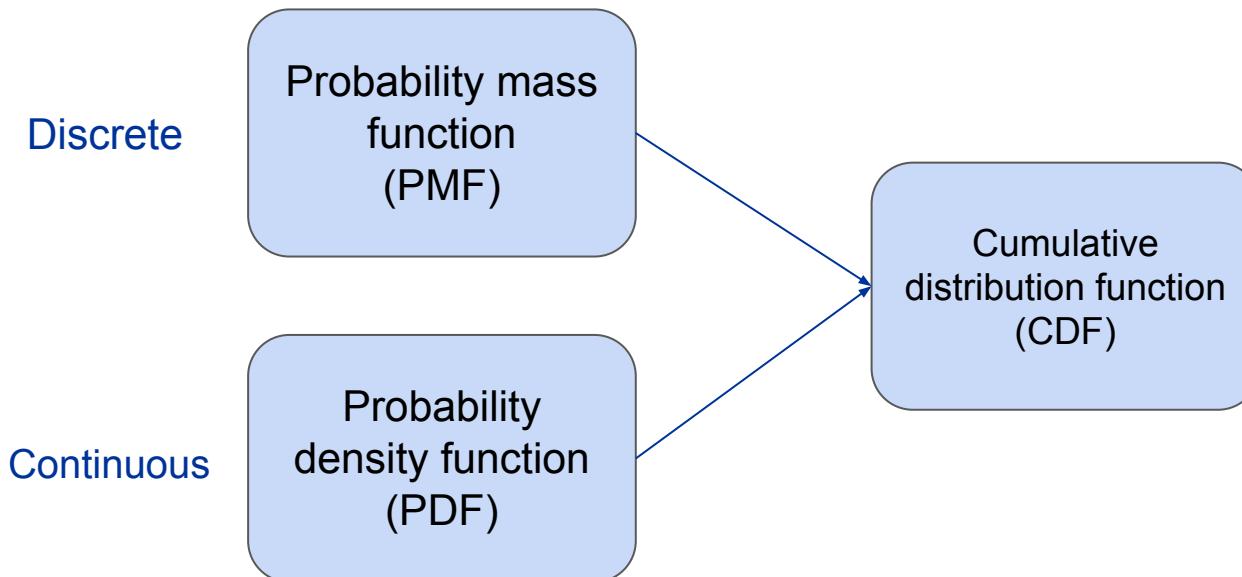
Probability mass  
function  
(PMF)

Continuous

Probability  
density function  
(PDF)

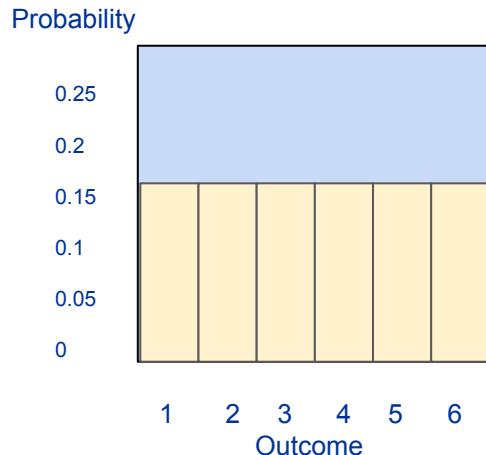


# Probability Distribution Function (PDFs ??)



**Discrete**

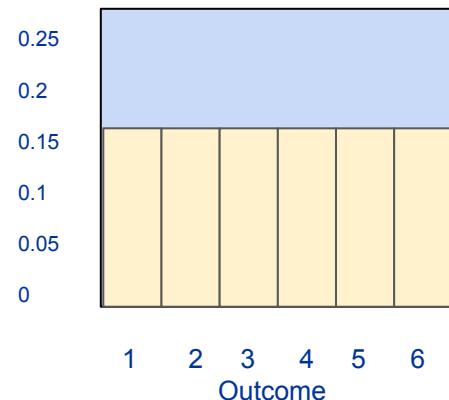
PMF



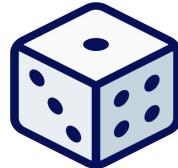
$$\frac{1}{6} = 0.167 \\ (\text{approx})$$

**Discrete**

Probability

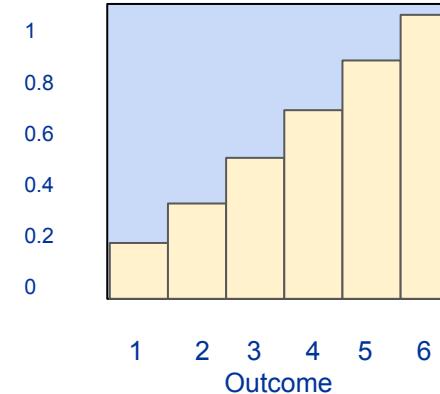


PMF



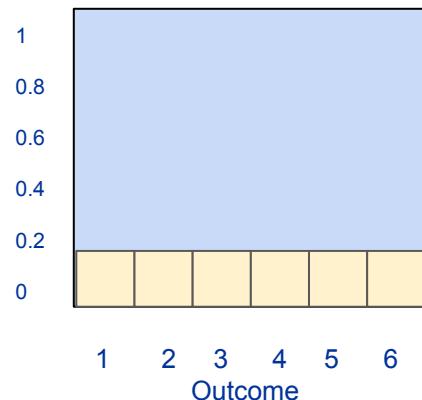
CDF

Cumulative Probability



**Discrete**

Probability

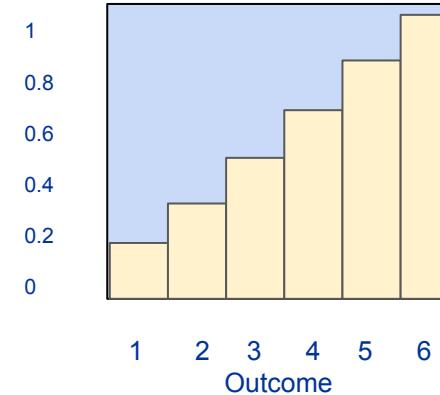


PMF



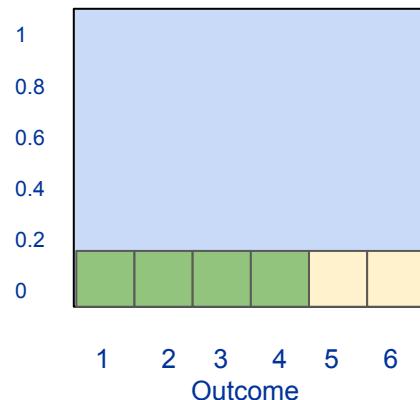
CDF

Cumulative Probability



**Discrete**

Probability

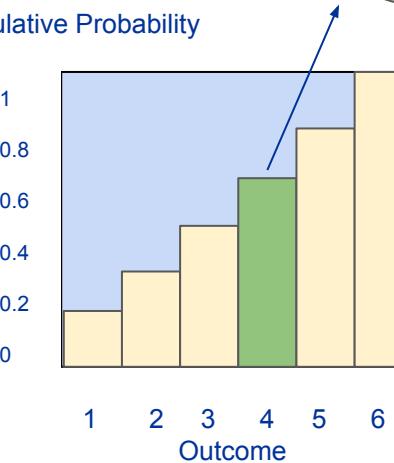


PMF



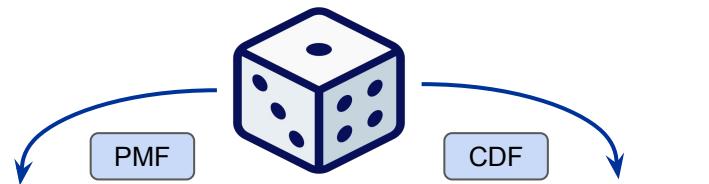
CDF

Cumulative Probability

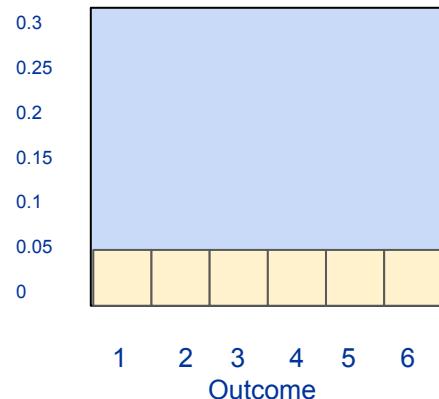


$$\begin{aligned}P(X \leq 4) &= \\&= P(X = 1) + P(X = 2) \\&\quad + P(X = 3) + P(X = 4)\end{aligned}$$

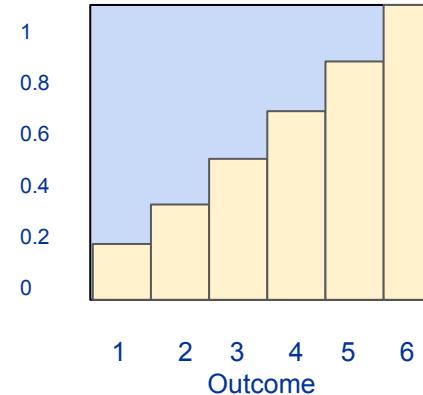
**Discrete**



Probability



Cumulative Probability



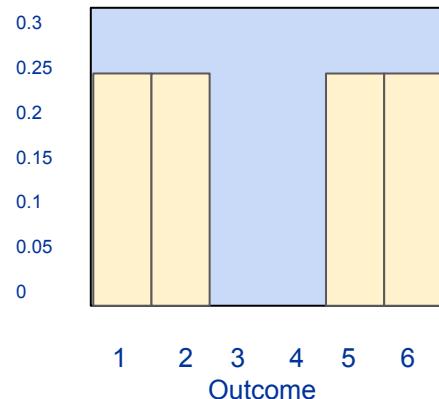
**Discrete**



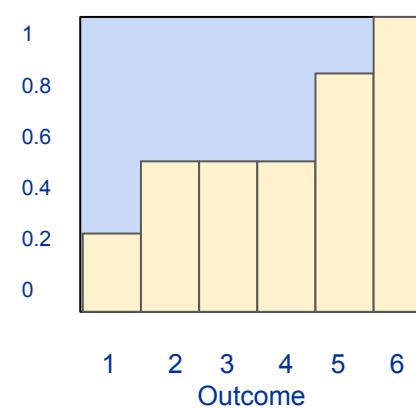
PMF

CDF

Probability

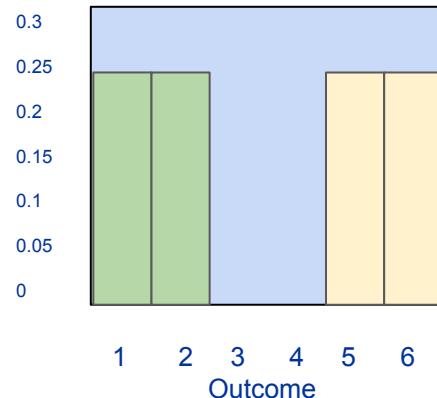


Cumulative Probability



**Discrete**

Probability

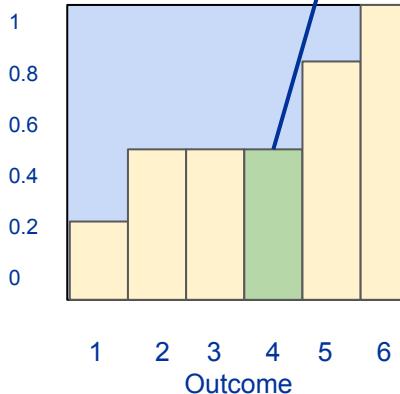


PMF



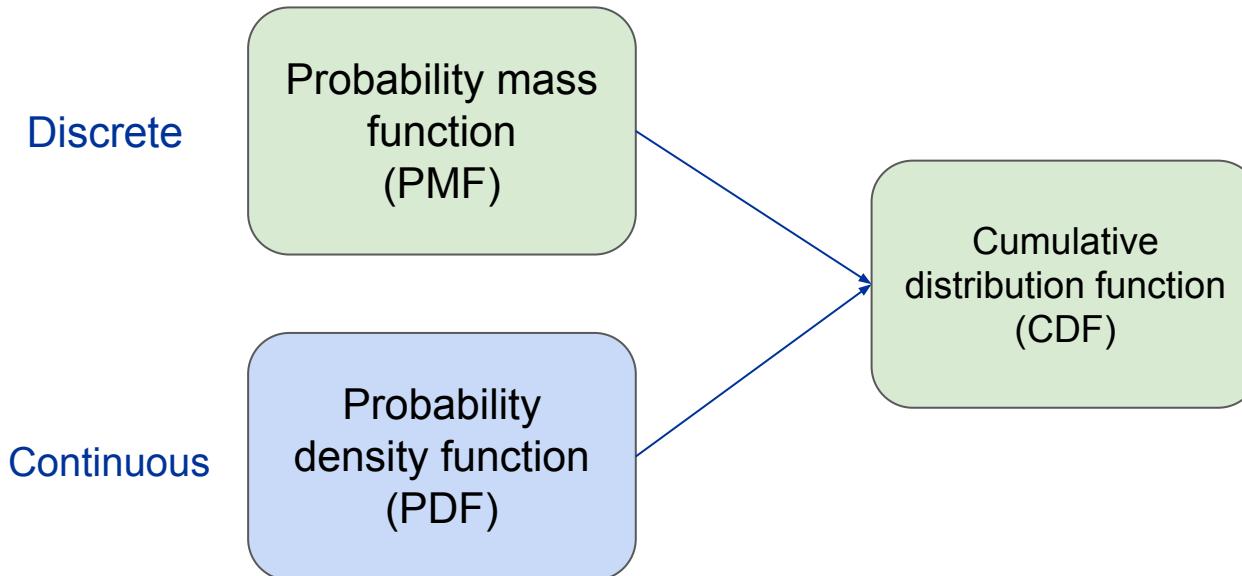
CDF

Cumulative Probability



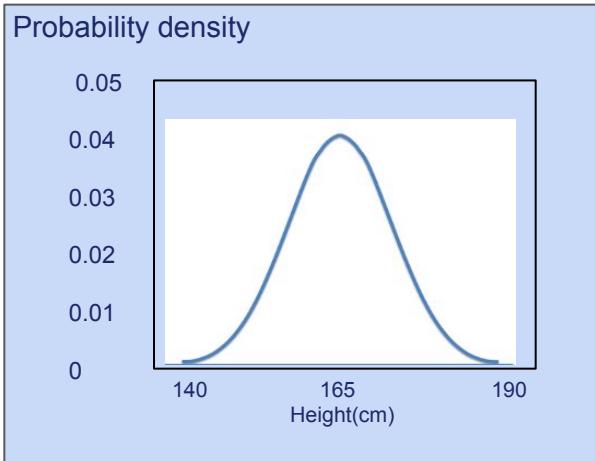
$$\begin{aligned}P(X \leq 4) &= \\&= P(X = 1) + P(X = 2) \\&\quad + P(X = 3) + P(X = 4)\end{aligned}$$

# Probability Distribution Function



**Continuous**

PDF

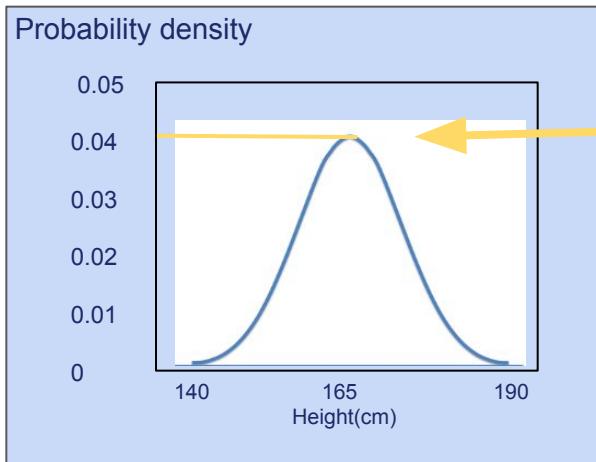


**165.387 cm**

**165.387684.....**

Continuous!

PDF

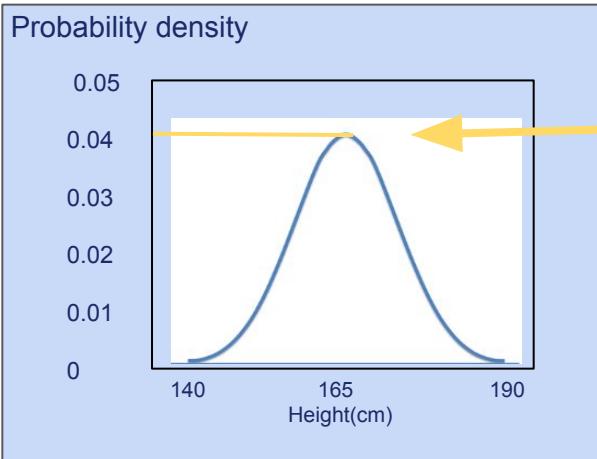


Probability  
density = 0.04  
????

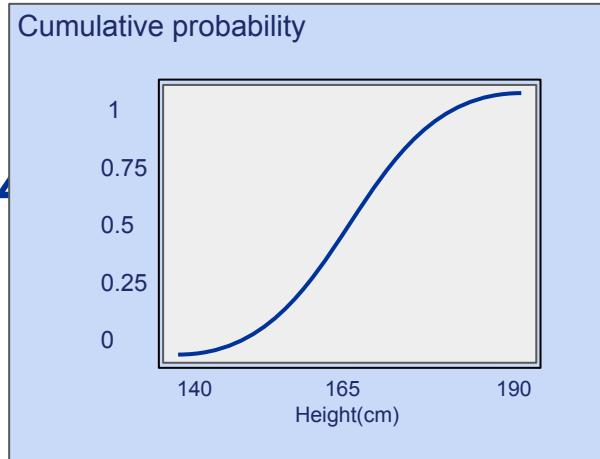
Continuous!

PDF

CDF



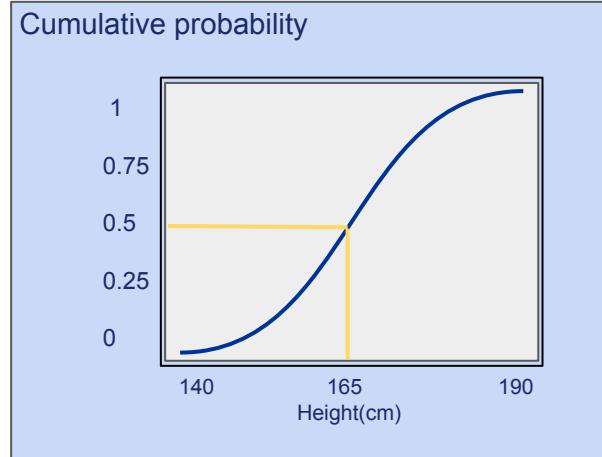
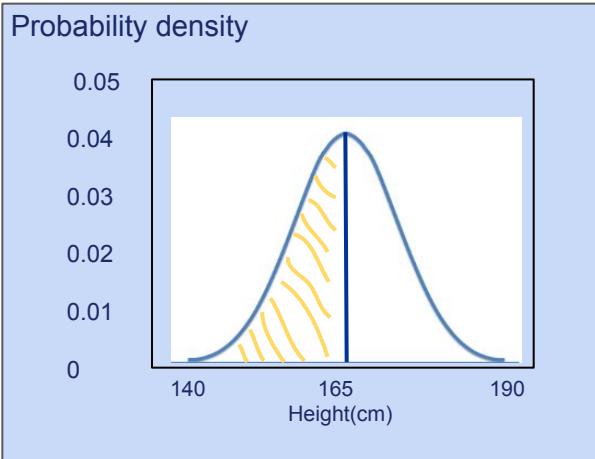
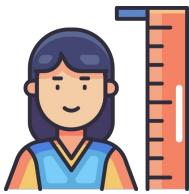
Probability  
density = 0.04  
????



**Continuous!**

PDF

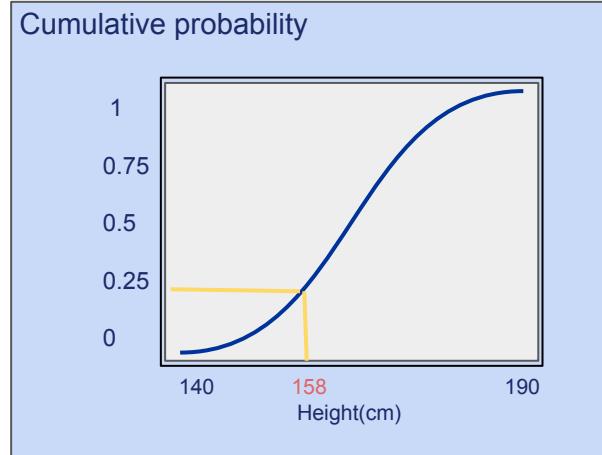
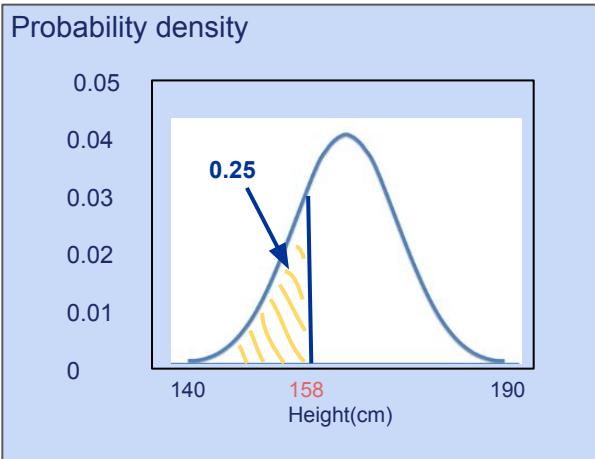
CDF

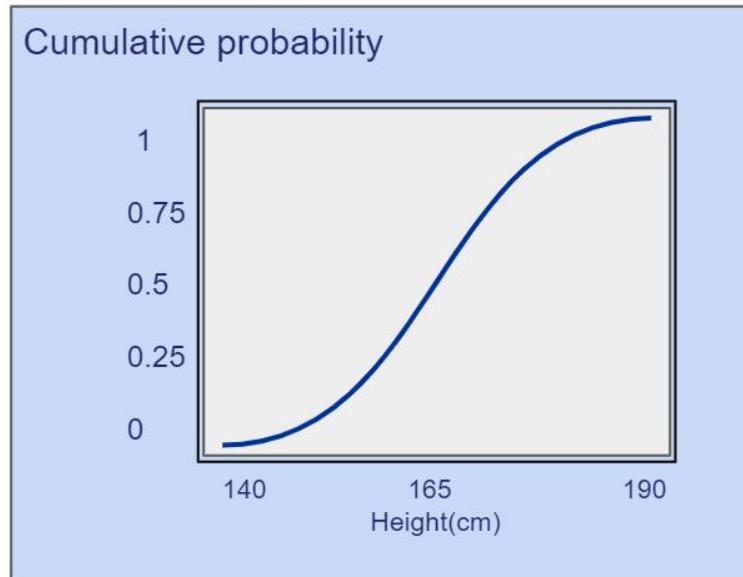


**Continuous!**

PDF

CDF





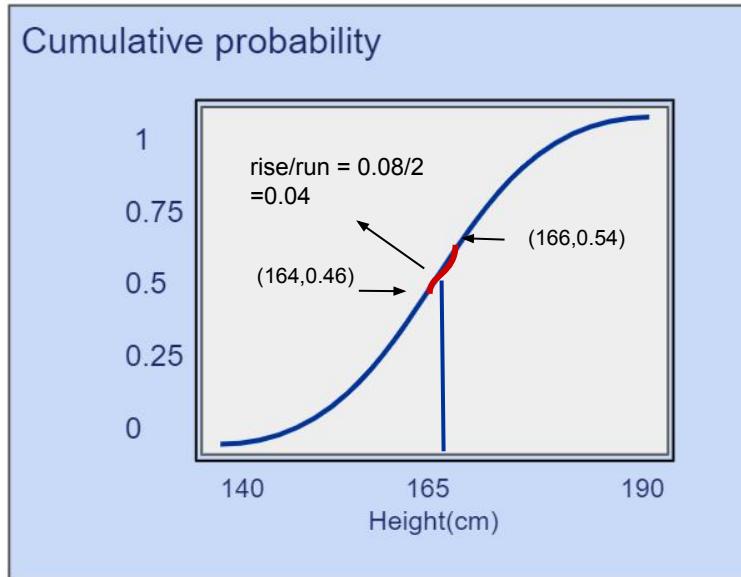
Higher  
gradient

=

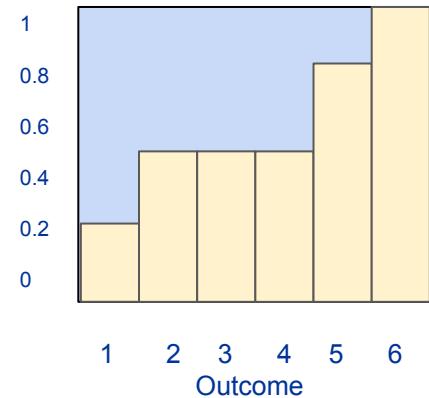
Higher  
density



CDF

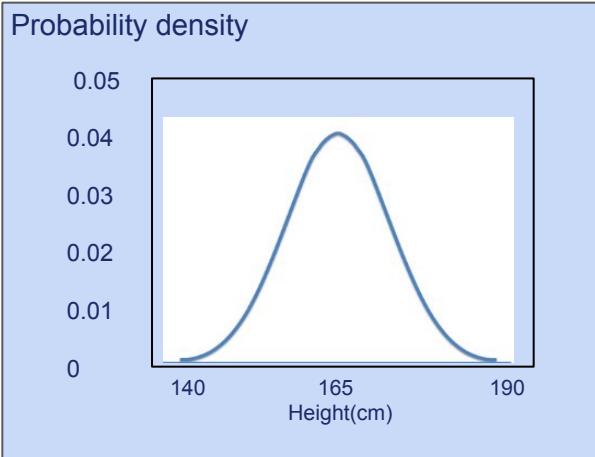


Cumulative Probability

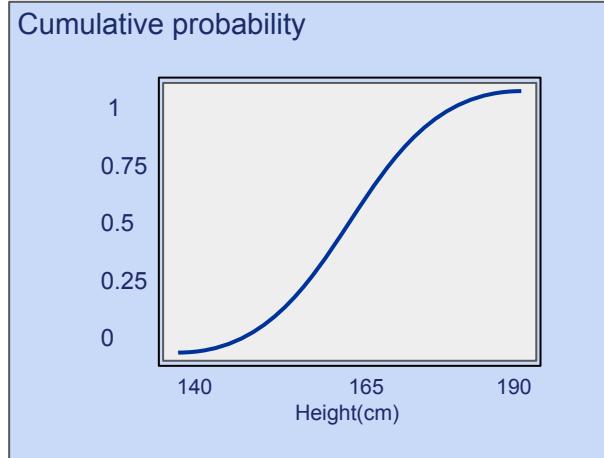


$f(x)$

$F(x)$



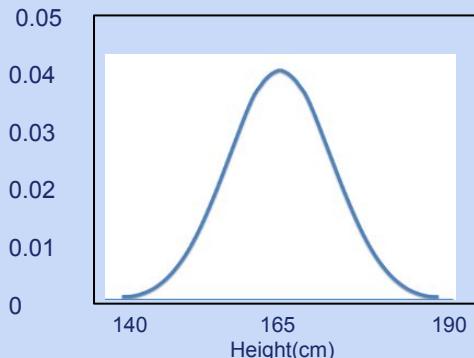
Gradient  
Area to the left



$f(x)$

$F(x)$

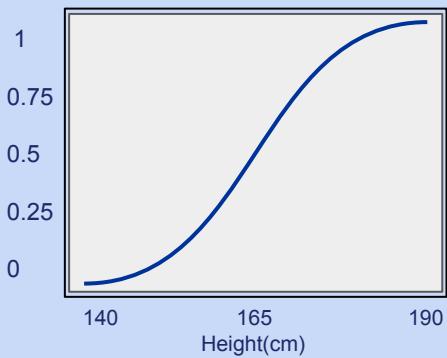
Probability density



$$dF(x)/dx = f(x)$$

$$\int_{-\infty}^x f(x)dx = F(x)$$

Cumulative probability



# ROC Curve

# ROC Curve

In some instances, a test provides several categories of response rather than simply providing positive or negative results. In other instances, the results of the test may be reported as a continuous variable. In either case, designation of a cutoff point for distinguishing a test result as positive versus negative is arbitrary.

---

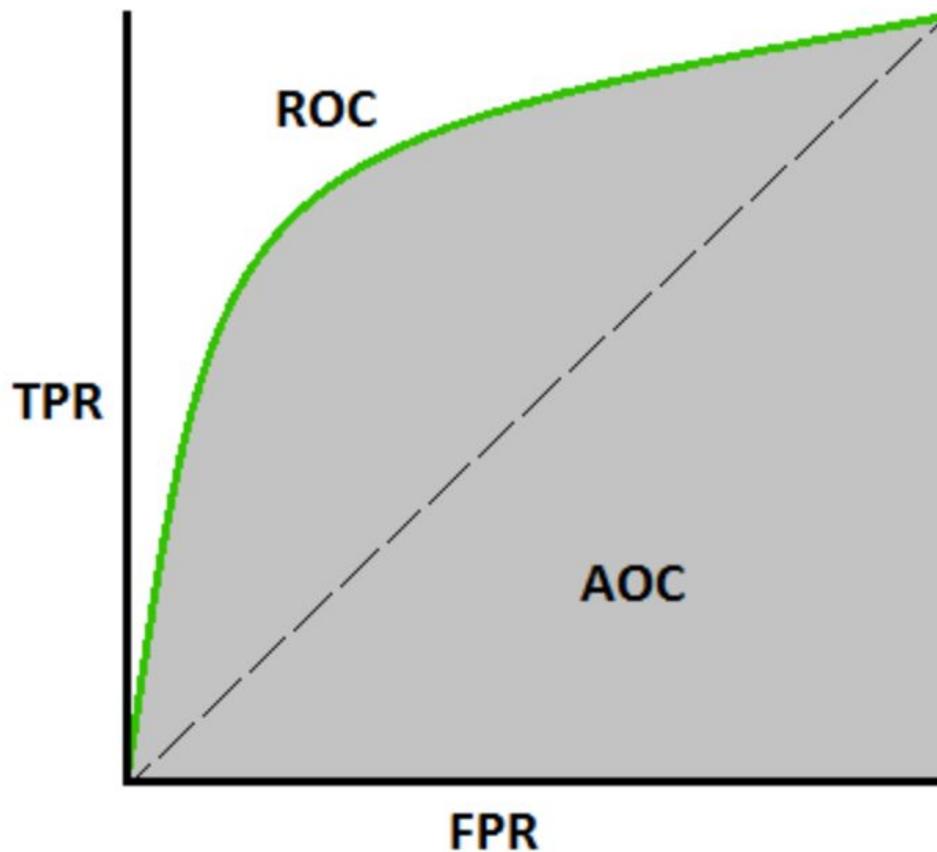
A **receiver operating characteristic (ROC) curve** is a plot of the sensitivity (on the y-axis) versus (1 – specificity) (on the x-axis) of a screening test, where the different points on the curve correspond to different cutoff points used to designate test-positive.

---

# What is the AUC - ROC Curve?

- AUC - ROC curve is a performance measurement for the classification problems at various threshold settings.
- ROC is a probability curve and AUC represents the degree or measure of separability.
- It tells how much the model is capable of distinguishing between classes.
- Higher the AUC, the better the model is at predicting 0s as 0s and 1s as 1s.
- By analogy, the Higher the AUC, the better the model is at distinguishing between patients with the disease and no disease.

# ROC Curve



# ROC Curve

$$\text{TPR / Recall / Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

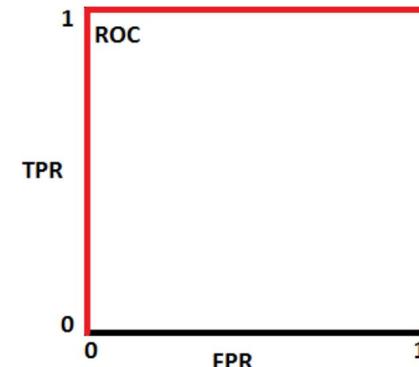
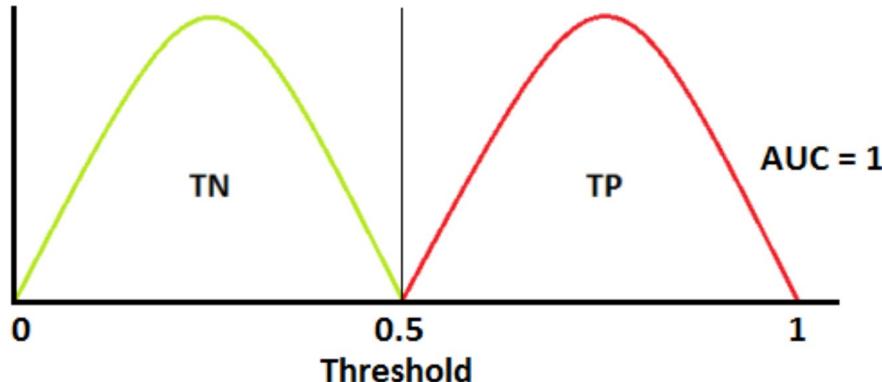
$$\text{FPR} = 1 - \text{Specificity}$$

$$= \frac{\text{FP}}{\text{TN} + \text{FP}}$$

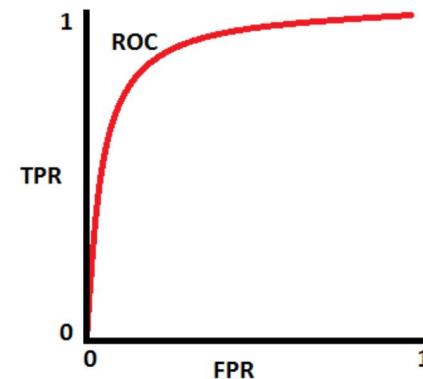
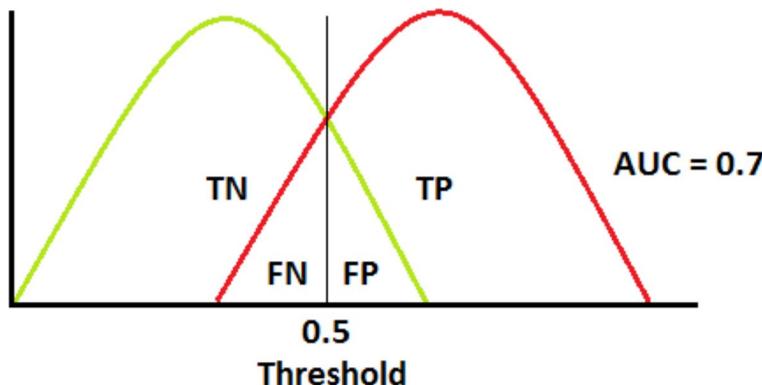
# How to speculate about the performance of the model?

- An excellent model has AUC near to the 1 which means it has a good measure of separability.
- A poor model has AUC near to the 0 which means it has the worst measure of separability.
- In fact, it means it is reciprocating the result. It is predicting 0s as 1s and 1s as 0s.
- And when AUC is 0.5, it means the model has no class separation capacity whatsoever.

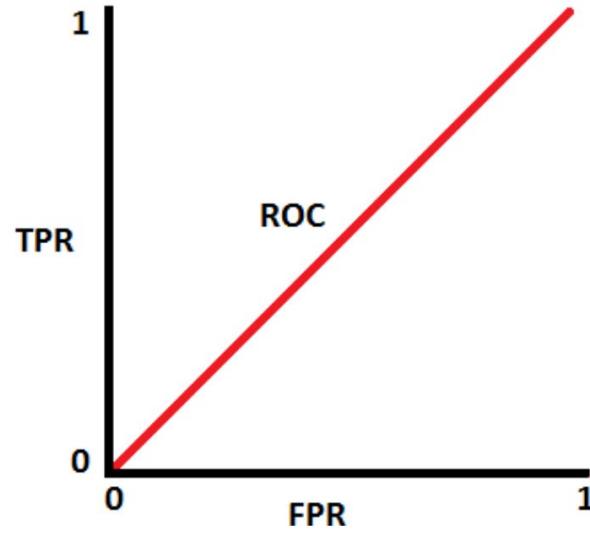
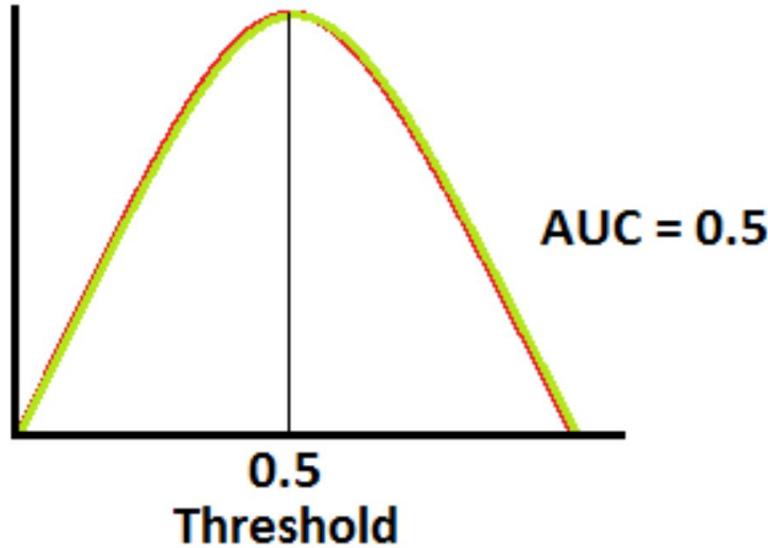
ROC is a curve of probability. So let's plot the distributions of those probabilities:



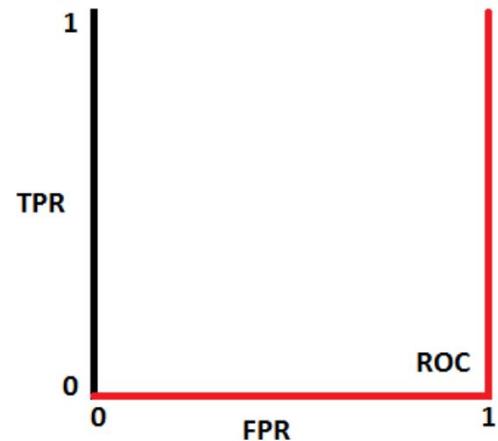
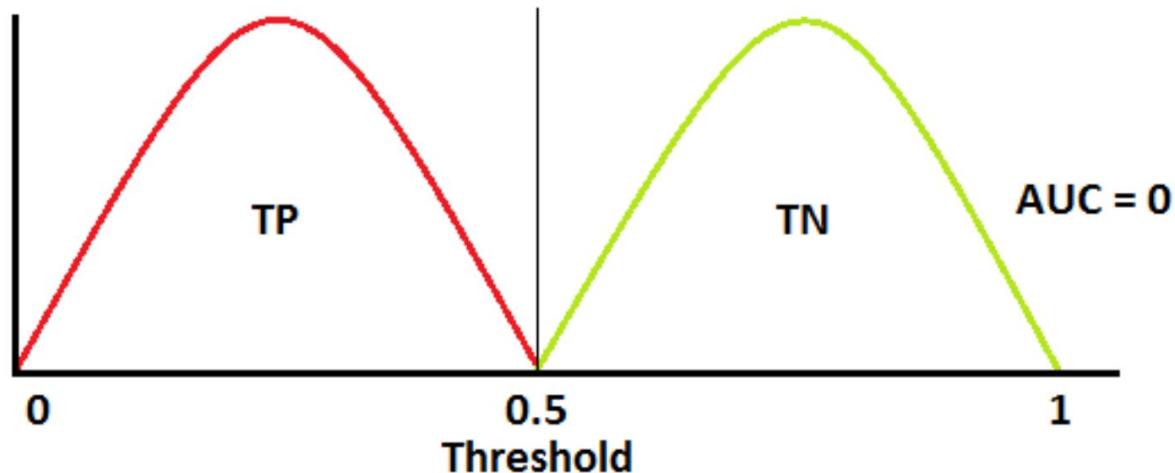
Figures 6 and 71 (Image courtesy My Photohopped Collection)



When two distributions overlap, we introduce type 1 and type 2 errors. Depending upon the threshold, we can minimize or maximize them. When AUC is 0.7, it means there is a 70% chance that the model will be able to distinguish between positive class and negative class.



When AUC is approximately 0, the model is actually reciprocating the classes. It means the model is predicting a negative class as a positive class and vice versa.



## Some Relationships

Sensitivity , Specificity  and Sensitivity ,  
Specificity 

TPR , FPR  and TPR , FPR 

# ROC Curve

**TABLE 3.3** Ratings of 109 CT images by a single radiologist vs. true disease status

True disease status	CT rating					Total
	Definitely normal (1)	Probably normal (2)	Questionable (3)	Probably abnormal (4)	Definitely abnormal (5)	
Normal	33	6	6	11	2	58
Abnormal	3	2	2	11	33	51
Total	36	8	8	22	35	109

**Radiology** The data in Table 3.3 provided by Hanley and McNeil [6], are ratings of computed tomography (CT) images by a single radiologist in a sample of 109 subjects with possible neurological problems. The true disease status is also known for each of these subjects. The data are presented in Table 3.3. How can we quantify the diagnostic accuracy of the test?

Unlike previous examples, this test has no obvious cutoff point to use for designating a subject as positive for disease based on the CT scan. For example, if we designate a subject as test-positive if he or she is either probably abnormal or definitely abnormal (a rating of 4 or 5, or 4+), then the sensitivity of the test is  $(11 + 33)/51 = 44/51 = .86$ , whereas the specificity is  $(33 + 6 + 6)/58 = 45/58 = .78$ . In Table 3.4, we compute the sensitivity and specificity of the radiologist's ratings according to different criteria for test-positive.

To display these data, we construct a receiver operating characteristic (ROC) curve.

---

A **receiver operating characteristic (ROC) curve** is a plot of the sensitivity (on the y-axis) versus  $(1 - \text{specificity})$  (on the x-axis) of a screening test, where the different points on the curve correspond to different cutoff points used to designate test-positive.

---

# ROC Curve

## EXAMPLE 3.33

**Radiology** Construct an ROC curve based on the data in Table 3.4.

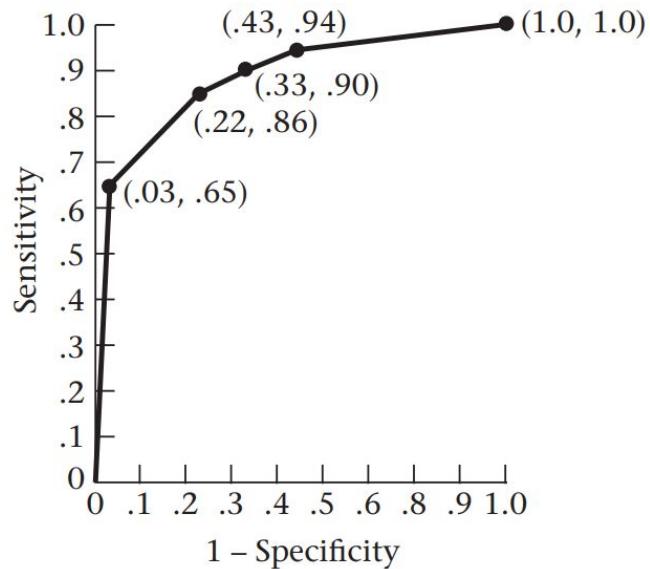
**Solution:** We plot sensitivity on the  $y$ -axis versus  $(1 - \text{specificity})$  on the  $x$ -axis using the data in Table 3.4. The plot is shown in Figure 3.7.

**TABLE 3.4** Sensitivity and specificity of the radiologist's ratings according to different test-positive criteria based on the data in Table 3.3

Test-positive criteria	Sensitivity	Specificity
1 +	1.0	0
2 +	.94	.57
3 +	.90	.67
4 +	.86	.78
5 +	.65	.97
6 +	0	1.0

# ROC Curve

ROC curve for the data in Table 3.4\*



\*Each point represents (1 – specificity, sensitivity) for different test-positive criteria.

# In Class Assessment!

**Radiology** Calculate the area under the ROC curve in Figure 3.7, and interpret what it means.

# ROC Curve

The area under the ROC curve is a reasonable summary of the overall diagnostic accuracy of the test. It can be shown [6] that this area, when calculated by the trapezoidal rule, corresponds to the probability that for a randomly selected pair of normal and abnormal subjects, the abnormal subject will have a higher CT rating. It is assumed that for untied ratings the radiologist designates the subject with the lower test score as normal and the subject with the higher test score as abnormal. For tied ratings, it is assumed that the radiologist randomly chooses one patient as normal and the other as abnormal.

# **PREVALENCE & INCIDENCE**

# PREVALENCE & INCIDENCE

In clinical medicine, the terms prevalence and incidence denote probabilities in a special context and are used frequently in this text.

---

The **prevalence** of a disease is the probability of currently having the disease regardless of the duration of time one has had the disease. Prevalence is obtained by dividing the number of people who currently have the disease by the number of people in the study population.

---

# PREVALENCE & INCIDENCE

**Hypertension** The prevalence of hypertension among adults (age 17 and older) was reported to be 20.3%, as assessed by the NHANES study conducted in 1999–2000 [7]. It was computed by dividing the number of people who had reported taking a prescription for hypertension and were 17 years of age and older (1225) by the total number of people 17 years of age and older in the study population (6044).

---

The **cumulative incidence** of a disease is the probability that a person with no prior disease will develop a new case of the disease over some specified time period.

---

# PREVALENCE & INCIDENCE

**Cancer** The cumulative-incidence rate of breast cancer in 40- to 44-year-old U.S. women over the time period 2002–2006 was approximately 118.4 per 100,000 [2]. This means that on January 1, 2002, about 118 in 100,000 women 40 to 44 years of age who had never had breast cancer would develop breast cancer by December 31, 2002.

# In Class Assessment!

- Suppose that of 25 students in a class, 5 are currently suffering from hay fever. Is the proportion 5 of 25 (20%) a measure of prevalence, incidence, or neither?
- Suppose 50 HIV-positive men are identified, 5 of whom develop AIDS over the next 2 years. Is the proportion 5 of 50 (10%) a measure of prevalence, incidence, or neither?

# In Class Assessment!

**Suppose a standard antibiotic kills a particular type of bacteria 80% of the time. A new antibiotic is reputed to have better efficacy than the standard antibiotic. Researchers propose to try the new antibiotic on 100 patients infected with the bacteria. Using principles of hypothesis testing, researchers will deem the new antibiotic “significantly better” than the standard one if it kills the bacteria in at least 88 out of the 100 infected patients.**

**Question-1:** Suppose there is a true probability (true efficacy) of 85% that the new antibiotic will work for an individual patient. Perform a “simulation study” on the computer, based on random number generation (using, for example, MINITAB, Excel, or R) for a group of 100 randomly simulated patients. Repeat this exercise 20 times with separate columns for each simulated sample of 100 patients. For what percentage of the 20 samples is the new antibiotic considered “significantly better” than the standard antibiotic? (This percentage is referred to as the statistical power of the experiment.) Compare results for different students in the class.

**Question-2:** Repeat the procedure in Problem 3.108 for each simulated patient, assuming the true efficacy of the new antibiotic is (a), 80%, (b) 90%, and (c) 95%, and compute the statistical power for each of (a), (b), and (c).

**Question-3:** Plot the statistical power versus the true efficacy. Do you think 100 patients is a sufficiently large sample to discover whether the new drug is “significantly better” if the true efficacy of the drug is 90%? Why or why not?

# Hypothesis Testing

# Hypothesis Testing

- An **objective** method of making decisions or **inferences** from sample data (evidence)
- Sample data used to choose between two choices i.e. **hypotheses** or statements about a population
- We typically do this by comparing what we have observed to what we expected if one of the statements (**Null Hypothesis**) was true



# Hypothesis Testing

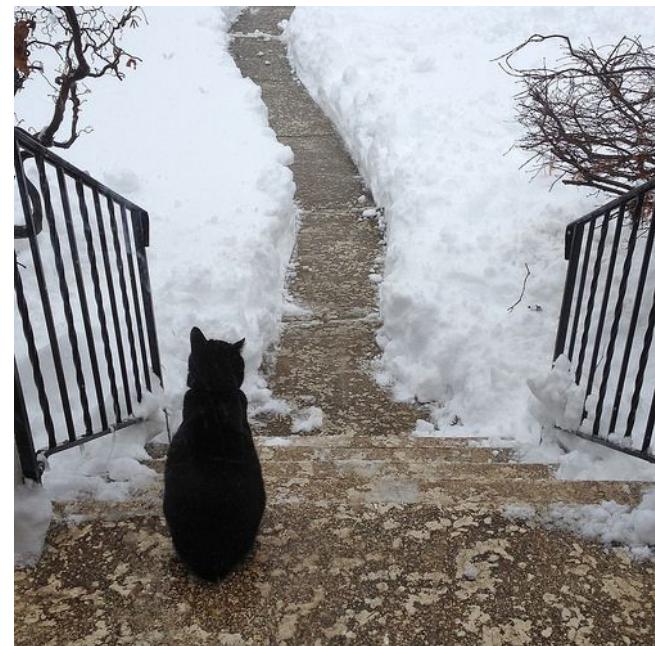
- Always two hypotheses:

$H_A$ : Research (Alternative) Hypothesis

- What we aim to gather evidence of
- Typically that there **is** a difference/effect/relationship etc.

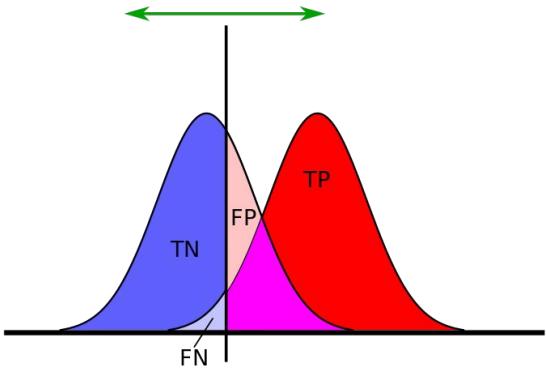
$H_0$ : Null Hypothesis

- What we assume is true to begin with
- Typically that there **is no** difference/effect/relationship etc.

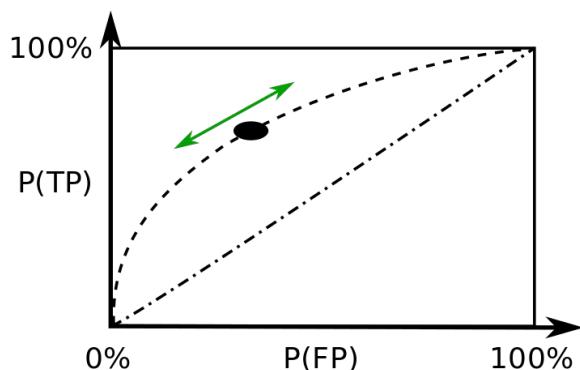


# Some Often asked Questions

- What is the **level of significance**?
- What is a **Type I error**? A **Type II error**?
- What is a **false positive**?
- What is **statistical power**?



TP	FP
FN	TN



The results obtained from positive sample (left curve) overlap with the results obtained from negative samples (right curve).

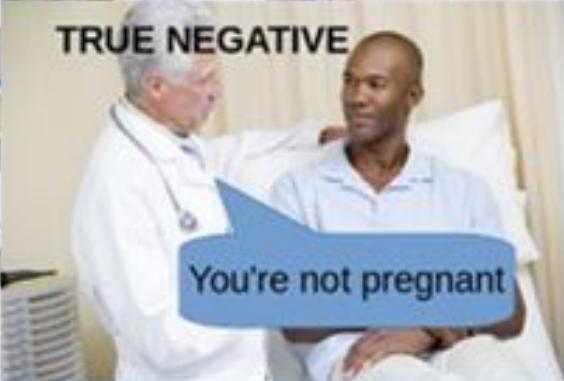
By moving the result cutoff value (vertical bar), the rate of false positives (FP) can be decreased, at the cost of raising the number of false negatives (FN), or vice-versa.

**Typically restrict to a 5% Risk  
= level of significance**

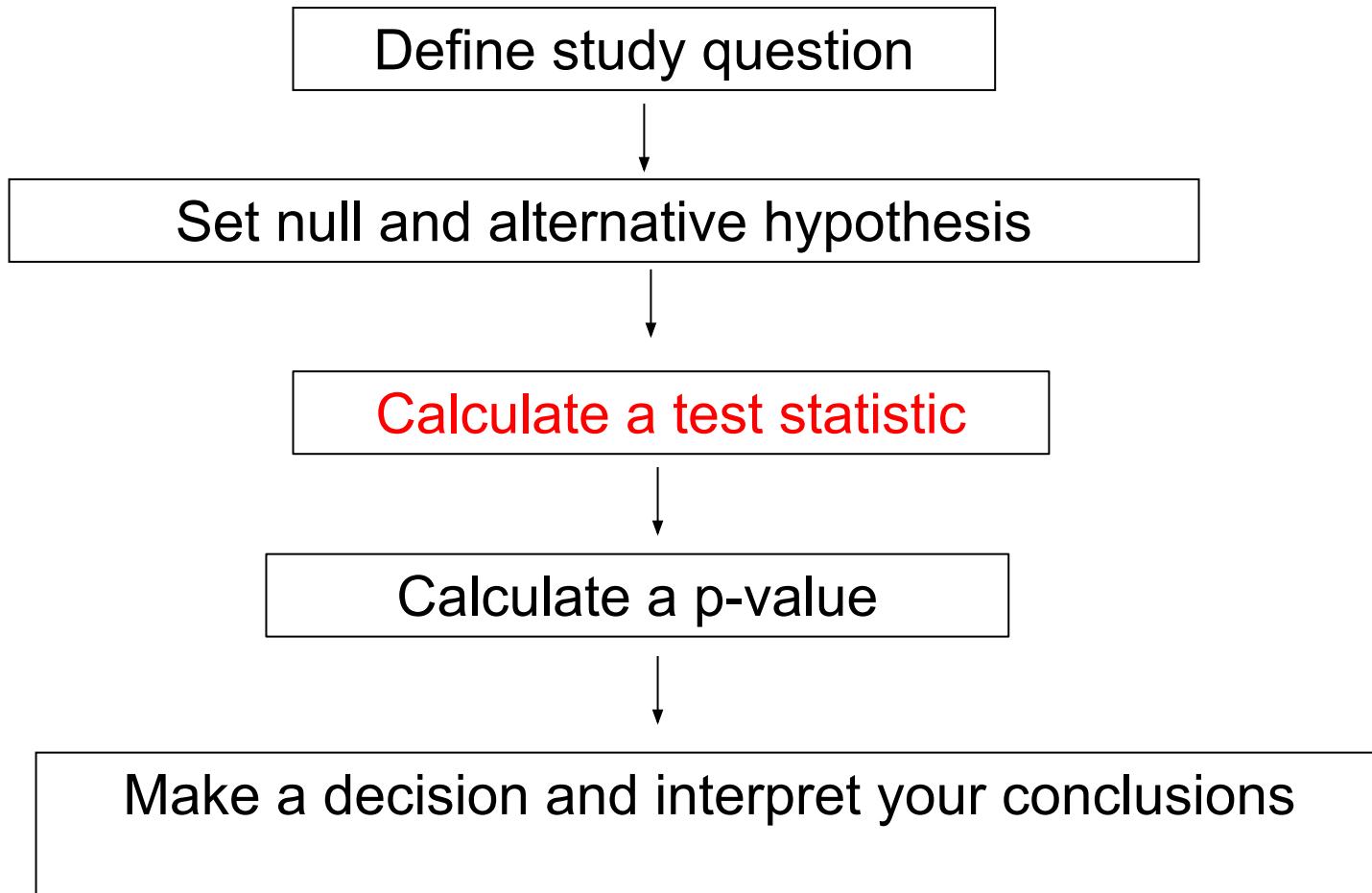
**Controlled via sample size  
(=1-Power of test)**

	Study reports <b>NO</b> difference (Do not reject $H_0$ )	Study reports <b>IS</b> a difference (Reject $H_0$ )
$H_0$ is true Difference Does <b>NOT</b> exist in population		<b>X</b> Type I Error
$H_A$ is true Difference <b>DOES</b> exist in population	<b>X</b>	

**Prob of this = Power of test**

		Actual Values
		1
Predicted Values	1	TRUE POSITIVE 
	0	FALSE POSITIVE 
0	1	FALSE NEGATIVE 
	0	TRUE NEGATIVE 

# Steps to undertaking a Hypothesis test



## Some Real World Problems



**The ship Titanic sank in 1912 with the loss of most of its passengers  
809 of the 1,309 passengers and crew died  
= 61.8%**

**Research question:** Did class (of travel) affect survival?

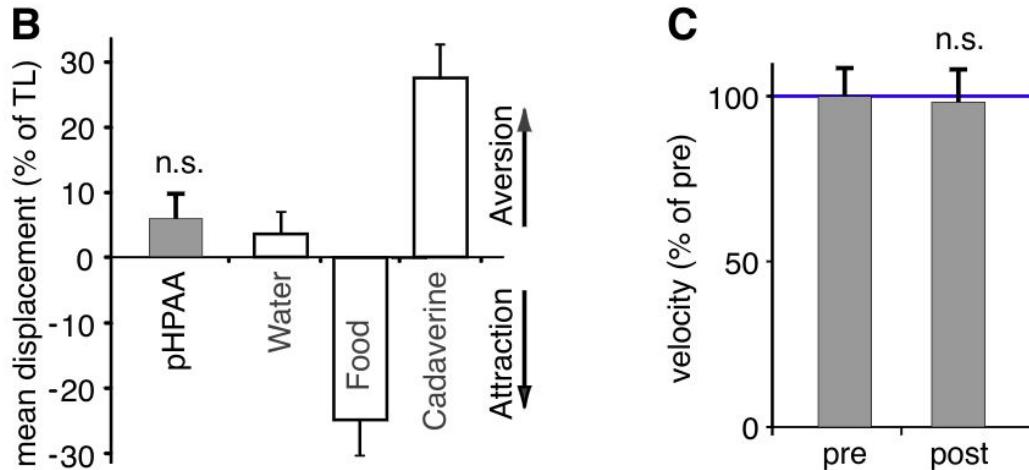
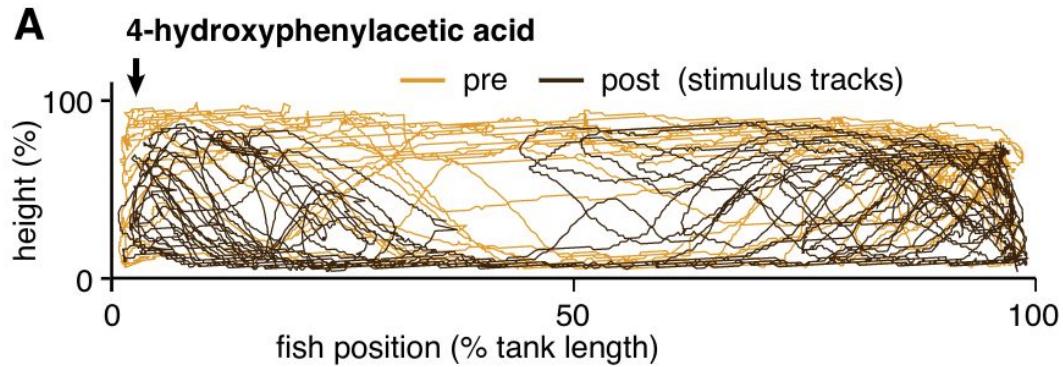
# Titanic

- **Null:** There is **NO** association between class and survival
- **Alternative:** There **IS** an association between class and survival

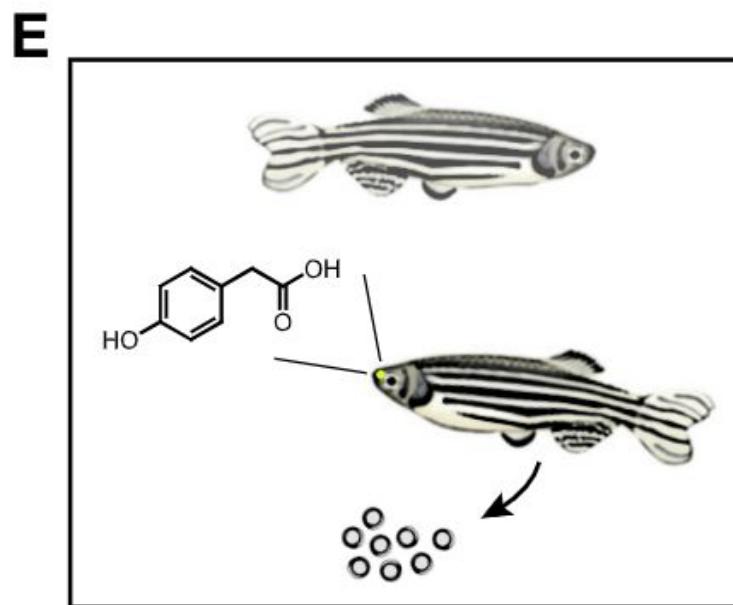
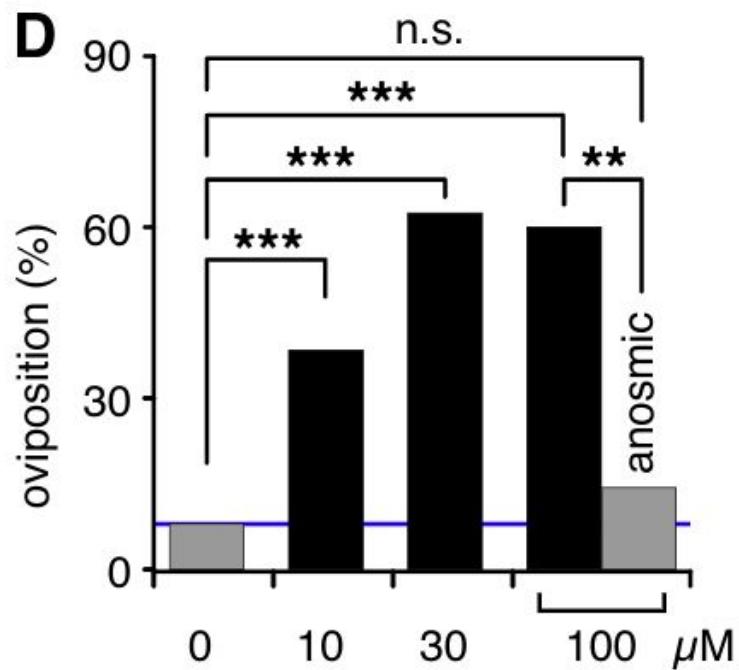
3 x 2 contingency table

		Count		Total	
		Survived?			
Class	1st	Died	Survived		
		123	200	323	
2nd		158	119	277	
3rd		528	181	709	
Total		809	500	1309	

# Real World Problem



# Chi Square Test



# **Statistical Tests**

# Data Analysis

**Statistics** - a powerful tool for analyzing data

1. **Descriptive Statistics** - provide an overview of the attributes of a data set. These include measurements of *central tendency* (frequency histograms, mean, median, & mode) and *dispersion* (range, variance & standard deviation)
  
2. **Inferential Statistics** - provide measures of how well your data support your hypothesis and if your data are generalizable beyond what was tested (*significance tests*)

# The Population: $\mu=5.314$

2	4	10	4	6	8	7	10	4	3	7	9	6	7	5	2	5	8	2	10
7	2	3	5	2	9	3	9	6	1	4	2	6	4	9	3	4	1	8	7
9	1	8	1	10	10	6	4	2	7	1	1	9	10	4	4	6	6	2	5
9	10	2	6	8	10	1	6	10	10	4	4	4	9	2	1	4	5	9	6
6	2	7	8	8	6	6	10	6	6	7	5	9	2	6	4	8	6	6	10
5	7	1	9	1	10	8	8	5	10	1	4	8	3	6	7	1	5	2	4
4	10	5	8	5	1	1	4	3	6	7	3	1	5	4	3	6	2	7	8
3	3	6	6	2	8	6	5	9	8	4	6	3	8	3	3	10	8	10	5
7	5	1	4	3	2	1	10	2	10	6	10	7	9	8	8	4	9	9	10
3	7	6	2	1	1	10	3	5	7	4	1	2	9	10	10	6	1	3	2
1	3	9	9	4	2	2	2	1	8	3	1	5	9	9	8	3	2	5	4
4	2	3	10	8	2	3	4	1	3	3	2	10	10	5	7	3	3	10	1
5	7	5	1	2	5	8	7	3	8	9	2	10	8	1	1	5	3	3	7
6	7	9	8	8	4	9	8	4	3	10	8	10	4	10	2	3	5	6	3
1	9	8	1	10	2	3	1	6	3	8	9	6	2	4	4	2	7	8	4
4	4	4	10	8	5	9	3	10	5	3	6	9	3	7	4	2	3	10	2
5	1	6	8	5	6	8	1	8	5	7	6	4	1	2	7	2	9	5	3
8	2	3	2	9	9	1	1	5	7	8	5	6	3	8	5	4	10	6	9
5	1	10	10	5	1	4	3	2	3	6	9	10	2	6	3	1	2	8	6
1	8	7	8	5	3	7	2	4	1	8	9	10	10	5	1	3	6	5	8
3	3	8	8	2	7	1	6	9	8	2	10	3	7	9	2	1	9	7	7
3	1	9	6	8	2	6	4	6	3	7	10	9	6	1	10	7	5	3	10
1	6	5	4	3	2	4	4	1	5	5	10	6	2	1	1	1	5	6	3
8	10	8	10	9	7	7	7	8	4	8	1	3	5	8	1	8	4	4	6
4	7	2	4	9	1	8	5	3	3	5	10	1	4	6	3	3	8	2	2

Population size = 500

# The Population: $\mu=5.314$

2	4	10	4	6	8	7	10	4	3	<b>7</b>	9	6	7	5	2	5	8	2	10
7	2	3	5	2	9	3	9	6	1	4	2	6	4	9	3	4	1	8	7
9	1	8	1	10	10	<b>6</b>	4	2	7	1	1	9	10	4	4	6	6	2	5
9	10	2	6	8	10	1	6	10	10	4	4	4	9	2	1	4	5	9	6
6	2	7	8	8	6	6	10	6	6	7	5	9	2	6	4	8	6	6	10
5	7	1	9	1	10	8	8	5	10	1	4	8	3	6	7	1	5	2	4
4	10	5	8	5	1	1	4	3	6	7	3	1	5	4	3	6	2	7	8
3	3	6	6	2	8	6	5	9	8	4	6	3	8	3	3	10	8	10	5
7	5	1	<b>4</b>	3	2	1	10	2	10	6	10	7	9	8	8	4	9	9	10
3	7	6	2	1	1	10	3	5	7	4	1	2	9	10	10	6	1	3	2
1	3	9	9	4	2	2	2	1	8	3	1	5	9	9	8	3	2	5	4
4	2	3	10	8	2	3	4	1	3	3	2	10	10	5	7	3	3	10	1
5	7	5	1	2	5	8	7	3	8	<b>9</b>	2	10	8	1	<b>1</b>	5	3	3	7
6	7	9	8	8	4	9	8	4	3	10	8	10	4	10	2	3	5	6	3
1	9	8	1	10	2	3	1	6	3	8	9	6	2	4	4	2	7	8	4
4	4	4	10	8	5	9	3	10	5	3	6	9	3	7	4	2	3	10	2
5	1	6	8	5	6	8	1	8	5	7	6	4	1	2	7	2	9	5	3
8	2	3	2	9	9	1	1	5	7	8	5	6	3	8	5	4	10	6	9
5	1	10	10	5	1	4	3	2	3	6	9	10	2	<b>6</b>	3	1	2	8	6
1	<b>8</b>	7	8	5	<b>3</b>	7	2	4	1	8	9	10	10	5	1	3	6	5	8
3	3	8	8	2	7	1	6	9	8	2	10	3	7	9	2	1	9	7	7
3	1	9	6	8	2	6	4	6	3	7	10	9	6	1	10	7	5	3	10
1	6	5	4	3	<b>2</b>	4	4	1	5	5	10	6	2	1	1	1	5	6	3
8	10	8	10	9	7	7	7	8	4	8	1	3	5	8	1	8	4	4	6
4	7	2	4	9	1	8	5	3	3	5	10	1	4	6	3	3	8	2	2

The Sample: 7, 6, 4, 9, 8, 3, 2, 6, 1  
mean = 5.111

# The Population: $\mu=5.314$

2	4	10	4	6	8	7	10	4	3	7	9	6	7	5	2	5	8	2	10
7	2	3	5	2	9	3	9	6	1	4	2	6	4	9	3	4	1	8	7
9	1	8	1	10	10	6	4	2	7	1	1	9	10	4	4	6	<b>6</b>	2	5
9	10	2	6	8	10	1	6	10	10	4	4	4	9	2	1	4	5	9	6
6	2	7	8	8	6	6	10	6	6	7	5	9	2	6	4	8	6	6	10
5	7	<b>1</b>	9	1	10	8	8	5	10	1	4	8	3	6	7	1	5	2	4
4	10	5	8	5	1	1	4	3	6	7	3	1	5	4	3	6	2	7	8
3	3	6	6	2	8	6	5	9	8	4	6	3	8	3	3	10	8	10	5
7	5	1	4	3	2	1	10	2	10	6	10	7	9	8	8	4	9	9	10
3	7	6	2	1	1	10	3	5	<b>7</b>	4	1	2	9	10	10	6	1	3	2
1	3	9	9	4	2	2	2	1	8	3	1	5	9	9	8	3	2	5	4
4	2	3	10	8	2	3	4	1	3	3	2	10	10	5	7	3	3	10	1
5	7	5	1	2	5	8	7	3	8	9	2	10	8	1	1	5	3	3	7
6	7	9	8	8	4	9	8	4	3	10	8	10	4	10	2	3	5	<b>6</b>	3
1	9	8	1	10	2	3	1	6	3	8	9	6	2	4	4	2	7	8	4
4	4	4	10	8	5	9	3	10	5	3	6	9	3	7	<b>4</b>	2	3	10	2
5	1	6	<b>8</b>	5	6	8	1	8	5	7	6	4	1	2	7	2	9	5	3
8	2	3	2	9	9	1	1	5	7	8	5	6	3	8	5	4	10	6	9
5	1	10	10	5	1	4	3	2	3	6	9	10	2	6	3	1	2	8	6
1	8	7	8	5	3	7	2	4	1	8	9	10	10	5	1	3	6	5	8
3	3	8	8	2	7	1	6	9	8	2	10	3	7	9	2	1	9	7	7
3	1	9	6	8	2	6	4	6	3	7	10	9	6	1	10	7	5	3	10
1	6	<b>5</b>	4	3	2	4	4	1	5	5	10	6	2	1	1	1	5	6	3
8	10	8	10	9	7	7	7	8	4	8	1	3	5	8	<b>1</b>	8	4	4	6
4	7	2	4	9	1	8	5	3	3	5	10	1	4	6	3	3	8	2	2

The Sample: 1, 5, 8, 7, 4, 1, 6, 6  
mean = 4.75

# Parametric or Non-parametric?

- Parametric tests are restricted to data that:

- 1) show a normal distribution
- 2) \* are independent of one another
- 3) \* are on the same *continuous* scale of measurement

- Non-parametric tests are used on data that:

- 1) show an other-than normal distribution
- 2) are dependent or conditional on one another
- 3) in general, do not have a continuous scale of measurement

e.g., the length and weight of something → parametric

vs.

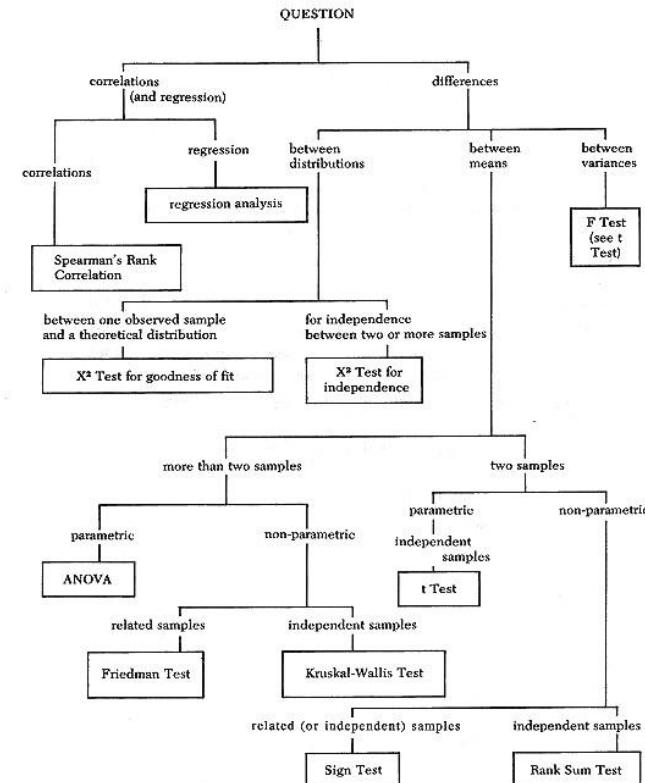
did the bacteria grow or not grow → non-parametric

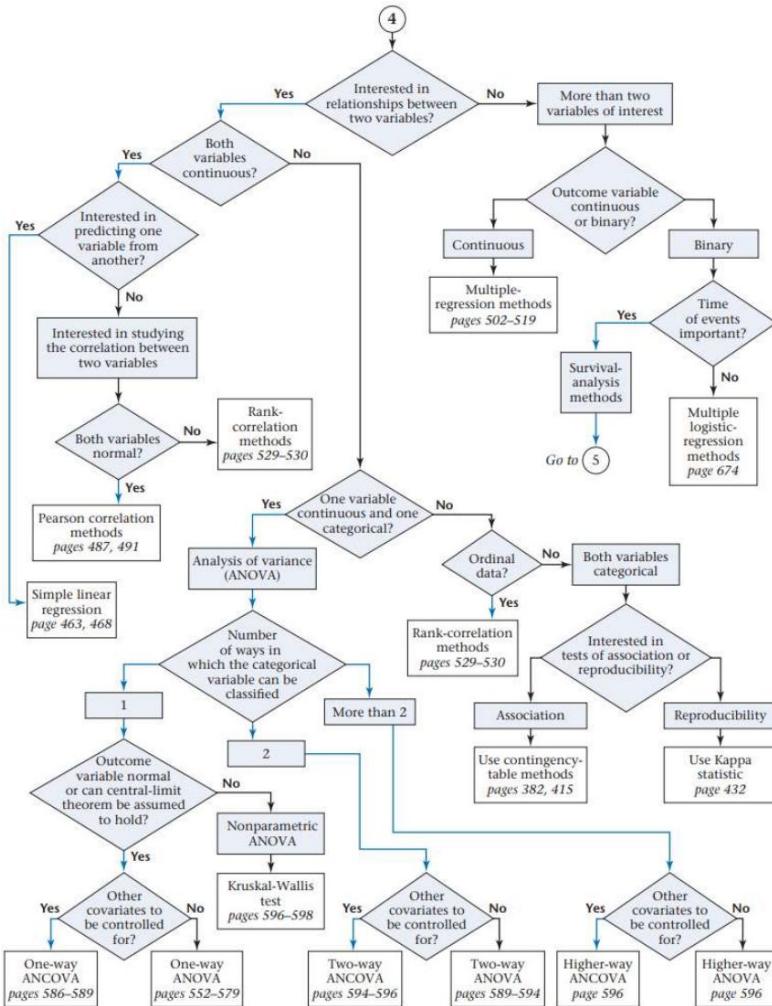
# The First Question

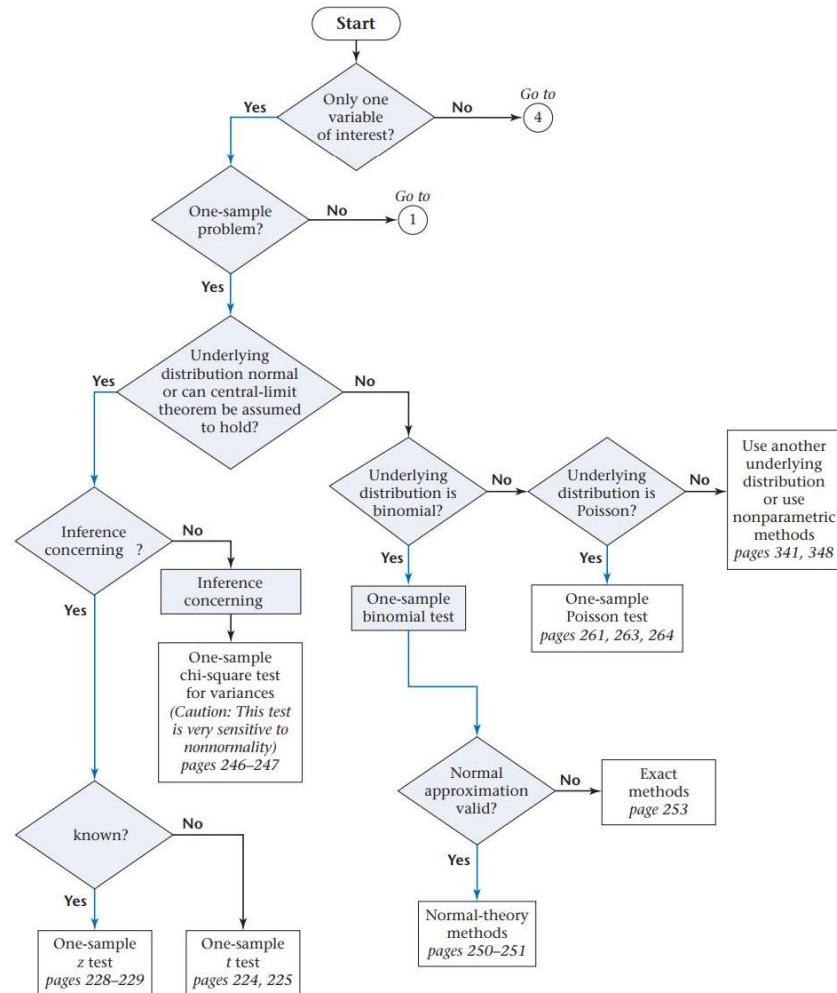
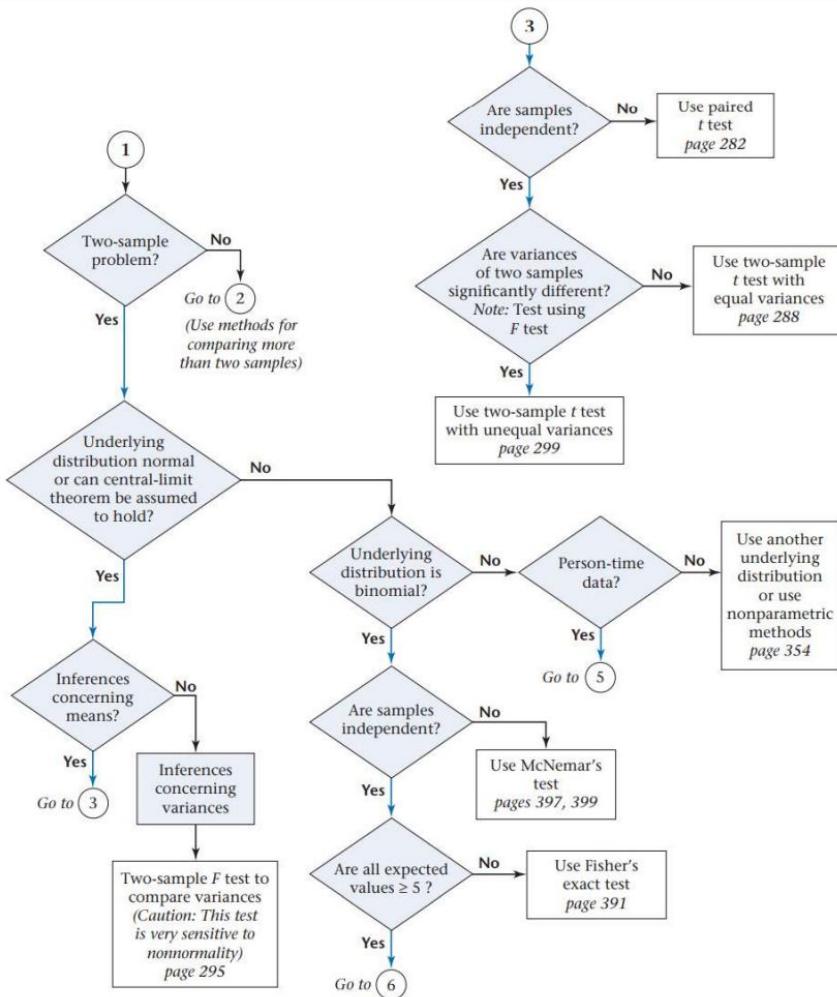
After examining your data, ask: does what you're testing seem to be a question of relatedness or a question of difference?

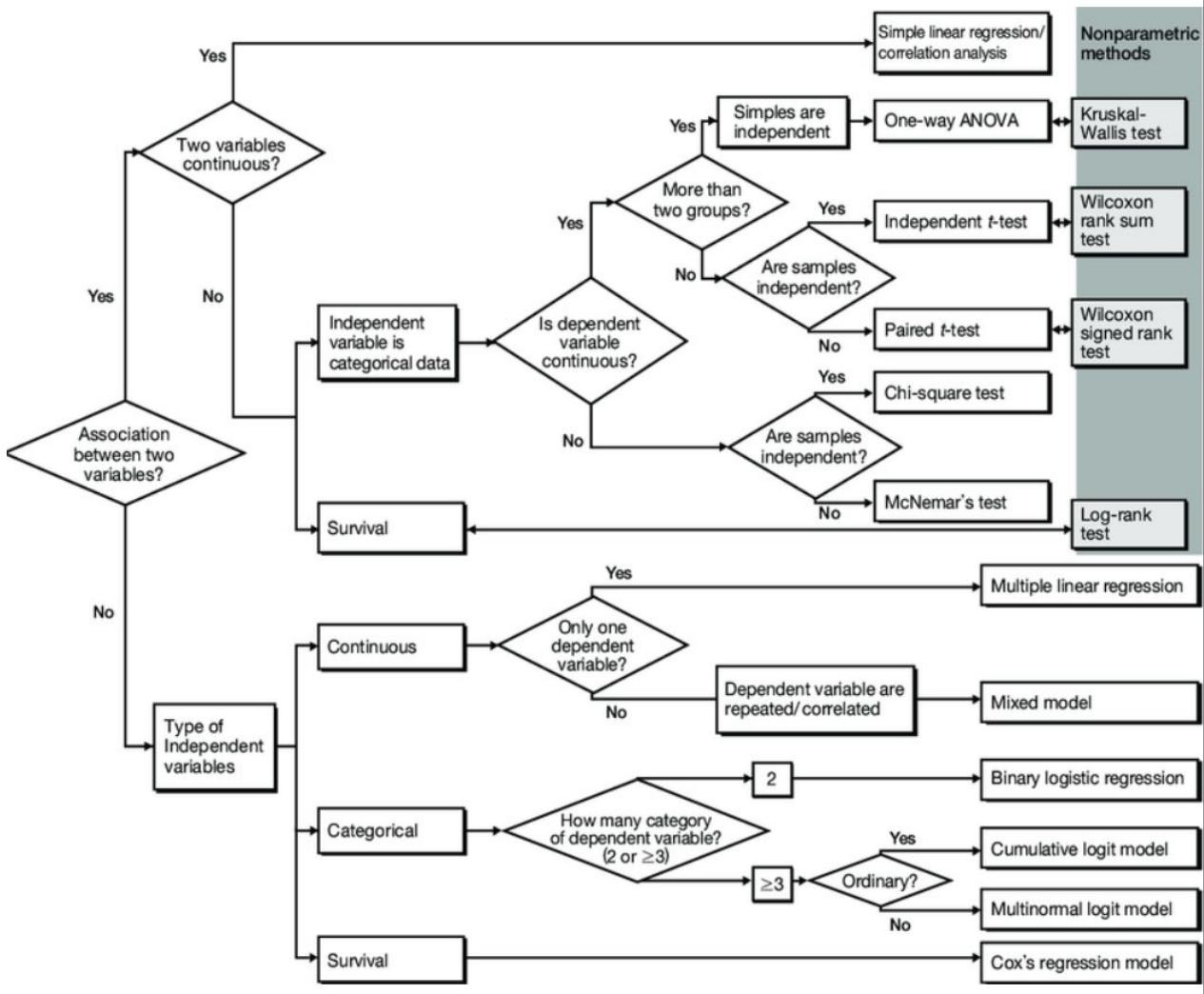
If **relatedness** (between your control and your experimental samples or between your dependent and independent variable), you will be using tests for correlation (positive or negative) or regression.

If **difference** (your control differs from your experimental), you will be testing for independence between distributions, means or variances. Different tests will be employed if your data show parametric or non-parametric properties.









# Tests for Differences

- Between **Means**

- t-Test - **P**
- ANOVA - **P**
- Friedman Test
- Kruskal-Wallis Test
- Sign Test
- Rank Sum Test

- Between **Distributions**

- Chi-square for goodness of fit
- Chi-square for independence

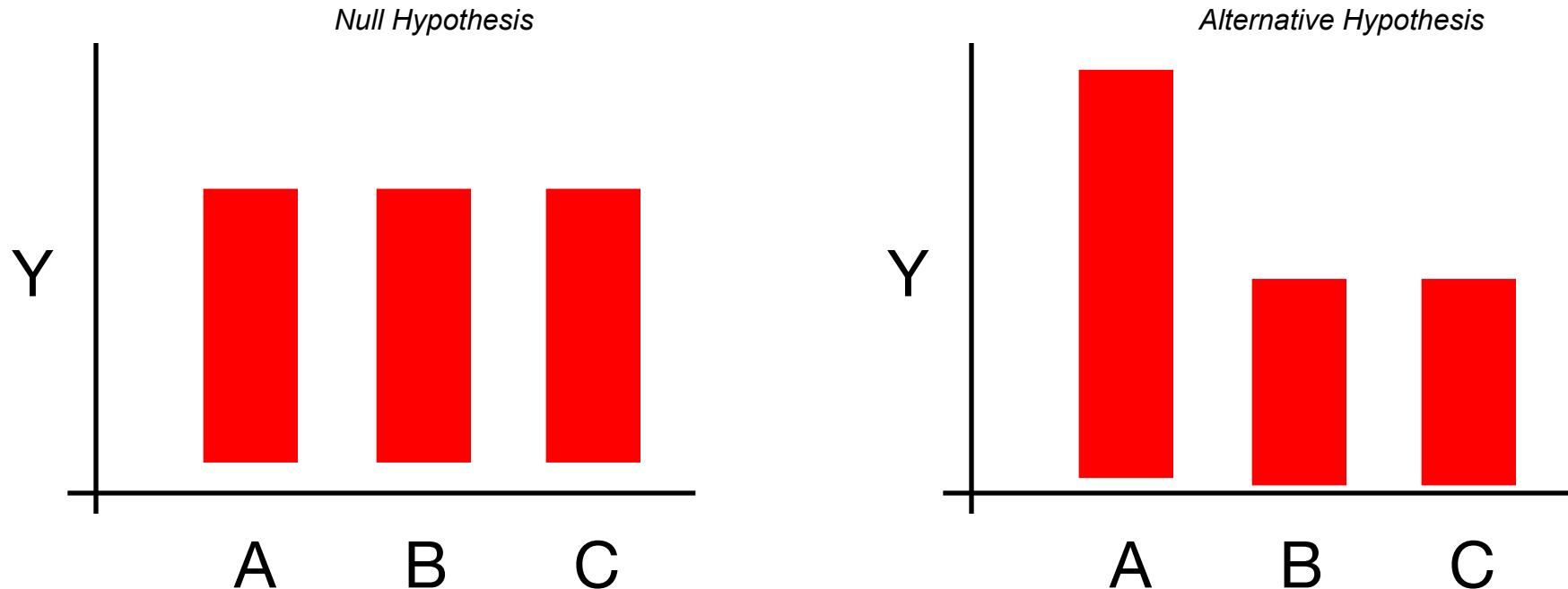
- Between **Variances**

- F-Test – **P**

**P** – parametric tests

# Differences Between Means

Asks whether samples come from populations with different means



*There are different tests if you have 2 vs more than 2 samples*

# Differences Between Means – Parametric Data

t-Tests compare the means of **two parametric** samples

E.g. Is there a difference in the mean height of men and women?

Question: A researcher compared the height of plants grown in high and low light levels. Her results are shown on right.

Use a T-test to determine whether there is a statistically significant difference in the heights of the two groups

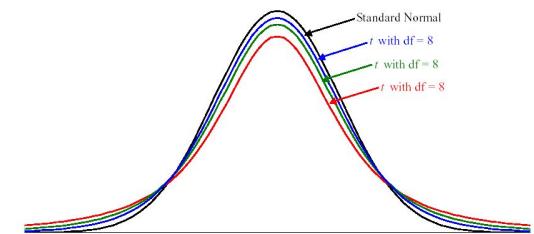
Low Light	High Light
49	45
31	40
43	59
31	58
40	55
44	50
49	46
48	53
33	43

# T-Test

Student's *t*-distribution

## Open Questions

- What is a T Test?
- The T Score
- T Values and P Values
- Calculating the T Test
- What is a Paired T Test (Paired Samples T Test)?



The **t** test tells you how significant the differences between groups are; In other words it lets you know if those differences (measured in means/averages) could have happened by chance.

Student's T-tests can be used in real life to compare means. For example, a drug company may want to test a new cancer drug to find out if it improves life expectancy. In an experiment, there's always a control group (a group who are given a placebo, or "sugar pill"). **The control group may show an average life expectancy of +5 years**, while the group taking the **new drug might have a life expectancy of +6 years**. It would seem that the drug might work. But it could be due to a fluke. To test this, researchers would use a Student's t-test to find out if the results are repeatable for an entire population.

# The T Score

The t score is a **ratio between the difference between two groups and the difference within the groups**. The larger the t score, the more difference there is between groups. The smaller the t score, the more similarity there is between groups. A t score of 3 means that the groups are three times as different from each other as they are within each other. When you run a t test, the bigger the t-value, the more likely it is that the results are repeatable.

- **A large t-score tells you that the groups are different.**
- **A small t-score tells you that the groups are similar.**

$$t = \frac{\frac{(\sum D)/N}{\sum D^2 - \left(\frac{(\sum D)^2}{N}\right)}}{\sqrt{\frac{(N-1)(N)}{(N-1)(N)}}$$

# T-Values and P-values

How big is “big enough”? Every t-value has a p-value to go with it.

1. A p-value is the probability that the results from your sample data occurred by chance.
2. P-values are from 0% to 100%. They are usually written as a decimal. For example, a p value of 5% is 0.05.
3. Low p-values are good; They indicate your data did not occur by chance.
4. For example, a p-value of .01 means there is only a 1% probability that the results from an experiment happened by chance. In most cases, a p-value of 0.05 (5%) is accepted to mean the data is valid.

# **T-Test**

There are three main types of t-test:

- An Independent Samples t-test compares the means for two groups.
- A Paired sample t-test compares means from the same group at different times (say, one year apart).
- A One sample t-test tests the mean of a single group against a known mean.

## t test and related tests for equal means

t test   F test   Mann-Whitney   Kolmogorov-Smirnov   Coeff. of var.

### Tests for equal means

#### *Males*

N: 25  
Mean: 6.012  
95% conf.: (5.7858 6.2382)  
Variance: 0.30027

#### *Females*

N: 25  
Mean: 5.028  
95% conf.: (4.8627 5.1933)  
Variance: 0.16043

Difference between means: 0.984

95% conf. interval (parametric): (0.71106 1.2569)  
95% conf. interval (bootstrap): (0.728 1.248)

*t* : 7.2486      *p* (same mean): 3.0602E-09

Uneq. var. *t* : 7.2486      *p* (same mean): 4.9785E-09

Monte Carlo permutation: *p* (same mean): 0.0001

## t test and related tests for equal means

### **t test**

The *t* test has null hypothesis

$H_0$ : The two samples are taken from populations with equal means.

The *t* test assumes normal distributions and equal variances.

From the standard error  $s_D$  of the difference of the means given above, the test statistic is

$$t = \frac{\bar{x} - \bar{y}}{s_D}$$

### **Unequal variance *t* test**

The unequal variance *t* test is also known as the Welch test. It can be used as an alternative to the basic *t* test when variances are very different, although it can be argued that testing for difference in the means in this case is questionable. The test statistic is

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\text{Var}(x)/n_1 + \text{Var}(y)/n_2}}$$

## F test for equal variances

t test	F test	Mann-Whitney	Kolmogorov-Smirnov	Coeff. of var.
Tests for equal variances				
<i>Males</i>		<i>Females</i>		
N:	12	N:	12	
Variance:	0.53356	Variance:	0.094545	
<i>F</i> :	5.6434	<i>p</i> (same var.):	0.0078543	
Monte Carlo permutation:		<i>p</i> (same var.):	0.0037	
Exact permutation:		<i>p</i> (same var.):	0.0041399	

The *F* test has null hypothesis

$H_0$ : The two samples are taken from populations with equal variance.

Normal distribution is assumed. The *F* statistic is the ratio of the larger variance to the smaller. The significance is two-tailed, with  $n_1$  and  $n_2$  degrees of freedom.

## Mann-Whitney test for equal medians

The two-tailed (Wilcoxon) Mann-Whitney  $U$  test can be used to test whether the medians of two independent samples are different. It is a non-parametric test and does not assume normal distribution, but does assume equal-shaped distribution in both groups. The null hypothesis is

$H_0$ : The two samples are taken from populations with equal medians.

t test	F test	Mann-Whitney	Kolmogorov-Smirnov	Coeff. of var.
Tests for equal medians				
<i>Males</i>				
N:	12	<i>Females</i>		
Mean rank:	8.75	N:	12	
		Mean rank:	3.75	
Mann-Whitn U : 12				
z :	-3.4427	p (same med.):	0.00057588	
Monte Carlo permutation:		p (same med.):	0.0001	
Exact permutation:		p (same med.):	0.00018194	

$$z = \frac{U - n_1 n_2 / 2 + 0.5}{\sqrt{\frac{n_1 n_2 \left( n^3 - n - \sum_g f_g^3 - f_g \right)}{12n(n-1)}}}$$

## Equal distributions

The Kolmogorov-Smirnov test is a nonparametric test for overall equal distribution of two univariate samples. In other words, it does not test specifically for equality of mean, variance or any other parameter. The null hypothesis is  $H_0$ : The two samples are taken from populations with equal distribution.

### **Anderson-Darling test for equal distributions**

The Anderson-Darling test is a nonparametric test for overall equal distribution of two univariate samples. It is an alternative to the Kolmogorov-Smirnov test.

With two samples  $x_1 \dots x_n$  and  $y_1 \dots y_m$ , the pooled sample size is  $N=n+m$ . The test statistic  $A^2_N$  is computed according to Pettitt (1976):

$$A^2_N = \frac{1}{mn} \sum_{i=1}^{N-1} \frac{(M_i N - ni)^2}{i(N-i)}$$

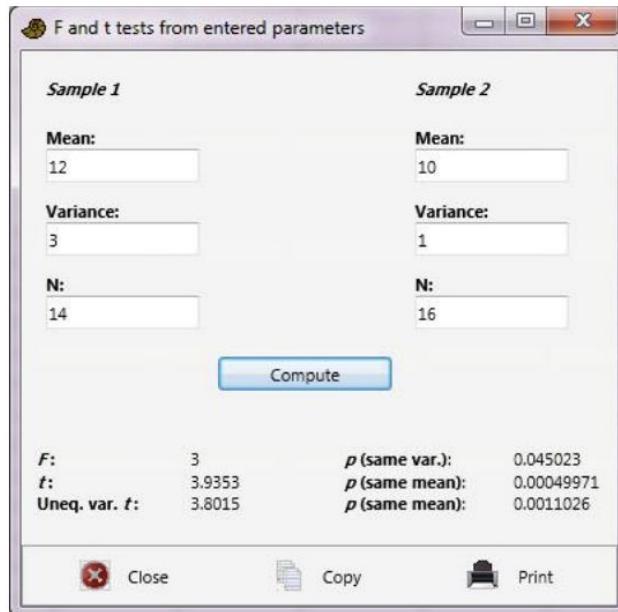
where  $M_i$  is the number of x's less than or equal to the  $i$ th smallest in the pooled sample.

This statistic is transformed to a statistic called  $Z$  according to Scholz & Stephens (1987). For our case with  $k=2$  samples, compute the variance of the statistic as follows:

## F and t tests from parameters

### **F and t tests from parameters**

Sometimes publications give not the data, but values for sample size, mean and variance for two samples. These can be entered manually using the 'F and t from parameters' option in the menu. This module does not use any data from the spreadsheet.

A screenshot of a software window titled "F and t tests from entered parameters". The window has two sections for "Sample 1" and "Sample 2". In "Sample 1", the Mean is 12, Variance is 3, and N is 14. In "Sample 2", the Mean is 10, Variance is 1, and N is 16. A "Compute" button is at the bottom left. Below the buttons, results are displayed: F: 3, t: 3.9353, p (same var.): 0.045023; t: 3.8015, p (same mean): 0.00049971; Uneq. var. t: 3.8015, p (same mean): 0.0011026. At the bottom are "Close", "Copy", and "Print" buttons.

Sample 1	Sample 2
Mean: 12	Mean: 10
Variance: 3	Variance: 1
N: 14	N: 16

Compute

**F:** 3      **p (same var.):** 0.045023  
**t:** 3.9353      **p (same mean):** 0.00049971  
**Uneq. var. t:** 3.8015      **p (same mean):** 0.0011026

 Close     Copy     Print

## Differences Between Means – Parametric Data

ANOVA (Analysis of Variance) compares the means of **two or more parametric** samples.

E.g. Is there a difference in the mean height of plants grown under red, green and blue light?

**A researcher fed pigs on four different foods. At the end of a month feeding, he weighed the pigs. Use an ANOVA test to determine if the different foods resulted in differences in growth of the pigs.**

weight of pigs fed different foods				
food 1	food 2	food 3	food 4	
60.8	68.7	102.6	87.9	
57.0	67.7	102.1	84.2	
65.0	74.0	100.2	83.1	
58.6	66.3	96.5	85.7	
61.7	69.8	69.8	90.3	



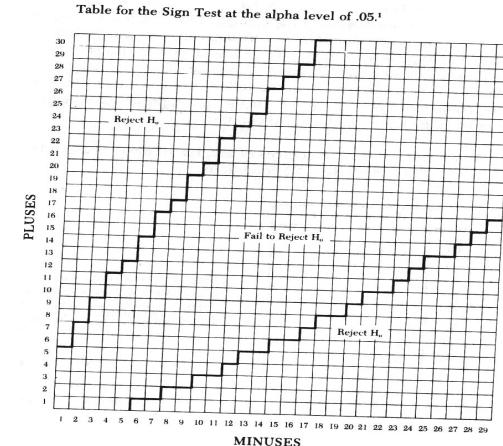
*Aplysia punctata* – the sea hare

## Differences Between Means – Non-Parametric Data

The Sign Test compares the means of **two “paired”, non-parametric** samples

E.g. Is there a difference in the gill withdrawal response of *Aplysia* in night versus day? Each subject has been tested once at night and once during the day → paired data.

Subject	Night Response	Day Response
1	2	5
2	1	3
3	2	2



<sup>1</sup> Constructed from values found in The Chemical Rubber Co., *Handbook of Probability and Statistics*, 2nd ed. (W.H. Beyer, ed.) The Chemical Rubber Company, Cleveland, Ohio, 1968.

## Differences Between Means – Non-Parametric Data

The Friedman Test is like the Sign test, (compares the means of “paired”, non-parametric samples) for **more than two samples**.

E.g. Is there a difference in the gill withdrawal response of *Aplysia* between morning, afternoon and evening? Each subject has been tested once during each time period → paired data

Subject	Morning Response	Afternoon Response	Evening. Response
1	4	3	2
2	5	2	1
3	3	4	3

## Differences Between Means – Non-Parametric Data

The Rank Sum test compares the means of **two *non-parametric*** samples

E.g. Is there a difference in the gill withdrawal response of *Aplysia* in night versus day? Each subject has been tested once, either during the night **or** during the day → unpaired data.

Subject	Night Response	Day Response
1		5
2	1	
3		2
4	3	
5		4
6	1	
7		5

## Differences Between Means – Non-Parametric Data

The Kruskal-Wallis Test compares the means of **more than two *non-parametric*, non-paired** samples

E.g. Is there a difference in the gill withdrawal response of *Aplysia* in night versus day? Each subject has been tested once, either during the morning, afternoon or evening → unpaired data.

Subject	Morning Response	Afternoon Response	Evening Response
1	4		
2	5		
3		4	
4		3	
5			2
6			3

# Hypothesis Testing with R

IBM

# A Quick Review

## C

Median	<code>median(mpg\$cty)</code>	17
--------	-------------------------------	----

## Spread

Measure	R	Result
Variance	<code>var(mpg\$cty)</code>	18.1130736
Standard Deviation	<code>sd(mpg\$cty)</code>	4.2559457
IQR	<code>IQR(mpg\$cty)</code>	5
Minimum	<code>min(mpg\$cty)</code>	9
Maximum	<code>max(mpg\$cty)</code>	35
Range	<code>range(mpg\$cty)</code>	9, 35

# A Quick Revision

- Histograms
- Barplots
- Boxplots
- Scatterplots

```
hist(mpg$cty,
      xlab = "Miles Per Gallon (City)",
      main = "Histogram of MPG (City)",
      breaks = 12,
      col = "dodgerblue",
      border = "darkorange")
```

```
barplot(table(mpg$drv),
        xlab = "Drivetrain (f =
        ylab = "Frequency",
        main = "Frequency Distribution by Drivetrain",
        col = "darkorange",
        border = "black")
plot(hwy ~ displ, data = mpg,
      xlab = "Engine Displacement (in Liters)",
      ylab = "Miles Per Gallon (Highway)",
      main = "MPG (Highway) vs Engine Displacement",
      pch = 20,
      col = "darkorange",
      border = "black")
boxplot(hwy ~ drv, data = mpg,
        xlab = "Drivetrain (f = FWD, r = RWD, 4 = 4WD)",
        ylab = "Miles Per Gallon (Highway)",
        main = "MPG (Highway) vs Drivetrain",
        pch = 20,
```

# Matrices

```
x = 1:9  
x
```

```
## [1] 1 2 3 4 5 6 7 8 9
```

```
X = matrix(x, nrow = 3, ncol = 3)  
X
```

```
##      [,1] [,2] [,3]  
## [1,]     1     4     7  
## [2,]     2     5     8  
## [3,]     3     6     9
```

```
rbind(x, rev(x), rep(1, 9))
```

```
##   [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9]
## x     1     2     3     4     5     6     7     8     9
##      9     8     7     6     5     4     3     2     1
##      1     1     1     1     1     1     1     1     1
```

```
cbind(col_1 = x, col_2 = rev(x), col_3 = rep(1, 9))
```

```
##       col_1 col_2 col_3
## [1,]     1     9     1
## [2,]     2     8     1
## [3,]     3     7     1
## [4,]     4     6     1
## [5,]     5     5     1
## [6,]     6     4     1
## [7,]     7     3     1
## [8,]     8     2     1
## [9,]     9     1     1
```

```
X = matrix(1:6, 2, 3)
X
```

```
##      [,1] [,2] [,3]
## [1,]     1     3     5
## [2,]     2     4     6
```

```
dim(X)
```

```
## [1] 2 3
```

```
rowSums(X)
```

```
## [1] 9 12
```

```
colSums(X)
```

```
## [1] 3 7 11
```

# Data Frames

We have previously seen vectors and matrices for storing data as we introduced R. We will now introduce a **data frame** which will be the most common way that we store and interact with data in this course.

```
example_data = data.frame(x = c(1, 3, 5, 7, 9, 1, 3, 5, 7, 9),  
                           y = c(rep("Hello", 9), "Goodbye"),  
                           z = rep(c(TRUE, FALSE), 5))
```

```
example_data$x
```

```
## [1] 1 3 5 7 9 1 3 5 7 9
```

```
str(example_data)
```

```
## 'data.frame':    10 obs. of  3 variables:  
##   $ x: num  1 3 5 7 9 1 3 5 7 9  
##   $ y: chr  "Hello" "Hello" "Hello" "Hello" ...  
##   $ z: logi  TRUE FALSE TRUE FALSE TRUE FALSE ...
```

```
nrow(example_data)
```

```
## [1] 10
```

```
ncol(example_data)
```

```
## [1] 3
```

```
dim(example_data)
```

```
## [1] 10  3
```

# Probability in R

The general naming structure of the relevant R functions is:

- `dname` calculates density (pdf) at input `x`.
- `pname` calculates distribution (cdf) at input `x`.
- `qname` calculates the quantile at an input probability.
- `rname` generates a random draw from a particular distribution.

Note that `name` represents the name of the given distribution.

# Probability in R

To calculate the value of the pdf at  $x = 3$ , that is, the height of the curve at  $x = 3$ , use:

```
dnorm(x = 3, mean = 2, sd = 5)
```

```
## [1] 0.07820854
```

# Probability in R

Or, to calculate the quantile for probability 0.975, use:

```
qnorm(p = 0.975, mean = 2, sd = 5)
```

```
## [1] 11.79982
```

```
dbinom(x = 6, size = 10, prob = 0.75)
```

```
## [1] 0.145998
```

# Probability in R

Command	Distribution
*binom	Binomial
*t	t
*pois	Poisson
*f	F
*chisq	Chi-Squared

# Hypothesis testing in R

- An overall model and related assumptions are made. (The most common being observations following a normal distribution.)
- The **null** ( $H_0$ ) and **alternative** ( $H_1$  or  $H_A$ ) hypothesis are specified. Usually the null specifies a particular value of a parameter.
- With given data, the **value** of the *test statistic* is calculated.
- Under the general assumptions, as well as assuming the null hypothesis is true, the **distribution** of the *test statistic* is known.
- Given the distribution and value of the test statistic, as well as the form of the alternative hypothesis, we can calculate a **p-value** of the test.
- Based on the **p-value** and pre-specified level of significance, we make a decision. One of:
  - Fail to reject the null hypothesis.
  - Reject the null hypothesis.

# One sample t-test

Suppose  $x_i \sim N(\mu, \sigma^2)$  and we want to test  $H_0 : \mu = \mu_0$  versus  $H_1 : \mu \neq \mu_0$ .

Assuming  $\sigma$  is unknown, we use the one-sample Student's  $t$  test statistic:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t_{n-1},$$

where  $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$  and  $s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$ .

A  $100(1 - \alpha)\%$  confidence interval for  $\mu$  is given by,

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

where  $t_{n-1}(\alpha/2)$  is the critical value such that  $P(t > t_{n-1}(\alpha/2)) = \alpha/2$  for  $n-1$  degrees of freedom.

# One sample t-test

```
capt_crisp = data.frame(weight = c(15.5, 16.2, 16.1, 15.8, 15.6, 16.0, 15.8, 15.9, 15.7))

x_bar = mean(capt_crisp$weight)
s      = sd(capt_crisp$weight)
mu_0   = 16
n      = 9

t = (x_bar - mu_0) / (s / sqrt(n))
t

## [1] -1.2

pt(t, df = n - 1)

## [1] 0.1322336
```

# One sample t-test

```
capt_test_results = t.test(capt_crisp$weight, mu = 16,
                           alternative = c("two.sided"), conf.level = 0.95)
```

This time we have stored the results. By doing so, we can directly access portions of the output from `t.test()`. To see what information is available we use the `names()` function.

```
names(capt_test_results)
```

```
## [1] "statistic"    "parameter"    "p.value"      "conf.int"     "estimate"
## [6] "null.value"   "stderr"        "alternative"  "method"       "data.name"
```

# Two sample t-test

Assume that the distributions of  $X$  and  $Y$  are  $N(\mu_1, \sigma^2)$  and  $N(\mu_2, \sigma^2)$ , respectively. Given the  $n = 6$  observations of  $X$ ,

```
x = c(70, 82, 78, 74, 94, 82)  
n = length(x)
```

and the  $m = 8$  observations of  $Y$ ,

```
y = c(64, 72, 60, 76, 72, 80, 84, 68)  
m = length(y)
```

we will test  $H_0 : \mu_1 = \mu_2$  versus  $H_1 : \mu_1 > \mu_2$ .

First, note that we can calculate the sample means and standard deviations.

# Two sample t-test

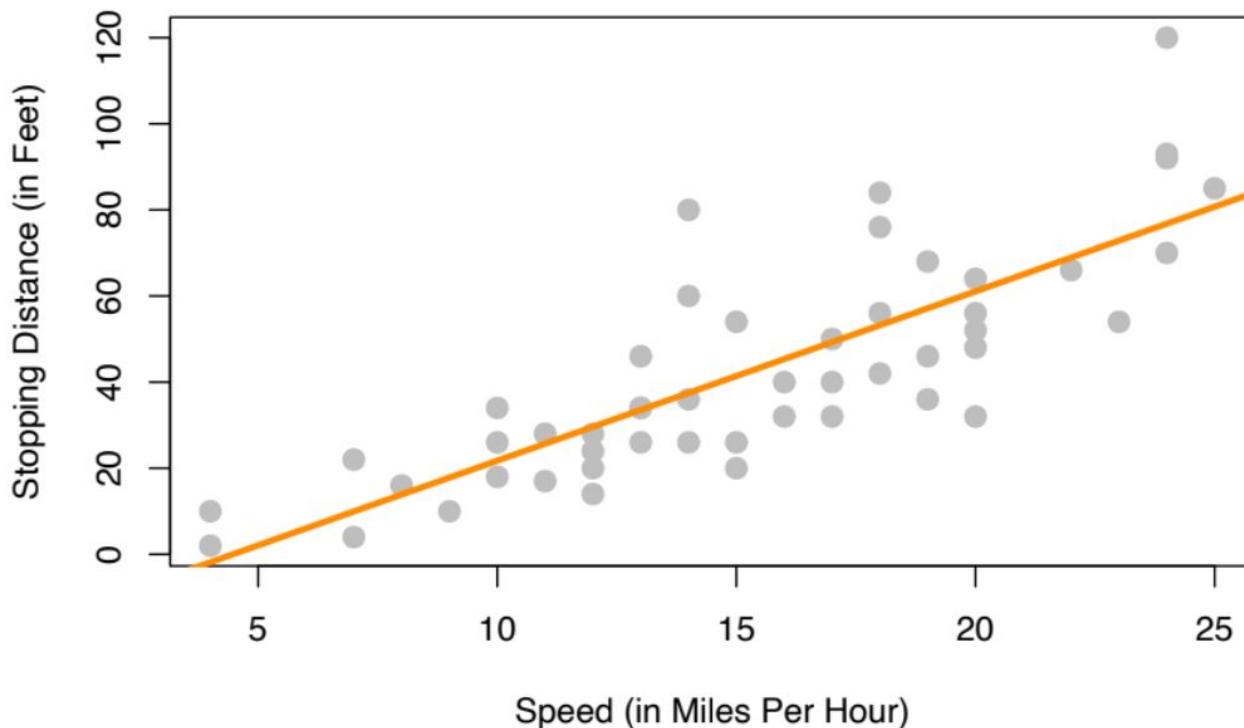
```
t.test(x, y, alternative = c("greater"), var.equal = TRUE)
```

```
##  
## Two Sample t-test  
##  
## data: x and y  
## t = 1.8234, df = 12, p-value = 0.04662  
## alternative hypothesis: true difference in means is greater than 0
```

# Linear Regression

$$Y = \beta_0 + \beta_1 X + \epsilon.$$

Stopping Distance vs Speed



```
x = cars$speed  
y = cars$dist
```

We then calculate the three sums of squares defined above.

```
Sxy = sum((x - mean(x)) * (y - mean(y)))  
Sxx = sum((x - mean(x))^ 2)  
Syy = sum((y - mean(y))^ 2)  
c(Sxy, Sxx, Syy)
```

```
## [1] 5387.40 1370.00 32538.98
```

```
beta_1_hat = Sxy / Sxx  
beta_0_hat = mean(y) - beta_1_hat * mean(x)  
c(beta_0_hat, beta_1_hat)
```

```
## [1] -17.579095 3.932409
```

# LR(Prediction)

```
unique(cars$speed)
```

```
## [1] 4 7 8 9 10 11 12 13 14 15 16 17 18 19 20 22 23 24 25
```

$$\hat{y} = -17.58 + 3.93 \times 50$$

```
beta_0_hat + beta_1_hat * 50
```

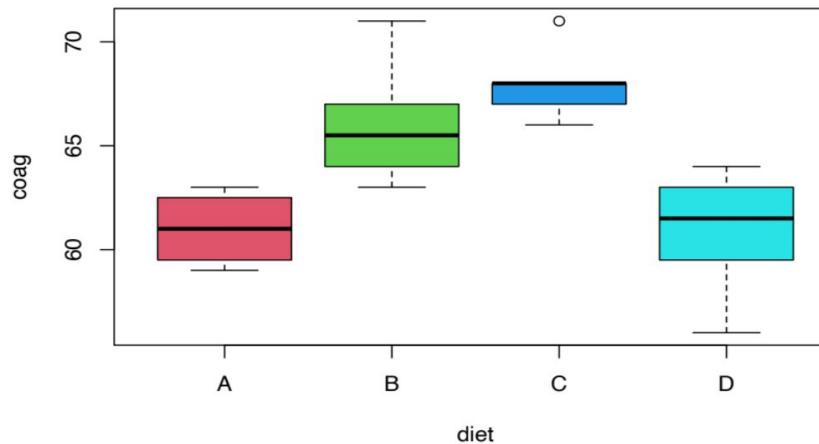
```
## [1] 179.0413
```

# One-way ANOVA

```
library(faraway)  
names(coagulation)
```

```
## [1] "coag" "diet"
```

```
plot(coag ~ diet, data = coagulation, col = 2:5)
```



```
coag_aov = aov(coag ~ diet, data = coagulation)
coag_aov
```

```
## Call:
##      aov(formula = coag ~ diet, data = coagulation)
##
## Terms:
##          diet Residuals
## Sum of Squares   228       112
## Deg. of Freedom    3        20
```

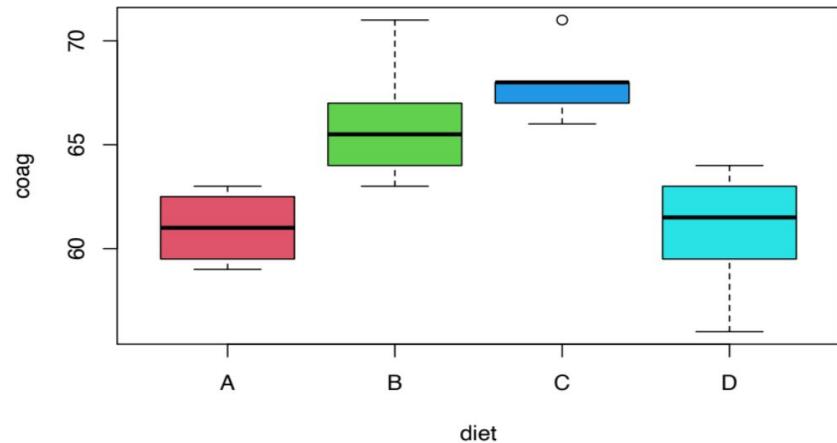
```
summary(coag_aov)
```

```
##          Df Sum Sq Mean Sq F value    Pr(>F)
## diet      3   228   76.0   13.57 4.66e-05 ***
## Residuals 20   112    5.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# One-way ANOVA

Suppose we reject the null hypothesis from the ANOVA test for equal means. That tells us that the means are different. But which means? All of them? Some of them? The obvious strategy is to test all possible comparisons of two means. We can do this easily in R.

## Post Hoc Testing



# Post Hoc Testing

```
with(coagulation, pairwise.t.test(coag, diet, p.adj = "none"))

##
##  Pairwise comparisons using t tests with pooled SD
##
## data: coag and diet
##
##      A          B          C
## B 0.00380  -         -
## C 0.00018  0.15878 -
## D 1.00000  0.00086 2.3e-05
##
## P value adjustment method: none
```

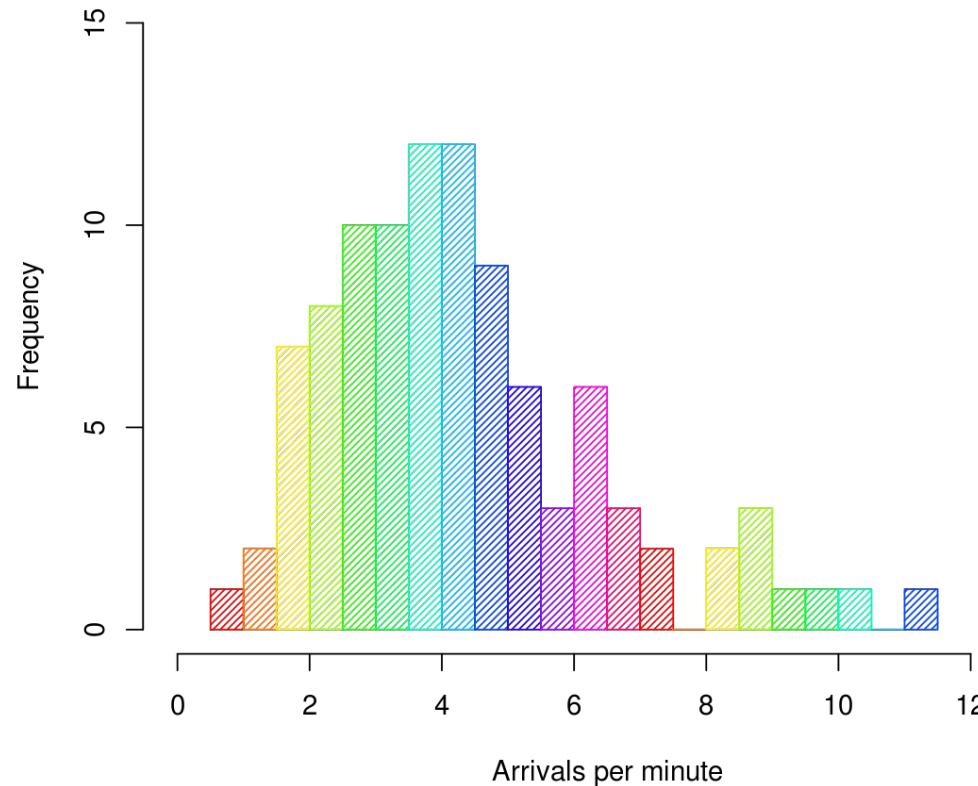
Also note that we are using the argument `p.adj = "none"`. What is this? An adjustment (in this case not an adjustment) to the p-value of each test. Why would we need to do this?

```
with(coagulation, pairwise.t.test(coag, diet, p.adj = "bonferroni"))

##
##  Pairwise comparisons using t tests with pooled SD
##
## data: coag and diet
##
##      A          B          C
## B 0.02282 -         -
## C 0.00108 0.95266 -
## D 1.00000 0.00518 0.00014
##
## P value adjustment method: bonferroni
```

# Distribution

Histogram of arrivals



# Shapiro-Wilk Test

Histograms and Q-Q Plots give a nice visual representation of the residuals distribution, however if we are interested in formal testing, there are a number of options available. A commonly used test is the **Shapiro-Wilk test**, which is implemented in R.

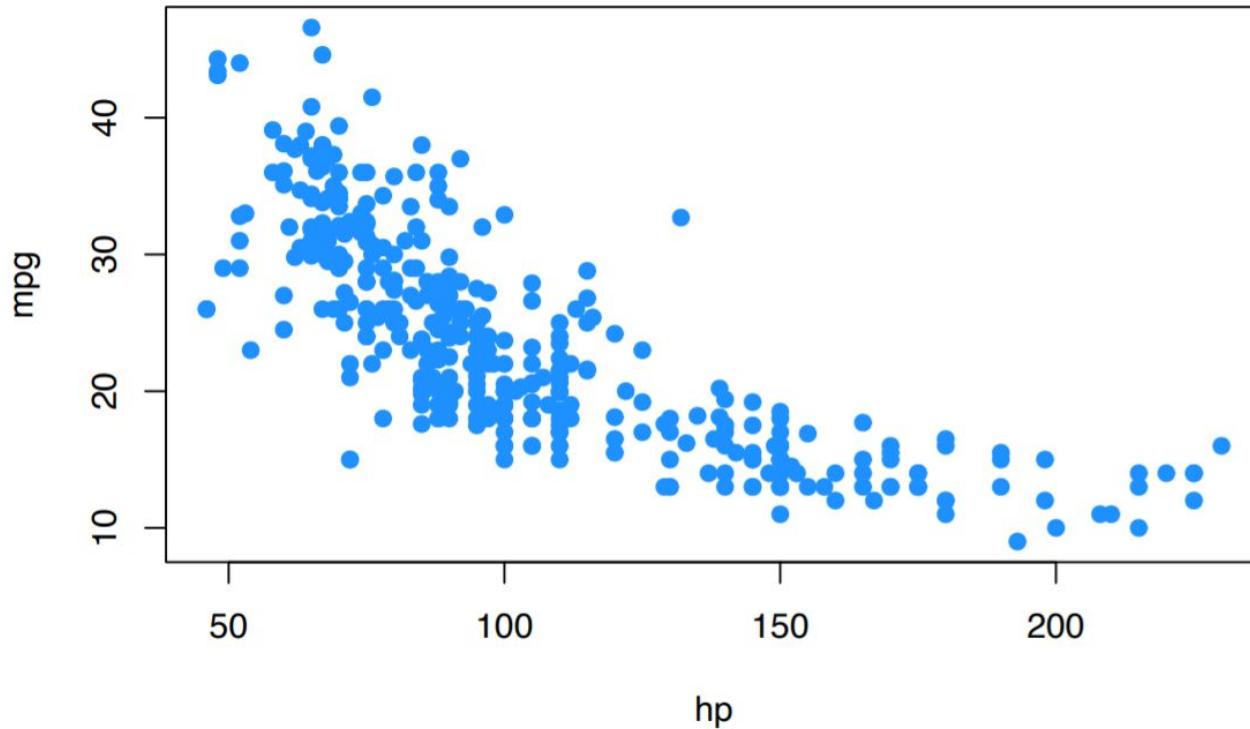
```
shapiro.test(rexp(25))
```

```
## Shapiro-Wilk normality test
##
## data: rexp(25)
## W = 0.71164, p-value = 1.05e-05
```

# Correlation

A word  
terms  
and co

re two  
*sation*



---

seatpos

---

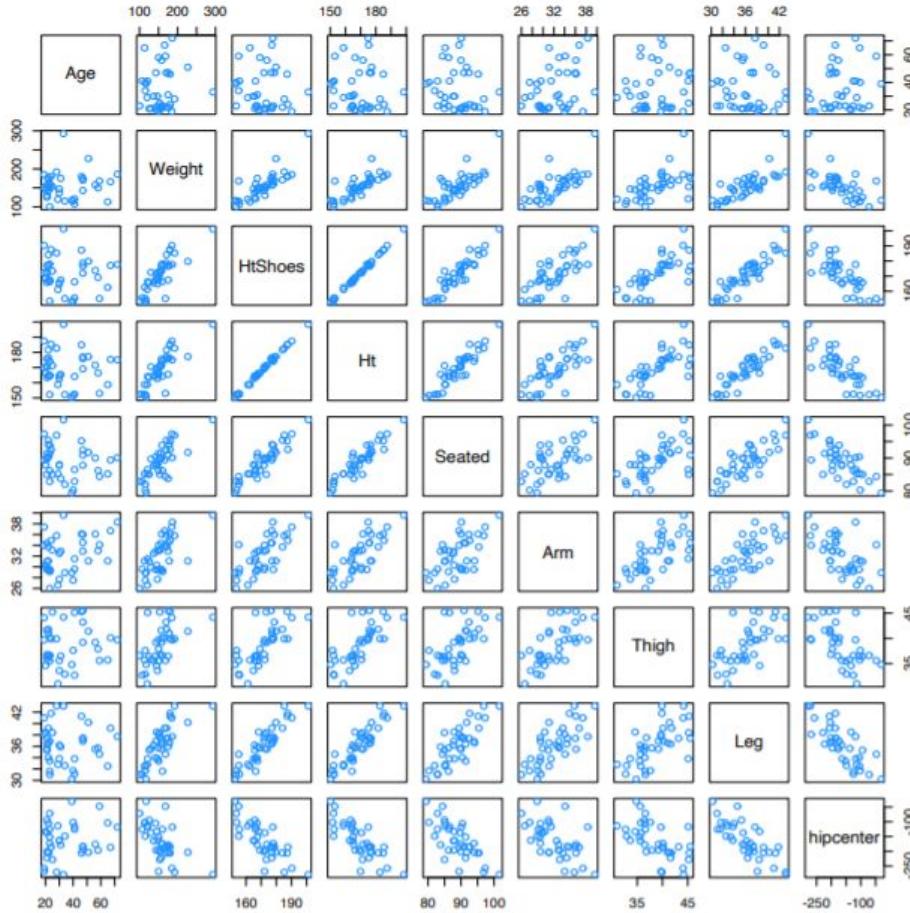
*Car seat position depending driver size*

---

## Description

Car drivers like to adjust the seat position for their own comfort. Car designers would find it helpful to know where different drivers will position the seat depending on their size and age. Researchers at the HuMoSim laboratory at the University of Michigan collected data on 38 drivers.

```
library(faraway)  
pairs(seatpos, col = "dodgerblue")
```



# Correlation

```
round(cor(seatpos), 2)
```

##	Age	Weight	HtShoes	Ht	Seated	Arm	Thigh	Leg	hipcenter
## Age	1.00	0.08	-0.08	-0.09	-0.17	0.36	0.09	-0.04	0.21
## Weight	0.08	1.00	0.83	0.83	0.78	0.70	0.57	0.78	-0.64
## HtShoes	-0.08	0.83	1.00	1.00	0.93	0.75	0.72	0.91	-0.80
## Ht	-0.09	0.83	1.00	1.00	0.93	0.75	0.73	0.91	-0.80
## Seated	-0.17	0.78	0.93	0.93	1.00	0.63	0.61	0.81	-0.73
## Arm	0.36	0.70	0.75	0.75	0.63	1.00	0.67	0.75	-0.59
## Thigh	0.09	0.57	0.72	0.73	0.61	0.67	1.00	0.65	-0.59
## Leg	-0.04	0.78	0.91	0.91	0.81	0.75	0.65	1.00	-0.79
## hipcenter	0.21	-0.64	-0.80	-0.80	-0.73	-0.59	-0.59	-0.79	1.00

# Chi-Squared

Command	Distribution
*binom	Binomial
*t	t
*pois	Poisson
*f	F
*chisq	Chi-Squared

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

$\chi^2$  = chi squared

$O_i$  = observed value

$E_i$  = expected value

### **Question 1**

Use the `iris` data frame.

We are interested to know if the Versicolor Irises have Sepals of a different width to Virginica Irises.

Carry out an appropriate statistical test.

### **Question 2**

Use the dataframe `sleep`

We are interested to know if the drug given affects the increase in sleep.

Carry out an appropriate hypothesis test. What conclusions can you draw?

Did you run into any problems along the way? How did you choose to address them?

(hint: look carefully at the patient ID column)

### **Question 3**

Install the package MASS

Access the package using

```
library(MASS)
```

inspect the data frame `chickwts` (this dataframe is part of the MASS package).

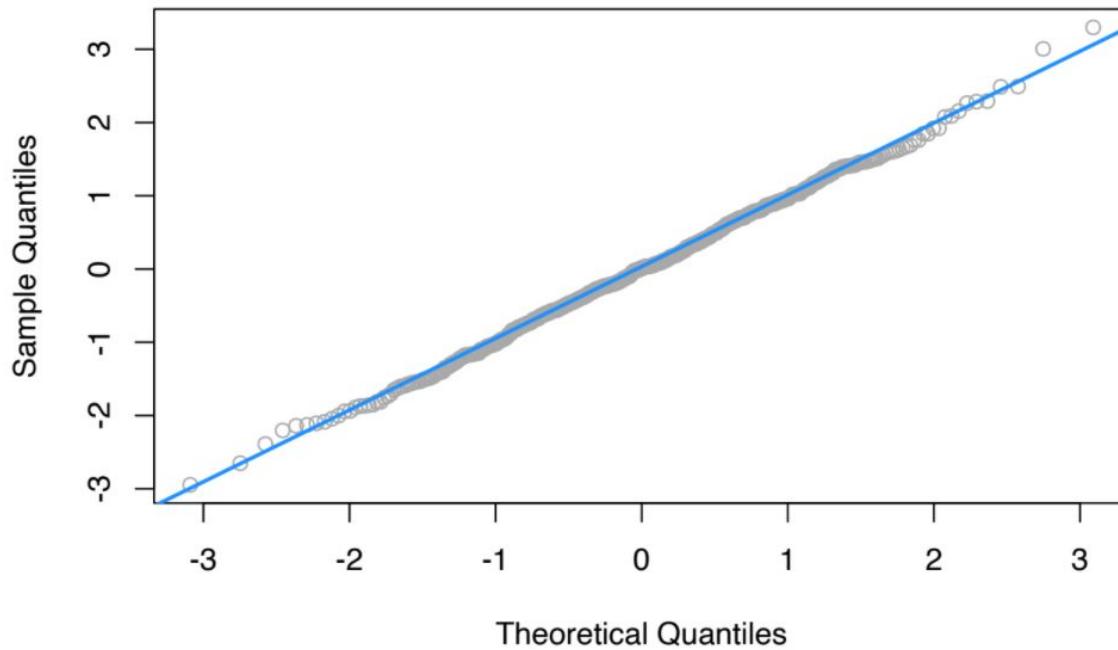
Carry out a test to see whether the average weight of chickens is affected by their feed.

# Q-Q plot

Another visual method for assessing the normality of errors, which is more powerful than a histogram, is a normal quantile-quantile plot, or **Q-Q plot** for short.

```
qqnorm(vector, main = "Normal Q-Q Plot, fit_1", col = "darkgrey")  
qqline(vector, col = "dodgerblue", lwd = 2)
```

### Normal Q-Q Plot, fit\_1

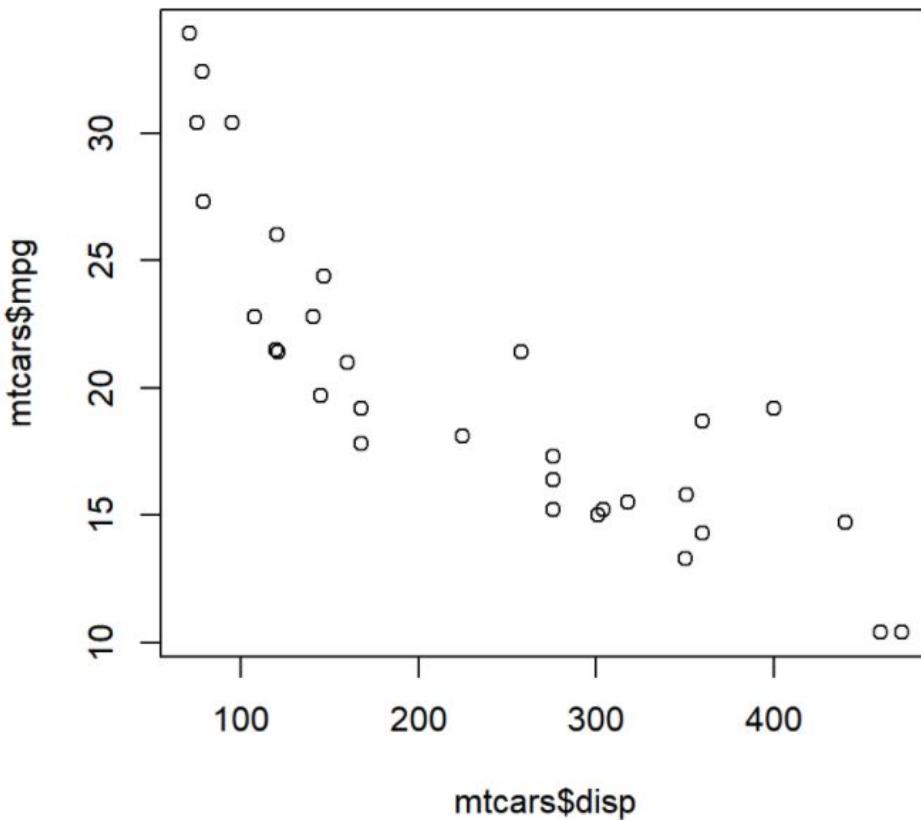


In short, if the points of the plot do not closely follow a straight line, this would suggest that the data do not come from a normal distribution.

# Principal Component Analysis

A:

	Attribute	Description	Type
In t	mpg	Miles Per Gallon	Continuous
use	cyl	Number of cylinders	Nominal
car	disp	Displacement	Continuous
car	hp	Horse Power	Continuous
car	drat	Real Wheel Axle Ratio	Continuous
	Wt	Weight	Continuous
	qsec	Time for 0.25 mile	Continuous
	vs	V/S	Nominal
	am	Transmission type	Nominal
	gear	Number of forward gears	Ordinal
	carb	Number of carburetors	Ordinal



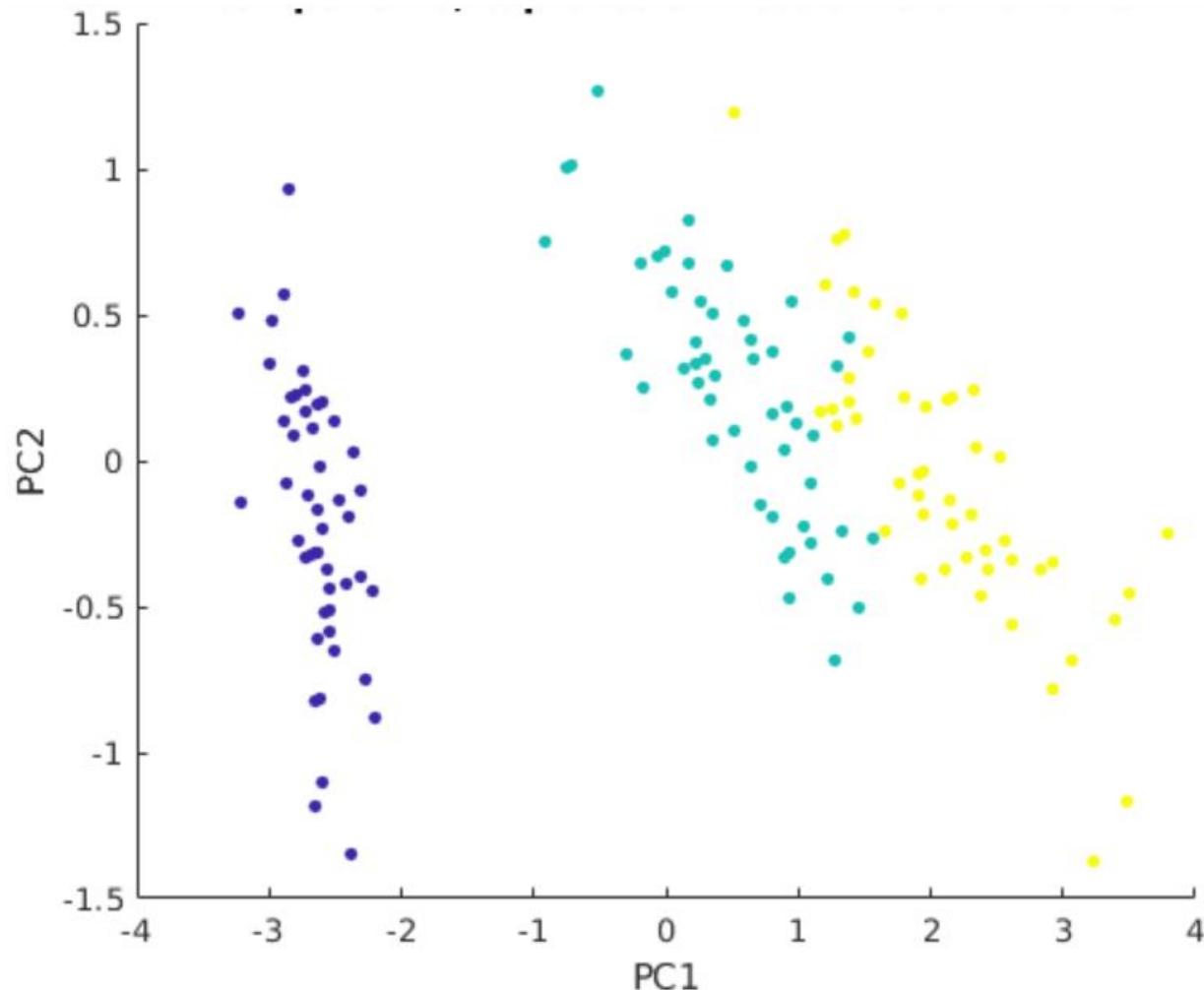
```
mtcars.pca <- prcomp(mtcars[,c(1:7,10,11)], center = TRUE,scale. = TRUE)

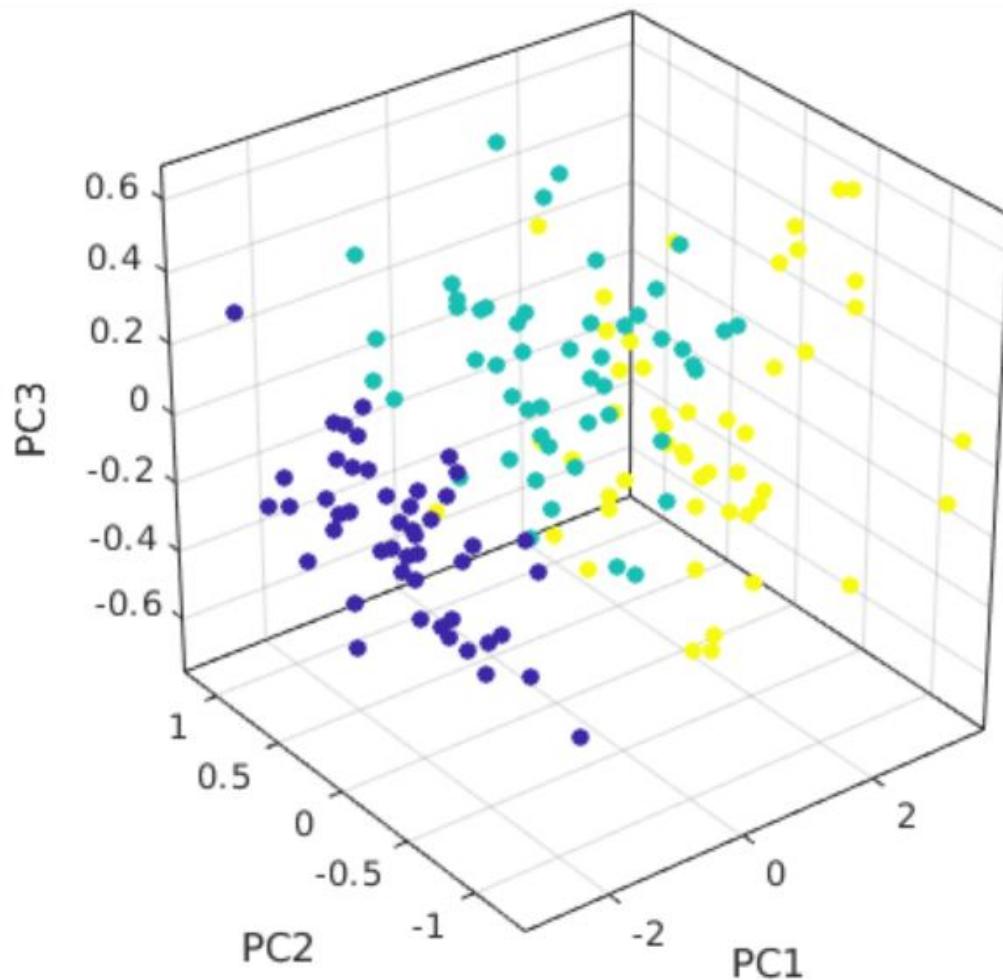
summary(mtcars.pca)

## Importance of components:
##                               PC1      PC2      PC3      PC4      PC5      PC6
## Standard deviation    2.3782  1.4429  0.71008 0.51481 0.42797 0.35184
## Proportion of Variance 0.6284  0.2313  0.05602 0.02945 0.02035 0.01375
## Cumulative Proportion  0.6284  0.8598  0.91581 0.94525 0.96560 0.97936
##                               PC7      PC8      PC9
## Standard deviation    0.32413 0.2419  0.14896
## Proportion of Variance 0.01167 0.0065  0.00247
## Cumulative Proportion  0.99103 0.9975  1.00000

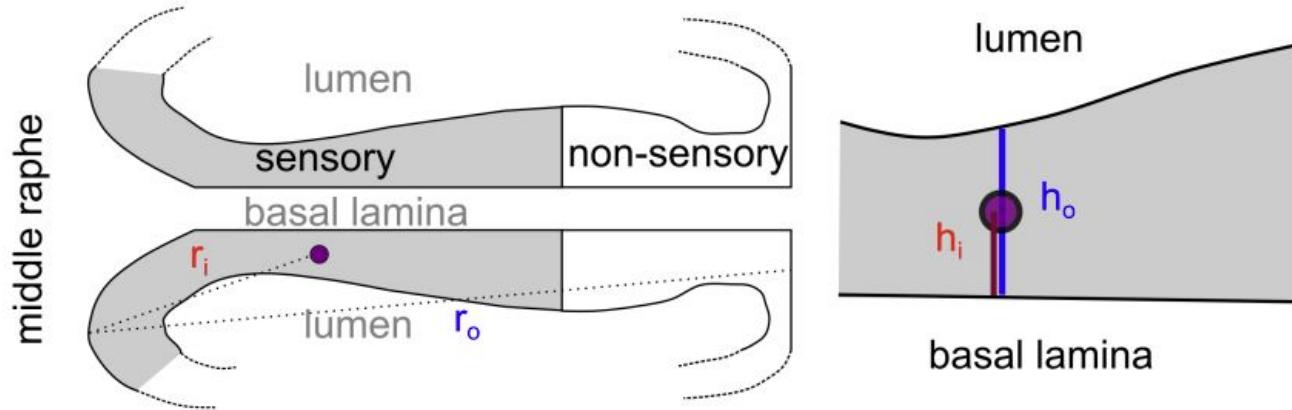
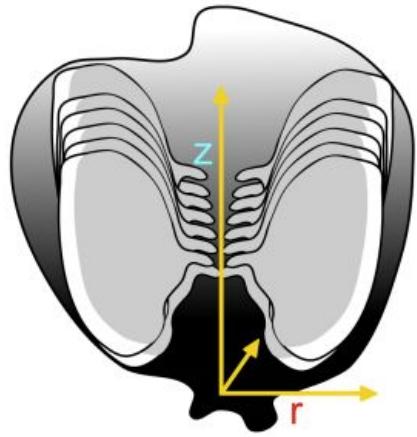
library(ggbiplot)

ggbiplot(mtcars.pca)
```

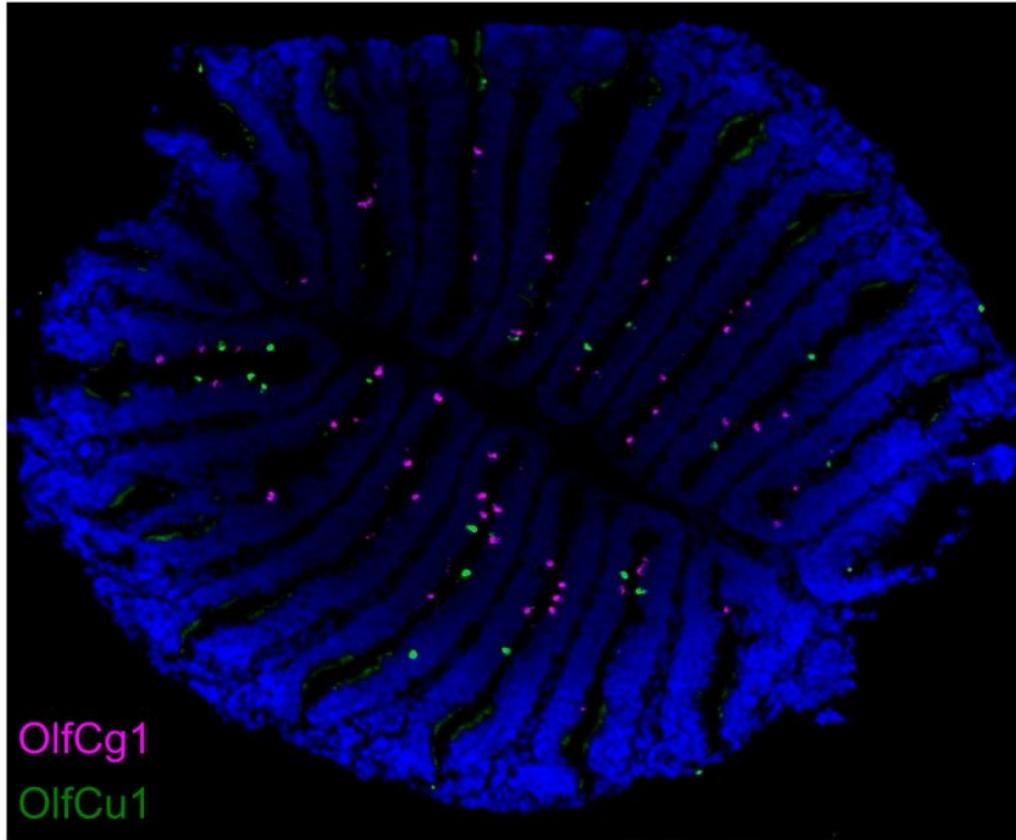


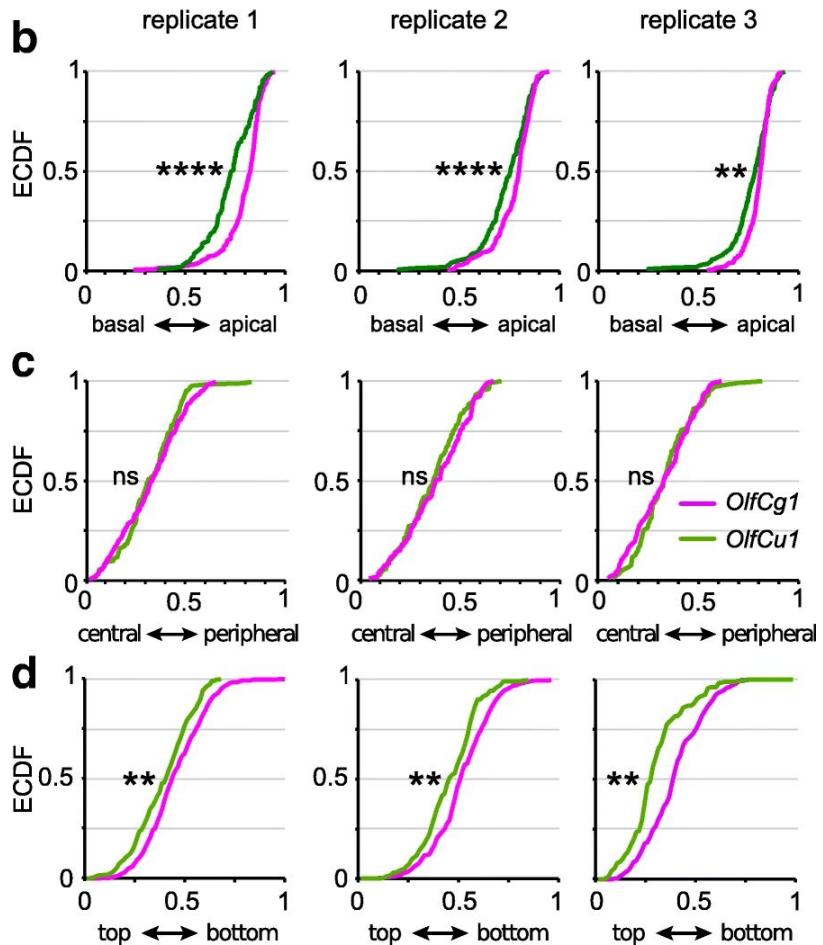


# **Differences Between Distributions**



**a**





# Some real world Problems & Explanation

## Exercise 1

Set a seed to 123 and plot a histogram of 1000 draws from a normal distribution with mean 10, standard deviation 2.

```
set.seed(seed)
```

Set the seed of R's random number generator, which is useful for creating simulations or random objects that can be reproduced.

## Usage

```
rnorm.acomp(n,mean,var)
rnorm.rcomp(n,mean,var)
rnorm.aplus(n,mean,var)
rnorm.rplus(n,mean,var)
rnorm.rmult(n,mean,var)
rnorm.ccomp(n,mean,var,lambda)
dnorm.acomp(x,mean,var,withJacobian=FALSE)
dnorm.aplus(x,mean,var,withJacobian=FALSE)
dnorm.rmult(x,mean,var)
```

```
set.seed(123)
normal_draws <- rnorm(1000, mean = 10, sd = 2)
hist(normal_draws)
```

## Some real world Problems & Explanation

### Exercise 2

Using a QQ plot. Assess the normality of your previously simulated draws.

The function `qnorm()`, which comes standard with R, aims to do: given an area, find the boundary value that determines this area.

```
qqnorm(normal_draws)  
qqline(normal_draws)
```

## Some real world Problems & Explanation

### Exercise 3

Using a t-test, test for a difference in means between your samples. 1000 samples of the Student's t-distribution, 10 degrees of freedom, and a delta value of 9. Report on your p-value and its significance at the 5% level.

#### Rt: Pseudo-Random Number Generation From T-Distribution

Hint: compare normal\_draws vs new T-distribution dataset

```
y <- rt(1000, df = 10, ncp = 9)
means_t_test <- t.test(normal_draws, y)
means_t_test$p.value

## [1] 0.03064702
```

## Some real world Problems & Explanation

### Exercise 4

Rewrite your t-test, testing now if your normal samples have a greater mean than your samples from the Student's t-distribution. Report on your new p-value.

```
greater_test <- t.test(normal_draws, y,  
alternative = "greater")  
greater_test$p.value  
  
## [1] 0.01532351
```

## Some real world Problems & Explanation

### Exercise 5

Putting these skills together now, calculate a two-sided t-test of equal means from two normal distributions. The first of mean 1, standard deviation 0.5, the second of mean 0.9, a standard deviation of 1. Hint: A function may become useful here.

```
test_func <- function(x, y) {  
  t_test <- t.test(x, y)  
  return(t_test$p.value)  
}  
test_func(x = rnorm(100, mean = 1, sd = 0.5), y = rnorm(100, mean = 0.9, sd = 1))  
## [1] 0.2236878
```

## Some real world Problems & Explanation

### Exercise 6

Replicate this t-test 1000 times and test for a standard uniform distribution, using a QQ plot.

```
p_reps <- replicate(1000, test_func(x = rnorm(100, mean = 1, sd = 0.5), y = rnorm(100, mean = 0.9, sd = 1)))
qqplot(p_reps, runif(1000))
abline(0,1)
```

# **Statistics Using R**

## **with Biological Examples**

# Qualitative Data

```
> library(MASS)      # load the MASS package
> painters
  Composition Drawing Colour Expression School
Da Udine          10      8    16        3     A
Da Vinci          15     16      4       14     A
Del Piombo         8     13    16        7     A
Del Sarto          12     16      9        8     A
Fr. Penni           0     15      8        0     A
Guilio Romano     15     16      4       14     A
.....
```

```
> painters$School
> help(painters)
```

Let's do the following

- Frequency Distribution of Qualitative Data
- Relative Frequency Distribution of Qualitative Data
- Bar Graph
- Pie Chart
- Category Statistics

# frequency distribution

```
> library(MASS)          # load the MASS package
> school = painters$School    # the painter schools
> school.freq = table(school)  # apply the table function

> school.freq
school
A   B   C   D   E   F   G   H
10   6   6  10   7   4   7   4

> cbind(school.freq)
  school.freq
A           10
B            6
C            6
D           10
E            7
F            4
G            7
H            4
```

# frequency distribution

```
> library(MASS)          # load the MASS package
> school = painters$School    # the painter schools
> school.freq = table(school)  # apply the table function
```

Then we find the sample size of painters with the nrow function, and divide the frequency distribution with it. Therefore the relative frequency distribution is:

```
> school.relfreq = school.freq / nrow(painters)

> school.relfreq
school
      A      B      C      D      E      F
0.185185 0.111111 0.111111 0.185185 0.129630 0.074074
      G      H
0.129630 0.074074

> old = options(digits=1)
> school.relfreq
school
      A      B      C      D      E      F      G      H
0.19  0.11  0.11  0.19  0.13  0.07  0.13  0.07
> options(old)
```

APPLY cbind function !

# Relative Frequency Distribution of Qualitative Data

```
> library(MASS)          # load the MASS package
> school = painters$School    # the painter schools
> school.freq = table(school)  # apply the table function
```

Then we find the sample size of painters with the nrow function, and divide the frequency distribution with it. Therefore the relative frequency distribution is:

```
> school.relfreq = school.freq / nrow(painters)

> school.relfreq
school
      A      B      C      D      E      F
0.185185 0.111111 0.111111 0.185185 0.129630 0.074074
      G      H
0.129630 0.074074

> old = options(digits=1)
> school.relfreq
school
      A      B      C      D      E      F      G      H
0.19  0.11  0.11  0.19  0.13  0.07  0.13  0.07
> options(old)
```

APPLY cbind function !

# Bar Graph

```
> library(MASS)          # load the MASS package
> school = painters$School    # the painter schools
> school.freq = table(school)  # apply the table function

> barplot(school.freq)      # apply the barplot function
```

```
#####
# colors = c("red", "yellow", "green", "violet",
+ "orange", "blue", "pink", "cyan")
> barplot(school.freq,        # apply the barplot function
+ col=colors)                # set the color palette
```

# Category Statistics

Suppose we would like to know which school has the highest mean composition score. We would have to first find out the mean composition score of each school. The following shows how to find the mean composition score of an arbitrarily chosen school.

1. Create a logical index vector for school C.

```
> library(MASS)           # load the MASS package  
> school = painters$School    # the painter schools  
> c_school = school == "C"      # the logical index vector
```

2. Find the child data set of painters for school C. For explanation, please consult the tutorial of [Data Frame Row Slice](#).

```
> c_painters = painters[c_school, ]  # child data set
```

3. Find the mean composition score of school C.

```
> mean(c_painters$Composition)  
[1] 13.167
```

# Category Statistics

## Alternative Solution

Instead of computing the mean composition score manually for each school, use the tapply function to compute them all at once.

```
> tapply(painters$Composition, painters$School, mean)
```

	A	B	C	D	E	F	G	H
mean	10.400	12.167	13.167	9.100	13.571	7.250	13.857	14.000

# Quantitative Statistics

**Quantitative data**, also known as **continuous** data, consists of numeric data that support arithmetic operations. This is in contrast with **qualitative data**, whose values belong to pre-defined classes with no arithmetic operation allowed.

```
> head(faithful)
  eruptions waiting
1      3.600      79
2      1.800      54
3      3.333      74
4      2.283      62
5      4.533      85
6      2.883      55

> duration = faithful$eruptions
> range(duration)
[1] 1.6 5.1

> breaks = seq(1.5, 5.5, by=0.5)      # half-integer sequence
> breaks
[1] 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0 5.5
```

Classify the eruption durations according to the half-unit-length sub-intervals with `cut`. As the intervals are to be closed on the left, and open on the right, we set the `right` argument as `FALSE`.

```
> duration.cut = cut(duration, breaks, right=FALSE)
```

Compute the frequency of eruptions in each sub-interval with the `table` function.

```
> duration.freq = table(duration.cut)
```

# Histogram

```
> duration = faithful$eruptions  
> hist(duration,      # apply the hist function  
+   right=FALSE)    # intervals closed on the left
```

To colorize the histogram, we select a color palette and set it in the col argument of hist. In addition, we update the titles for readability.

```
> colors = c("red", "yellow", "green", "violet", "orange",  
+   "blue", "pink", "cyan")  
  
> hist(duration,      # apply the hist function  
+   right=FALSE,      # intervals closed on the left  
+   col=colors,        # set the color palette  
+   main="Old Faithful Eruptions", # the main title  
+   xlab="Duration minutes")       # x-axis label
```

# Relative Frequency Distribution of Quantitative Data

We first find the frequency distribution of the eruption durations as follows. Further details can be found in the [Frequency Distribution](#) tutorial.

```
> duration = faithful$eruptions  
> breaks = seq(1.5, 5.5, by=0.5)  
> duration.cut = cut(duration, breaks, right=FALSE)  
> duration.freq = table(duration.cut)
```

Then we find the sample size of faithful with the nrow function, and divide the frequency distribution with it. As a result, the relative frequency distribution is:

```
> duration.relfreq = duration.freq / nrow(faithful)
```

# Cumulative Frequency Distribution

We first find the frequency distribution of the eruption durations as follows. Further details can be found in the [Frequency Distribution](#) tutorial.

```
> duration = faithful$eruptions  
> breaks = seq(1.5, 5.5, by=0.5)  
> duration.cut = cut(duration, breaks, right=FALSE)  
> duration.freq = table(duration.cut)
```

We then apply the cumsum function to compute the cumulative frequency distribution.

```
> duration.cumfreq = cumsum(duration.freq)
```

## Answer

The cumulative distribution of the eruption duration is:

```
> duration.cumfreq  
[1.5,2) [2,2.5) [2.5,3) [3,3.5) [3.5,4) [4,4.5) [4.5,5)  
51 92 97 104 134 207 268  
[5,5.5)  
272
```

# Stem-and-Leaf Plot

```
> duration = faithful$eruptions  
> stem(duration)
```

The decimal point is 1 digit(s) to the left of the |

```
16 | 070355555588  
18 | 000022233333335577777777888822335777888  
20 | 00002223378800035778  
22 | 0002335578023578  
24 | 00228  
26 | 23  
28 | 080  
30 | 7  
32 | 2337  
34 | 250077  
36 | 0000823577  
38 | 2333335582225577  
40 | 0000003357788888002233555577778  
42 | 03335555778800233333555577778  
44 | 02222335557780000000023333357778888  
46 | 0000233357700000023578  
48 | 00000022335800333  
50 | 0370
```

# Stem-and-Leaf Plot

```
> duration = faithful$eruptions      # the eruption durations
> waiting = faithful$waiting        # the waiting interval
> head(cbind(duration, waiting))
   duration waiting
[1,]    3.600     79
[2,]    1.800     54
[3,]    3.333     74
[4,]    2.283     62
[5,]    4.533     85
[6,]    2.883     55
```

We apply the plot function to compute the scatter plot of eruptions and waiting.

```
> duration = faithful$eruptions      # the eruption durations
> waiting = faithful$waiting        # the waiting interval
> plot(duration, waiting,           # plot the variables
+       xlab="Eruption duration",    # x-axis label
+       ylab="Time waited")         # y-axis label
```

We can generate a linear regression model of the two variables with the lm function, and then draw a trend line with abline.

```
> abline(lm(waiting ~ duration))
```

# RUV-Seq

In this document, we show how to conduct a differential expression (DE) analysis that controls for “unwanted variation”, e.g., batch, library preparation, and other nuisance effects, using the between-sample normalization methods

- *RUVg uses negative control genes, assumed to have constant expression across samples;*
- *RUVs uses centered (technical) replicate/negative control samples for which the covariates of interest are constant;*
- *RUVr uses residuals, e.g., from a first-pass GLM regression of the counts on the covariates of interest.*

```
library(RUVSeq) # Install it if not installed.  
  
#also install ggplot2 and DESeq2 packages  
  
library(zebrafishRNASeq)  
  
data(zfGenes)  
  
head(zfGenes)  
  
tail(zfGenes)  
  
##Filtering and exploratory data analysis
```

We filter out non-expressed genes, by requiring more than 5 reads in at least two samples for each gene.

```
filter <- apply(zfGenes, 1, function(x) length(x[x>5])>=2)  
  
filtered <- zfGenes[filter,] genes <- rownames(filtered)[grep("^ENS", rownames(filtered))]  
  
spikes <- rownames(filtered)[grep("^ERCC", rownames(filtered))]
```

Don't forget to use dim command before and after filtering!!!

# RUV-Seq

```
x <- as.factor(rep(c("Ctl", "Trt"), each=3))

set <- newSeqExpressionSet(as.matrix(filtered), phenoData = data.frame(x, row.names=colnames(filtered)))

Set

The boxplots of relative log expression (RLE = log-ratio of read count to median read count across sample) and plots of
principal components (PC)

library(RColorBrewer)

colors <- brewer.pal(3, "Set2") plotRLE(set, outline=FALSE, ylim=c(-4, 4), col=colors[x])
plotPCA(set, col=colors[x], cex=1.2)

We can use the betweenLaneNormalization function of EDASeq to normalize the data using upper-quartile (UQ) normalization

set <- betweenLaneNormalization(set, which="upper")

plotRLE(set, outline=FALSE, ylim=c(-4, 4), col=colors[x])
plotPCA(set, col=colors[x], cex=1.2)
```

# RUVg: Estimating the factors of unwanted variation using control genes

To estimate the factors of unwanted variation, we need a set of negative control genes.

Here, we use the ERCC spike-ins as controls and we consider  $k = 1$  factors of unwanted variation

```
set1 <- RUVg(set, spikes, k=1)

pData(set1)

plotRLE(set1, outline=FALSE, ylim=c(-4, 4), col=colors[x])

plotPCA(set1, col=colors[x], cex=1.2)
```

# RUVg: Estimating the factors of unwanted variation using control genes

```
library(DESeq2)

dds <- DESeqDataSetFromMatrix(countData = counts(set1), colData = pData(set1), design = ~ W_1 + x)
```

```
dds <- DESeq(dds) res <- results(dds)
```

```
Res
```

Note that this will perform by default a Wald test of significance of the last variable in the design formula, in this case  $x$ . If one wants to perform a likelihood ratio test, she needs to specify a reduced model that includes  $W$  (see the DESeq2 vignette for more details on the test statistics).

```
dds <- DESeq(dds, test="LRT", reduced=as.formula(~ W_1))
```

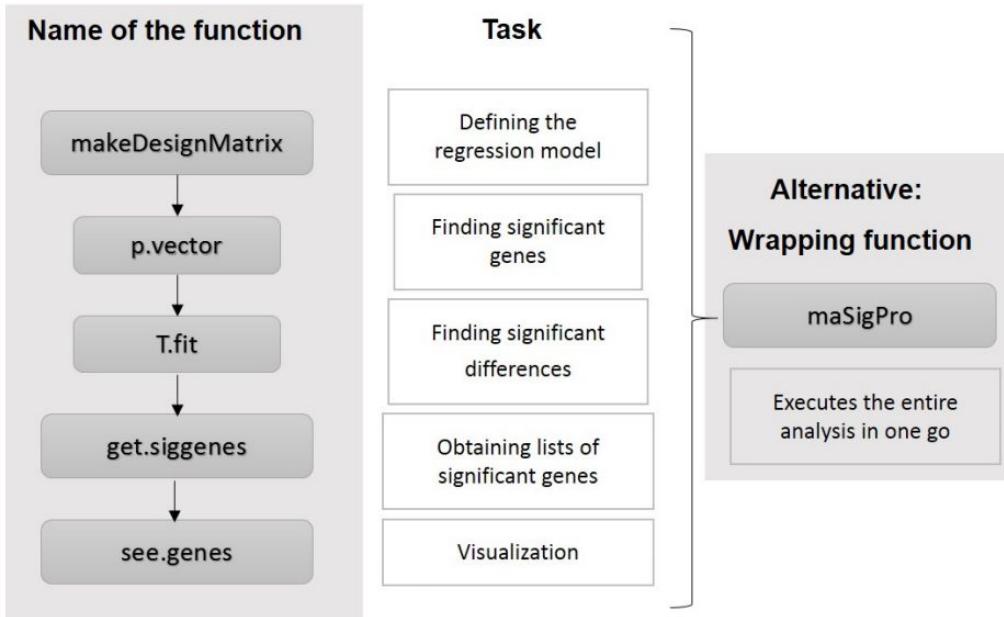
```
res <- results(dds)
```

```
res
```

# Time Series Data Analysis in R

The maSigPro package can be obtained from the Bioconductor repository.

```
> library(maSigPro) # load maSigPro library  
>help(package="maSigPro") #for package help  
>?p.vector #for function help
```



# Time Series Data Analysis in R

The maSigPro package can be obtained from the Bioconductor repository.

```
> data(data.abiotic)
> data(edesign.abiotic)
#The edesign.abiotic object describes the experimental design of this experiment in maSigPro format.
```

```
> edesign.abiotic
> colnames(data.abiotic)
> rownames(edesign.abiotic)
> colnames(edesign.abiotic)
> rownames(data.abiotic)
```

Defining the regression model

```
design <- make.design.matrix(edesign.abiotic, degree = 2)
> design$groups.vector
```

Finding significant genes

```
##The next step is to compute a regression fit for each gene. This is done by the function p.vector(). This function also
computes the p-value associated to the F-Statistic of the model, which is used to select significant genes
```

```
> fit <- p.vector(data.abiotic, design, Q = 0.05, MT.adjust = "BH", min.obs = 20)
```

# Time Series Data Analysis in R

p.vector() returns a list of values:

```
> fit$i # returns the number of significant genes  
  
> fit$alfa # gives p-value at the Q false discovery control level  
  
> fit$SELEC # is a matrix with the significant genes and their expression values
```

Finding significant differences

```
tstep <- T.fit(fit, step.method = "backward", alfa = 0.05)
```

```
##T.fit() executes stepwise regression. The step.method can be "backward" or "forward" indicating whether the step procedure starts from the model with all or none variables.
```

For each selected gene the following values are given:

1. p-value of the regression ANOVA
2. R-squared of the model
3. p-value of the regression coefficients of the selected variables

# Time Series Data Analysis in R

Obtaining lists of significant genes

This is done by the function `get.siggenes()`.

This function has two major arguments, `rsq` and `vars`.

`rsq`: is a cutt-off value for the R-squared of the regression model.

`vars`: is used to indicate how to group variables to show results.

```
> sigs <- get.siggenes(tstep, rsq = 0.6, vars = "groups")

> names(sigs)

> names(sigs$sig.genes) [

> names(sigs$sig.genes$ColdvsControl)

> suma2Venn(sigs$summary[, c(2:4)])

> suma2Venn(sigs$summary[, c(1:4)])

> see.genes() #Use see.genes() to visualize the result of a group of genes, for example, to visualize the significant genes obtained as significant in the previous step in ControlvsSalt

> see.genes(sigs$sig.genes$ColdvsControl, show.fit = T, dis =design$dis, cluster.method="hclust" ,cluster.data = 1, k = 9)
```

# Time Series Data Analysis in R

**k:** number of clusters for data partitioning. By default it is 9. Mclust cluster method can compute an optimal k, choosing k.mclust=TRUE.

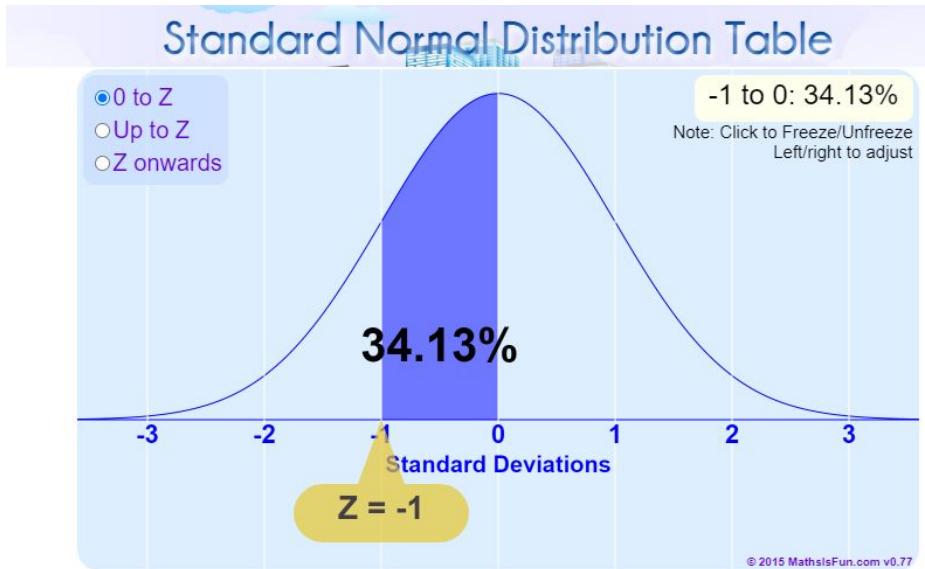
**cluster.method:** clustering method for data partitioning. hclust, kmeans and Mclust are supported.

**distance:** distance measurement function when cluster.method is hclust. By default it is 'cor' to compute a distance based on the correlation because we are interested in similar trends or changes.

**agglo.method:** aggregation method used when cluster.method is hclust. By default ward.D.

```
> STMDE66 <- data.abiotic[rownames(data.abiotic)=="STMDE66",]  
  
> PlotGroups (STMDE66, edesign = edesign.abiotic)  
  
> PlotGroups (STMDE66, edesign = edesign.abiotic, show.fit = T, + dis = design$dis, groups.vector =  
design$groups.vector)
```

# Standard Normal Distribution



This is the "bell-shaped" curve of the Standard Normal Distribution.

It is a [Normal Distribution](#) with mean 0 and [standard deviation](#) 1.

It shows you the percent of population:

- between 0 and Z (option "0 to Z")
- less than Z (option "Up to Z")
- greater than Z (option "Z onwards")

It only display values to 0.01%

<https://www.mathsisfun.com/data/standard-normal-distribution-table.html>

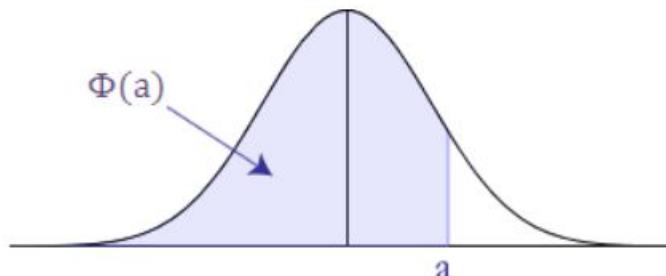
# How to do Normal Distributions Calculations

The most common form of standard normal distribution table that you see is a table similar to the one below:

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5357	0.5398
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753	0.5793
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141	0.6179
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6444	0.6481	0.6518	0.6554
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879	0.6915
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7191	0.7224	0.7257
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7548	0.7578
0.7	0.7578	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7793	0.7821	0.7849	0.7876
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8105	0.8131	0.8156

# How to do Normal Distributions Calculations

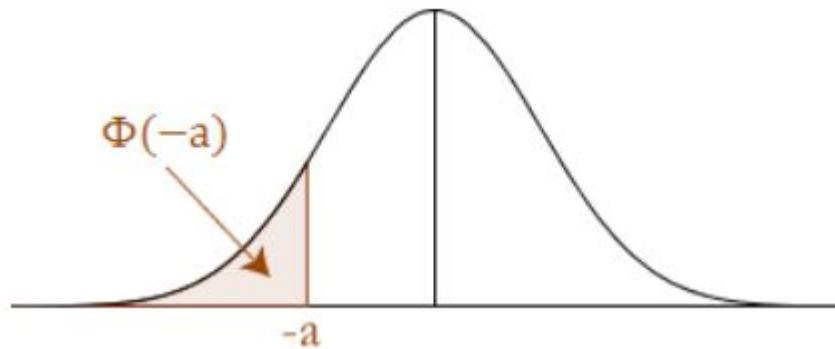
- The standard normal distribution table provides the probability that a normally distributed random variable  $Z$ , with mean equal to 0 and variance equal to 1, is less than or equal to  $z$ .
- It does this for positive values of  $z$  only (i.e.,  $z$ -values on the right-hand side of the mean).
- What this means in practice is that if someone asks you to find the probability of a value being less than a specific, positive  $z$ -value, you can simply look that value up in the table.
- We call this area  $\Phi$ . Thus, for this table,  $P(Z < a) = \Phi(a)$ , where  $a$  is positive.



Probability less than a z-value

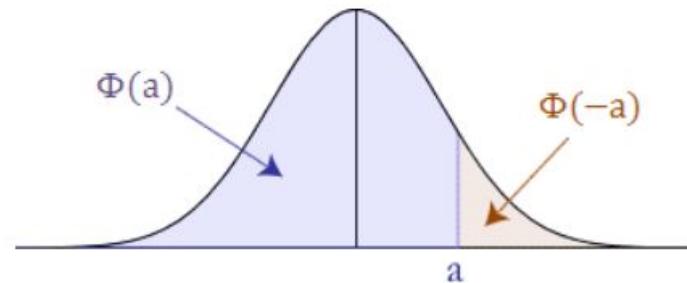
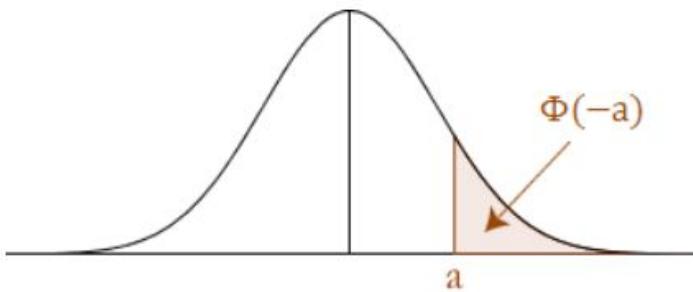
$$P(Z < -a)$$

As explained above, the standard normal distribution table only provides the probability for values less than a positive z-value (i.e., z-values on the right-hand side of the mean). So how do we calculate the probability below a negative z-value (as illustrated below)?



# How to do Normal Distributions Calculations

- We start by remembering that the standard normal distribution has a total area (probability) equal to 1 and it is also symmetrical about the mean.
- Thus, we can do the following to calculate negative z-values: we need to appreciate that the area under the curve covered by  $P(Z > a)$  is the same as the probability less than  $-a$   $\{P(Z < -a)\}$  as illustrated below:



# How to do Normal Distributions Calculations

From this illustration, and from our knowledge that the area under the standard normal distribution is equal to 1, we can conclude that the two areas add up to 1. We can, therefore, make the following statements:

$$\Phi(a) + \Phi(-a) = 1$$

$$\therefore \Phi(-a) = 1 - \Phi(a)$$

Thus, we know that to find a value less than a negative z-value we use the following equation:

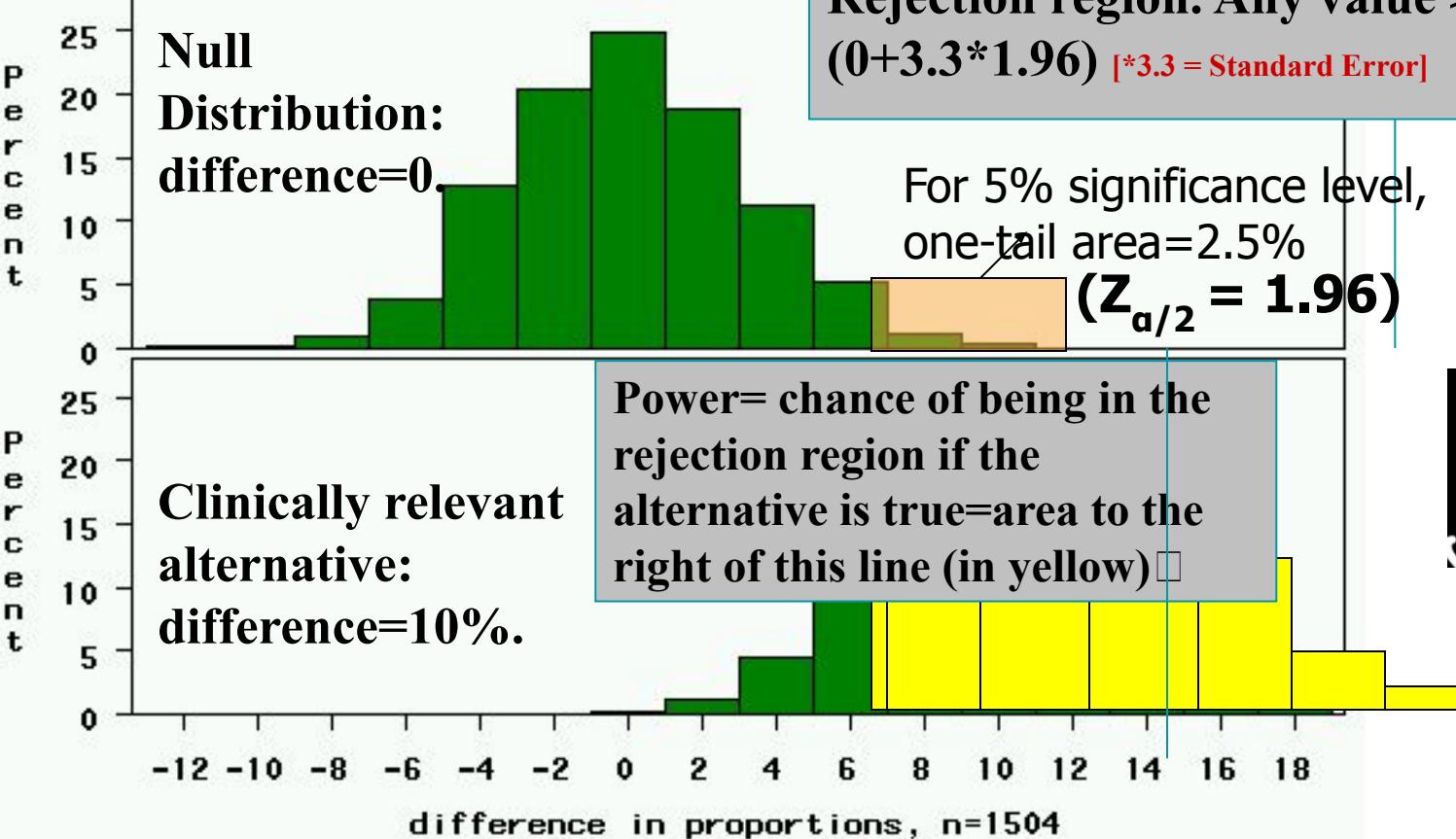
$$\Phi(-a) = 1 - \Phi(a), \quad \text{e.g. } \Phi(-1.43) = 1 - \Phi(1.43)$$

# Introduction to sample size and power calculations

How much chance do we have to reject the null hypothesis when the alternative is in fact true?  
(what's the probability of detecting a real effect?)

Can we quantify how much power we have for given sample sizes?

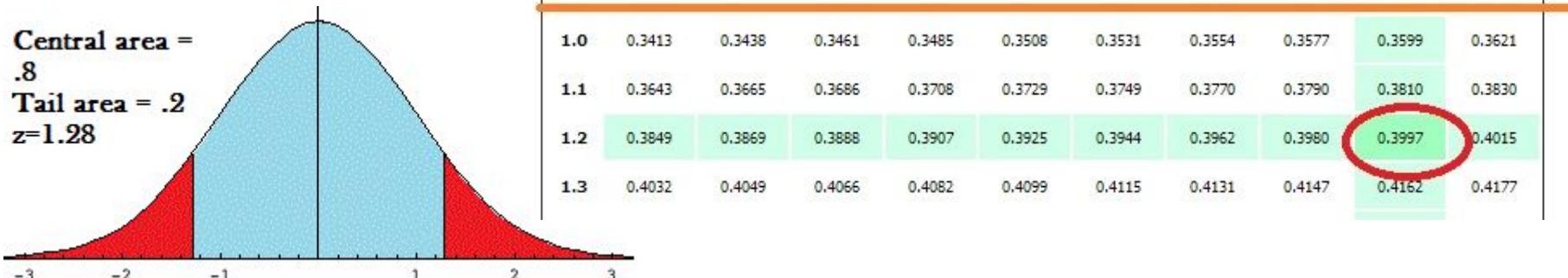
# study 1: 263 cases, 1241 controls



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

# What is a critical value?

- A critical value is the value of the test statistic which defines the upper and lower bounds of a confidence interval, or which defines the threshold of statistical significance in a statistical test. It describes how far from the mean of the distribution you have to go to cover a certain amount of the total variation in the data (i.e. 90%, 95%, 99%).
- If you are constructing a 95% confidence interval and are using a threshold of statistical significance of  $p = 0.05$ , then your critical value will be identical in both cases.
- A critical value is a line on a graph that splits the graph into sections. One or two of the sections is the “rejection region”; if your test value falls into that region, then you reject the null hypothesis.



study 1: 263 cases, 1241 controls

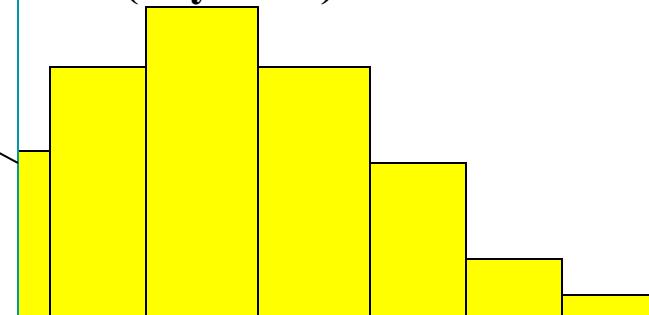
Rejection region.

Any value  $\geq 6.5$

$$(0+3.3*1.96)$$

Power= chance of being in the rejection region if the alternative is true=area to the right of this

line (in yellow)

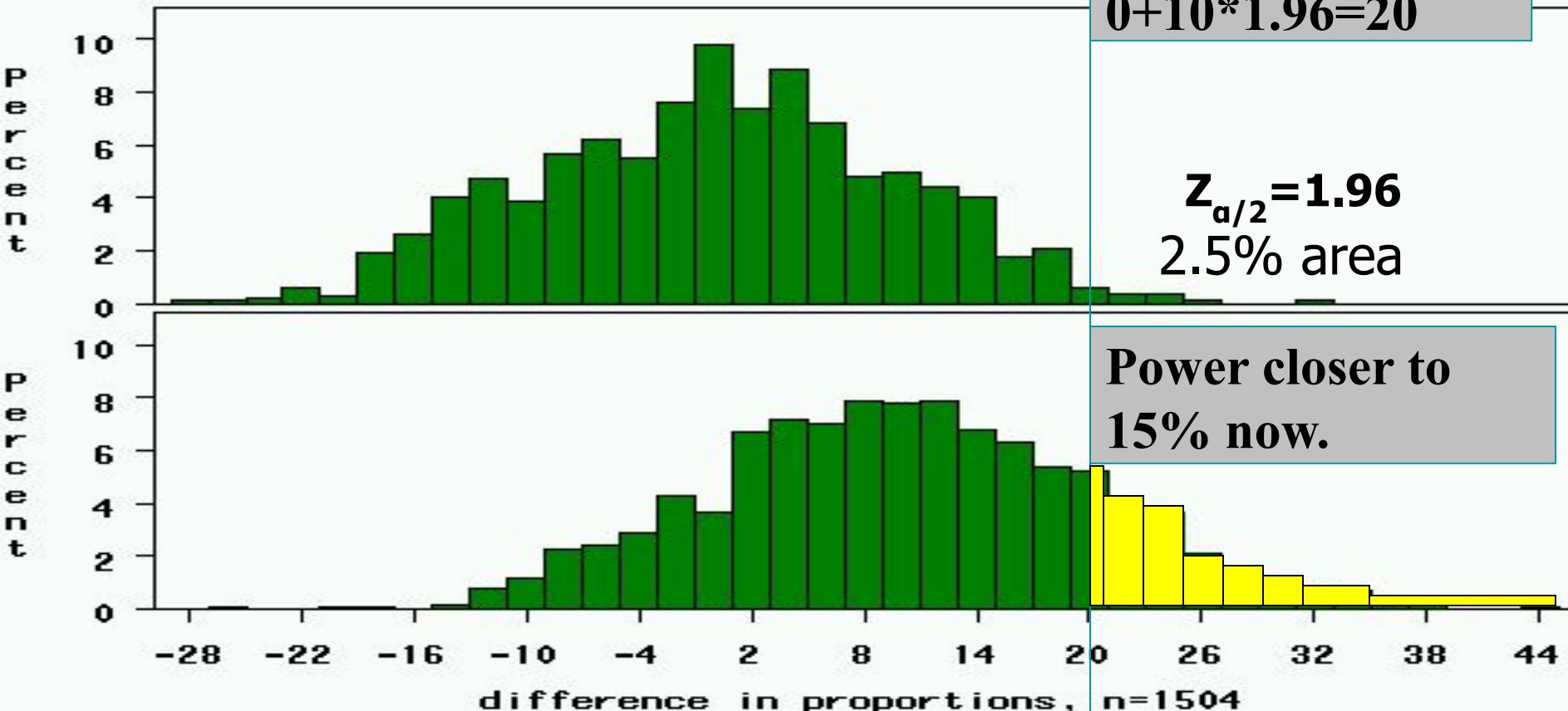


Power here:

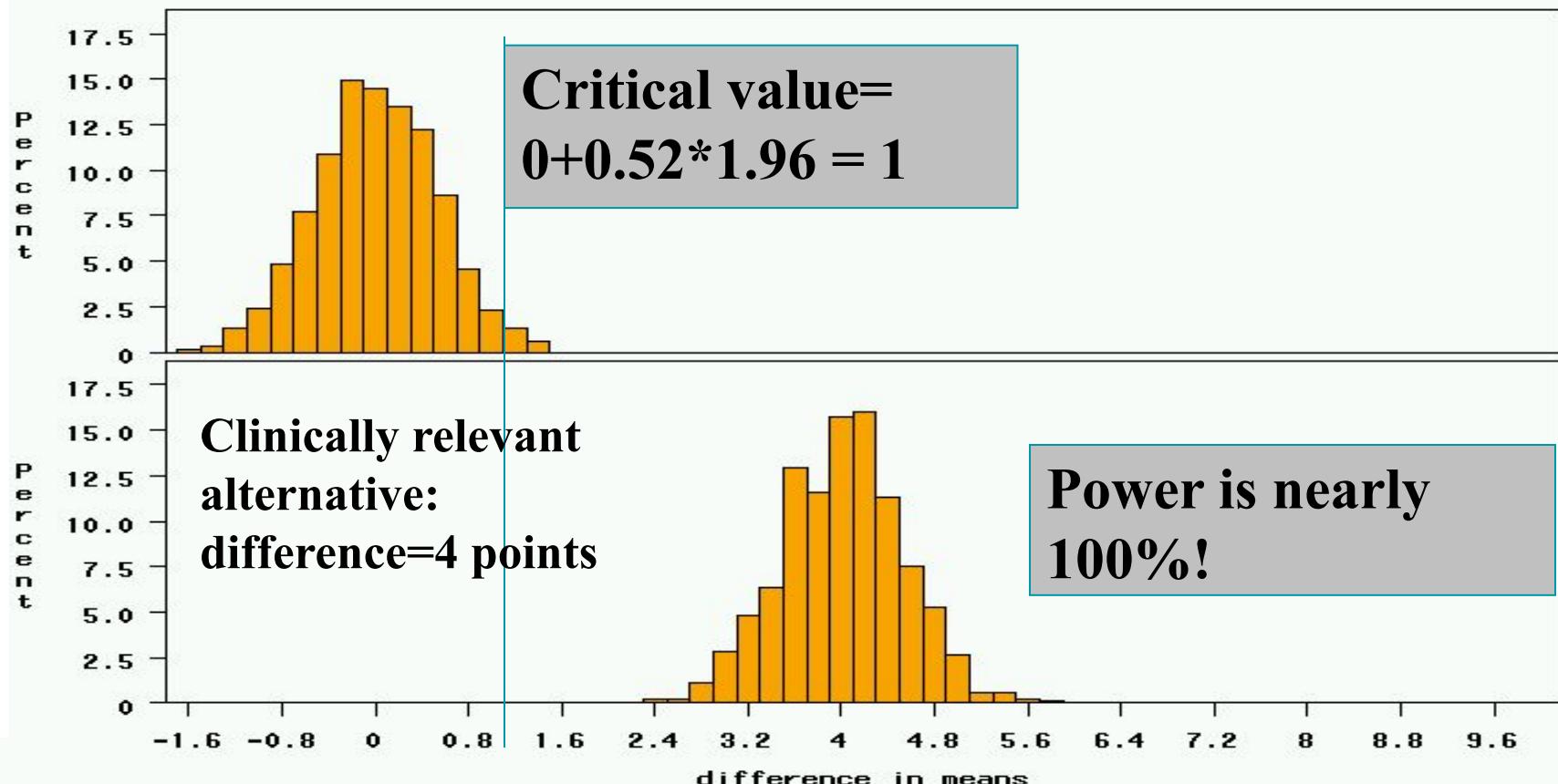
$$P(Z > \frac{6.5 - 10}{3.3}) =$$

$$P(Z > -1.06) = 85\%$$

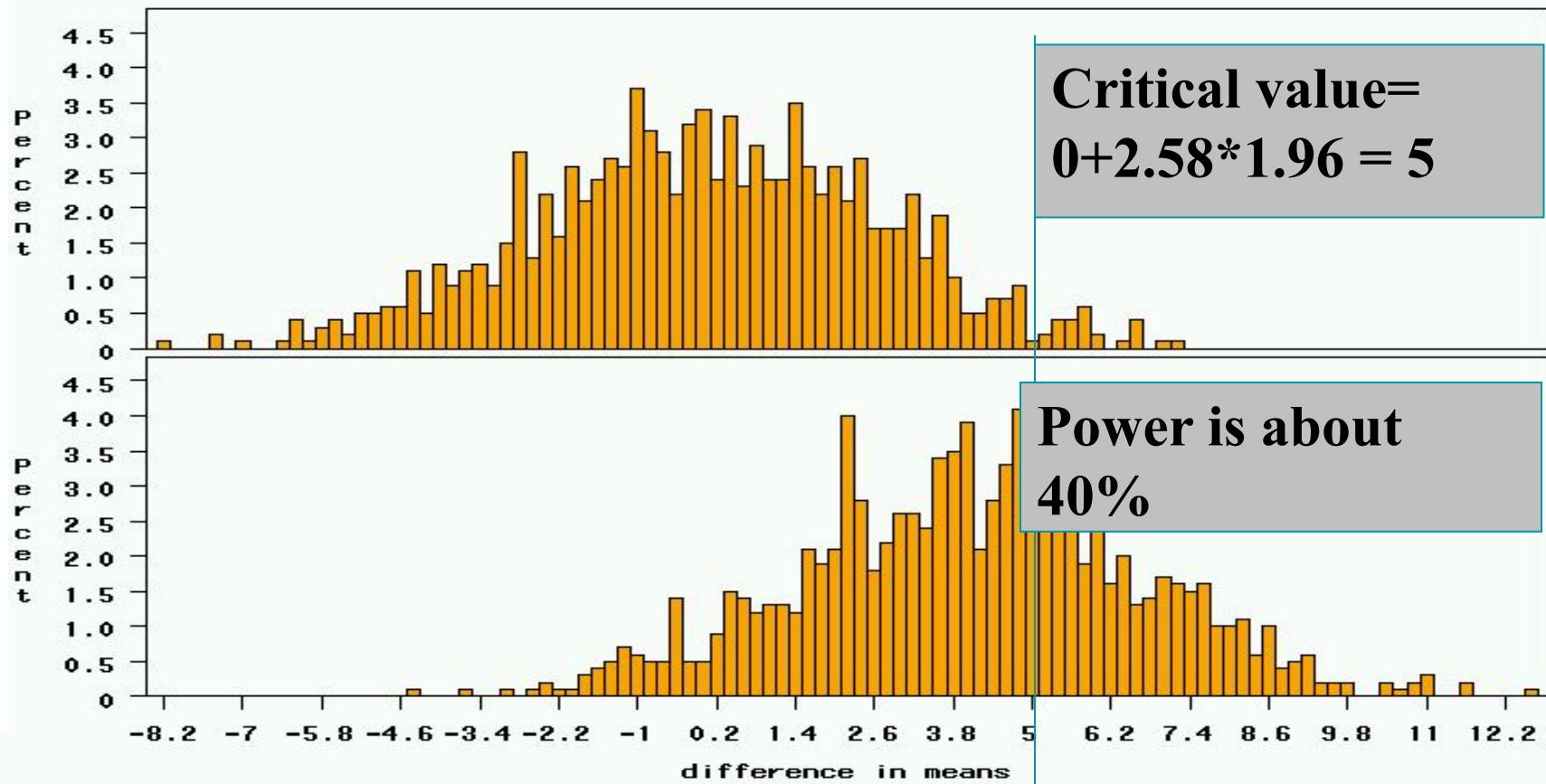
## study 1: 50 cases, 50 controls



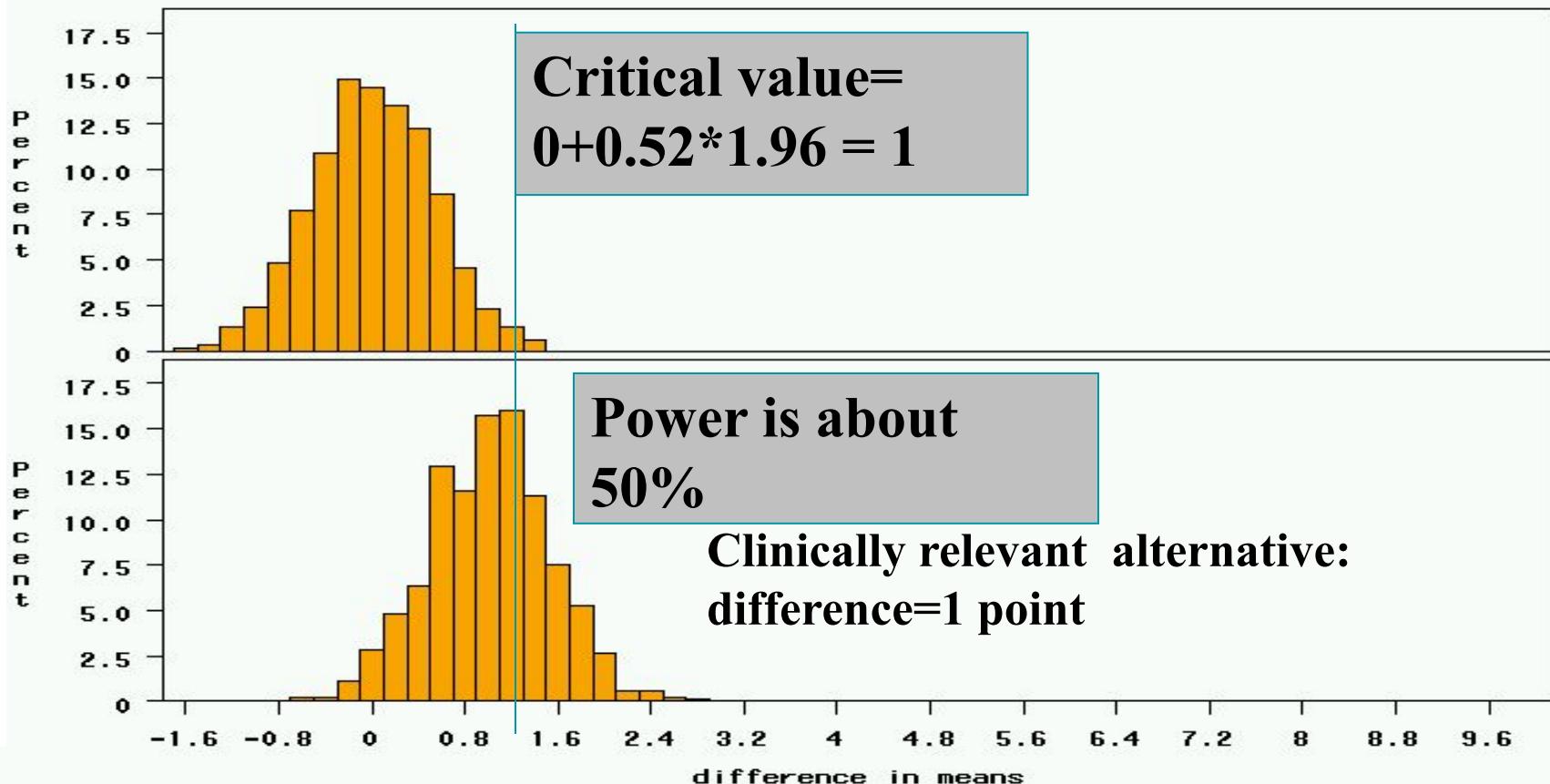
## Study 2: 18 treated, 72 controls, STD DEV = 2



## Study 2: 18 treated, 72 controls, STD DEV=10



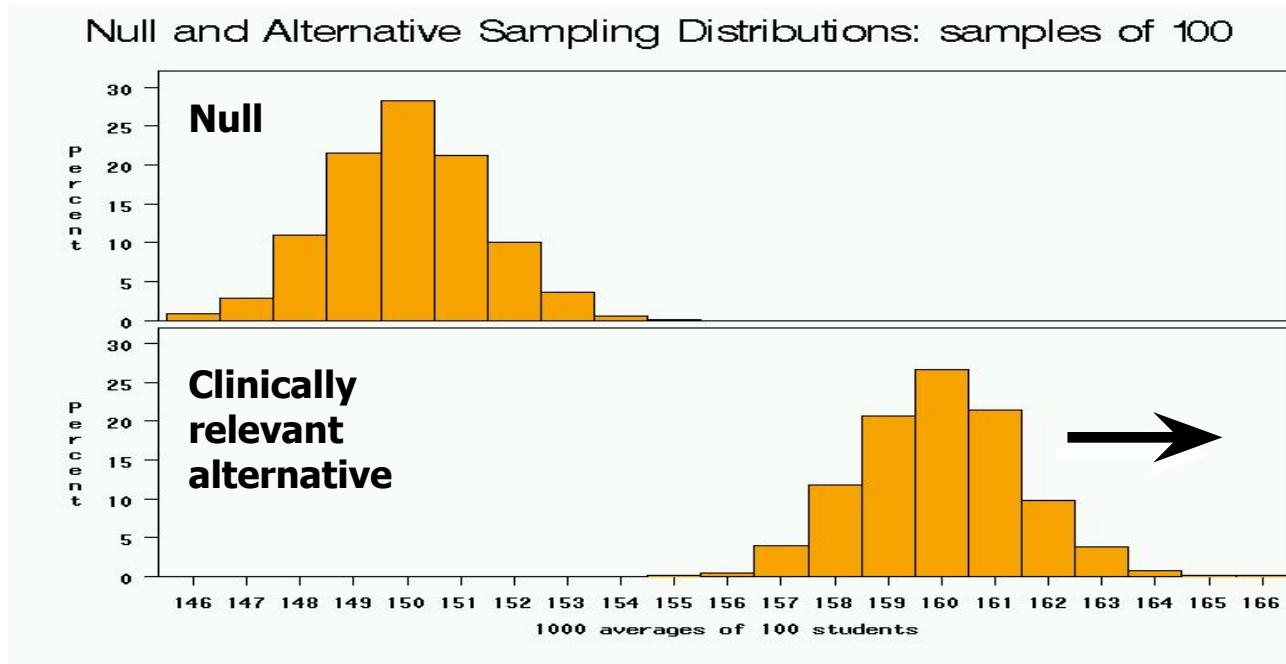
## Study 2: 18 treated, 72 controls, effect size=1.0



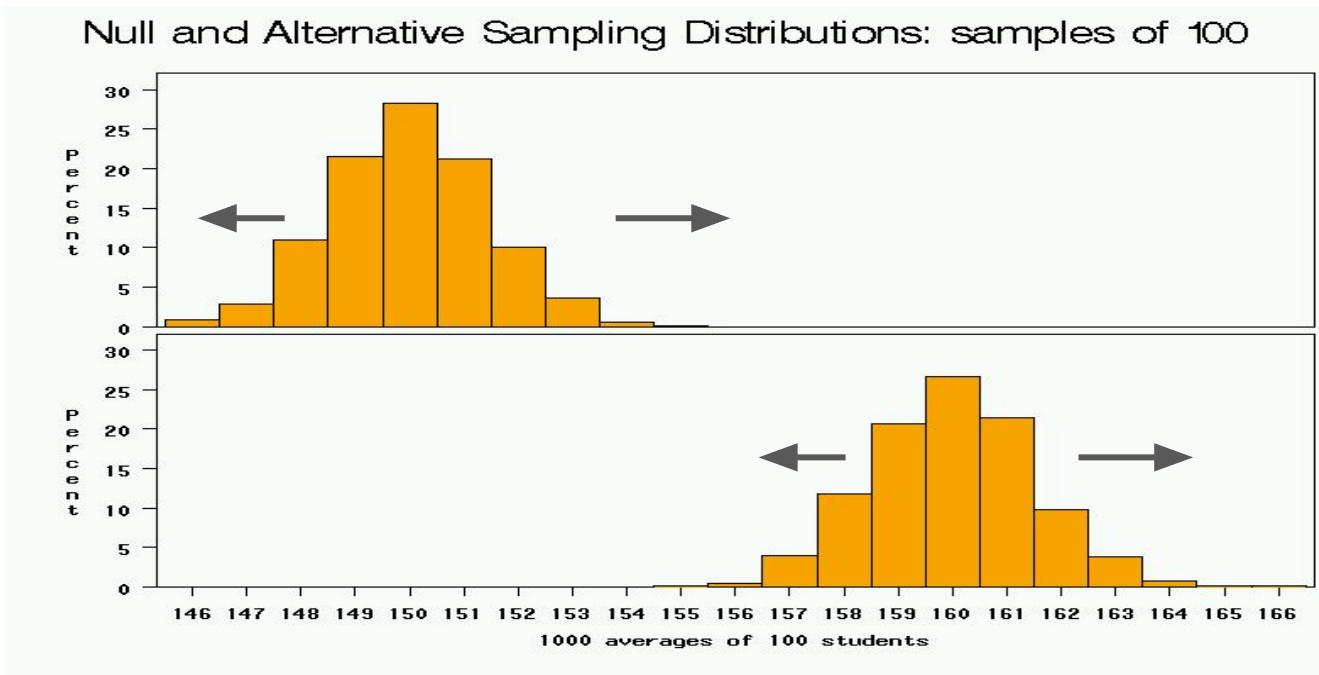
# Factors Affecting Power

1. Size of the effect ↑
2. Standard deviation of the characteristic ↘
3. Bigger sample size ↑
4. Significance level desired ↘

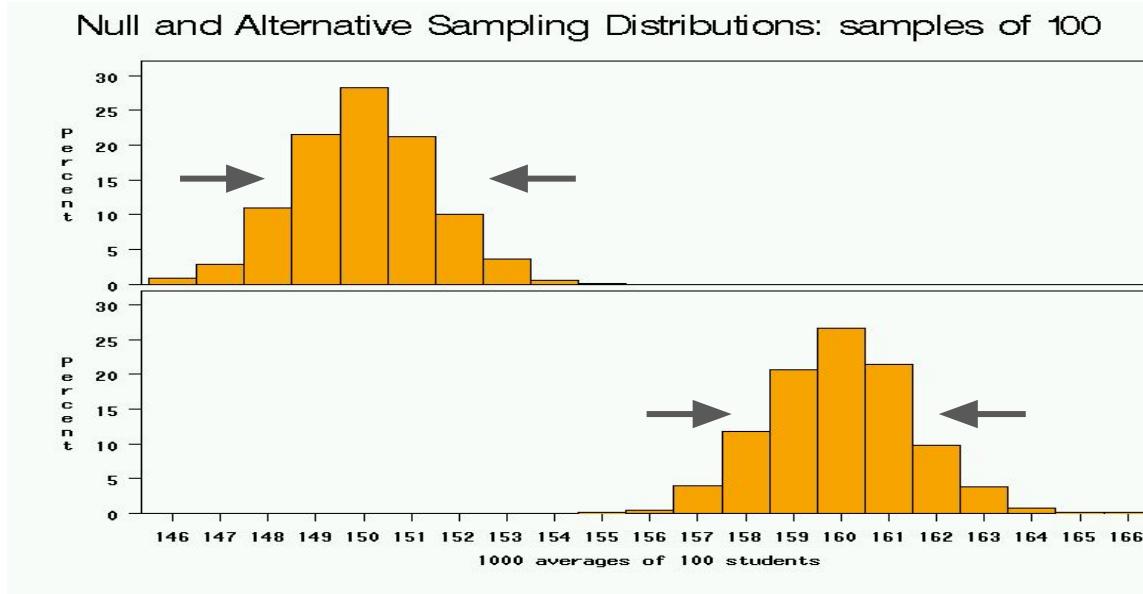
# 1. Bigger difference from the null mean



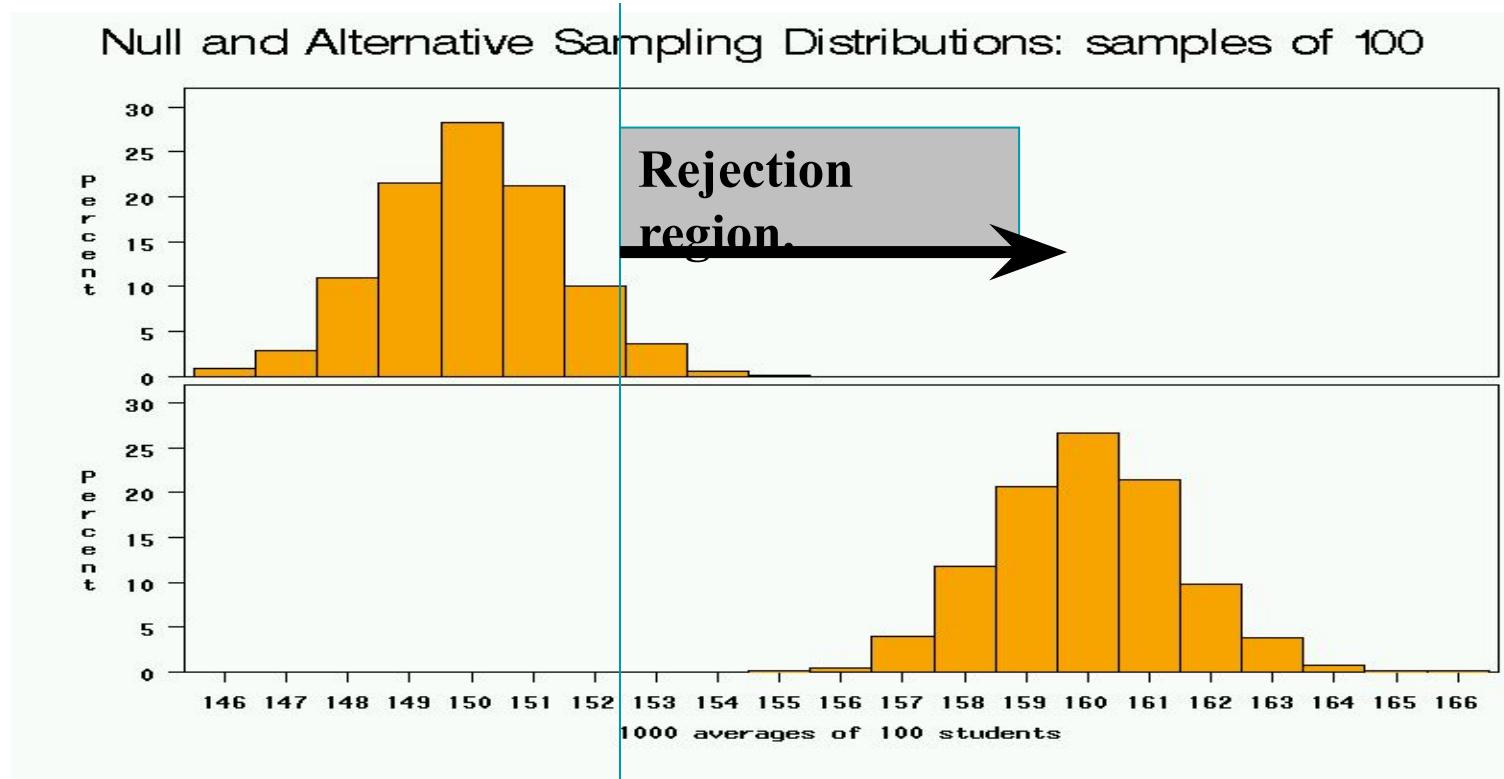
## 2. Bigger standard deviation



### 3. Bigger Sample Size



## 4. Higher significance level



# Sample size calculations

- Based on these elements, you can write a formal mathematical equation that relates power, sample size, effect size, standard deviation, and significance level...
- **\*\*WE WILL DERIVE THESE FORMULAS FORMALLY SHORTLY\*\***

# Simple formula for difference in means

**Sample size** in each group  
(assumes equal sized groups)

Represents the **desired power** (typically .84 for 80% power).

$$n = \frac{2\sigma^2(Z_{\beta} + Z_{\alpha/2})^2}{difference^2}$$

**Standard deviation** of the outcome variable

**Effect Size** (the difference in means)

Represents the desired **level of statistical significance** (typically 1.96).

# Simple formula for difference in proportions

**Sample size** in each group  
(assumes equal sized groups)

Represents the **desired power** (typically .84 for 80% power).

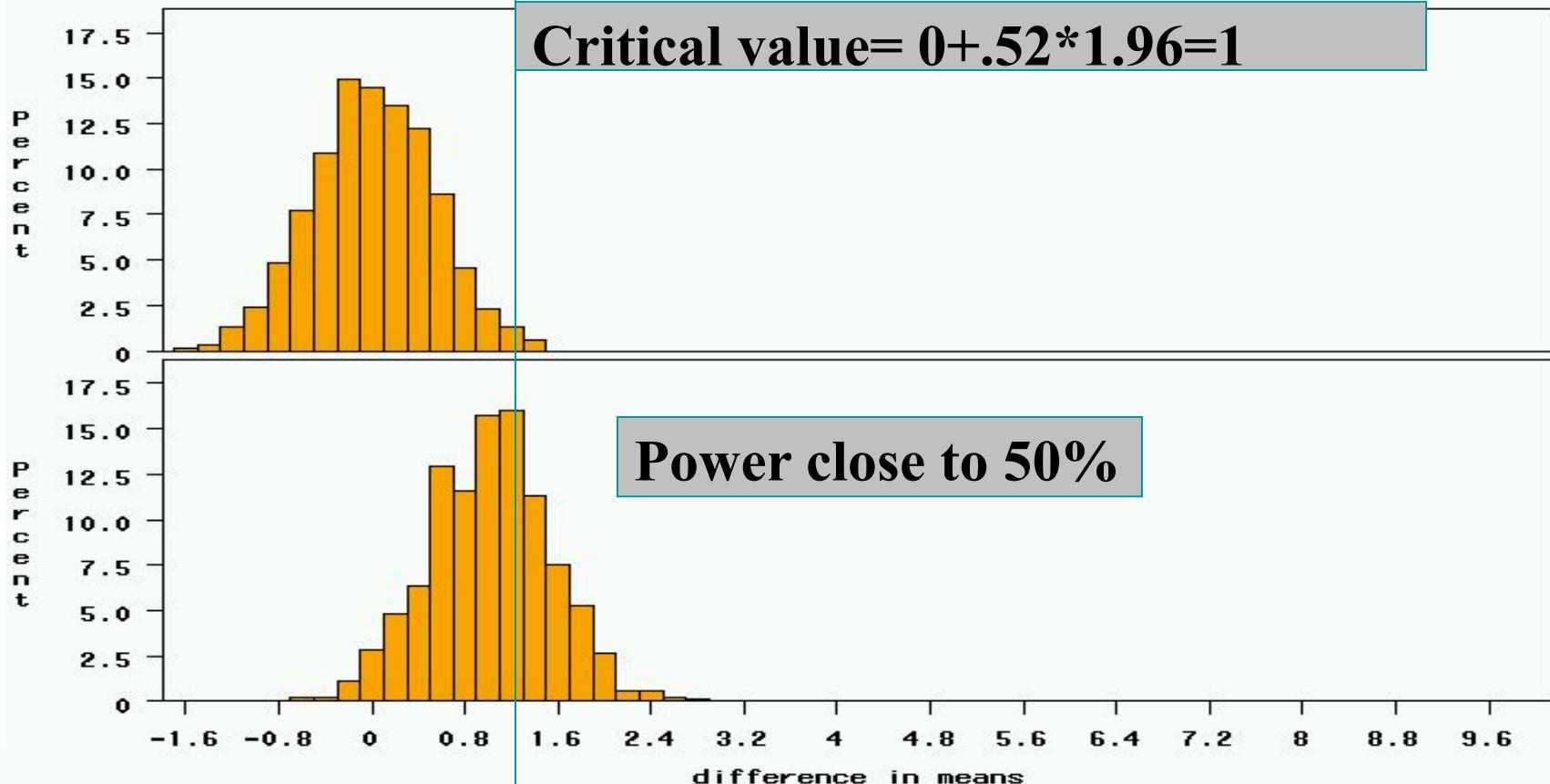
$$n = \frac{2(\bar{p})(1 - \bar{p})(Z_{\beta} + Z_{\alpha/2})^2}{(p_1 - p_2)^2}$$

A measure of **variability**  
(similar to standard deviation)

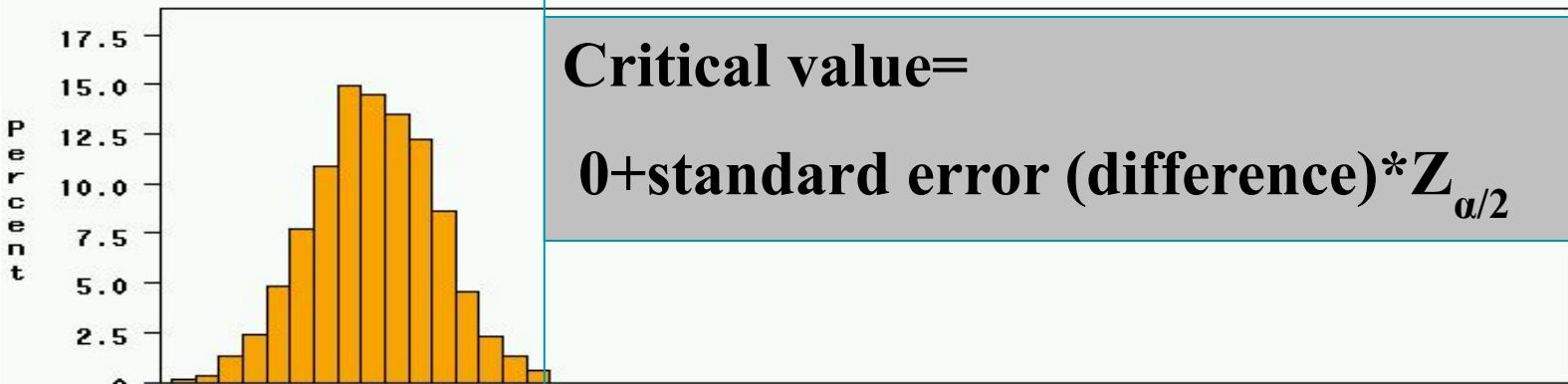
**Effect Size** (the difference in proportions)

Represents the desired **level of statistical significance** (typically 1.96).

## Study 2: 18 treated, 72 controls, effect size=1.0

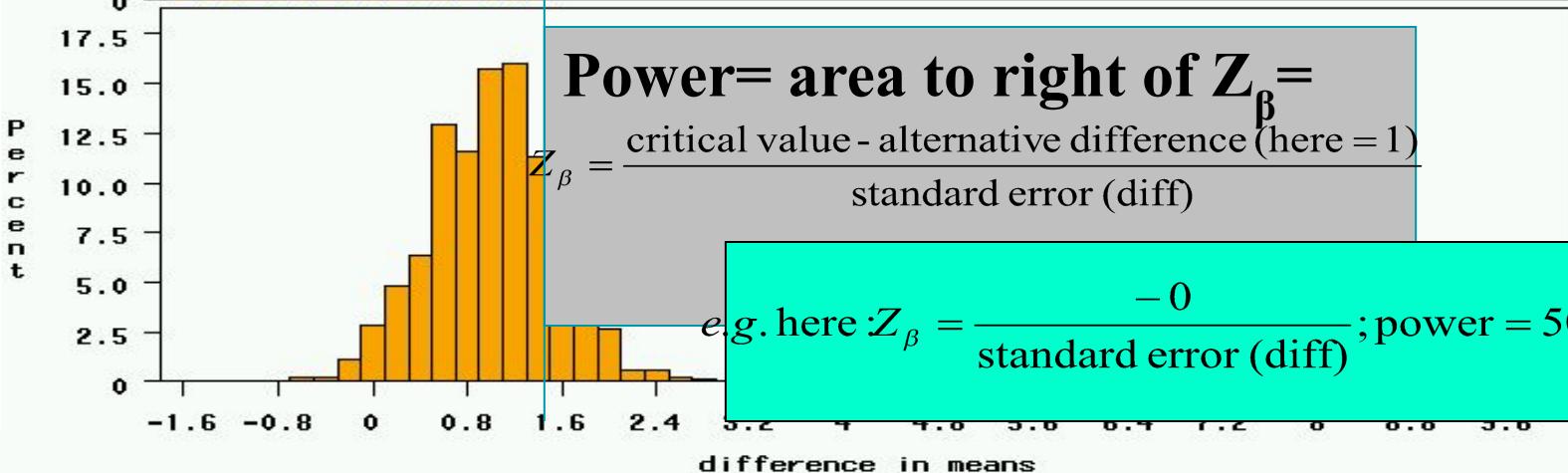


# SAMPLE SIZE AND POWER FORMULAS



**Critical value=**

$$0 + \text{standard error (difference)} * Z_{\alpha/2}$$



**Power= area to right of  $Z_\beta$ =**

$$Z_\beta = \frac{\text{critical value} - \text{alternative difference}}{\text{standard error (diff)}}$$

$$Z_\beta = \frac{Z_{\alpha/2} * \text{standard error (diff)} - \text{difference}}{\text{standard error}(diff)}$$

$$Z_\beta = Z_{\alpha/2} - \frac{\text{difference}}{\text{standard error}(diff)}$$

$$-Z_\beta = \frac{\text{difference}}{\text{standard error}(diff)} - Z_{\alpha/2}$$

$$Z_{power} = -Z_\beta \longrightarrow$$

the area to the left of  $Z_{power}$  = the area to the right of  $Z_\beta$

Power is the area to the *right* of  $Z_\beta$ . OR power is the area to the *left* of  $-Z_\beta$ . Since normal charts give us the area to the left by convention, we need to use  $-Z_\beta$  to get the correct value. Most textbooks just call this " $Z_\beta$ "; I'll use the term  $Z_{power}$  to avoid confusion.

# All-purpose power formula...

$$Z_{power} = \frac{\text{difference}}{\text{standard error(difference)}} - Z_{\alpha/2}$$

# Survival analysis

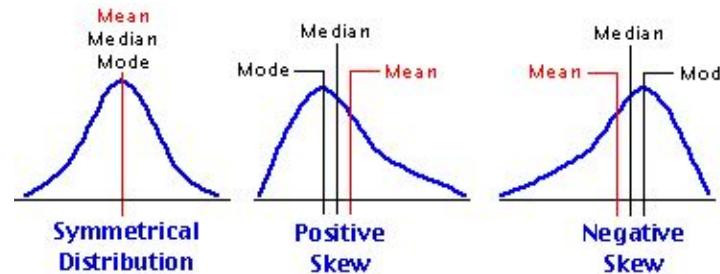
# Survival analysis

- Time to event data.
- Censoring.
- Survivor function – Kaplan-Meier plot.
- Log-rank test.
- Hazard function and Hazard ratio .

# Time to event data: examples

- Time to death.
- Time to progression of cancer.
- Time to development of diabetes.
- Time to recovery from diarrhea.
- Time to event data typically collected in
  - cohort studies (time between study baseline and event of interest).
  - clinical trials (time between randomisation and event of interest).
- Also known as **survival data**.

# Features of time to event data

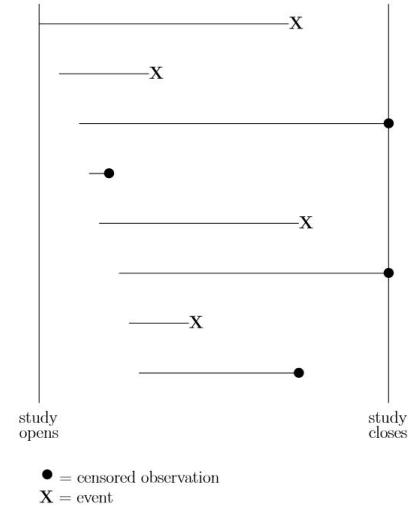


- Non-negative values.
- Not normally distributed (usually positively skewed).
- Event not usually observed for all individuals during the study.
- An observation is **censored** if individual does not experience event during the study.
- **Censoring time:** time from baseline/randomisation until latest date at which individual is known to be still alive and event-free.

# Censoring

- Definition: Event of interest not observed for all individuals.
- **Fixed censoring:** event has not occurred when study has ended or data analysis is performed.
- **Loss to follow-up:** individual has been lost to follow-up (e.g. he/she no longer wishes to take part in study).
- Survival analysis methods make use of information from censored observations.
- Assume censoring is **non-informative**, i.e. if an individual is censored, his/her subsequent risk of the event of interest is unaffected.

Illustration of survival data



# Example of time to event data

Weeks to death or censoring (\*) in 20 adults with recurrent astrocytoma:

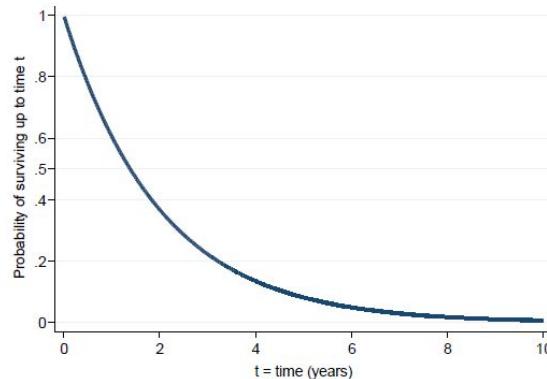
6	13	21	30	31*	37	38	47*	49	50
63	79	80*	82*	82*	86	98	149	202	219

ID	death	weeks
1	1	6
2	1	13
3	1	21
4	1	30
5	0	31
6	1	37
7	1	38
8	0	47
9	1	49
10	1	50
11	1	63
12	1	79
13	0	80
14	0	82
15	0	82
16	1	86
17	1	98
18	0	149
19	1	202
20	1	219

# Aims of survival analysis

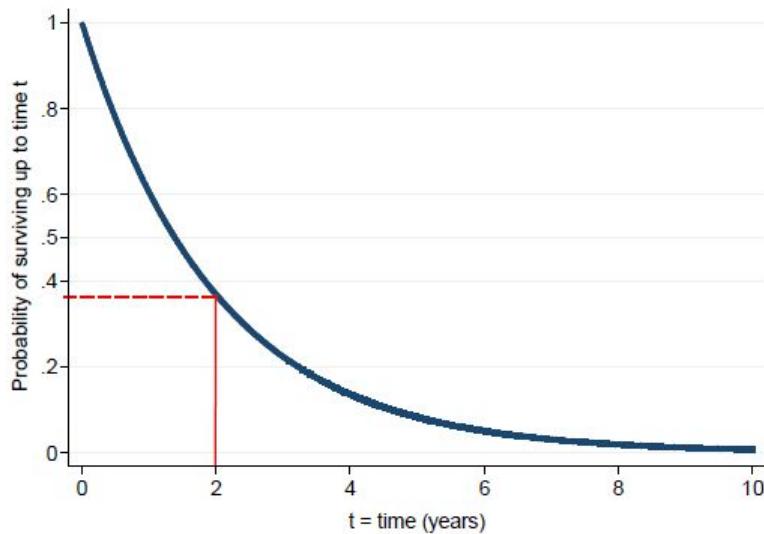
- To estimate probability of not experiencing event of interest (not dying = “surviving”) over any given time period (e.g. 5 year survival rate).
- To compare overall survival experience between different groups of individuals (e.g. between groups in a randomised clinical trial).
- **Survivor function:** Probability of not experiencing event of interest (“surviving”) up to time  $t$ .

Example:



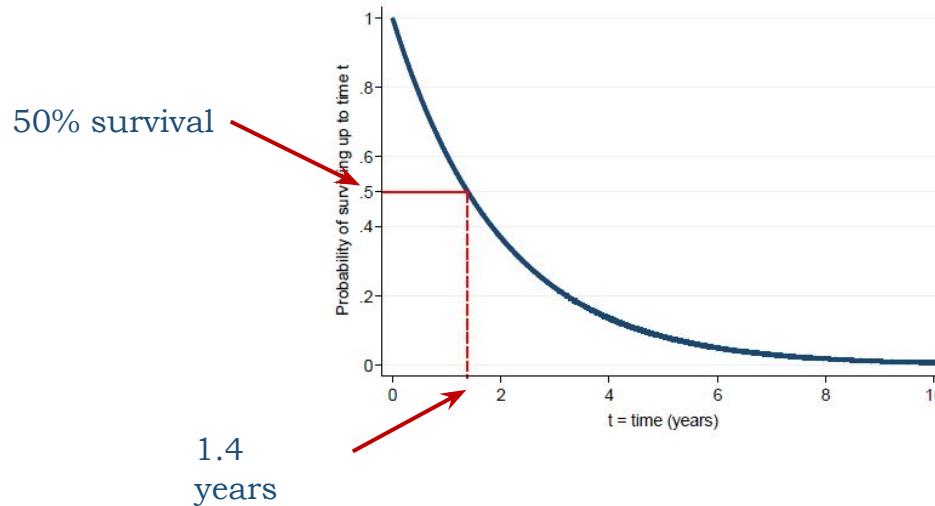
# Estimating a survival rate

- Probability of surviving up to 2 years = 0.37.



# Median survival time

- It is the time (expressed in months or years) when half the patients are expected to be alive. It means that the chance of surviving beyond that time is 50%.
- Median survival time = 1.4 years, since the probability of surviving up to 1.4 years is 0.5.

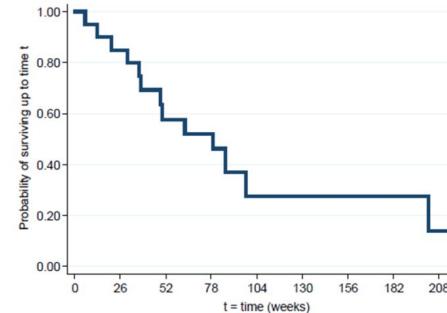


# Kaplan-Meier (KM) estimation of survivor function

## First death

6	13	21	30	31*	37	38	47*	49	50
63	79	80*	82*	82*	86	98	149	202	219

- 20 individuals in study at  $t=0$ .
- First death at  $t=6$  weeks.
- No individuals censored before  $t=6$ .
- Probability of death for each individual:  $1/20=0.05$
- Therefore probability of surviving beyond  $t=6$  is  $(1-0.05)=0.95=19/20$



Weeks in follow-up ( $t$ )	N at risk at time $t$	N of deaths at time $t$	Prob. of death at time $t$	Prob. of no death at time $t$	Prob. of surviving up to and including time $t$
0	20	0	0	1	1
6	20	1	0.05	0.95	$1 \times 0.95 = 0.95$

"Risk set" at time  $t$

$1/20$

$19/20$

# K-M estimation of survivor function

## Second death

	13	21	30	31*	37	38	47*	49	50
63	79	80*	82*	82*	86	98	149	202	219

- 19 individuals in study between  $t=6$  and  $t=13$ .
- Second death at  $t=13$ .
- No individuals censored between  $t=6$  and  $t=13$ .  $\frac{19}{2}$        $\frac{18}{1}$
- Probability of death for each individual:  $1/19 = 0.053$
- Therefore probability of surviving beyond  $t=13$  is  $0.95 \times 0.947 = 0.90$ .

- with 0

Weeks in follow-up (t)	N at risk at time t	N of deaths at time t	Prob. of death at time t	Prob. of no death at time t	Prob. of surviving up to and including time t
6	20	1	0.05	0.95	0.95
13	19	1	0.053	0.947	$0.95 \times 0.947 = 0.90$

$1/19$

$1 - (1/19) = 18/19$

# K-M estimation of survivor function

## Third and fourth death

		21	30	31*	37	38	47*	49	50
63	79	80*	82*	82*	86	98	149	202	219

- **18** individuals in study between  $t=13$  and  $t=21$ .
- Probability of death for each individual:  $1/18=0.056$
- Probability of surviving beyond  $t=21$  is  $0.90 \times (1-(1/18)) = 0.85$ .
- **17** individuals in study between  $t=21$  and  $t=30$ .
- Probability of death for each individual:  $1/17=0.059$
- Probability of surviving beyond  $t=30$  is  $0.85 \times (1-(1/17)) = 0.80$ .

From  $t=13$ :  
 $0.95 \times 0.947$

Weeks in follow-up (t)	N at risk at time t	N of deaths at time t	Prob. of death at time t	Prob. of no death at time t	Prob. of surviving up to and including time t
13	19	1	$1/19= 0.053$	0.947	0.90
21	18	1	$1/18= 0.056$	0.944	0.85
30	17	1	$1/17= 0.059$	0.941	0.80

# K-M estimation of survivor function

## Fifth and sixth death

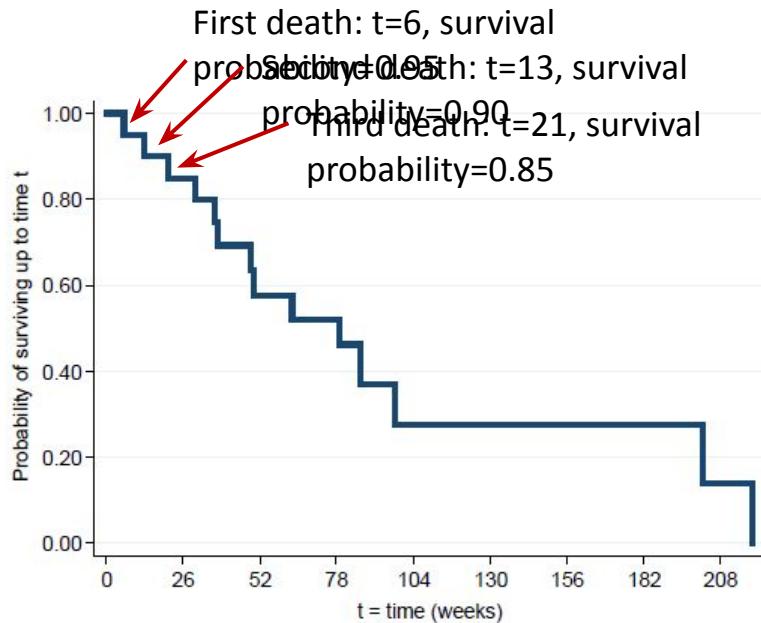
			31*	37	38	47*	49	50	
63	79	80*	82*	82*	86	98	149	202	219

- 16 individuals in study between  $t=30$  and  $t=31$ .
- 1 individual censored at  $t=31$ .
- **Probability of surviving beyond  $t=31$  remains at 0.80.**
- 15 individuals in study between  $t=31$  and  $t=37$ .
- Probability of surviving beyond  $t=37$  is  **$0.80 \times (1-(1/15)) = 0.747$** .

Weeks in follow-up (t)	N at risk at time t	N of deaths at time t	Prob. of death at time t	Prob. of no death at time t	Prob. of surviving up to and including time t
30	17	1	0.059	0.941	0.80
31	16	0	0	1	$0.80 \times 1 = 0.80$
37	15	1	$1/15 = 0.067$	0.933	$0.80 \times 0.933 = 0.747$

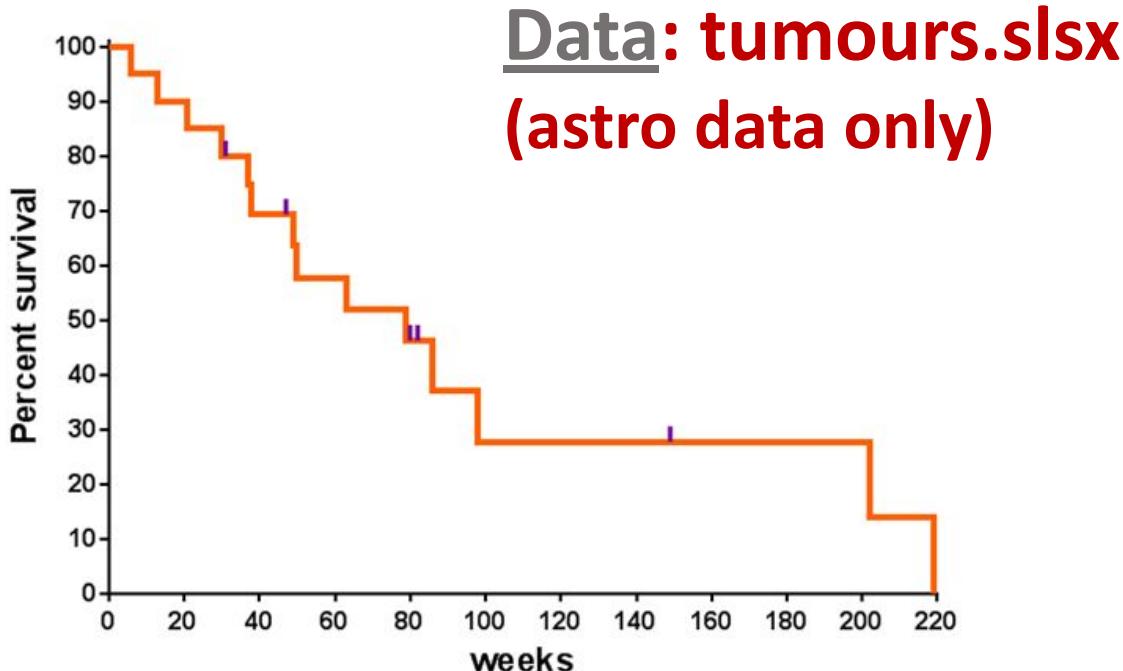
# K-M plot of survivor function

- Continue these calculations until reaching the longest event time.
- K-M plot drawn as a step function:



# K-M plot of survivor function

- Add ticks to indicate where censoring occurred.



# Comparing 2 groups

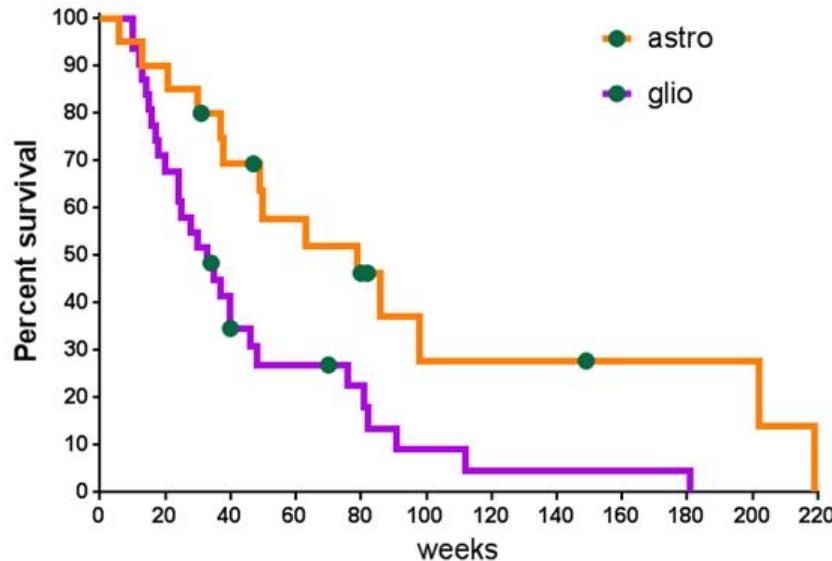
- Weeks to death or censoring (\*) in **20 adults** with recurrent astrocytoma:

6	13	21	30	31*	37	38	47*	49	50
63	79	80*	82*	82*	86	98	149	202	219

- Weeks to death or censoring (\*) in **31 adults** with recurrent glioblastoma:

10	10	12	13	14	15	16	17	18	20
24	24	25	28	30	33	34*	35	37	40
40	40*	46	48	70*	76	81	82	91	112
181									

# K-M plot of survivor function by tumour type



- Survival chances appear better in individuals with astrocytoma than with glioblastoma, but is the **difference between groups statistically significant?**

# Comparing 2 samples

- Could compare **median survival time**, or **probability of surviving** up to any particular time.
- Better to use a test which compares survivor functions over whole follow-up period.
- **Log rank test:** tests null hypothesis of no difference between samples in probability of an event (death in this example) at any time point during follow-up.
- **Log rank test statistic:**
  - based on calculating expected number of events that would occur under null hypothesis at each event time, and comparing to observed number of events.
  - under null hypothesis has a  $\text{Chi}^2$  distribution with 1 degree of freedom.

# Log rank test to compare 2 groups

Astro	Death (=1)	Glio	Death (=1)
6	1	10	1
13	1	10	1
21	1	12	1
30	1	13	1
31	0	14	1
37	1	15	1
38	1	16	1
47	0	17	1
49	1	18	1
50	1	20	1
63	1	24	1
79	1	24	1
80	0	25	1
82	0	28	1
82	0	30	1
86	1	33	1
98	1	34	0
149	0	35	1
202	1	37	1
219	1	40	1
=14		40	1
deaths		40	0
		46	1
		48	1
		70	0
		76	1
		81	1
		82	1
		91	1
		112	1
		181	1
		=28	deaths

Week	Overall Observed Deaths	Expected Deaths –	Expected Deaths –	Observed Remainder – Astro	Observed Remainder – Glio
		Astro	Glio		
6	1/51	0.392157	0.607843	19	31
10	2/50	0.76	1.24	19	29
12					
13					
14		(19/50)*2			(31/50)*2
15					
...					
	Total (Expected)	Sum	Sum		
	Total (Observed)	14	28		

Log rank test statistic has a Chi<sup>2</sup> distribution:

$$Z = \frac{\sum_{j=1}^J (O_{1j} - E_{1j})}{\sqrt{\sum_{j=1}^J V_j}}$$

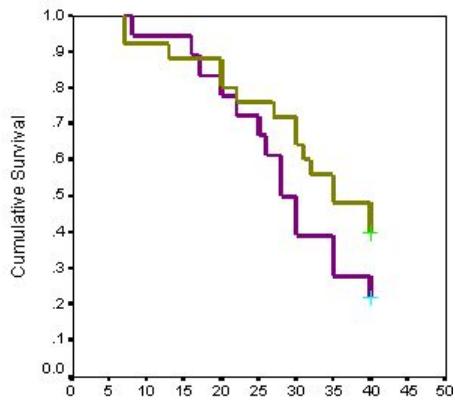
$O_j = d_{1j}$  : observed number of failures

$E_j = d_j \frac{Y_1(\tau_j)}{Y(\tau_j)}$  : expected number of failures

$V_j = \frac{Y_0(\tau_j)Y_1(\tau_j)d_j(Y(\tau_j)-d_j)}{Y(\tau_j)^2(Y(\tau_j)-1)}$  : variance of the observed number of failures

# Log rank test

- Unlikely to detect a difference between Groups if survivor functions cross over during follow-up.
- Assumes **non-informative censoring**
- Can be extended to compare more than 2 groups.

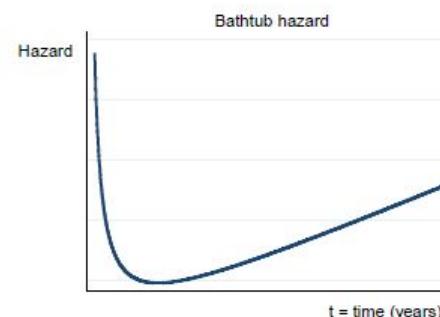
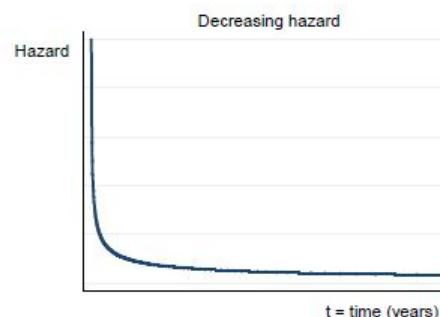
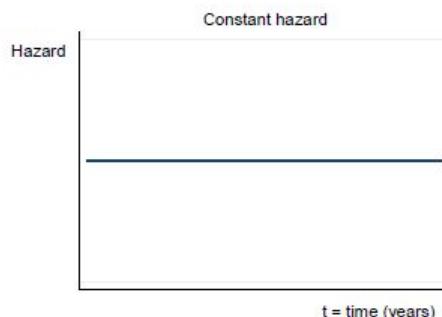


But

- Only provides a p-value, not an estimate of size of difference between groups or a confidence interval.
  - Estimate of size of difference = **Hazard Ratio**

# Hazard function

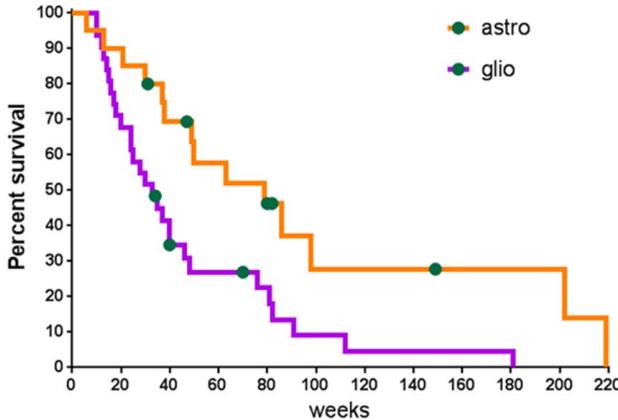
- **Hazard** is defined as the slope of the survival curve :a measure of how rapidly subjects are dying.
- Hazard function describes how hazard varies over time.



# Hazard Ratio (HR) for comparing 2 samples

- Hazards may vary over time, but assume that **HR is constant over time**.
- The hazard ratio is not directly related to the ratio of median survival times.
- When comparing 2 groups (a and b):
  - observed events (deaths) in each group: **O<sub>a</sub>** and **O<sub>b</sub>**,
  - expected events (deaths) in each group: **E<sub>a</sub>** and **E<sub>b</sub>**,
  - assuming a null hypothesis of no difference in survival.
- **HR= (O<sub>a</sub>/E<sub>a</sub>)/(O<sub>b</sub>/E<sub>b</sub>)**
- No assumption is needed about shape of hazard functions or underlying distribution of time to event data.
- HR is obtained from **Cox regression**

# Hazard Ratio (HR)



- **HR = 2.3 (95% CI [1.32;4.44])**
- At any point in time, hazard (i.e. instantaneous rate) of dying in individuals with recurrent glioblastoma is **2.3 times** higher than in individuals with recurrent astrocytoma.

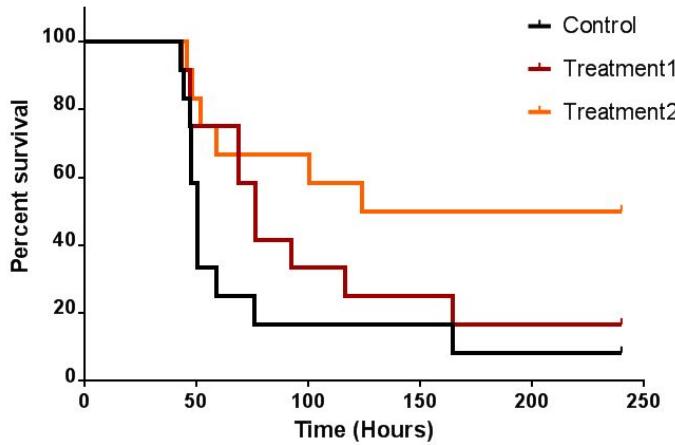
Survival		
Curve comparison		
1	Comparison of Survival Curves	
2		
3	Log-rank (Mantel-Cox) test	
4	Chi square	7.497
5	df	1
6	P value	0.0062
7	P value summary	**
8	Are the survival curves sig different?	Yes
9		
10	Gehan-Breslow-Wilcoxon test	
11	Chi square	5.828
12	df	1
13	P value	0.0158
14	P value summary	*
15	Are the survival curves sig different?	Yes
16		
17	Median survival	
18	astro	79
19	glio	33
20	Ratio (and its reciprocal)	2.394
21	95% CI of ratio	1.26 to 4.547
22		0.2199 to 0.7934
23	Hazard Ratio (Mantel-Haenszel)	
24	A/B	B/A
25	Ratio (and its reciprocal)	0.4132
26		2.42
27	95% CI of ratio	0.2194 to 0.7779
28		1.286 to 4.557
29	Hazard Ratio (logrank)	
30	A/B	B/A
28	Ratio (and its reciprocal)	0.4341
29	95% CI of ratio	2.304
		0.2367 to 0.7961
		1.256 to 4.224

# Comparing more than 2 samples

- **Issue with GraphPad:** cannot compare more than 2 groups directly
  - As in: does not run post-hoc pairwise comparisons
- **So how do we do it?**
  - Step 1: All groups comparisons (equivalent omnibus step in ANOVA)
  - Step 2: Make all pairwise comparisons of interest
  - Step 3: Apply Bonferroni correction
- **Example dataset: Lung infection**
  - Mice are infected with *Streptococcus pneumoniae*
    - 3 groups: Control, treatment 1 and treatment 2

# Comparing more than 2 groups

- Step 1: All groups comparisons



Comparison of Survival Curves	
Log-rank (Mantel-Cox) test (recommended)	
Chi square	7.112
df	2
P value	0.0286
P value summary	*
Are the survival curves sig different?	Yes
Logrank test for trend (recommended)	
Chi square	7.044
df	1
P value	0.0080
P value summary	**
Sig. trend?	Yes
Gehan-Breslow-Wilcoxon test	
Chi square	6.743
df	2
P value	0.0343
P value summary	*
Are the survival curves sig different?	Yes

- There is an overall difference in survival between the 3 groups but which group is different from which?

# Comparing more than 2 groups

- Step 2: Make all pairwise comparisons of interest

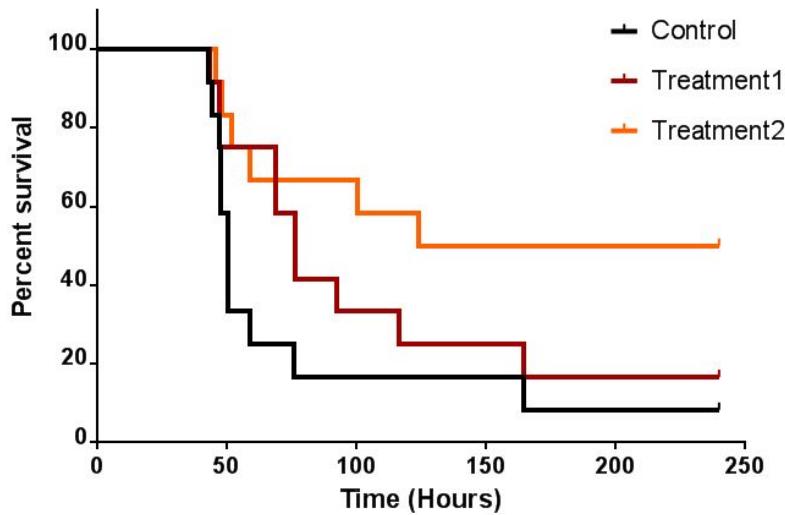
Control vs. T1	
Comparison of Survival Curves	
Log-rank (Mantel-Cox) test	
Chi square	1.800
df	1
P value	0.1798
P value summary	ns
Are the survival curves sig different?	No
Adjusted p-value =	<b>0.5394</b>
Gehan-Breslow-Wilcoxon test	
Chi square	2.227
df	1
P value	0.1356
P value summary	ns
Are the survival curves sig different?	No
Median survival	
Control	50.50
Treatment1	76.50
Ratio (and its reciprocal)	0.6601 1.515
95% CI of ratio	0.2804 to 1.554 0.6433 to 3.567
Hazard Ratio (Mantel-Haenszel)	A/B B/A
Ratio (and its reciprocal)	1.898 0.5270
95% CI of ratio	0.7443 to 4.838 0.2067 to 1.344
Hazard Ratio (logrank)	A/C C/A
Ratio (and its reciprocal)	1.720 0.5813
95% CI of ratio	0.7895 to 4.560 0.2193 to 1.267

Control vs. T2	
Comparison of Survival Curves	
Log-rank (Mantel-Cox) test	
Chi square	6.101
df	1
P value	0.0135
P value summary	*
Are the survival curves sig different?	Yes
Adjusted p-value =	<b>0.0405</b>
Gehan-Breslow-Wilcoxon test	
Chi square	5.825
df	1
P value	0.0158
P value summary	*
Are the survival curves sig different?	Yes
Median survival	
Control	50.50
Treatment2	182.0
Ratio (and its reciprocal)	0.2775 3.604
95% CI of ratio	0.1026 to 0.7503 1.333 to 9.745
Hazard Ratio (Mantel-Haenszel)	A/C C/A
Ratio (and its reciprocal)	3.642 0.2746
95% CI of ratio	1.306 to 10.16 0.09847 to 0.7658
Hazard Ratio (logrank)	A/C C/A
Ratio (and its reciprocal)	3.130 0.3195
95% CI of ratio	1.360 to 9.751 0.1026 to 0.7353

T1 vs. T2	
Comparison of Survival Curves	
Log-rank (Mantel-Cox) test	
Chi square	2.214
df	1
P value	0.1367
P value summary	ns
Are the survival curves sig different?	No
Adjusted p-value =	<b>0.4101</b>
Gehan-Breslow-Wilcoxon test	
Chi square	1.528
df	1
P value	0.2164
P value summary	ns
Are the survival curves sig different?	No
Median survival	
Treatment1	76.50
Treatment2	182.0
Ratio (and its reciprocal)	0.4203 2.379
95% CI of ratio	0.1528 to 1.157 0.8647 to 6.546
Hazard Ratio (Mantel-Haenszel)	B/C C/B
Ratio (and its reciprocal)	2.151 0.4649
95% CI of ratio	0.7843 to 5.899 0.1695 to 1.275
Hazard Ratio (logrank)	B/C C/B
Ratio (and its reciprocal)	2.084 0.4797
95% CI of ratio	0.8024 to 5.767 0.1734 to 1.246

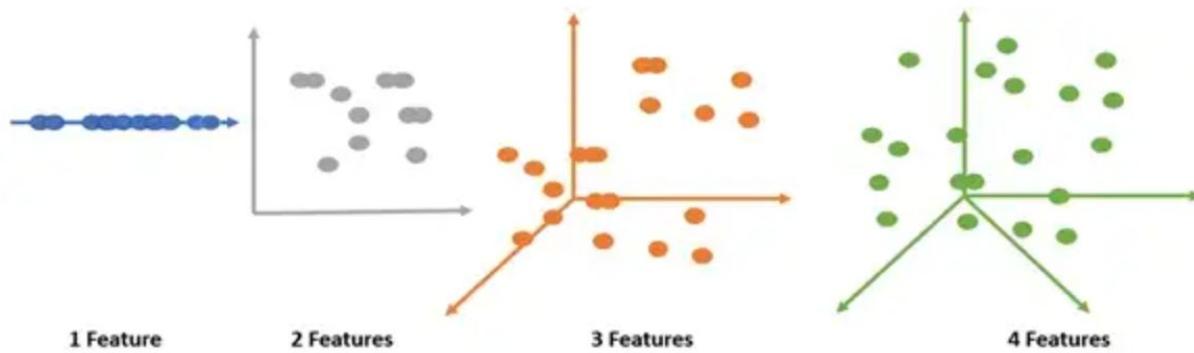
- Step 3: Apply Bonferroni correction:  $0.05/3=0.06$  or initial p-values\*3

# Comparing more than 2 groups



- At any point in time, hazard of dying in mice with lung infection is:
  - almost 2 times higher in the control than in the treatment 1 group ( $p=0.54$ )
  - 3.6 times higher in the control than in the treatment 1 group ( $p=0.04$ )

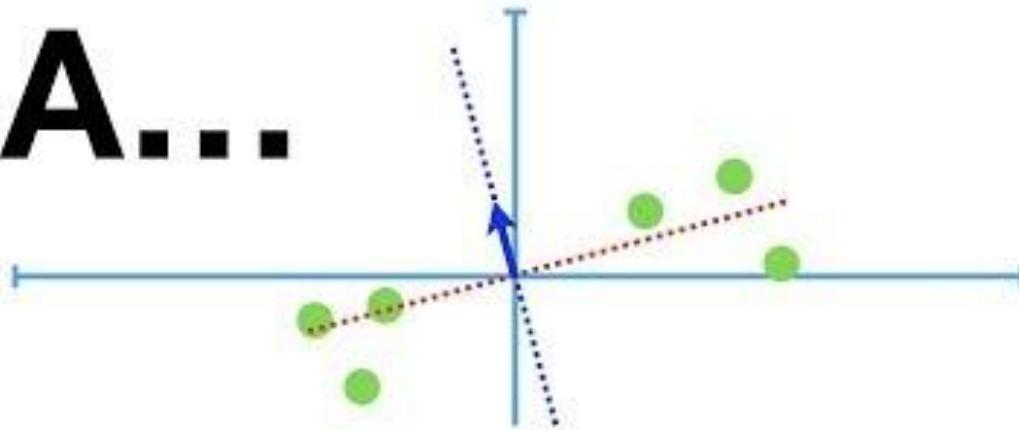
# Dimensionality Reduction Techniques





<https://www.metaboanalyst.ca/>

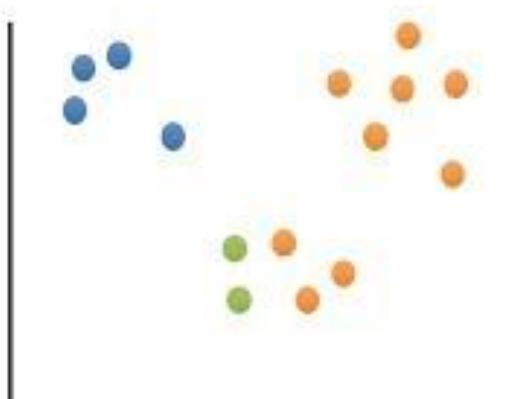
# PCA...



# Step-by-Step!!!



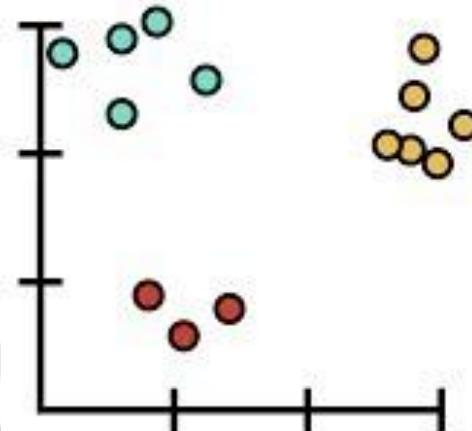
# K-Means Clustering...



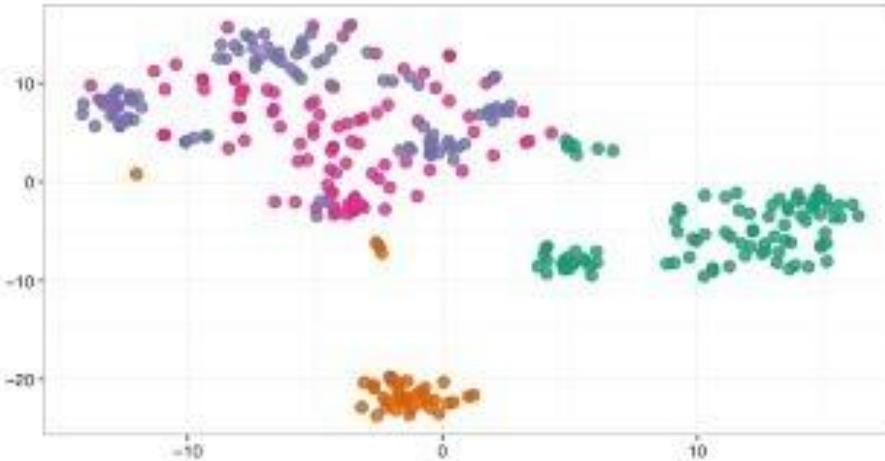
...clearly  
explained!!!

# MDS and PCoA...

...Clearly  
Explained!!!



**t-SNE...**



**Clearly Explained!!!**

# Probability

## Types of Distributions



simplilearn

# TYPES OF DISTRIBUTION IN STATISTICS



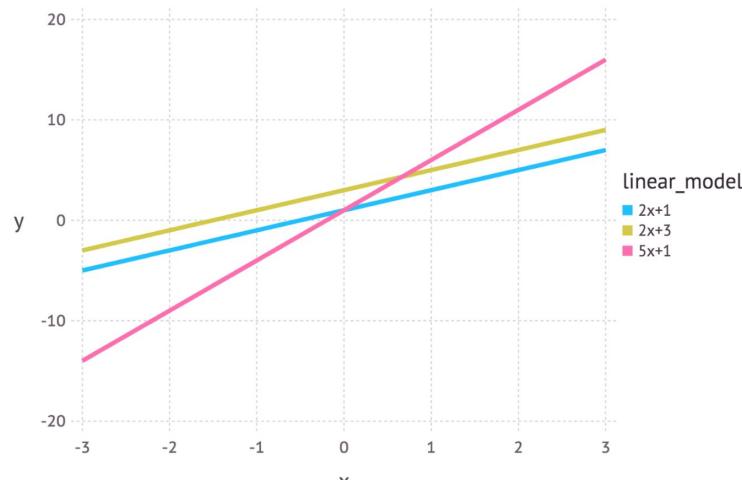
# Maximum likelihood estimation

## What are parameters?

Often in machine learning we use a model to describe the process that results in the data that are observed. For example, we may use a random forest model to classify whether customers may cancel a subscription from a service (known as [churn modelling](#)) or we may use a linear model to predict the revenue that will be generated for a company depending on how much they may spend on advertising (this would be an example of [linear regression](#)). Each model contains its own set of parameters that ultimately defines what the model looks like.

# Maximum likelihood estimation

For a linear model we can write this as  $y = mx + c$ . In this example  $x$  could represent the advertising spend and  $y$  might be the revenue generated.  $m$  and  $c$  are parameters for this model. Different values for these parameters will give different lines (see figure below).



Three linear models with different parameter values.

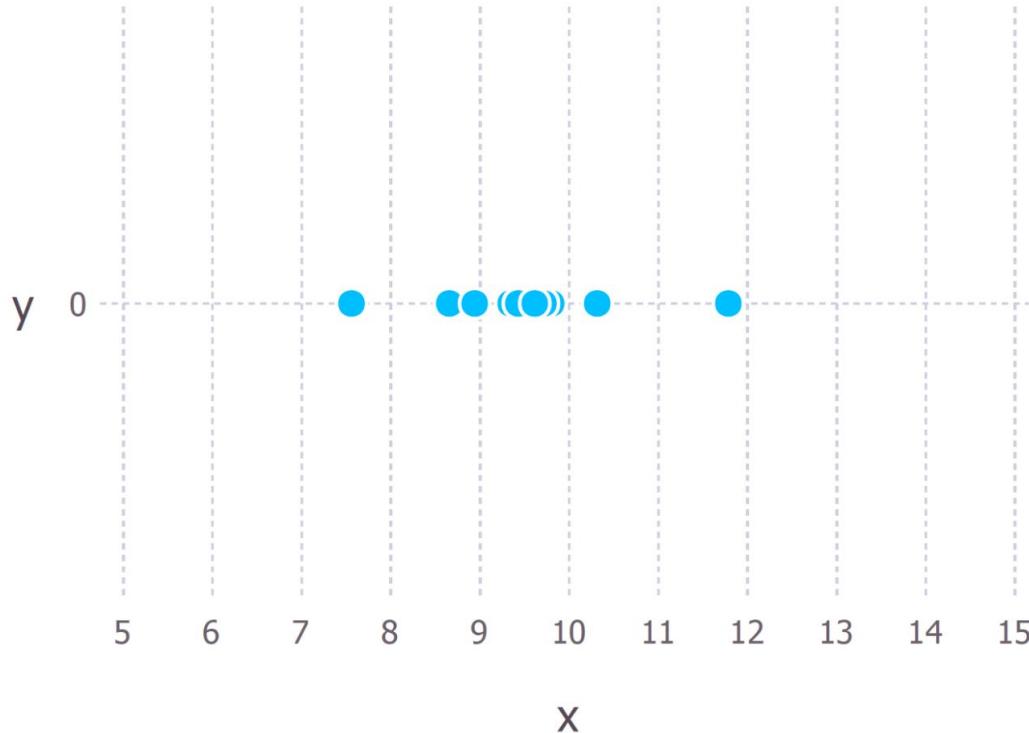
# Maximum likelihood estimation

So parameters define a blueprint for the model. It is only when specific values are chosen for the parameters that we get an instantiation for the model that describes a given phenomenon.

# Maximum likelihood estimation

Maximum likelihood estimation is a method that determines values for the parameters of a model. The parameter values are found such that they maximise the likelihood that the process described by the model produced the data that were actually observed. **Let's suppose we have observed 10 data points from some process. For example, each data point could represent the length of time in seconds that it takes a student to answer a specific exam question. These 10 data points are shown in the figure below**

# Maximum likelihood estimation

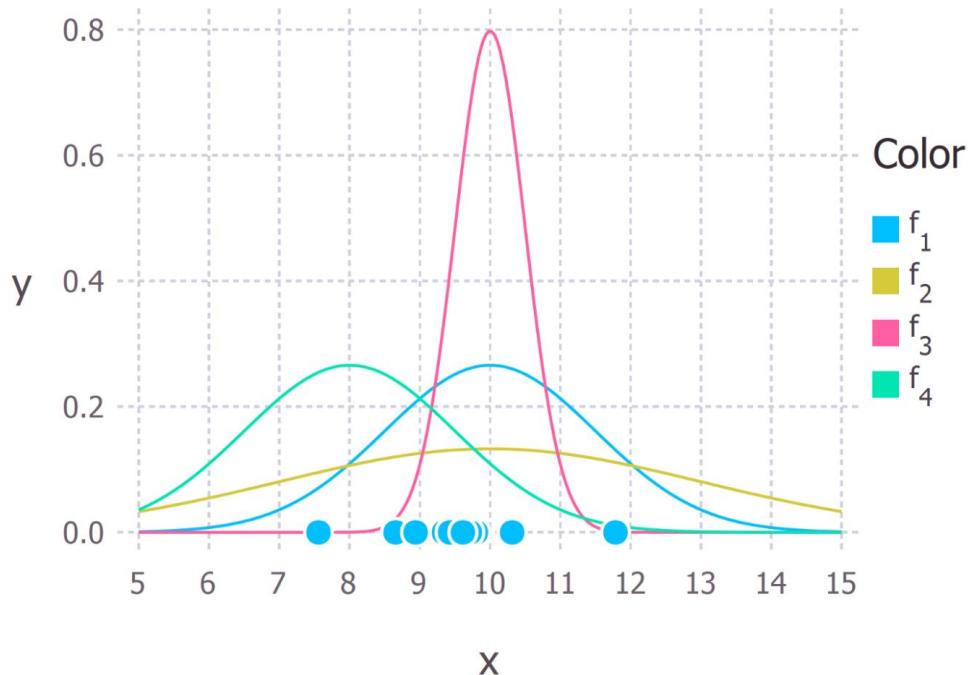


The 10 (hypothetical) data points that we have observed

# Maximum likelihood estimation

For these data we'll assume that the data generation process can be adequately described by a Gaussian (normal) distribution. Visual inspection of the figure above suggests that a Gaussian distribution is plausible because most of the 10 points are clustered in the middle with few points scattered to the left and the right. (Making this sort of decision on the fly with only 10 data points is ill-advised but given that I generated these data points we'll go with it).

# Maximum likelihood estimation



The 10 data points and possible Gaussian distributions from which the data were drawn. f<sub>1</sub> is normally distributed with mean 10 and variance 2.25 (variance is equal to the square of the standard deviation), this is also denoted  $f_1 \sim N(10, 2.25)$ .  $f_2 \sim N(10, 9)$ ,  $f_3 \sim N(10, 0.25)$  and  $f_4 \sim N(8, 2.25)$ . The goal of maximum likelihood is to find the parameter values that give the distribution that maximise the probability of observing the data.

The true distribution from which the data were generated was  $f_1 \sim N(10, 2.25)$ , which is the blue curve in the figure above.

# Calculating the Maximum Likelihood Estimates

How do we calculate the maximum likelihood estimates of the parameter values of the Gaussian distribution  $\mu$  and  $\sigma$ ?

The probability density of observing a single data point  $x$ , that is generated from a Gaussian distribution is given by:

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

# Calculating the Maximum Likelihood Estimates

How do we calculate the maximum likelihood estimates of the parameter values of the Gaussian distribution  $\mu$  and  $\sigma$ ?

In our example the total (joint) probability density of observing the three data points is given by:

$$P(9, 9.5, 11; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(9-\mu)^2}{2\sigma^2}\right) \times \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(9.5-\mu)^2}{2\sigma^2}\right) \\ \times \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(11-\mu)^2}{2\sigma^2}\right)$$

# When is least squares minimisation the same as maximum likelihood estimation?

**Least squares minimisation is another common method for estimating parameter values for a model in machine learning. It turns out that when the model is assumed to be Gaussian as in the examples above, the MLE estimates are equivalent to the least squares method.**

# Hands On and Live Assignment

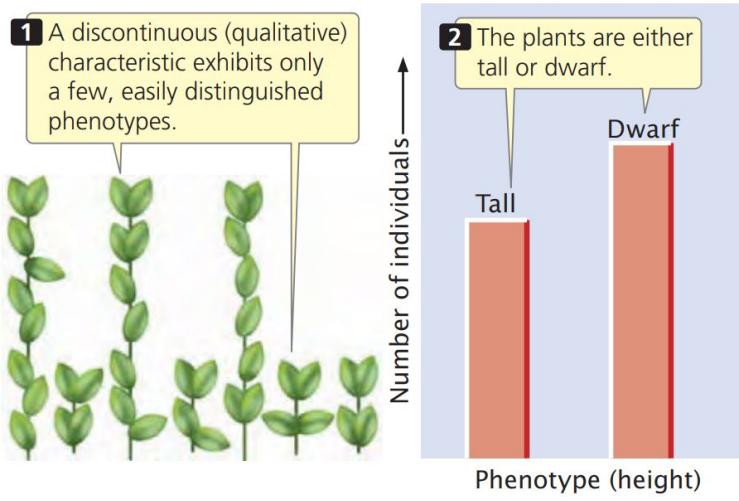
[https://drive.google.com/file/d/1YGsVgv8d-RjRifoFjGqwVDgGuPcCooW-/vie  
w?usp=share\\_link](https://drive.google.com/file/d/1YGsVgv8d-RjRifoFjGqwVDgGuPcCooW-/view?usp=share_link)

<https://statsandr.com/blog/a-shiny-app-for-inferential-statistics-by-hand/>

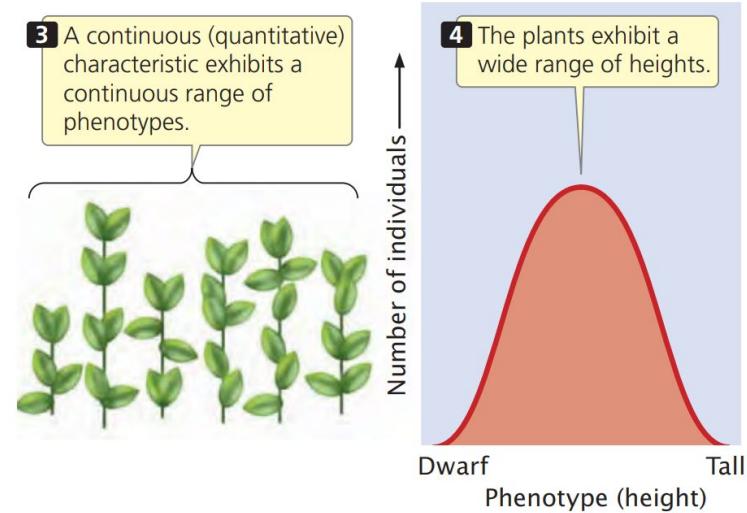
- Quantitative Genetics: Relationship between genotype and phenotype,
- Types of Quantitative Characteristics
- Nilsson-Ehle's cross.

# Quantitative Genetics: Relationship between genotype and phenotype

## (a) Discontinuous characteristic



## (b) Continuous characteristic



How quantitative characteristics are often influenced by many genes, each of which has a small effect on the phenotype.

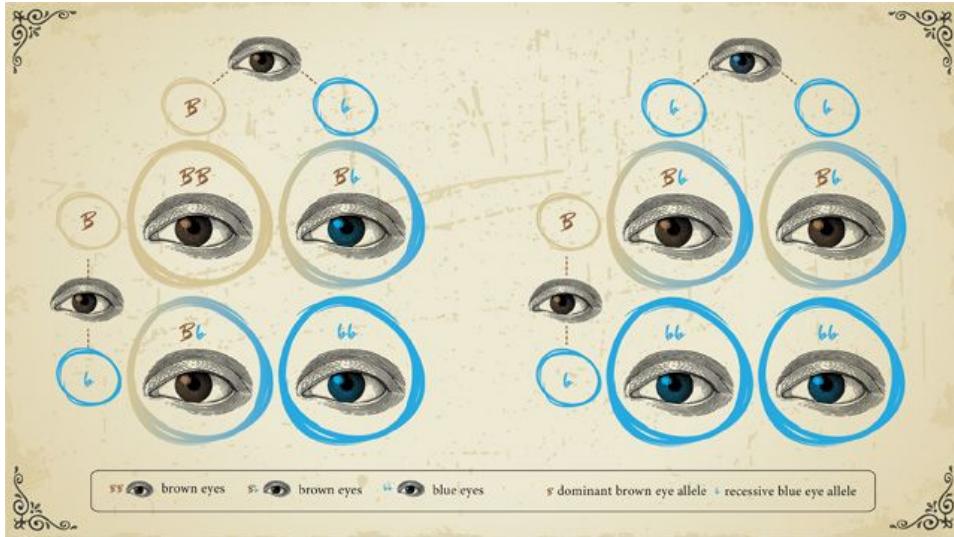
# Quantitative Genetics: Relationship between genotype and phenotype

- Qualitative, or discontinuous, characteristics possess only a few distinct phenotypes.
- However, many characteristics vary continuously along a scale of measurement with many overlapping phenotypes
- They are referred to as continuous characteristics; they are also called quantitative characteristics because any individual's phenotype must be described with a quantitative measurement.
- Quantitative characteristics might include height, weight, and blood pressure in humans, growth rate in mice, seed weight in plants, and milk production in cattle

Let's learn first the difference between a **Phenotype** and **Genotype**

# Genotype and Phenotype

Phenotype	Genotype
Purple	$PP$ (homozygous)
Purple	$Pp$ (heterozygous)
Purple	
White	$pp$ (homozygous)



Ratio 3:1

Ratio 1:2:1

# Quantitative Characteristics

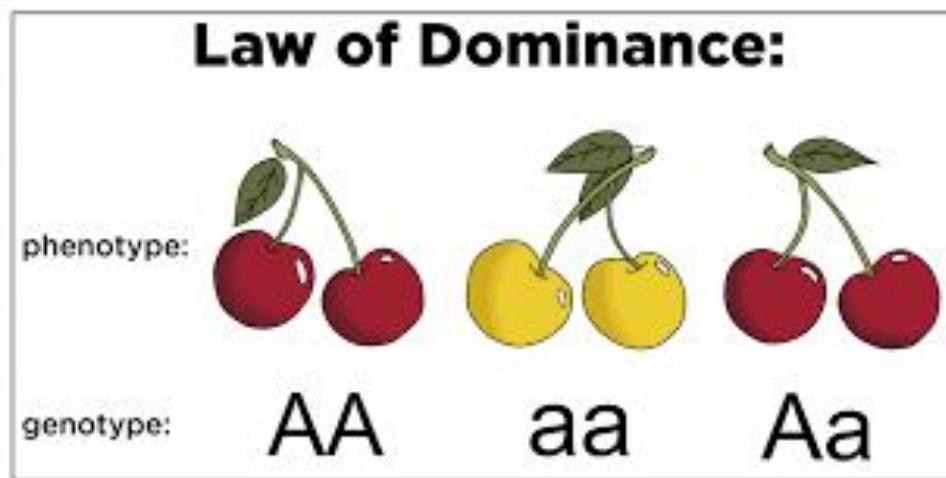
Quantitative characteristics arise from two phenomena.

- First, many are polygenic: they are influenced by genes at many loci. If many loci take part, many genotypes are possible, each producing a slightly different phenotype.
- Second, quantitative characteristics often arise when environmental factors affect the phenotype because environmental differences result in a single genotype producing a range of phenotypes.

**Most continuously varying characteristics are both polygenic and influenced by environmental factors, and these characteristics are said to be multifactorial.**

# The Relation Between Genotype and Phenotype

- For many discontinuous characteristics, the relation between genotype and phenotype is straightforward. Each genotype produces a single phenotype, and most phenotypes are encoded by a single genotype.



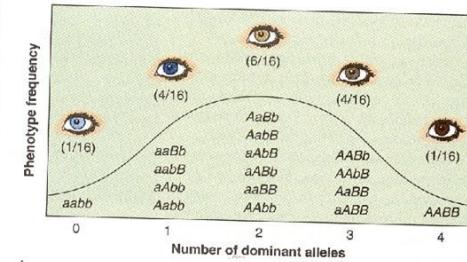
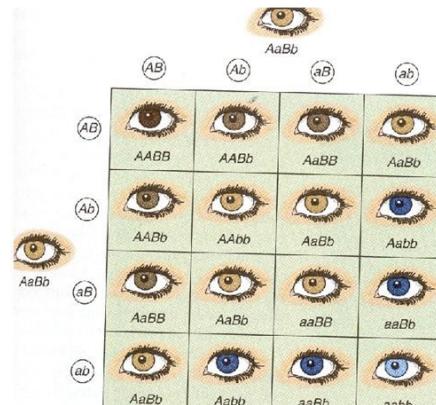
# Quantitative Characteristics

## Quantitative Genetics – role of genetics

- For quantitative characteristics, the relation between genotype and phenotype is often more complex. If the characteristic is polygenic, many different genotypes are possible, several of which may produce the same phenotype.

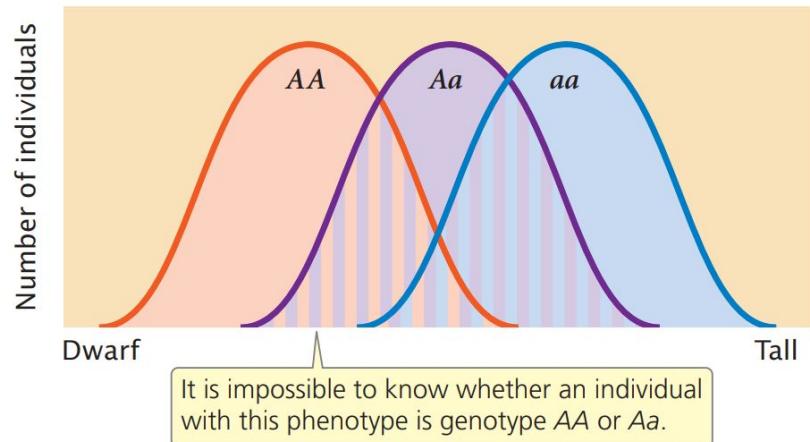
**Pure Polygenic Traits** are those not influenced by the environment

Range of phenotypes due to additive effects of loci, e.g. eye color.



# Environment further complicates the relation between genotype and phenotype.

Because of environmental effects, the same genotype may produce a range of potential phenotypes. The phenotypic ranges of different genotypes may overlap, making it difficult to know whether individuals differ in phenotype because of genetic or environmental differences



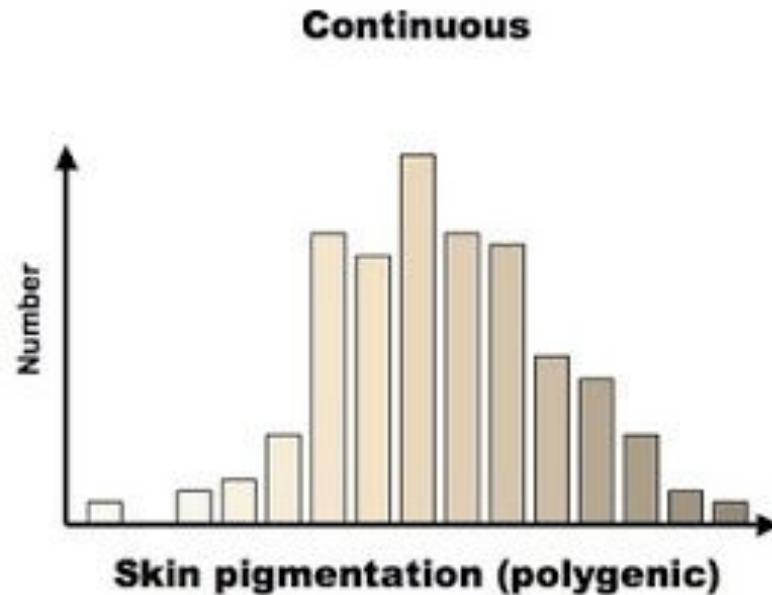
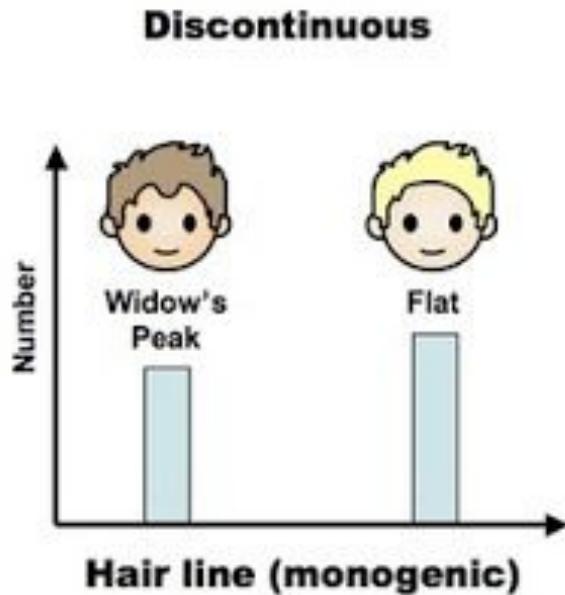
# Types of Quantitative Characteristics

## continuous characteristic

- A continuous characteristic can theoretically assume any value between two extremes; the number of phenotypes is limited only by our ability to precisely measure the phenotype.
- Human height is a continuous characteristic because, within certain limits, people can theoretically have any height.
- Although the number of phenotypes possible with a continuous characteristic is infinite, we often group similar phenotypes together for convenience; we may say that two people are both 5 feet 11 inches tall, but careful measurement may show that one is slightly taller than the other

# Types of Quantitative Characteristics

- continuous characteristic

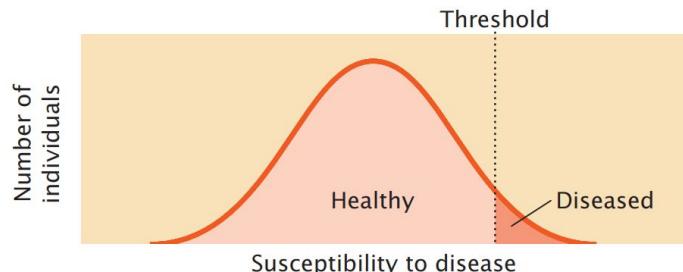


## Meristic characteristics

- Some characteristics are not continuous but are nevertheless considered quantitative because they are determined by multiple genetic and environmental factors.
- Meristic characteristics, for instance, are measured in whole numbers. An example is litter size: a female mouse may have 4, 5, or 6 pups but not 4.13 pups.
- A meristic characteristic has a limited number of distinct phenotypes, but the underlying determination of the characteristic may still be quantitative.

# Threshold Characteristic

- It is simply present or absent.
- For example, the presence of some diseases can be considered a threshold characteristic. Although threshold characteristics exhibit only two phenotypes, they are considered quantitative because they, too, are determined by multiple genetic and environmental factors.



**24.3 Threshold characteristics display only two possible phenotypes—the trait is either present or absent—but they are quantitative because the underlying susceptibility to the characteristic varies continuously.** When the susceptibility exceeds a threshold value, the characteristic is expressed.

# Nilsson-Ehle's cross

Nilsson-Ehle obtained several homozygous varieties of wheat that differed in color. Like Mendel, he performed crosses between these homozygous varieties and studied the ratios of phenotypes in the progeny. In one experiment, he crossed a variety of wheat that possessed white kernels with a variety that possessed purple (very dark red) kernels and obtained the following results:

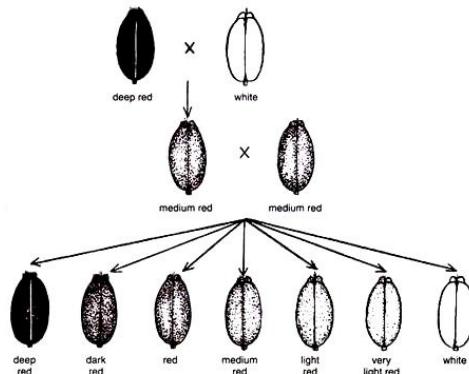
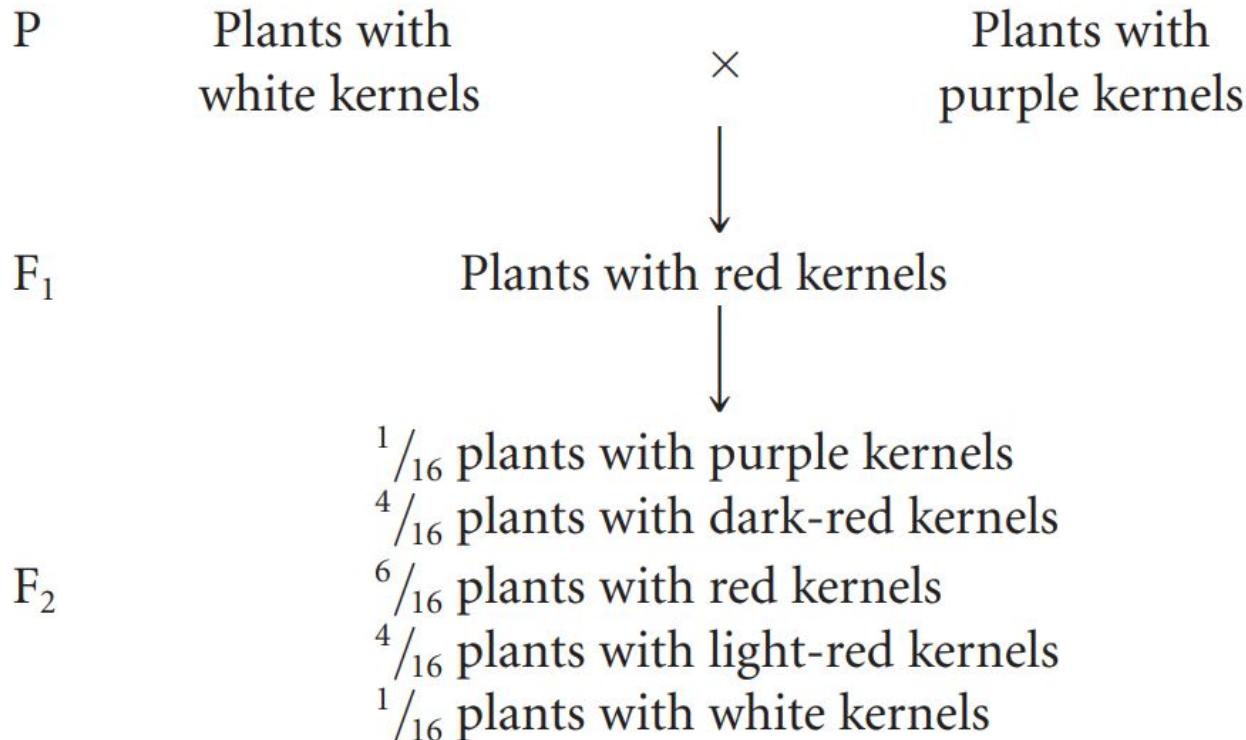


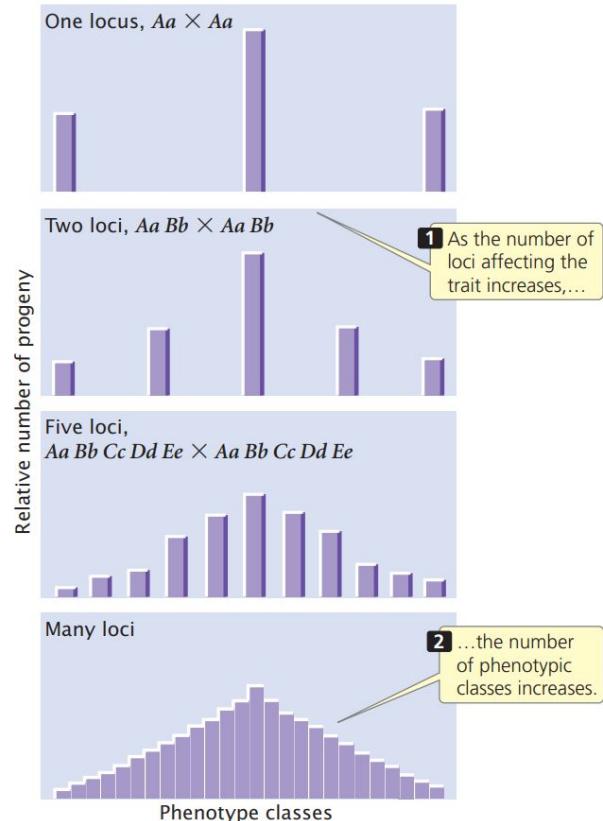
Fig. 4.1 Cross made by Nilsson-Ehle between red and white varieties of wheat.

## Nilsson-Ehle's cross



## Conclusions and implications

Nilsson-Ehle's crosses demonstrated that the difference between the inheritance of genes influencing quantitative characteristics and the inheritance of genes influencing discontinuous characteristics is in the number of loci that determine the characteristic. When multiple loci affect a character, more genotypes are possible; so the relation between the genotype and the phenotype is less obvious. As the number of loci affecting a character increases, the number of phenotypic classes in the F<sub>2</sub> increases.

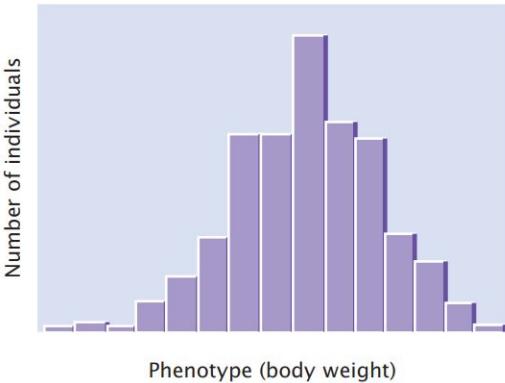


24.5 The results of crossing individuals heterozygous for different numbers of loci affecting a characteristic.

# Statistical Methods Required for Analyzing Quantitative Characteristics.

Because quantitative characteristics are described by a measurement and are influenced by multiple factors, their inheritance must be analyzed statistically.

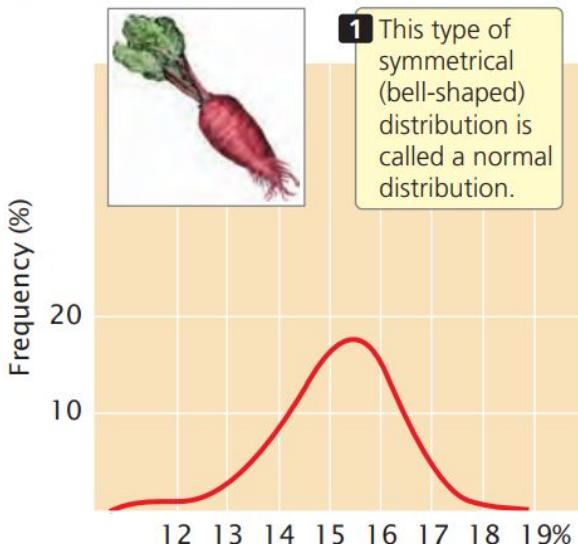
## DISTRIBUTIONS



Phenotypic variation in a group can be conveniently represented by a frequency distribution, which is a graph of the frequencies (numbers or proportions) of the different phenotypes

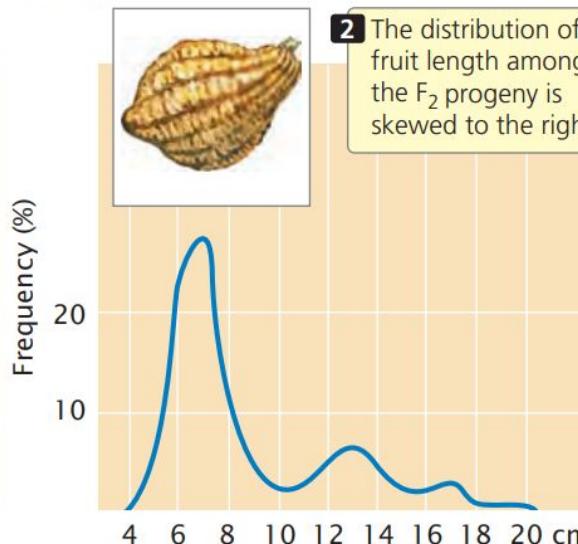
# Statistical Methods Required for Analyzing Quantitative Characteristics.

(a) Sugar beet percentage of sucrose



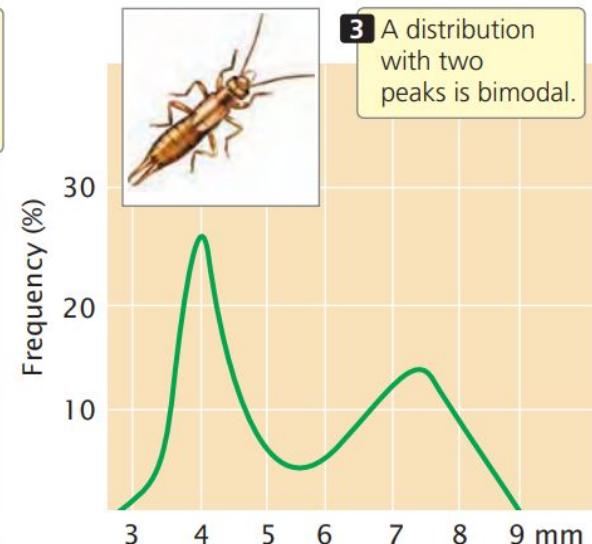
1 This type of symmetrical (bell-shaped) distribution is called a normal distribution.

(b) Squash fruit length



2 The distribution of fruit length among the F<sub>2</sub> progeny is skewed to the right.

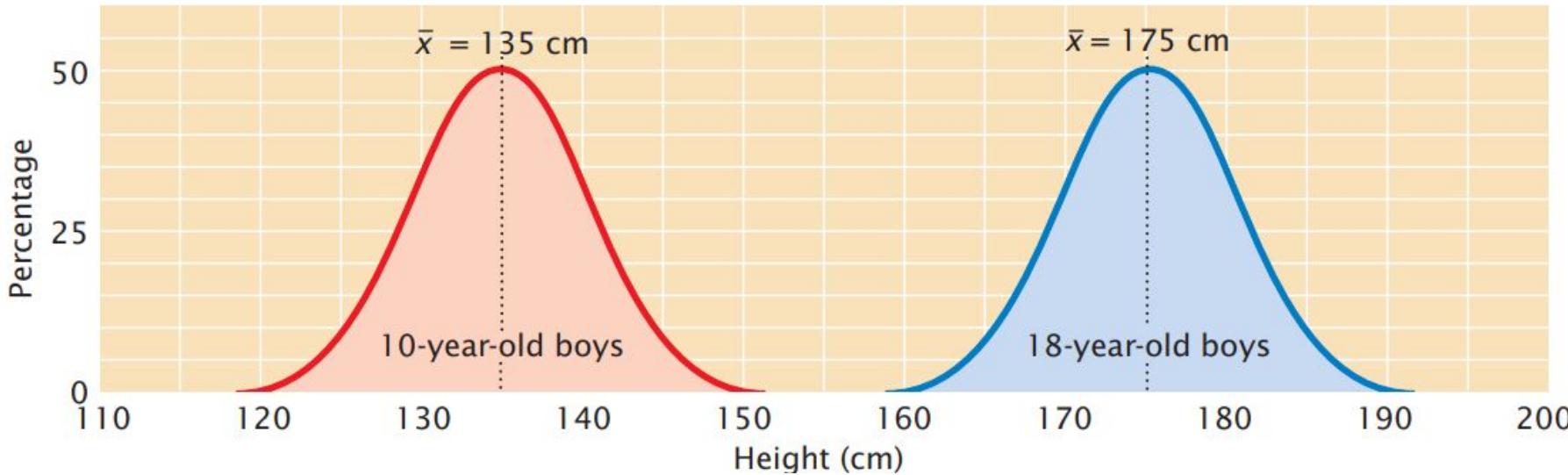
(c) Earwig forceps length



3 A distribution with two peaks is bimodal.

Normal distributions arise when a large number of independent factors contribute to a measurement, as is often the case in quantitative characteristics.

## Samples and Populations

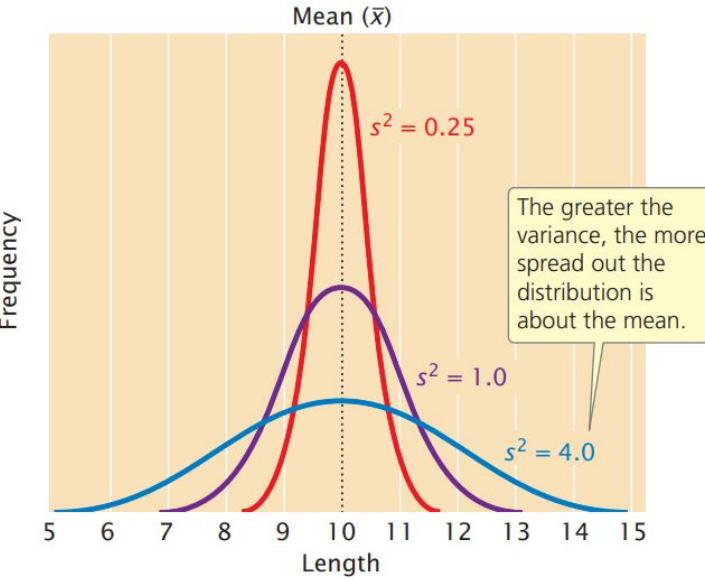


**The mean provides information about the center of a distribution.**

# Samples and Populations

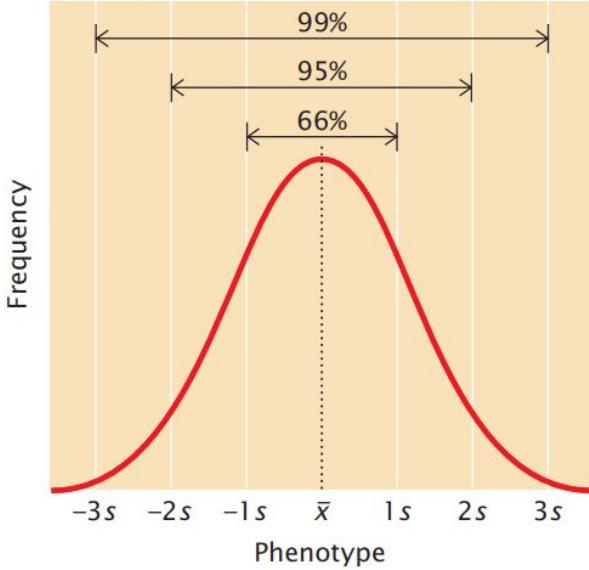
The variance ( $s^2$ ) is defined as the average squared deviation from the mean:

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} \quad (24.4)$$



**The variance provides information about the variability of a group of phenotypes.**

# Samples and Populations



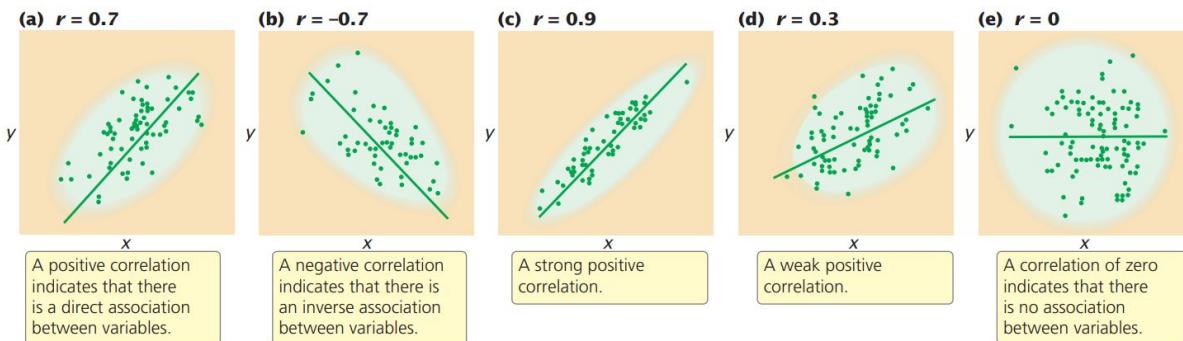
The proportions of a normal distribution occupied by plus or minus one, two, and three standard deviations from the mean.

# correlation coefficient

The mean and the variance can be used to describe an individual characteristic, but geneticists are frequently interested in more than one characteristic.

Often, two or more characteristics vary together.

$$\text{COV}_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$



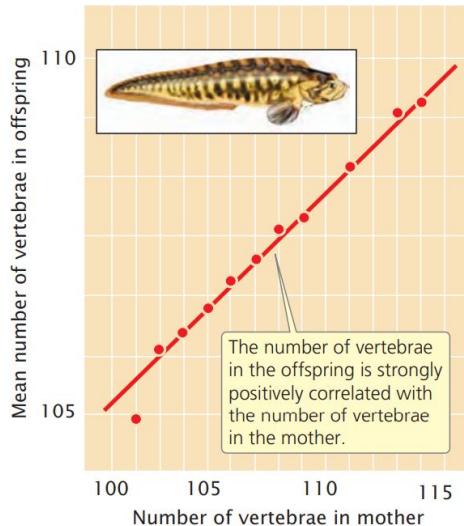
$$r = \frac{\text{COV}_{xy}}{s_x s_y}$$

24.11 The correlation coefficient describes the relation between two or more variables.

## correlation coefficient

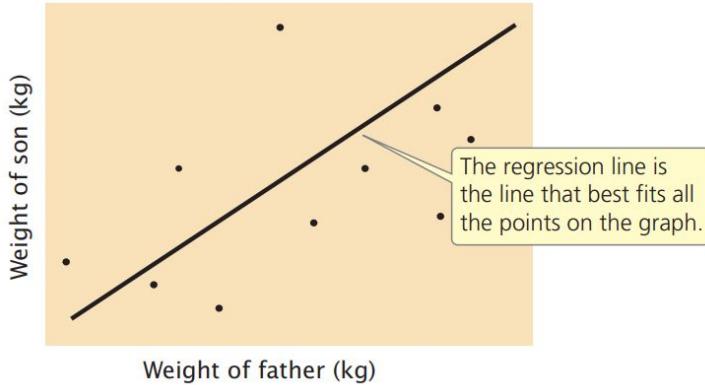
A correlation coefficient can be computed for two variables measured for the same individual, such as height (x) and weight (y).

A correlation coefficient can also be computed for a single variable measured for pairs of individuals. For example, we can calculate for fish the correlation between the number of vertebrae of a parent (x) and the number of vertebrae of its offspring (y),



**A correlation coefficient can be computed for a single variable measured for pairs of individuals.**

# Statistical prediction: regression.



Correlation provides information only about the strength and direction of association between variables. However, we often want to know more than just whether two variables are associated; we want to be able to predict the value of one variable, given a value of the other.

Above Figure depicts a regression of the weights of fathers against the weights of sons. Each father-son pair is represented by a point on the graph: the x value of a point is the father's weight and the y value of the point is the son's weight.

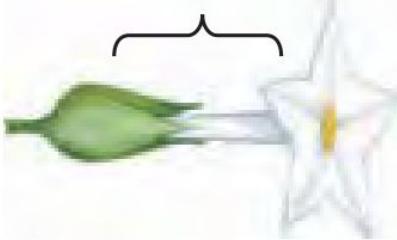
# Applying Statistics to the Study of a Polygenic Characteristics

Edward East carried out one early statistical study of polygenic inheritance on the length of flowers in tobacco (*Nicotiana longiflora*).

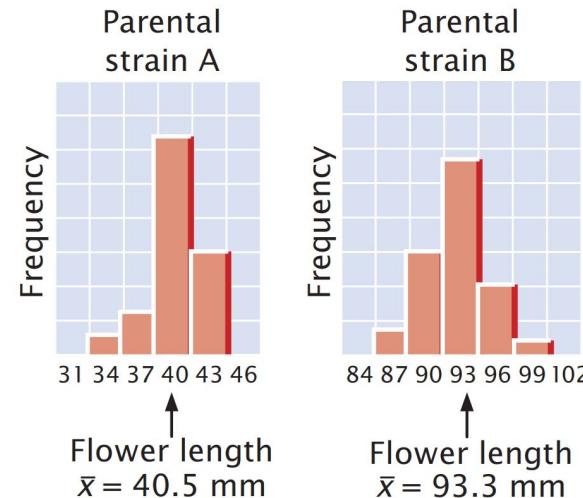
- Two varieties had been inbred for many generations and were homozygous at all loci contributing to flower length.
- Thus, there was no genetic variation in the original parental strains; the small differences in flower length within each strain were due to environmental effects on flower length.

# EAST experiment

Flower length



## P generation

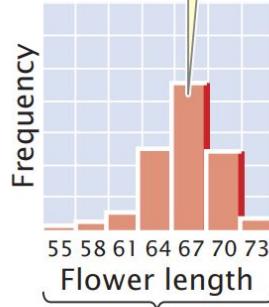


Two varieties had been inbred for many generations and were homozygous at all loci contributing to flower length.

# EAST experiment

## F<sub>1</sub> generation

- 1 Flower length in the F<sub>1</sub> was about halfway between that in the two parents,...



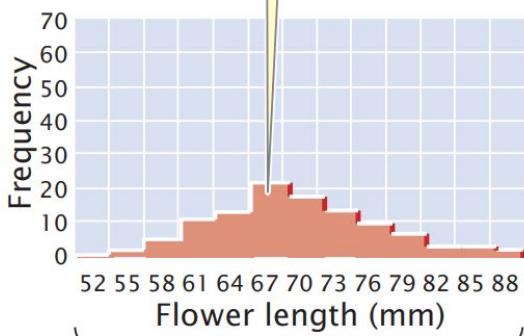
- 2 ...and the variance in the F<sub>1</sub> was similar to that seen in the parents.

The variance of flower length in the F1 was similar to that seen in the parents, because all the F1 had the same genotype, as did each parental strain

# EAST experiment

## F<sub>2</sub> generation

- 3 The mean of the F<sub>2</sub> was similar to that observed for the F<sub>1</sub>,...



This greater variability indicates that not all of the F<sub>2</sub> progeny had the same genotype.

- 4 ...but the variance in the F<sub>2</sub> was greater, indicating the presence of different genotypes among the F<sub>2</sub> progeny.

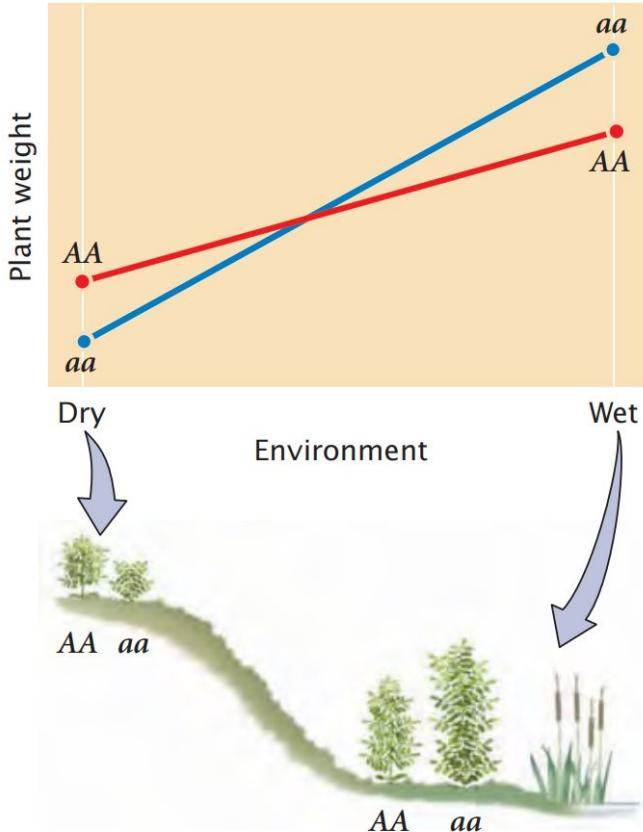
# Phenotypic Variance

- To determine how much of phenotypic differences in a population is due to genetic and environmental factors, we must first have some quantitative measure of the phenotype under consideration.
- Consider a population of wild plants that differ in size.
- We could collect a representative sample of plants from the population, weigh each plant in the sample, and calculate the mean and variance of plant weight.
- **This phenotypic variance is represented by VP.**
- Components of phenotypic variance First, some of the phenotypic variance may be due to differences in genotypes among individual members of the population. **These differences are termed the genetic variance and are represented by VG.**

# Phenotypic Variance

- Second, some of the differences in phenotype may be due to environmental differences among the plants; **these differences are termed the environmental variance, VE.**
- Third, **genetic-environmental interaction variance (VGE)** arises when the effect of a gene depends on the specific environment in which it is found.

# Phenotypic Variance



Genetic-environmental interaction variance is obtained when the effect of a gene depends on the specific environment in which it is found.

In summary, the total phenotypic variance can be apportioned into three components:

$$V_P = V_G + V_E + V_{GE} \quad (24.11)$$

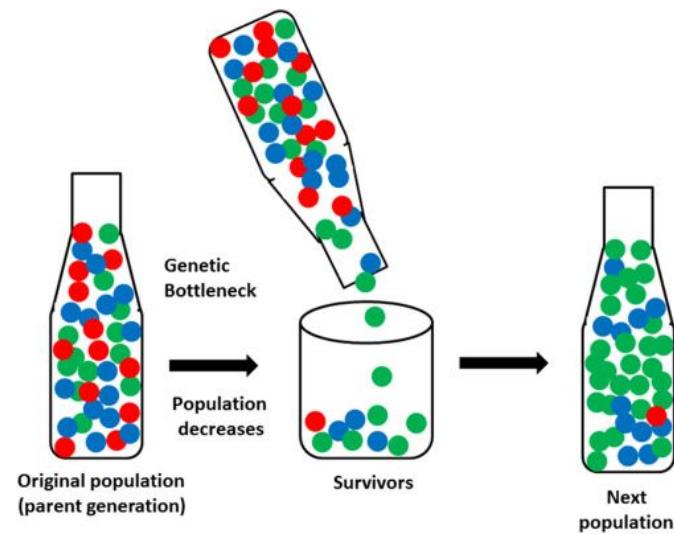
# Population Genetics

# Population Genetics



*Rocky Mountain bighorn sheep (*Ovis canadensis*)*

## Genetic Drift



## **Genotypic and Allelic Frequencies Are Used to Describe the Gene Pool of a Population**

- An obvious and pervasive feature of life is variability.
- Genetic Variation is more complex wrt to Phenotypic Variation e.g. two members of a population can produce the same protein even if their DNA sequences are different. DNA sequences between the genes and introns within genes do not encode proteins; much of the variation in these sequences also has little effect on the phenotype.



## **Calculating Genotypic Frequencies**

- A frequency is simply a proportion or a percentage, usually expressed as a decimal fraction. For example, if 20% of the alleles at a particular locus in a population are A, we would say that the frequency of the A allele in the population is 0.20.
- For large populations, for which a determination of the genes of all individual members is impractical, a sample of the population is usually taken and the genotypic and allelic frequencies are calculated for this sample
- The genotypic and allelic frequencies of the sample are then used to represent the gene pool of the population

## Calculating Genotypic Frequencies

- Genotypic frequency, we simply add up the number of individuals possessing the genotype and divide by the total number of individuals in the sample (N).
- For a locus with three genotypes AA, Aa, and aa, the frequency ( $f$ ) of each genotype is

$$f(AA) = \frac{\text{number of } AA \text{ individuals}}{N}$$

$$f(Aa) = \frac{\text{number of } Aa \text{ individuals}}{N}$$

$$f(aa) = \frac{\text{number of } aa \text{ individuals}}{N}$$

**The sum of all the genotypic frequencies always equals 1.**

# Calculating Allelic Frequencies

- The gene pool of a population can also be described in terms of the allelic frequencies.
- There are always fewer alleles than genotypes; so the gene pool of a population can be described in fewer terms when the allelic frequencies are used.

**Allelic frequencies can be calculated from (1) the numbers or (2) the frequencies of the genotypes.**

$$\text{frequency of an allele} = \frac{\text{number of copies of the allele}}{\text{number of copies of all alleles at the locus}}$$

For a locus with only two alleles (A and a), the frequencies of the alleles are usually represented by the symbols p and q, and can be calculated as follows:

$$p = f(A) = \frac{2n_{AA} + n_{Aa}}{2N}$$

$$q = f(a) = \frac{2n_{aa} + n_{Aa}}{2N}$$

where  $n_{AA}$ ,  $n_{Aa}$ , and  $n_{aa}$  represent the numbers of AA, Aa, and aa individuals, and N represents the total number of individuals in the sample. We divide by  $2N$  because each diploid individual has two alleles at a locus. The sum of the allelic frequencies always equals 1 ( $p + q = 1$ ); so, after  $p$  has been obtained,  $q$  can be determined by subtraction:  $q = 1 - p$ .

## **Allelic frequencies can be calculated from the genotypic frequencies**

To calculate an allelic frequency from genotypic frequencies, we add the frequency of the homozygote for each allele to half the frequency of the heterozygote (because half of the heterozygote's alleles are of each type):

$$p = f(A) = f(AA) + \frac{1}{2}f(Aa)$$

$$q = f(a) = f(aa) + \frac{1}{2}f(Aa)$$

**We obtain the same values of p and q whether we calculate the allelic frequencies from the numbers of genotypes or from the genotypic frequencies.**

## **The Hardy–Weinberg Law Describes the Effect of Reproduction on Genotypic and Allelic Frequencies**

The primary goal of population genetics is to understand the processes that shape a population's gene pool.

First, we must ask what effects reproduction and Mendelian principles have on the genotypic and allelic frequencies: How do the segregation of alleles in gamete formation and the combining of alleles in fertilization influence the gene pool ?

The law is actually a mathematical model that evaluates the effect of reproduction on the genotypic and allelic frequencies of a population. It makes several simplifying assumptions about the population and provides two key predictions if these assumptions are met.

**Assumptions** If a population is large, randomly mating, and not affected by mutation, migration, or natural selection, then:

- **Prediction 1** the allelic frequencies of a population do not change; and
- **Prediction 2** the genotypic frequencies stabilize (will not change) after one generation in the proportions  $p^2$  (the frequency of AA),  $2pq$  (the frequency of Aa), and  $q^2$  (the frequency of aa), where p equals the frequency of allele A and q equals the frequency of allele a.

**The Hardy–Weinberg law indicates that, when the assumptions are met, reproduction alone does not alter allelic or genotypic frequencies and the allelic frequencies determine the frequencies of genotypes.**

		Sperm	
		$A$	$a$
		$p$	$q$
Eggs	$A$	$AA$ $p \times p = p^2$	$Aa$ $q \times p = pq$
	$a$	$Aa$ $p \times q = pq$	$aa$ $q \times q = q^2$

$f(A) = p$   
 $f(a) = q$   
  
 $f(AA) = p^2$   
 $f(Aa) = 2pq$   
 $f(aa) = q^2$

**Conclusion:** Random mating will produce genotypes of the next generation in proportions  $p^2(AA)$ ,  $2pq(Aa)$ , and  $q^2(aa)$

**25.2 Random mating produces genotypes in the proportions  $p^2$ ,  $2pq$ , and  $q^2$ .**