Babraham Institute

Introduction to Statistics with R

Anne Segonds-Pichon

v2019-07

Babraham Bioinformatics

Outline of the course

- Short introduction to Power analysis
- Analysis of qualitative data:
  - **Chi-square test**
- Analysis of quantitative data:
  - **Student's *t*-test, One-way ANOVA and correlation**

R packages needed

beanplot

pastecs

plotrix

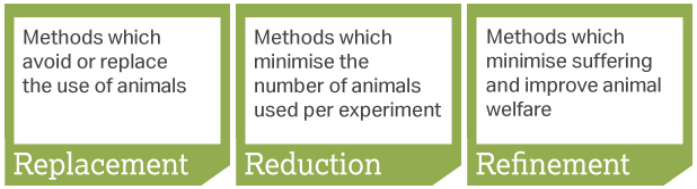
reshape2

Power analysis

- **Definition of power**: probability that a statistical test will reject a false null hypothesis (H_0).
  - **Translation**: the probability of detecting an effect, given that the effect is really there.

- **In a nutshell**: the bigger the experiment (big sample size), the bigger the power (more likely to pick up a difference).

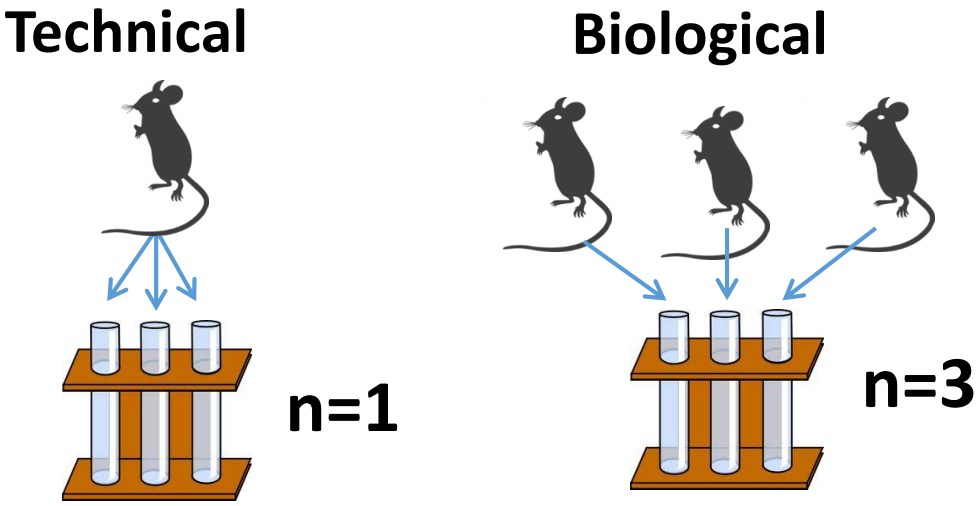
- Main output of a **power analysis**:
  - Estimation of an appropriate **sample size**
    - **Too big**: waste of resources,
    - **Too small**: may miss the effect ($p>0.05$)+ waste of resources,
    - **Grants**: justification of sample size,
    - **Publications:** reviewers ask for power calculation evidence,
    - **Home office**: the 3 Rs: Replacement, **Reduction** and Refinement.



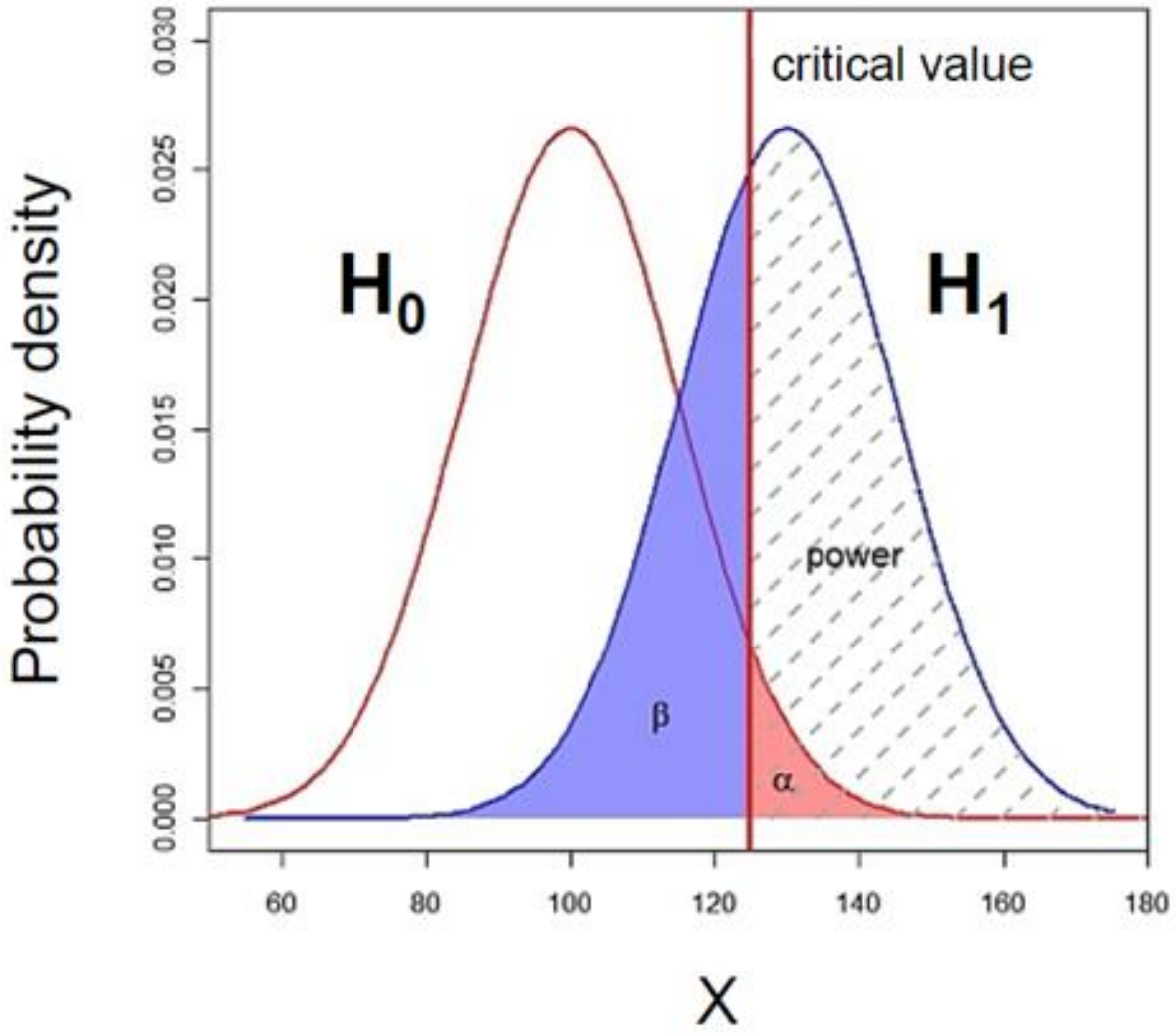
Experimental design

Think stats!!

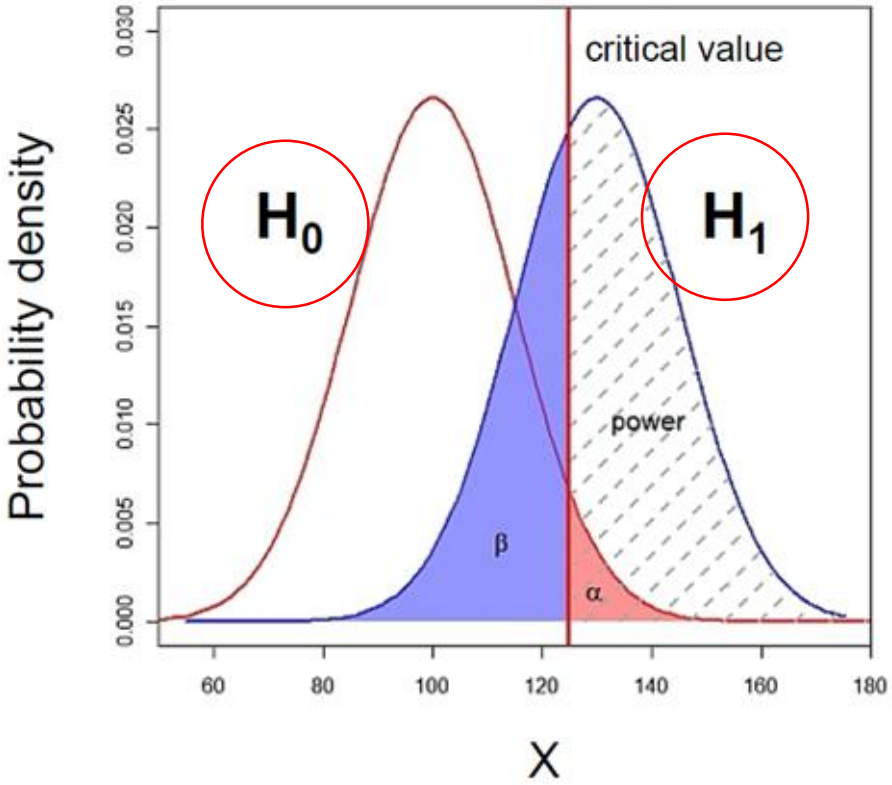
- Translate the hypothesis into statistical questions:
  - What type of data?
  - What statistical test ?
  - **What sample size?**
- Very important: Difference between **technical** and **biological** replicates.



What does Power look like?

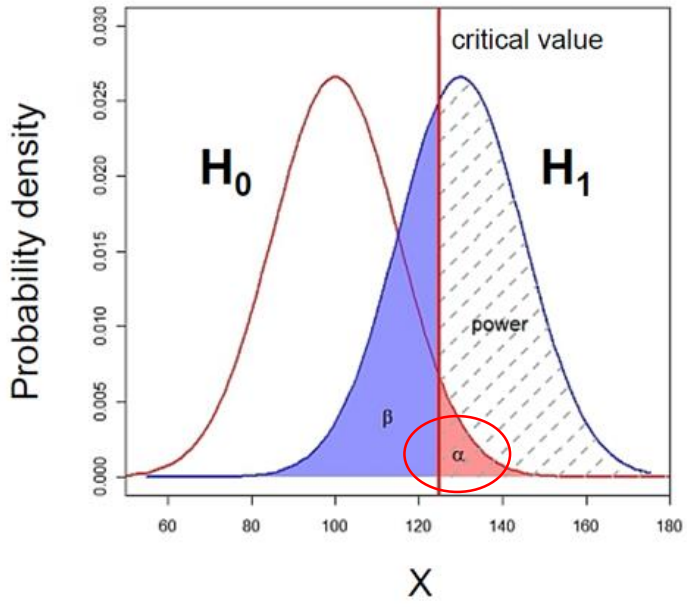


What does Power look like? Null and alternative hypotheses



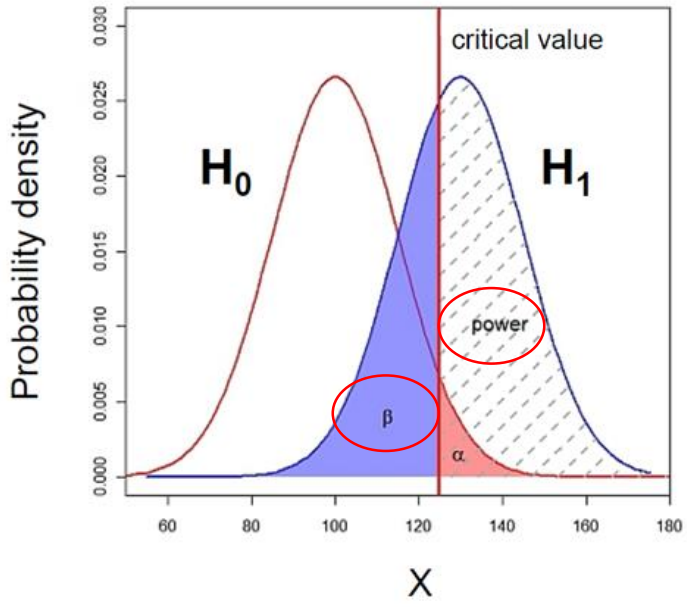
- Probability that the observed result occurs if $\mathbf{H_0}$ is true
  - H_0 : **Null hypothesis** = absence of effect
  - H_1: **Alternative hypothesis** = presence of an effect

What does Power look like? Type I error α



- **α :** the threshold value that we measure p-values against.
  - For results with 95% level of confidence: **α = 0.05**
  - = probability of **type I error**
- **p-value**: probability that the observed statistic occurred by chance alone
- **Statistical significance**: comparison between **α** and the **p-value**
  - p-value < 0.05: reject H_0 and p-value > 0.05: fail to reject H_0

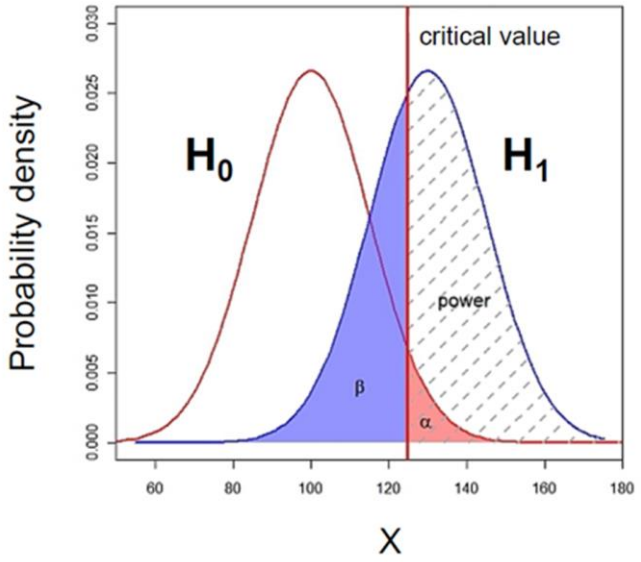
What does Power look like? Power and Type II error β



- **Type II error (β)** is the failure to reject a <u>false</u> H_0
  - Probability of missing an effect which is really there.
  - **Power**: probability of detecting an effect which is really there

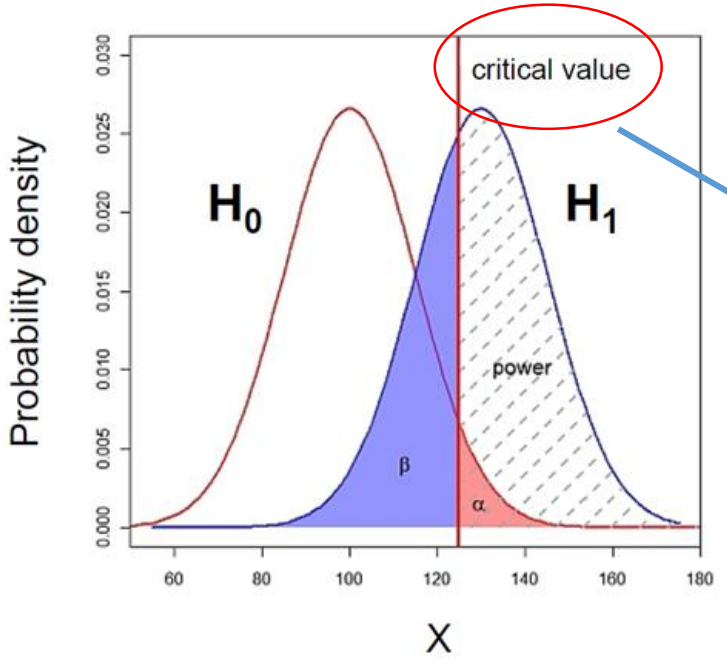
  - Direct relationship between **Power** and **type II error**:
    - **Power** = 1 – **β**

What does Power look like? Power = 80%



- **Type II error (β)** is the failure to reject a <u>false</u> H_0
  - Probability of missing an effect which is really there.
  - **Power**: probability of detecting an effect which is really there
    - Direct relationship between **Power** and type II error:
    - if **Power** = 0.8 then β = 1- **Power** = 0.2 (20%)
  - Hence a true difference will be missed 20% of the time
  - **General convention: 80%** but could be more
  - Cohen (1988):
    - For most researchers: Type I errors are four times more serious than Type II errors so **0.05 * 4 = 0.2**
    - Compromise: 2 groups comparisons:
      - 90% = +30% sample size
      - 95% = +60%s sample size

What does Power look like? Critical value



df	0.20	0.10	0.05	0.02	0.01	0.001
1	3.0777	6.3138	12.7062	31.8205	63.6567	636.6192
2	1.8856	2.9200	4.3027	6.9646	9.9248	31.5991
3	1.6377	2.3534	3.1824	4.5407	5.8409	12.9240
4	1.5332	2.1318	2.7764	3.7469	4.6041	8.6103
5	1.4759	2.0150	2.5706	3.3649	4.0321	6.8688
6	1.4398	1.9432	2.4469	3.1427	3.7074	5.9588
7	1.4149	1.8946	2.3646	2.9980	3.4995	5.4079
8	1.3968	1.8595	2.3060	2.8965	3.3554	5.0413
9	1.3830	1.8331	2.2622	2.8214	3.2498	4.7809
10	1.3722	1.8125	2.2281	2.7638	3.1693	4.5869
11	1.3634	1.7959	2.2010	2.7181	3.1058	4.4370
12	1.3562	1.7823	2.1788	2.6810	3.0545	4.3178
13	1.3502	1.7709	2.1604	2.6503	3.0123	4.2208
14	1.3450	1.7613	2.1448	2.6245	2.9768	4.1405
15	1.3406	1.7531	2.1314	2.6025	2.9467	4.0728

Example: 2-tailed t-test with n=15 (df=14)

- In **hypothesis testing**, a **critical value** is a point on the test distribution that is compared to the **test statistic** to determine whether to reject the null **hypothesis**
  - Example of test statistic: t-value
- If the absolute value of your **test statistic** is greater than the **critical value**, you can declare statistical significance and reject the null **hypothesis**
  - Example: t-value > critical t-value

To recapitulate:

- The null hypothesis (H_0): H_0 = no effect

- The aim of a statistical test is to reject or not H_0.

Statistical decision	True state of H_0	
	H_0 True (no effect)	**H_0 False (effect)**
Reject H_0	**Type I error α** **False Positive** 	Correct **True Positive** 
Do not reject H_0	Correct **True Negative** 	**Type II error β** **False Negative** 

- Traditionally, a test or a difference are said to be "**significant**" if the probability of type I error is: **α =< 0.05**
- **High specificity** = low **False Positives** = low **Type I error**
- **High sensitivity** = low **False Negatives** = low **Type II error**

Sample Size: Power Analysis

The power analysis depends on the relationship between 6 variables:

- the **difference** of biological interest
- the **variability** in the data (**standard deviation**)

} **Effect size**

- the significance level (5%)
- the desired power of the experiment (80%)
- the **sample size**
- the alternative hypothesis (ie one or two-sided test)

The difference of biological interest

- This is to be determined scientifically, not statistically.
  - minimum meaningful effect of biological relevance
  - the larger the effect size, the smaller the experiment will need to be to detect it.
- **How to determine it?**
  - Substantive knowledge, previous research, pilot study …

The Standard Deviation (SD)

- Variability of the data
- **How to determine it?**
  - Substantive knowledge, previous research, pilot study …

- In ‘power context’: **effect size**: combination of both:
  - e.g.: **Cohen’s d** = (Mean 1 – Mean 2)/Pooled SD

Power Analysis

The power analysis depends on the relationship between 6 variables:

- the **difference** of biological interest
- the **standard deviation**
- **the significance level (5%) ($p < 0.05$) α**
- **the desired power of the experiment (80%) β**
- the **sample size**
- the alternative hypothesis (ie one or two-sided test)

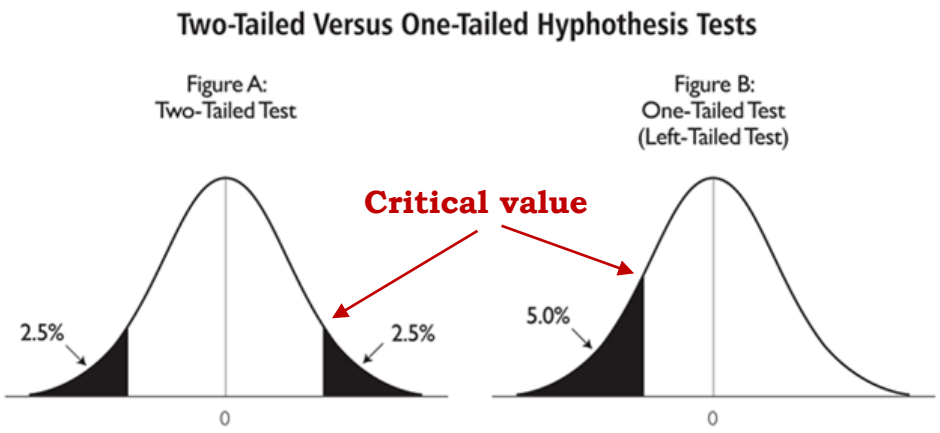
Power Analysis

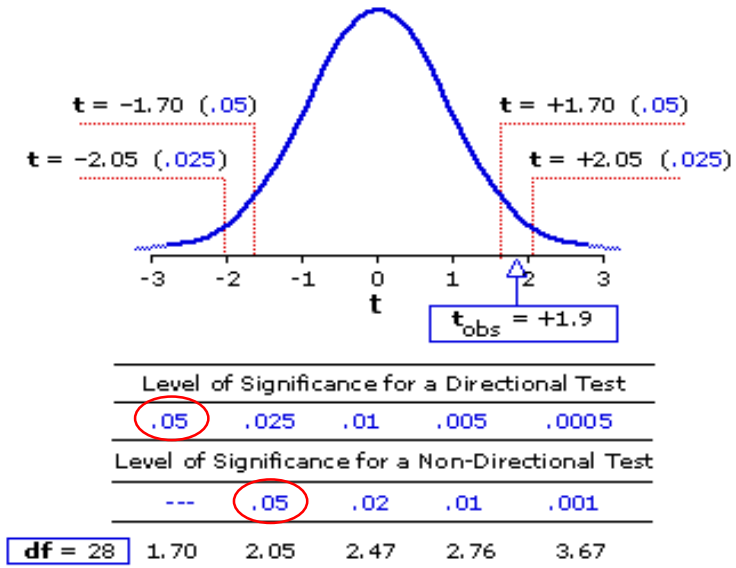
The power analysis depends on the relationship between 6 variables:

- the **effect size** of biological interest
- the **standard deviation**
- **the significance level (5%)**
- **the desired power of the experiment (80%)**
- the **sample size**
- the alternative hypothesis (ie one or two-sided test)

The alternative hypothesis: what is it?

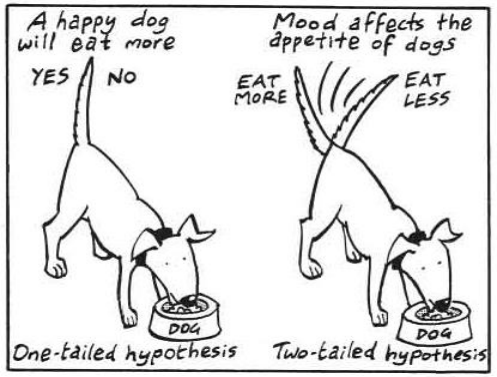
- One-tailed or 2-tailed test? One-sided or 2-sided tests?





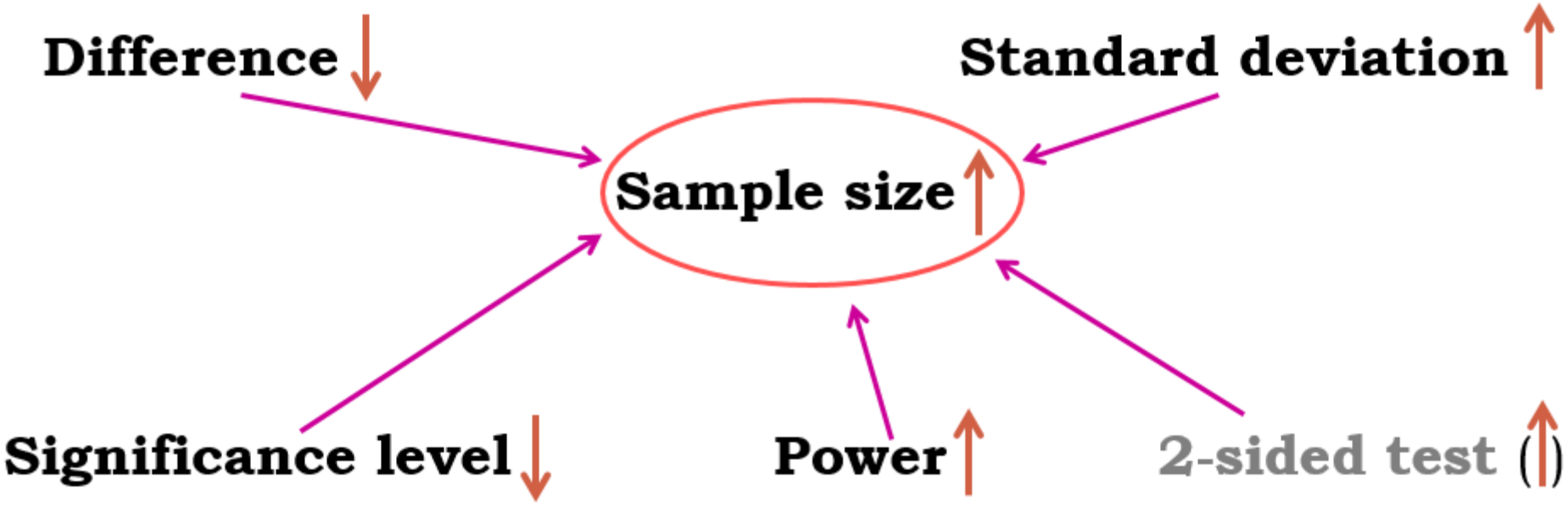
- Is the question:
  - Is the there a difference?
  - Is it bigger than or smaller than?

- Can rarely justify the use of a one-tailed test
- Two times easier to reach significance with a one-tailed than a two-tailed
  - Suspicious reviewer!



- **Fix any five of the variables and a mathematical relationship can be used to estimate the sixth**.

e.g. What sample size do I need to have a 80% probability (**power**) to detect this particular effect (**difference** and **standard deviation**) at a 5% **significance level** using a **2-sided test**?



- **Good news**:

there are packages that can do the power analysis for you ... providing you have some prior knowledge of the key parameters!

difference + standard deviation = effect size

- **Free packages**:
  - **R**
  - G*Power and InVivoStat
  - Russ Lenth's power and sample-size page:
    - http://www.divms.uiowa.edu/~rlenth/Power/

- Cheap package: StatMate (~ $95)

- Not so cheap package: MedCalc (~ $495)

Sample → **Statistical inference** → Population

Difference → Meaningful? → Yes → Real?

Real? → Statistical test → Statistic (e.g. t, F …) → Big enough?

Statistic = **Difference** + Noise + Sample

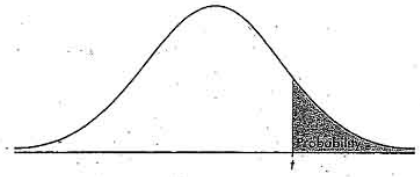


TABLE B: t-DISTRIBUTION CRITICAL VALUES

	Tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	.765	.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	.741	.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	.727	.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	.718	.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	.711	.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	.706	.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	.703	.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	.700	.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	.697	.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	.695	.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	.694	.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	.692	.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	.691	.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	.690	.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	.689	.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	.688	.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	.688	.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	.687	.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	.686	.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	.686	.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792

Qualitative data

Qualitative data

- = not numerical
- = values taken = usually names (also *nominal*)
    - e.g. causes of death in hospital
- Values can be numbers but not numerical
    - e.g. group number = numerical label but not unit of measurement
- Qualitative variable with intrinsic order in their categories = *ordinal*
- Particular case: qualitative variable with 2 categories: ***binary*** or *dichotomous*
    - e.g. alive/dead or presence/absence

Fisher's exact and Chi^2

Example: cats.dat

- Cats trained to line dance
- 2 different rewards: food or affection
- **Question**: Is there a difference between the rewards?

- **Is there a significant relationship between the 2 variables?**
  - does the reward significantly affect the likelihood of dancing?

- To answer this type of question:
  - **Contingency table**
  - **Fisher's exact or Chi^2 tests**

	Food	Affection
Dance	?	?
No dance	?	?

But first: **how many cats** do we need?

Power analysis: Fisher's test

- Preliminary results from a pilot study: **25%** line-danced after having received affection as a reward vs. **70%** after having received food.

```
power.prop.test(n = NULL, p1 = NULL, p2 = NULL , sig.level = NULL, power = NULL , alternative
= c("two.sided", "one.sided")
```

- Exactly one of the parameters `n`, `p1`, `p2`, `power` and `sig.level` must be passed as NULL, and that parameter is determined from the others. "two-sided" is the default.

```
power.prop.test(p1 = 0.25, p2 = 0.7, sig.level = 0.05, power = 0.8)
```

```
     Two-sample comparison of proportions power calculation

              n = 18.10585
             p1 = 0.25
             p2 = 0.7
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

Providing the effect size observed in the experiment is similar to the one observed in the pilot study, we will need 2 samples of about **18 cats** to reach significance ($p<0.05$) with a Fisher's exact test.

Plot ‘cats.dat’ (From raw data)

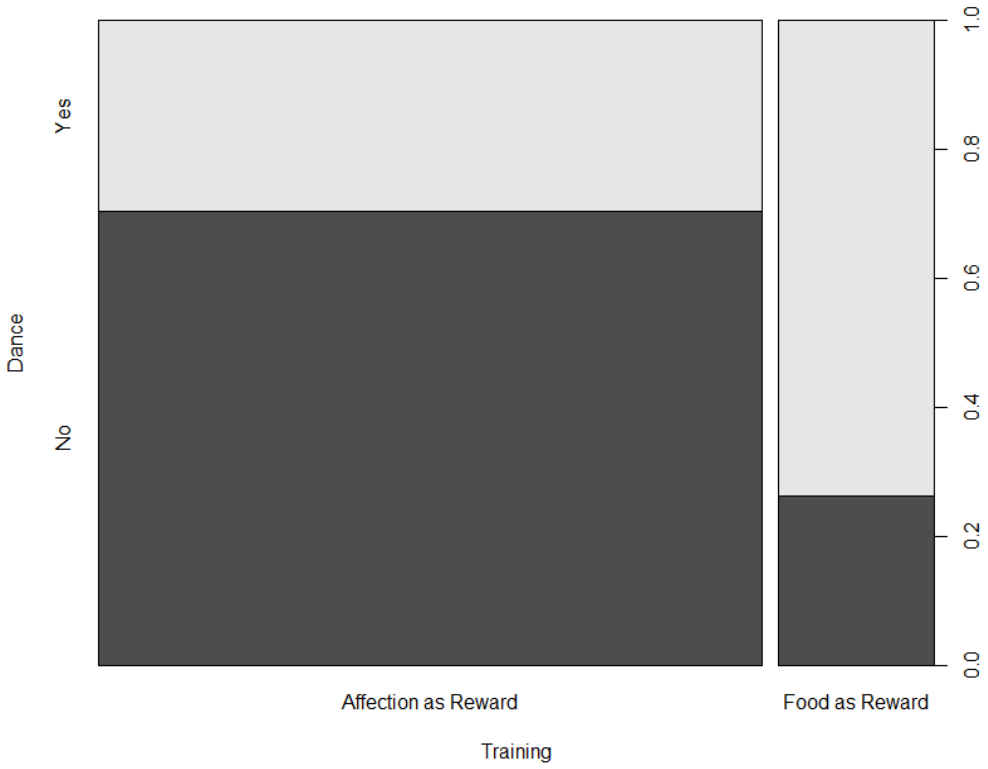
```
head(cats.data)
```

```
        Training Dance
1 Food as Reward   Yes
2 Food as Reward   Yes
3 Food as Reward   Yes
4 Food as Reward   Yes
5 Food as Reward   Yes
6 Food as Reward   Yes
```

```
plot(cats.data$Training, cats.data$Dance, xlab = "Training", ylab = "Dance")
```

```
table(cats.data)
```

```
                        Dance
Training                 No Yes
  Affection as Reward 114  48
  Food as Reward       10  28
```



Plot cats data (From raw data)

```
contingency.table <- table(cats)
contingency.table <- prop.table(contingency.table,1)
contingency.table100 <- round(contingency.table*100)
contingency.table100
```

```
                      Dance
Training                No Yes
  Affection as Reward 114  48
  Food as Reward       10  28
```

→

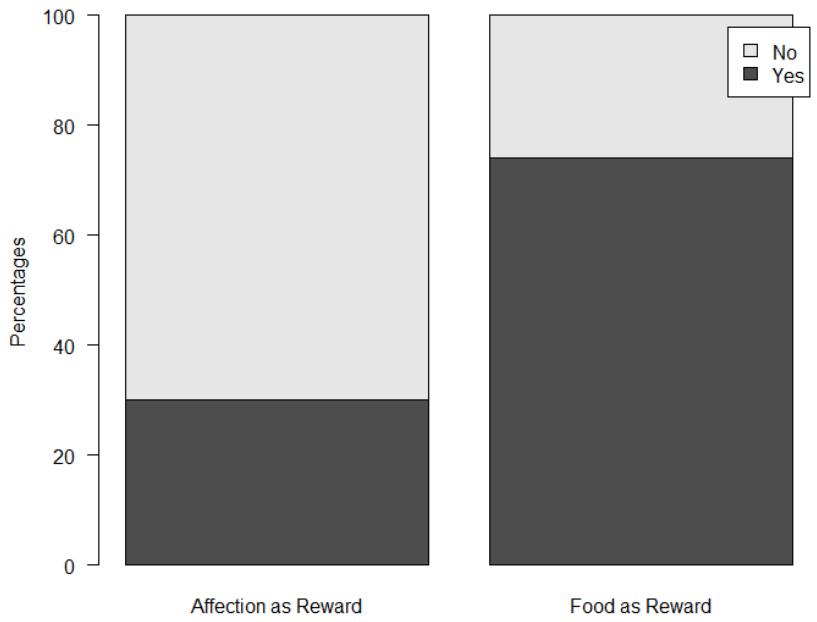
```
                          Dance
Training                     No       Yes
  Affection as Reward 0.7037037 0.2962963
  Food as Reward      0.2631579 0.7368421
```

→

```
                      Dance
Training                No Yes
  Affection as Reward 70  30
  Food as Reward      26  74
```

```
contingency.table100<-cbind(contingency.table100[,"Yes"],contingency.table100[,"No"])
colnames(contingency.table100) <- c("Yes", "No")
contingency.table100
```

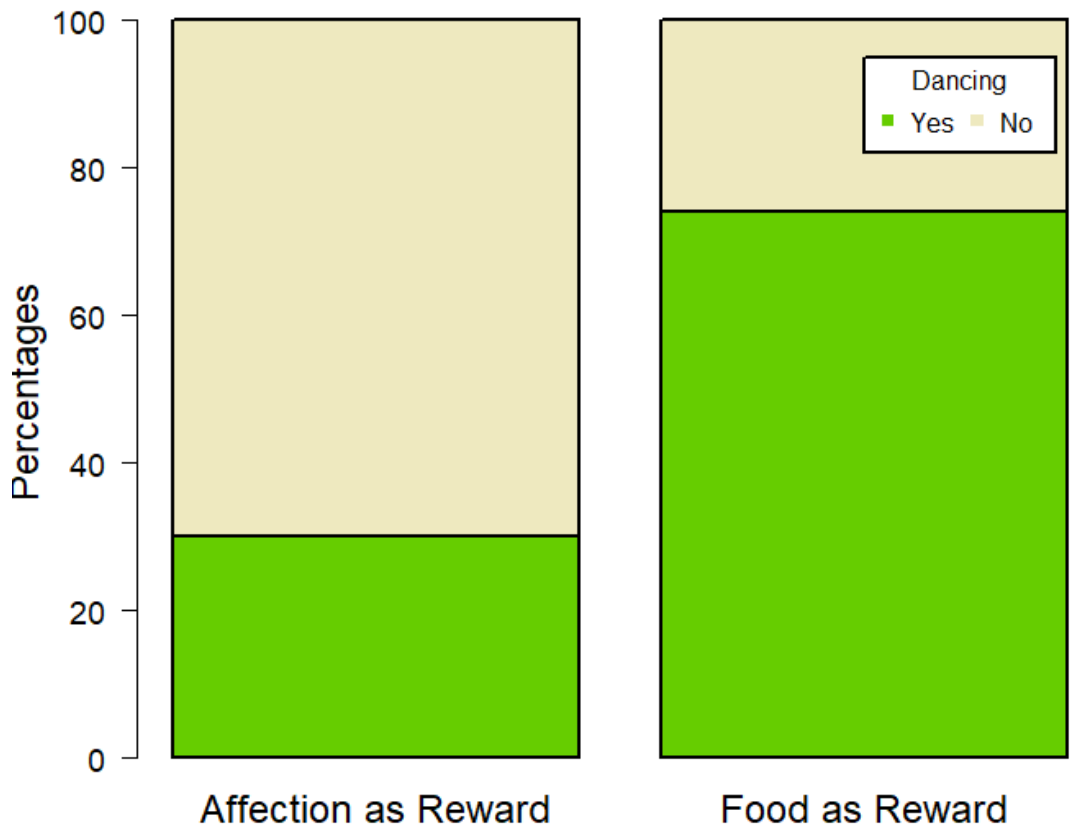
```
barplot(t(contingency.table100),
        legend.text=TRUE,
        ylab = "Percentages",
        las = 1
        )
```



Plot cats data (From raw data) Prettier!

```
barplot(t(contingency.table100),
        col=c("chartreuse3","lemonchiffon2"),
        cex.axis=1.2,
        cex.names=1.5,
        cex.lab=1.5,
        ylab = "Percentages",
        las=1)
```

```
legend("topright",
       title="Dancing",
       inset=.05,
       c("Yes","No"),
       horiz=TRUE,
       pch=15,
       col=c("chartreuse3","lemonchiffon2"))
```



Chi-square and Fisher's tests

- Chi^2 test very easy to calculate by hand but Fisher's very hard
- Many software will not perform a Fisher's test on tables > 2x2
- **Fisher's test more accurate** than Chi^2 test on **small samples**
- **Chi^2 test more accurate** than Fisher's test on **large samples**
- Chi^2 test assumptions:
  - 2x2 table: no expected count <5
  - Bigger tables: all expected > 1 and no more than 20% < 5
- Yates's continuity correction
  - All statistical tests work well when their assumptions are met
  - When not: probability Type 1 error increases
  - <u>Solution</u>: corrections that increase p-values
    - Corrections are dangerous: no magic
    - Probably best to avoid them

Chi-square test

- In a chi-square test, the observed frequencies for two or more groups are compared with expected frequencies by chance.

$$\chi^2 = \Sigma \frac{(Observed\ Frequency - Expected\ Frequency)^2}{Expected\ Frequency}$$

  - With observed frequency = collected data

- **Example with 'cats.dat'**

Chi-square test

- Formula for Expected frequency = **(row total)*(column total)/grand total**

Total Observations in Table: 200

cat.data$Training	cat.data$Dance No	Yes	Row Total
Affection as Reward	114 100.440 70.370%	48 61.560 29.630%	162 81.000%
Food as Reward	10 23.560 26.316%	28 14.440 73.684%	38 19.000%
Column Total	124	76	200

Example: expected frequency of cats line dancing after having received food as a reward:

Expected = (38*76)/200=14.44

Alternatively:
Probability of line dancing: 76/200
Probability of receiving food: 38/200

(76/200)*(38/200)=0.072

Expected: 7.2% of 200 = 14.44

$Chi^2 = (114-100.4)^2/100.4 + (48-61.6)^2/61.6 + (10-23.6)^2/23.6 + (28-14.4)^2/14.4$
$= \mathbf{25.35}$

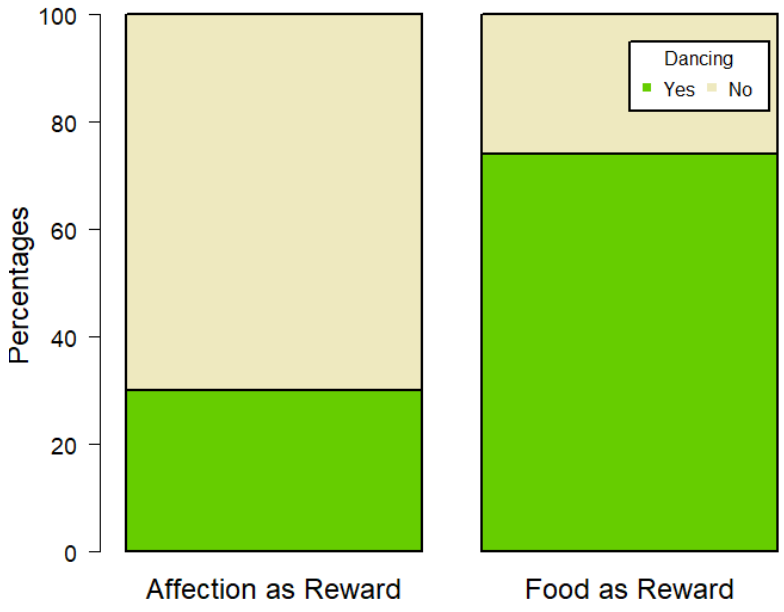
Is 25.35 big enough for the test to be significant?

Chi-square and Fisher's Exact tests

```
> chisq.test(contingency.table)

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

data:  contingency.table
X-squared = 23.52, df = 1, p-value = 1.236e-06
```

- without the correction:

```
> chisq.test(contingency.table, correct=F)

        Pearson's Chi-squared test

data:  contingency.table
X-squared = 25.356, df = 1, p-value = 4.767e-07
```




```
> fisher.test(contingency.table)

        Fisher's Exact Test for Count Data

data:  contingency.table
p-value = 1.312e-06
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
  2.837773 16.429686
sample estimates:
odds ratio
  6.579265
```

Ratio of the odds

Training	Dance No	Dance Yes
Affection as Reward	114	48
Food as Reward	10	28

Odds of dancing
48/114 = affection
28/10 = food

$$\frac{\text{food}}{\text{affection}} = 6.6$$

Answer: Training significantly affects the likelihood of cats line dancing (p=4.8e-07).

Quantitative data

Quantitative data

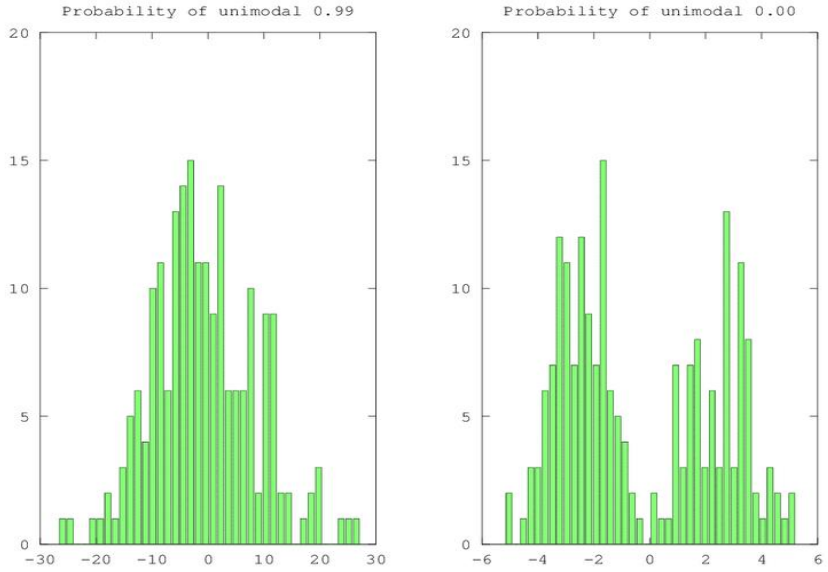
- They take numerical values (units of measurement)

- Discrete: obtained by counting
  - Example: number of students in a class
  - values vary by finite specific steps
- or continuous: obtained by measuring
  - Example: height of students in a class
  - any values

- They can be described by a series of parameters:
  - Mean, variance, standard deviation, standard error and confidence interval

Measures of central tendency
Mode and Median

- **Mode:** most commonly occurring value in a distribution



- **Median:** value exactly in the middle of an ordered set of numbers

Example 1: 18 27 34 52 54 59 61 68 78 82 85 87 91 93 100, Median = 68
Example 2: 18 27 27 34 52 52 59 61 68 68 85 85 85 90, Median = 60

Measures of central tendency
Mean

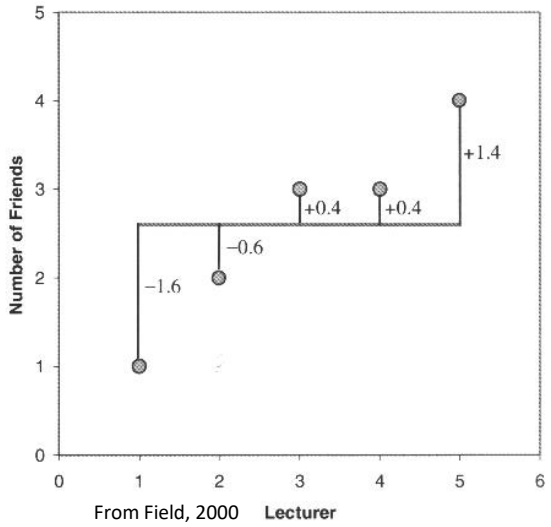
- Definition: average of all values in a column

- It can be considered as a model because it summaries the data
  - Example: a group of 5 lecturers: number of friends of each members of the group: 1, 2, 3, 3 and 4
    - Mean: (1+2+3+3+4)/5 = 2.6 friends per person
      - Clearly an hypothetical value

- How can we know that it is an accurate model?
  - Difference between the real data and the model created

Measures of dispersion

- Calculate the magnitude of the differences between each data and the mean:



From Field, 2000

- Total error = sum of differences

$$= 0 = \Sigma(x_i - \overline{x}) = (-1.6)+(-0.6)+(0.4)+(1.4) = 0$$

No errors !

- Positive and negative: they cancel each other out.

Sum of Squared errors (SS)

- To avoid the problem of the direction of the errors: we square them
  - Instead of sum of errors: sum of squared errors (SS):

$$(SS) = \Sigma(x_i - \overline{x})(x_i - \overline{x})$$

$$= (1.6)^2 + (-0.6)^2 + (0.4)^2 + (0.4)^2 + (1.4)^2$$

$$= 2.56 + 0.36 + 0.16 + 0.16 + 1.96$$

$$= 5.20$$

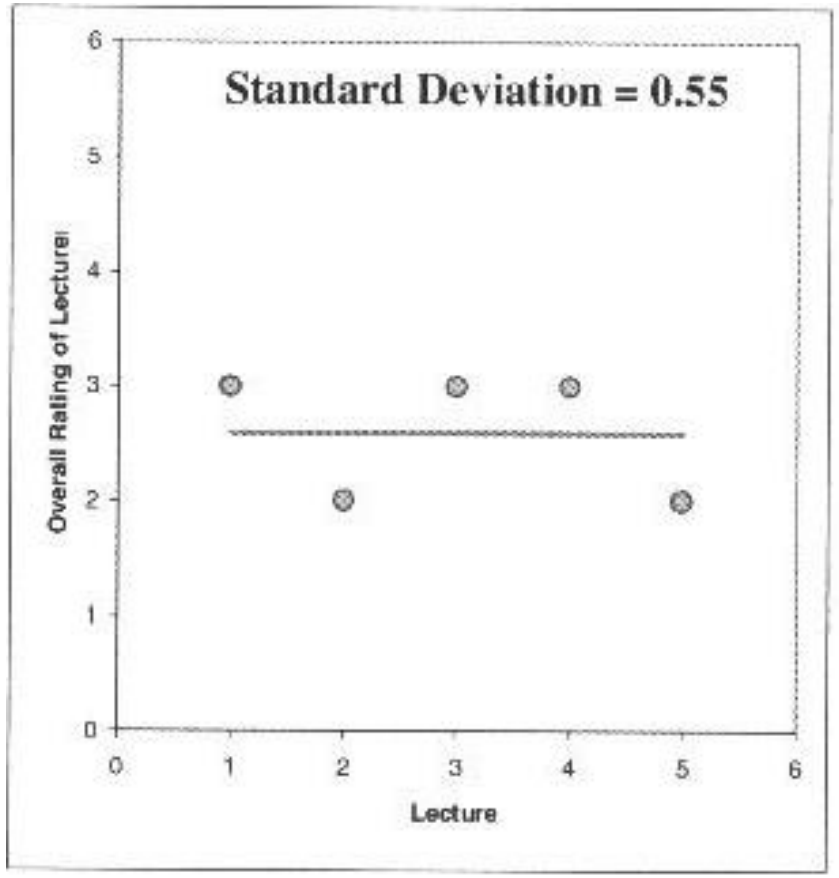
- SS gives a good measure of the accuracy of the model
  - But: dependent upon the amount of data: the more data, the higher the SS.
  - <u>Solution</u>: to divide the SS by the number of observations (N)
    - As we are interested in measuring the error in the sample to estimate the one in the population we divide the SS by N-1 instead of N and we get the **variance** (S^2) = SS/N-1

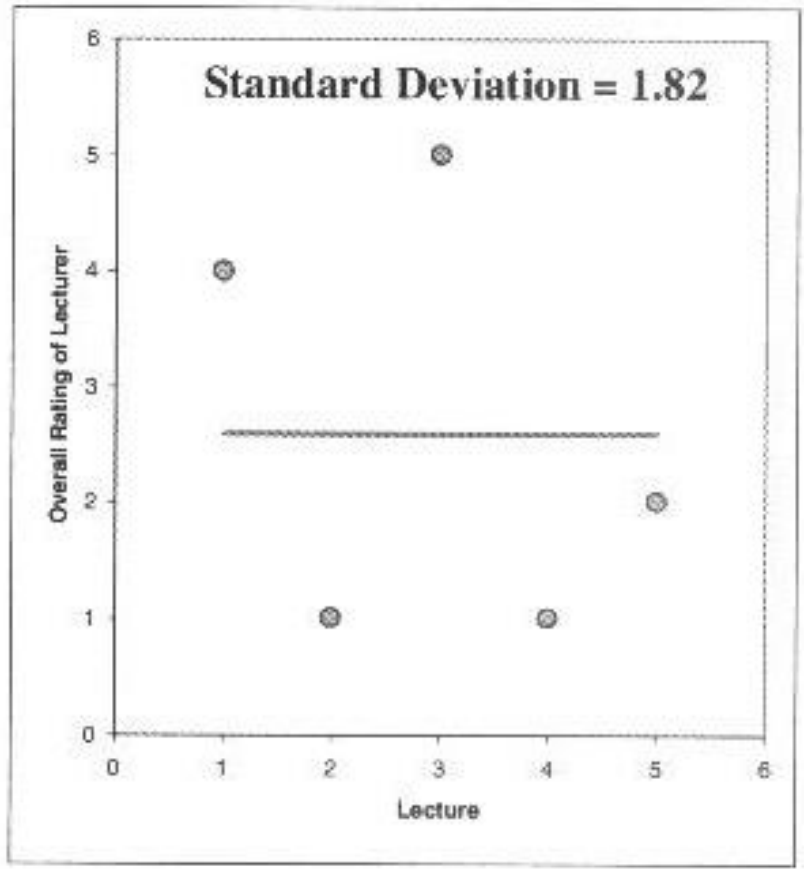
Variance and standard deviation

- $variance\ (s^2) = \frac{SS}{N-1} = \frac{\Sigma\,(x_i - \overline{x})^2}{N-1} = \frac{5.20}{4} = 1.3$

- Problem with variance: measure in squared units
  - For more convenience, the square root of the variance is taken to obtain a measure in the same unit as the original measure:
    - the **standard deviation**
      - S.D. = √(SS/N-1) = √(s^2) = s = $\sqrt{1.3} = 1.14$
    - The **standard deviation** is a measure of how well the mean represents the data.

Standard deviation





Small S.D.:
data close to the mean:
mean is a good fit of the data

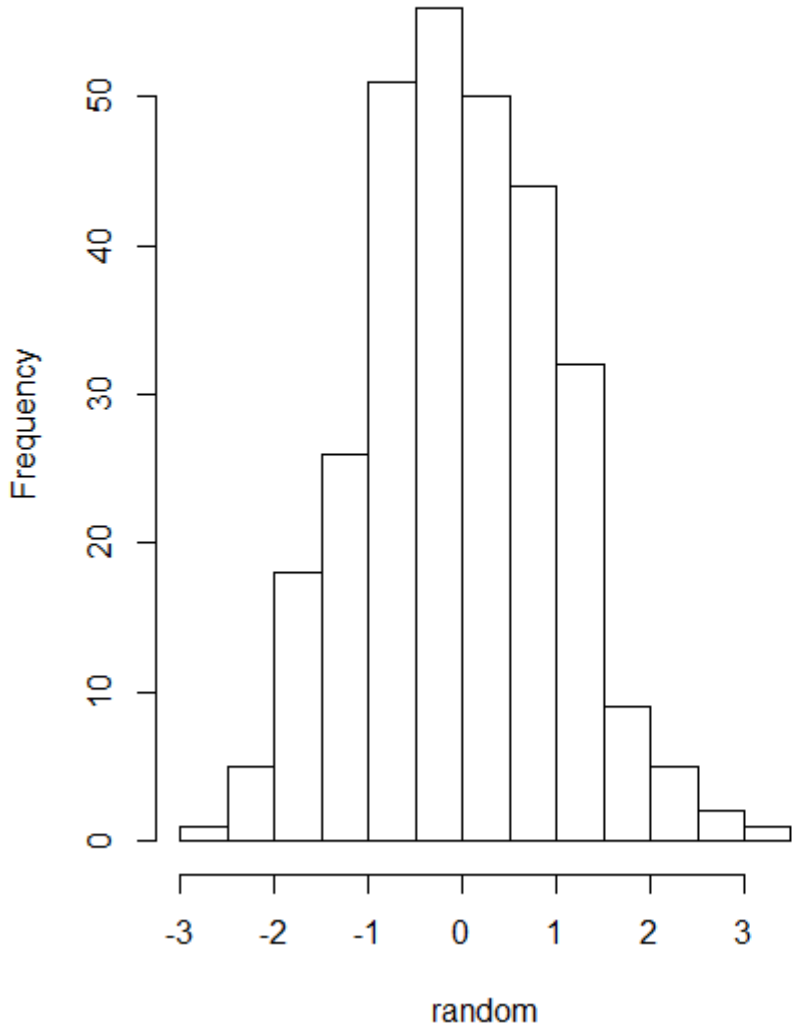
Large S.D.:
data distant from the mean:
mean is not an accurate representation

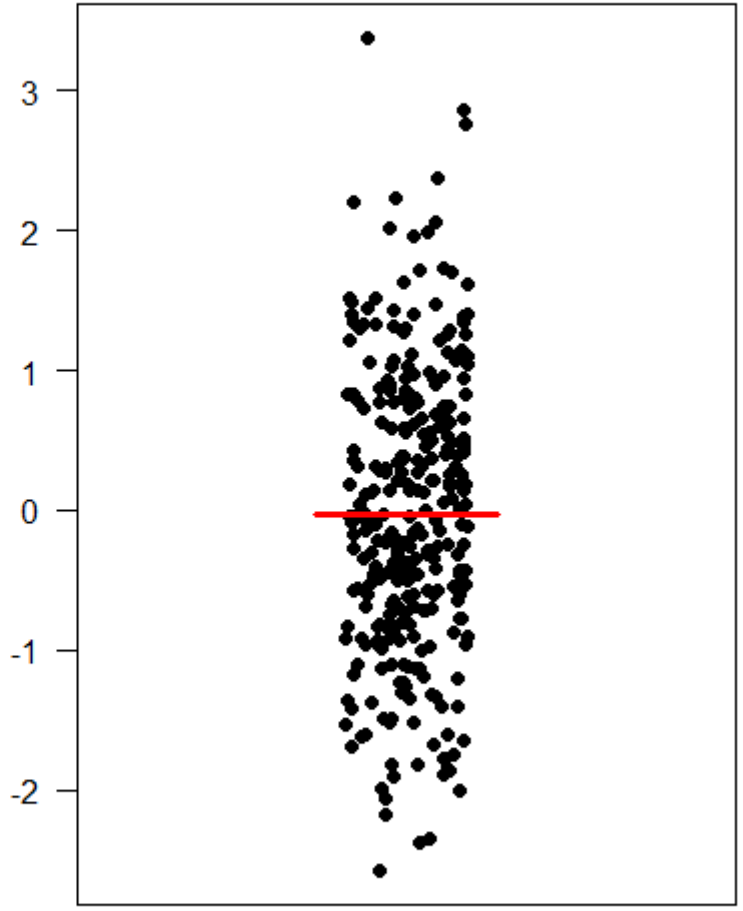
SD and SEM (SEM = SD/√N)

- What are they about?
  - The **SD** quantifies **how much the values vary** from one another: **scatter or spread**
    - The SD does not change predictably as you acquire more data.
  - The **SEM** quantifies **how accurately** you know the **true mean** of the population.
    - Why? Because it takes into account: **SD + sample size**
    - The SEM gets smaller as your sample gets larger
      - Why? Because the mean of a large sample is likely to be closer to the true mean than is the mean of a small sample.

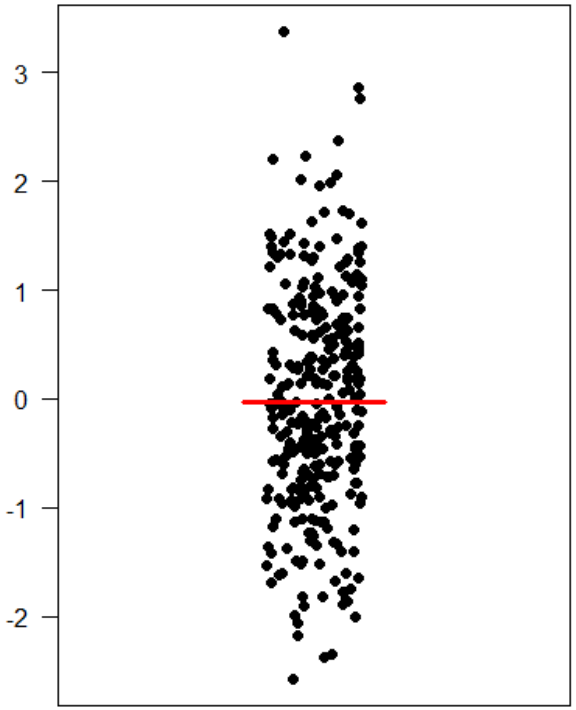
The SEM and the sample size

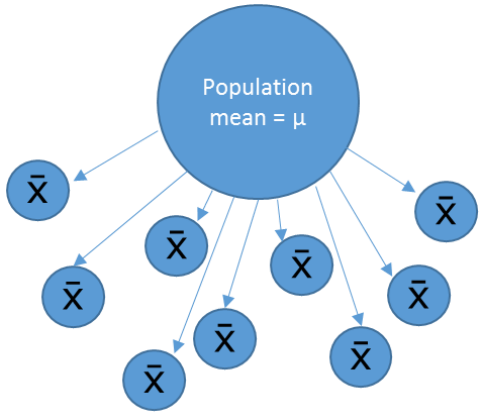
A population





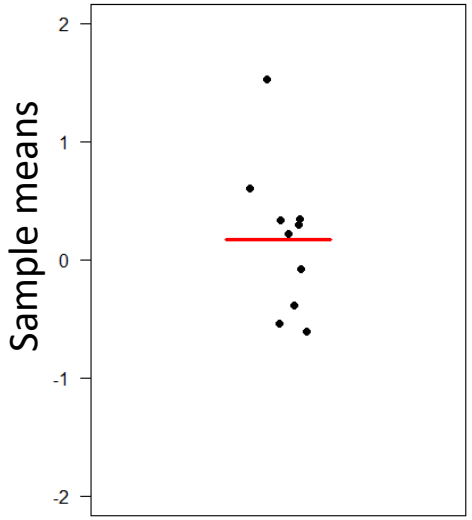
The SEM and the sample size



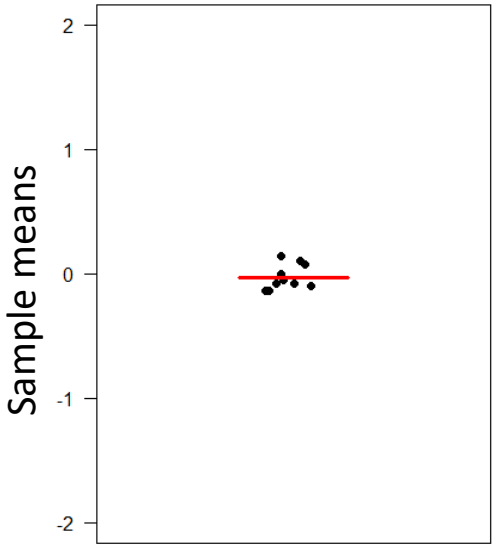


'Infinite' number of samples
Samples means = $\bar{X}$

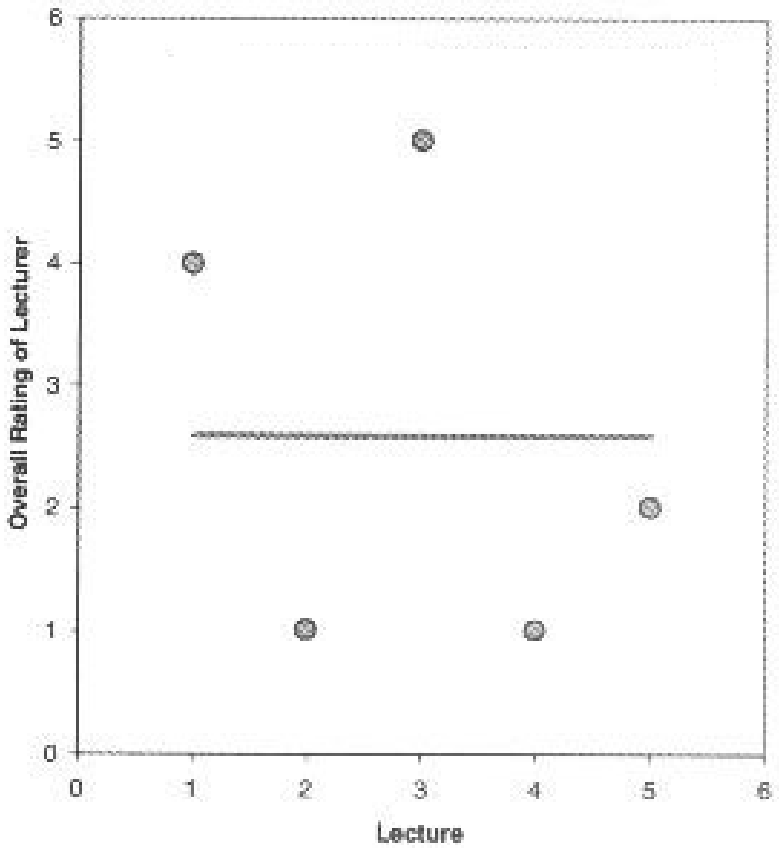
Small samples (n=3)



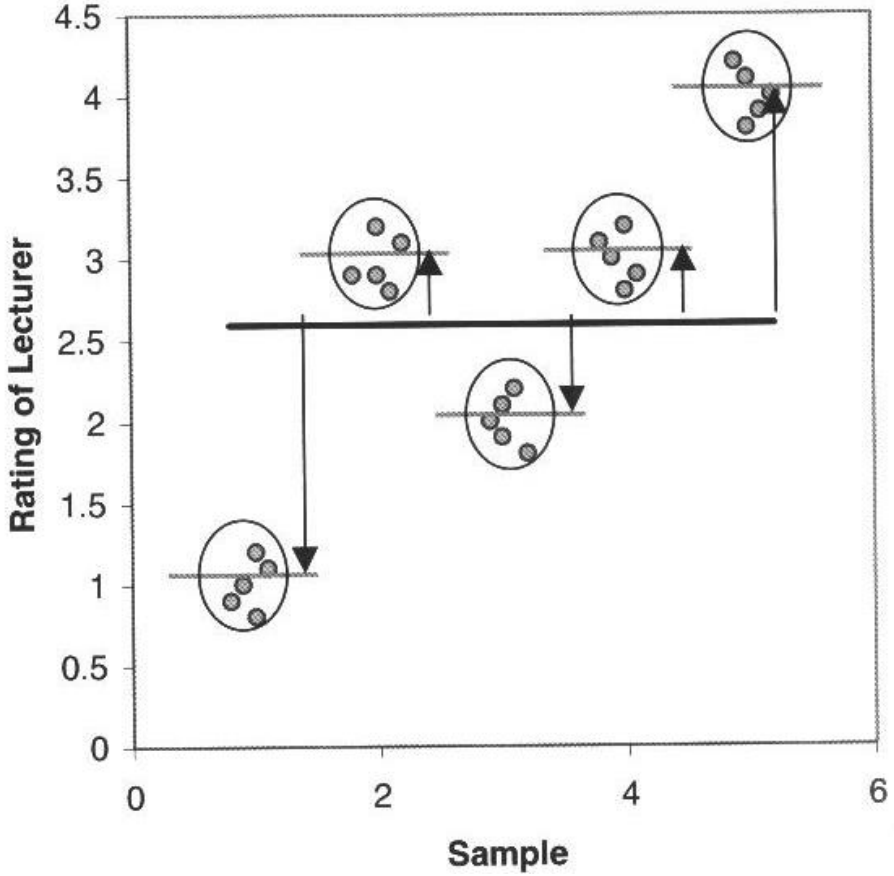
Big samples (n=30)



SD and SEM



The SD quantifies the scatter of the data.



The SEM quantifies the distribution of the sample means.

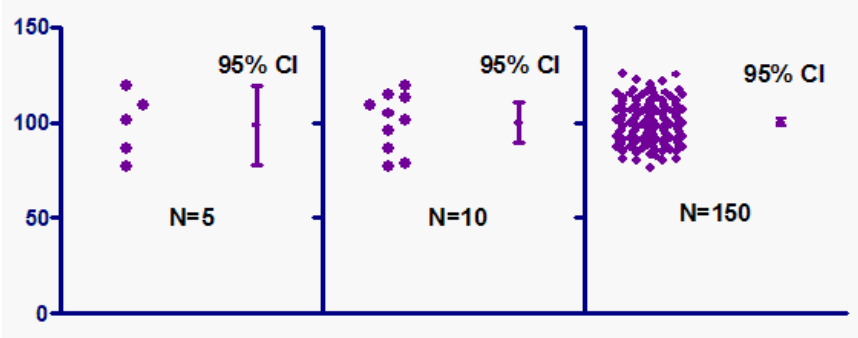
SD or SEM ?

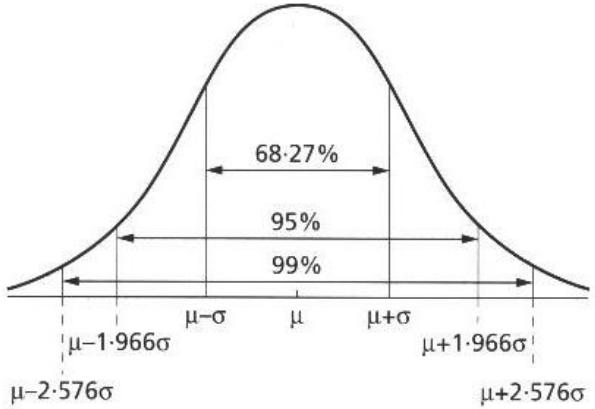
- If the scatter is caused by biological variability, it is important to show the variation.
  - Report the SD rather than the SEM.
    - Better even: show a graph of all data points.

- If you are using an in vitro system with no biological variability, the scatter is about experimental imprecision (no biological meaning).
  - Report the SEM to show how well you have determined the mean.

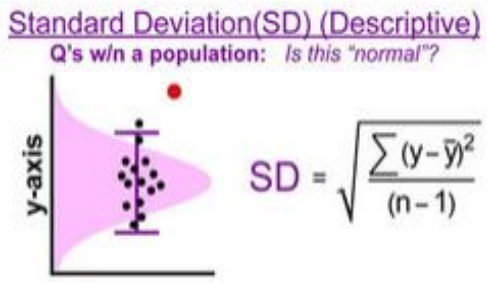
Confidence interval

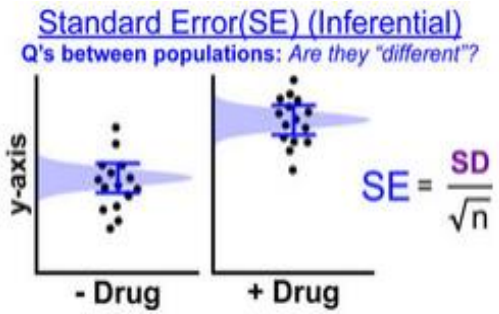
- Range of values that we can be 95% confident contains the true mean of the population.
  - So limits of 95% CI: **[Mean - 1.96 SEM; Mean + 1.96 SEM]** (SEM = SD/√N)





Error bars	Type	Description
Standard deviation	Descriptive	Typical or average difference between the data points and their mean.
Standard error	Inferential	A measure of how variable the mean will be, if you repeat the whole study many times.
Confidence interval usually 95% CI	Inferential	A range of values you can be 95% confident contains the true mean.





Analysis of Quantitative Data

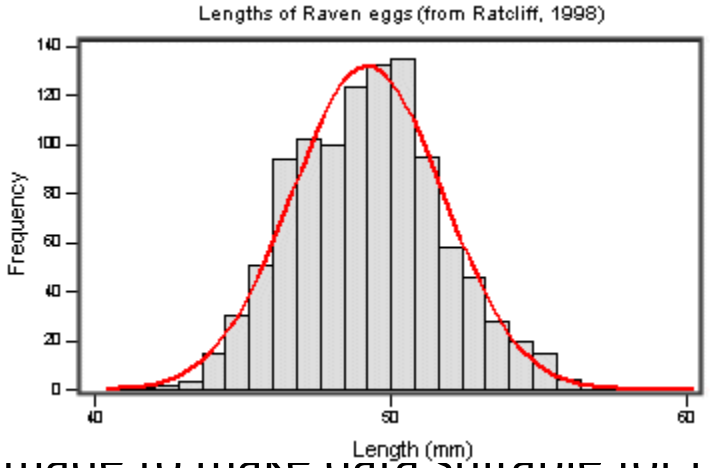
- Choose the correct statistical test to answer your question:
  - They are 2 types of statistical tests:
    - **Parametric tests** with 4 assumptions to be met by the data,
    - **Non-parametric tests** with no or few assumptions (e.g. Mann-Whitney test) and/or for qualitative data (e.g. Fisher's exact and χ^2 tests).

Assumptions of Parametric Data

- All parametric tests have 4 basic assumptions that must be met for the test to be accurate.

1) <u>Normally distributed data</u>

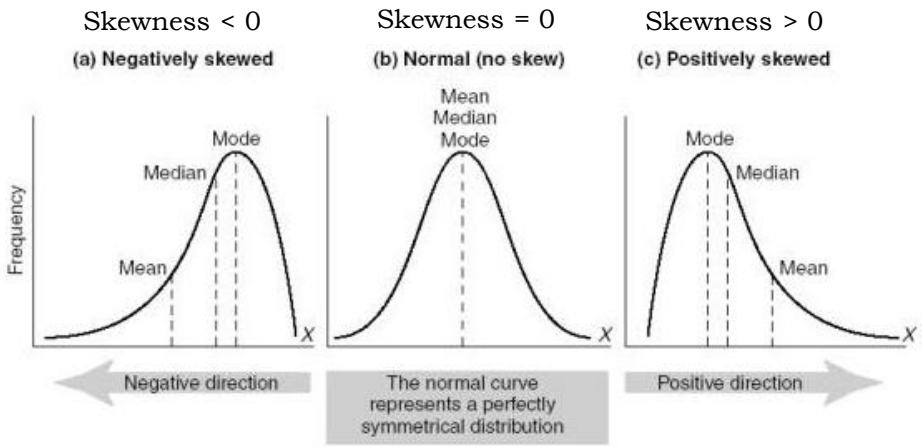
- Normal shape, bell shape, Gaussian shape



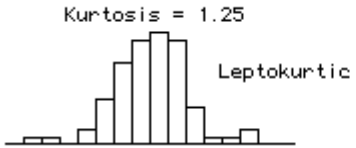
- Transformations can be made to make data suitable for parametric analysis.

Assumptions of Parametric Data

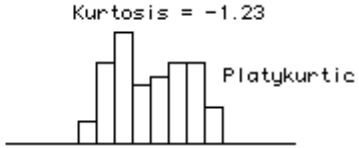
- Frequent departures from normality:
  - Skewness: lack of symmetry of a distribution



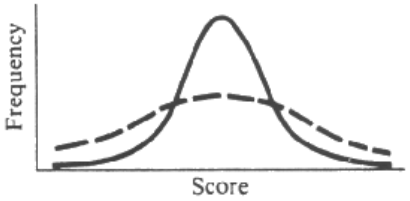
  - Kurtosis: measure of the degree of 'peakedness' in the distribution
    - The two distributions below have the same variance approximately the same skew, but differ markedly in kurtosis.



More peaked distribution: kurtosis > 0



Flatter distribution: kurtosis < 0



(e) Platykurtic and leptokurtic

Assumptions of Parametric Data

2) ***Homogeneity in variance***

- The variance should not change systematically throughout the data

3) ***Interval data (linearity)***

- The distance between points of the scale should be equal at all parts along the scale.

4) ***Independence***

- Data from different subjects are independent
  - Values corresponding to one subject do not influence the values corresponding to another subject.
  - Important in repeated measures experiments

Analysis of Quantitative Data

- **Is there a difference between my groups regarding the variable I am measuring?**
  - e.g. are the mice in the group A heavier than those in group B?
    - Tests with 2 groups:
      - Parametric: **Student's *t*-test**
      - Non parametric: Mann-Whitney/Wilcoxon rank sum test
    - Tests with more than 2 groups:
      - Parametric: **Analysis of variance (one-way ANOVA)**
      - Non parametric: Kruskal Wallis
- **Is there a relationship between my 2 (continuous) variables?**
  - e.g. is there a relationship between the daily intake in calories and an increase in body weight?
    - Test: **Correlation** (parametric) and **curve fitting**

Sample → Statistical inference → Population

Difference → Meaningful? → Yes → Real?

Real? → Statistical test → Statistic (e.g. t, F …) → Big enough?

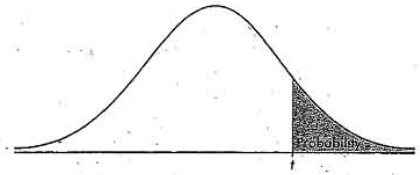


TABLE B: t-DISTRIBUTION CRITICAL VALUES

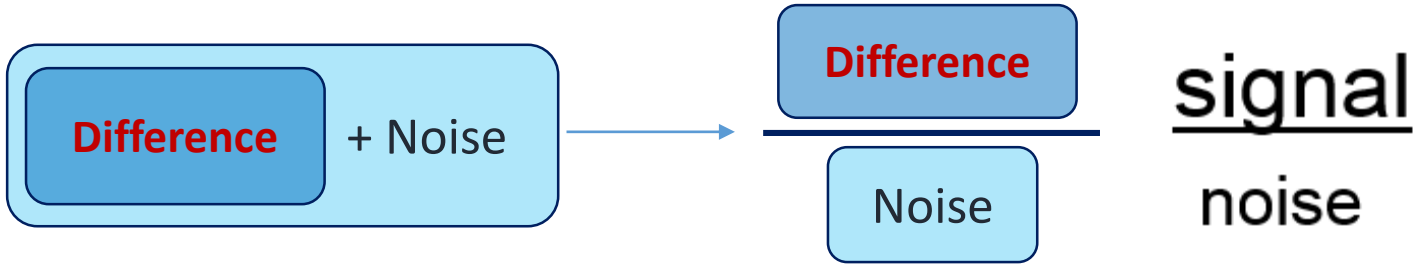
	Tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	.765	.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	.741	.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	.727	.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	.718	.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	.711	.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	.706	.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	.703	.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	.700	.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	.697	.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	.695	.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	.694	.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	.692	.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	.691	.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	.690	.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	.689	.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	.688	.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	.688	.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	.687	.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	.686	.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	.686	.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792

=

Difference + Noise + Sample

Signal-to-noise ratio

- Stats are all about understanding and controlling variation.



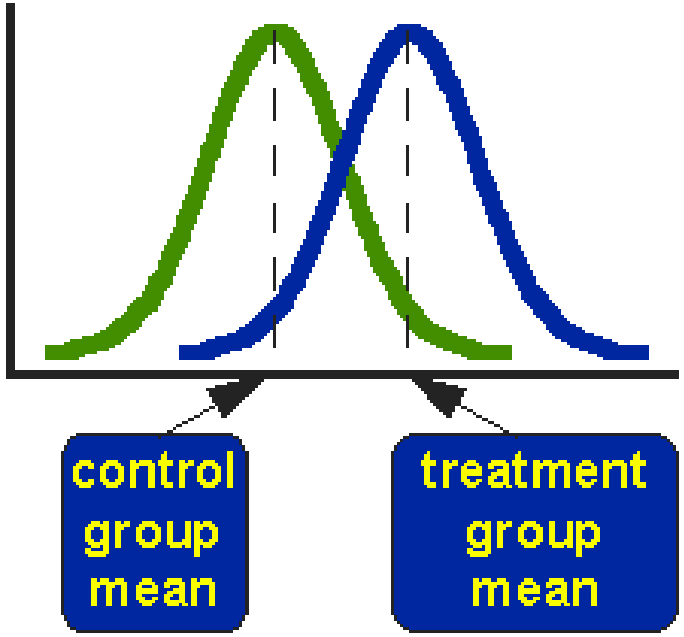
$\frac{\text{signal}}{\text{noise}}$ If the **noise is low** then the **signal is detectable** …
= statistical significance

$\frac{\text{signal}}{\text{noise}}$ … but if the **noise** (i.e. interindividual variation) **is large** then the **same signal will not be detected**
= no statistical significance

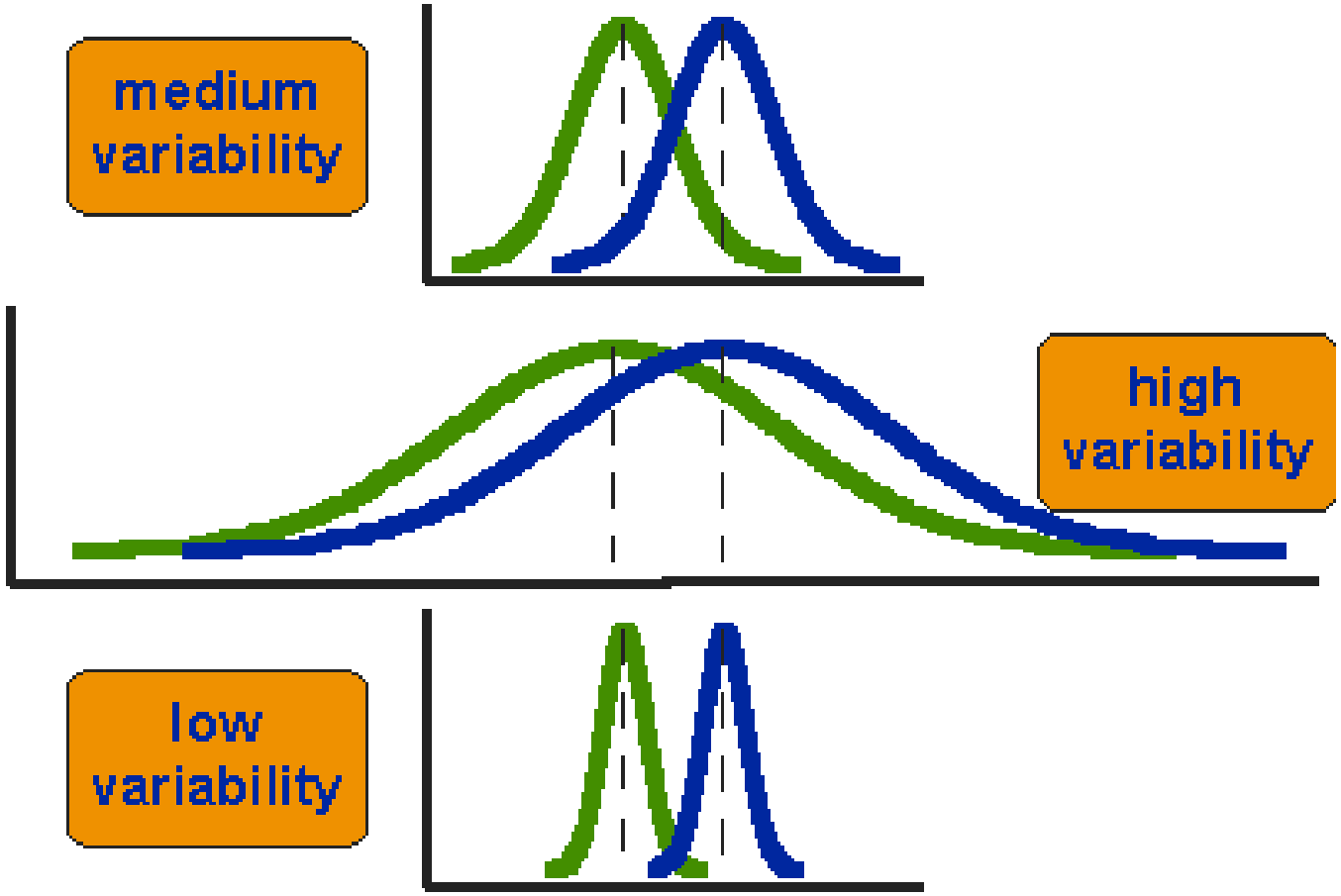
- In a statistical test, the ratio of signal to noise determines the significance.

Comparison between 2 groups:
Student's *t*-test

- **Basic idea**:
  - When we are looking at the differences between scores for 2 groups, we have to judge the difference between their means relative to the spread or variability of their scores.
    - Eg: comparison of 2 groups: control and treatment



Student's *t*-test

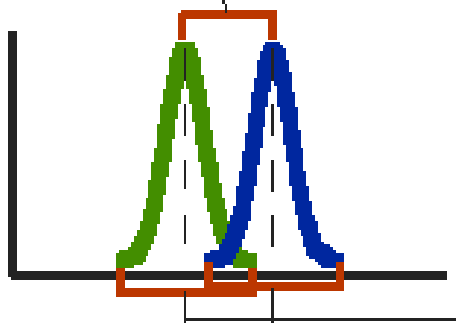


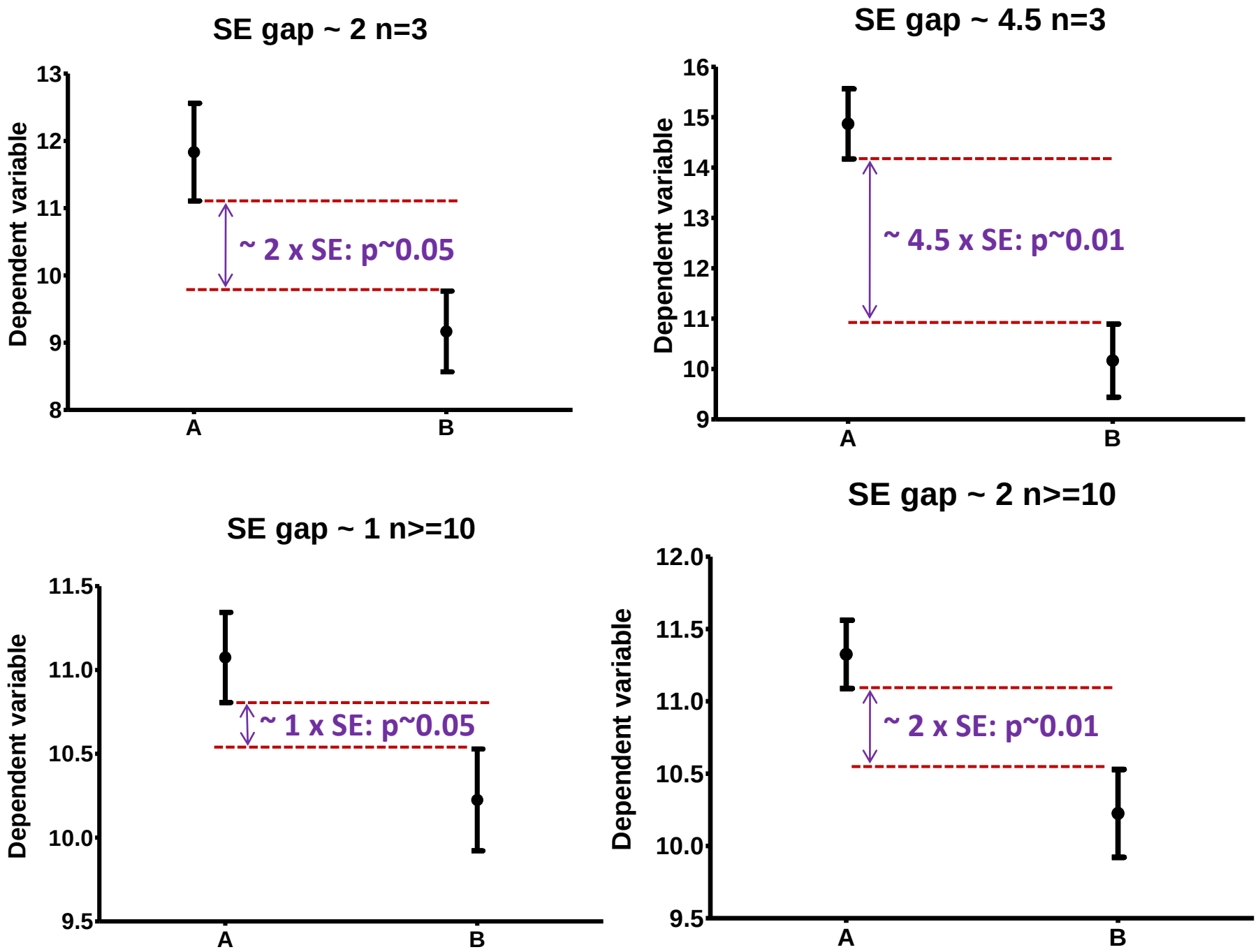
Student's *t*-test

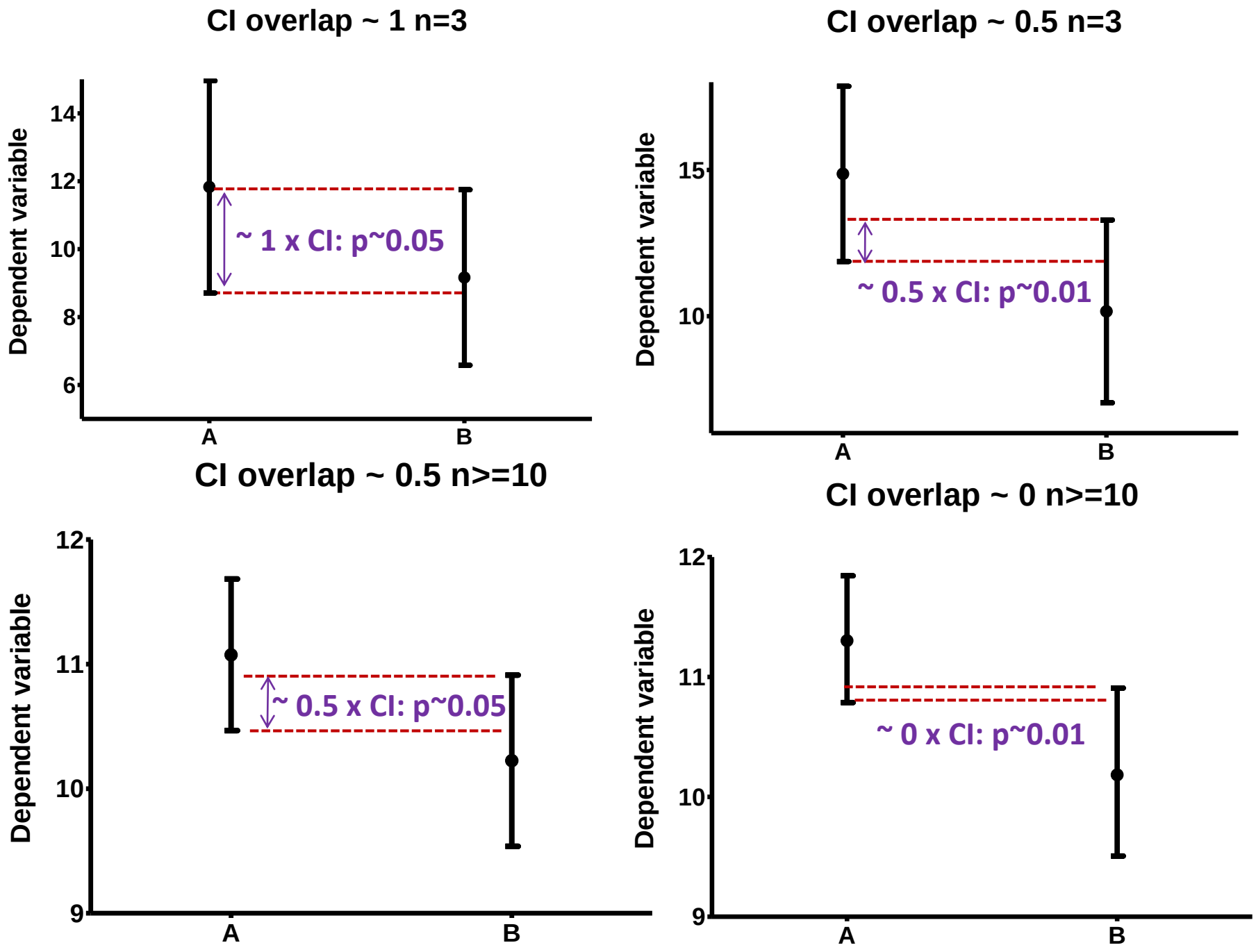
$$\frac{\text{signal}}{\text{noise}} = \frac{\text{difference between group means}}{\text{variability of groups}}$$

$$= \frac{\bar{x}_T - \bar{x}_C}{\sqrt{\frac{var_T}{n_T} + \frac{var_C}{n_C}}}$$

$$= \text{t-value}$$




SE gap ~ 2 n=3
Dependent variable
13
12
11
10
9
8
A
B
~ 2 x SE: p~0.05
SE gap ~ 4.5 n=3
Dependent variable
16
15
14
13
12
11
10
9
A
B
~ 4.5 x SE: p~0.01
SE gap ~ 1 n>=10
Dependent variable
11.5
11.0
10.5
10.0
9.5
A
B
~ 1 x SE: p~0.05
SE gap ~ 2 n>=10
Dependent variable
12.0
11.5
11.0
10.5
10.0
9.5
A
B
~ 2 x SE: p~0.01


CI overlap ~ 1 n=3
Dependent variable
14
12
10
8
6
~ 1 x CI: p~0.05
A
B
CI overlap ~ 0.5 n=3
Dependent variable
15
10
~ 0.5 x CI: p~0.01
A
B
CI overlap ~ 0.5 n>=10
Dependent variable
12
11
10
9
~ 0.5 x CI: p~0.05
A
B
CI overlap ~ 0 n>=10
Dependent variable
12
11
10
9
~ 0 x CI: p~0.01
A
B

Student's *t*-test

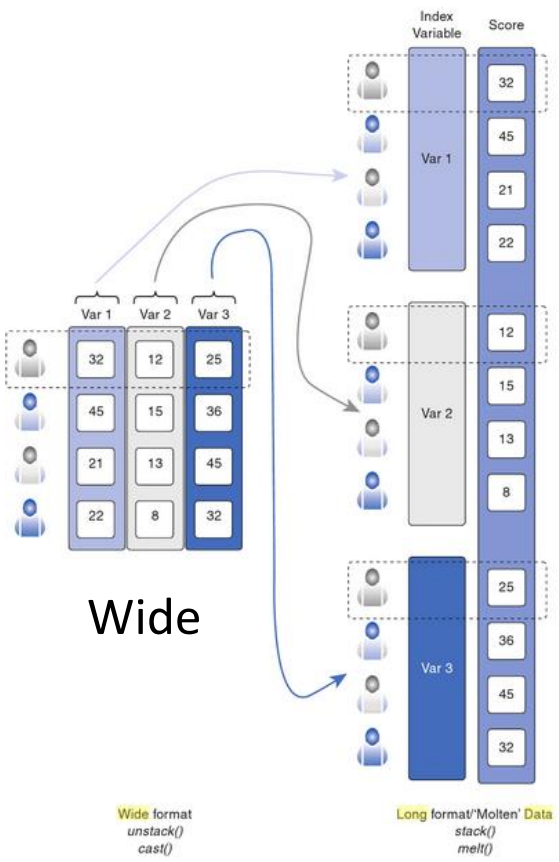
- <u>3 types</u>:
  - **Independent t-test**
    - compares means for two independent groups of cases.
  - **Paired t-test**
    - looks at the difference between two variables for a single group:
      - the second 'sample' of values comes from the same subjects (mouse, petri dish …).
  - One-Sample t-test
    - tests whether the mean of a single variable differs from a specified constant (often 0)

Before going any further

- **Data format**: `melt()` wide vs long (molten) format
- **Some extra R**:
  - `tapply()`
  - `par(mfrow)`
  - `y~x`

Data file format

- Wide vs long (molten) format



Wide

Long

Predictor

Outcome

cond A	cond B
5	2
8	5
9	0
4	2
3	3

condition	measure
A	5
A	8
A	9
A	4
A	3
B	2
B	5
B	0
B	2
B	3

In R: `melt()` ## reshape2 package ##

Extra R: `tapply()`

- Want to compute summaries of variables? **`tapply()`**
  - break up a vector into groups defined by some classifying factor,
  - compute a function on the subsets,
  - and return the results in a convenient form.

- **`tapply`**`(data,groups,function)`

```
tapply(some.data$measure, some.data$condition, mean)
```

```
  cond.A cond.B
     5.8    2.4
```

Some.data

Condition	Measure
Cond.A	5
Cond.A	8
Cond.A	9
Cond.A	4
Cond.A	3
Cond.B	2
Cond.B	5
Cond.B	0
Cond.B	2
Cond.B	3

(Long format)

Extra R: `par(mfrow)`

- Want to create a multi-paneled plotting window? **`par(mfrow)`**
  - Rather **`par(mfrow=c(row,col))`**
  - Will plot a window with x rows and y columns

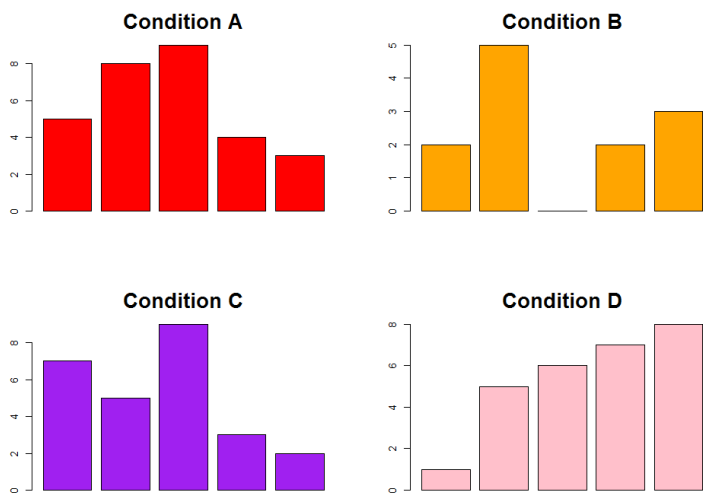
- We want to plot conditions A, B, C and D on the same panel

```
par(mfrow=c(2,2))
```
so that's 2 row and 2 columns

```
barplot(some.data$cond.A, main = "Condition A", col="red")
barplot(some.data$cond.B, main = "Condition B", col="orange")
barplot(some.data$cond.C, main = "Condition C", col="purple")
barplot(some.data$cond.D, main = "Condition D", col="pink")
dev.off()
```

Some.data

Cond A	Cond B	Cond C	Cond D
5	2	7	1
8	5	5	5
9	0	9	6
4	2	3	7
3	3	2	8

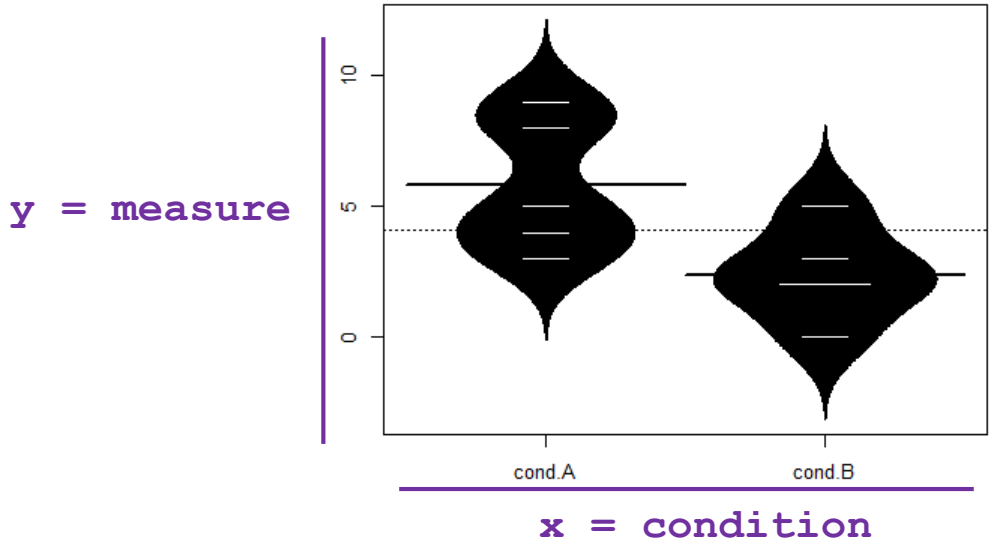


Extra R: `y~x`

- Want to plot and do stats on long-format file? **`y~x`**
  - break up a vector into groups defined by some classifying factor,
  - compute a function on thesubsets
  - creates a functional link between x and y, a **model**
  - does what **`tapply`** does but in different context.

- **`function(y~x)`** : y explained/predicted by x, y=f(x)

```
beanplot(some.data$measure~some.data$condition)
```



Some.data

Condition	Measure
Cond.A	5
Cond.A	8
Cond.A	9
Cond.A	4
Cond.A	3
Cond.B	2
Cond.B	5
Cond.B	0
Cond.B	2
Cond.B	3

Example: coyote.csv

- Question: do male and female coyotes differ in size?

- **Sample size**
- **Data exploration**
- **Check the assumptions for parametric test**
- **Statistical analysis: Independent t-test**

Power analysis

No data from a pilot study but we have found some information in the literature.

In a study run in similar conditions as in the one we intend to run, **male coyotes** were found to measure: **92cm+/- 7cm (SD**).

We expect a **5% difference** between genders.

- **smallest biologically meaningful difference**

```
power.t.test(n = NULL, delta = NULL, sd = 1, sig.level = NULL, power = NULL,
type = c("two.sample", "one.sample", "paired"),alternative = c("two.sided","one.sided"))
```

Power analysis

Independent t-test

A priori Power analysis

Example case:

We don't have data from a pilot study but we have found some information in the literature.

In a study run in similar conditions as in the one we intend to run, **male coyotes** were found to measure:
92cm+/- 7cm (SD)

We expect a **5% difference** between genders with a similar variability in the female sample.

```
power.t.test(n = NULL, delta = NULL, sd = 1, sig.level = NULL,
power = NULL, type = c("two.sample", "one.sample", "paired"),
alternative = c("two.sided", "one.sided"))
```

Mean 1 = 92
Mean 2 = 87.4 (5% less than 92cm)

delta = 92 – 87.4
sd = 7

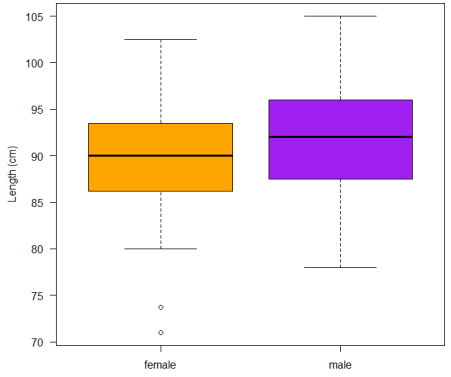
```
power.t.test(delta=92-87.4, sd = 7,
sig.level = 0.05, power = 0.8)
```

```
     Two-sample t test power calculation

              n = 37.33624
          delta = 4.6
             sd = 7
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

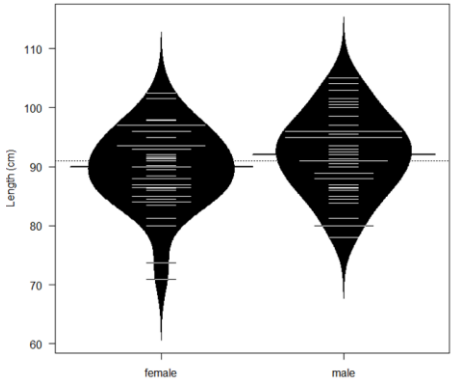
We need a sample size of **n~76 (2*38)**

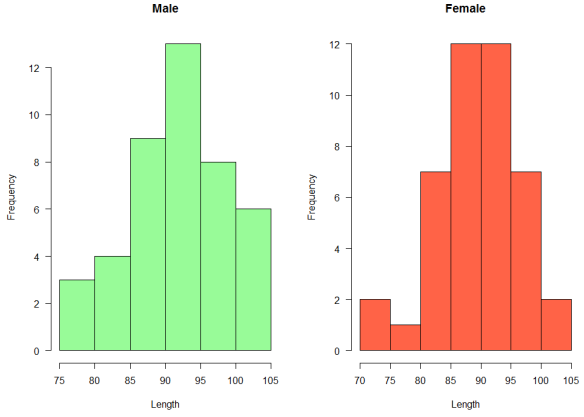
Data exploration ≠ plotting data

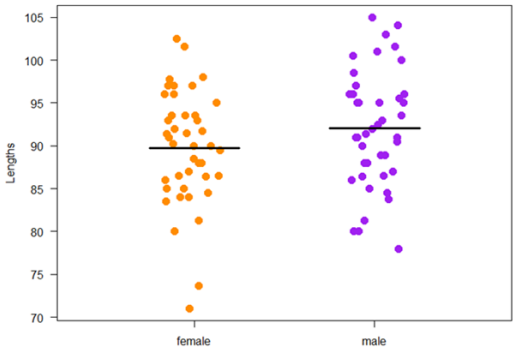
- Download: **coyote.csv**
- Explore data using 4 different representations: **boxplot**, **histogram**, **beanplot** and **stripchart**



`function(y~x)`

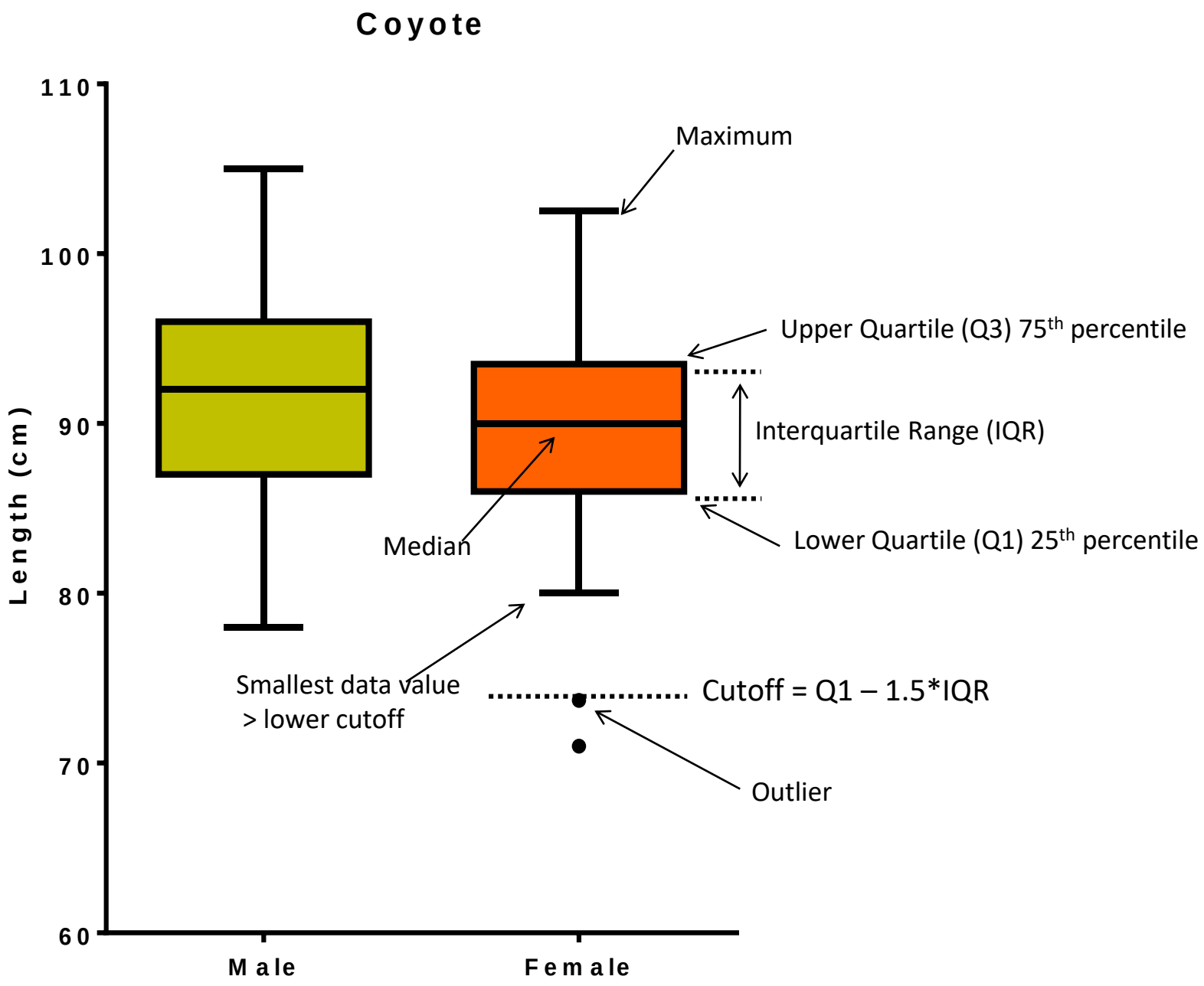


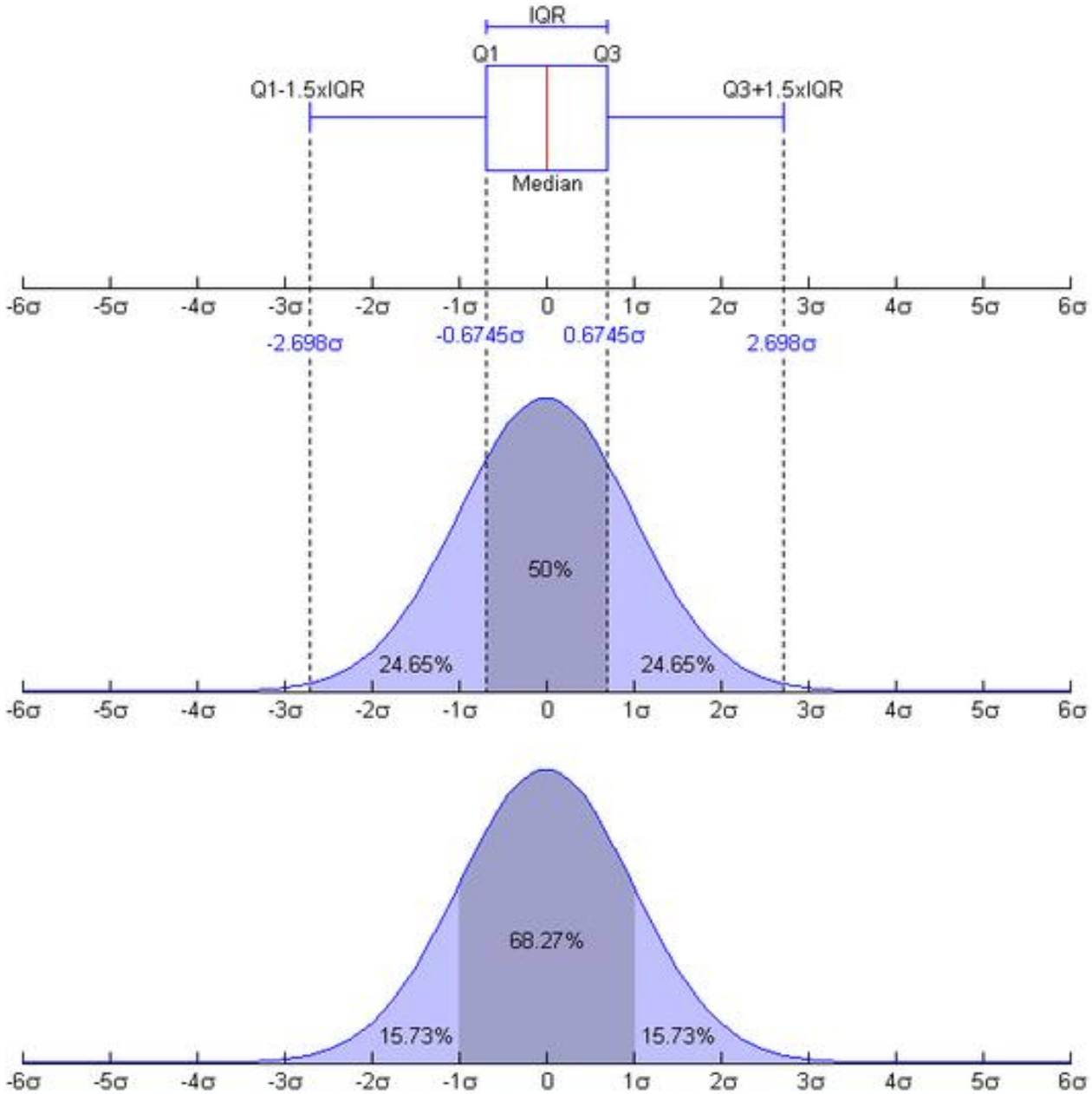




`tapply()`
`segment()`

```
par(mfrow=c(?,?))
coyote[only female]$length
coyote[only male]$length
```


Coyote
110
100
90
80
70
60
Length (cm)
Male
Female
Maximum
Upper Quartile (Q3) 75th percentile
Interquartile Range (IQR)
Lower Quartile (Q1) 25th percentile
Median
Smallest data value
> lower cutoff
Cutoff = Q1 – 1.5*IQR
Outlier

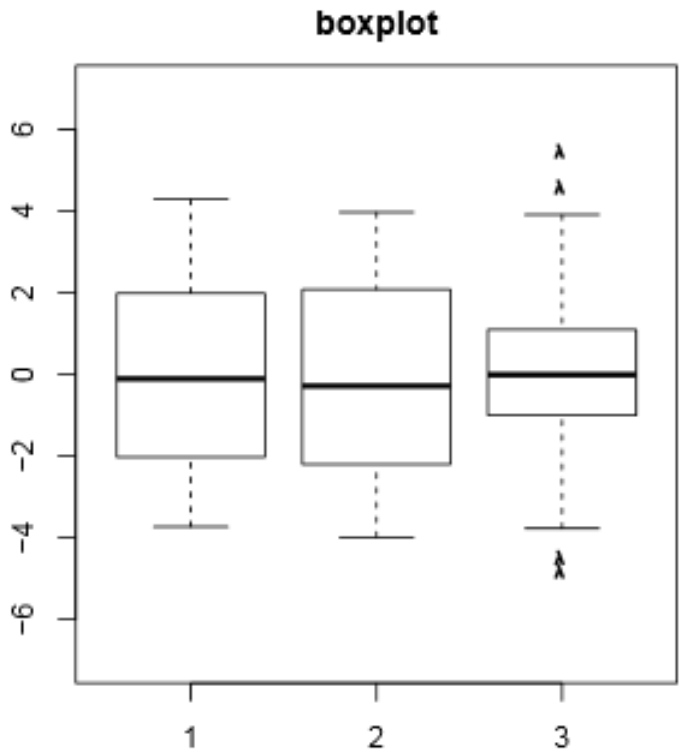

IQR
Q1
Q3
Q1-1.5xIQR
Q3+1.5xIQR
Median
-6σ -5σ -4σ -3σ -2σ -1σ 0 1σ 2σ 3σ 4σ 5σ 6σ
-2.698σ
-0.6745σ
0.6745σ
2.698σ
50%
24.65%
24.65%
-6σ -5σ -4σ -3σ -2σ -1σ 0 1σ 2σ 3σ 4σ 5σ 6σ
68.27%
15.73%
15.73%
-6σ -5σ -4σ -3σ -2σ -1σ 0 1σ 2σ 3σ 4σ 5σ 6σ

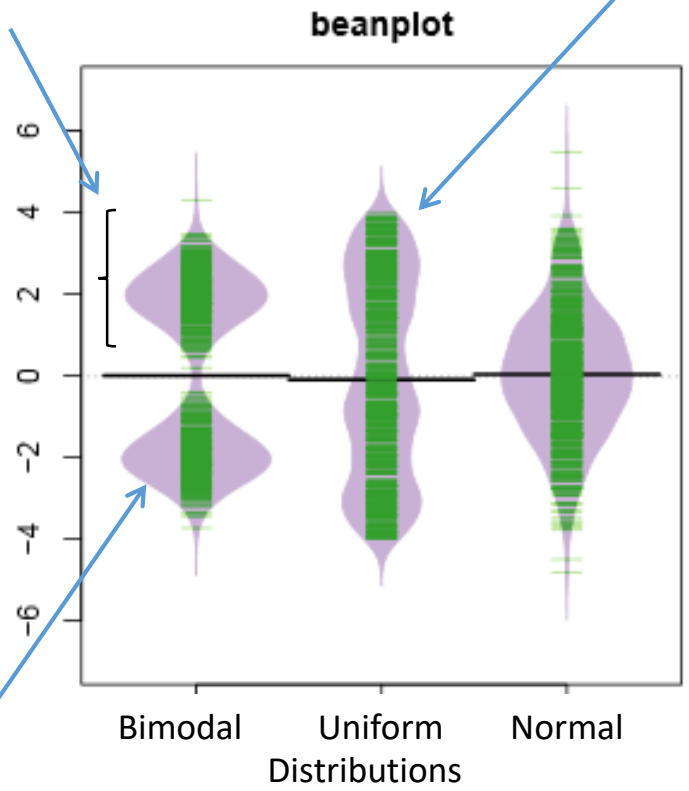
Exploring data: quantitative data

Boxplots <u>or</u> beanplots

Scatterplot shows individual data

A bean= a 'batch' of data

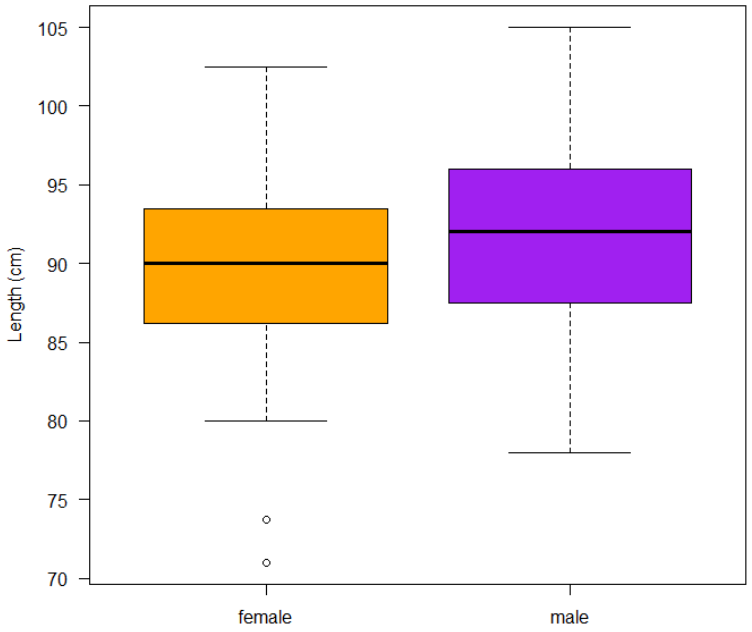


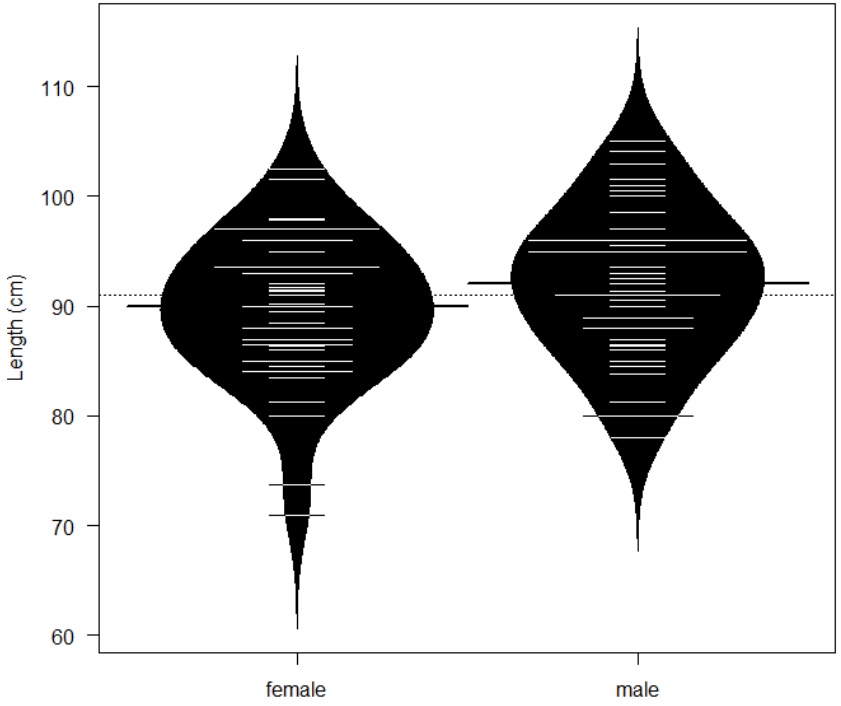


Data density mirrored by the shape of the polygon

Boxplots and beanplots

```
boxplot(coyote$length~coyote$gender,
        col=c("orange","purple"),
        las=1,
        ylab="Length (cm)")
```

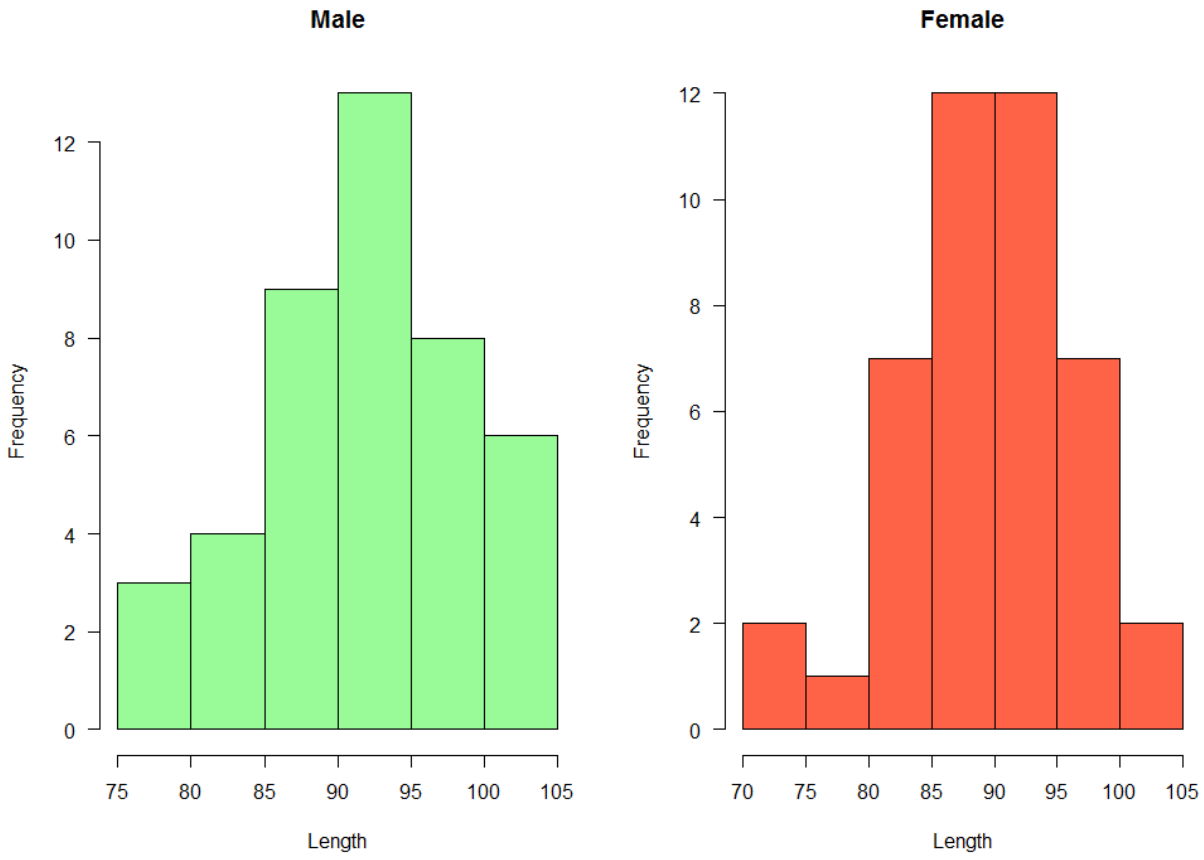





```
beanplot(coyote$length~coyote$gender,
        las=1,
        ylab="Length (cm)")
## beanplot package ##
```

Histograms

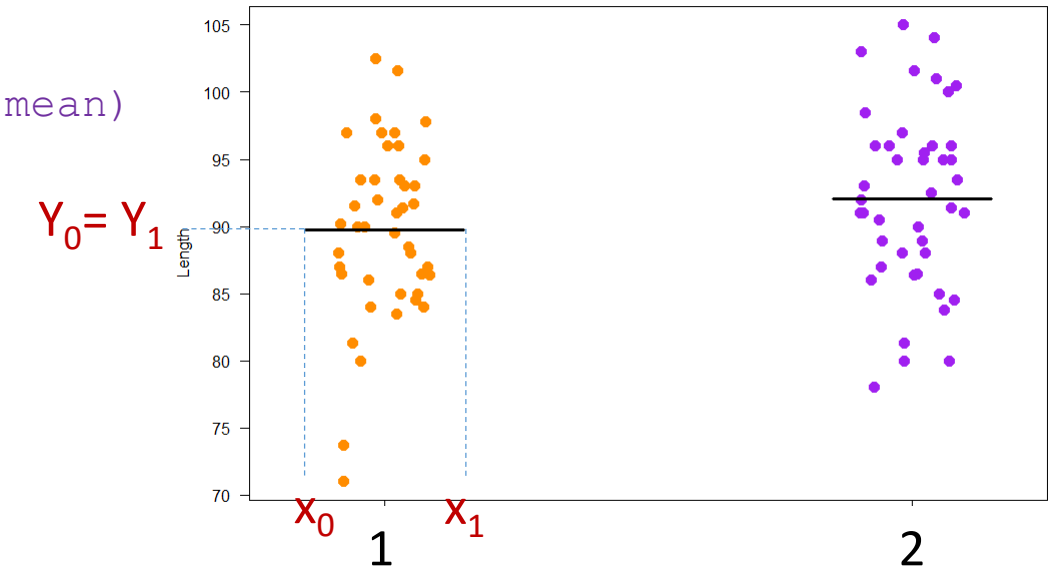
```
par(mfrow=c(1,2))
hist(coyote[coyote$gender=="male",]$length, main="Male", xlab="Length", col="lightgreen", las=1)
hist(coyote[coyote$gender=="female",]$length, main="Female", xlab="Length", col="tomato1", las=1)
```



Stripcharts

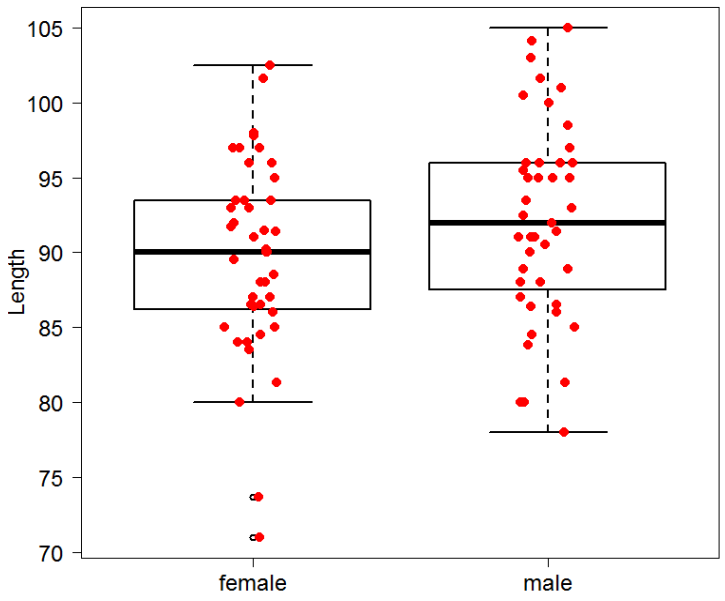
```
stripchart(coyote$length~coyote$gender,
        vertical=TRUE,
        method="jitter",
        las=1,
        ylab="Length",
        pch=16,
        col=c("darkorange","purple"),
        cex=1.5
        )

length.means <- tapply(coyote$length, coyote$gender, mean)

segments(x0, y0, x1, y1)

segments( x0=1:2-0.15,
        y0=length.means,
        x1=1:2+0.15,
        y1=length.means,
        lwd=3
        )
```

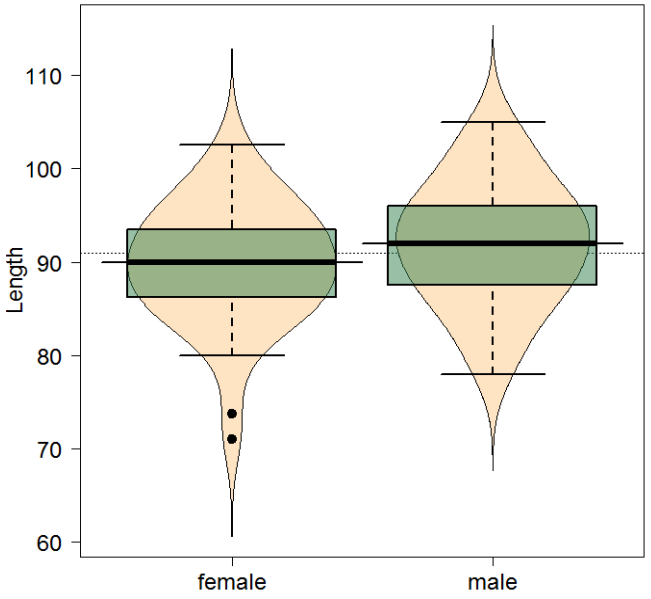
```
> 1:2-0.15
[1] 0.85 1.85
> 1:2+0.15
[1] 1.15 2.15
```



Graphs combinations

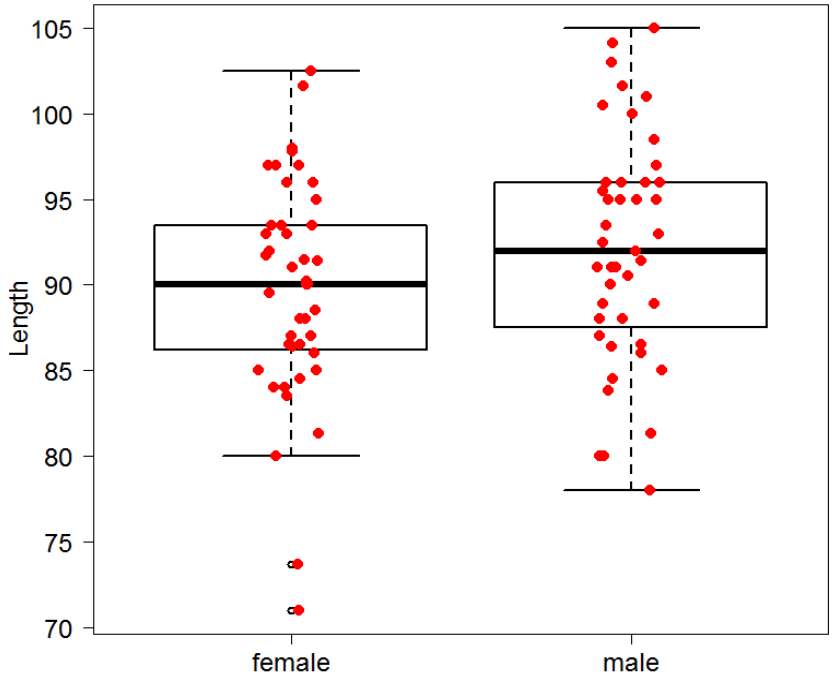
```
boxplot(coyote$length~coyote$gender,
          lwd = 2,
          ylab = "Length",
          cex.axis=1.5,
          las=1,
          cex.lab=1.5)

stripchart(coyote$length~coyote$gender,
          vertical = TRUE,
          method = "jitter",
          pch = 20,
          col = 'red',
          cex=2,
          add = TRUE)
```





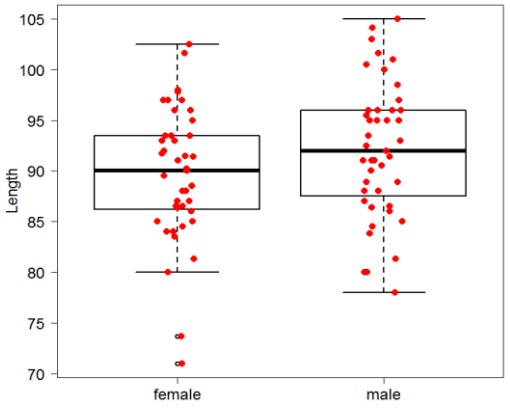

```
beanplot(coyote$length~coyote$gender,
         las=1, overallline = "median",
         ylab = 'Length',
         cex.lab=1.5,
         col="bisque",
         what = c(1, 1, 1, 0),
         cex.axis=1.5)

boxplot(coyote$length~coyote$gender,
        col=rgb(0.2,0.5,0.3, alpha=0.5),
        pch = 20,
        cex=2,
        lwd=2,
        yaxt='n',
        xaxt='n',
        add=TRUE)
```

Assumptions of Parametric Data



- First assumption: <u>Normality</u>
  - ❖ Shapiro-Wilk test `shapiro.test()`
- Second assumption: <u>Homoscedasticity</u>
  - ❖ Bartlett test `bartlett.test()`

Assumptions of Parametric Data



- First assumption: <u>Normality</u>
  - ❖ Shapiro-Wilk test `shapiro.test()`
- Second assumption: <u>Homoscedasticity</u>
  - ❖ Bartlett test `bartlett.test()`

```
tapply(coyote$length,coyote$gender, shapiro.test)
```

```
> tapply(coyote$length,coyote$gender, shapiro.test)
$`female`

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.97001, p-value = 0.3164

$male

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.98446, p-value = 0.819
```

Normality ☑

```
bartlett.test(coyote$length~coyote$gender)
```

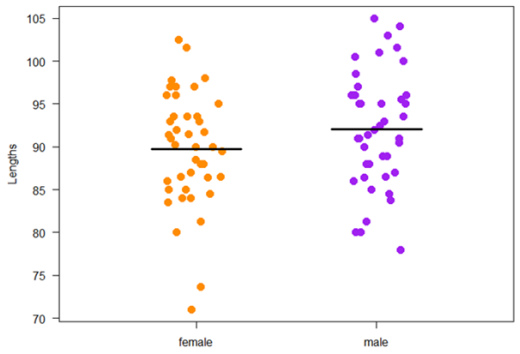
```
> bartlett.test(coyote$length~coyote$gender)

        Bartlett test of homogeneity of variances

data:  coyote$length by coyote$gender
Bartlett's K-squared = 0.02021, df = 1, p-value = 0.887
```

Homogeneity in variance ☑

Independent Student's *t*-test

```
t.test(coyote$length~coyote$gender, var.equal=T)
```

```
 Two Sample t-test

data:  coyote$length by coyote$gender
t = -1.6411, df = 84, p-value = 0.1045
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.184747  0.496375
sample estimates:
mean in group female   mean in group male
            89.71163             92.05581
```



Answer: males coyote are longer than females but not significantly so (p=0.1045).

- How many more coyotes to reach significance?

```
power.t.test(delta=92-89.7, sd = 7, sig.level = 0.05, power = 0.8)
```

```
     Two-sample t test power calculation

              n = 146.3712
          delta = 2.3
             sd = 7
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

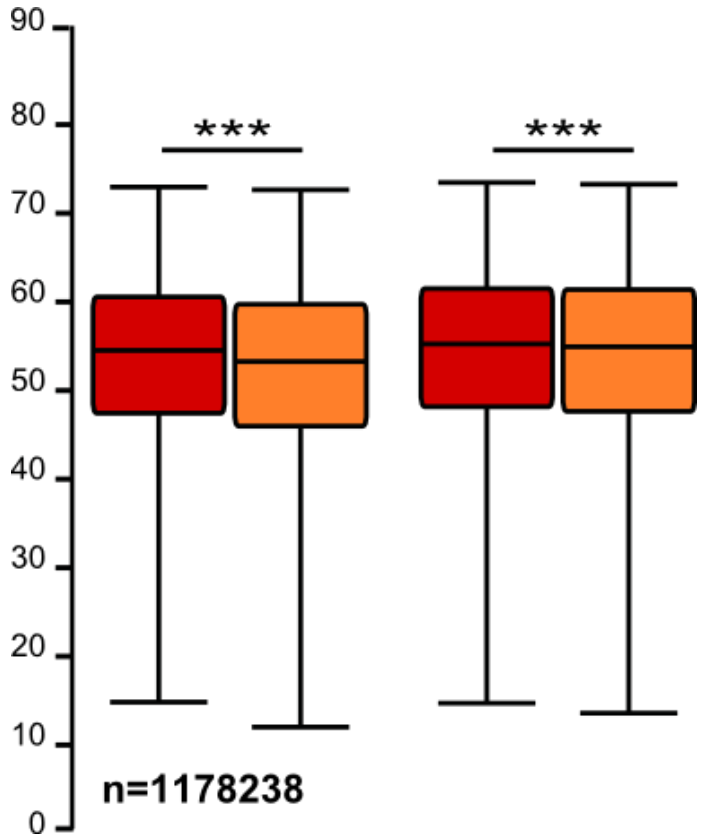
But does it make sense?

The sample size: the bigger the better?

- It takes huge samples to detect tiny differences but tiny samples to detect huge differences.

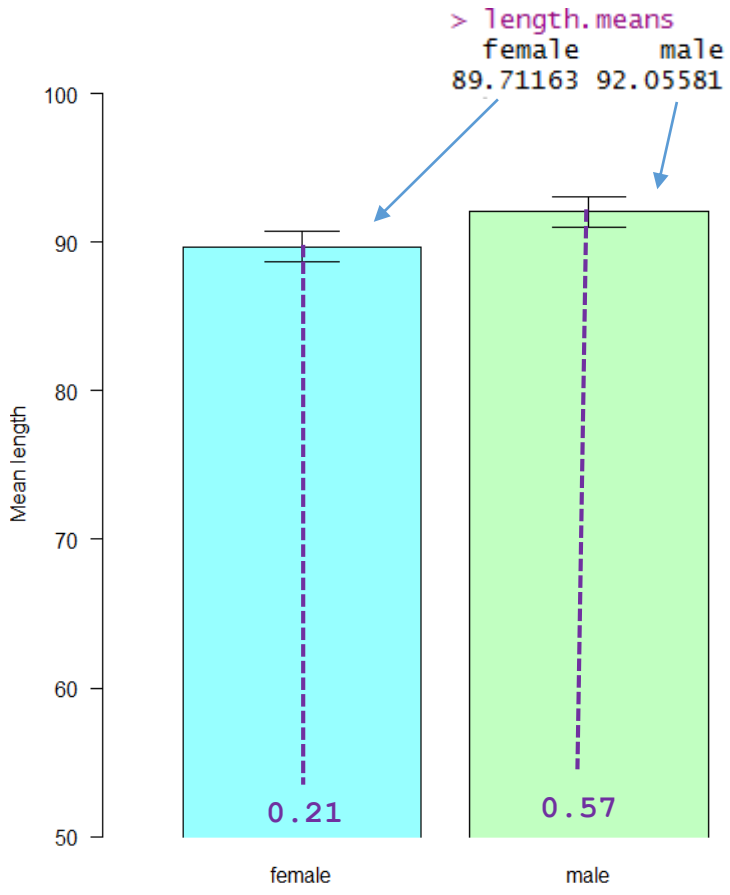
- What if the tiny difference is meaningless?
  - Beware of **overpower**
  - Nothing wrong with the stats: it is all about interpretation of the results of the test.

- Remember the important first step of power analysis
  - **What is the effect size of biological interest?**



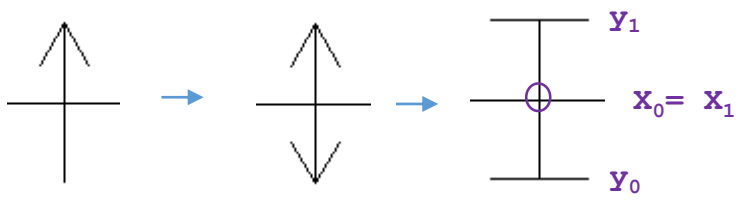
Plot ‘coyote.csv’ data

```
bar.length<-barplot(length.means,
         col=c("darkslategray1","darkseagreen1"),
         ylim=c(50,100),
         beside=TRUE,
         xlim=c(0,1),
         width=0.3,
         ylab="Mean length",
         las=1,
         xpd=FALSE)
```




```
length.se <- tapply(coyote$length,coyote$gender,std.error)
## plotrix package ##
                          female      male
                        0.9988377 1.0211241

bar.length      [,1]
           [1,] 0.21
           [2,] 0.57
```

```
arrows(x0=bar.length,
       y0=length.means-length.se,
       x1=bar.length,
       y1=length.means+length.se,
       length=0.3,
       angle=90,
       code=3)
```



Dependent or Paired *t*-test

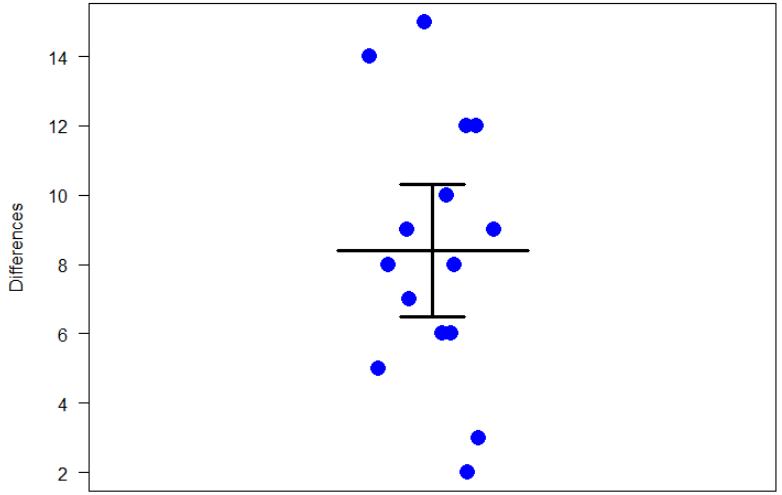
working.memory.csv

- A researcher is studying the effects of dopaminedepletion on working memory in rhesus monkeys.
- **Question**: does dopamine affect working memory in rhesus monkeys?
  - Load **working.memory.csv and** use `head()` to get to know the structure of the data.
  - Work out the difference: DA.depletion – placebo and assign the difference to a column: working.memory$difference
  - Plot the difference as a stripchart with a mean
  - Add confidence intervals as error bars
    - *Clue 1: you need* `std.error()` *from* # plotrix package #
    - *Clue 1 alternative: write a function to calculate the SEM (SD/√N)*
    - *Clue 2: interval boundaries: mean+/-1.96*SEM*
  - Run the paired *t*-test.

Dependent or Paired *t*-test - *Answers*

```
working.memory<-read.csv("working.memory.csv", header=T)
head(working.memory)

working.memory$difference <- working.memory$placebo-working.memory$DA.depletion

stripchart(working.memory$difference,
          vertical=TRUE,
          method="jitter",
          las=1,
          ylab="Differences",
          pch=16,
          col="blue",
          cex=2)

diff.mean <- mean(working.memory$difference)
centre<-1
segments(centre-0.15,diff.mean, centre+0.15, diff.mean, col="black", lwd=3)

diff.se <- std.error(working.memory$difference) ## plotrix package ##
lower<-diff.mean-1.96*diff.se
upper<-diff.mean+1.96*diff.se

arrows(x0=centre,
          y0=lower,
          x1=centre,
          y1=upper,
          length=0.3,
          code=3,
          angle=90,
          lwd=3)
```

	Subject	Placebo	DA.depletion
1	M1	9	7
2	M2	10	7
3	M3	15	10
4	M4	18	12
5	M5	19	13
6	M6	22	15

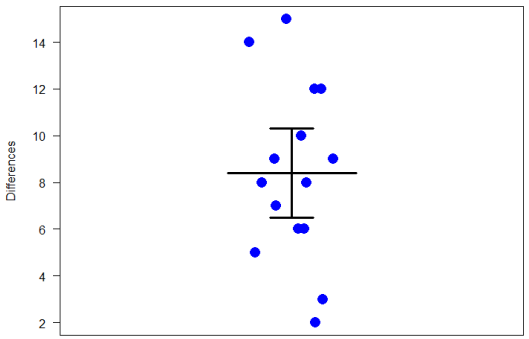


Alternative to using the plotrix package:

```
length.se<-tapply(coyote$length,coyote$gender,
          function(x) sd(x)/sqrt(length(x)))
```

Dependent or Paired *t*-test - *Answers*

Question: does dopamine affect working memory in rhesus monkeys?




```
> apply(working.memory[,3:4], 2, shapiro.test)
$`DA.depletion`

        Shapiro-Wilk normality test

data:  newX[, i]
W = 0.94274, p-value = 0.4181


$Difference

        Shapiro-Wilk normality test

data:  newX[, i]
W = 0.97727, p-value = 0.9474

> shapiro.test(working.memory$Difference)

        Shapiro-Wilk normality test

data:  working.memory$Difference
W = 0.97727, p-value = 0.9474
```

```
t.test(working.memory$placebo, working.memory$DA.depletion,paired=T)

        Paired t-test

data:  working.memory$Placebo and working.memory$DA.depletion
t = 8.6161, df = 14, p-value = 5.715e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  6.308997 10.491003
sample estimates:
mean of the differences
                    8.4
```

Answer: the injection of a dopamine-depleting agent significantly affects working memory in rhesus monkeys (t=8.62, df=14, p=5.715e-7).

Comparison of more than 2 means

- Running multiple tests on the same data increases the **familywise error rate**.

- What is the familywise error rate?
  - The error rate across tests conducted on the same experimental data.

- One of the basic rules ('laws') of probability:
  - The Multiplicative Rule: The probability of the joint occurrence of 2 or more independent events is the product of the individual probabilities.

$$\boxed{P(A,B) = P(A) \times P(B)}$$

For example:

$$P(2\ \text{Heads}) = P(\text{head}) \times P(\text{head}) = 0.5 \times 0.5 = 0.25$$

Familywise error rate

- **Example**: All pairwise comparisons between 3 groups A, B and C:
  - A-B, A-C and B-C

- Probability of making the Type I Error: **5%**
  - The probability of not making the Type I Error is 95% (=1 – 0.05)

- Multiplicative Rule:
  - Overall probability of no Type I errors is: 0.95 * 0.95 * 0.95 = 0.857

- So the probability of making at least one Type I Error is 1-0.857 = 0.143 or **14.3%**
    - The probability has increased from 5% to 14.3%

- Comparisons between 5 groups instead of 3, the familywise error rate is **40%** (=1-$(0.95)^n$)

Familywise error rate

- Solution to the increase of familywise error rate: correction for multiple comparisons
  - **Post-hoc tests**

- Many different ways to correct for multiple comparisons:
  - Different statisticians have designed corrections addressing different issues
    - e.g. unbalanced design, heterogeneity of variance, liberal vs conservative

- However, they all have **one thing in common**:
  - the more tests, the higher the familywise error rate: the more stringent the correction

- Tukey, Bonferroni, Sidak, Benjamini-Hochberg …
  - Two ways to address the multiple testing problem
    - **Familywise Error Rate** (FWER) vs. **False Discovery Rate** (FDR)

Multiple testing problem

- **FWER**: **Bonferroni**: α_{adjust} = 0.05/n comparisons e.g. 3 comparisons: 0.05/3=0.016
  - Problem: very conservative leading to loss of power (lots of false negative)
  - 10 comparisons: threshold for significance: 0.05/10: 0.005
  - Pairwise comparisons across 20.000 genes ☹

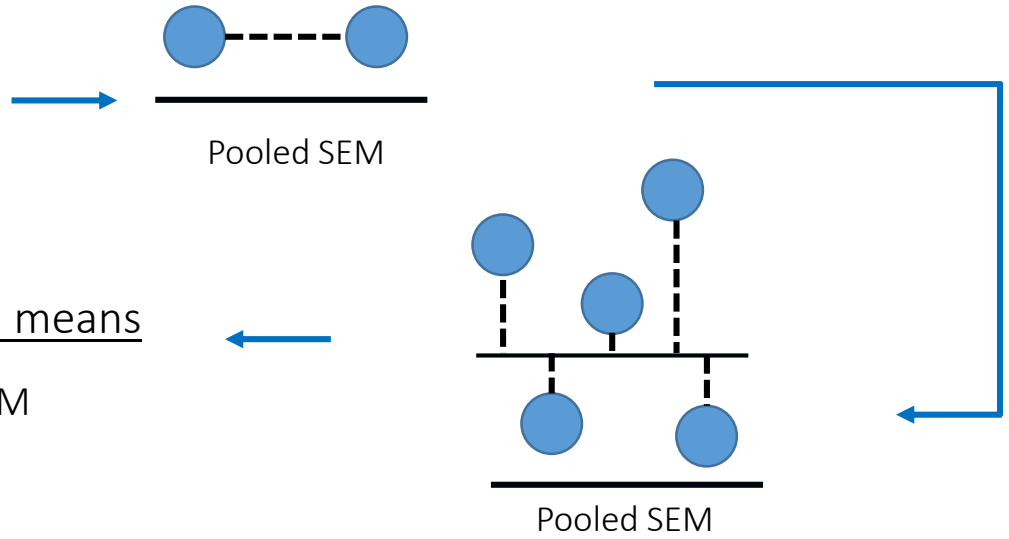
- **FDR**: **Benjamini-Hochberg**: the procedure controls the expected proportion of "discoveries" (significant tests) that are false (false positive).
  - Less stringent control of Type I Error than FWER procedures which control the probability of at least one Type I Error
  - More power at the cost of increased numbers of Type I Errors.

- **Difference between FWER and FDR**:
  - a p-value of 0.05 implies that 5% of all tests will result in false positives.
  - a FDR adjusted p-value (or **q-value**) of 0.05 implies that 5% of significant tests will result in false positives.

Analysis of variance

- Extension of the 2 groups comparison of a *t*-test but with a slightly different logic:

- *t*-test = $\frac{\text{mean1} - \text{mean2}}{\text{Pooled SEM}}$



- ANOVA = $\frac{\text{variance between means}}{\text{Pooled SEM}}$

- ANOVA compares variances:
  - If variance between the several means > variance within the groups (random error) then the means must be more spread out than it would have been by chance.

Analysis of variance

- The statistic for ANOVA is the **F ratio**.

- $F = \dfrac{\text{Variance between the groups}}{\text{Variance within the groups (individual variability)}}$

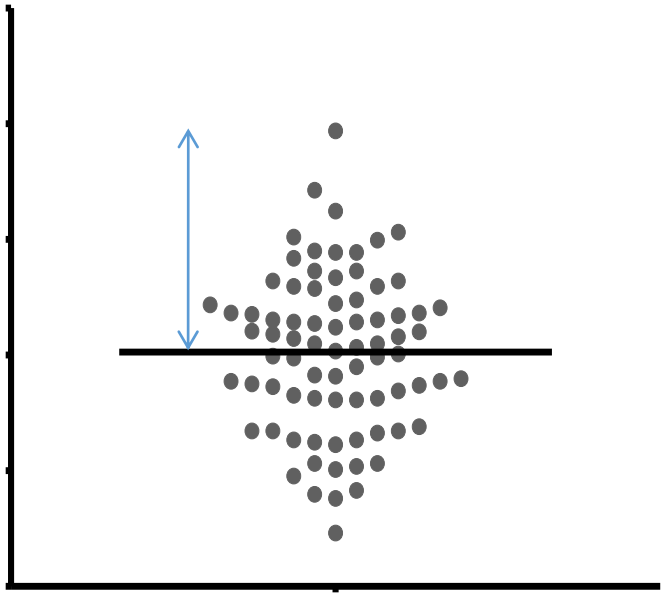
- $F = \dfrac{\text{Variation explained by the model (= systematic)}}{\text{Variation explained by unsystematic factors (= random variation)}}$

- If the variance amongst sample means is greater than the error/random variance, then F>1
  - In an ANOVA, we test whether F is significantly higher than 1 or not.

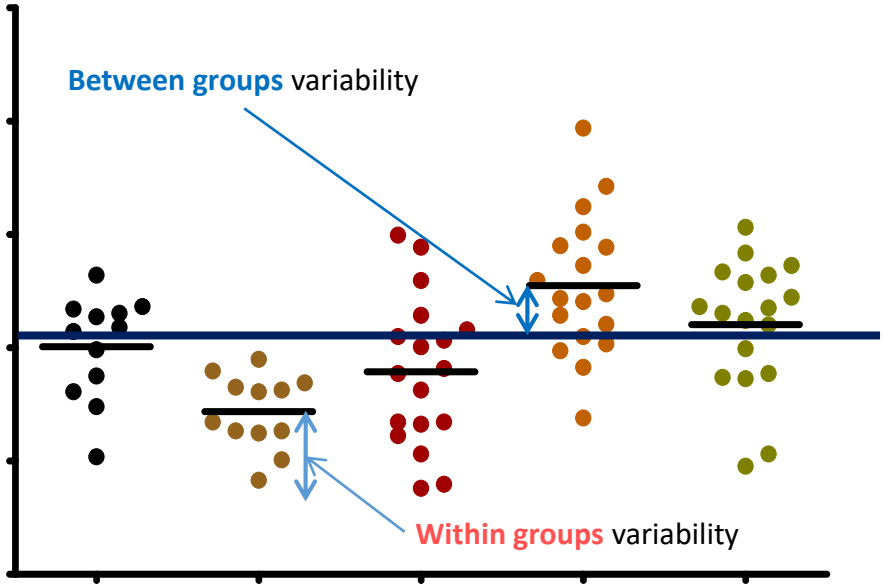
Analysis of variance

Source of variation	Sum of Squares	df	Mean Square	F	p-value
Between Groups	2.665	4	0.6663	8.423	<0.0001
Within Groups	5.775	73	0.0791		
Total	8.44	77			

- Variance (= SS / N-1) is the mean square
  - df: degree of freedom with df = N-1



Total sum of squares



<u>Example</u>: One-way ANOVA: protein.expression.csv

- **<u>Question</u>**: is there a difference in protein expression between the 5 cell lines?

- **1 Plot the data**
- **2 Check the assumptions for parametric test**
- **3 Statistical analysis: ANOVA**

<u>Example</u>: One-way ANOVA: protein.expression.csv

- **Question**: Difference in protein expression between 5 cell types?
  - Load **protein.expression.csv**
  - Restructure the file: wide to long
    - Clue: `melt()` ## reshape2 ##
  - Rename the columns: `"line"` and `"expression"`
    - Clue: `colnames()`
  - Remove the NAs
    - Clue: `na.omit`
  - Plot the data using at least 2 types of graph

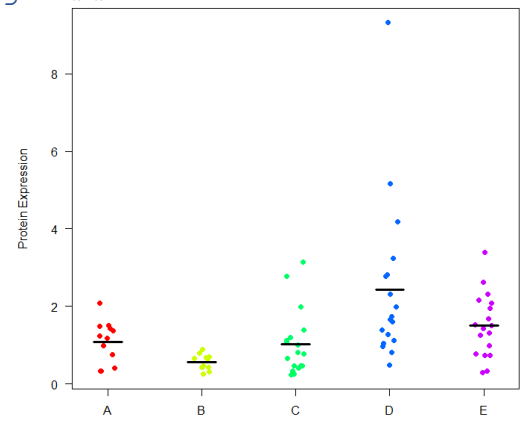
Example: One-way ANOVA: protein.expression.csv

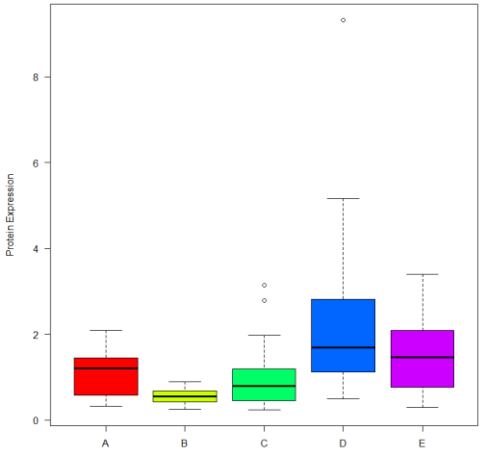
```
protein<-read.csv("protein.expression.csv",header=T)
protein.stack<-melt(protein) ## reshape2 package ##
colnames(protein.stack)<-c("line","expression")
protein.stack.clean <- na.omit(protein.stack)
head(protein.stack.clean)
```

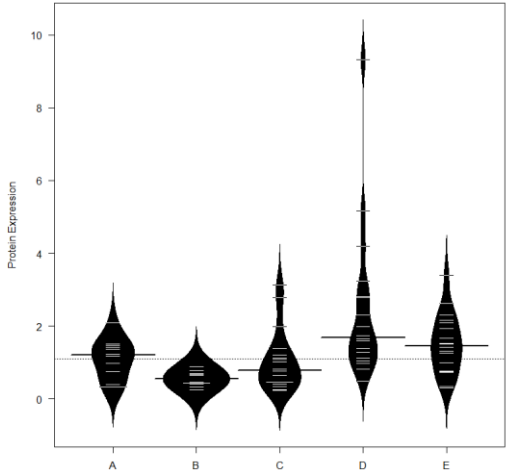
```
     A    B    C    D    E
1 0.40 0.26 0.24 1.04 0.74
2 1.50 0.47 0.25 2.78 0.99
3 0.98 0.42 1.01 0.82 1.26
4 0.33 0.64 0.77 1.65 1.50
5 0.75 0.32 0.47 0.49 0.30
6 1.48 0.65 0.47 0.97 0.34
```

```
  line expression
1    A       0.40
2    A       1.50
3    A       0.98
4    A       0.33
5    A       0.75
6    A       1.48
```

```
stripchart(protein.stack.clean$expression~protein.stack.clean$line,vertical=TRUE, method="jitter", las=1,
ylab="Protein Expression",pch=16,col=1:5)
expression.means<-tapply(protein.stack.clean$expression,protein.stack.clean$line,mean)
segments(1:5-0.15,expression.means, 1:5+0.15, expression.means, col="black", lwd=3)

boxplot(protein.stack.clean$expression~protein.stack.clean$line,col=rainbow(5),ylab="Protein Expression",las=1)

beanplot(protein.stack.clean$expression~protein.stack.clean$line, log="",ylab="Protein Expression",las=1)
## beanplot package ##
```







Assumptions of Parametric Data

```
tapply(protein.stack.clean$expression,protein.stack.clean$line, shapiro.test)
```

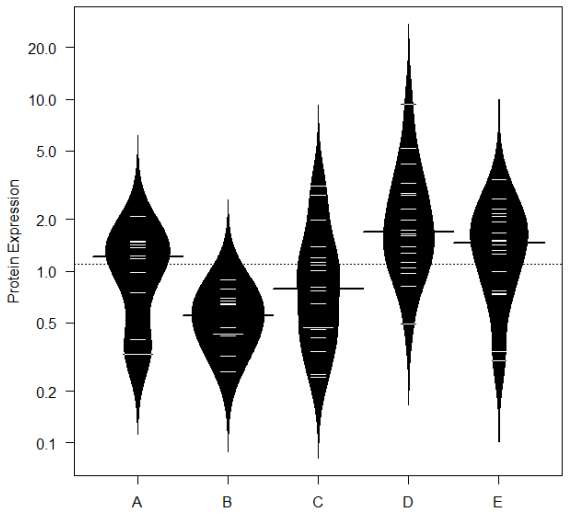
```
$`A`

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.92957, p-value = 0.3755


$B

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.95351, p-value = 0.6888


$C

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.81968, p-value = 0.002921


$D

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.75307, p-value = 0.0003549


$E

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.96707, p-value = 0.7411
```

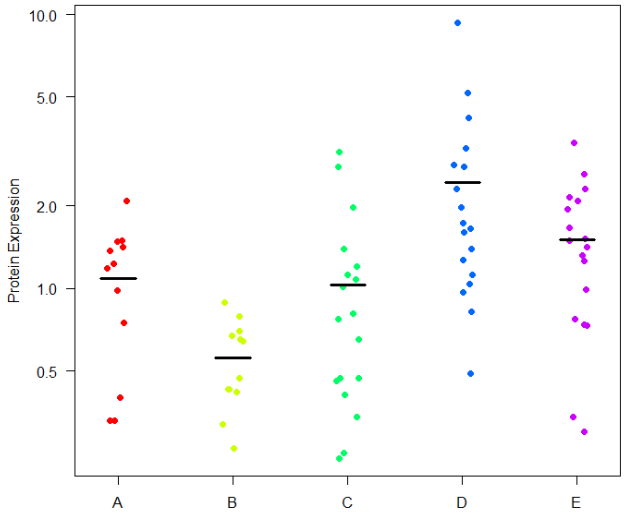
```
protein.stack.clean$log10.expression<-log10(protein.stack.clean$expression)
```

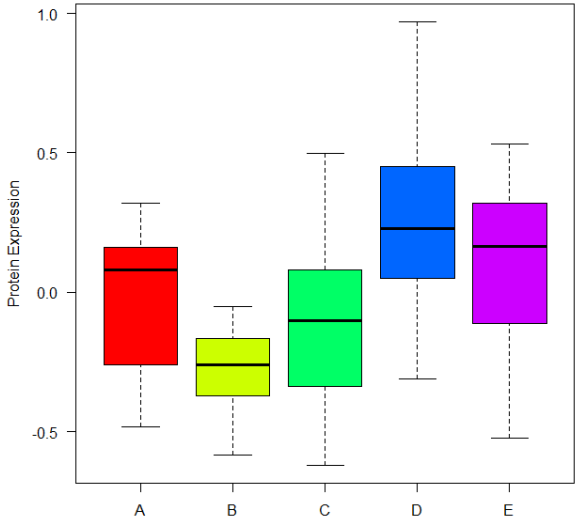
Plot ‘protein.expression.csv’ data

Log transformation

```
beanplot(protein.stack.clean$expression~protein.stack.clean$line, ylab="Protein Expression", las=1)

stripchart(protein.stack.clean$expression~protein.stack.clean$line,vertical=TRUE,
method="jitter", las=1, ylab="Protein Expression",pch=16,col=rainbow(5),log="y")

expression.means<-tapply(protein.stack.clean$expression,protein.stack.clean$line,mean)
segments(1:5-0.15,expression.means, 1:5+0.15, expression.means, col="black", lwd=3)

boxplot(protein.stack.clean$log10.expression~protein.stack.clean$line,col=rainbow(5), ylab="Protein
Expression",las=1)
```







Assumptions of Parametric Data

```
tapply(protein.stack.clean$log10.expression,protein.stack.clean$line,shapiro.test)
```

```
$`A`

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.85425, p-value = 0.04144


$B

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.94584, p-value = 0.5773


$C

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.96571, p-value = 0.7142


$D

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.98684, p-value = 0.9935


$E

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.93134, p-value = 0.205
```

Normality ☑-ish

```
bartlett.test(protein.stack.clean$log10.expression~protein.stack.clean$line)
```

```
        Bartlett test of homogeneity of variances

data:  protein.stack.clean$log10.expression by protein.stack.clean$line
Bartlett's K-squared = 5.8261, df = 4, p-value = 0.2125
```

Homogeneity in variance ☑

Analysis of variance: Post hoc tests

- The ANOVA is an “omnibus” test: it tells you that there is (or not) a difference between your means but not exactly which means are significantly different from which other ones.
  - To find out, you need to apply **post hoc** tests.
  - These post hoc tests should only be used when the ANOVA finds a significant effect.

Analysis of variance

```
anova.log.protein<-aov(log10.expression~line,data=protein.stack.clean)
summary(anova.log.protein)
```

```
            Df Sum Sq Mean Sq F value   Pr(>F)
line         4  2.691  0.6728   8.123 1.78e-05 ***
Residuals   73  6.046  0.0828
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
pairwise.t.test(protein.stack.clean$log10.expression,protein.stack.clean$line, p.adj = "bonf")
```

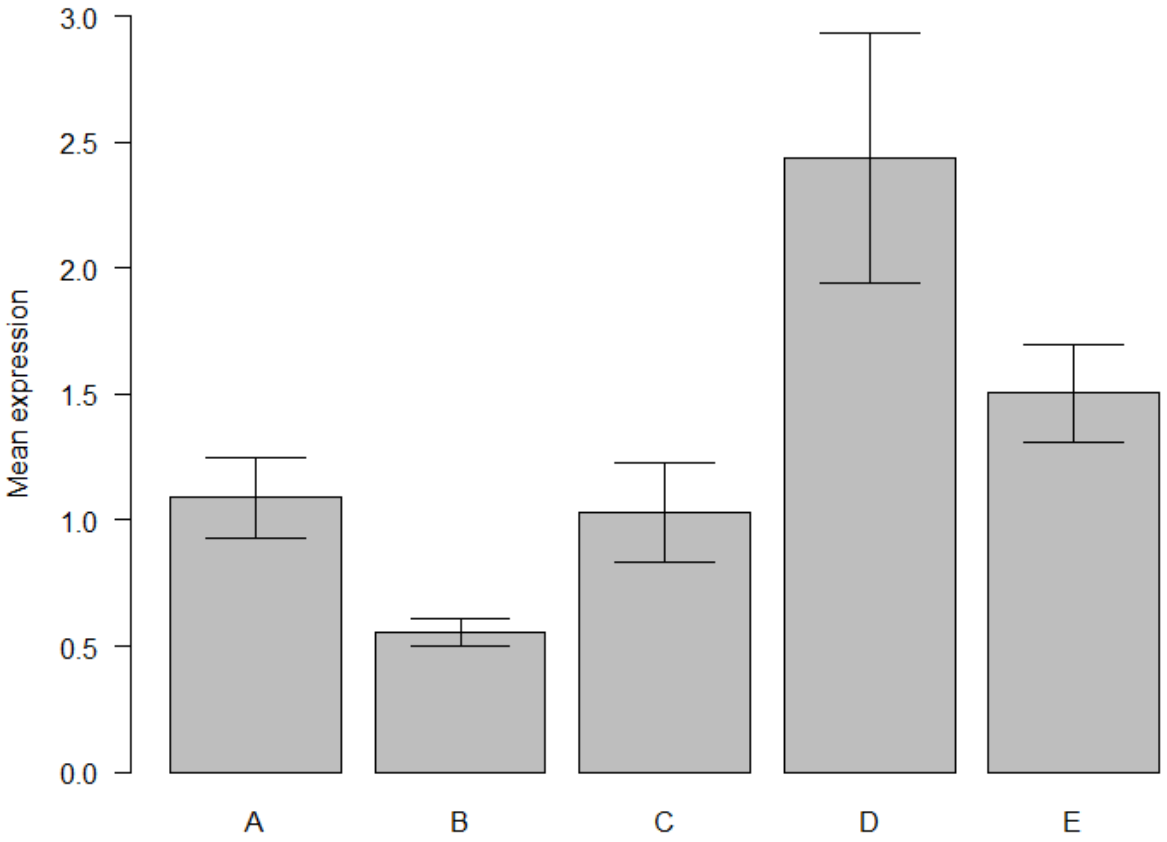
```
        Pairwise comparisons using t tests with pooled SD

data:  protein.stack.clean$log10.expression and protein.stack.clean$line

  A      B       C      D
B 0.3655 -       -      -
C 1.0000 1.0000  -      -
D 0.0571 1.9e-05 0.0017 -
E 1.0000 0.0062  0.3318 0.7675

P value adjustment method: bonferroni
```

```
TukeyHSD(anova.log.protein,"line")
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = log10.expression ~ line, data = protein.stack.clean)

$line
           diff          lwr        upr     p adj
B-A -0.25024832 -0.578882494 0.07838585 0.2187264
C-A -0.07499724 -0.374997820 0.22500335 0.9560187
D-A  0.30549397  0.005493391 0.60549456 0.0438762
E-A  0.13327517 -0.166725416 0.43327575 0.7265567
C-B  0.17525108 -0.124749499 0.47525167 0.4809387
D-B  0.55574230  0.255741712 0.85574288 0.0000183
E-B  0.38352349  0.083522904 0.68352407 0.0054767
D-C  0.38049121  0.112162532 0.64881989 0.0015431
E-C  0.20827240 -0.060056276 0.47660108 0.2023355
E-D -0.17221881 -0.440547487 0.09610987 0.3841989
```

Analysis of variance

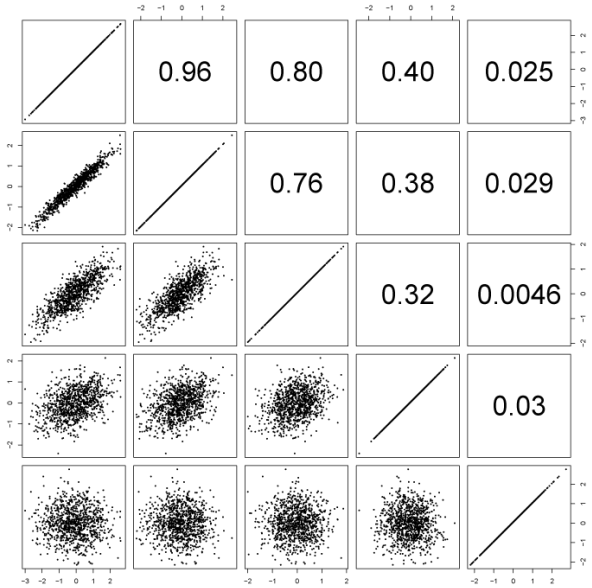
```
bar.expression<-barplot(expression.means, beside=TRUE, ylab="Mean expression", ylim=c(0, 3), las=1)

expression.se <- tapply(protein.stack.clean$expression,protein.stack.clean$line,std.error)

arrows(x0=bar.expression, y0=expression.means-expression.se,
x1=bar.expression, y1=expression.means+expression.se, length=0.2, angle=90,code=3)
```

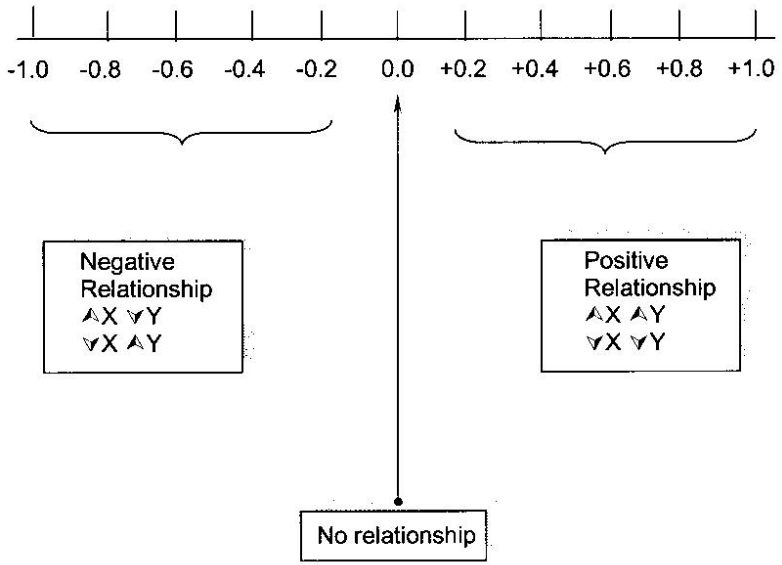


Association between 2 continuous variables

Correlation

- A correlation coefficient is an index number that measures:
  - The <u>magnitude</u> and the <u>direction</u> of the relation between 2 variables
  - It is designed to range in value between -1 and +1





Correlation

- Most widely-used correlation coefficient:
  - Pearson product-moment correlation coefficient “r”

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

    - The 2 variables do not have to be measured in the same units but they have to be proportional (meaning linearly related)

  - Coefficient of determination:
    - r is the correlation between X and Y
    - r^2 is the coefficient of determination:
      - It gives you the proportion of variance in Y that can be explained by X, in percentage.

Correlation

- <u>Assumptions for correlation</u>
  - Regression and linear Model (lm)

- **Linearity**: The relationship between X and the mean of Y is linear.
- **Homoscedasticity**: The variance of residual is the same for any value of X.
- **Independence:** Observations are independent of each other.
- **Normality:** For any fixed value of X, Y is normally distributed.

Correlation

- <u>Assumptions for correlation</u>
  - Regression and linear Model (lm)

- **Outliers**: the observed value for the point is very different from that predicted by the regression model.

- **Leverage points**: A leverage point is defined as an observation that has a value of x that is far away from the mean of x.

- **Influential observations**: change the slope of the line. Thus, have a large influence on the fit of the model.

One method to find influential points is to compare the fit of the model with and without each observation.

- Bottom line: **influential outliers** are problematic.

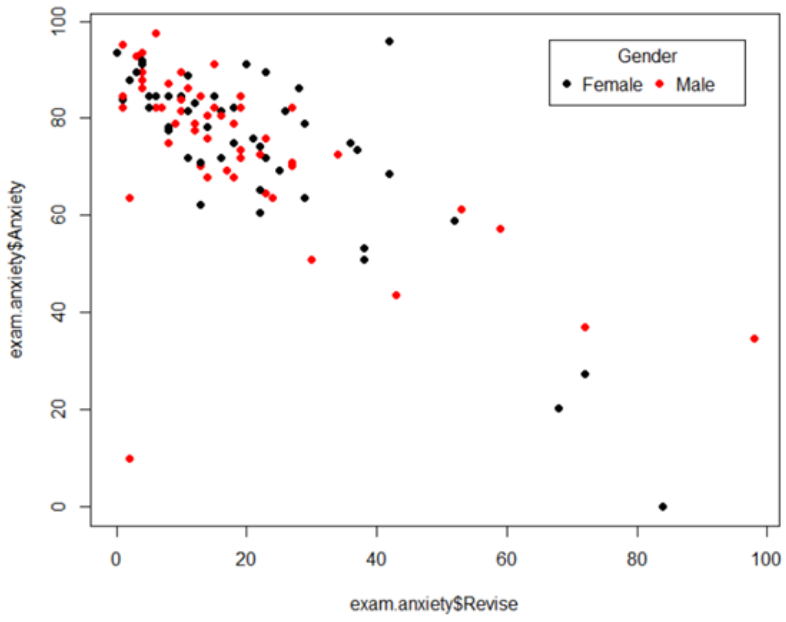
Correlation: exam.anxiety.dat

- Is there a relationship between time spent revising and exam anxiety?

```
exam.anxiety<-read.table("Exam Anxiety.dat", sep="\t",header=T)
head(exam.anxiety)
```

```
  Code Revise Exam Anxiety Gender
1    1      4   40  86.298   Male
2    2     11   65  88.716 Female
3    3     27   80  70.178   Male
4    4     53   80  61.312   Male
5    5      4   40  89.522   Male
6    6     22   70  60.506 Female
```

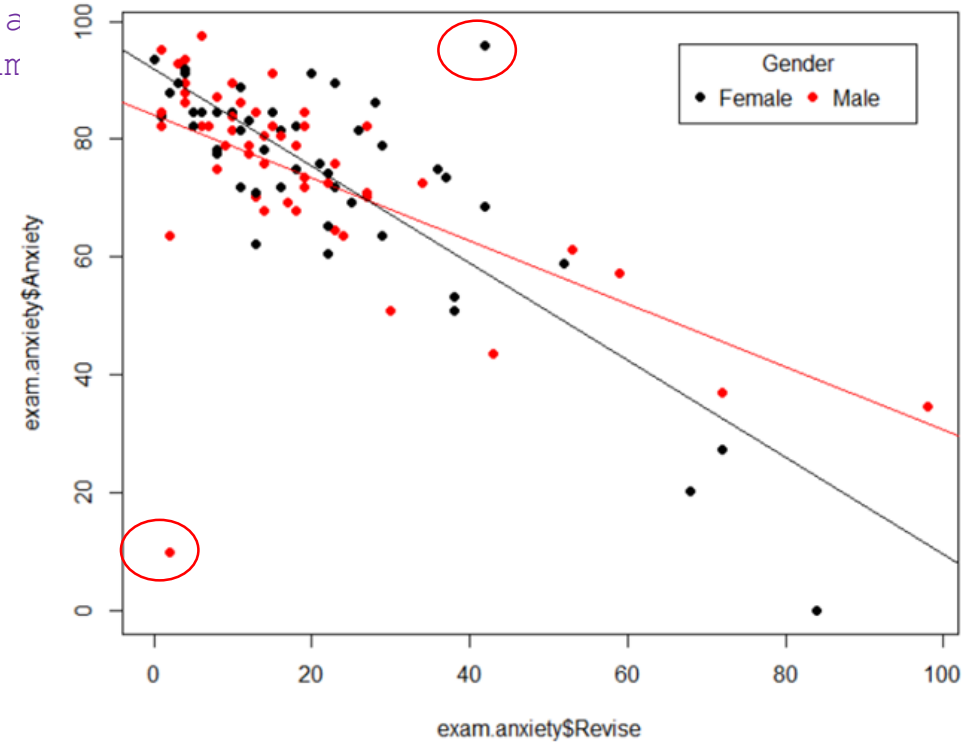
```
plot(exam.anxiety$Revise,exam.anxiety$Anxiety,col=exam.anxiety$Gender,pch=16)

legend("topright", title="Gender",inset=.05, c("Female","Male"), horiz=TRUE, pch=16,col=1:2)
```



Correlation: exam anxiety.dat

- Is there a relationship between time spent revising and exam anxiety?
  - `lm()` linear modelling
  - model(x) = y (e.g. mean(3, 5, 6) = 4.7)
  - lm(outcome ~ predictor) (e.g. in mammals: lm(weight ~ sex)

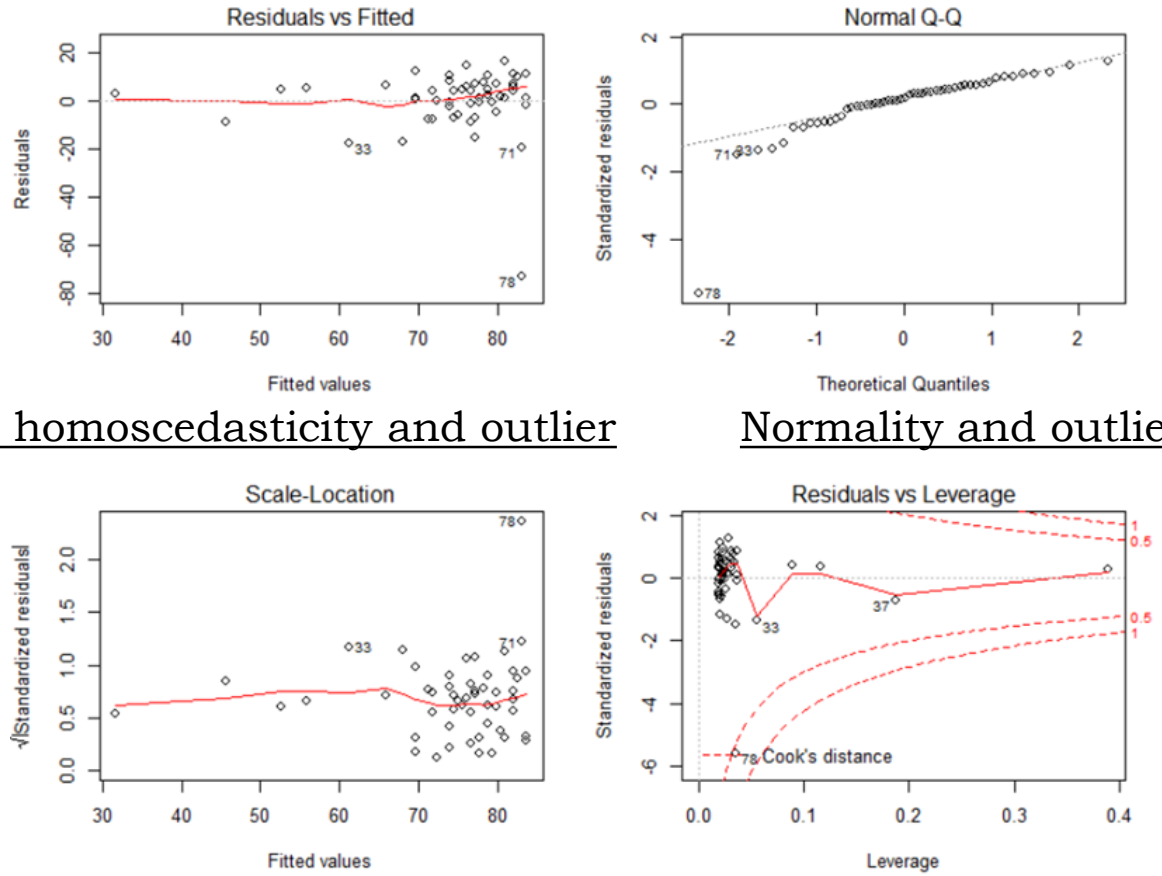
```
fit.male<-lm(Anxiety~Revise,data=exam.a
fit.female<-lm(Anxiety~Revise,data=exam
abline((fit.male), col="red")
abline((fit.female), col="black")
```



Correlation: exam anxiety.dat

Assumptions, outliers and influential cases

```
par(mfrow=c(2,2))
plot(fit.male)
```



Linearity, homoscedasticity and outlier

Normality and outlier

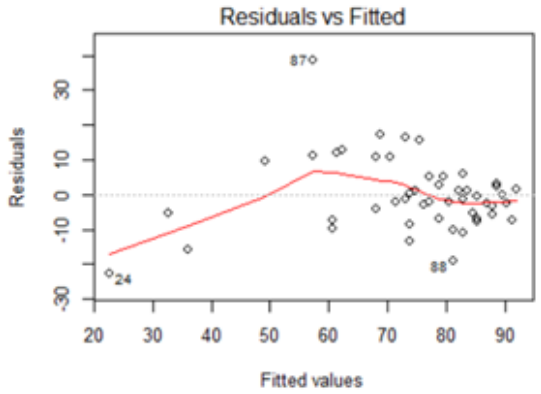
Homoscedasticity

Influential cases

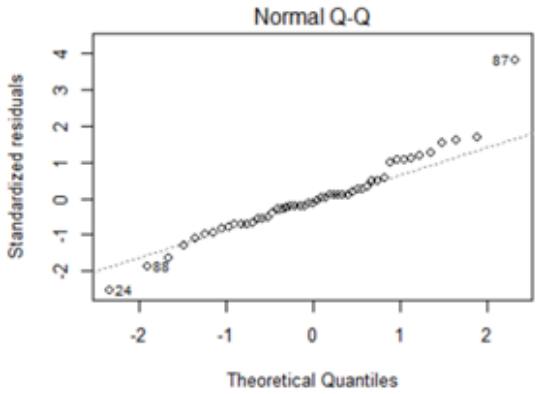
Correlation: exam anxiety.dat

Assumptions, outliers and influential cases

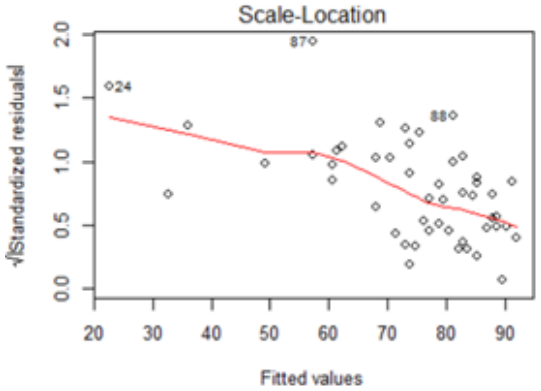
```
plot(fit.female)
```



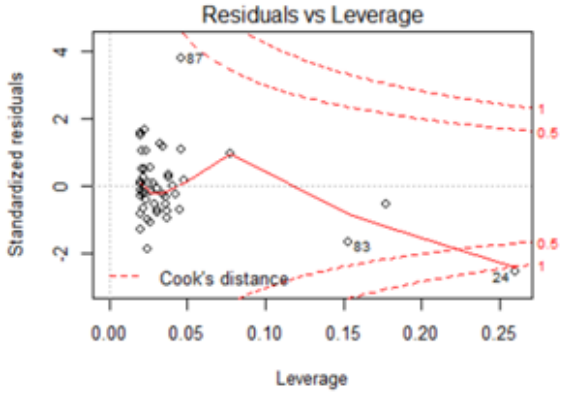
Linearity, homoscedasticity and outlier



Normality and outlier

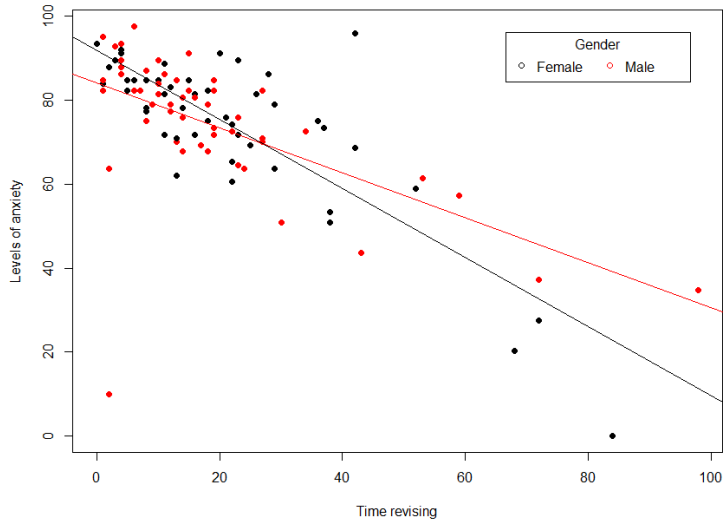


Homoscedasticity



Influential cases

Correlation: exam anxiety.dat




```
cor(exam.anxiety[exam.anxiety$Gender=="Male",
c("Exam","Anxiety","Revise")])
```

```
              Exam    Anxiety     Revise
Exam     1.0000000 -0.5056874  0.3593981
Anxiety -0.5056874  1.0000000 -0.5973682
Revise   0.3593981 -0.5973682  1.0000000
```

```
cor(exam.anxiety[exam.anxiety$Gender == "Female",
c("Exam","Anxiety","Revise")])
```

```
              Exam    Anxiety     Revise
Exam     1.0000000 -0.3813845  0.4399865
Anxiety -0.3813845  1.0000000 -0.8213698
Revise   0.4399865 -0.8213698  1.0000000
```

```
summary(fit.male)
```

Anxiety=84.19-0.53*Revise

```
Call:
lm(formula = Anxiety ~ Revise, data = exam.anxiety[exam.anxiety$Gender ==
    "Male", ])

Residuals:
    Min      1Q  Median      3Q     Max
-73.124  -2.900   2.221   6.750  16.600

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  84.1941     2.6213  32.119  < 2e-16 ***
Revise       -0.5353     0.1016  -5.267 2.94e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.3 on 50 degrees of freedom
Multiple R-squared:  0.3568, Adjusted R-squared:  0.344
F-statistic: 27.74 on 1 and 50 DF,  p-value: 2.937e-06
```

Anxiety=91.94-0.82*Revise

```
summary(fit.female)
```

```
Call:
lm(formula = Anxiety ~ Revise, data = exam.anxiety[exam.anxiety$Gender ==
    "Female", ])

Residuals:
    Min      1Q  Median      3Q     Max
-22.687  -6.263  -1.204   4.197  38.628

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 91.94181    2.27858   40.35  < 2e-16 ***
Revise      -0.82380    0.08173  -10.08 1.54e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.42 on 49 degrees of freedom
Multiple R-squared:  0.6746, Adjusted R-squared:  0.668
F-statistic: 101.6 on 1 and 49 DF,  p-value: 1.544e-13
```

Correlation: exam anxiety.dat

Influential outliers (fit2)

```
exam.anxiety.filtered <- exam.anxiety[c(-78,-87),]
```

```
> fit.male2 <- lm(Anxiety~Revise, data=exam.anxiety.filtered[exam.anxiety.filtered$Gender=="Male",])
> summary(fit.male2)

Call:
lm(formula = Anxiety ~ Revise, data = exam.anxiety.filtered[exam.anxiety.filtered$Gender ==
    "Male", ])

Residuals:
     Min       1Q   Median       3Q      Max
-22.0296  -3.8704   0.5626   6.0786  14.2525

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 86.97461    1.64755  52.790  < 2e-16 ***
Revise      -0.60752    0.06326  -9.603 7.59e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.213 on 49 degrees of freedom
Multiple R-squared:  0.653,	Adjusted R-squared:  0.6459
F-statistic: 92.22 on 1 and 49 DF,  p-value: 7.591e-13
```

Anxiety=86.97-0.61*Revise

```
> fit.female2 <- lm(Anxiety~Revise, data=exam.anxiety.filtered[exam.anxiety.filtered$Gender=="Female",])
> summary(fit.female2)

Call:
lm(formula = Anxiety ~ Revise, data = exam.anxiety.filtered[exam.anxiety.filtered$Gender ==
    "Female", ])

Residuals:
     Min       1Q   Median       3Q      Max
-18.7518  -5.7069  -0.7782   3.2117  18.5538

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 92.24536    1.93591   47.65   <2e-16 ***
Revise      -0.87504    0.07033  -12.44   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.849 on 48 degrees of freedom
Multiple R-squared:  0.7633,	Adjusted R-squared:  0.7584
F-statistic: 154.8 on 1 and 48 DF,  p-value: < 2.2e-16
```

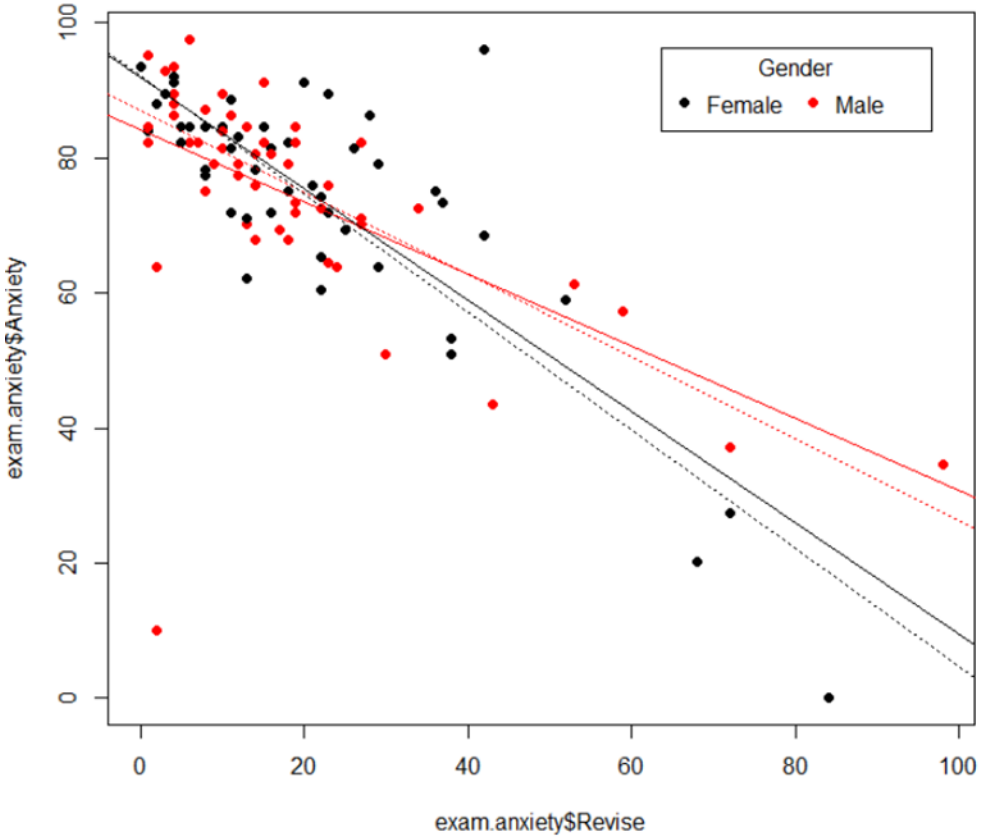
Anxiety=92.25-0.86*Revise

```
> cor(exam.anxiety.filtered[exam.anxiety.filtered$Gender=="Male",c("Exam","Anxiety","Revise")])
              Exam    Anxiety     Revise
Exam     1.0000000 -0.4653914  0.4028863
Anxiety -0.4653914  1.0000000 -0.8080950
Revise   0.4028863 -0.8080950  1.0000000
> cor(exam.anxiety.filtered[exam.anxiety.filtered$Gender=="Female",c("Exam","Anxiety","Revise")])
              Exam    Anxiety     Revise
Exam     1.0000000 -0.4070663  0.4312691
Anxiety -0.4070663  1.0000000 -0.8736684
Revise   0.4312691 -0.8736684  1.0000000
```

Correlation

without the outlier/influential case

```
plot(exam.anxiety$Revise,exam.anxiety$Anxiety,col=exam.anxiety$Gender,pch=16)

legend("topright", title="Gender",inset=.05, c("Female","Male"), horiz=TRUE, pch=16,col=1:2)
abline((fit.male), col="red")
abline((fit.female), col="black")
abline((fit.male2), col="red",lty=3)
abline((fit.female2), col="black",lty=3)
```



My email address if you need some help with GraphPad:

anne.segonds-pichon@babraham.ac.uk

Slides and manual available on:

https://www.bioinformatics.babraham.ac.uk/training.html