Data Science in the Wild

Lecture 7: Analyzing Experiments

Eran Toch

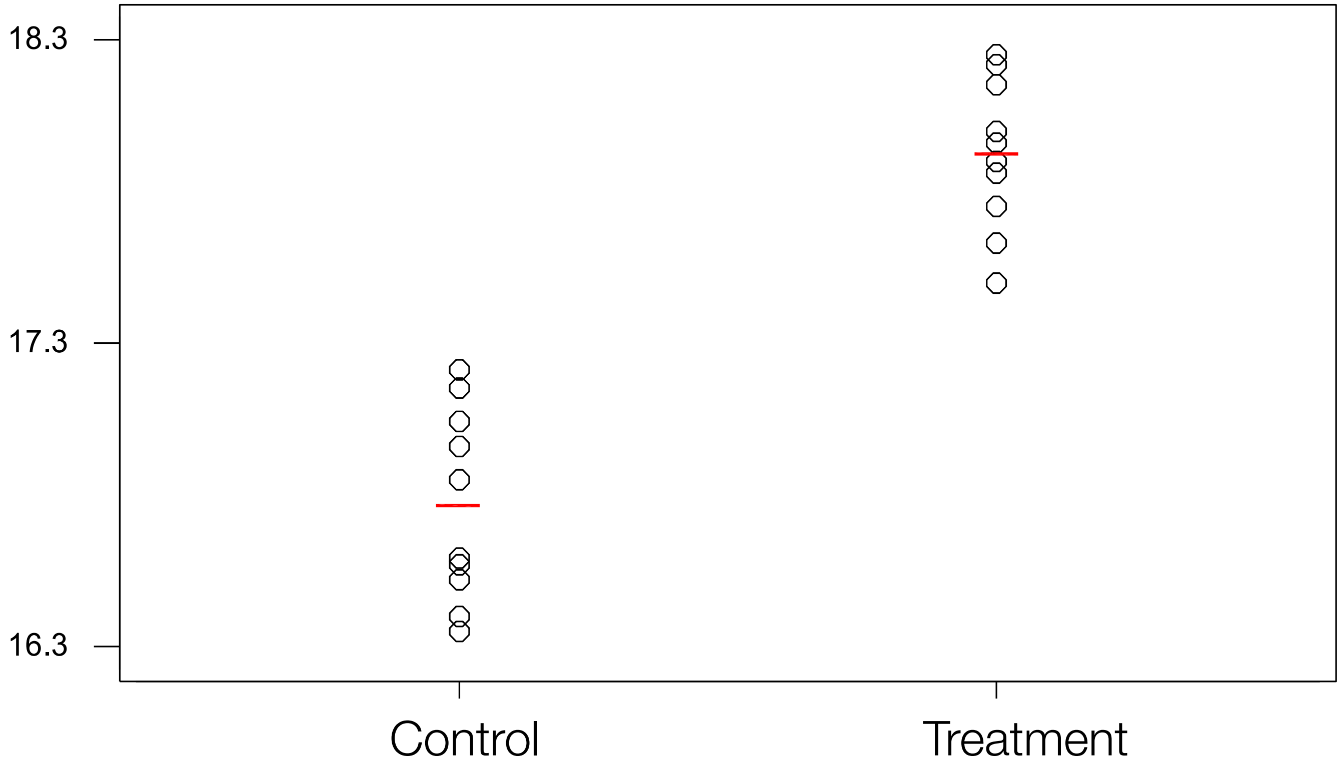
CORNELL TECH

Agenda

1. Statistical Tests and the t-Test
2. Running the t-Test
3. t-Test assumptions
4. Analyzing Inferential Statistics
5. Find the test that works for you
6. Non-Parametric Mean Comparison
7. Categorical Tests

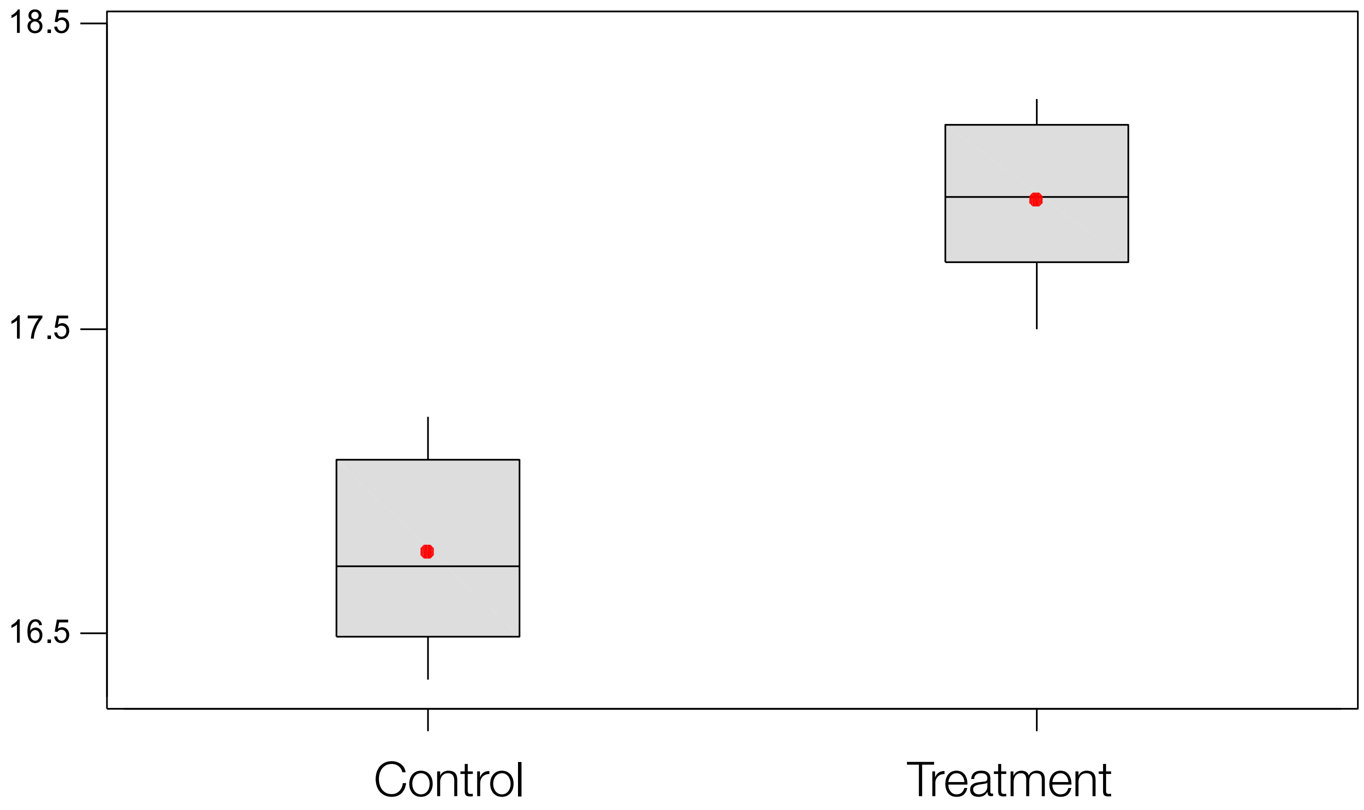
(1) Statistical Tests and the t-Test

Experiment data



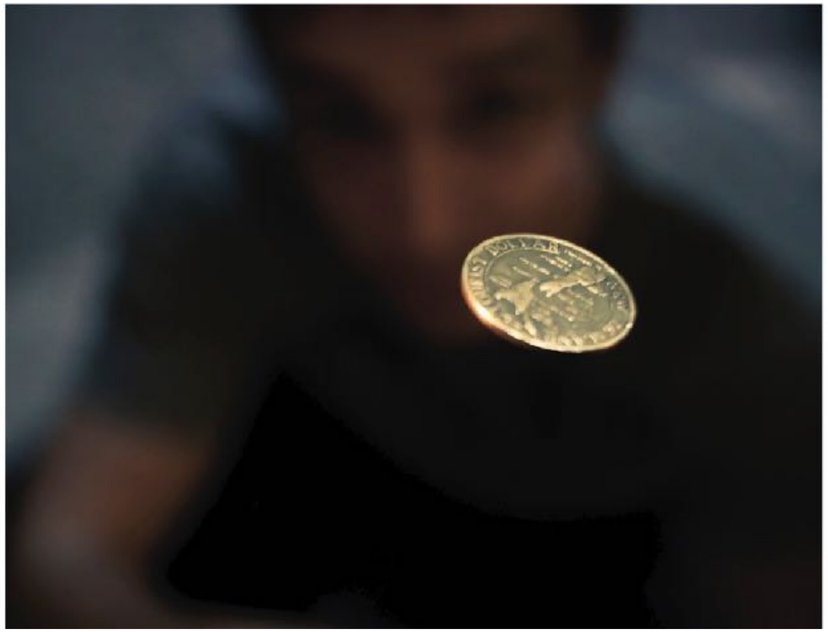
Graphical representation

Is there real difference between the means?



Statistical Tests

- How do we know that a statistical statement is correct with regard to the population?
- Is it significance or due to mere chance?
- The “chance” is the null hypothesis (H_0) and the non-chance hypothesis the alternate hypothesis (H_A)



Hypothesis testing

There are two types of errors one can make in statistical hypothesis testing:

	"accept Ho"	"reject H_0"
H_0 is true	Correct decision	Type I error
H_1 is true	Type II error	Correct decision

Too confident

Cowards

Test statistics

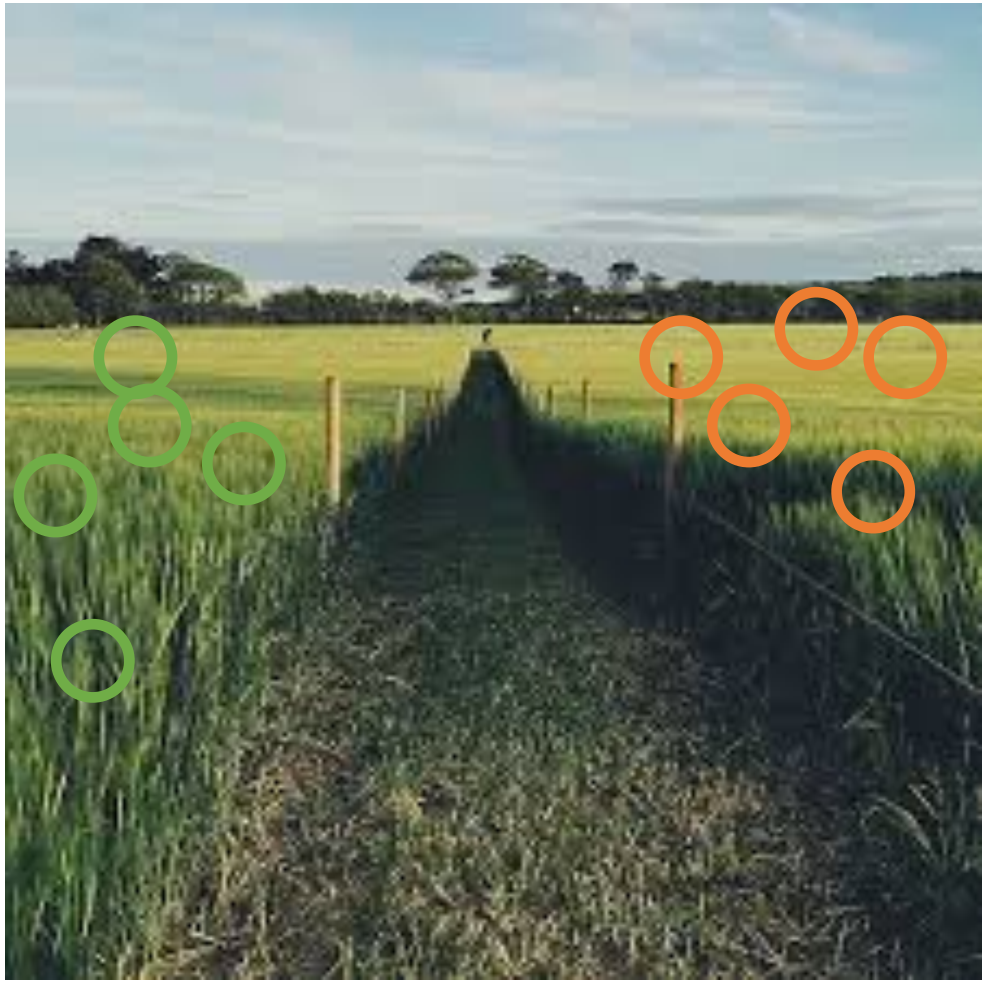
- To create a statistical test, we first need some test statistics
- It tells us the ration between signal to noise in a given statistics

William S. Gosset

A

B

Sampling

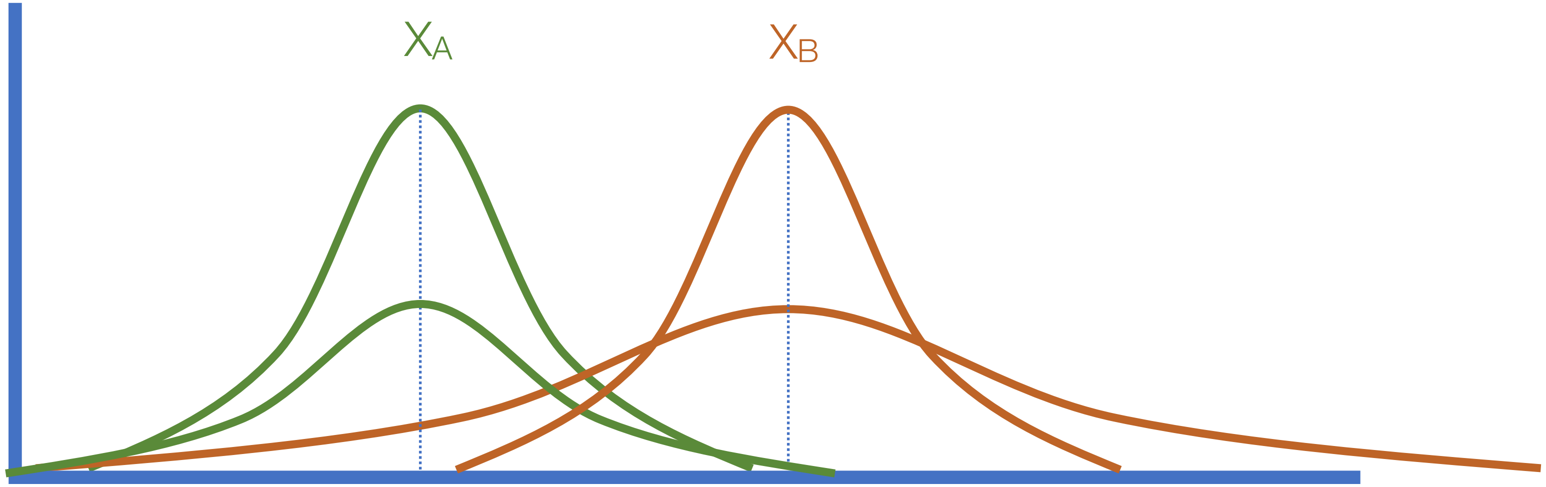


How can we infer a different in the yield of two fields from the samples alone?

T-value

A

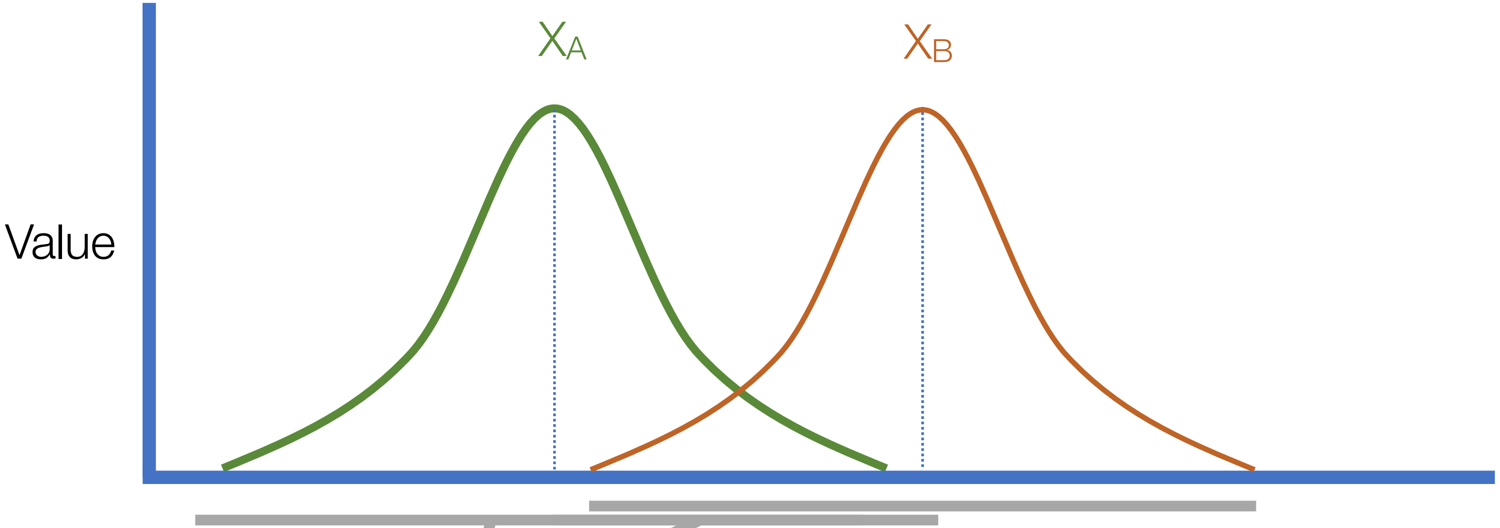
B



T-value

A

B



$$\frac{\text{Signal}}{\text{Noise}} = \frac{\text{Difference between means}}{\text{Variability}} = \frac{X_A - X_B}{\sqrt{\frac{S_A^2}{n_A} + \frac{S_B^2}{n_B}}}$$

T-Value: Intuition

- The larger the t-value, the more difference there is between groups
- The smaller the t-value, the more similarity there is between groups
- A t-value of 3 means that the groups are three times as different from each other as they are within each other
- The significance test relies on the t-value and the number of samples

Statistical tests

- After calculating a test statistic (t-value), we can use it to test whether we can reject the null hypothesis
- By comparing its value to a critical value (α)Measure of how likely the test statistic value is under the null hypothesis
  - t-value ≤ α ⇒ Reject H_0 at level α
  - t-value > α ⇒ Do not reject H_0 at level α
- In a different phrasing, we generate a p-value according to the level of t-value

Calculating the t-Value

	Area to the right (a)								
df	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
1	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.3	636.6
2	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.33	31.60
3	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.21	12.92
4	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781

- In many domains, 5% probability is an arbitrary (and problematic) cut-off for rejecting the null hypothesis
- Calculating the p-Value is based on the degrees of freedom:
  - the minimum amount of data necessary to calculate the statistics
  - Df = $n_A + n_B - 2$

Summary

- Inferential statistics
- Test statistics
- t-value
- Critical value and p-value

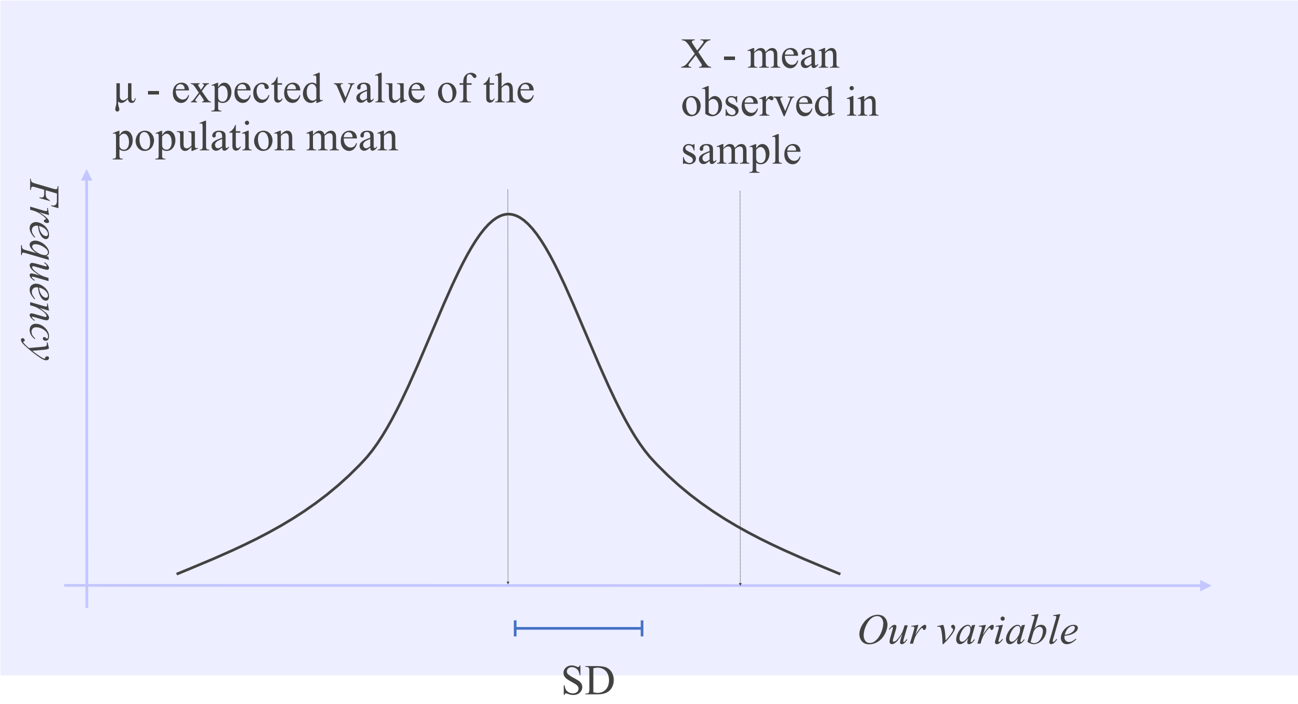
(2) Running t-Tests

Test of difference – T-Test

- t-test
  - Compares means
  - Interval or ratio variable
  - Assumes normal frequency distribution
- Types of t-tests:
  - one sample t-test: comparing a sample to a hypothetical mean
  - two independent sample t-test
  - paired t-test

1 Sided T-Test

- In a 1 sided t-test, we want to compare a value we observed to a known mean.
- We want to see if we have a new phenomenon worth reporting.



Calculating t statistics

$$t = \frac{sample\ mean - population\ mean}{standard\ error}$$

Let us assume we want to check whether our sample of gas-per-mile for various cars is different than a 23 mpg average

$$t = \frac{\bar{X} - \mu}{SD/\sqrt{n}} \quad = \frac{20.09 - 23}{6.023/\sqrt{32}} = -2.73$$

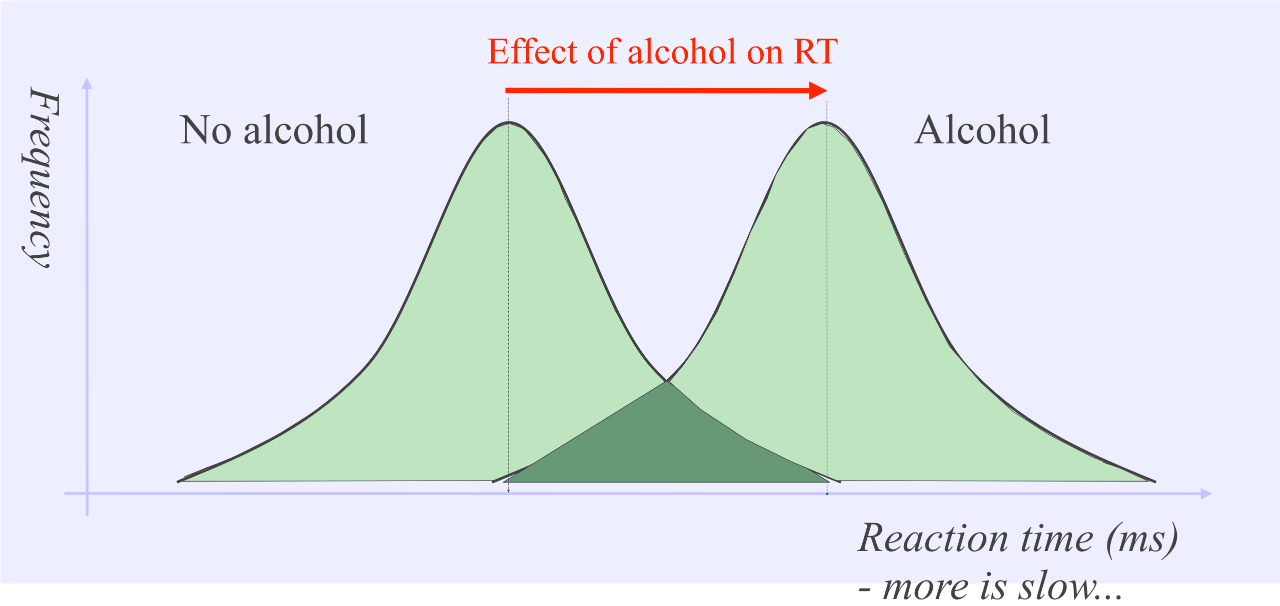
If our t-value is higher than the critical value? This is actually the t-test

Two Sample t-test

Hypothesis test: 'Alcohol' vs 'No alcohol' condition

Hypothesis true (reaction time slower in 'alcohol' condition)

Hypothesis false (reaction time faster in 'alcohol' condition)



Code Example

```
df = pd.read_csv("https://raw.githubusercontent.com/Opensourcefordatascience/
Data-sets/master//Iris_Data.csv")
setosa = df[(df['species'] == 'Iris-setosa')]
setosa.reset_index(inplace= True)

versicolor = df[(df['species'] == 'Iris-versicolor')]
versicolor.reset_index(inplace= True)

stats.ttest_ind(setosa['sepal_width'], versicolor['sepal_width'])
Ttest_indResult(statistic=9.282772555558111, pvalue=4.362239016010214 3e-15)
```

Descriptive Statistics

```
rp.summary_cont(df.groupby("species")['sepal_width'])
```

N	Mean	SD	SE	95% Conf.	Interval	
species						
Iris-setosa	50	3.418	0.381024	0.053885	3.311313	3.524687
Iris-versicolor	50	2.770	0.313798	0.044378	2.682136	2.857864

Boxplots

Boxplot grouped by species
sepal_width
4.5
4.0
3.5
3.0
2.5
2.0
Iris-setosa
Iris-versicolor
species

t-Test results

```
descriptives, results =
rp.ttest(setosa['sepal_width'],
versicolor['sepal_width'])

results
```

Independent t-test	results	
0	Difference (sepal_width - sepal_width) =	0.6480
1	Degrees of freedom =	98.0000
2	t =	9.2828
3	Two side test p value =	0.0000
4	Mean of sepal_width > mean of sepal_width p va...	1.0000
5	Mean of sepal_width < mean of sepal_width p va...	0.0000
6	Cohen's d =	1.8566
7	Hedge's g =	1.8423
8	Glass's delta =	1.7007
9	r =	0.6840

Paired vs. Unpaired

- Unpaired means that you simply compare the two groups. So, you will build a model for each group (calculate the mean and variance), and see whether there is a difference.
- Paired means that you will look at the differences between the two groups.
- In which study design paired t-test should be used?

Paired vs. Unpaired

Paired

	Subject	Before diet	After diet
Diet 1	A	100	70
	B	90	89
	C	89	70
Diet 2	D	100	101
	E	100	98
	F	90	87

Unpaired

	Subject	Weight Change
Diet 1	A	-30
	B	-1
	C	-19
Diet 2	D	+1
	E	-2
	F	-3

(3) t-Test Assumptions

Assumptions

- Independence
- Homogeneity of variance
- t-tests works only with data that distributes normally
- t-tests works best with smaller datasets
    - For larger datasets, Z-statistics is often used

Homogeneity of variance

- The independent t-test assumes the variances of the two groups measured are equal in the population
- The assumption of homogeneity of variance can be tested using Levene's Test of Equality of Variances
- The Levene’s F Test for Equality of Variances is the most commonly used statistic to test the assumption of homogeneity of variance

Levene Test

- This test for homogeneity provides a statistic and a significance value (*p*-value)
- If the p-value is greater than 0.05 (i.e., $p > .05$), the group variances can be treated as equal
- However, if $p < 0.05$, we have unequal variances and we have violated the assumption of homogeneity of variances

```
stats.levene(setosa['sepal_width'], versicolor['sepal_width'])

LeveneResult(statistic=0.66354593329432332, pvalue=0.41728596812962038)
```

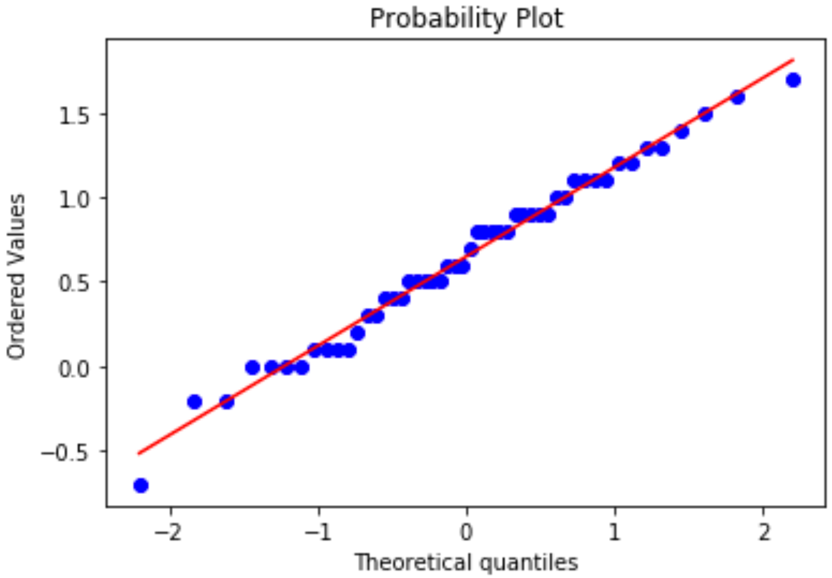
Normality Assumption

- T-tests require that the residuals needs to be normally distributed
- To calculate the residuals between the groups, subtract the values of one group from the values of the other group

```
diff = setosa['sepal_width'] - versicolor['sepal_width']
```

- Checking for normality is done with a visual comparison and with a statistical test

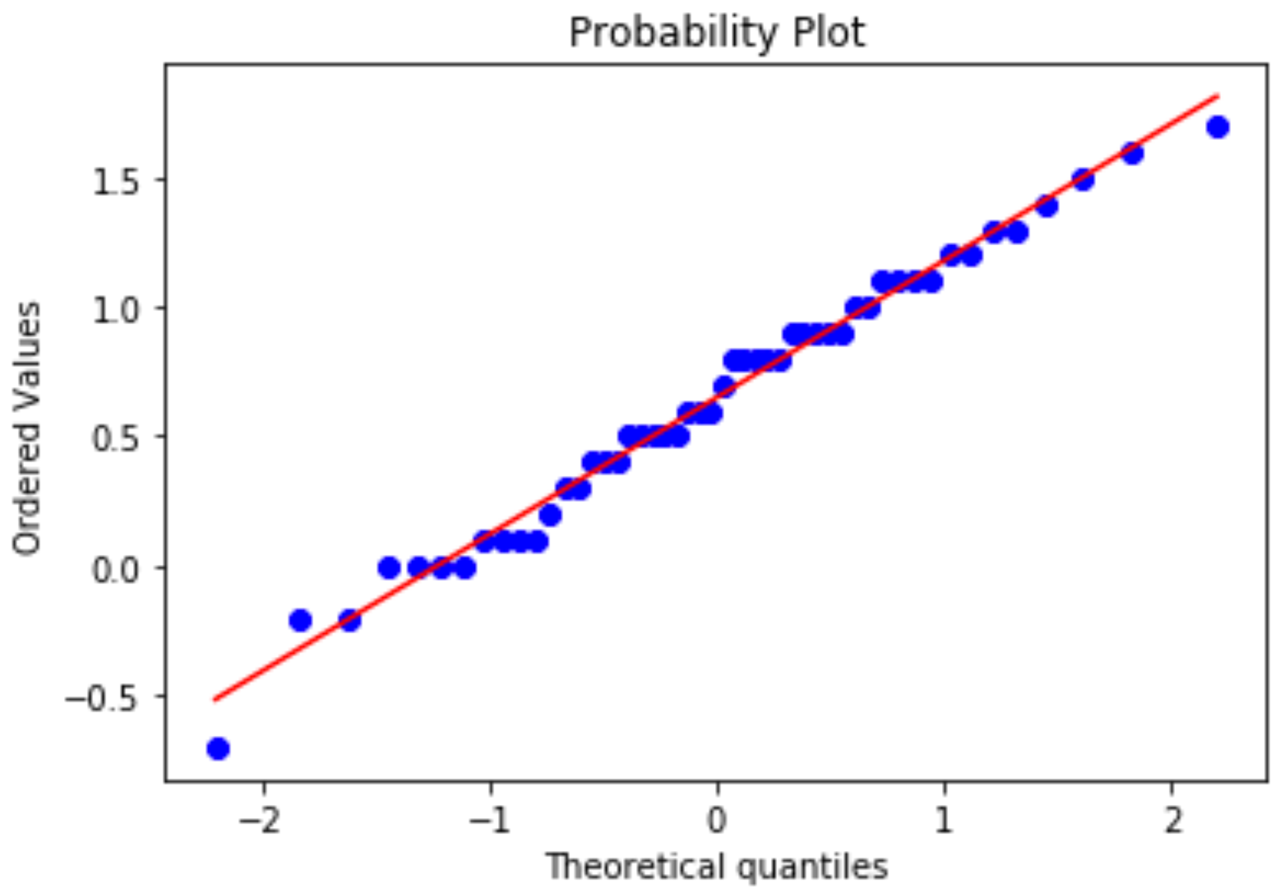
Q–Q (quantile-quantile)



- a Q–Q (quantile-quantile) plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other
- Normal data in a q-q plot will show the dots should fall on the red line. If the dots are not on the red line then it's an indication that there is deviation from normality
- Some deviations from normality is fine, as long as it's not severe

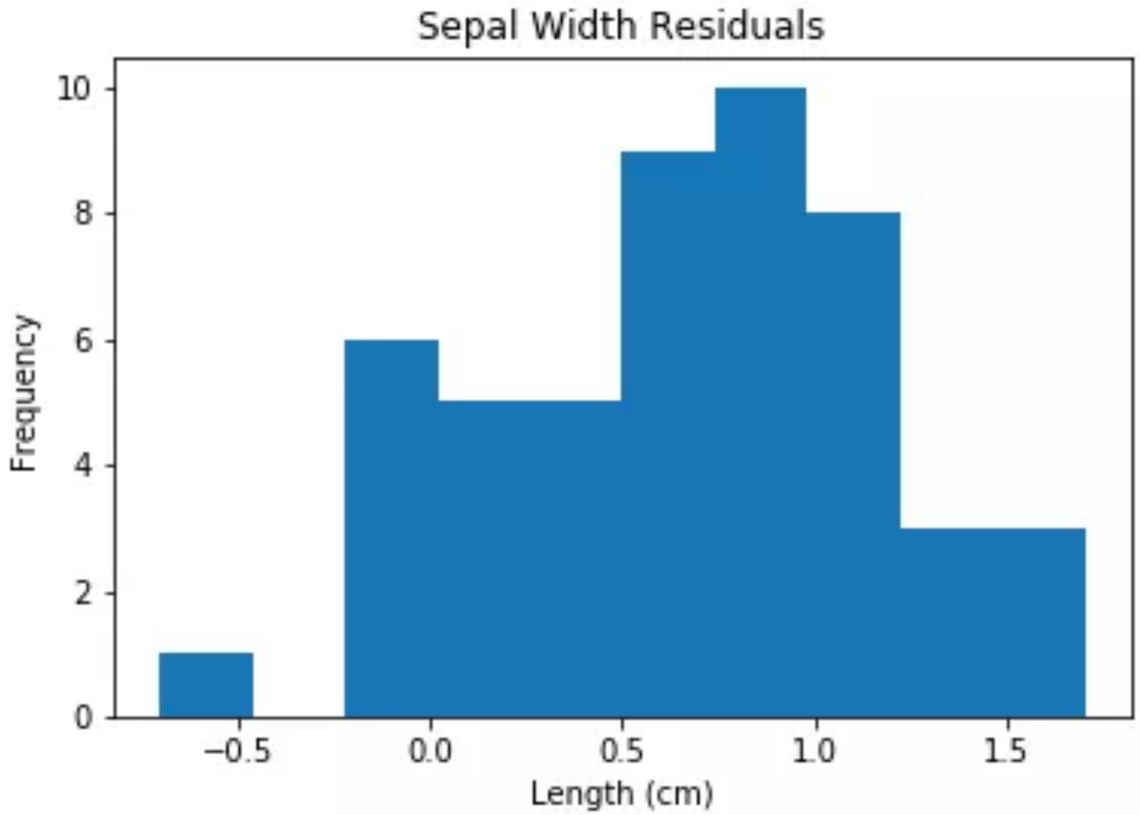
Q-Q Plot

```
import pylab
stats.probplot(diff,
dist="norm", plot=pylab)
pylab.show()
```



Histogram

```
diff.plot(kind= "hist", title=
"Sepal Width Residuals")
plt.xlabel("Length (cm)")
plt.savefig("Residuals Plot of
Sepal Width.png")
```



The Shapiro–Wilk Test

- The Shapiro–Wilk test tests the null hypothesis that a sample x_1, ..., x_n came from a normally distributed population

```
stats.shapiro(diff)
```

```
(0.9859335422515869, 0.8108891248703003)
```

- The first value is the W test statistic and the second value is the p-value
- Since the test statistic does not produce a significant p-value, the data is indicated to be normally distributed

(4) Analyzing Inferential Statistics

Effect Size

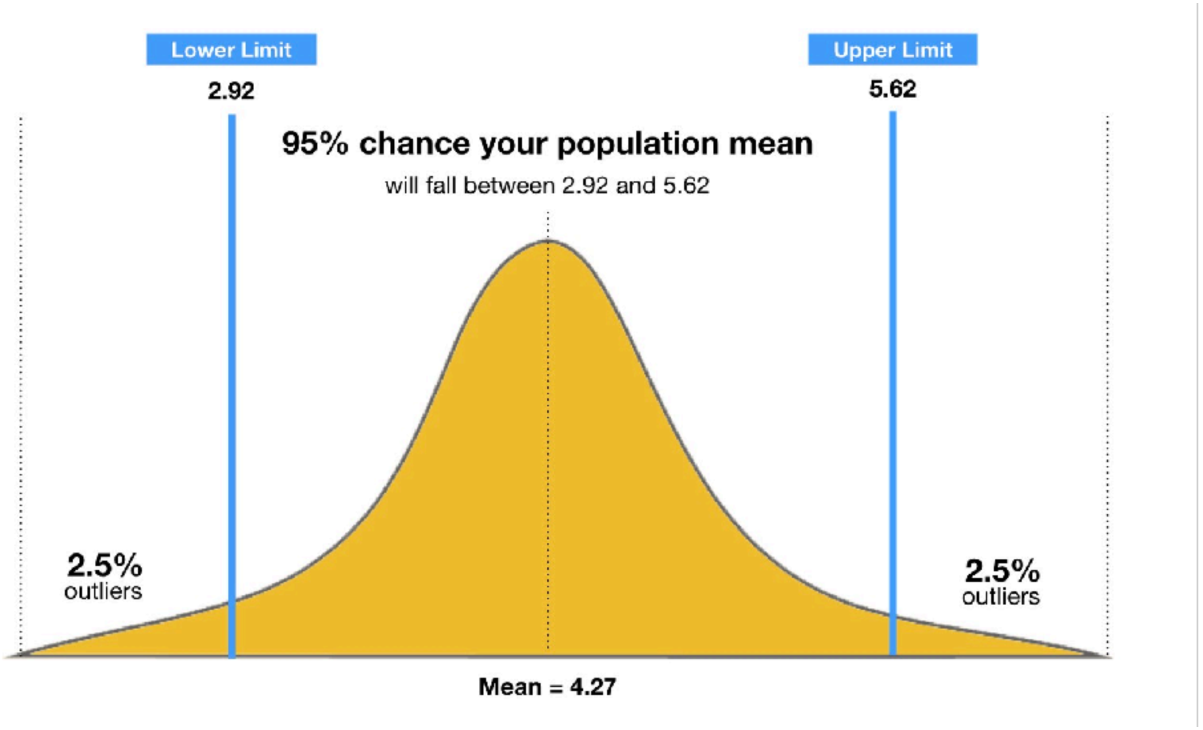
- Effect size: a measure of the size to the effect observed in the statistics
- There are many ways to determine the effect size, dependent on the assumptions about the data
- In t-tests, Cohen's d is often used
- It is determined by calculating the mean difference between your two groups, and then dividing the result by the pooled standard deviation

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s}$$

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

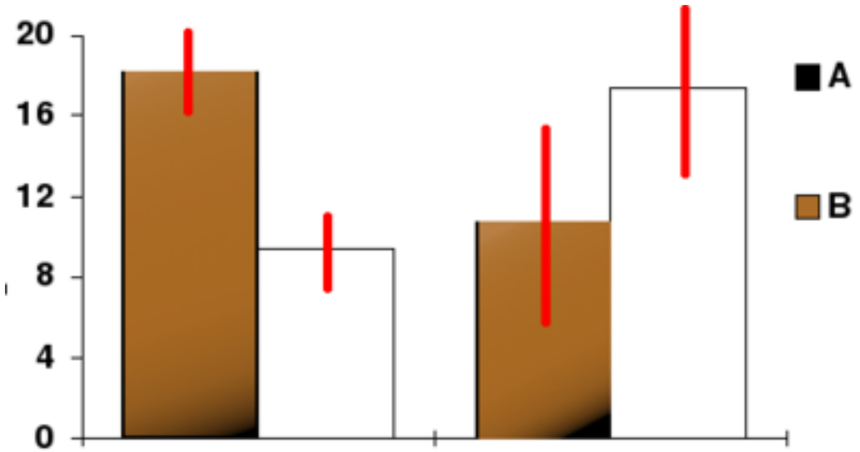
Confidence interval

An interval that contains the estimated population parameter (e.g., mean), within a certain degree of confidence (e.g., 95%)



Example

- "The results from the poll stated that the confidence level was 95% +/-3, which means that if the pool would be repeated over and over, using the same techniques, 95% of the time the results would fall within the published results."
- The 95% is the confidence level and the +/-3 is called a margin of error



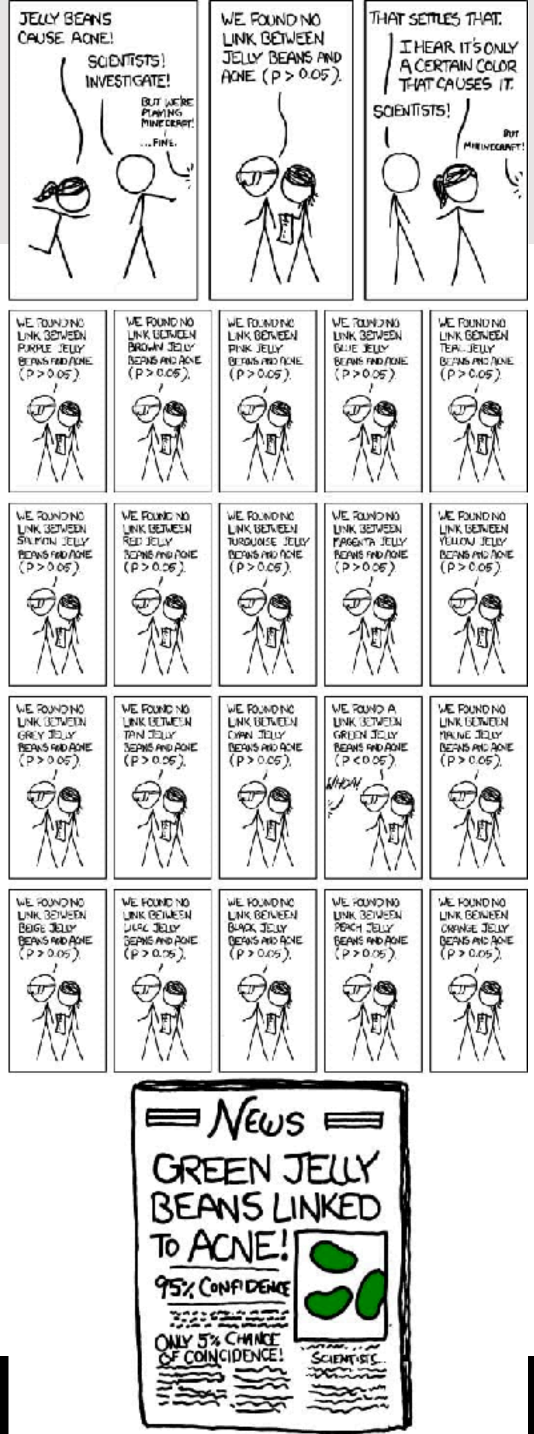
Calculating CI for a given test statistics

- t - the t-value, taken according to the critical value table (if we look for 95% confidence, we should pick the 0.05 critical value)
- s - the standard deviation
- n - the sample size

$$\overline{X} \pm t \frac{s}{\sqrt{n}}$$

Limitations of Inferential Statistics

- Criticisms against threshold-based tests:
    - A critical value of 0.05 for a p-value is totally arbitrary
    - Statistical significance is very problematic in large data sets
    - And can lead to p-Hacking



p-Hacking

1. Stop collecting data when you hit $p < 0.05$
2. Analyze many measures, but report only those with $p<.05$. 3.
3. Collect and analyze many conditions, but only report those with $p<.05$.
4. Use covariates to reach $p < 0.05$
5. Exclude participants to reach $p < 0.05$
6. Transform the data to get $p<.05$.

Leif D. Nelson, False-Positives, p-Hacking, Statistical Power, and Evidential Value

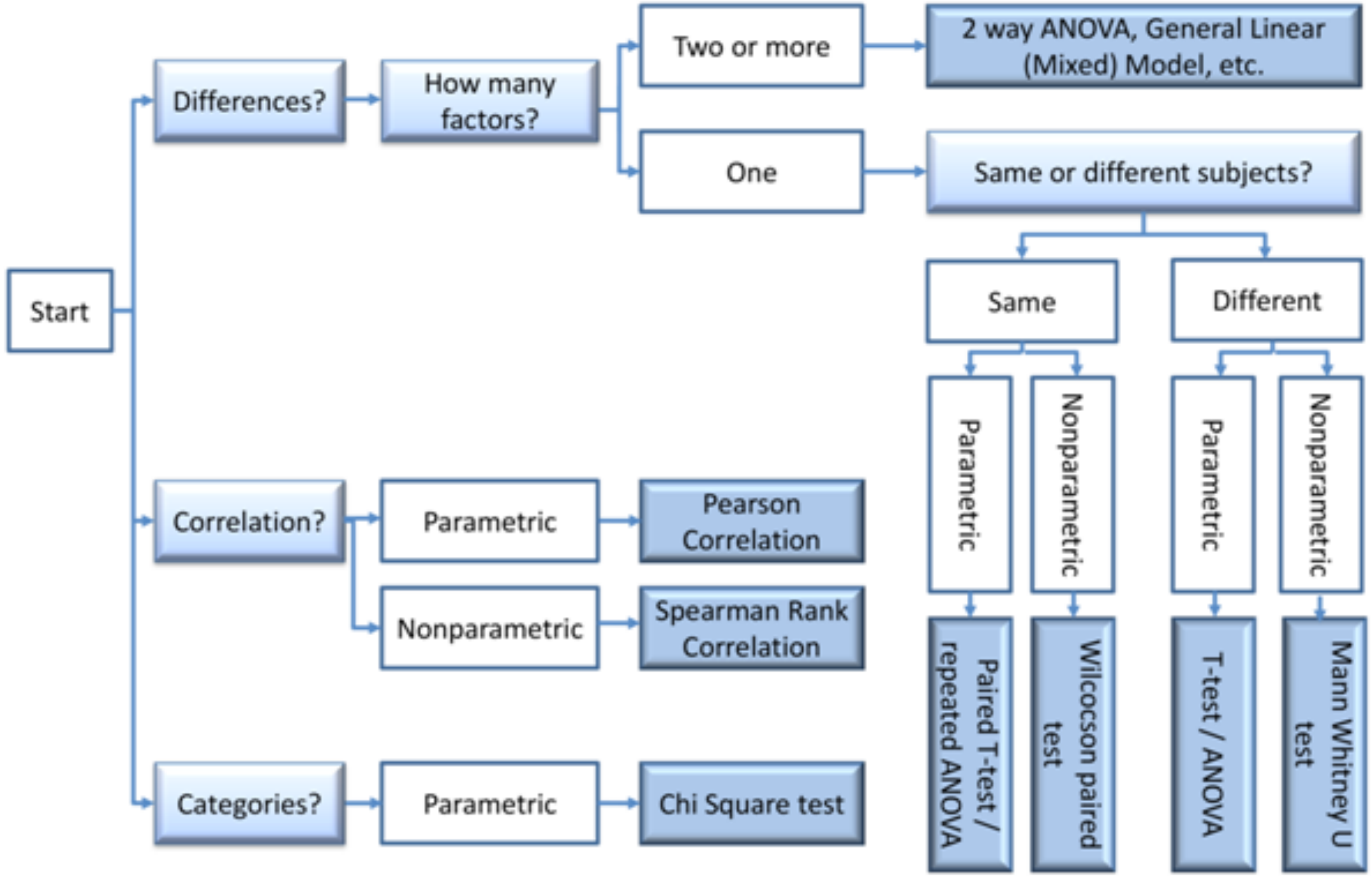
How to think about statistics

- “when a measure become a target, it is no longer a measure“ Goodhart’s law.
- Report everything to provide a better overview of the results
- Use train/test paradigm

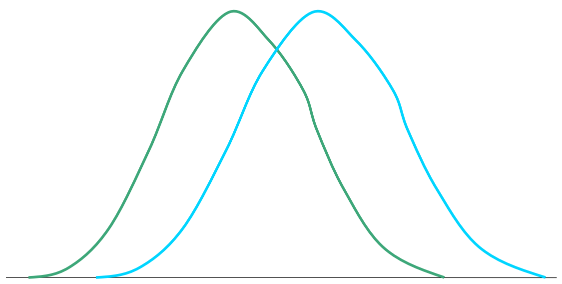
(5) Find the test that works for you

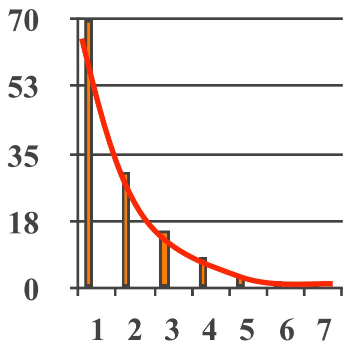
Types of Tests

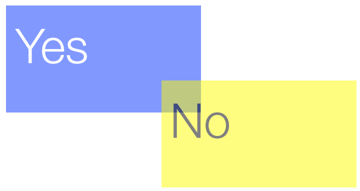
- Parametric vs. Non-Parametric
- Difference vs. Correlation
- Categorical vs. Differential
- Number of samples



Parametric vs. Non-Parametric







Parametric tests for data

- Continuous, and
- normal distribution, and
- independent

E.g., time to complete task, number of errors

Non-parametric

- Discrete, or
- non normal, or
- dependent

E.g., whether users found the system useful or not

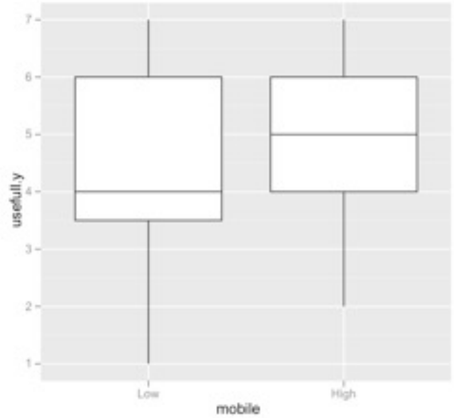
Categorical vs. Differential

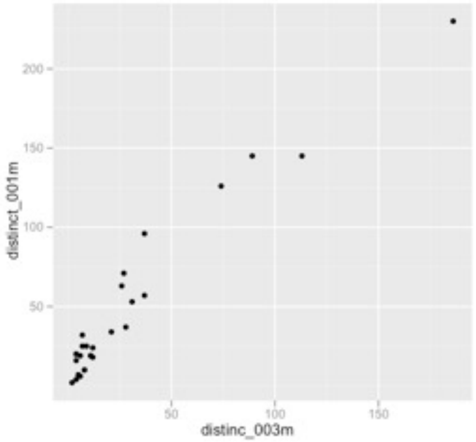
- Differences - compares two groups in terms of a ‘score’
- Frequency - compares frequency of membership of one category with another (nominal or ordinal)

	4	6	8
0	3	4	12
1	8	3	2

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225	105	2.76	3.460	20.22	1	0	3	1

Difference vs. Correlation





Difference

- Finding differences between variables
- Using tests for differences between means, variance, distribution

Correlation

- Finding relations between variables
- Using tests for correlation & regressions

(6) Non-Parametric Mean Comparison

Mann–Whitney *U* test

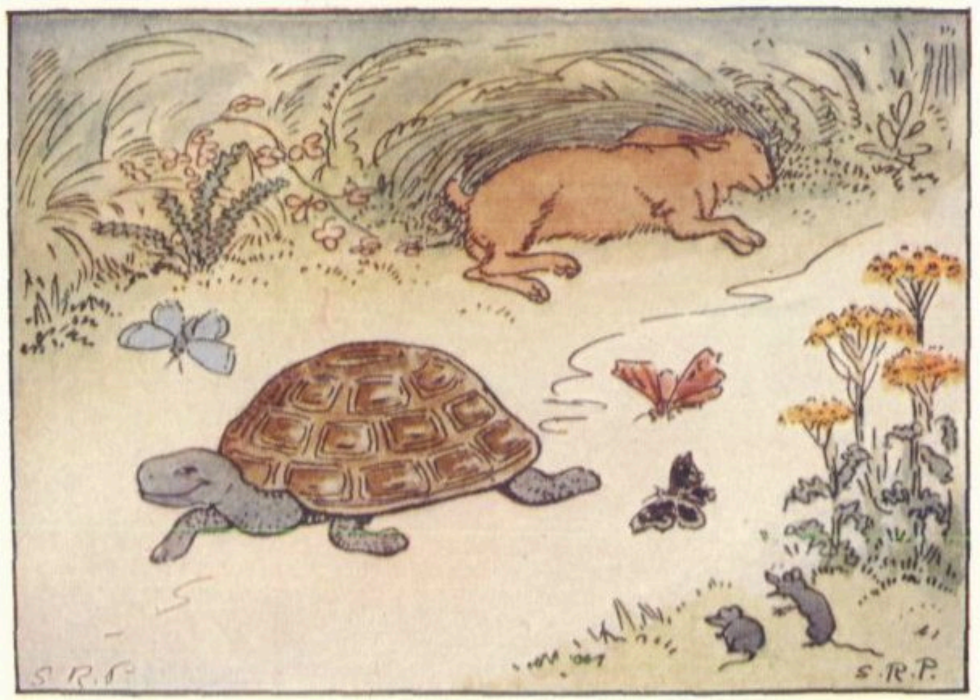
- The Mann–Whitney *U* test (aka Wilcoxon Rank-Sum test) relaxes many of the t-test assumptions
- Used to compare one or two samples of non-parametric independent values
    - All the observations from both groups are independent of each other
    - The responses are ordinal (i.e., one can at least say, of any two observations, which is the greater)
    - Under the null hypothesis H0, the distributions of both populations are equal
    - The alternative hypothesis H1 is that the distributions are not equal
- A similar nonparametric test used on dependent samples is the Wilcoxon signed-rank test

Calculating the U value

- For each observation in one set, U is the the number of times this first value wins over any observations in the other set
- Count 0.5 for any ties
- The sum of wins and ties is U for the first set
- U for the other set is the converse
- It's a little more complicated for larger sets

Classic example

- Suppose we want to see if tortoises win over hares
- This is the in which they reach the finishing post (their rank order, from first to last crossing the finish line) is as follows, writing T for a tortoise and H for a hare:
    - T H H H H H T T T T T H
- Tortoises win at: 6, 1, 1, 1, 1, 1, so $U_T = 11$
- For Hares, the wins are: 5, 5, 5, 5, 5, 0, so $U_H = 25$
- Is $U_H > U_T$? That depends on the statistical test…



THE TORTOISE AND THE HARE

Why shouldn't we compare medians?

- H H H H H H H H H T T T T T T T T T **T H** H H H H H H H H H T T T T T T T T T
- The median tortoise is faster than the median hare
- But $U_H = 19*9 + 10*9 = 261$ and $U_T = 100$
- The U value reflects skewness and not just variance

Running Mann–Whitney *U* test

```
import scipy.stats

# u : Mann-Whitney test statistic
# p : p-value
u, p = scipy.stats.mannwhitneyu(x, y)
```

(7) Categorical Tests

Categorical Tests

These tests are for summaries of categorical (nominal) data:

		Behaviour:		
		No	*Yes*	*Total:*
Gender:	*Male*	**60**	**120**	180
	Female	**70**	**20**	90
	Total:	130	140	270

$\chi^2 = 6.5$, $p=0.011$

χ^2 Test

- The Chi-square test is intended to test how likely it is that an observed distribution is due to chance
- It is also called a "goodness of fit" statistic, because it measures how well the observed distribution of data fits with the distribution that is expected if the variables are independent
- Thus, if we have 40 observations and four categories or groups, we expect 10 observations in each group

The χ^2 Value

- Where:
  - Oi - Observed Data
  - Ei - Expected Values
- The null hypothesis is that there is no statistical significance between the observed and the expected

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Example

	Vegan	Vegitatian	Total
Male	20 (25)	30 (25)	50
Female	30 (25)	20 (25)	50
Total	50	50	100

χ^2 = ((20-25)^2/25) + ((30-25)^2/25) + ((30-25)^2/25) + ((20-25)^2/25) = (25/25) + (25/25) + (25/25) + (25/25) = 4

DF = (r-1)(c-1)
Where
DF = Degree of freedom
r = number of rows
c = number of columns

Critical values of the Chi-square distribution with *d* degrees of freedom

	Probability of exceeding the critical value						
d	0.05	0.01	0.001	*d*	0.05	0.01	0.001
1	3.841	6.635	10.828	11	19.675	24.725	31.264
2	5.991	9.210	13.816	12	21.026	26.217	32.910
3	7.815	11.345	16.266	13	22.362	27.688	34.528
4	9.488	13.277	18.467	14	23.685	29.141	36.123
5	11.070	15.086	20.515	15	24.996	30.578	37.697
6	12.592	16.812	22.458	16	26.296	32.000	39.252
7	14.067	18.475	24.322	17	27.587	33.409	40.790
8	15.507	20.090	26.125	18	28.869	34.805	42.312
9	16.919	21.666	27.877	19	30.144	36.191	43.820
10	18.307	23.209	29.588	20	31.410	37.566	45.315

INTRODUCTION TO POPULATION GENETICS, **Table D.1**

When to use χ^2

- The samples are taken independently or are unpaired
  - If not, use McNemar's test.
- If the sample is really small (<50), use Fisher's exact test

Summary

- Inferential Statistics
- T-tests
- Statistical tests zoo:
  - Parametric vs. Non Parametric
  - Categorical vs. Nominal
  - Pairs vs. Unpaired