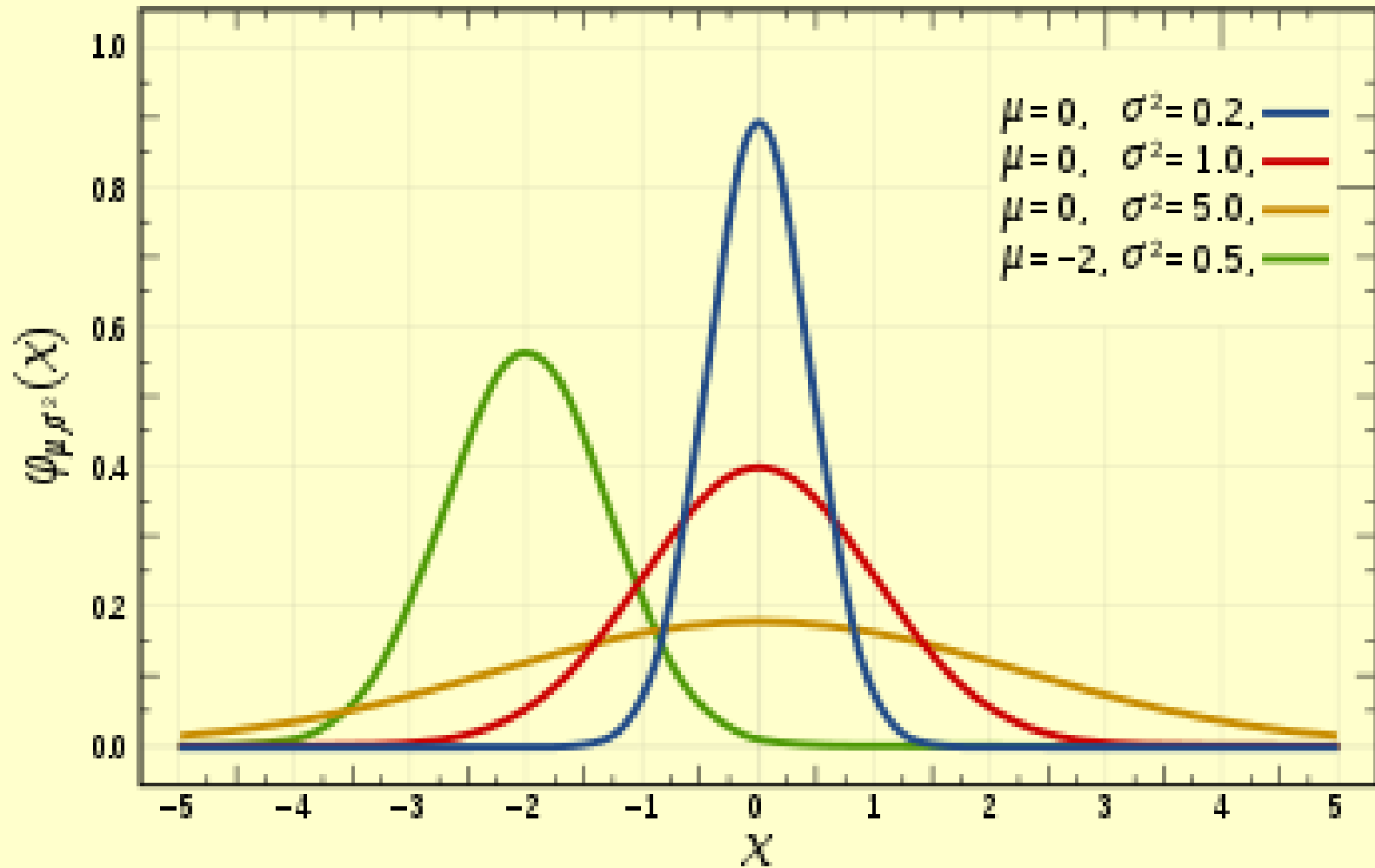
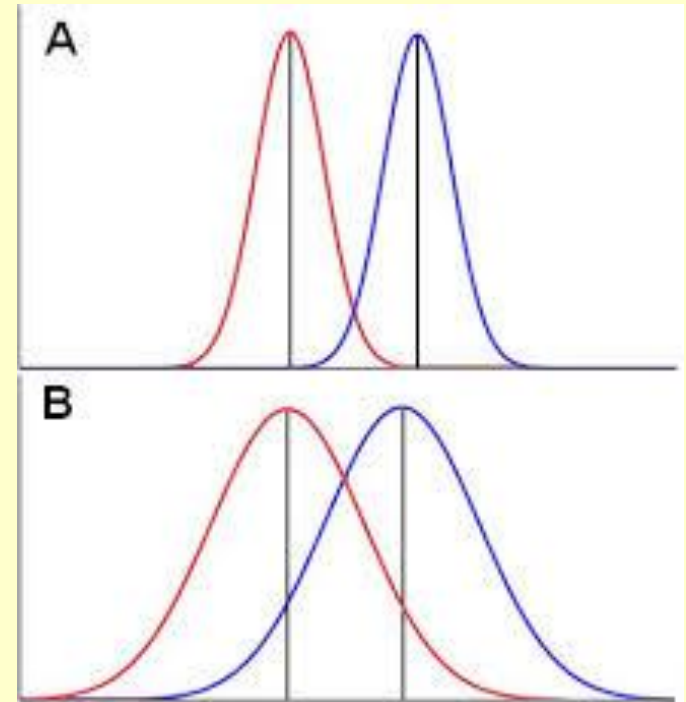


# Data Distributions and Normality



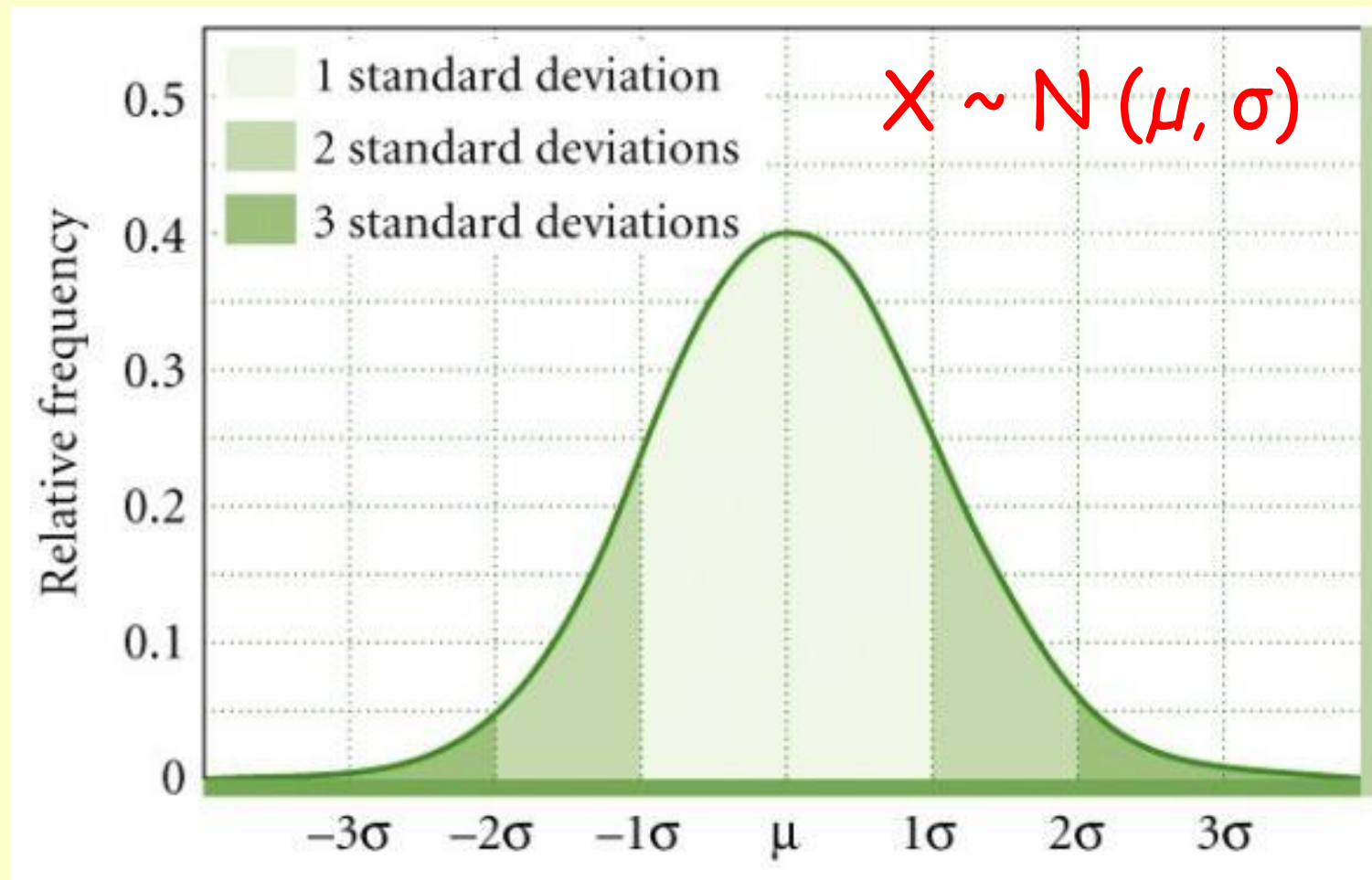
# Definition - (Non)Parametric

**Parametric statistics** assume that data come from a normal distribution, and make inferences about parameters of that distribution. These statistical tests are based on comparing the means (central tendency) of the distributions, as a function of their variability (spread).



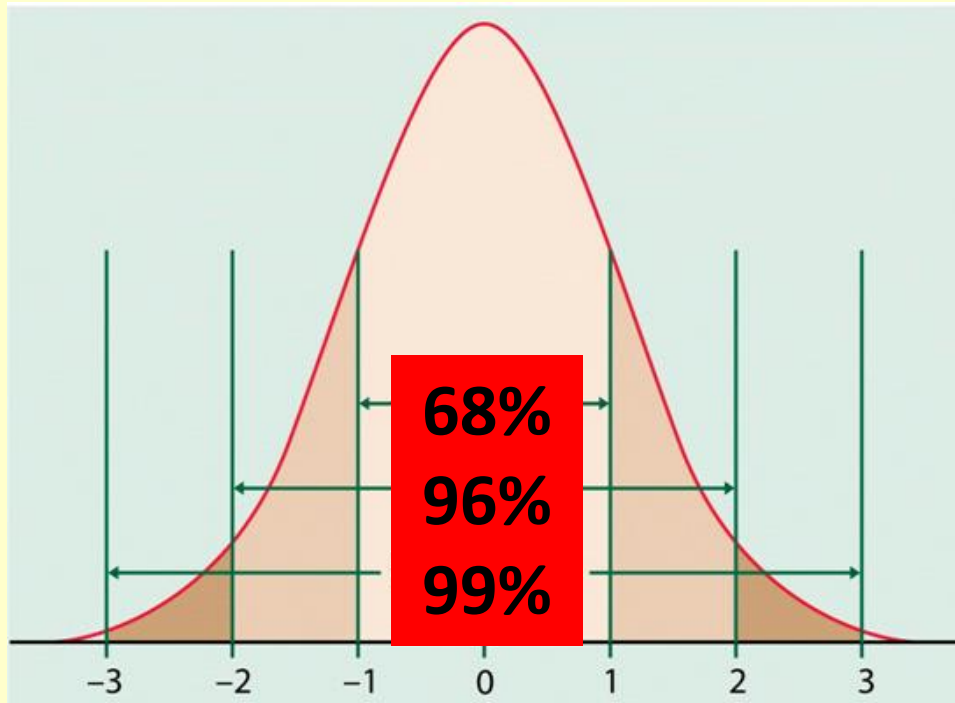
**Non-parametric statistics** do not depend on fitting a parameterized distribution, based on normality. These statistical tests are based on comparing the medians (50 % of data distributions) and the ranks of the observations amongst the samples.

# The Normal Distribution



Every Normal Distribution can be described using only two parameters: Mean and S.D.

# Is the Basis of Parametric Statistics



Parametric statistical methods require that numerical variables approximate a **normal distribution**.

They compare the **means & S.D.s**

In a normal distribution:

- ~ 68% observations within 1 standard deviation of mean
- ~ 96% within 2 standard deviations
- ~ 99% within 3 standard deviations

# Assessing Normality

- Three ways to assess the normality of the data
  - 1) Graphical Displays
    - Histogram, Density plot, Boxplot, Q-Q Plot
  - 2) Skewness / Kurtosis
    - Are they different from 0? (normal distribution)
    - Rule of Thumb: Too Large ( $> 1$ ) or too small ( $< -1$ )
  - 3) Shapiro - Wilk Tests
    - Tests if data differ from a normal distribution
    - Significant = non-Normal data
    - Non-Significant = Normal data

# Assessing Normality

- Three ways to assess the normality of the data
  - 1) Graphical Displays
    - Histogram, Density plot, Boxplot

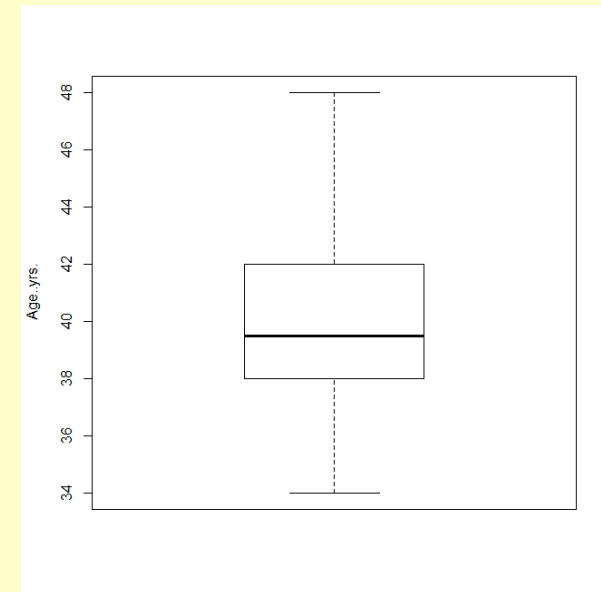
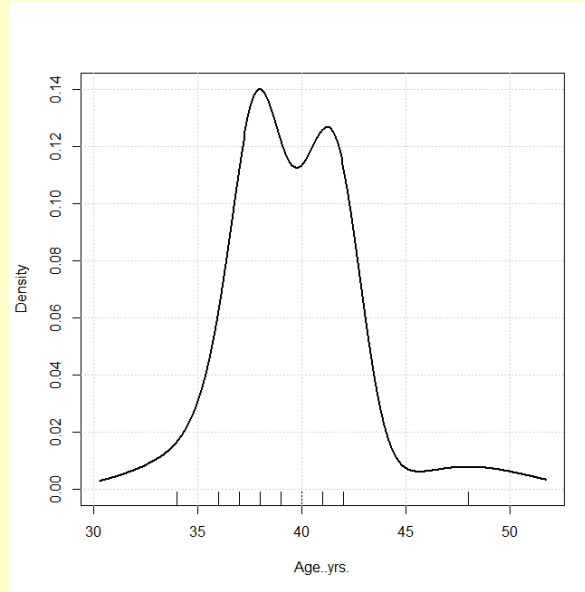
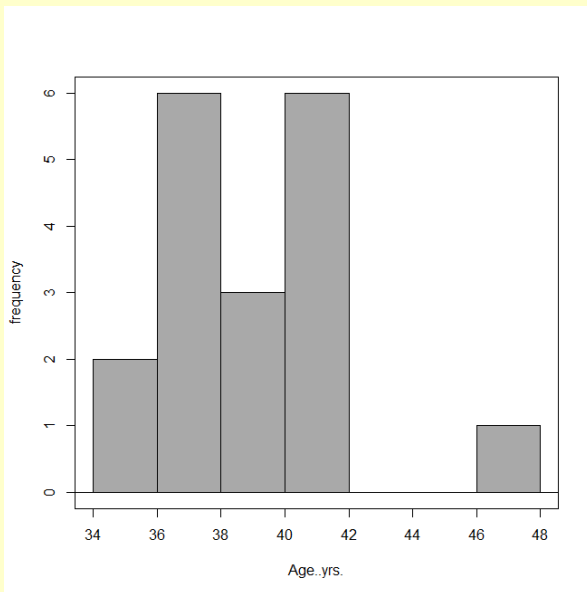
Graphs	Models	Distributions
Color palette...		
Index plot...		
Dot plot...		
Histogram...		
Plot discrete numeric variable...		
Density estimate...		
Stem-and-leaf display...		
Boxplot...		
Quantile-comparison plot...		
Symmetry boxplot...		

Graphs	Models	Distributions
Color palette...		
Index plot...		
Dot plot...		
Histogram...		
Plot discrete numeric variable...		
Density estimate...		
Stem-and-leaf display...		
Boxplot...		
Quantile-comparison plot...		
Symmetry boxplot...		

Graphs	Models	Distributions
Color palette...		
Index plot...		
Dot plot...		
Histogram...		
Plot discrete numeric variable...		
Density estimate...		
Stem-and-leaf display...		
Boxplot...		
Quantile-comparison plot...		
Symmetry boxplot...		

# Assessing Normality

- Three ways to assess the normality of the data
  - 1) Graphical Displays
    - Histogram, Density plot, Boxplot



# Assessing Normality

## ➤ 1) More Graphical Displays

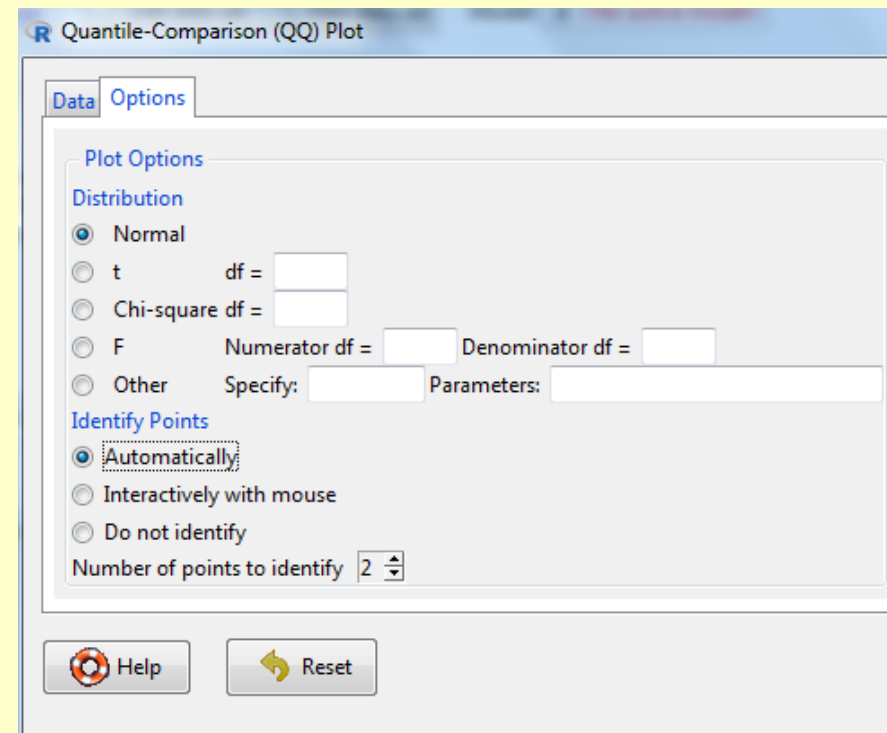
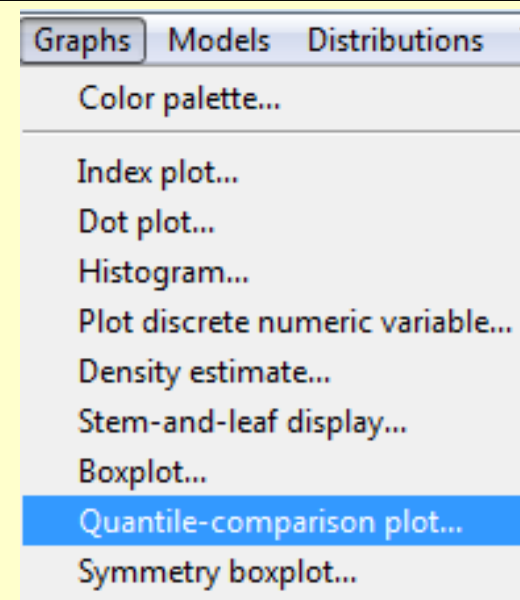
– Q-Q Plot: quantile / quantile plot

compares observed data and theoretical data, from a normal distribution

**OPTIONS** tab:

Select the type and the parameters of theoretical data distribution.

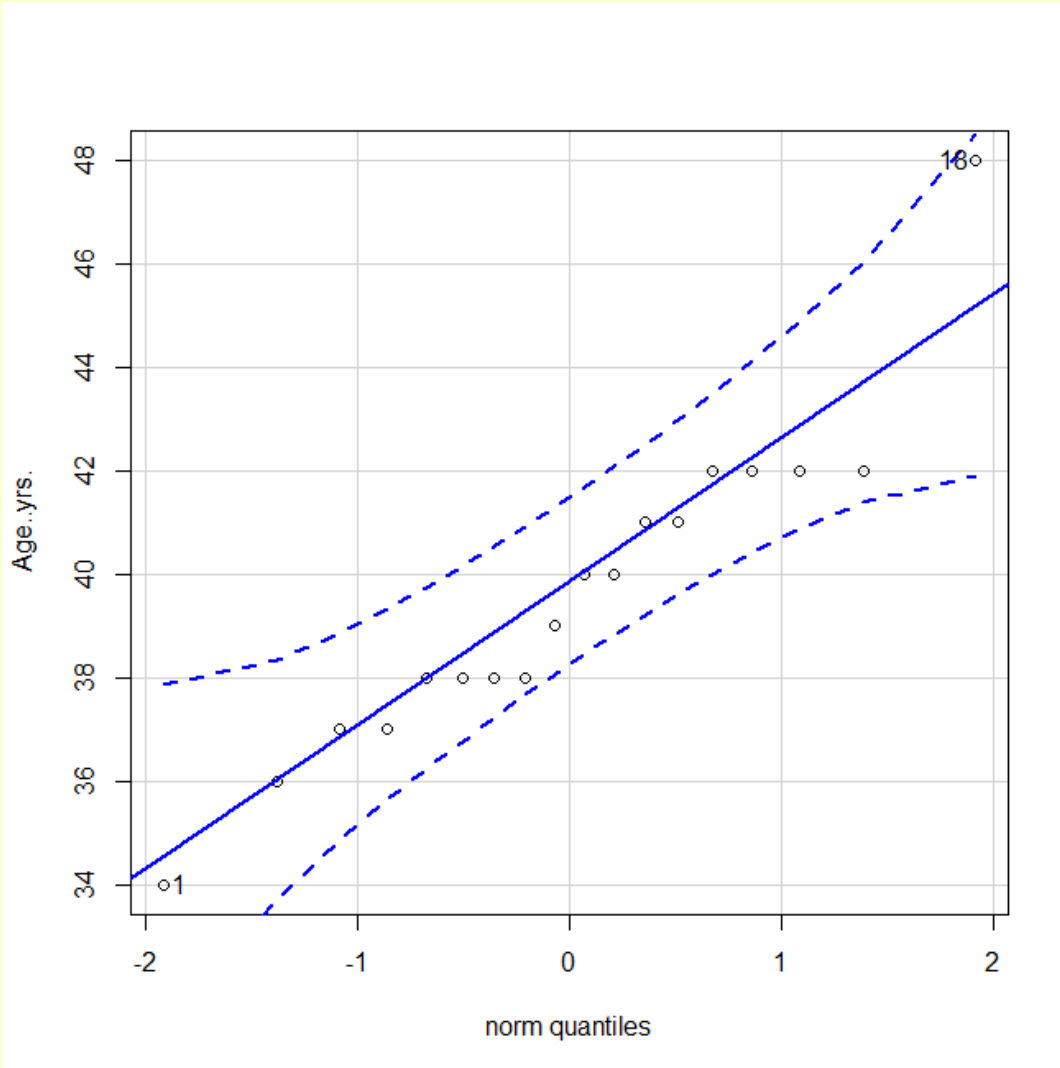
Default: "Normal"





# Assessing Normality

➤ Q-Q Plot: quantile / quantile plot



Things to Look For:

How many points plotted?

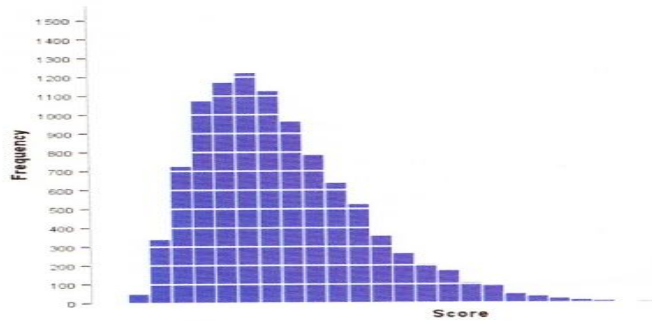
Are there any outliers?

# Quantifying Distributions

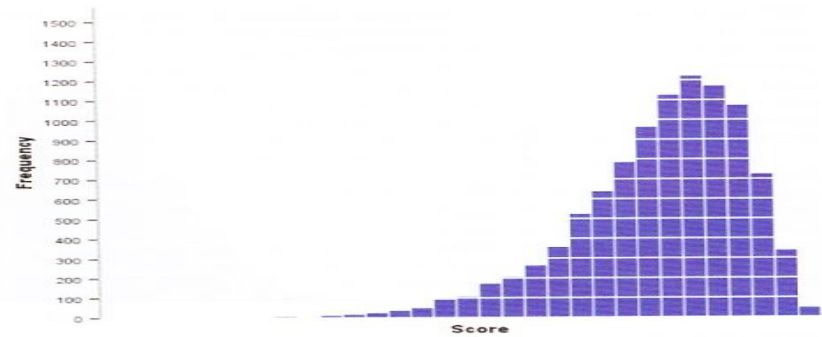
## 2) Skewness: Distribution symmetry (skew)

Skew: Measure of the symmetry of a distribution.

Symmetric distributions have a skew = 0.



Positive skew:  
the mean is larger  
than the median,  
skewness  $> 0$

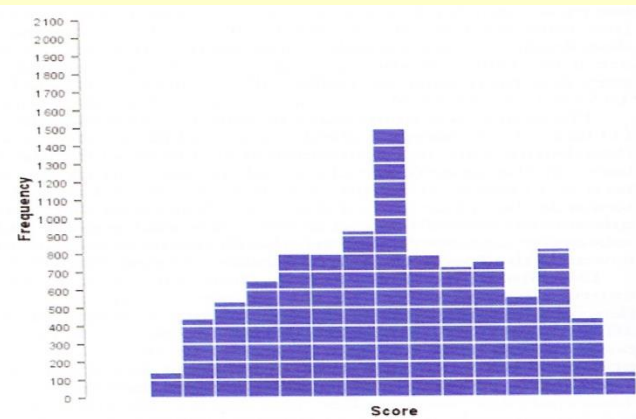
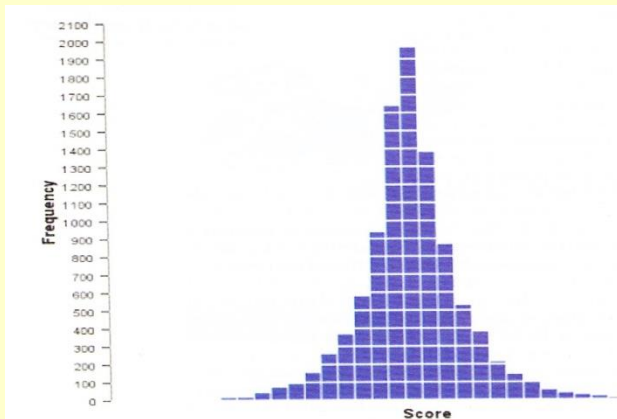


Negative skew:  
the mean is smaller  
than the median,  
skewness  $< 0$

# Quantifying Distributions

## 2) Kurtosis: Distribution of data in peak / tails

Kurtosis: Measure of the degree to which observations cluster in the tails or the center of the distribution.



### Positive kurtosis:

Less values in tails and more values close to mean.  
Leptokurtic.

### Negative kurtosis:

More values in tails and less values close to mean.  
Platykurtic.

# Assessing Normality - Example

- Use "Normality.Example.xls" Dataset (posted on class web-site)
- Follow along this example using Rcmdr
- Open Rstudio and activate Rcmdr
- Import dataset and start exploring

# An Example in Estimation

How old is your professor ?

**N = 18 guesses**

**Range = 34 – 48**

Age (yrs)
34
36
37
37
38
38
38
38
39
40
40
41
41
42
42
42
42
48

# An Example in Estimation

How old is your professor ?

**N = 18 guesses**

**What is the  
Midpoint Value =**

Age (yrs)
34
36
37
37
38
38
38
38
39
40
40
41
41
42
42
42
42
48

# An Example in Estimation

**N = 18 guesses**

**Mean = 39.6**

**Median = 39.5**

**S.D. = 3.1**

value	frequency	relative frequency
34	1	0.056
35	0	0.000
36	1	0.056
37	2	0.111
38	4	0.222
39	1	0.056
40	2	0.111
41	2	0.111
42	4	0.222
43	0	0.000
44	0	0.000
45	0	0.000
46	0	0.000
47	0	0.000
48	1	0.056
<b>sum</b>	<b>18</b>	<b>1</b>

# An Example in Estimation

**N = 18 guesses**

**50% = 39.5**

**5% = 34**

**25% = 38**

**75% = 42**

**95% = 48**

value	relative freq.	cumulative freq.
34	0.056	0.056
35	0.000	0.056
36	0.056	0.111
37	0.111	0.222
38	0.222	0.444
39	0.056	0.500
40	0.111	0.611
41	0.111	0.722
42	0.222	0.944
43	0.000	0.944
44	0.000	0.944
45	0.000	0.944
46	0.000	0.944
47	0.000	0.944
48	0.056	1.000
sum	1	9.389

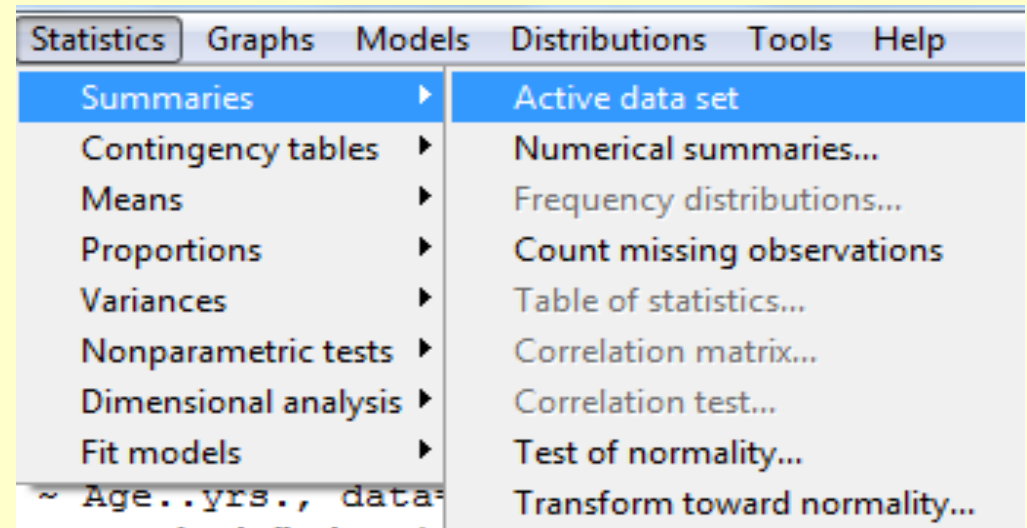


# Data Summary with Rcmdr

## Summaries:

- Active data set

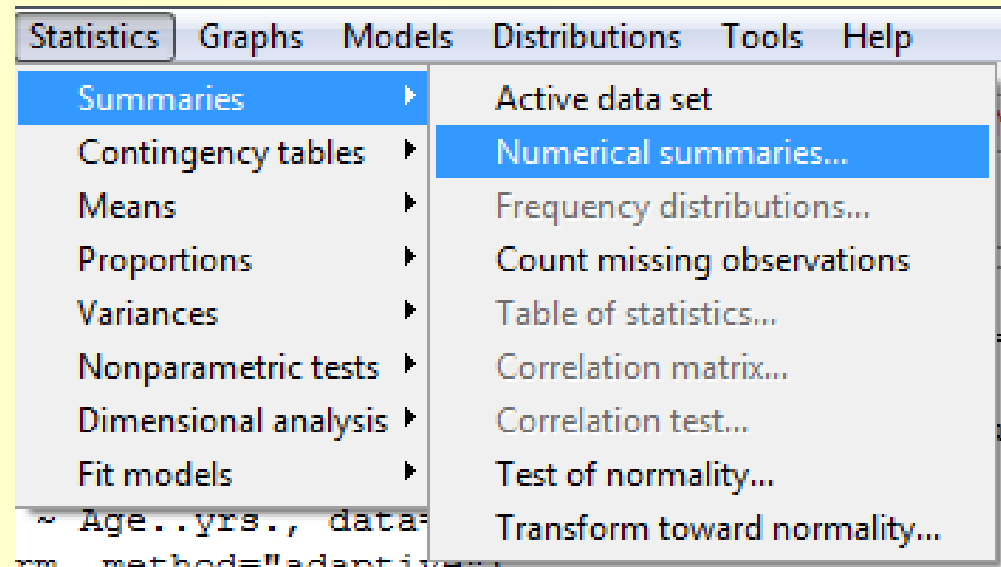
```
Rcmdr> summary(age)
  Age..yrs.
Min.      :34.00
1st Qu.  :38.00
Median   :39.50
Mean     :39.61
3rd Qu.  :41.75
Max.     :48.00
```



# Data Summary with Rcmdr

## Summaries:

- Numerical summaries

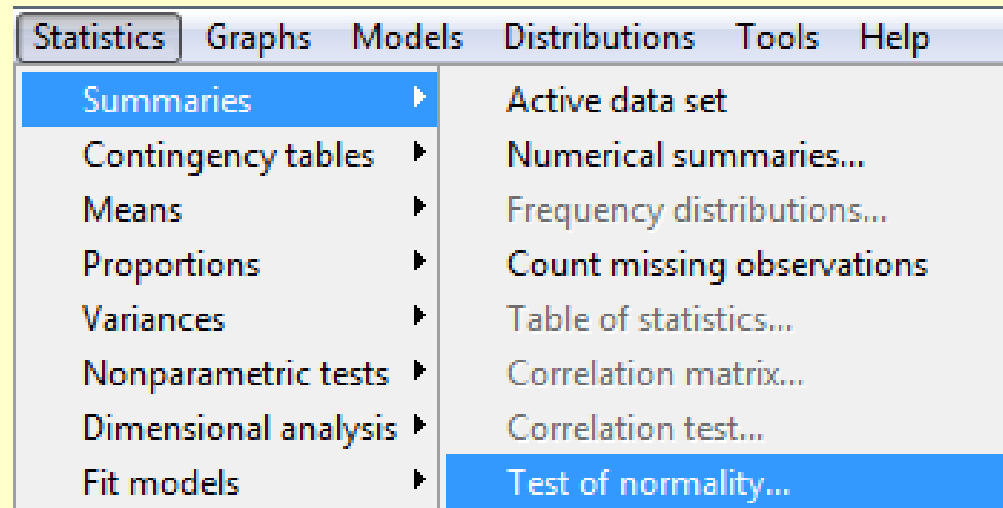


0%	25%	50%	75%	100%	n	se(mean)
34	38	39.5	41.75	48	18	0.7371509

mean	sd	IQR	skewness	kurtosis
39.61111	3.127466	3.75	0.8091976	1.970161

# Normality Test with Rcmdr

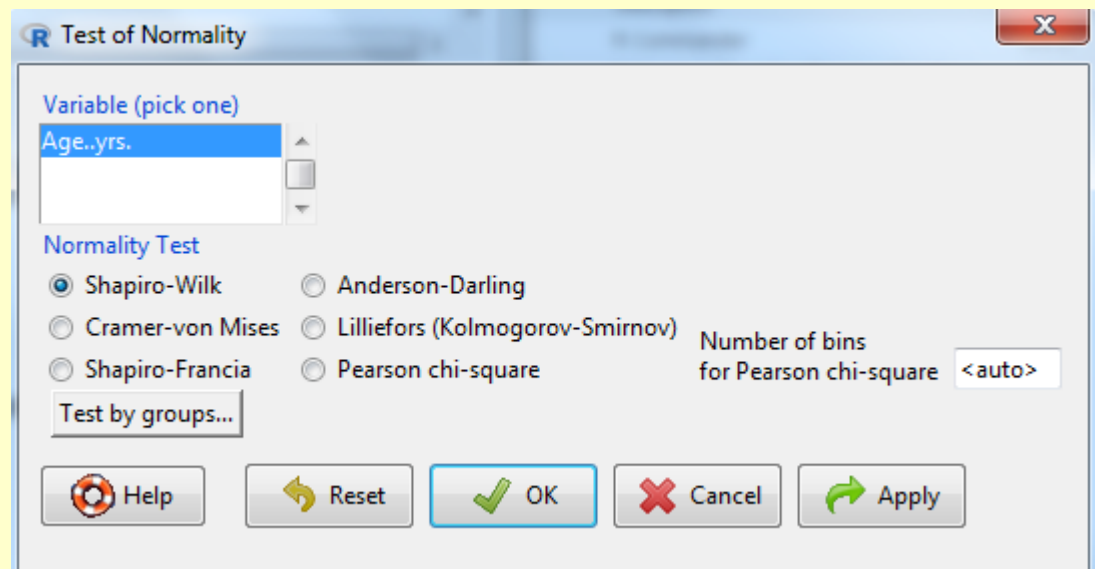
## Test of Normality



Select data

Use Shapiro-Wilk

Test multiple data  
using "by groups"



# Normality Test with Rcmdr

Test of Normality: SW (Wilk Sidak) Test

Null Hypothesis: Data ARE Normal

Alternate Hypothesis: Data ARE NOT Normal

```
Rcmdr> normalityTest(~Age..yrs., test="shapiro.test", data=age)
      Shapiro-wilk normality test

data:  Age..yrs.
W = 0.931, p-value = 0.2022
```

# Normality Test with Rcmdr

Test of Normality: SW (Wilk Sidak) Test

```
Rcmdr> normalityTest(~Age..yrs., test="shapiro.test", data=age)
      shapiro-wilk normality test

data:  Age..yrs.
W = 0.931, p-value = 0.2022
```

Is this Result Significant ? How Can You Tell ?

P value > 0.05 (alpha). Result is NOT Significant

Null is not Rejected. Data ARE Normally Distributed

What do you Need to Report ?

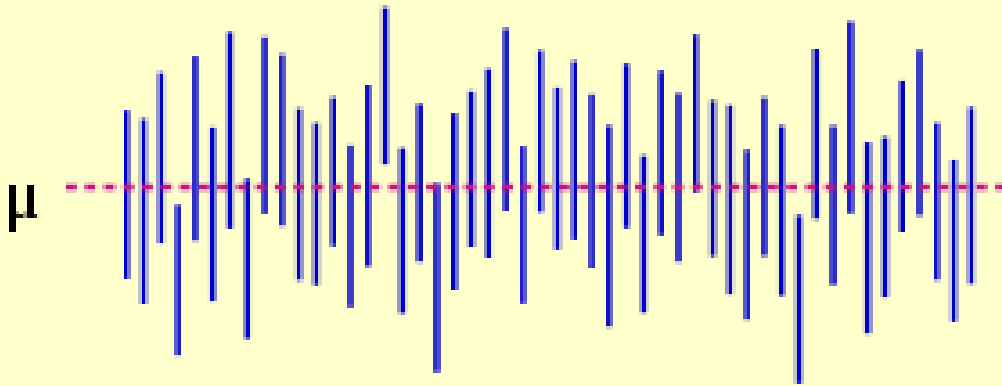
Test Name, Sample Size (n OR df), test statistic, p value

# Confidence Intervals - Many Tests

Formulation = 95% confidence intervals

Lower bound:  $\text{Mean} - (1.96 * \text{SE})$

Upper bound:  $\text{Mean} + (1.96 * \text{SE})$



By definition: 95% of the confidence intervals (from different experiments) will overlap the real parameter  $\mu$

# NOTE: Estimates Depend on Sample Size

C.I. Formulation: Mean +/- (Z score \* SE)  
Mean +/- (1.96 \* SE)

$$\text{S.E.} = \text{S.D.} / \text{sqrt}(n) = \\ 3.127466 / (\text{sqrt}(18)) = \mathbf{0.737151}$$

n	mean	SD	sqrt(n)	SE	95% CI
3	38.3	1.5	1.7	0.9	1.7
6	40.2	4.4	2.4	1.8	3.5
9	40.1	3.5	3.0	1.2	2.3
12	39.9	3.2	3.5	0.9	1.8
15	39.7	3.0	3.9	0.8	1.5
18	39.6	3.1	4.2	0.7	1.4

# NOTE: Estimates are influenced by chance

Age Estimate: 39.6 years (SD = 3.1)

C.I. Formulation: Mean +/- (Z score \* SE)  
Mean +/- (1.96 \* SE)

$$S.E. = S.D. / \text{sqrt}(n)$$

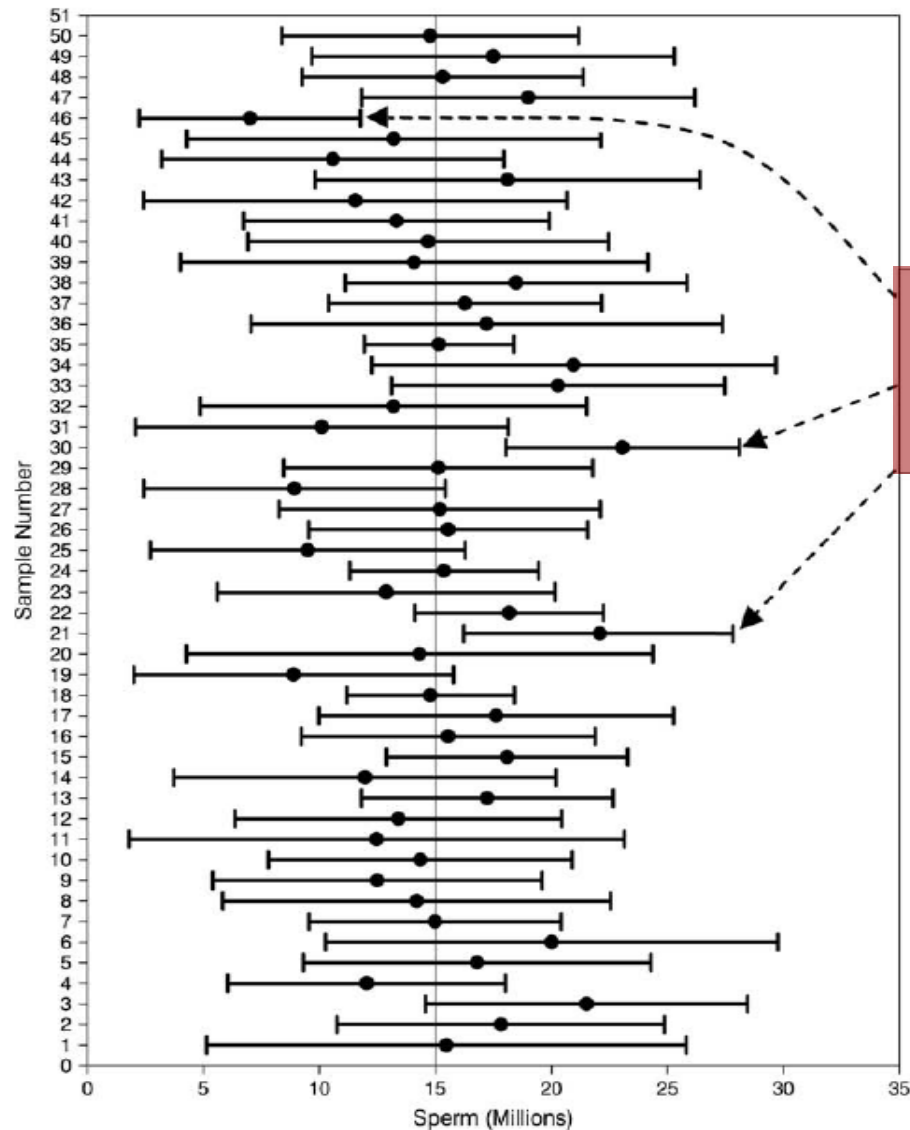
n	mean	SD	sqrt(n)	SE	95% CI	lower	upper
9	40.1	3.5	3.0	1.2	2.3	37.8	42.4
9	39.1	2.8	3.0	0.9	1.8	37.3	40.9

Are these two samples from the same population ?



# Interpreting Confidence Intervals

The (CI) is the interval that includes the estimated parameter, with a probability determined by confidence level (usually 95%).



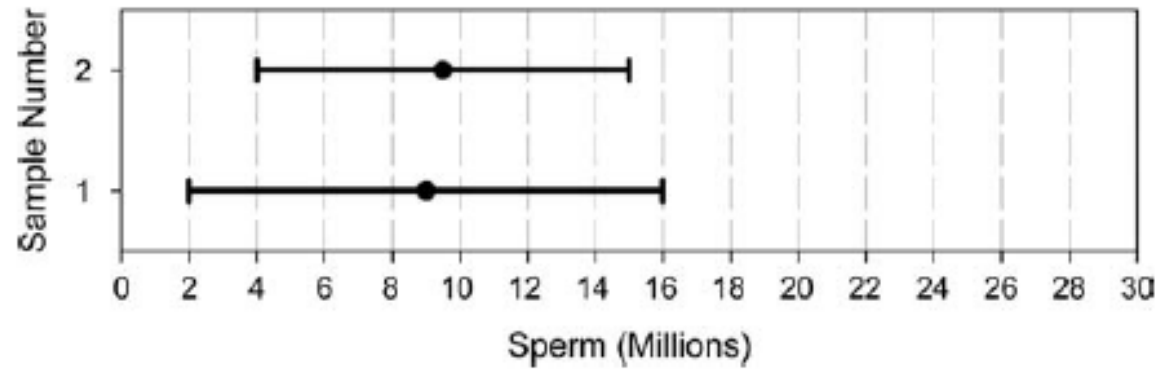
**NOTE**

These intervals don't contain the 'true' value of the mean

# Interpreting Confidence Intervals

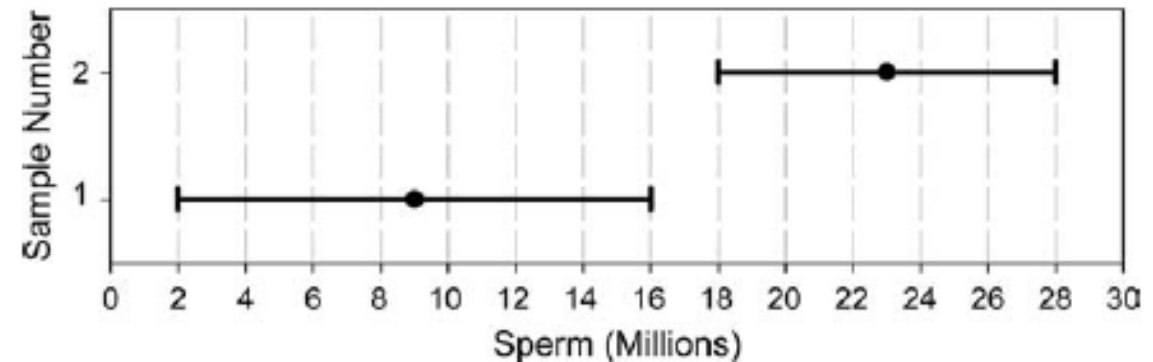
## Case 1.

Two samples indistinguishable. They are from same population



## Case 2.

Two samples different. They are not from same population



# Summary - Parametric Statistics

## Benefits and Costs:

- Parametric methods make more assumptions than non-parametric methods. If the extra assumptions are correct, parametric methods have more statistical power (produce more accurate and precise estimates.)
- However, if those assumptions are incorrect, parametric methods can be very misleading. They can cause false positives (type -I errors). Thus, they are often not considered robust.

# Summary - Normality

➤ Indicators of a normal (Gaussian) distribution

A. Mean = Median = Mode

B. Skewness:

Measures asymmetry of the distribution. A value of zero indicates symmetry. Skewness absolute value  $> 1$  indicates non-normal skewed distribution.

C. Kurtosis:

Measures the distribution of mass in the distribution. A value of zero indicates a normal distribution. Kurtosis absolute value  $> 1$  indicates non-normal unbalanced distribution.

# Summary - Approach

## Suggested Approach:

- Use parametric tests - whenever possible.
- Take care to examine diagnostic statistics and to determine if extra assumptions are met.
- If you are in doubt...  
Perform the matching non-parametric test and compare results.

If they agree: go with results of normal test

If they disagree: what caused the disagreement