Data Science in the Wild

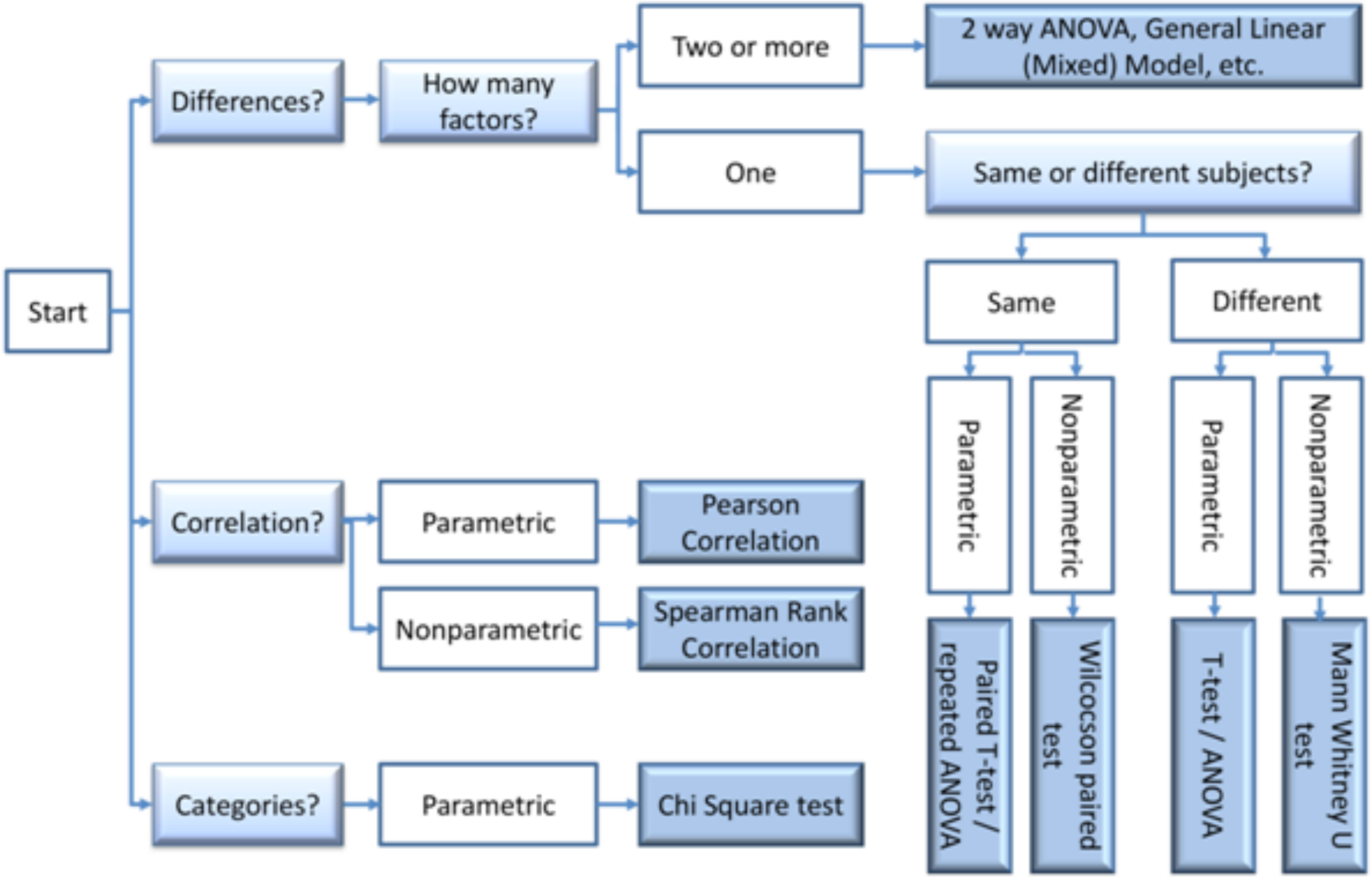
Lecture 8: Advanced Experimental Analysis

Eran Toch

CORNELL TECH

Types of Tests

- Parametric vs. Non-Parametric
- Difference vs. Correlation
- Categorical vs. Differential
- Number of samples



Agenda

1. Introduction
2. ANOVA
3. Post-hoc tests
4. Correlation tests
5. Sampling

Analysis of Variance - ANOVA

Why not t-tests?

- Every time you conduct a t-test there is a chance that you will make a Type I error with a probability of $\alpha = 0.05$
- By running two t-tests on the same data you will have increased your chance of making a mistake to about 0.1
- ANOVA controls for these errors, keeping the confidence level to 0.95

Example

- If you are comparing 3 groups (A, B, C), than you can do a 3 total comparisons

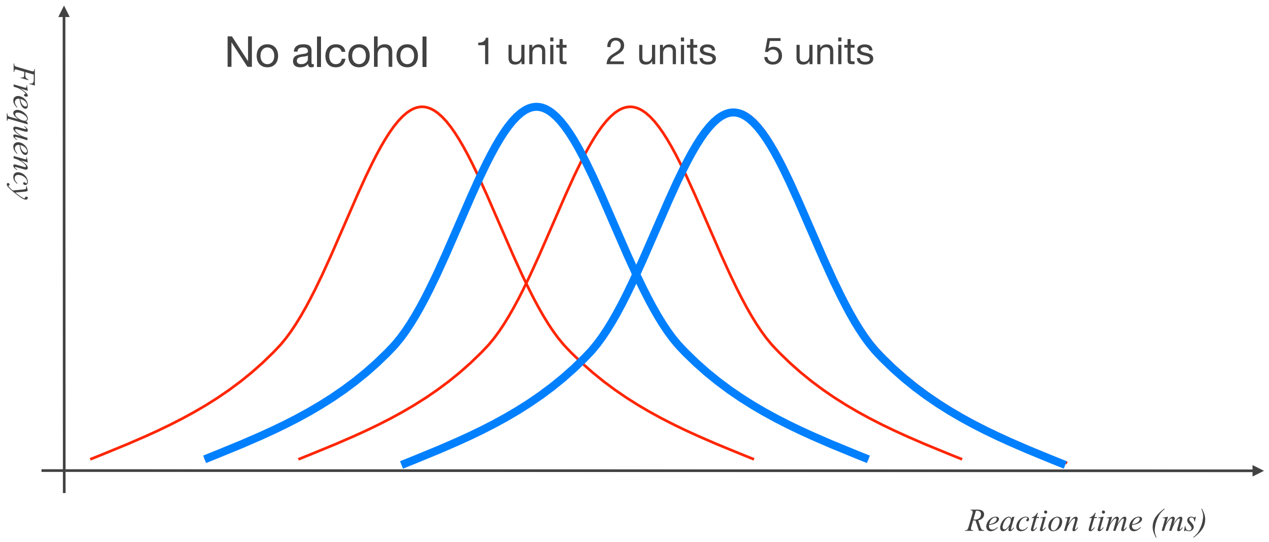
  A – B
  A – C
  B – C

- The experiment-wise error rate without any adjustments would be:

  $\alpha_e = 1 - (1-\alpha)^c$
  $= 1 - (1-.05)^3$
  $= 1 - .95^3$
  $= 1 - 0.86$
  $= .14$

ANOVA

- ANOVA will tell us if one condition is significantly different to one or more of the others
- But it won't tell us which conditions are different
- We can compare one (or more) against one (or more) of the others



ANOVA

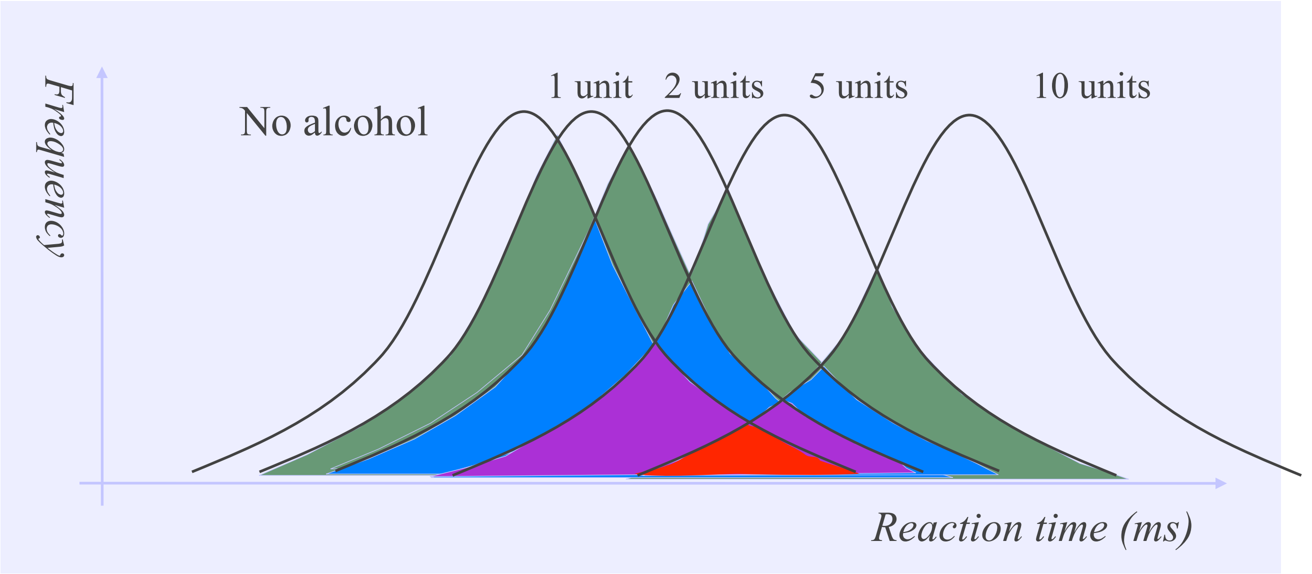
- Analysis of variance (ANOVA) is used to determine whether groups of data are the same or different
- It incorporates means and variances to determine its test statistics, called the F-ratio
- What is the null hypothesis?
    - $H_{0:}$ $x_1 = x_2 = x_3 = x_4 = \ldots x_k$ (x - group mean, k - number of groups)

Conditions

- The dependent variable is normally distributed in each group
- Homogeneity of variances:
  - The variance in each group should be similar enough.
  - For example, using the Bartlett test
- Data type: The dependent variable must be interval or ratio (e.g., time or error rates)

Analysis of Variance (ANOVA)

$$\text{F-ratio} = \frac{\text{Area of non-overlap (hypothesis true)}}{\text{Area of overlap (hypothesis false)}}$$



- Large “F” means significant differences
- Large “F” means evidence in support of hypothesis
- We need to calculate the size of all these areas

F ratio

- Mean square
- MS error - the variance not accounted for by the variable
- F ratio is a variance ratio or ‘signal to noise’ ratio
- Large F means large differences accounted for by the variable

$$MS = \frac{SS}{df}$$

$$F = \frac{MS_{within}}{MS_{between}}$$

Where:

MS - mean square

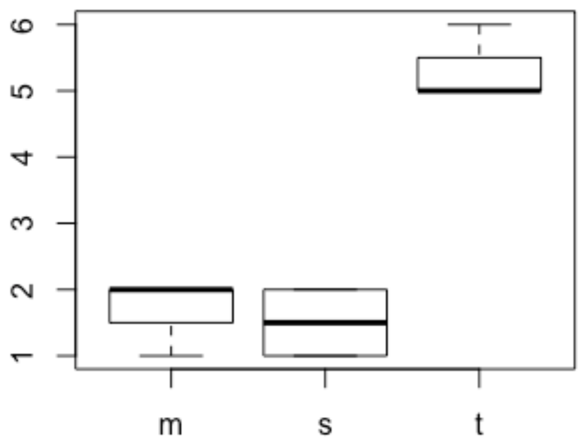
SS - sum of squares

df - degrees of freedom

One-way Anova

- One-way ANOVA is used to determine whether there are any statistically significant differences between the means of three or more independent groups
- Suits a simple between-subject design with one independent variable

Participant	Condition	Values
1	Mouse	1
2	Mouse	2
3	Mouse	2
4	Touch	5
5	Touch	6
6	Touch	5
7	Speech	2
8	Speech	1



Model

$$Y_{ij} = \mu + A_i + \varepsilon_{ij}$$

- An observation Y_{ij} is given by the average performance of the users (μ), the effect of the treatment (A_i) and an error for each participant and condition ε_{ij}
- Our goal is to test if the hypothesis $A_1 = A_2 = A_3 = A_4 = \ldots A_k = 0$ is plausible

Calculation

- Means:
  - M_{mouse} = (1 + 2 + 2) / 3 = 1.667
  - M_{touch} = (5 + 6 + 5) / 3 = 5.33
  - M_{Speech} = (1 + 2) / 2 = 1.5

- The grand mean is calculated as follows:
  - $\mu\hat{} = (1 + 2 ++ 2 + 5 + 6 + 5 + 2 + 1) / 8 = 3$

Estimated Effect

- The estimated effects, $A^\wedge_i$, are the difference between the estimated overall mean and the estimated treatment mean:

  $$A^\wedge_i = M_i - \mu^\wedge$$

- Therefore, we get:
  - $A_{mouse} = 1.667 - 3 = -1.33$
  - $A_{touch} = 5.333 - 3 = 2.333$
  - $A_{Speech} = 1.5 - 3 = -1.5$

Degrees of Freedom

- Calculating the degrees of freedom (just minus 1, actually)
- $df_{between} = 3 - 1 = 2$
- $df_{within} = 8 - 1 = 7$

Sum of Squares

- $SS_{between}$: Sum of squares between conditions

$$\sum \hat{A}_i^2 \cdot \#measures$$

$$= (-1.33)^2 * 3 + (2.33)^2 * 3 + (1.4)^2 * 2 = 26.17$$

- SS_{within}: Sum of squares within conditions

$$\sum_i \sum_j (y_{ij} - y_i)2$$

$$= [(1-1.667)^2 + (2 - 1.6667)^2 + (2 - 1.6667)^2] + [0.667] + [0.5] = 1.83$$

Calculating the Mean Square

- $MS = SS / df$
- $MS_{between} = \frac{SS_{between}}{df_{between}} = 26.17 / 2 = 13.08$
- $MS_{within} = \frac{SS_{within}}{df_{within}} = 1.83 / 5 = 0.37$
- $F = \frac{MS_{between}}{MS_{within}} = 13.08 / 0.37 = 35.68$

Interpretation

- The F–value says us how far away we are from the hypothesis of indistinguishability between the error and the conditions (treatment)
- A large F-value implies that the effect of the treatment (conditions) is relevant
- We calculate the critical value for the level α = 5% with degrees of freedom 2 and 5.
  - p = 0.011 => We can reject the hypothesis that $A_{mouse} = A_{touch} = A_{speech} = 0$

F	Degrees of Freedom in the Numerator				
α = 0.05	6	7	8	9	10
Degress of Freedom in the Denominator: 1	234.0	236.8	238.9	240.5	241.9
2	19.33	19.35	19.37	19.38	19.40
3	8.94	8.89	8.85	8.81	8.79
4	6.16	6.09	6.04	6.00	5.96
5	4.95	4.88	4.82	4.77	4.74
6	4.28	4.21	4.15	4.10	4.06
7	3.87	3.79	3.73	3.68	3.64
8	3.58	3.50	3.44	3.39	3.35
9	3.37	3.29	3.23	3.18	3.14
10	3.22	3.14	3.07	3.02	2.98
12	3.00	2.91	2.85	2.80	2.75
15	2.79	2.71	2.64	2.59	2.54
20	2.60	2.51	2.45	2.39	2.35
25	2.49	2.40	2.34	2.28	2.24

Python Code

```
from scipy import stats

F, p = stats.f_oneway(d_data['ctrl'], d_data['trt1'],
d_data['trt2'])
```

Factorial ANOVA

- Factorial ANOVA (two-way) measures whether a combination of independent variables predict the value of a dependent variable
- Suits between-group design, with multiple conditions

Observation	Gender	Dosage	Alertness
1	m	a	8
2	m	a	12
3	m	a	13
4	m	a	12
5	m	b	6
6	m	b	7
7	m	b	23
8	m	b	14
9	f	a	15
10	f	a	12
11	f	a	22
12	f	a	14
13	f	b	15
14	f	b	12
15	f	b	18
16	f	b	22

Python Code

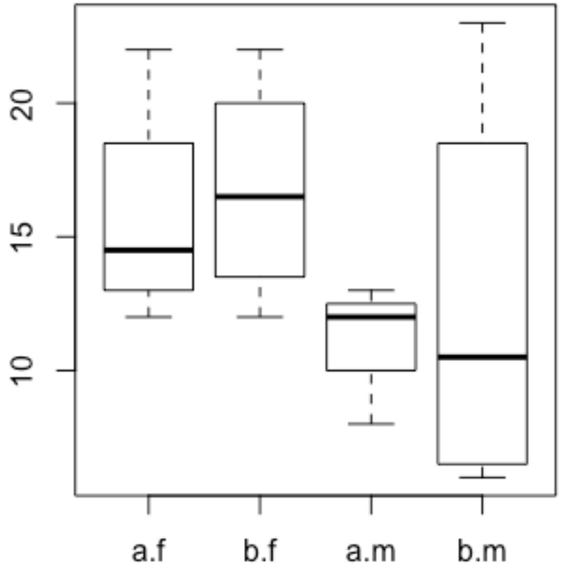
```
formula = 'len ~ C(supp) + C(dose) + C(supp):C(dose)'
model = ols(formula, data).fit()
aov_table = anova_lm(model, typ=2)
```

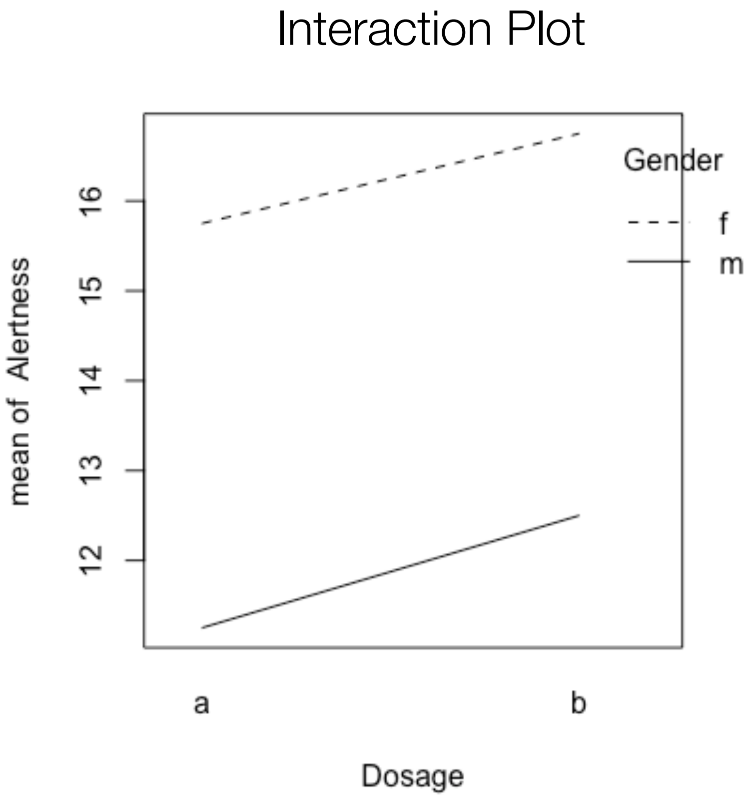
```
from pyvttbl import DataFrame
df=DataFrame()
df.read_tbl(datafile)
df['id'] = xrange(len(df['len']))

print(df.anova('len', sub='id', bfactors=['supp', 'dose']))
```

https://www.marsja.se/three-ways-to-carry-out-2-way-anova-with-python/

Visualization





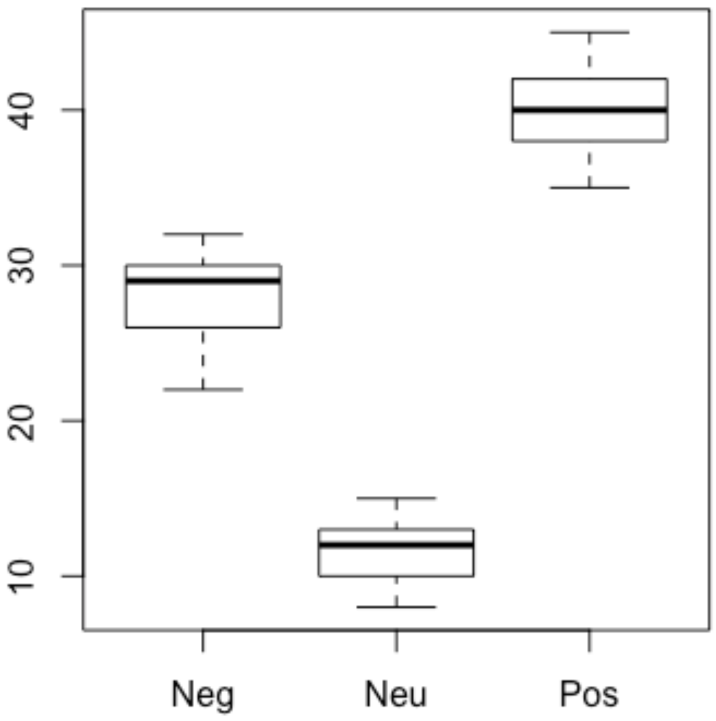
Repeated Measure ANOVA

- In repeated measure ANOVA, we test the same entity in several conditions
  - One independent variable: one way
  - Several independent variables: two way
- Suits a within-subject study with multiple conditions
- The design should be balanced: without missing values in some conditions

```
aov = df.anova('rt', sub='Sub_id', wfactors=['condition'])
print(aov)
```

Observation	Subject	Valence	Recall
1	Jim	Neg	32
2	Jim	Neu	15
3	Jim	Pos	45
4	Victor	Neg	30
5	Victor	Neu	13
6	Victor	Pos	40
7	Faye	Neg	26
8	Faye	Neu	12
9	Faye	Pos	42
10	Ron	Neg	22
11	Ron	Neu	10
12	Ron	Pos	38
13	Jason	Neg	29
14	Jason	Neu	8
15	Jason	Pos	35

Visual Representation



Kruskal-Wallis rank sum test

We can use the Kruskal-Wallis rank sum test to compare the means of non-parametric groups

```
# Kruskal-Wallis H-test
from numpy.random import seed
from numpy.random import randn
from scipy.stats import kruskal
# seed the random number generator
seed(1)
# generate three independent samples
data1 = 5 * randn(100) + 50
data2 = 5 * randn(100) + 50
data3 = 5 * randn(100) + 52
# compare samples
stat, p = kruskal(data1, data2, data3)
print('Statistics=%.3f, p=%.3f' % (stat, p))
```

Summary

- ANOVA uses general analysis of variance to
- F-value as the main inferential statistics
- One way / two way / repeated measures

Post-Hoc Tests

Limits of ANOVA

- Analysis of variance just tells us there is at least one level that is significantly different than the other
- It does not tell us which level is different and how
- t-tests would not keep the alpha level in the confidence interval

Types of Post-Hoc Tests

- Fisher's least significant difference (LSD)
- The Bonferroni procedure
- Holm–Bonferroni method
- Tukey's procedure
- And many more…

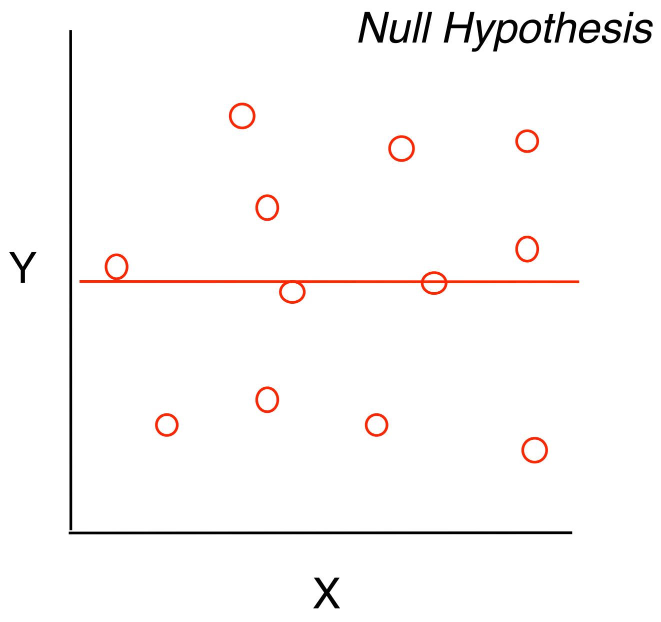
Tukey's HSD (honest significant difference)

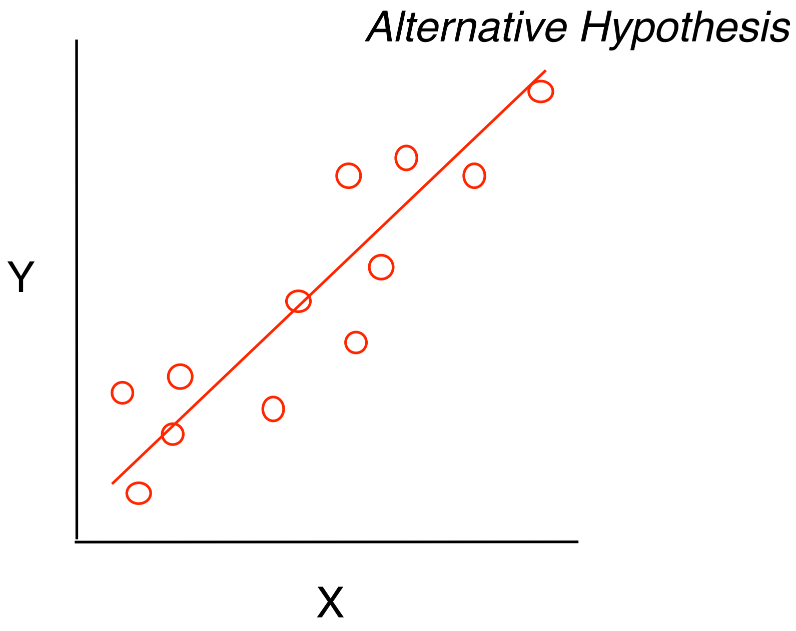
- Tukey's test is based on a formula very similar to that of the t-test, except that it corrects for family-wise error rate
- When there are multiple comparisons being made, the probability of making a Type I error within at least one of the comparisons, increases — Tukey's test corrects for that
- It is suitable for multiple comparisons than a number of t-tests would be

Correlation Tests

Correlation

- A correlation measures the "degree of association" between two variables





Correlation Tests

- Correlation: Two factors are correlated if there is a relationship between them
- For parametric data, the most common test is the Pearson's product moment correlation coefficient test.
    - Pearson's r: ranges between -1 to 1
    - Pearson's r square represents the proportion of the variance shared by the two variables

Types of Tests

- Pearson's product-moment coefficient
  - Tests linearity for parametric data, but pretty robust
  - The sample is independently and randomly drawn
  - A linear relationship between the two variables is present
  - When plotted, the lines form a line and is not curved
  - There is homogeneity of variance
- Spearman test
  - Tests non-parametric data.

Pearson Correlation

- Given paired data $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ consisting of n pairs, r_{xy} is defined as

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

- where:
  - n is sample size
  - x_i, y_i are the individual sample points indexed with i
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the the sample mean

Effect Size

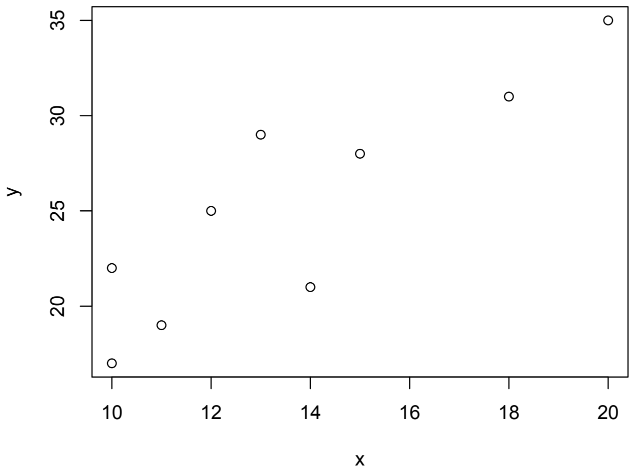
Correlation is measured in:

- “r” (parametric, Pearson’s)
- “ρ” - rho (non-parametric, Spearman’s)
- Both range in [-1,1], where 0 is no correlation

	small size	**medium size**	**large size**
Pearson's r	0.1	0.3	0.5

Example

```
df['carat'].corr(df['price'])
df['carat'].corr(df['price'], method= 'spearman')
```



Non-Parametric

```
> cor.test(x,y,method="spearman")

    Spearman's rank correlation rho

data:  x and y
S = 22.5933, p-value = 0.00789
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.8117226
```

Summary

- Multi-factor analyses
- One way / Two way
- Repeated measures
- Posthoc tests
- Correlation tests