

Statistical Essentials for Health Data Science

Dileepa Ediriweera

MBBS, MSc(BioStat), MSc(Biomed Info), PhD

Outline

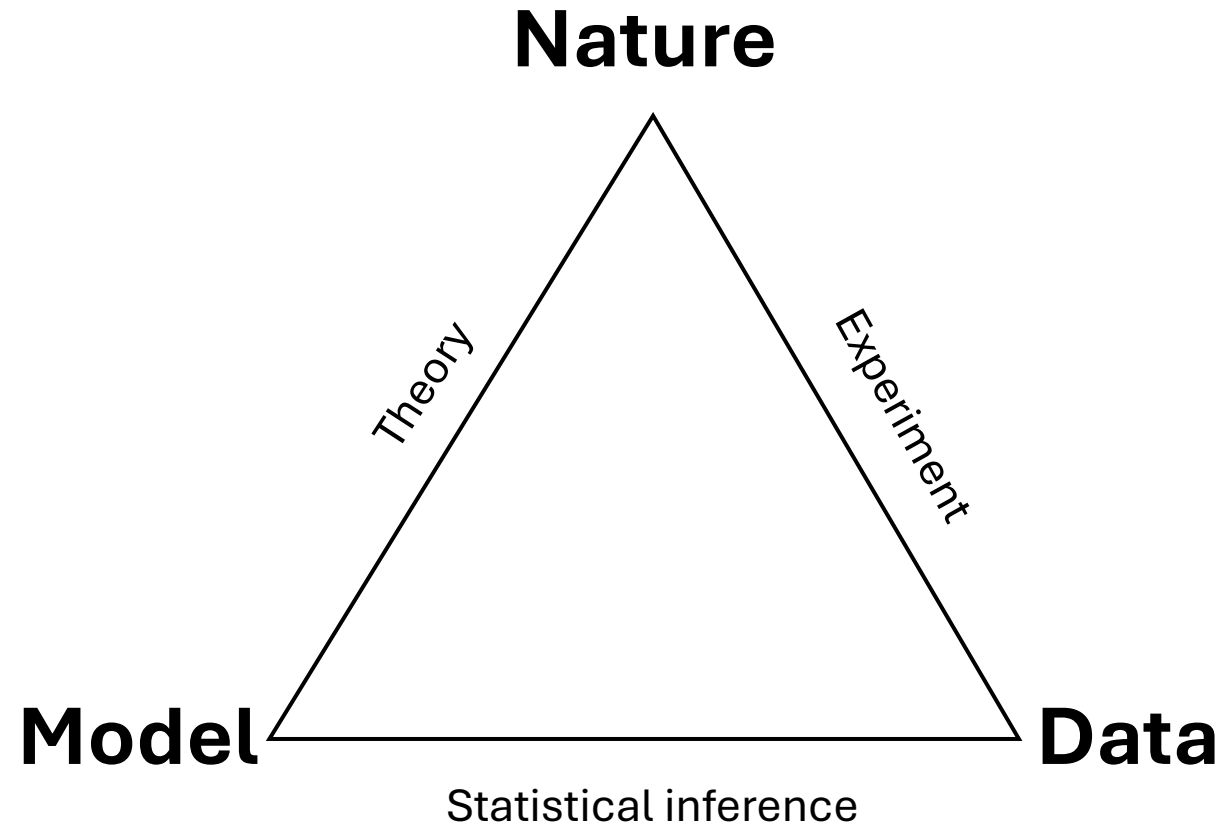
- Session 1: Describing the Health Data
- Session 2: Making Informed Decisions
- Session 3: Exploring Relationships and Prediction

Why we need science?

Science

- Goal is to understand nature
- Two pillars of the scientific method
 - Theory
 - Observations
- Theory – predicts how a natural process should behave
- Observations (controlled experiment or direct observation of the natural world) – tells us whether the theory is correct

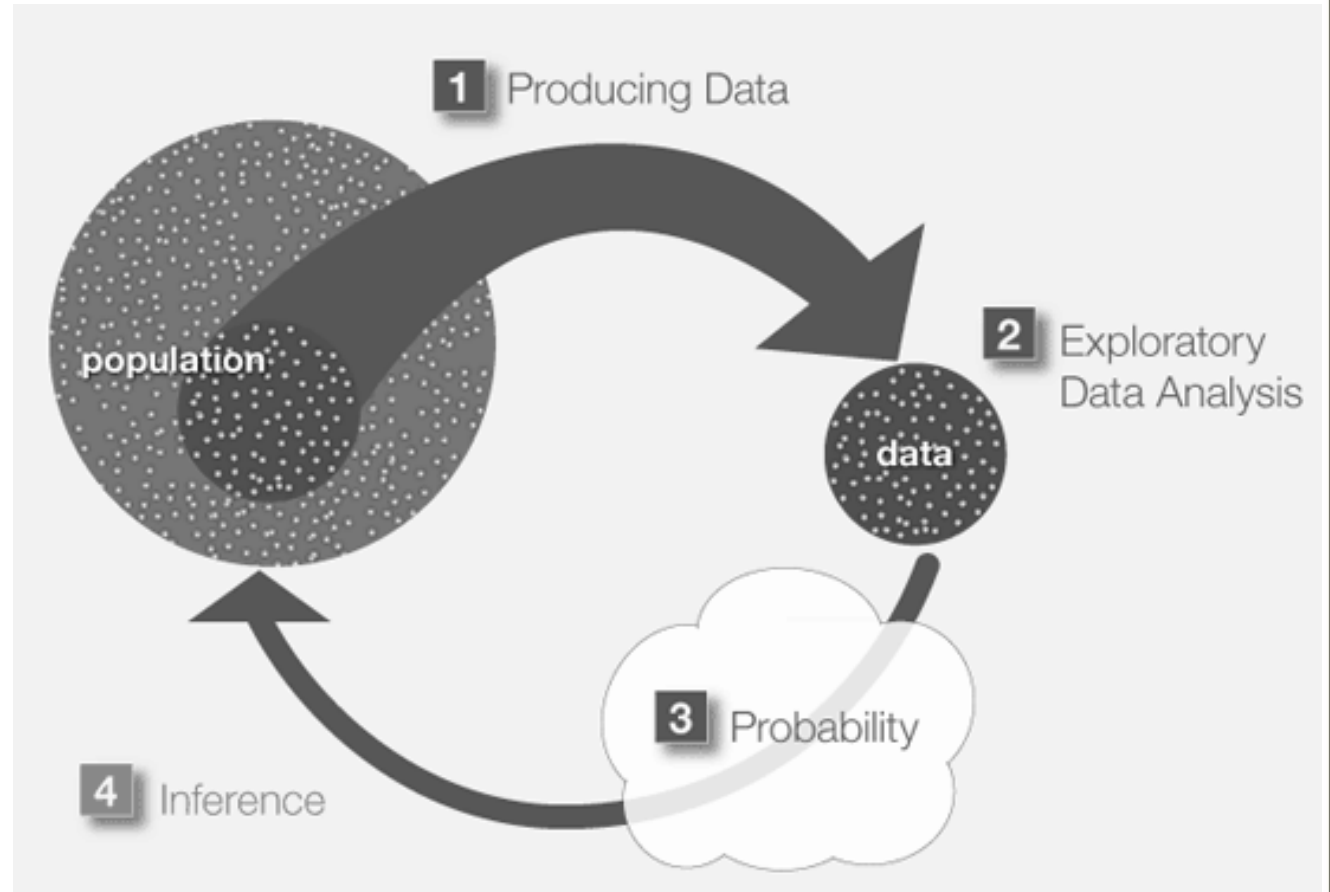
Role of statistics within scientific method



Session 1: Describing the Health Data

Statistics

- Collection of data
- Analysis
- Making inference



Areas of statistics

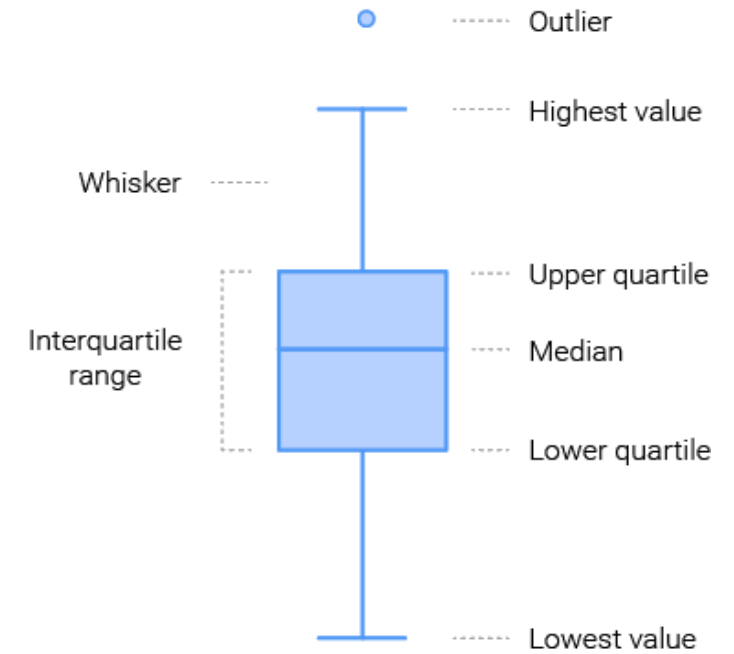
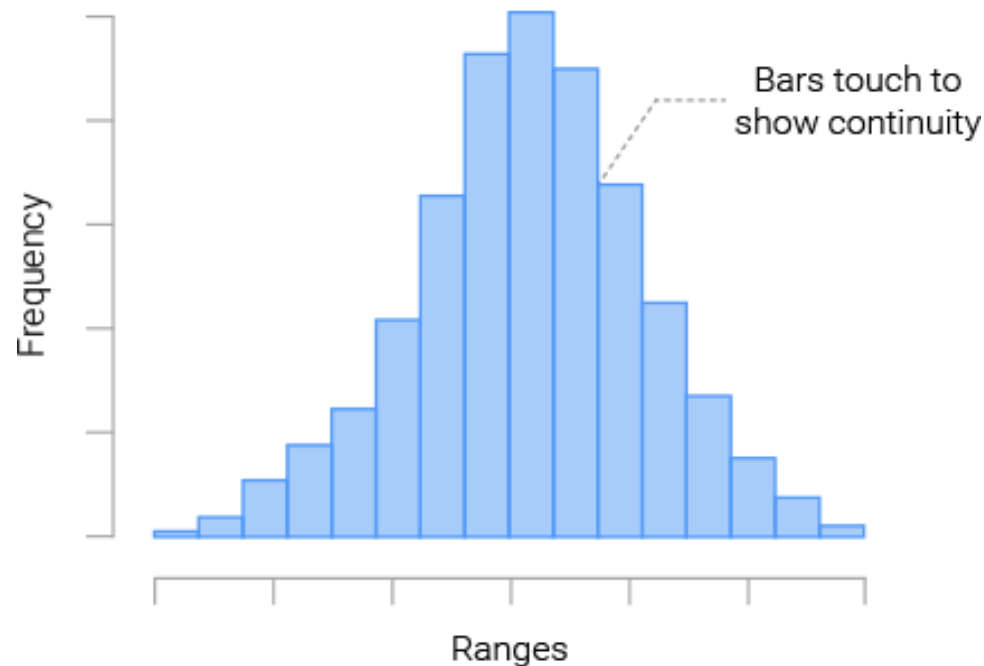
- Descriptive statistics
 - Used to summarize, organize and present data in a convenient and communicable form
 - Average science marks of a class
 - Number of students for a given age groups
- Inferential statistics
 - Techniques that allow us to make **inferences or conclusions about a population** based on data that are gathered from a sample
 - This is done either;
 - Estimate parameters (e.g. population mean)
 - Hypothesis testing (e.g. effectiveness of a new drug)

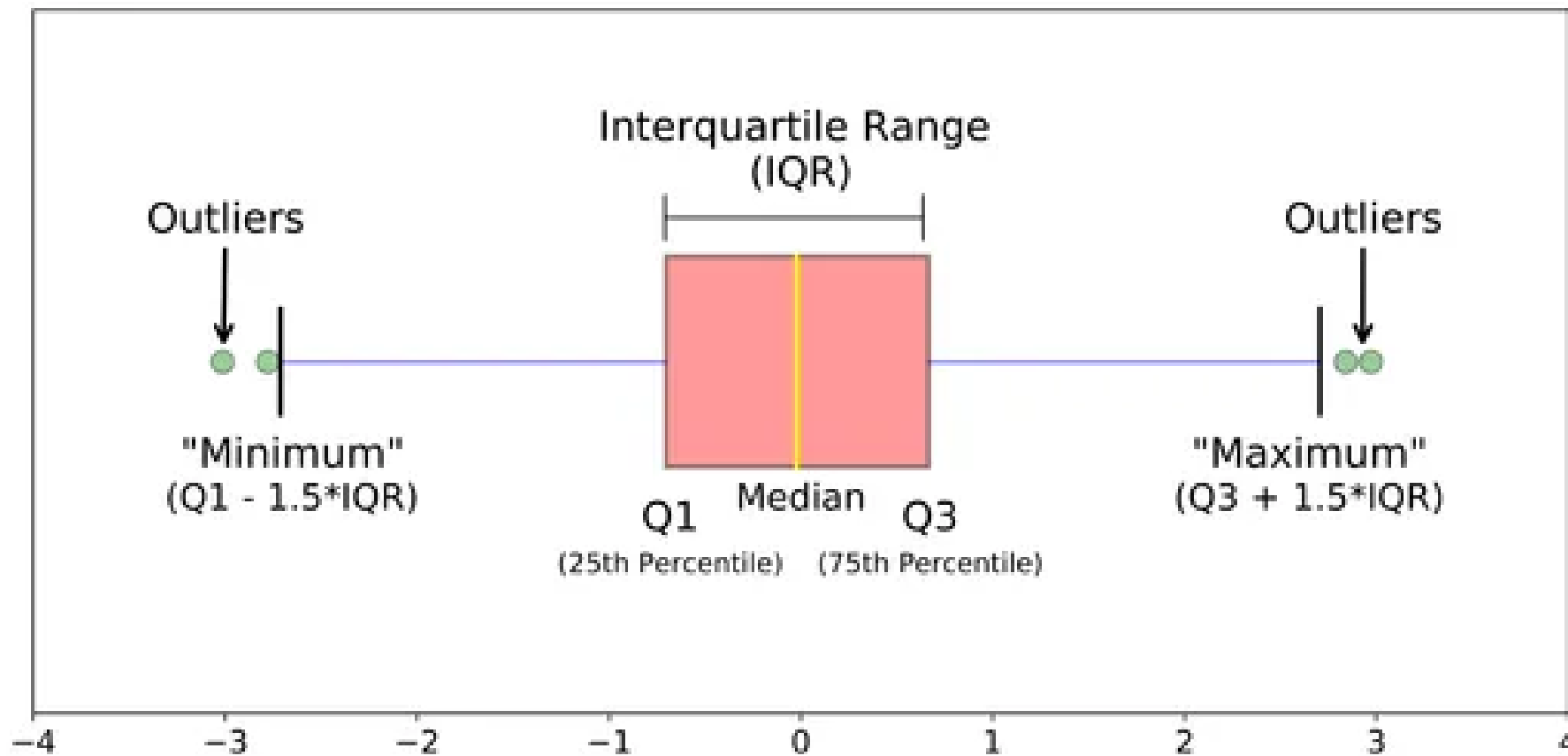
Descriptive statistics

- To describe the distribution of sample data (what data show)
- Graphically
 - Histogram, boxplot, bar charts, pie charts
- Numerically
 - Distribution (frequency distribution table)
 - Central tendency (centre of location)
 - Mean, median, mode
 - Dispersion
 - Standard deviation, range, interquartile range

Graphical illustrations – single variable

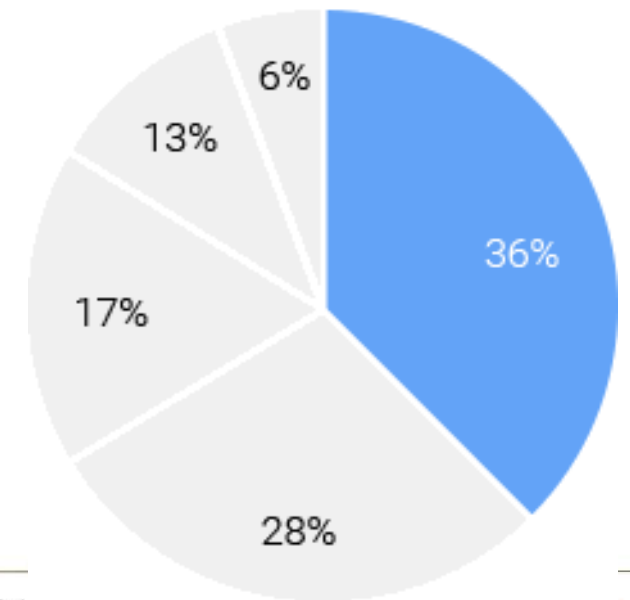
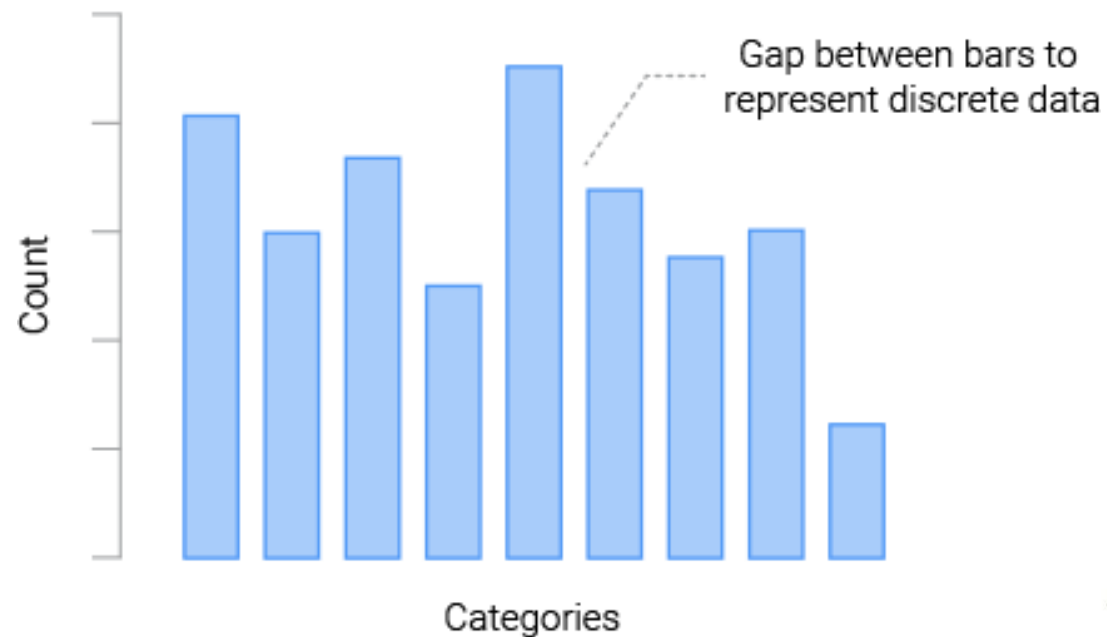
- Continuous data?





Graphical illustrations – single variable

- Categorical data ?



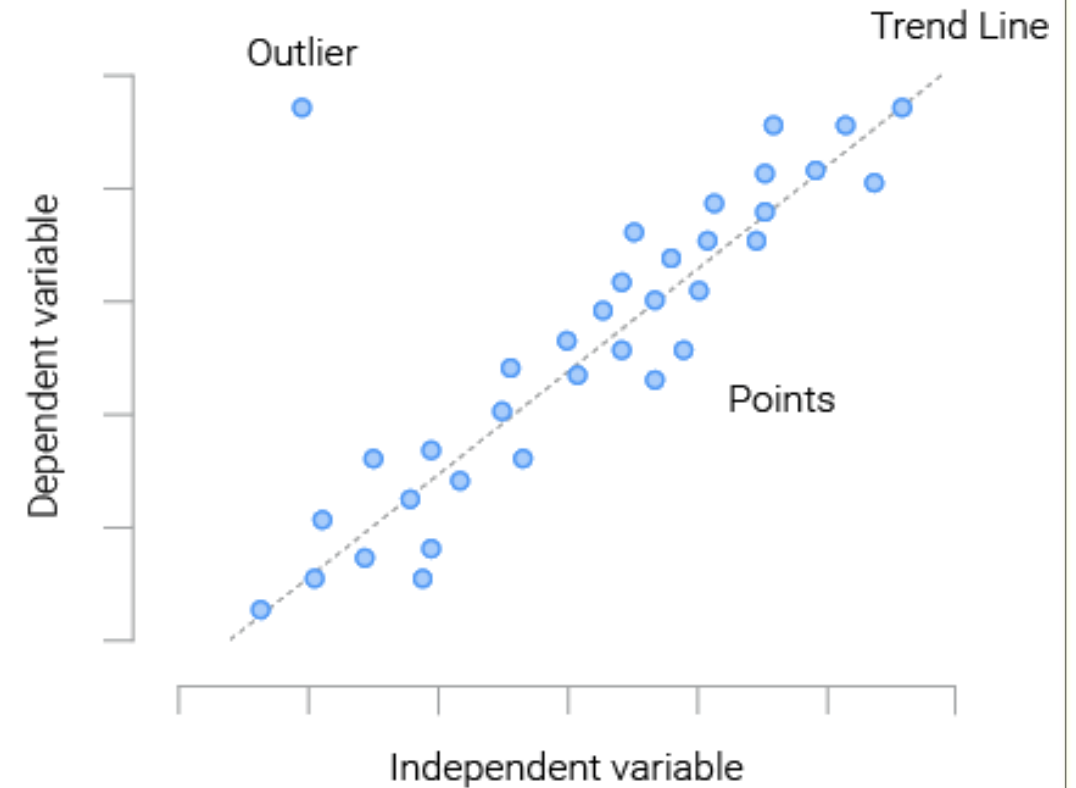
Frequency distribution table

Degree	Frequency
High School	2
Bachelor's	7
MBA	20
Master's	3
Law	4
PhD	4

Range	Frequency
0-39	12
40-79	6
80-119	2
120-159	3
160-199	5
200-240	2

Graphical illustrations – two variables

- How are two variables correlated?

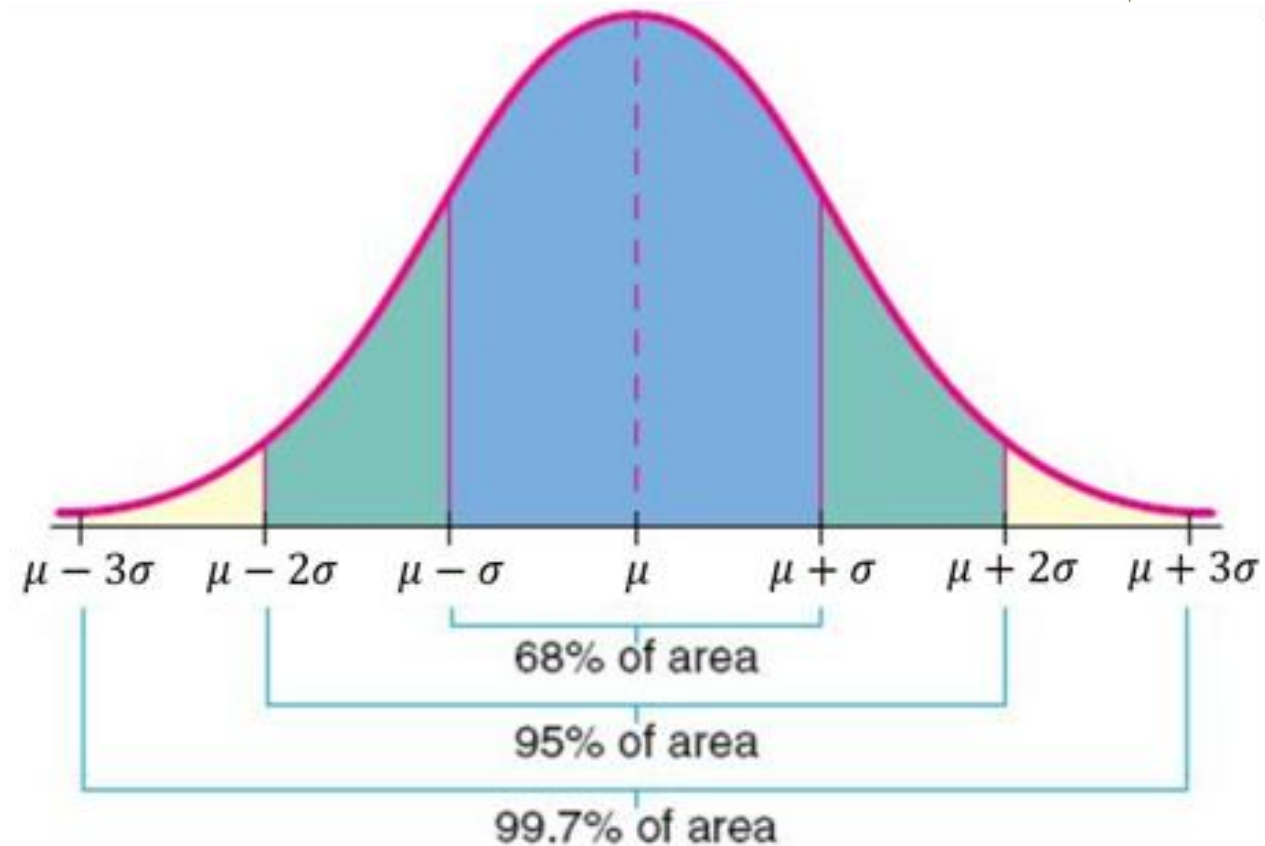


Data distributions

- We need to identify the distribution that best fit the data (and to specify the parameters)
 - Theoretical framework
- Continuous distributions
 - Normal, Lognormal, Exponential, Uniform, Cauchy, Weibull
- Discrete distributions
 - Binomial, Poisson, Negative Binomial, Discrete uniform, Geometric

Normal Distribution (theoretical)

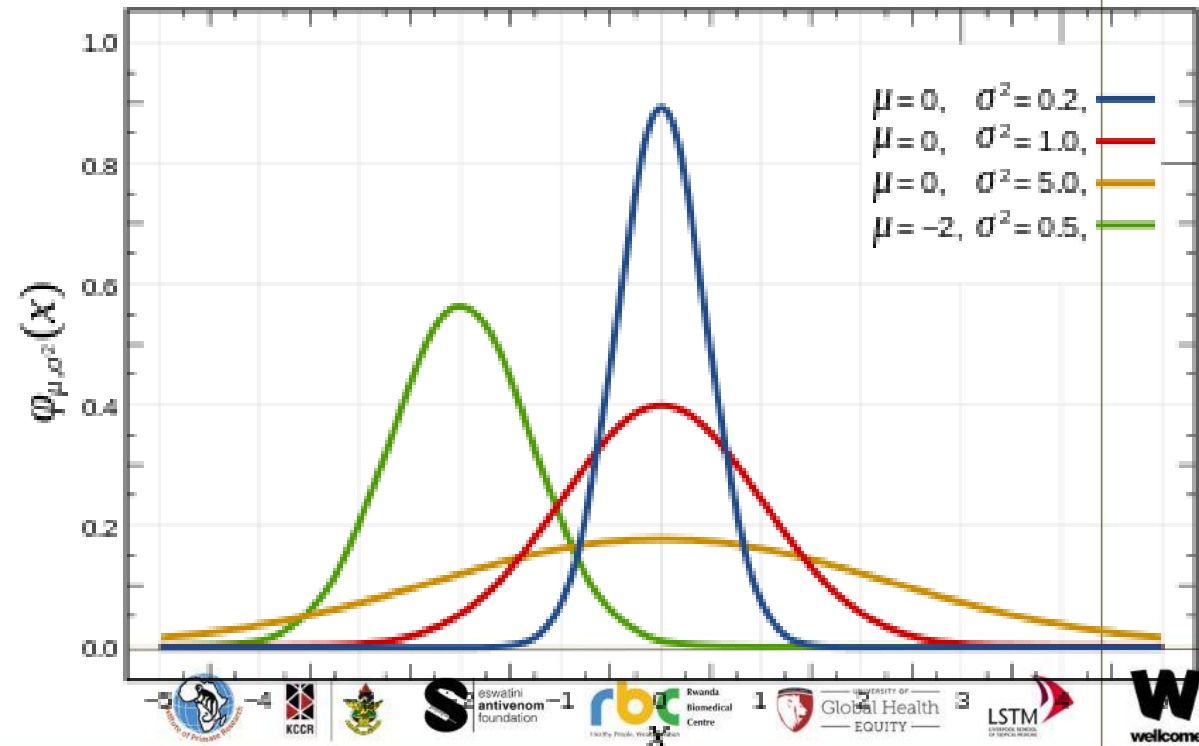
- Bell shape ("Bell Curve")
- Goes from $-\infty$ to $+\infty$
- Mean = Median = Mode
- Symmetry around the centre
 - 50% of values less than the mean
 - 50% of values greater than the mean
 - 68% of values are within 1 standard deviation of the mean
 - 95% of values are within 2 standard deviations of the mean
 - 99.7% of values are within 3 standard deviations of the mean



Statistical distribution has parameters

- Parameters describe the shape of a distribution
 - E.g. Normal distribution
 - $X \sim N(\mu, \sigma)$

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



github.com/DileepaE

Measures of Central Tendency

- Mean
 - Add up all the numbers and divide by number of observations
- Median
 - The middle number (order the numbers and find the actual middle number or average of the two numbers if not)
- Mode
 - Most commonly occurring number
- Activity
 - 17, 18, 20, 21, 21, 24, 23, 21, 15, 19
 - 17, 18, 23, 20, 21, 24, 23, 20, 20, 15, 19, 20
 - 17, 18, 23, 20, 21, 24, 23, 20, 20, 15, 19, 20, 60

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i$$

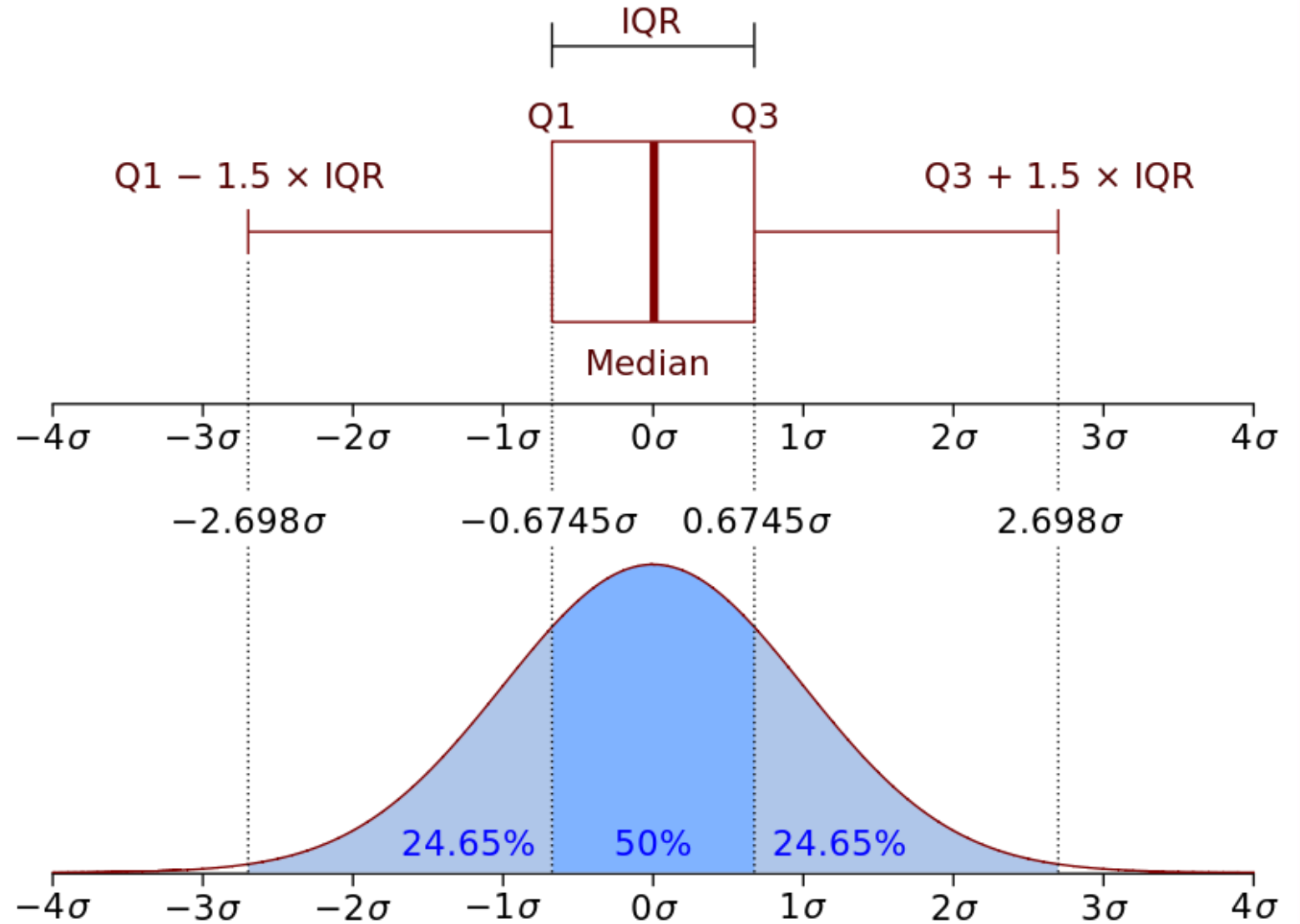
Measures of Dispersion

- Range
 - Maximum – minimum (this is different to reporting the max and min)
- Variance (σ^2)
 - Average squared deviation from the mean
- Standard Deviation (σ)
 - Square root of variance

$$\sigma^2 = \frac{\sum (x - \mu)^2}{N}$$

Measures of Dispersion

- Quartile :
 - A division of observations into four (quarters) of equal size
- Interquartile range
 - First quartile (Q1) : middle number between the smallest and median
 - Second quartile (Q2) : median (50% of the data lies below this point)
 - Third quartile (Q3) : middle value between the median and the highest
 - Q1 – Q3 (this is reporting Q1 and Q3)



Selecting most suitable statistic

- If data has a normal distribution
 - Mean (SD)
 - Sensitive to outliers
- If data does not have a normal distribution?
 - Median (IQR)
 - Not sensitive to outliers

Introduction to data simulation

- Assume mean (SD) of FBS in general population is 100 (20) mg/dl
- Simulate 20 FBS values from this distribution

```
rnorm(n, mean = 0, sd = 1)
```

n	number of observations. If length(n) > 1, the length is taken to be the number required.
mean	vector of means.
sd	vector of standard deviations.

```
y <- rnorm(20,100,20)
```

```
y
```

```
> y
```

```
[1] 124.33053 74.75877 69.40745 149.12174 143.60878 107.27091 106.09180 93.16760 65.99009 110.09239  
[11] 92.44645 124.89783 93.32005 74.17525 67.33445 120.53812 101.79306 114.33832 99.01170 103.23623
```

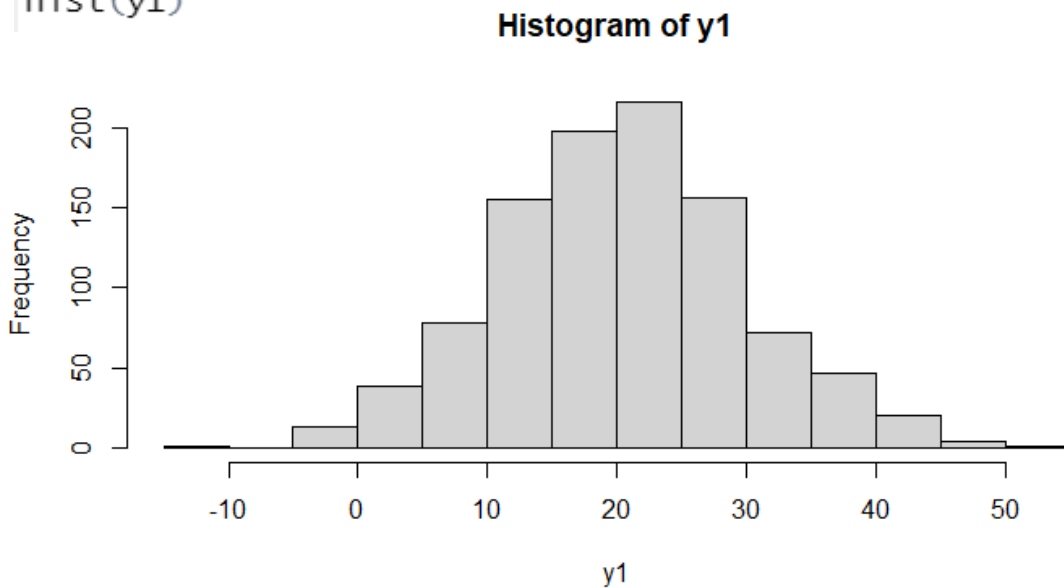

Testing for normality

- To determine if a data set is well-modeled by a normal distribution
- Graphical methods
 - Histogram
 - Quantile-quantile (QQ) plot of the standardized data against the standard normal distribution
- Statistical tests
 - Shapiro–Wilk test

Testing for normality

```
y1 <- rnorm(1000,20,10)
```

```
hist(y1)
```

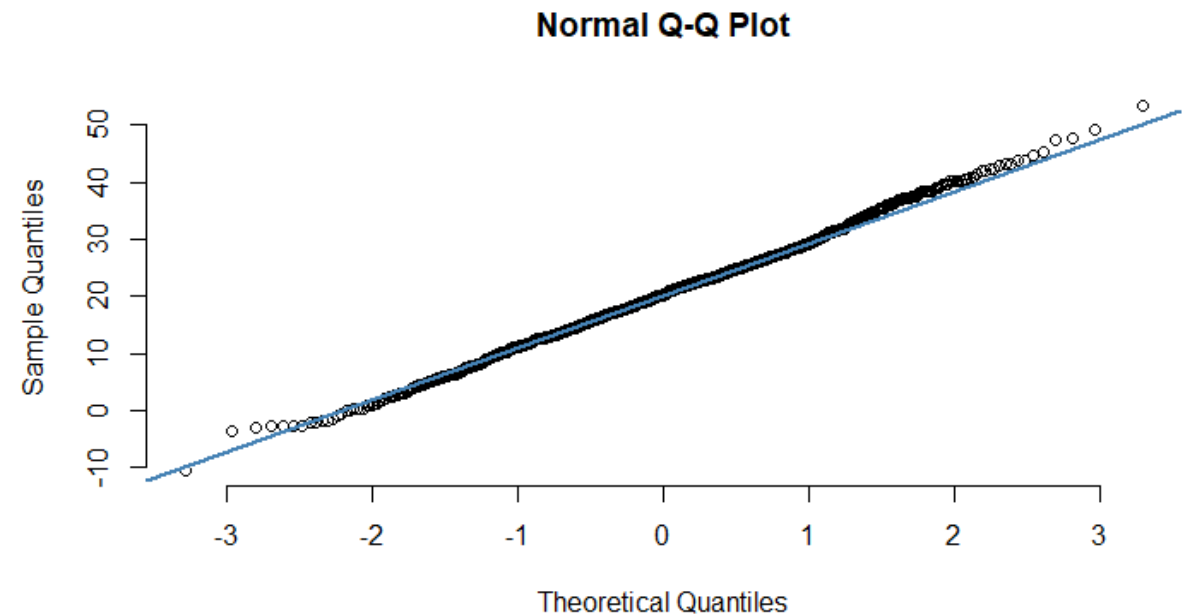


```
> shapiro.test(y1)
```

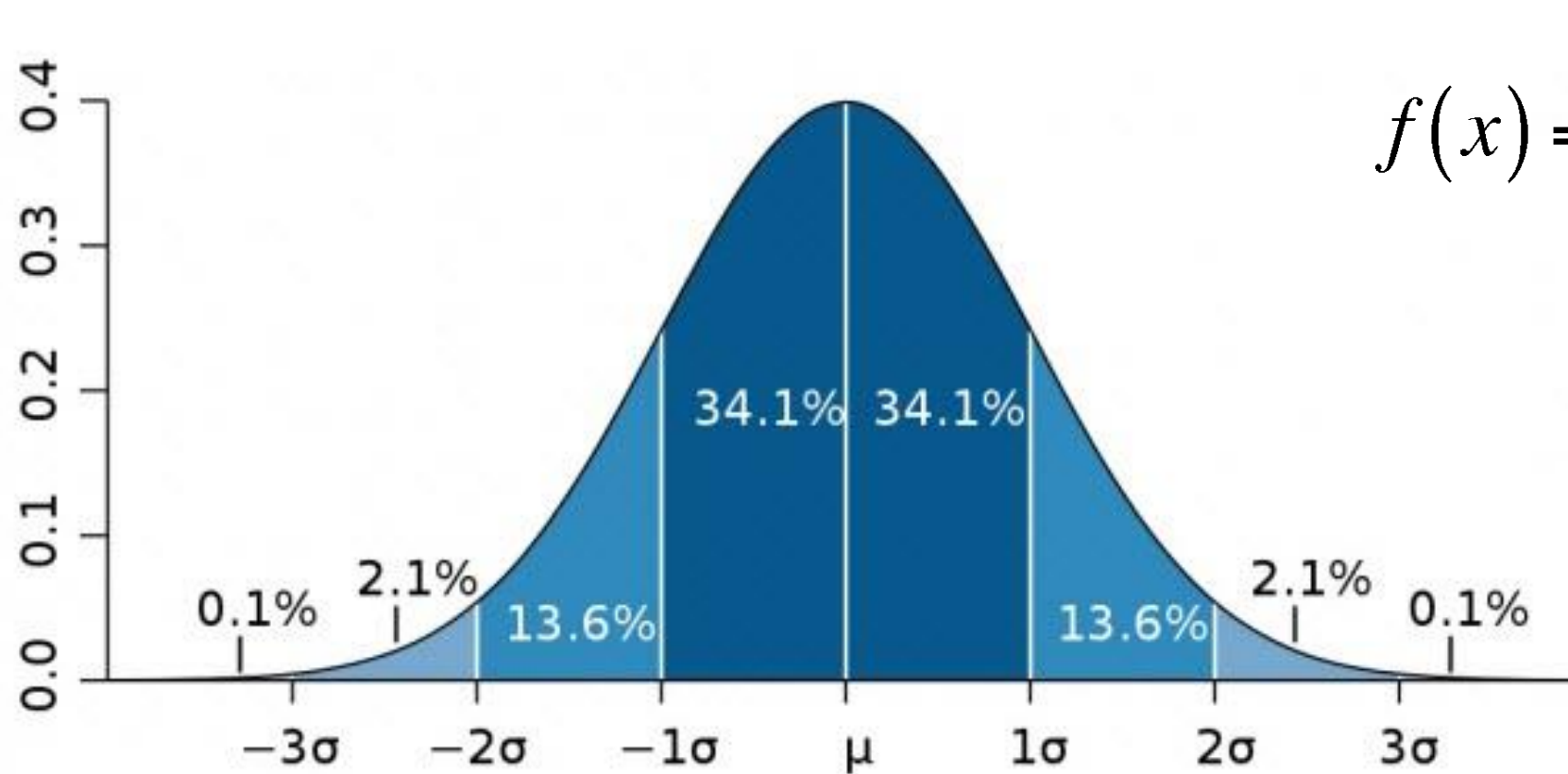
Shapiro-Wilk normality test

```
data: y1  
W = 0.98454, p-value = 0.2939
```

```
qqnorm(y1, pch = 1, frame = FALSE)  
qqline(y1, col = "steelblue", lwd = 2)
```



Normal distribution



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

Standard normal distribution

- Is a normal distribution
- Bell shape
- Total area under curve = 1
- Area:
 - $-1 < Z < +1 \rightarrow 68\%$
 - $-2 < Z < +2 \rightarrow 95\%$
- Probability
 - $P(Z < 0) = 0.5$

Why SND?

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

$$= \frac{1}{1 \times \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-0}{1}\right)^2} \quad -\infty \leq x \leq \infty$$

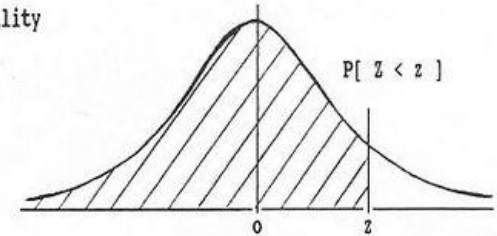
$$= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z)^2} \quad -\infty \leq z \leq \infty$$

$$\left(\frac{x-0}{1} = z \right)$$

1. Areas under the Normal Distribution

The table gives the cumulative probability up to the standardised normal value z i.e.

$$P[Z < z] = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right) dz$$

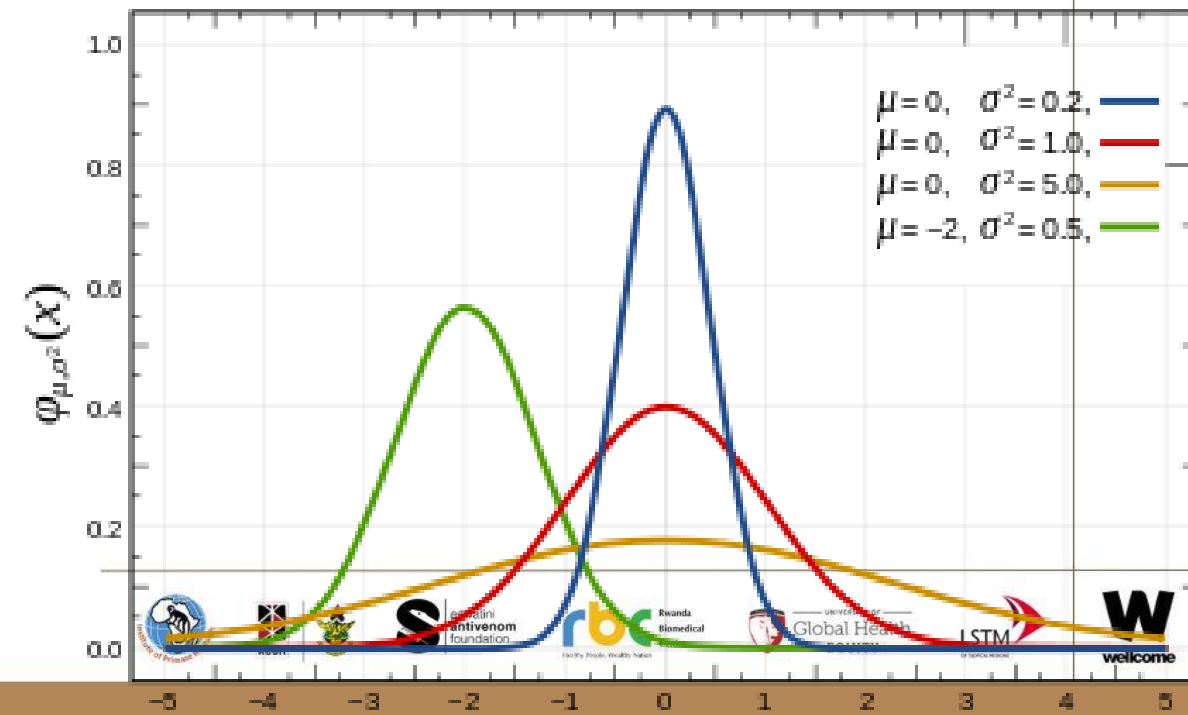


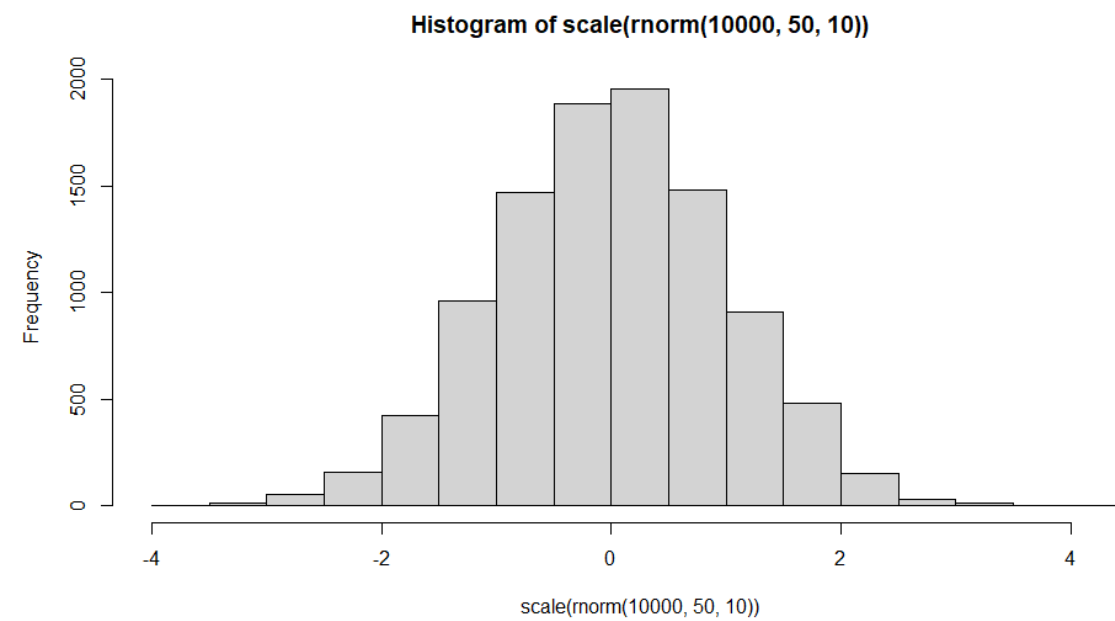
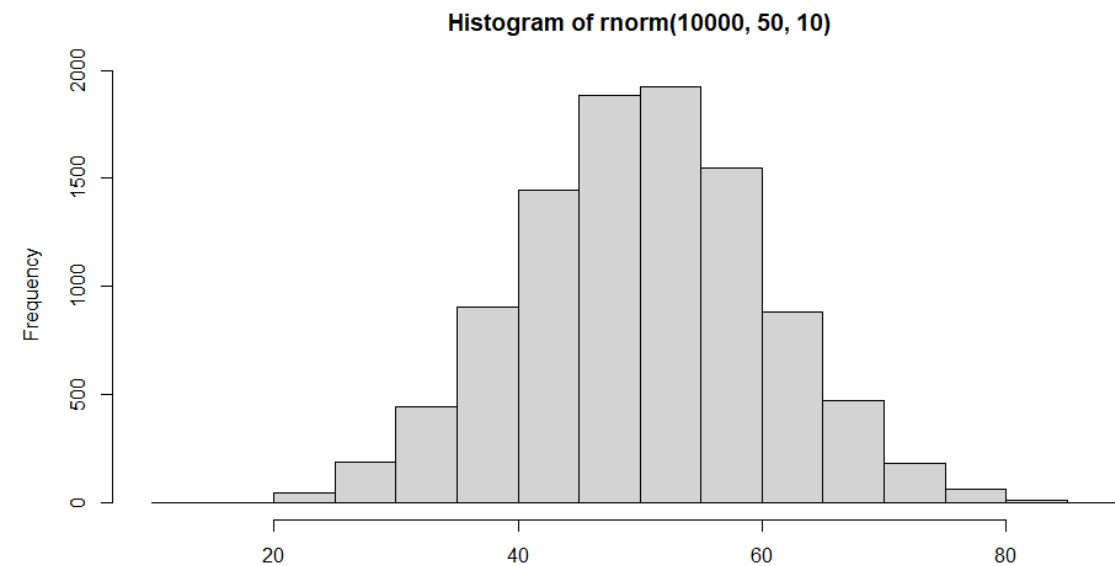
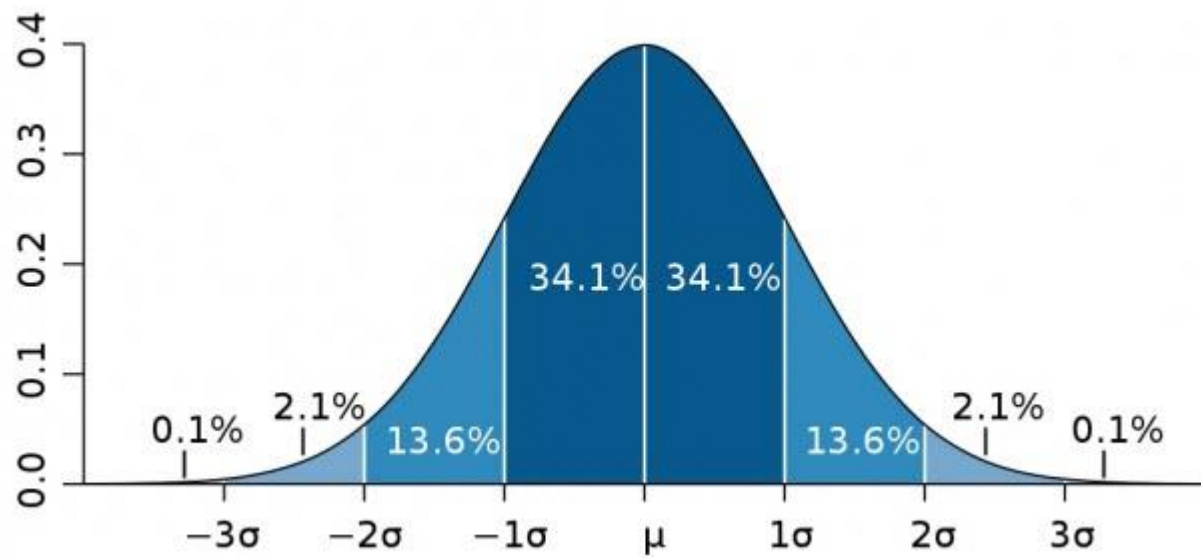
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5159	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7854
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8804	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9773	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9865	0.9868	0.9871	0.9874	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9924	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9980	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
z	3.00	3.10	3.20	3.30	3.40	3.50	3.60	3.70	3.80	3.90
P	0.9986	0.9990	0.9993	0.9995	0.9997	0.9998	0.9998	0.9999	0.9999	1.0000

Example

- Blood pressure $X \sim (120, 10)$
 - $P(X > 140)$?
- Blood pressure $X \sim (150, 12)$
 - $P(X > 140)$?

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





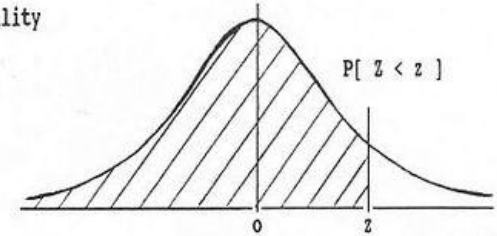
Exercise

- $P(Z < 0.5)$
- $P(Z > 0.5)$
- $P(Z < 0.4)$
- $P(Z < 1.35)$
- $P(0.4 < Z < 1.35)$
- $P(Z < -1)$

1. Areas under the Normal Distribution

The table gives the cumulative probability up to the standardised normal value z i.e.

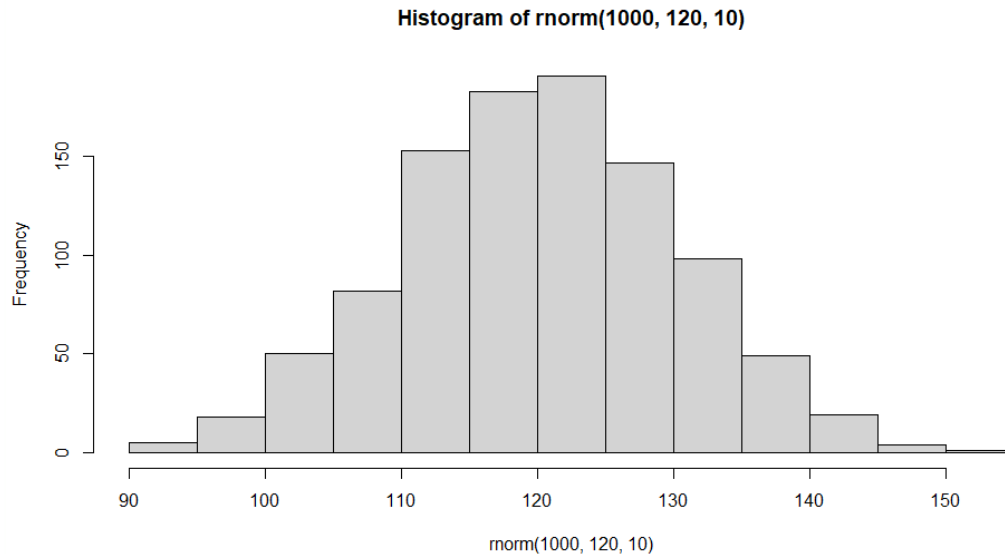
$$P[Z < z] = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}z^2) dz$$



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5159	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7854
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8804	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9773	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9865	0.9868	0.9871	0.9874	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9924	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9980	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
z	3.00	3.10	3.20	3.30	3.40	3.50	3.60	3.70	3.80	3.90
P	0.9986	0.9990	0.9993	0.9995	0.9997	0.9998	0.9998	0.9999	0.9999	1.0000

Example

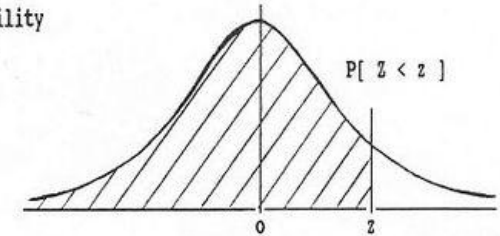
- Blood pressure $X \sim N(120, 10)$
 - $P(X > 140)$?



1. Areas under the Normal Distribution

The table gives the cumulative probability up to the standardised normal value z i.e.

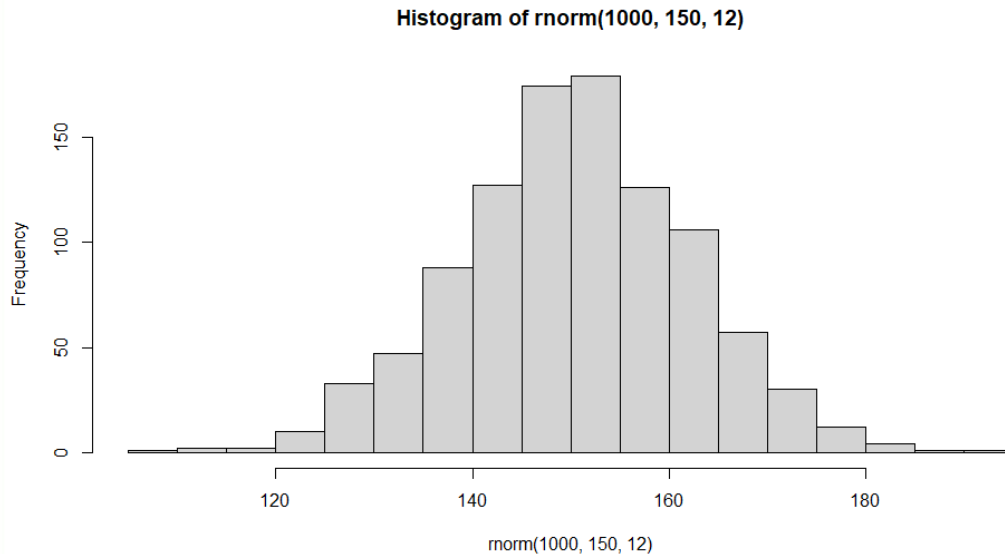
$$P[Z < z] = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right) dz$$



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5159	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7854
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8804	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9773	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9865	0.9868	0.9871	0.9874	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9924	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9980	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
z	3.00	3.10	3.20	3.30	3.40	3.50	3.60	3.70	3.80	3.90
P	0.9986	0.9990	0.9993	0.9995	0.9997	0.9998	0.9998	0.9999	0.9999	1.0000

Example

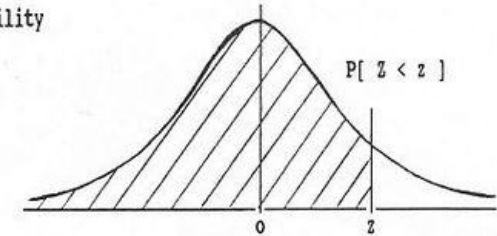
- Blood pressure $X \sim N(150, 12)$
 - $P(X > 140)$?



1. Areas under the Normal Distribution

The table gives the cumulative probability up to the standardised normal value z i.e.

$$P[Z < z] = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right) dz$$



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5159	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7854
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8804	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9773	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9865	0.9868	0.9871	0.9874	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9924	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9980	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
z	3.00	3.10	3.20	3.30	3.40	3.50	3.60	3.70	3.80	3.90
P	0.9986	0.9990	0.9993	0.9995	0.9997	0.9998	0.9998	0.9999	0.9999	1.0000

Data types

Activity

- Identifying data types from common health datasets

Scale of measurement (types) of data

1. Nominal/categorical
2. Ordinal
3. Interval
4. Ratio

Offers:	Nominal	Ordinal	Interval	Ratio
The sequence of variables is established	–	Yes	Yes	Yes
Mode	Yes	Yes	Yes	Yes
Median	–	Yes	Yes	Yes
Mean	–	–	Yes	Yes
Difference between variables can be evaluated	–	–	Yes	Yes
Addition and Subtraction of variables	–	–	Yes	Yes
Multiplication and Division of variables	–	–	–	Yes
Absolute zero	–	–	–	Yes



Session 2:

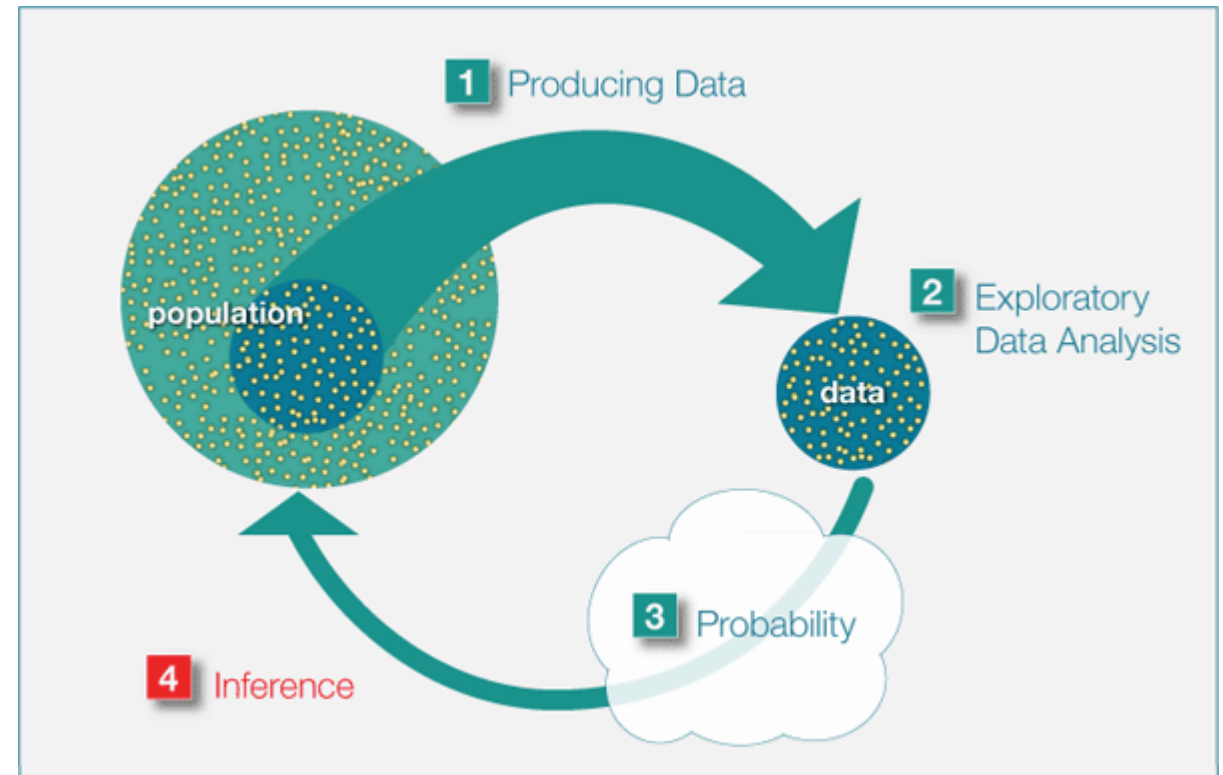
Making Informed Decisions

Inferential statistics : Part I

Inferential statistics

Statistics

- Collection of data
- Analysis
- Making inference



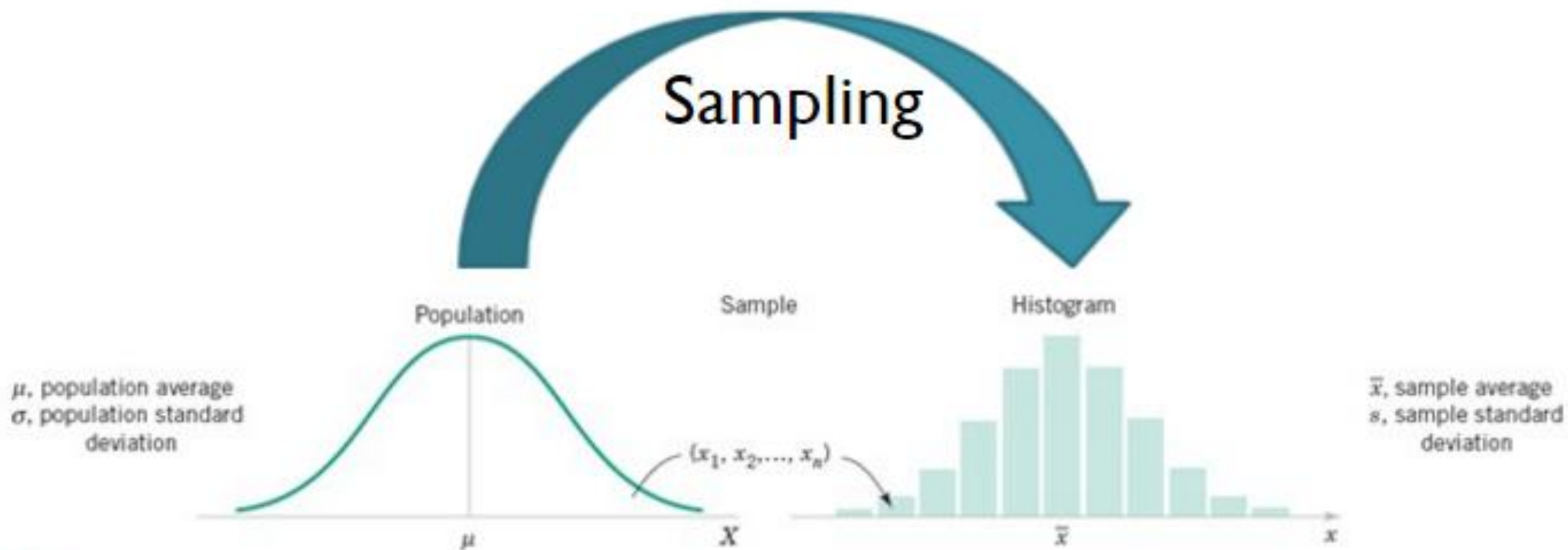
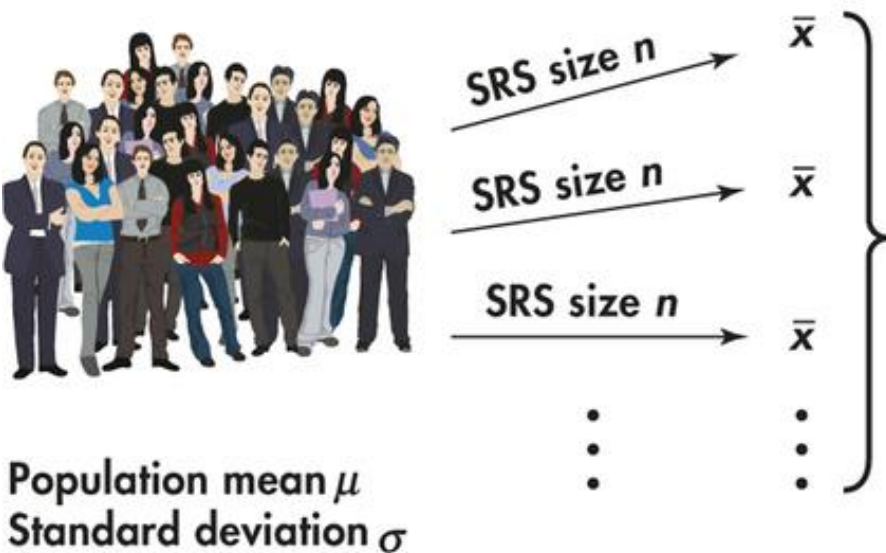


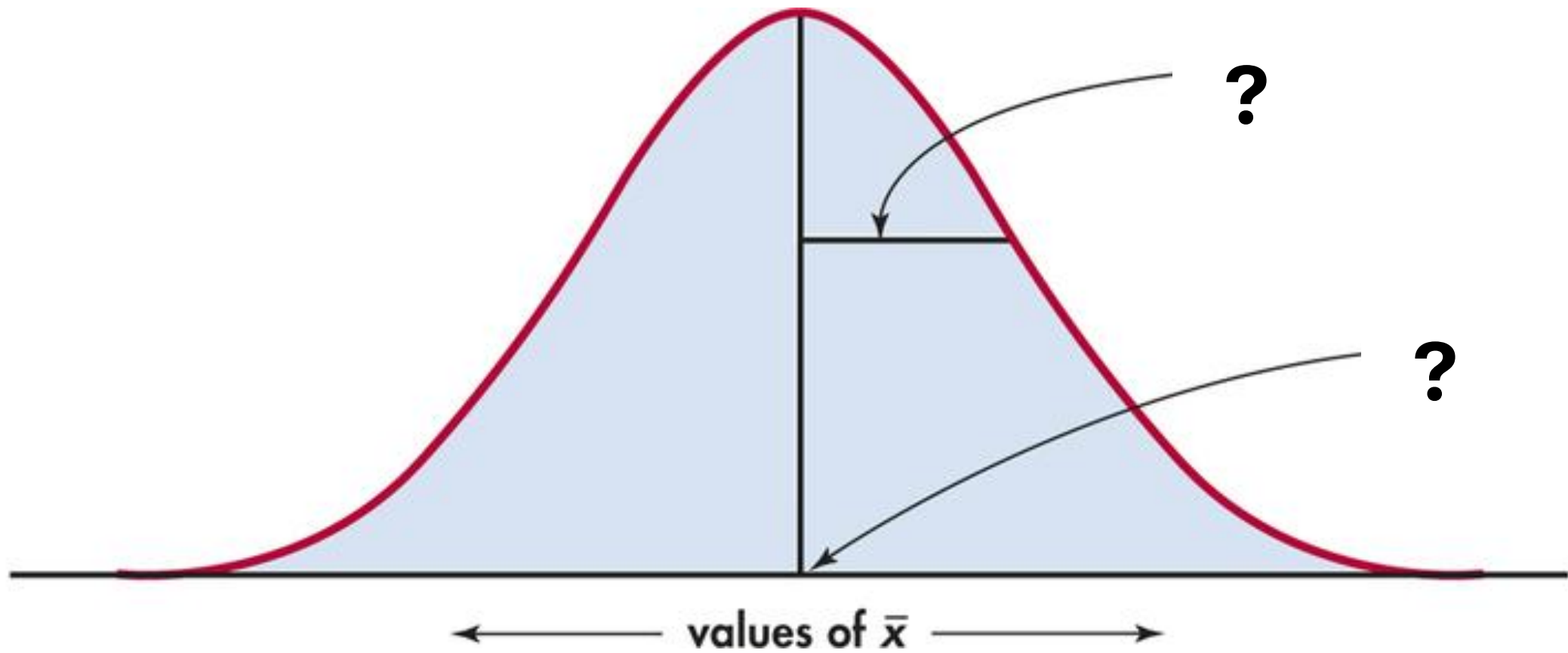
Figure 4-1 Relationship between a population and a sample.

Population vs Sample

- Population value vs Sample value
 - Parameter vs statistic
- Notations
 - Greek vs English



Sampling distribution



Variables (experimental designs)

- Independent
 - ?
- Dependent
 - ?

Variables (experimental designs)

- Independent
 - Cause (independent of other variables in the study)
 - Assumed to have a direct effect on the dependent variable
 - Experimenter has control (we can manipulate)
- Dependent
 - Effect/outcome (depends on changes in the independent variable)
 - Experimenter does not have control (we observe)

Outcome data (dependent variable)

- Continuous
 - Infinite number of possibilities within finite interval
 - E.g.
 - Height/weight/time
- Discrete
 - Fixed number of possibilities
 - E.g.
 - Outcome of a coin/disease status
- Basis for the statistical tests

Learning outcomes

- By the end of this session, you should be able to explain the methods of hypothesis testing
- By the end of this session, you should be able to test hypothesis in below situations
 - Categorical independent variable and continuous dependent variable
 - Categorical independent variable and categorical dependent variable
 - Continuous independent variable and continuous dependent variable
 - Continuous independent variable and categorical dependent variable

Hypothesis testing

- Start with an idea/imaginary value
- Do a study and test it
- Two ways
 - Draw confidence intervals
 - Perform a test

Inferential statistics

- Part 1: Independent variable : Categorical
- Part 2: Independent variable : Continuous

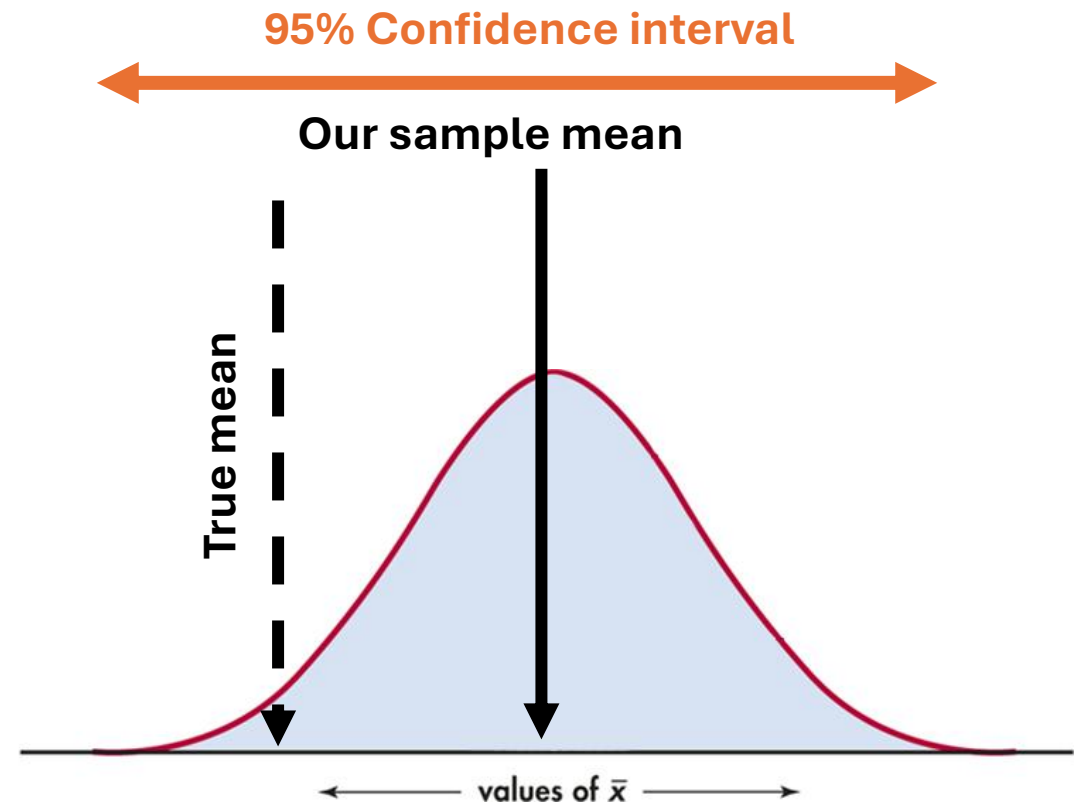
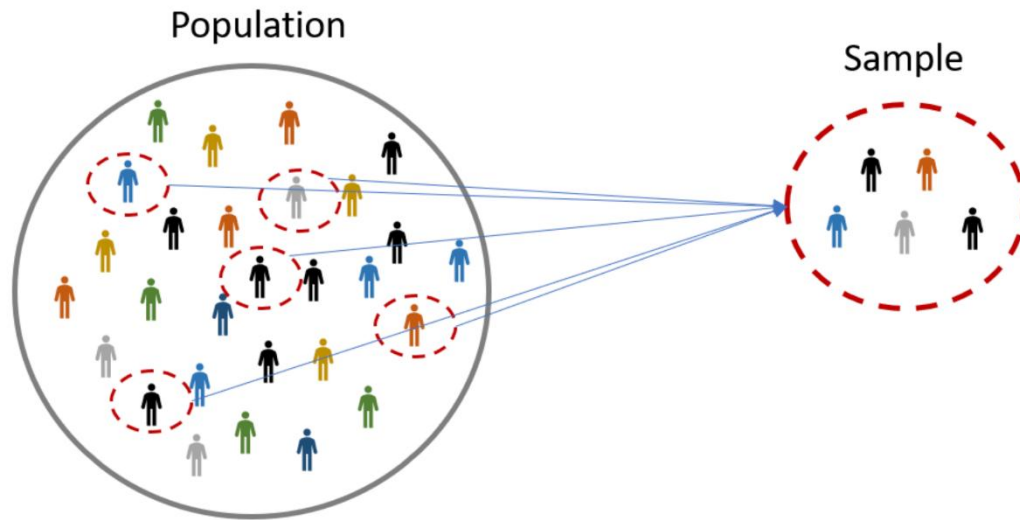
Part 1: Independent variable : Categorical

- What are the examples for categorical type independent variable?
- What are the examples for
 - Continuous type dependent (outcome) variables related to a categorical type independent variable?
 - Categorical type dependent (outcome) variables related to a categorical type independent variable?

Independent variable : Categorical (2 groups)

- Continuous outcomes (e.g. FBS)
 - To estimate a mean (one sample)
 - To compare a mean with a given value (one sample)
 - To compare two or more means (two or more samples)
- Categorical outcomes (e.g. DM)
 - To estimate a proportion (one sample)
 - To compare a proportion with a given value (one sample)
 - To compare two or more proportion (two or more samples)

One sample scenario



One sample – Estimate population parameter

- Continuous outcomes (e.g. estimate mean FBS level in a population)

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

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$$

$$\bar{x} \pm z \frac{s}{\sqrt{n}}$$

Task:

Imagine that you have randomly sampled 10000 people from SL. Assume the mean FBS of them is 115 and standard deviation is 20. Estimate the mean FBS value of the above population with 95% confidence. Hint: $z = 1.96$

One sample – Hypothesis testing

Continuous outcomes (e.g. Mean FBS in a population is 112 mg/dl?)

- 95% CI:

$$\bar{x} \pm z \frac{s}{\sqrt{n}}$$

Task:

Does mean FBS in the previous example (i.e. in the population where the 10000 people were selected) equal to 112 mg/dl?

- One sample t test :

- Null hypothesis $H_0 : \mu = m_0$
- Alternative Hypothesis $H_1 : \mu \neq m_0$

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$$

t.test {stats}

Student's t-Test

Description

Performs one and two sample t-tests on vectors of data.

Usage

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

```
## Default S3 method:
```

```
t.test(x, y = NULL,  
       alternative = c("two.sided", "less", "greater"),  
       mu = 0, paired = FALSE, var.equal = FALSE,  
       conf.level = 0.95, ...)
```

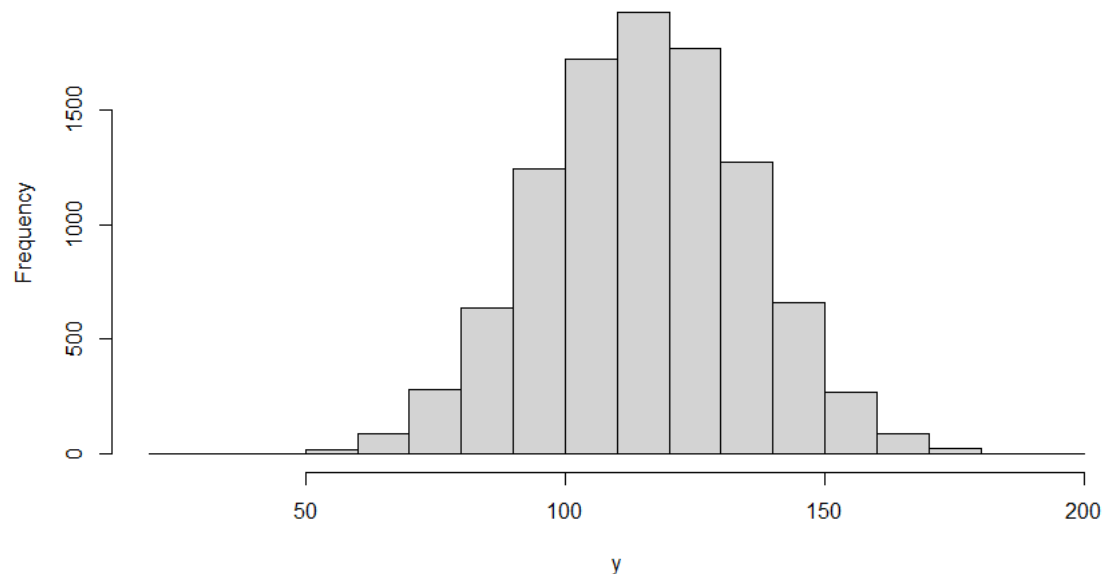
```
## S3 method for class 'formula'
```

```
t.test(formula, data, subset, na.action, ...)
```

Simulation example

```
y <- rnorm(10000, 115, 20)
```

Histogram of y



```
> summary(y)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
27.6	101.5	115.2	115.2	128.9	194.0

```
> t.test(y, mu=112)
```

One sample t-test

```
data: y
t = 15.373, df = 9999, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 112
95 percent confidence interval:
 114.6744 115.4561
sample estimates:
mean of x
 115.0653
```

What happens when sample size changes?

- Imagine this time you randomly sampled 100 people from SL.
- Assume the mean FBS is 115 and standard deviation is 20.
- Estimate the mean FBS value of the above population with 95% confidence.
- Does mean FBS in the previous example (i.e. in the population where the 100 people were selected) equal to 112 mg/dl?
- Hint: $z = 1.96$

```
> y <- rnorm(100,115,20)
> t.test(y,mu=112)
```

One sample t-test

```
data: y
t = 0.33569, df = 99, p-value = 0.7378
alternative hypothesis: true mean is not equal to 112
95 percent confidence interval:
 108.8652 116.4114
sample estimates:
mean of x
 112.6383
```

One sample – Estimate population parameter

- Binary outcomes (e.g. estimate the prevalence of DM in SL)

$$\hat{p} = \frac{M}{N}$$

$$SE = \sqrt{\frac{p(1 - p)}{n}}$$

$$\hat{p} \pm z \sqrt{\frac{\hat{p} (1 - \hat{p})}{n}}$$

Task:

Imagine that you have randomly sampled 10000 people from SL. Assume there are 3000 patients with diabetes. Estimate the prevalence of DM of the above population with 95% confidence. Hint: $z = 1.96$

One sample – Hypothesis testing

Binary outcomes: (e.g. DM prevalence in a population is 25%?)

- 95% CI:

$$\hat{p} \pm z \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

Task:

Does DM prevalence in the previous example (i.e. in the population where the 10000 people were selected) equal to 25%?

One sample proportion test

- Null hypothesis $H_0: p = p_0$
- Alternative Hypothesis $H_1: p \neq p_0$

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

Simulation example

```
> prop.test(x=3000,n=10000,p=0.25)
```

1-sample proportions test with continuity correction

```
data: 3000 out of 10000, null probability 0.25
x-squared = 133.07, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.25
95 percent confidence interval:
 0.2910470 0.3091075
sample estimates:
 p
0.3
```

```
> prop.test(x=30,n=100,p=0.25)
```

1-sample proportions test with continuity correction

```
data: 30 out of 100, null probability 0.25
x-squared = 1.08, df = 1, p-value = 0.2987
alternative hypothesis: true p is not equal to 0.25
95 percent confidence interval:
 0.2145426 0.4010604
sample estimates:
 p
0.3
```

prop.test {stats}

R Documentation

Test of Equal or Given Proportions

Description

prop.test can be used for testing the null that the proportions (probabilities of success) in several groups are the same, or that they equal certain given values.

Usage

```
prop.test(x, n, p = NULL,
          alternative = c("two.sided", "less", "greater"),
          conf.level = 0.95, correct = TRUE)
```

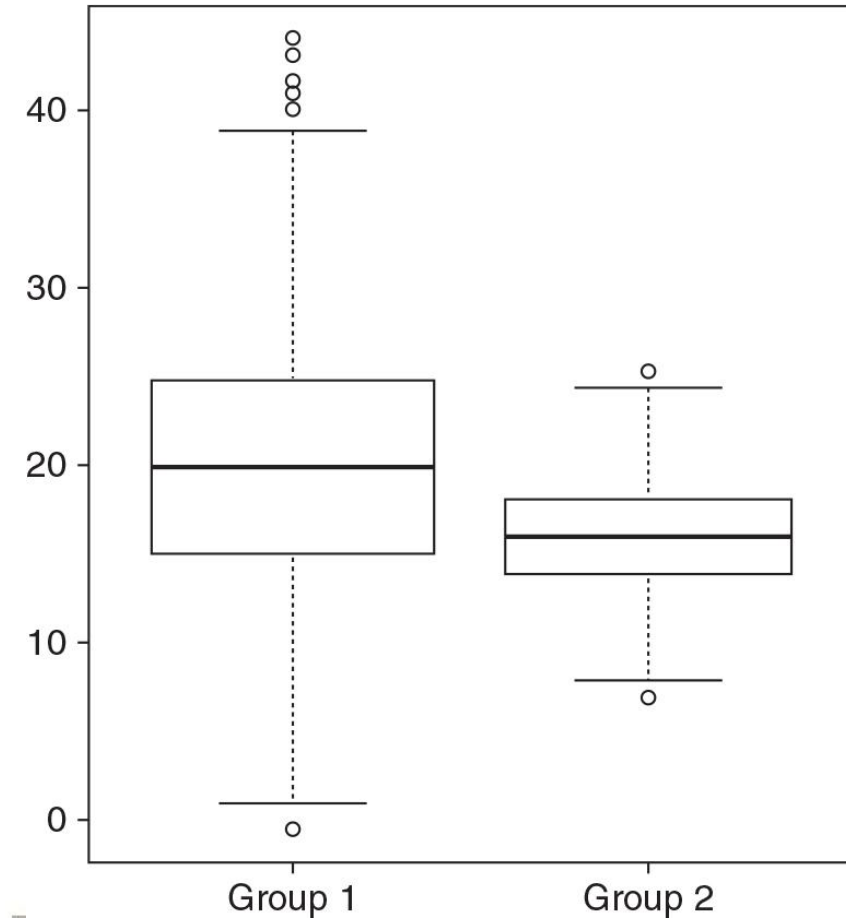
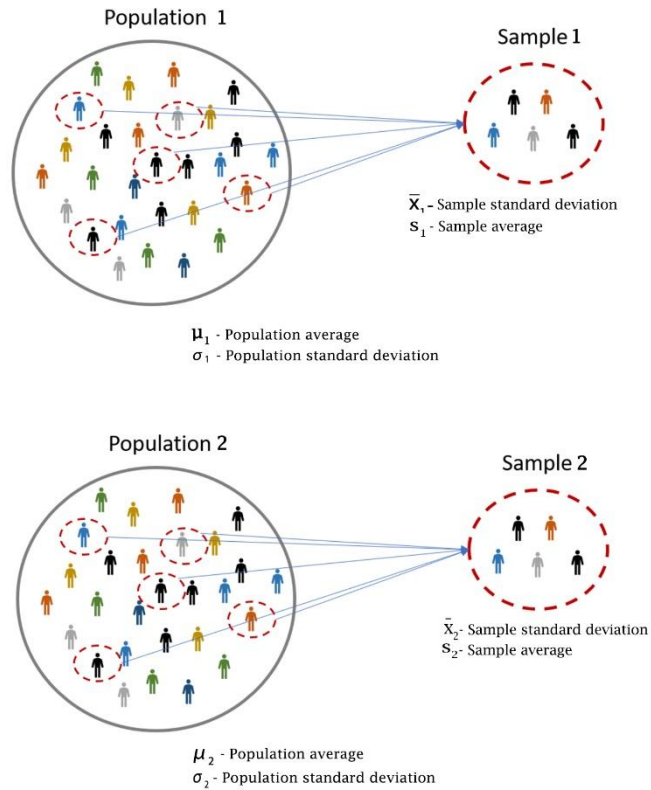
Hint:

n = number of trials

x = number of
successes

$p = H_0$

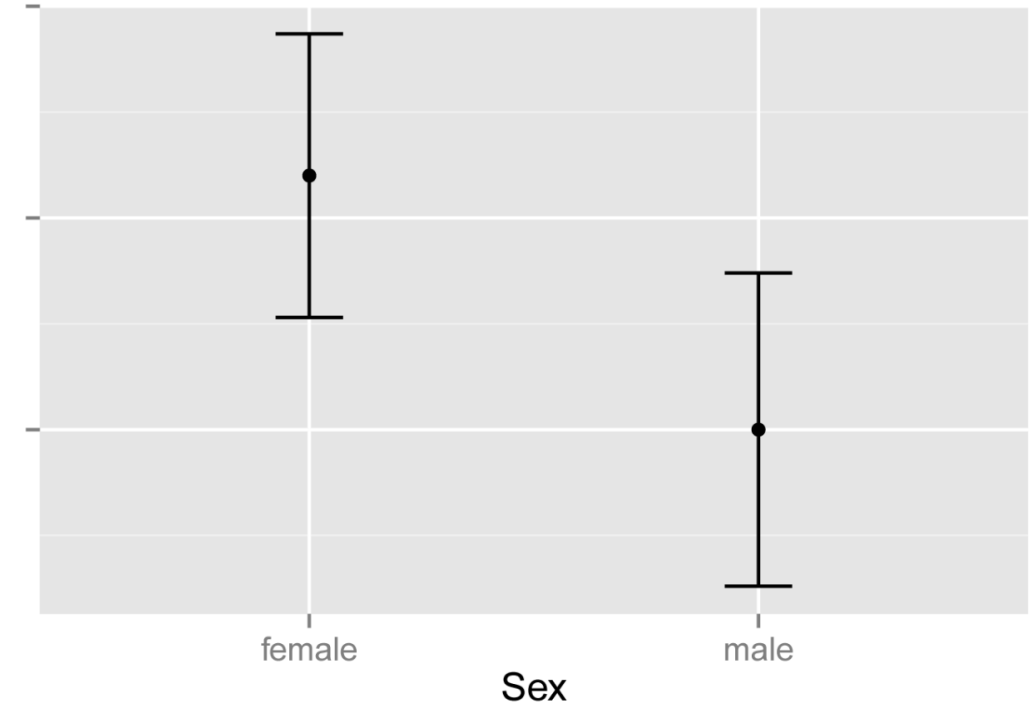
Two samples



Two sample – Hypothesis testing by 95%CI

- Continuous variables:
 - Example?
- How?
 - Construct 95% CI for both groups and see for overlaps

$$\bar{x} \pm z \frac{s}{\sqrt{n}}$$



Task:

Imagine that you have randomly sampled 100 people from two places. Assume the mean FBS readings are 115 (SD = 20) and 100 (SD = 17) in sample 1 and 2 respectively. Are the mean values different? Hint: $z = 1.96$

Two samples – Hypothesis testing by test

Continuous variables

- Null hypothesis $H_0 : \mu_1 = \mu_2$
- Alternative Hypothesis $H_1 : \mu_1 \neq \mu_2$
- Independent or related (paired) samples
 - Independent sample t test (different samples)
 - Paired t test (one group - pre and post data)
- Equal or unequal variances between groups
 - If equal : Pooled t test (exact t distribution)
 - If unequal : Welch's t test

Simulation example

```
> y1 <- rnorm(100,25,10)
> y2 <- rnorm(100,30,10)
```

```
> summary(y1)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-0.385  17.832  25.423  25.193  32.750  51.133
```

```
> summary(y2)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 4.646  23.948  30.129  29.228  36.211  47.921
```

```
> var.test(y1,y2)
```

F test to compare two variances

```
data: y1 and y2
F = 1.4029, num df = 99, denom df = 99, p-value = 0.09373
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.9439417 2.0850641
sample estimates:
ratio of variances
 1.402918
```

t.test {stats}

Student's t-Test

Description

Performs one and two sample t-tests on vectors of data.

Usage

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

```
## Default S3 method:
```

```
t.test(x, y = NULL,  
       alternative = c("two.sided", "less", "greater"),  
       mu = 0, paired = FALSE, var.equal = FALSE,  
       conf.level = 0.95, ...)
```

```
## S3 method for class 'formula'
```

```
t.test(formula, data, subset, na.action, ...)
```

```
> t.test(y1,y2)
```

Welch Two Sample t-test

```
data: y1 and y2
t = -2.7886, df = 192.59, p-value = 0.005825
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -6.888075 -1.180903
sample estimates:
mean of x mean of y
 25.19331  29.22780
```

```
> t.test(y1,y2,var.equal = T)
```

Two Sample t-test

```
data: y1 and y2
t = -2.7886, df = 198, p-value = 0.00581
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -6.887581 -1.181396
sample estimates:
mean of x mean of y
 25.19331  29.22780
```

```
> t.test(y1,y2, paired = T)
```

Paired t-test

```
data: y1 and y2
t = -2.8289, df = 99, p-value = 0.005656
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -6.864361 -1.204617
sample estimates:
mean of the differences
 -4.034489
```

t.test {stats}

Student's t-Test

Description

Performs one and two sample t-tests on vectors of data.

Usage

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

```
## Default S3 method:
```

```
t.test(x, y = NULL,
       alternative = c("two.sided", "less", "greater"),
       mu = 0, paired = FALSE, var.equal = FALSE,
       conf.level = 0.95, ...)
```

```
## S3 method for class 'formula'
```

```
t.test(formula, data, subset, na.action, ...)
```



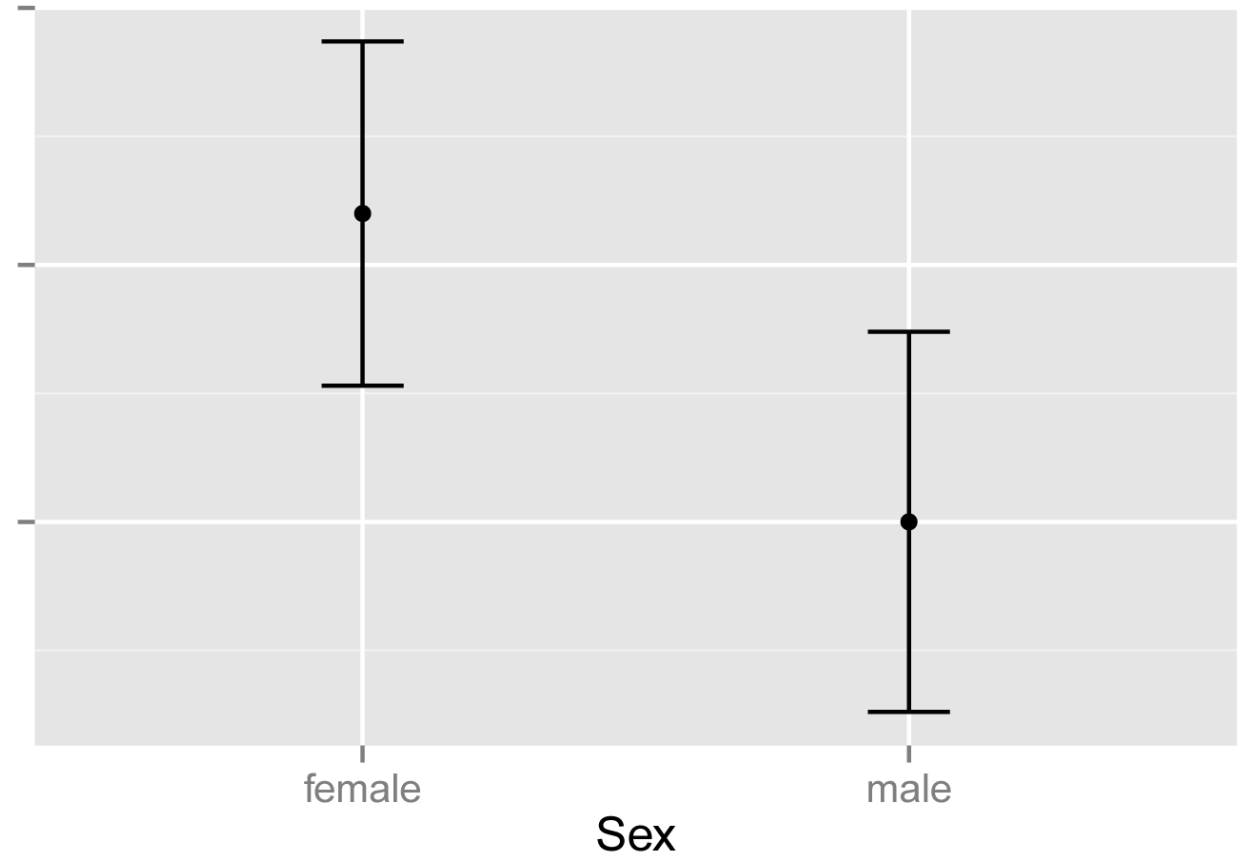
eswatini
antivenom
foundation



Two sample – Hypothesis testing by 95%CI

- Binary outcomes:
 - Example?
- How?
 - 95% CI for both groups and see for overlaps

$$\hat{p} \pm z \sqrt{\frac{\hat{p} (1 - \hat{p})}{n}}$$



Two samples – Hypothesis testing by test

Binary outcomes

Two (Independent) sample proportion test

- Null hypothesis $H_0: P_1 = P_2$
- Alternative Hypothesis $H_1: P_1 \neq P_2$

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}(1 - \hat{p}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

McNemar's test (paired)

- Null hypothesis $H_0: P_1 = P_2$
- Alternative Hypothesis $H_1: P_1 \neq P_2$

$$\chi^2 = \frac{(b - c)^2}{b + c}$$

Simulation example

prop.test {stats}

Test of Equal or Given Proportions

Description

`prop.test` can be used for testing the null that the proportions (probabilities of success) in several groups are the same, or that they equal certain given values.

Usage

```
prop.test(x, n, p = NULL,
          alternative = c("two.sided", "less", "greater"),
          conf.level = 0.95, correct = TRUE)
```

```
> prop.test(x = c(45,50), n = c(100,100))
```

2-sample test for equality of proportions with continuity correction

```
data: c(45, 50) out of c(100, 100)
X-squared = 0.3208, df = 1, p-value = 0.5711
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.19824347  0.09824347
sample estimates:
prop 1 prop 2
 0.45   0.50
```

```
> prop.test(x = c(30,50), n = c(100,100))
```

2-sample test for equality of proportions with continuity correction

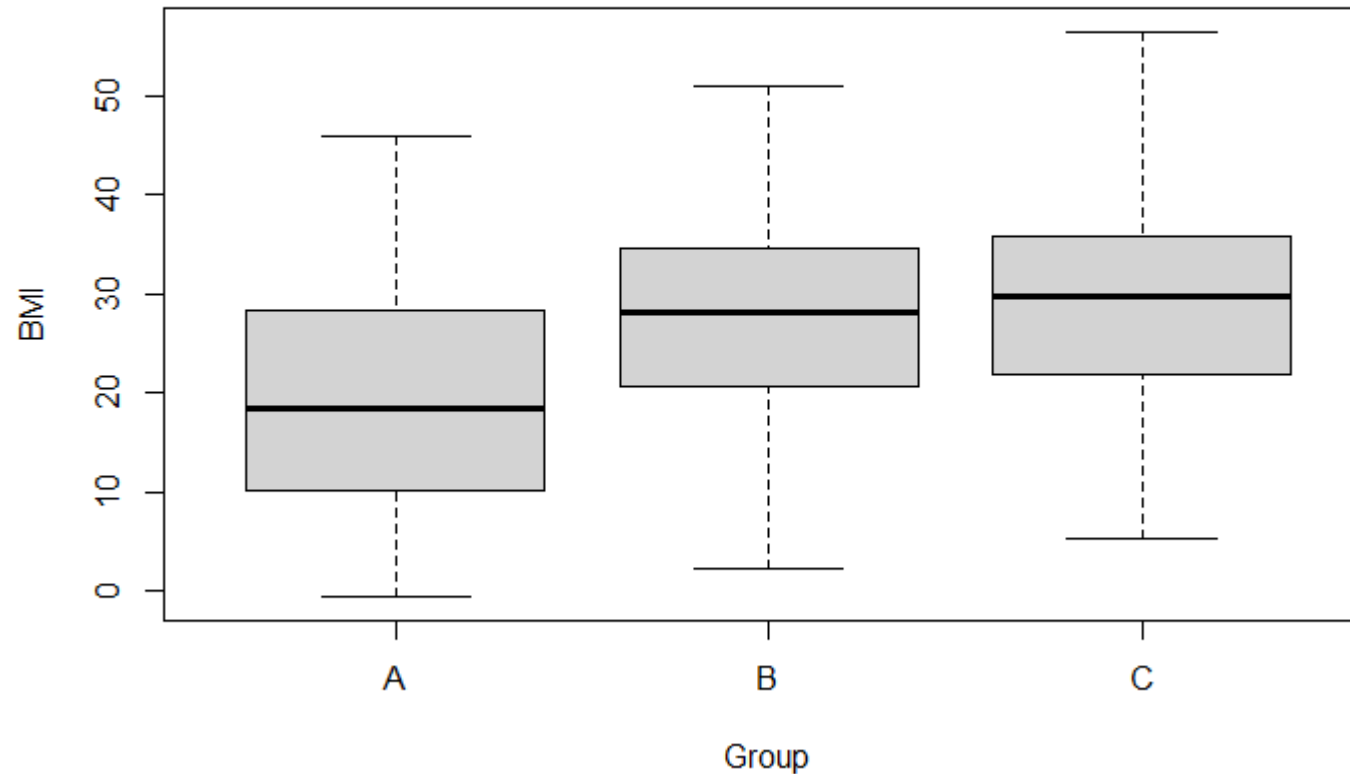
```
data: c(30, 50) out of c(100, 100)
X-squared = 7.5208, df = 1, p-value = 0.006099
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.34293122 -0.05706878
sample estimates:
prop 1 prop 2
 0.3    0.5
```

Hint:

x = number of successes (2 groups)

n = number of trials (2 groups)

More than two samples/groups



More than two samples

- Construct 95% CI and see for overlaps
- Do a test
 - Independent samples : ANOVA and Post Hoc test
 - Dependent samples : Repeated measures ANOVA
(Longitudinal data analysis)

Analysis of variability

- Total variability = Variability due to treatment + Natural variability
- How to calculate?

ANOVA table

Source	df	SS	MS	F
Model	p	$\sum (\hat{y}_i - \bar{y})^2$	$SS_{\text{Model}} / df_{\text{Model}}$	$MS_{\text{Model}} / MS_{\text{Error}}$
Error	N - p - 1	$\sum (y_i - \hat{y}_i)^2$	$SS_{\text{Error}} / df_{\text{Error}}$	
Total	N - 1	$\sum (y_i - \bar{y})^2$		

Simulation example

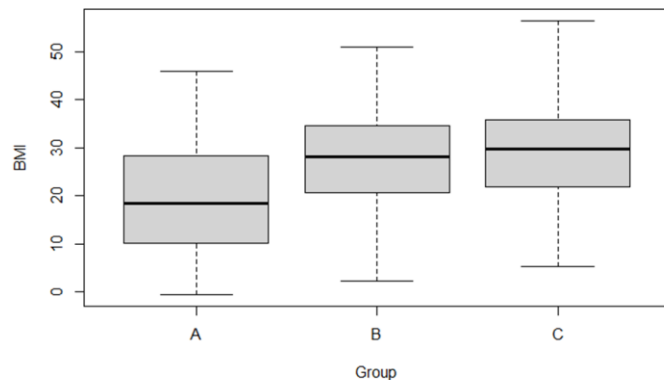
```
y1 <- rnorm(100,20,10)
y2 <- rnorm(100,28,10)
y3 <- rnorm(100,30,10)

data <- data.frame(group = c(rep("A",100), rep("B",100),rep("C",100)),
                    BMI = c(y1,y2,y3))

> aov.out <- aov(data$BMI ~ data$group)
> summary(aov.out)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
data\$group	2	5995	2997.5	28.12	6.56e-12 ***
Residuals	297	31663	106.6		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1



```
> TukeyHSD(aov.out)
Tukey multiple comparisons of means
95% family-wise confidence level

Fit: aov(formula = data$BMI ~ data$group)

$`data$group`
      diff      lwr      upr      p adj
B-A  8.732144  5.292587 12.171702 0.0000000
C-A 10.087824  6.648266 13.527381 0.0000000
C-B  1.355679 -2.083878  4.795237 0.6227532
```

- Categorical outcomes
 - Give an example?
 - Pearson's chi-square test

Exposure	Disease status			
		Disease	No Disease	
	Exposed	12	12	24
	Unexposed	11	32	43
		25	36	67

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

$$\chi^2 = \text{the test statistic} \quad \sum = \text{the sum of}$$

Pearson's Chi-squared Test for Count Data

Description

chisq.test performs chi-squared contingency table tests and goodness-of-fit tests.

Usage

```
chisq.test(x, y = NULL, correct = TRUE,
           p = rep(1/length(x), length(x)), rescale.p = FALSE,
           simulate.p.value = FALSE, B = 2000)
```

Simulation example

```
Smoking <- sample(c("Yes", "No"), 100, replace = T)
Cancer <- sample(c("Yes", "No"), 100, replace = T)
```

```
> table(Smoking ,Cancer)
```

	Cancer	
Smoking	No	Yes
No	18	16
Yes	32	34

```
> prop.table(table(Smoking ,Cancer),1)*100
```

	Cancer	
Smoking	No	Yes
No	52.94118	47.05882
Yes	48.48485	51.51515

```
> chisq.test(table(Smoking ,Cancer))
```

Pearson's Chi-squared test with Yates' continuity correction

```
data: table(Smoking, Cancer)
X-squared = 0.044563, df = 1, p-value = 0.8328
```

- Categorical outcomes (If sample size is very small in any cell : eg <5)
 - Fisher`s exact test

```
Smoking <- sample(c("Yes","No"),100,replace = T)
Cancer <- sample(c("Yes","No"),100,prob = c(0.1,0.9),replace = T)
table(Smoking ,Cancer)
```

```
> table(Smoking ,Cancer)
```

	Cancer	
Smoking	No	Yes
No	32	2
Yes	55	11

```
> prop.table(table(Smoking ,Cancer),1)*100
```

	Cancer	
Smoking	No	Yes
No	94.117647	5.882353
Yes	83.333333	16.666667

```
> chisq.test(table(Smoking ,Cancer))
```

Pearson's Chi-squared test with Yates'

```
data: table(Smoking, Cancer)
X-squared = 1.4525, df = 1, p-value = 0.2281
```

Warning message:
In chisq.test(table(Smoking, Cancer)) :
Chi-squared approximation may be incorrect

```
> fisher.test(table(Smoking ,Cancer))
```

Fisher's Exact Test for Count Data

```
data: table(Smoking, Cancer)
p-value = 0.209
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.631179 31.207926
sample estimates:
odds ratio
 3.168927
```

More than 2 groups

Surgical Apgar Score	No morbidity	Minor morbidity	Major morbidity or mortality
0-4	21	20	16
5-6	135	71	35
7-10	158	62	35

- Still can do a Pearson's chi-square test
 - This is a global test (like ANOVA)
 - If there is no difference – then ok
 - If there is a difference – need further analysis (loglinear models)

Session 3: Exploring Relationships and Prediction

Inferential statistics : Part II

Independent variable : Continuous

- What are the examples for continuous type independent variable?
 - What are the examples for continuous type dependent (outcome) variables related to a continuous type independent variable?
 - What are the examples for categorical type dependent (outcome) variables related to a continuous type independent variable?

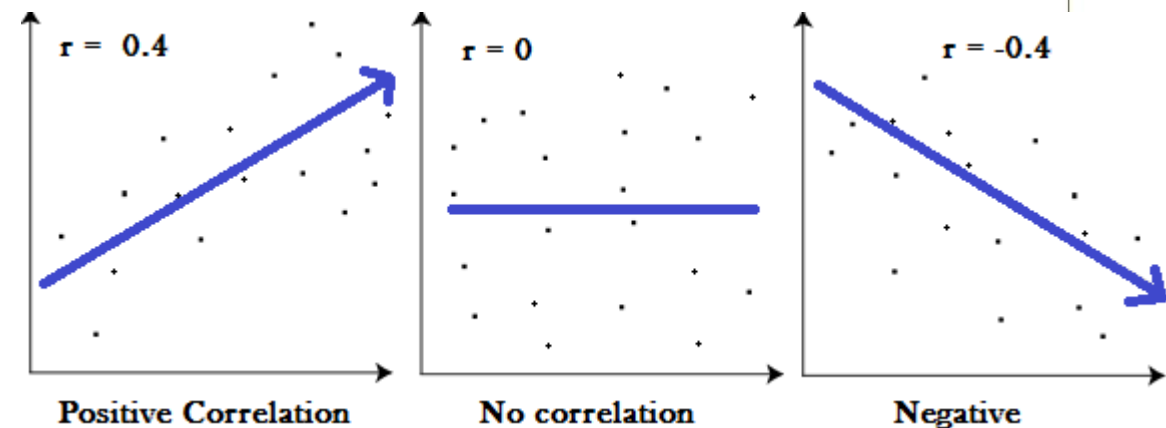
Independent variable : Continuous

- Continuous outcomes (e.g. FBS with age)
 - Correlation analysis
 - Linear regression analysis
- Categorical outcomes (e.g. DM prevalence with age)
 - Logistic regression
 - Multinomial regression

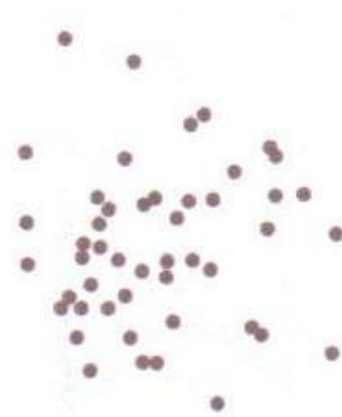
Pearson's correlation (Pearson's R)

- Strength of relationship (How strong a relationship is ?)
- Not to find the relationship (by regression)
- Correlation coefficient

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



Correlation



Correlation $r = 0$



Correlation $r = -0.3$



Correlation $r = 0.5$



Correlation $r = -0.7$



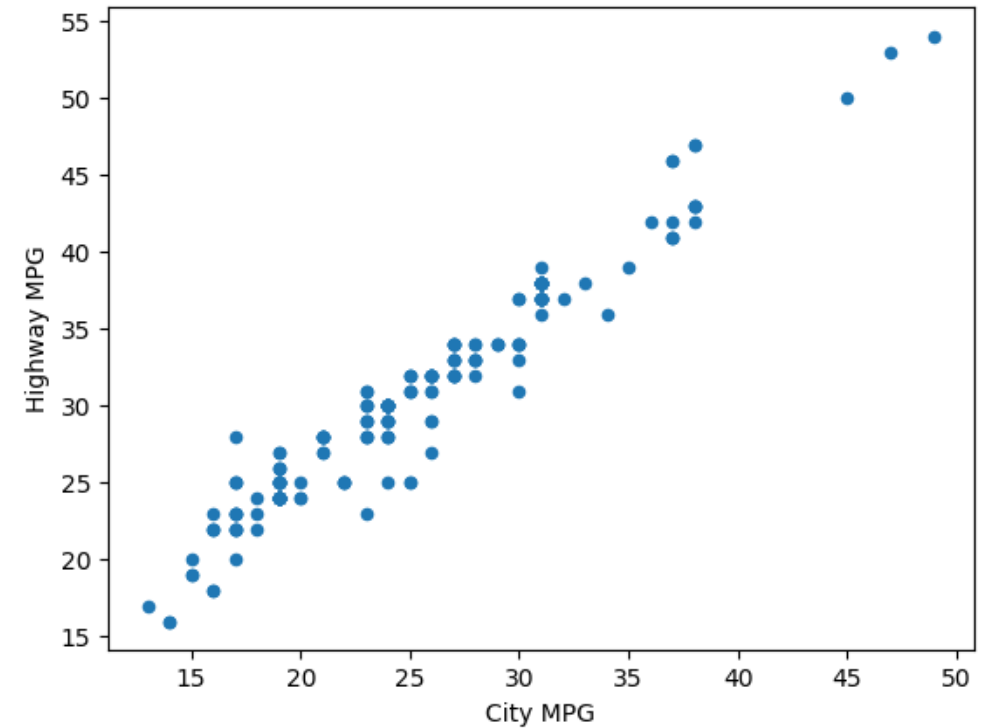
Correlation $r = 0.9$



Correlation $r = -0.99$

Regression analysis

- Finding the relationship between variables
- Why we need to know?
 - To predict



Types

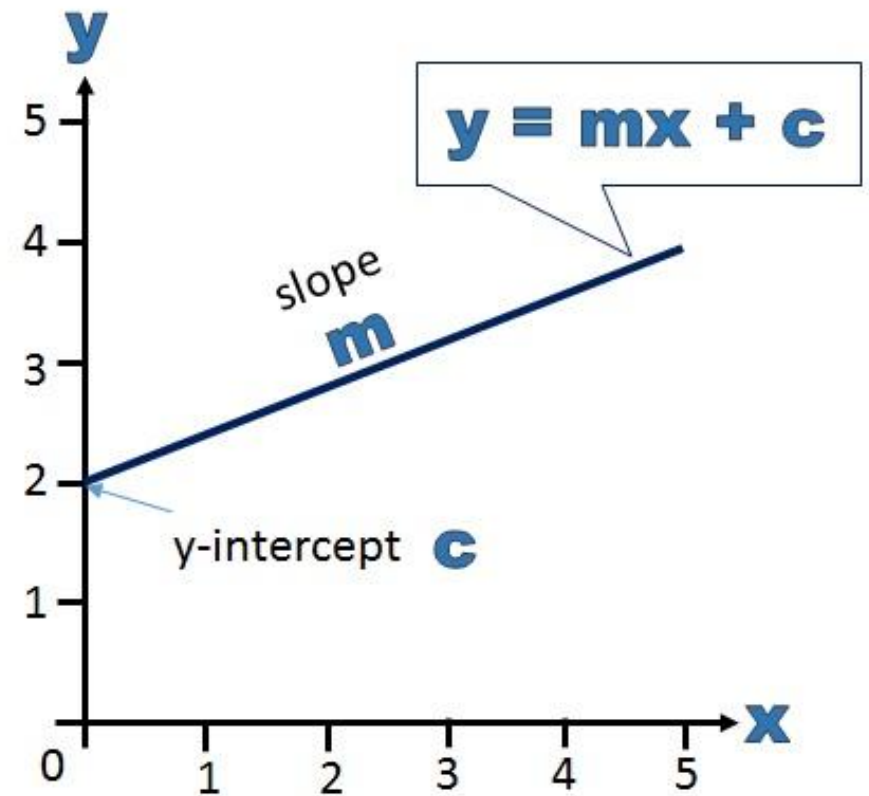
- Linear regression
- Binary logistic regression
- Poisson regression (loglinear models)
- Cox proportional hazard models

Regression analysis

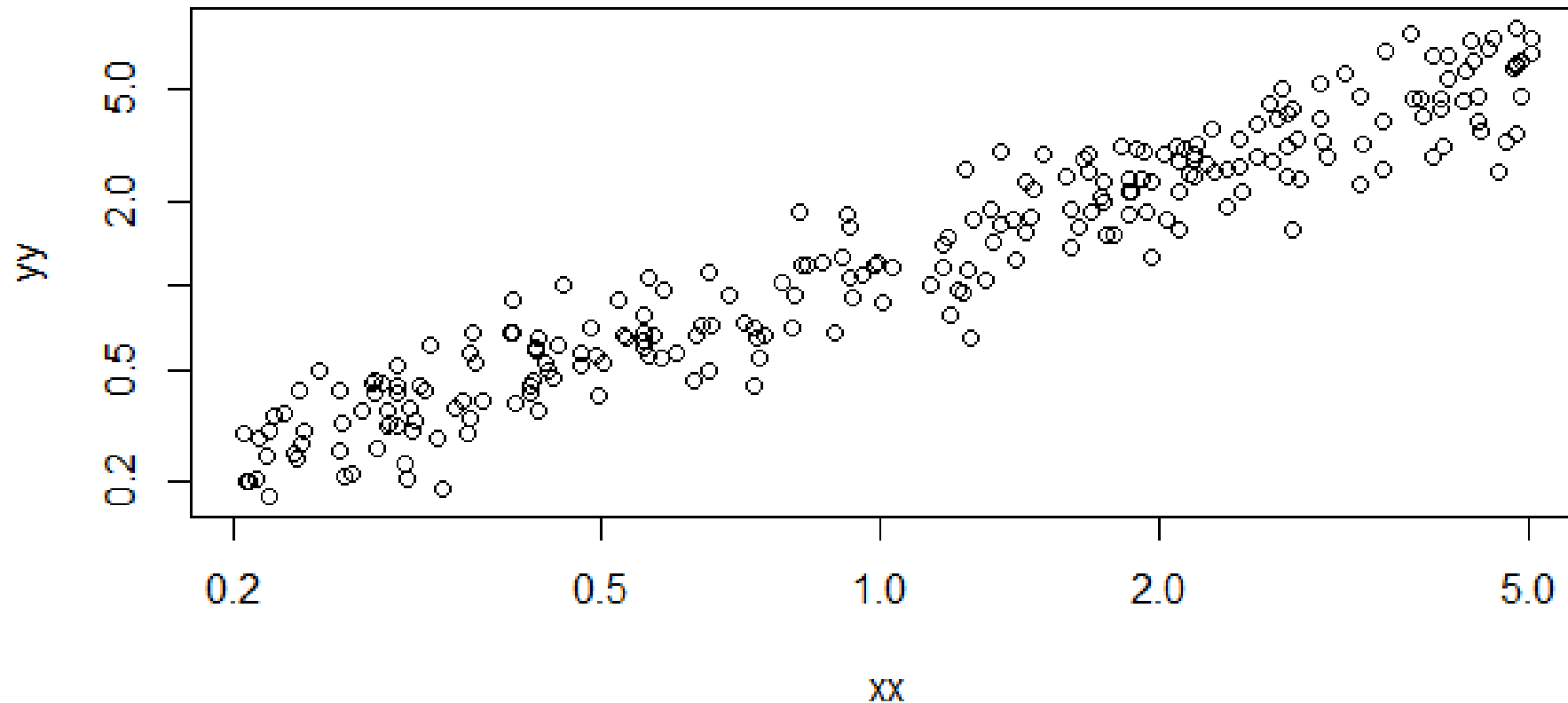
- Linear regression makes several key assumptions:
 - Linear relationship
 - Multivariate normality
 - No or little multicollinearity (ie. independent)
 - No auto-correlation (value of $y(x+1)$ is not dependent on the value of $y(x)$)
 - Homoscedasticity (that is the error terms along the regression are equal)

Mathematics

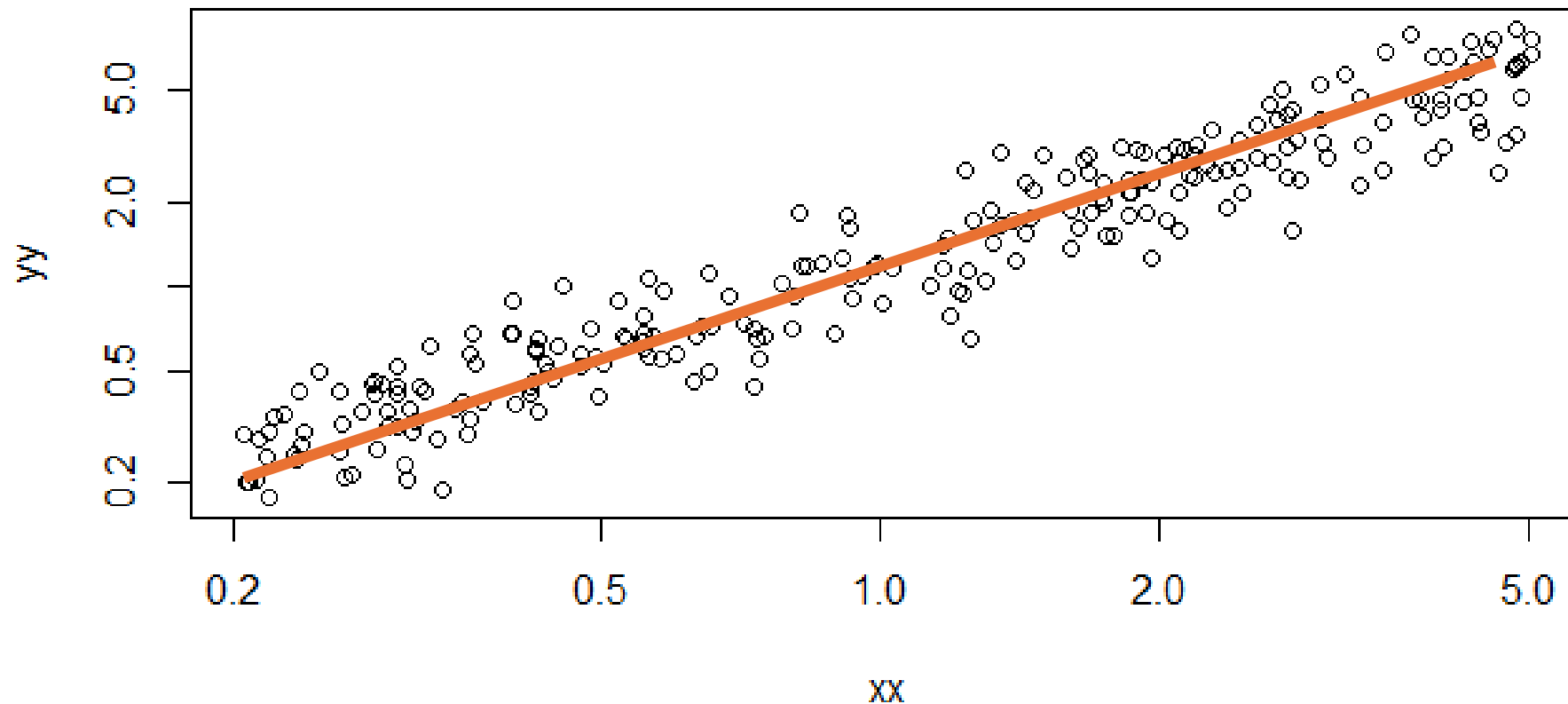
- How to write this mathematically?
 - $y_i = m x_i + c$

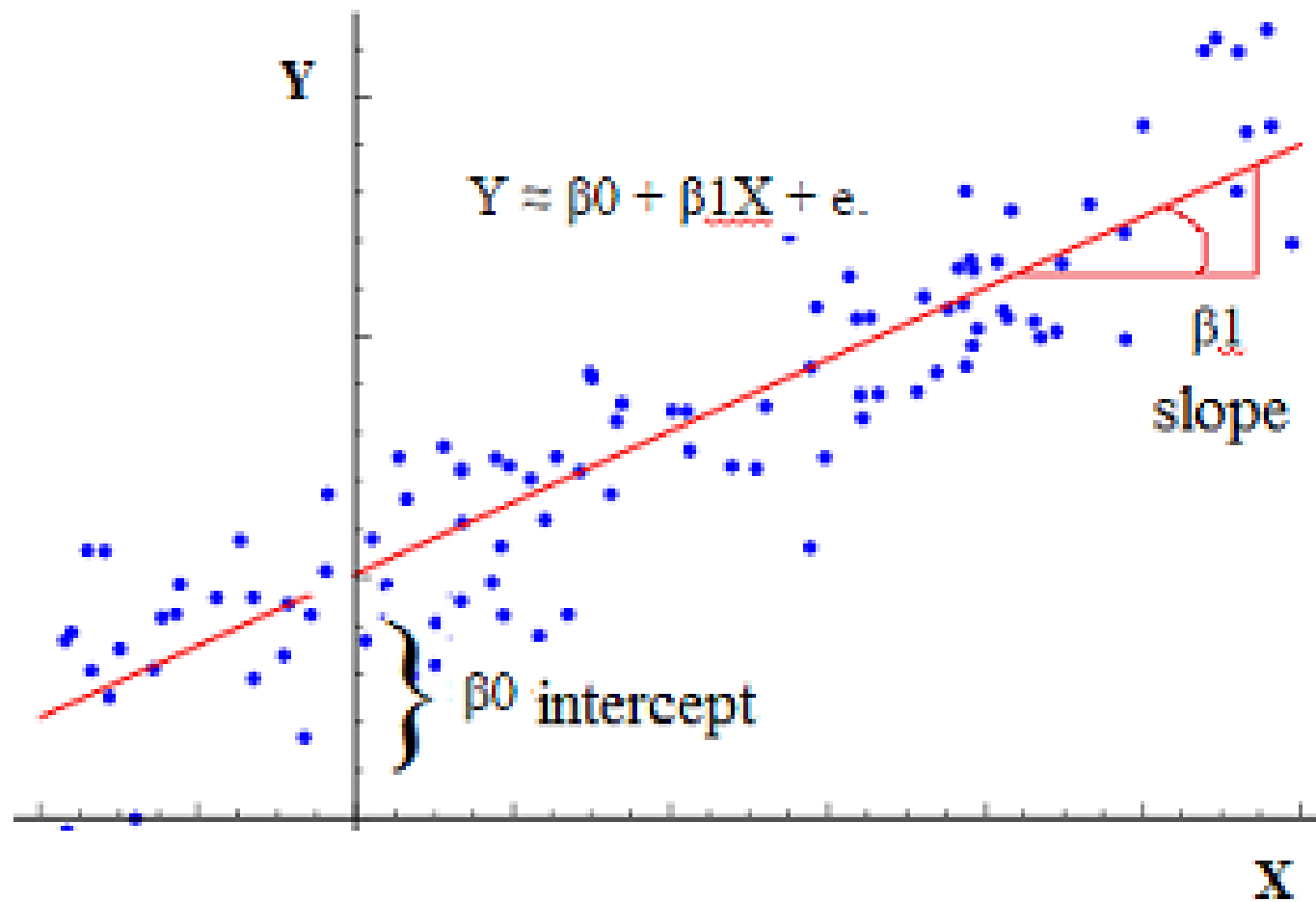


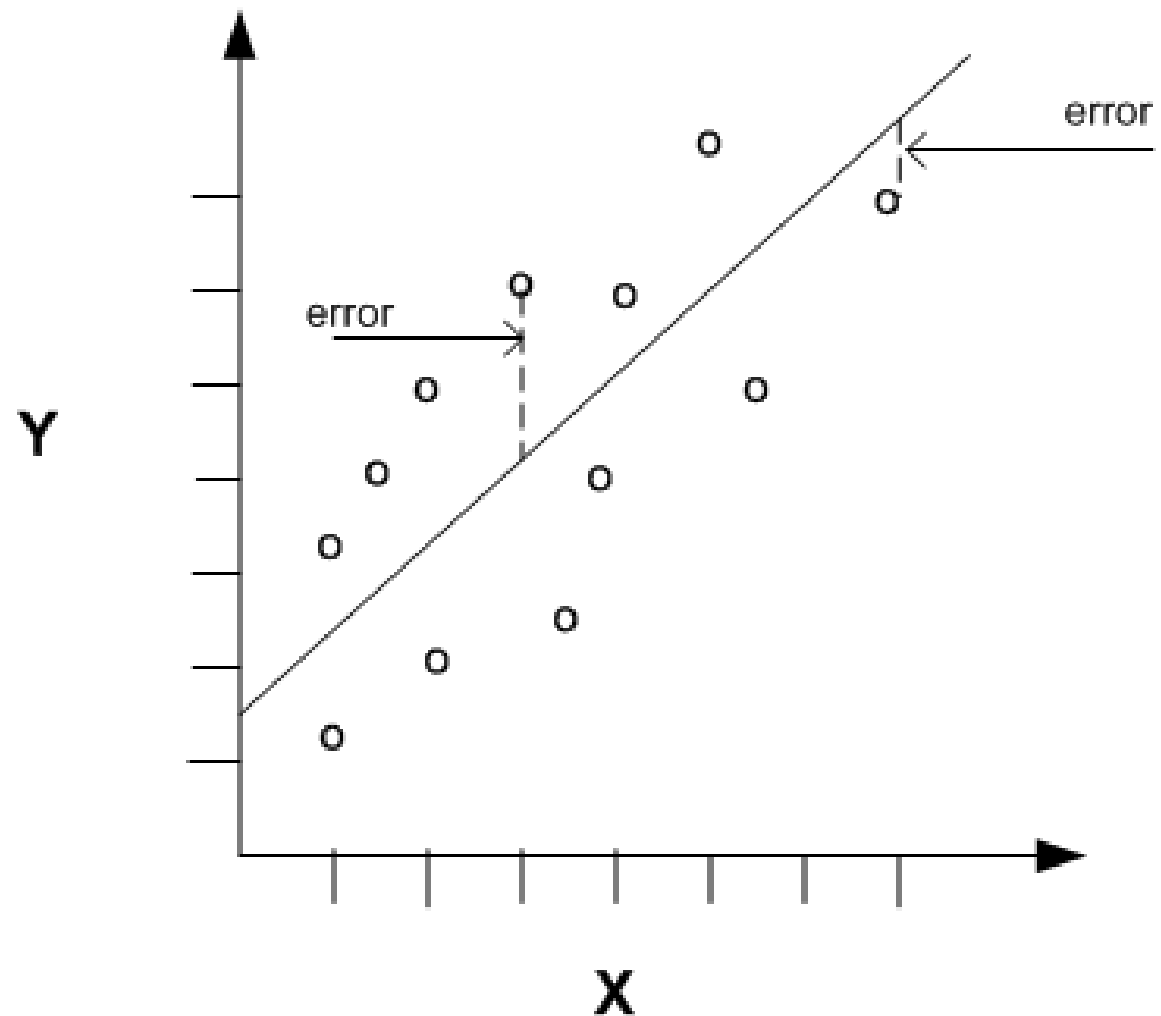
Real world example



Real world example

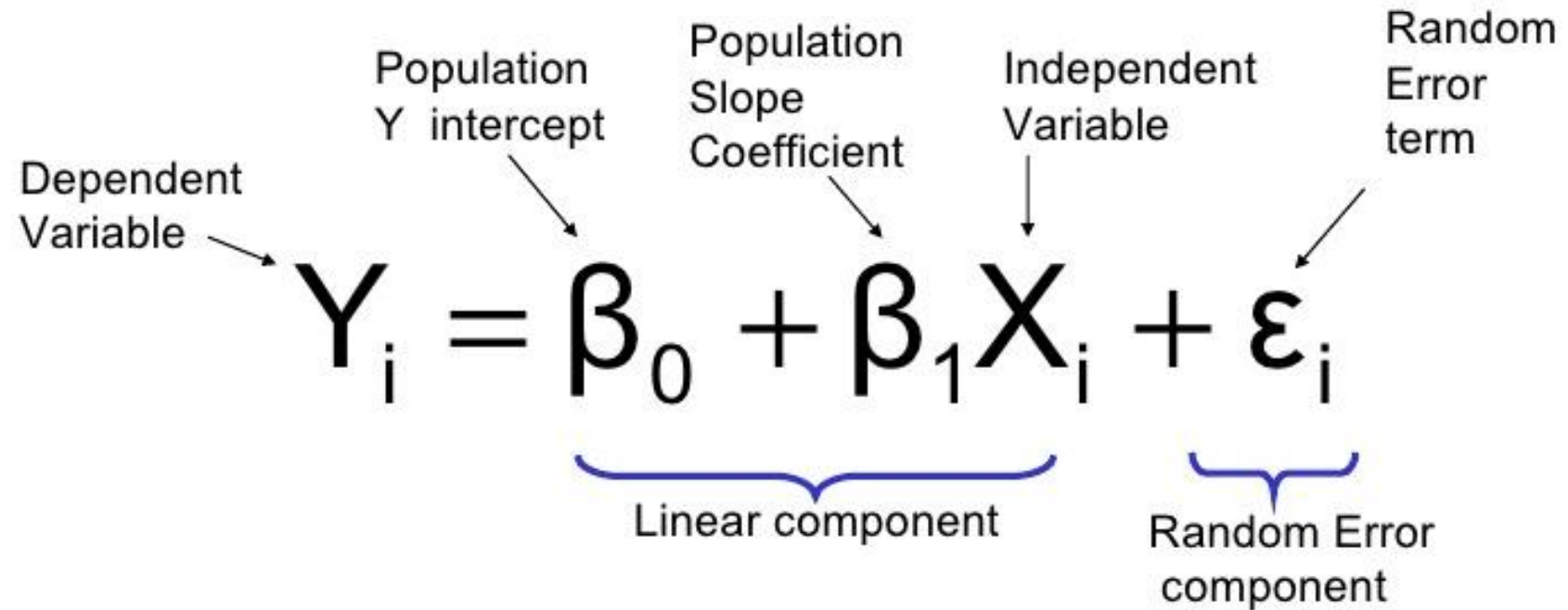






Conceptual model:

- Population regression model


$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

Dependent Variable

Population Y intercept

Population Slope Coefficient

Independent Variable

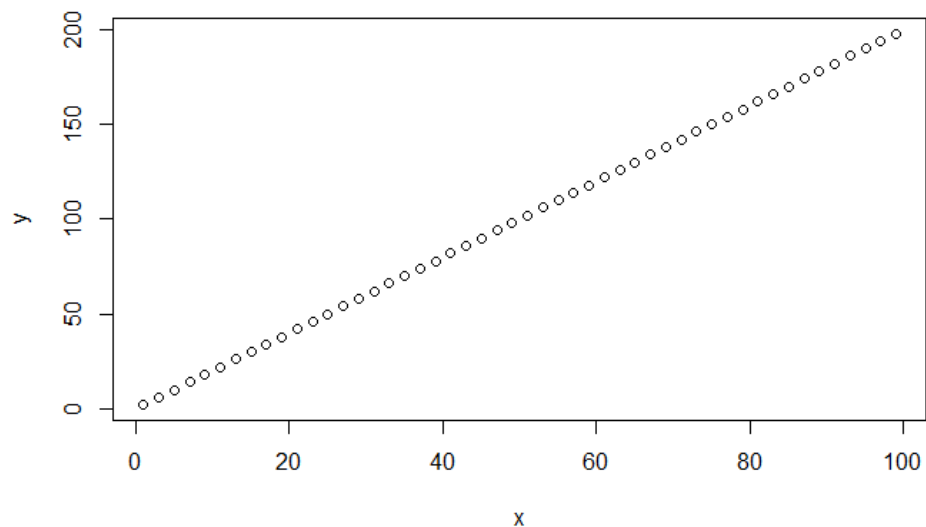
Random Error term

Linear component

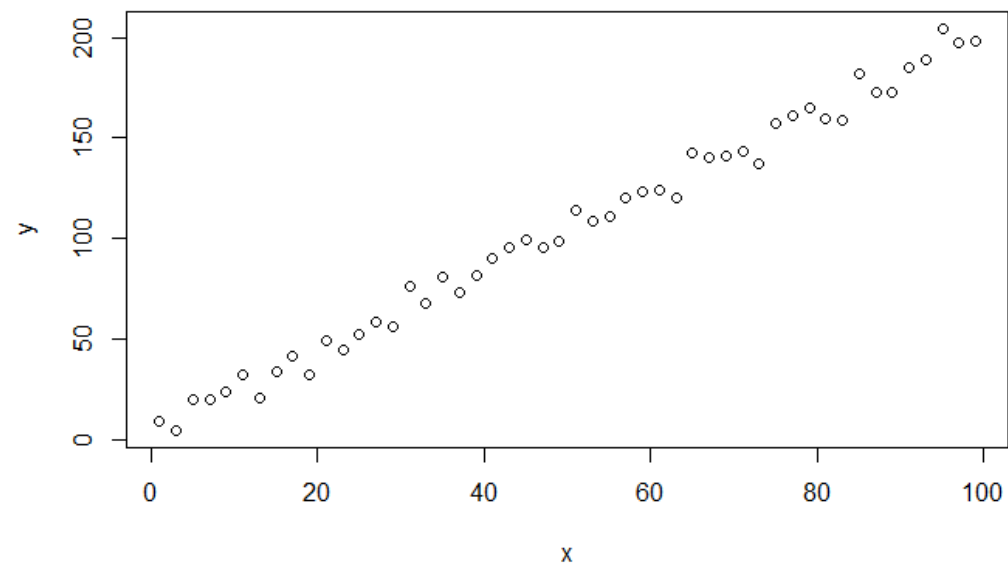
Random Error component

Simulation example

```
x <- seq(1,100,by=2)
y <- 2*x
plot(x,y)
```

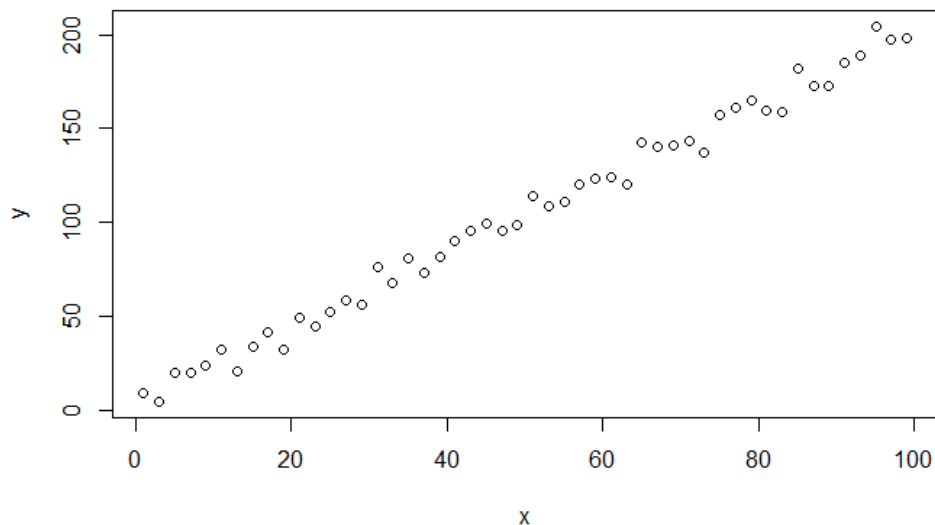


```
x <- seq(1,100,by=2)
y <- 2*x + rnorm(50,5,5)
plot(x,y)
```



Simulation example

```
x <- seq(1,100,by=2)
y <- 2*x + rnorm(50,5,5)
plot(x,y)
```



```
> summary(lm(y~x))
```

```
Call:
lm(formula = y ~ x)
```

```
Residuals:
```

	Min	1Q	Median	3Q	Max
Residuals	-7.8008	-2.9175	-0.0975	2.9962	9.9960

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	4.18140	1.29732	3.223	0.00228 **
x	2.01687	0.02247	89.753	< 2e-16 ***

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 4.586 on 48 degrees of freedom
```

```
Multiple R-squared:  0.9941,    Adjusted R-squared:  0.994
```

```
F-statistic: 8056 on 1 and 48 DF,  p-value: < 2.2e-16
```


Expanding simple linear regression

Simple
Linear
Regression

$$y = b_0 + b_1x_1$$

Multiple
Linear
Regression

$$y = b_0 + b_1x_1 + b_2x_2 + \dots + b_nx_n$$

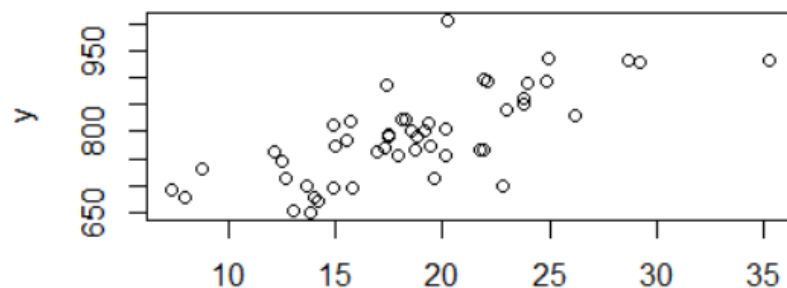
Polynomial
Linear
Regression

$$y = b_0 + b_1x_1 + b_2x_1^2 + \dots + b_nx_1^n$$

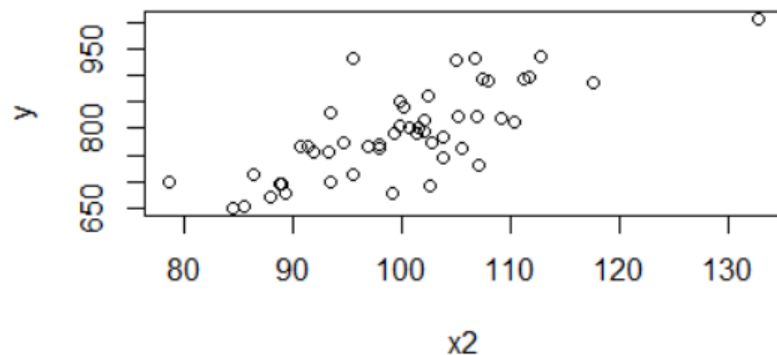
Example - multiple linear regression

```
x1 <- rnorm(50,20,5)
x2 <- rnorm(50,100,10)
y <- 10*x1 + 6*x2 + rnorm(50,5,5)
```

plot(x1,y)



plot(x2,y)



```
> summary(lm(y~x1+x2))
```

Call:
lm(formula = y ~ x1 + x2)

Residuals:

Min	1Q	Median	3Q	Max
-7.9808	-3.0874	-0.0921	2.3243	11.4903

Coefficients:

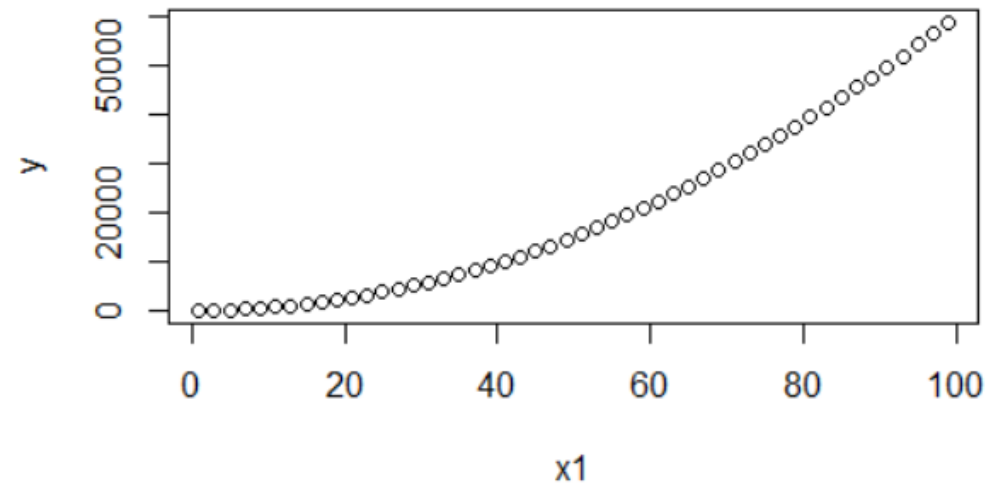
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.70147	6.81788	0.25	0.804
x1	10.07570	0.11660	86.41	<2e-16 ***
x2	6.02172	0.06662	90.38	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.412 on 47 degrees of freedom
Multiple R-squared: 0.9973, Adjusted R-squared: 0.9972
F-statistic: 8728 on 2 and 47 DF, p-value: < 2.2e-16

Example – polynomial linear regression

```
x1 <- seq(1,100,by=2)
x2 <- x1^2
y <- 2*x1 + 6*x2 + rnorm(50,5,5)
plot(x1,y)
```



```
> summary(lm(y~x1+x2))
```

Call:

```
lm(formula = y ~ x1 + x2)
```

Residuals:

	Min	1Q	Median	3Q	Max
Residuals	-15.1871	-3.3959	0.3643	3.7373	11.6549

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	8.266053	2.256018	3.664	0.00063	***
x1	1.799061	0.104224	17.262	< 2e-16	***
x2	6.001591	0.001009	5946.932	< 2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.314 on 47 degrees of freedom

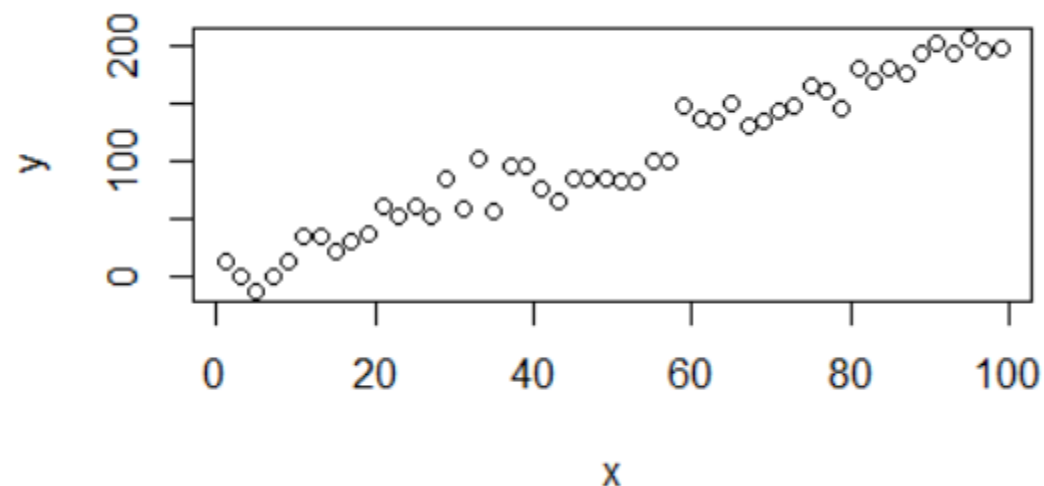
Multiple R-squared: 1, Adjusted R-squared: 1

F-statistic: 2.849e+08 on 2 and 47 DF, p-value: < 2.2e-16

Goodness of a fit

- Outcome vs predictors (independent variables)
- Outcome vs fitted value plot
- Residual vs fitted value plot
- Coefficient of determination (R^2)
- Deviance statistic
- Autocorrelation plot

```
x <- seq(1,100,by=2)
y <- 2*x + rnorm(50,5,15)
plot(x,y)
```



```
> fit<-lm(y~x)
> summary(fit)
```

```
Call:
lm(formula = y ~ x)
```

Residuals:

Min	1Q	Median	3Q	Max
-26.541	-8.061	-3.050	10.570	35.286

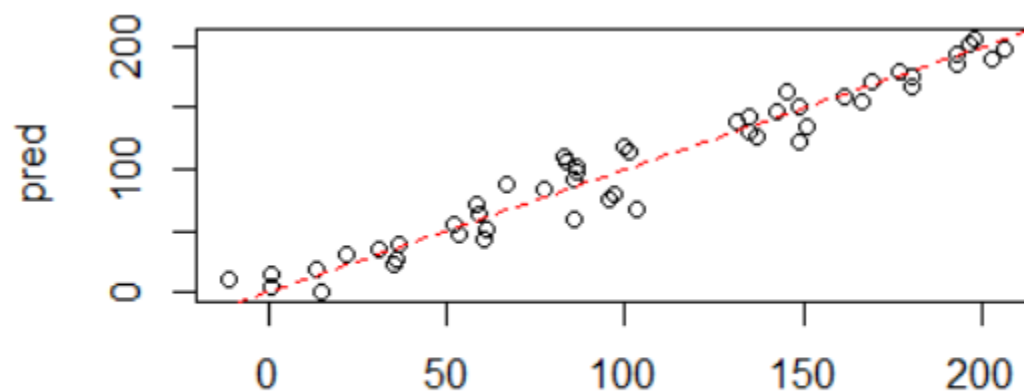
Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.51530	3.95677	-0.13	0.897
x	2.07919	0.06854	30.34	<2e-16 ***

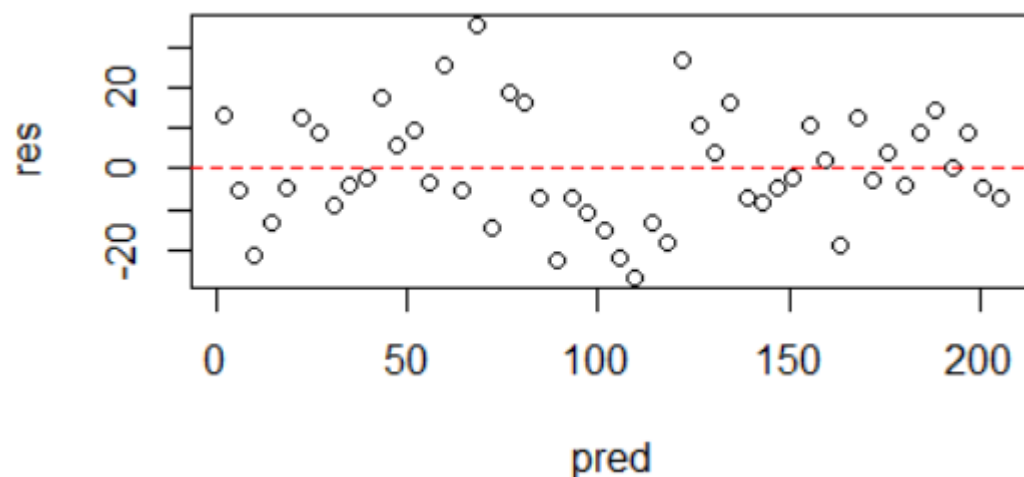
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

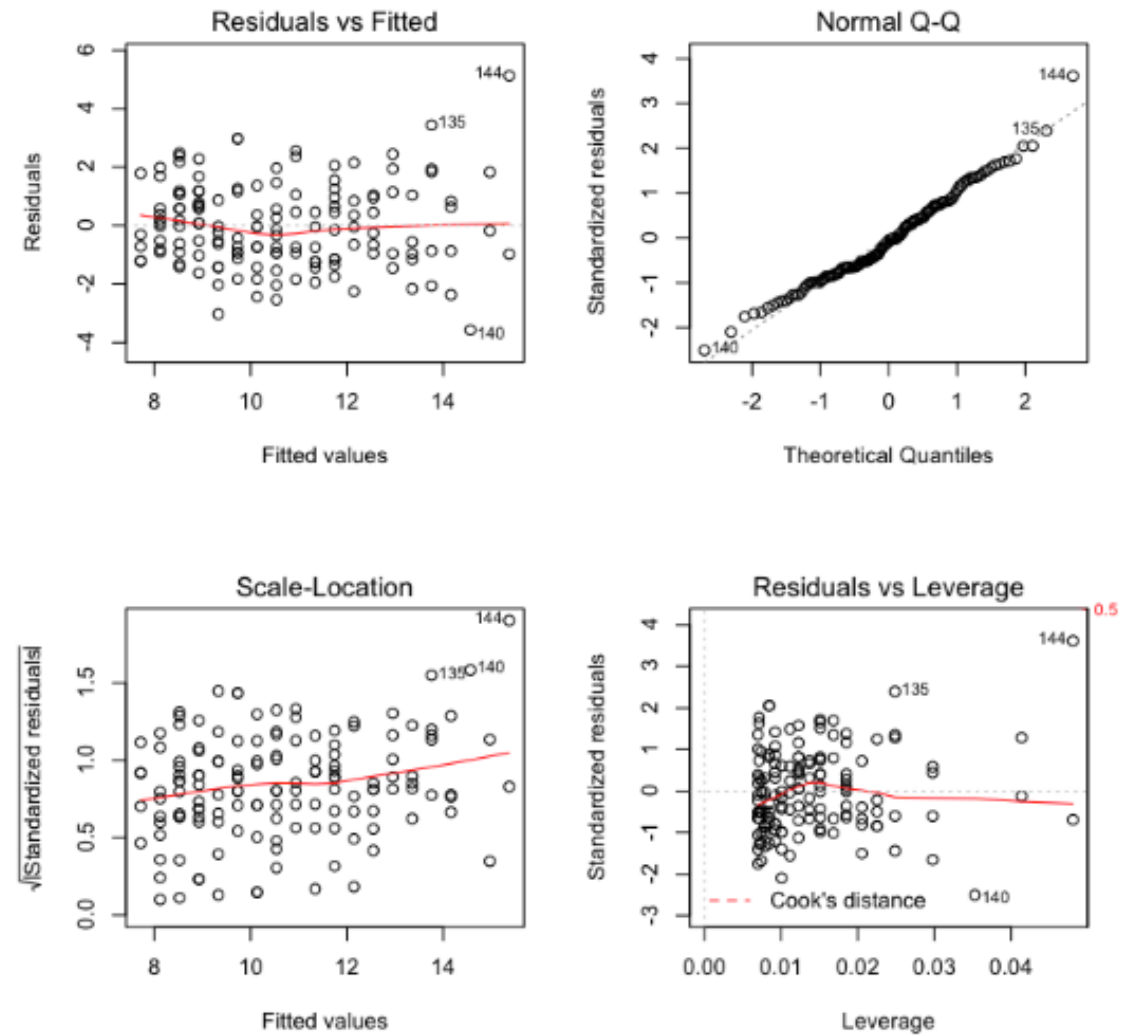
Residual standard error: 13.99 on 48 degrees of freedom
 Multiple R-squared: 0.9504, Adjusted R-squared: 0.9494
 F-statistic: 920.3 on 1 and 48 DF, p-value: < 2.2e-16

```
pred <- -0.51530+2.07919*x
plot(y, pred)
abline(0,1,col="Red",lty="dashed")
```



```
res <- y - pred
plot(pred,res)
```





Logistic regression

- Outcome – binary (1 or 0)
- Need to transform the outcome variable (continuous variable)

```
> log(0)
[1] -Inf
> 1/0
[1] Inf
> log(Inf)
[1] Inf
.
```

$$\text{Logit Function} = \log \left(\frac{p}{1-p} \right)$$

Logistic regression

$$\log \left(\frac{p(X)}{1 - p(X)} \right) = \beta_0 + \beta_1 X$$

$$\log \left(\frac{p(X)}{1 - p(X)} \right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$$

Example : obtaining predictions

Table 5 Fitted linear logistic regression model to predict the probability of having OV on screening UGIE

Variable	Estimate	Std error	Z value	P value
Intercept	-0.189	0.652	-0.290	0.771
Small platelet count	-0.046	0.015	-0.310	0.002
CTP class B (compared to Class A)	2.852	0.944	3.021	0.003
CTP class C (compared to Class A)	3.695	1.229	3.005	0.003

The prediction formula;

Log odds (presence of OV) = $-0.189 - 0.046 * \%SP + 2.9$ [if CTP class B, otherwise zero] $+ 3.7$ [if CTP class C, otherwise zero]

The prediction formula;

Log odds (presence of OV) = $-0.189 - 0.046 * \%SP + 2.9$ [if CTP class B, otherwise zero] $+ 3.7$ [if CTP class C, otherwise zero]

$$\ln\left(\frac{P}{1-P}\right) = a + bX$$

$$\frac{P}{1-P} = e^{a+bX}$$

$$P = \frac{e^{a+bX}}{1 + e^{a+bX}}$$

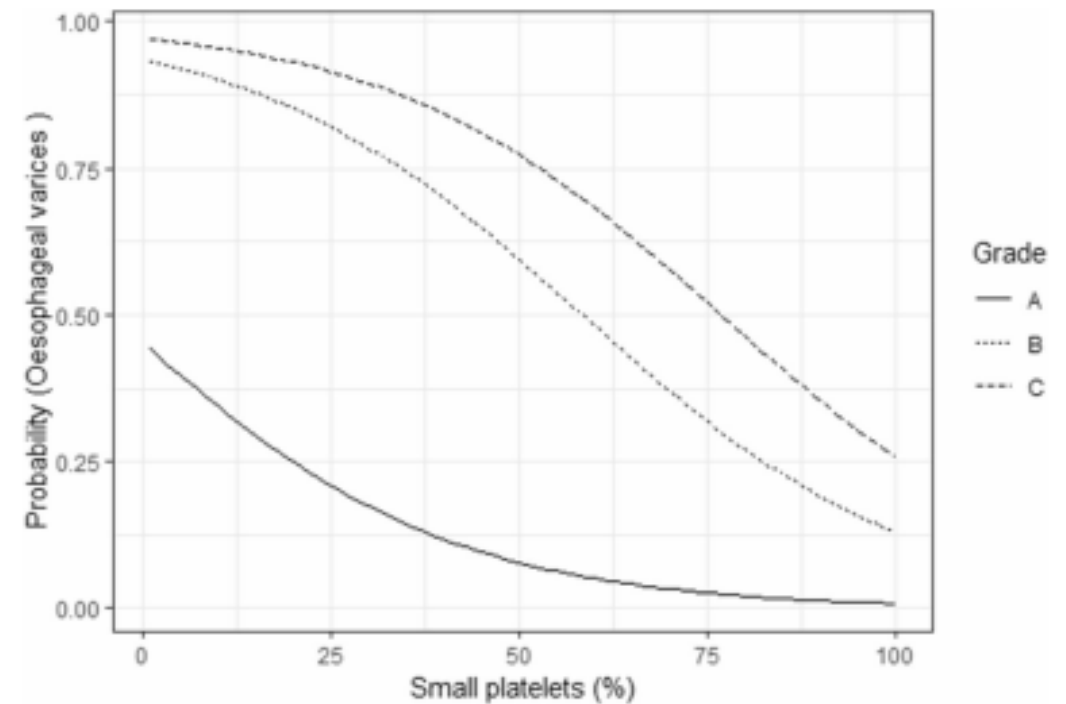
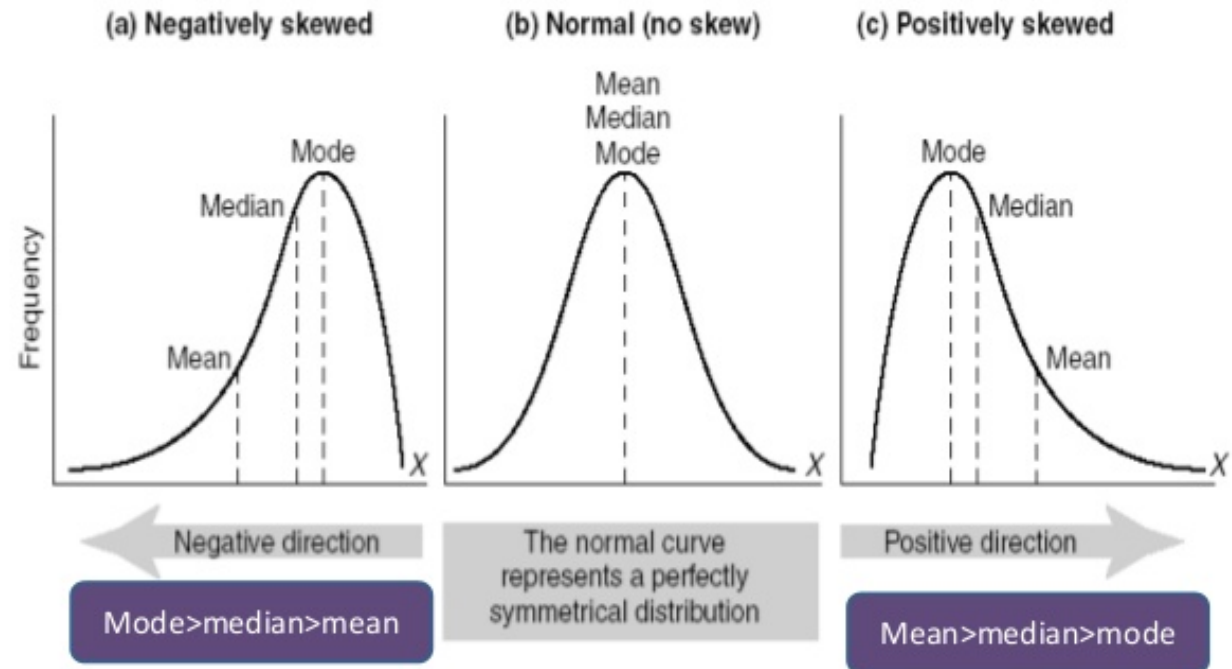


Fig. 1 Probability of having OV along with the number of small platelets for each CTP class

When normality is not there?

Position of mean median mode



Parametric vs Non-parametric tests

- **Parametric** statistical test is one that makes assumptions about the parameters (defining properties) of the population distribution(s) from which one's data are drawn
- **Non-parametric** test is one that makes no such assumptions.
 - In this strict sense, "non-parametric" is essentially a null category, since virtually all statistical tests assume one thing or another about the properties of the source population(s).

Parametric tests

One sample

One sample *t*-test

Two-sample

Paired *t*-test

Unpaired *t*-test

K-sample

ANOVA

Two-way ANOVA (repeated
measure ANOVA)

Pearson correlation coefficient (*r*)

ANOVA – Analysis of variance

Non-parametric tests

Sign test

Wilcoxon's signed rank test

Sign test

Wilcoxon's signed rank test

Mann-Whitney U-test

Kolmogorov-Smirnov test

Kruskal-Wallis test

Jonckheere test

Friedman test

Spearman rank order (ρ)



Contact: dileepa@kln.ac.lk

• bit.ly/stat_workshop_feedback

Feedback

1. what did you like about this session?
2. what didn't you like about this session?
3. what did you learn from this session?



Thank You

