Human-Computer Interaction

Statistics II

Introduction to Inferential Statistics

Professor Bilge Mutlu

Today's Agenda

» Topic overview: *introduction to inferential statistics*

» Hands-on activity

Recap: *Why do we need to use statistics?*

Statistical methods enable us to analyze quantitative data, specifically (1) to inspect data quality and characteristics and (2) to discover relationships (e.g., causal) among experimental variables or to estimate population characteristics.

1 ↠ **Descriptive** statistics

2 ↠ **Inferential** statistics

Recap: *What is the difference between **descriptive** and **inferential** statistics?*

A **descriptive statistic** is a summary statistic that quantitatively describes or summarizes features of collected data, while **descriptive statistics** is the process of using and analyzing those statistics.[1]

Inferential statistics, or statistical inference (or modeling), is the process making propositions about a population using data drawn from the population through sampling.[2]

Simply put, using descriptive statistics, we summarize a sample of data; using inferential statistics, we make propositions about the population.

[1] Wikipedia: Desciptive Statistics

[2] Wikipedia: Inferential Statistics

Recap: *When do we use descriptive and inferential statistics?*

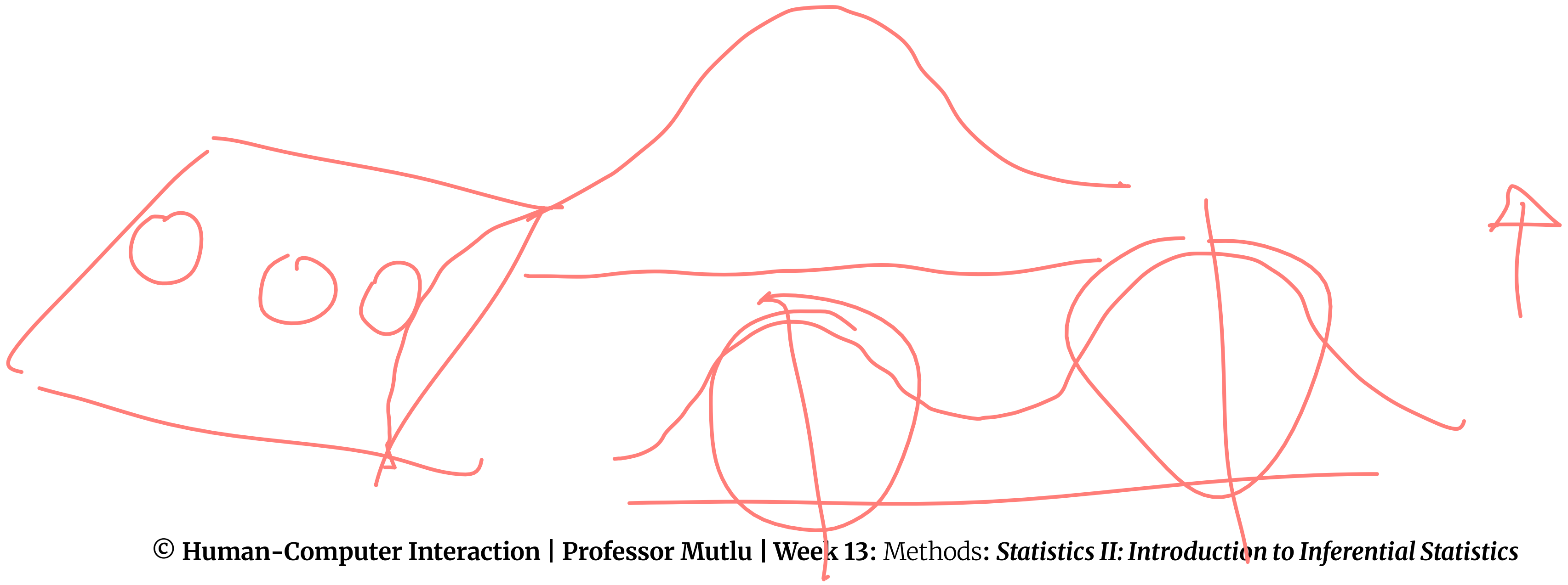
Usually, descriptive and inferential statistics are used together.

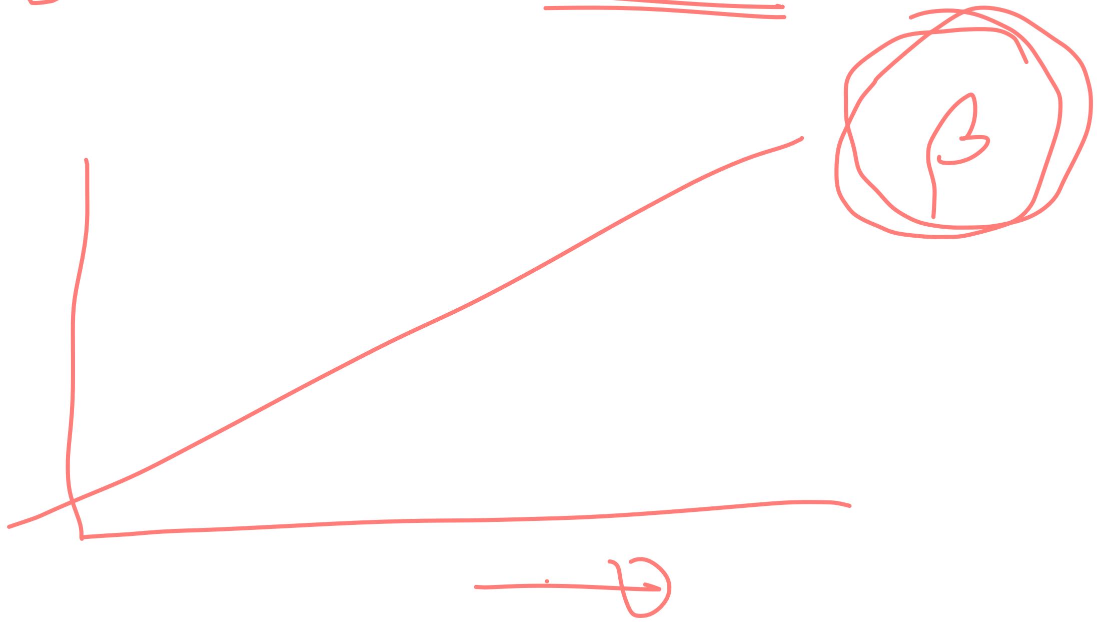
Descriptive statistics:

- To assess data quality and structure
- To describe population characteristics
- To assess dependence among variables

Inferential statistics:

- To test hypotheses
- To estimate parameters
- To perform clustering or classification





*How do we apply **inferential statistics**?*

Inferential statistics involves families of **statistical tests** that aim to establish *statistically significant* differences between distributions.

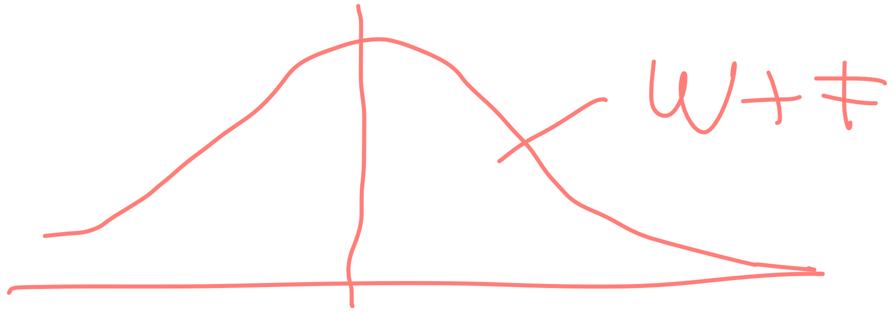
What is a statistical test?

Definition: A statistical test is a mechanism for assessing whether data provides support for particular hypotheses.

How do we test a hypothesis?

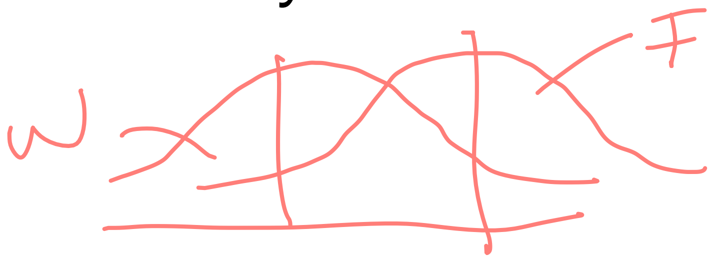
Hypotheses are provisional statements about relationships among concepts. In hypothesis testing, we seek to determine *which* statement data is consistent with.

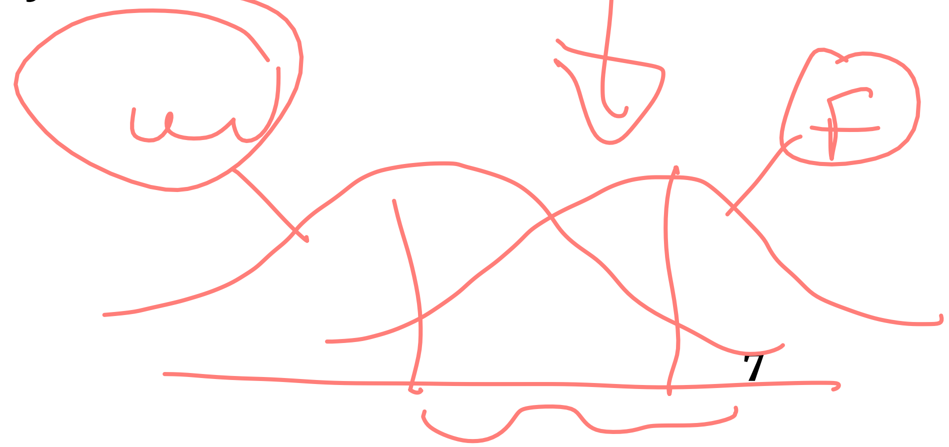
How many hypotheses do we have consider?



Two mutually exclusive hypotheses/statements about a population:

1. **Null Hypothesis**: Denoted by H_0, it states that a population parameter (e.g., the mean) is equal to a hypothesized value.
2. **Alternative Hypothesis** (or Research Hypothesis): Denoted by H_1 or H_A, it states that the population parameter is smaller, greater, or simply different than the hypothesized value in the null hypothesis.
   - **One-sided hypothesis**: H_1 where the population parameter differs in a particular direction, e.g., higher or lower.
   - **Two-sided hypothesis**: H_1 where the population parameter simply differs in a nondirectional way.





Can you identify what type of hypotheses these are?

- The SUS scores of Google Maps and Apple Maps will not differ.

  Null, not commonly done

- The usability of Google Docs and Microsoft Word will be rated differently by users.

- Users will file their taxes faster using TurboTax 2020 than they will using TurboTax 2019.

- Users will reach targets faster using a mouse than a joystick and fastest using a touchpad.

  H_{A1} – M > J

  H_{A2} – T > M, T > J

So how do we determine what test to use?

The appropriate test for a given hypothesis-testing scenario is determined by the *data types* of the **input** and **output** variables.

Recap: Data types include:

- Nominal
- Ordinal
- Interval
- Ratio

The distribution of internal and ratio data can be *normal* or *non-normal*.

	Nominal	Categorical (2+)	Ordinal	Quantitative Discrete	Quantitative Non-Normal	Quantitative Normal
Nominal	Chi-squared, Fisher's	Chi-squared	Chi-squared Trend, Mann-Whitney	Mann-Whitney	Mann-Whitney, log-rank *	Student's *t*
Categorical (2+)	Chi-squared	Chi-squared	Kruskal-Wallis**	Kruskal-Wallis**	Kruskal-Wallis**	ANOVA***
Ordinal	Chi-squared Trend, Mann-Whitney	*****	Spearman rank	Spearman rank	Spearman rank	Spearman rank, linear regression
Quantitative Discrete	Logistic regression	*****	*****	Spearman rank	Spearman rank	Spearman rank, linear regression
Quantitative Non-Normal	Logistic regression	*****	*****	*****	Plot data-Pearson, Spearman rank	Plot data-Pearson, Spearman rank & linear regression
Quantitative Normal	Logistic regression	*****	*****	*****	Linear regression****	Pearson, linear regression

Footnotes

Rows are *input* variables, columns are *output* variables.[3]

∗ If data are censored.

** The Kruskal-Wallis test is used for comparing ordinal or non-Normal variables for more than two groups, and is a generalisation of the Mann-Whitney U test. The technique is beyond the scope of this book, but is described in more advanced books and is available in common software (Epi-Info, Minitab, SPSS).

*** Analysis of variance is a general technique, and one version (one way analysis of variance) is used to compare Normally distributed variables for more than two groups, and is the parametric equivalent of the Kruskal-Wallis test.

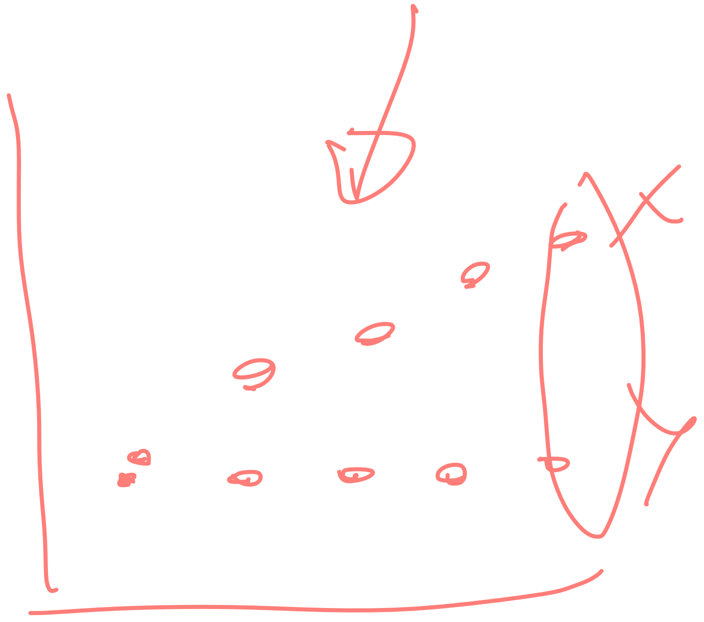
[3] Hinton, 2014, Statistics explained

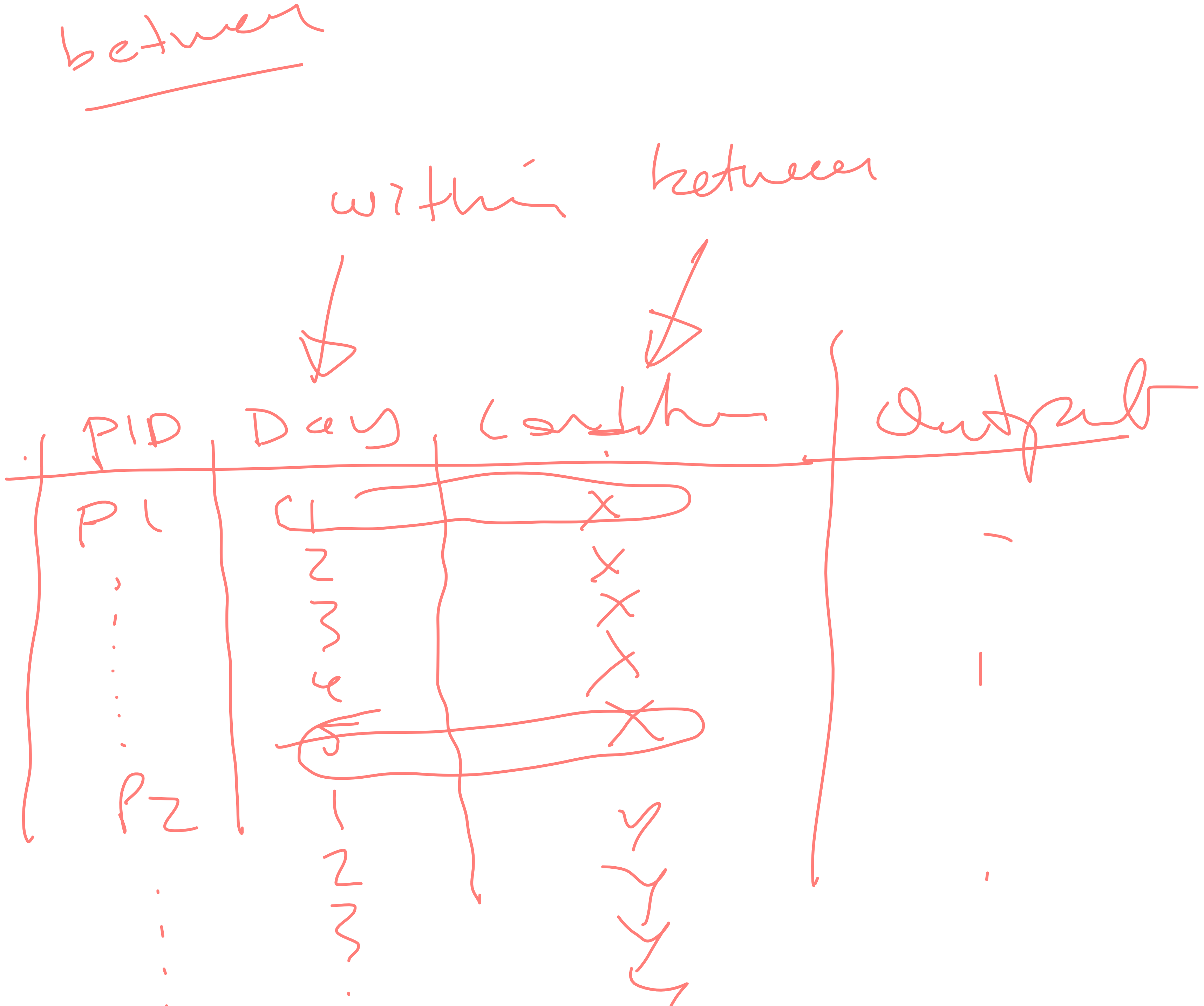
**** If the outcome variable is the dependent variable, then provided the residuals (see) are plausibly Normal, then the distribution of the independent variable is not important.

***** There are a number of more advanced techniques, such as Poisson regression, for dealing with these situations. However, they require certain assumptions and it is often easier to either dichotomise the outcome variable or treat it as continuous.

Which methods will we cover in this class?

- χ^2
- Student's t
- ANOVA
- Regression





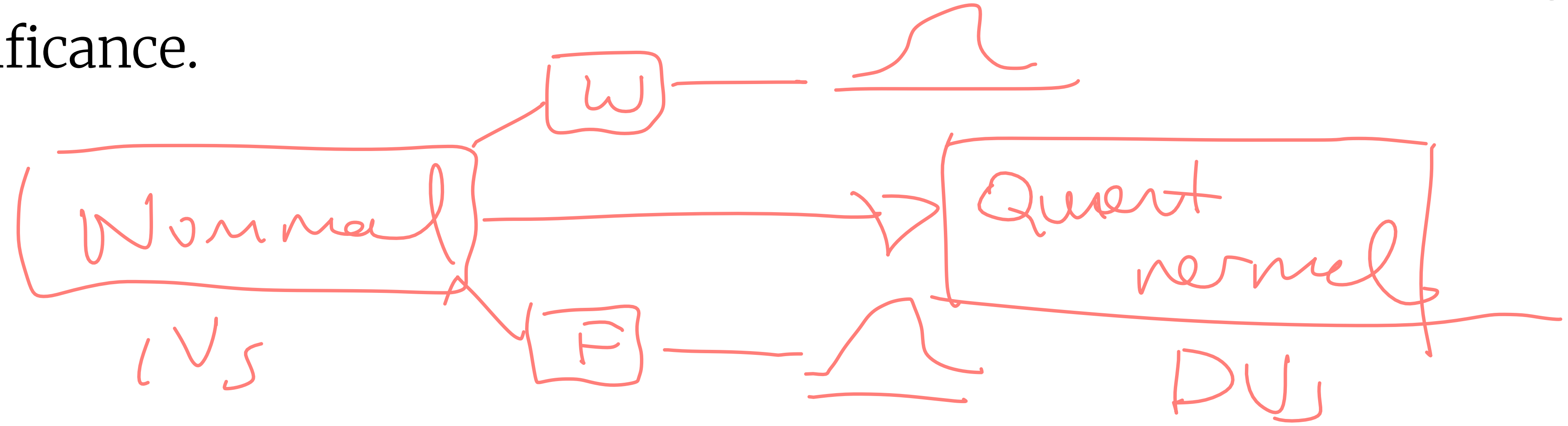
How do we conduct a t-test?

The *Student's t-test* assesses whether the means of two groups are statistically different.

What does it mean for something to be statistically significant?

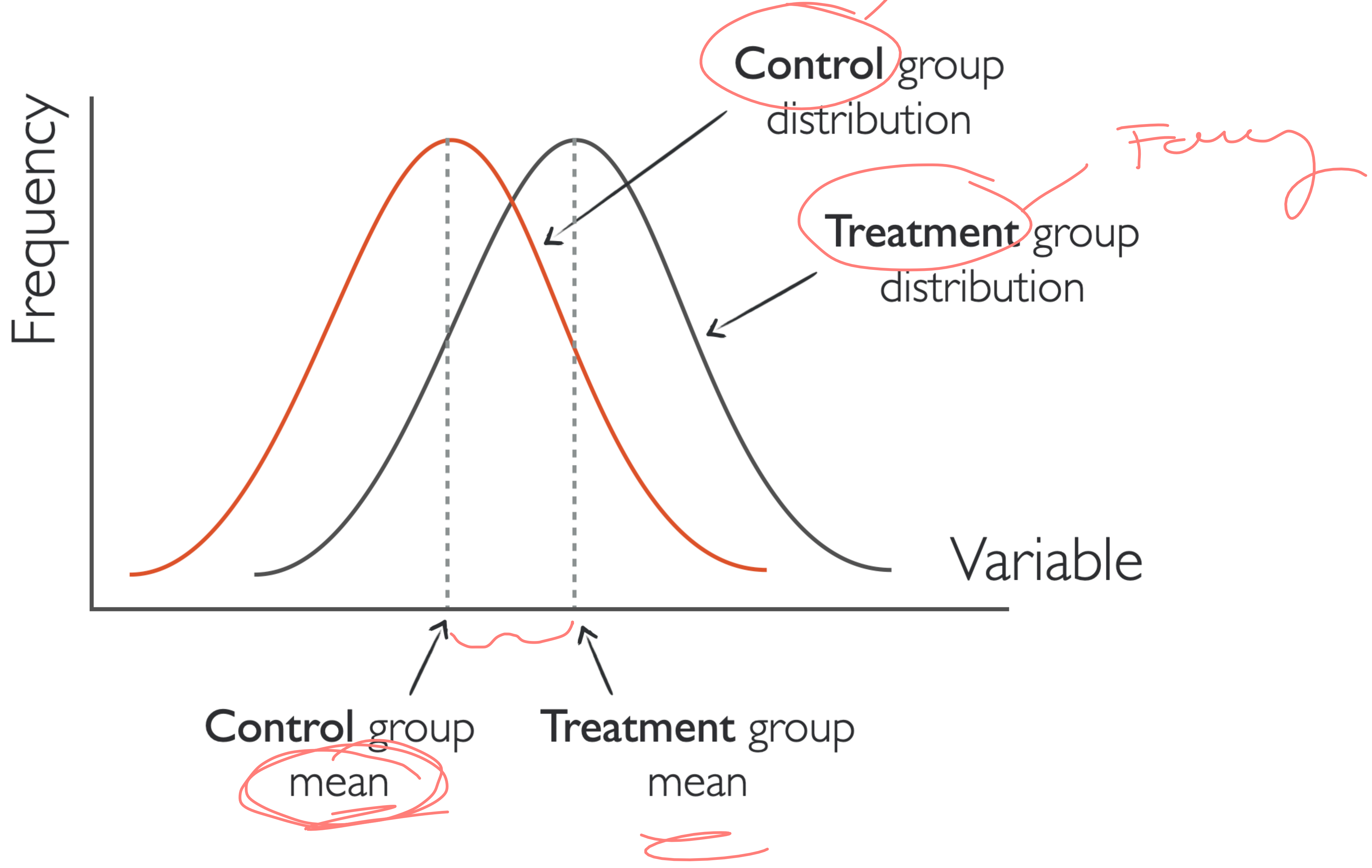
When a difference is *statistically significant*, the likelihood of it occurring by change is low, determined by a margin, called α level.

In HCI research, $\alpha = .05$ is used, thus the probability, p, that the difference is occurring by change has to be $p > .05$ for significance.

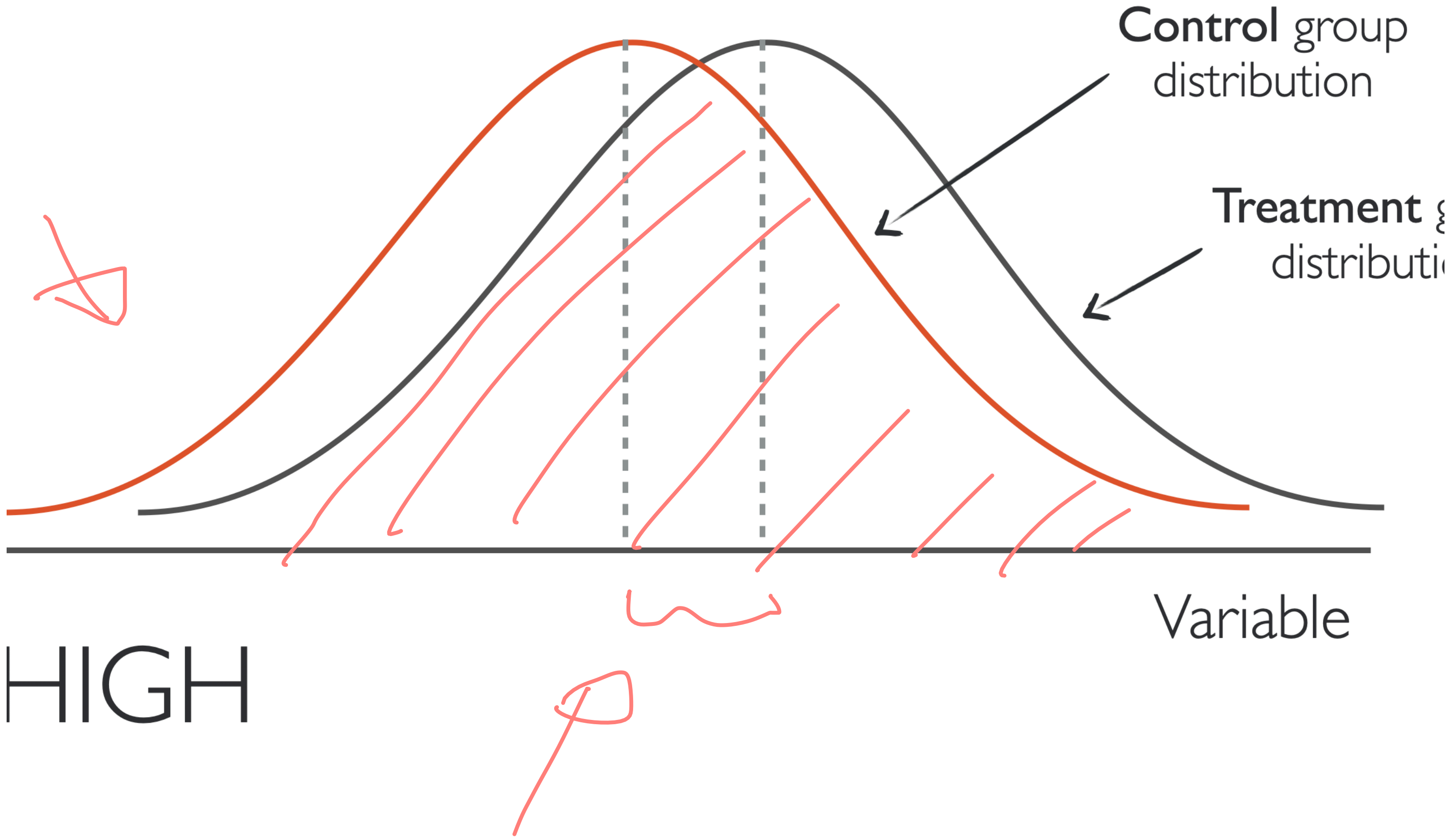


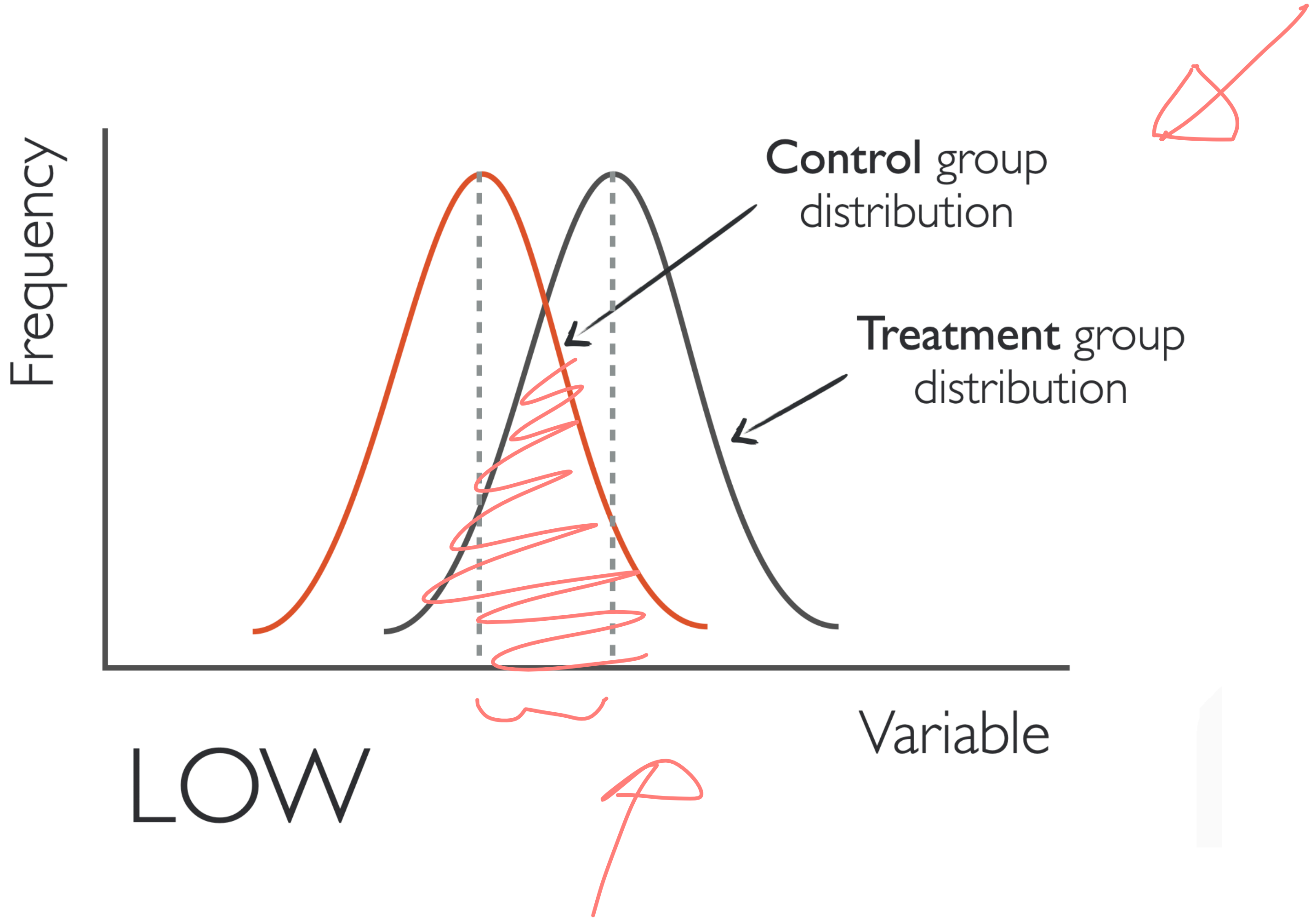
So, how do we conduct a t-test?

We look at two things: *difference in means* and *variability*.



Which two distributions are more likely to be statistically significant?

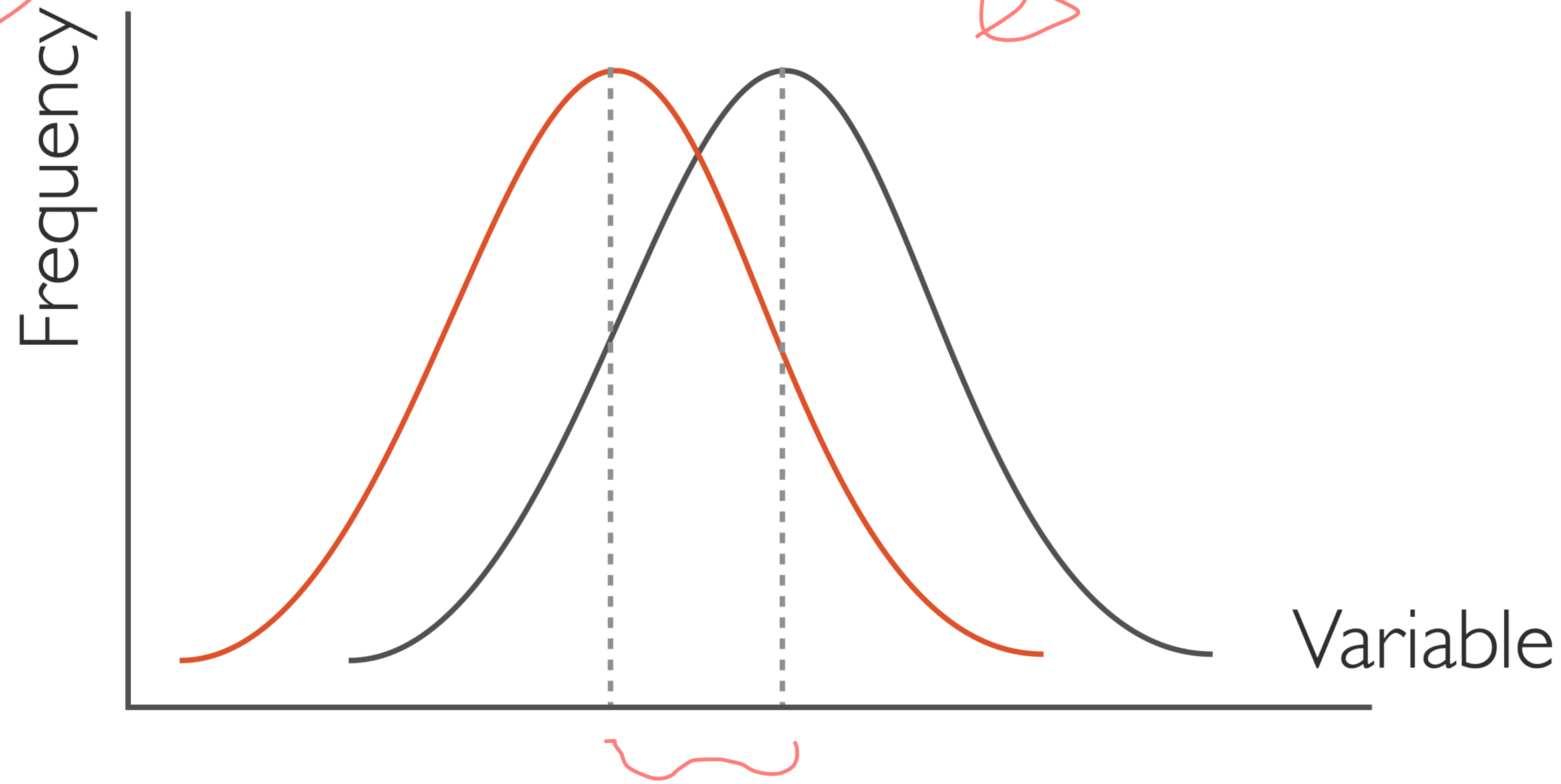




We need to calculate the t-statistic:

$$t = \frac{signal}{noise} = \frac{difference}{variability} = \frac{\mu_t - \mu_c}{\sqrt{\frac{\sigma_t}{n_t} + \frac{\sigma_c}{n_c}}}$$

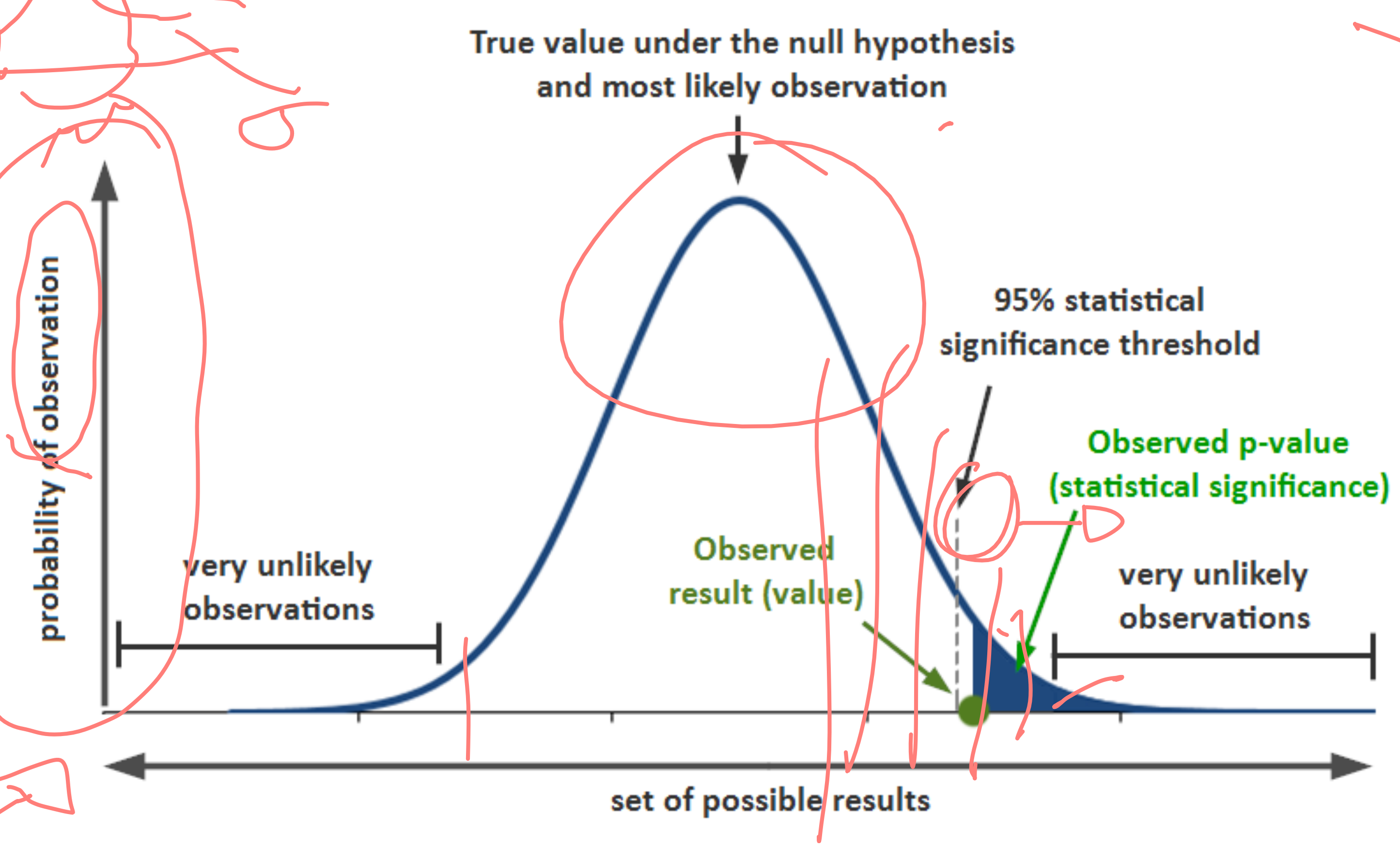
μ_t and σ_t are mean and variance of the treatment group, μ_c and σ_c are mean and variance of the control group.

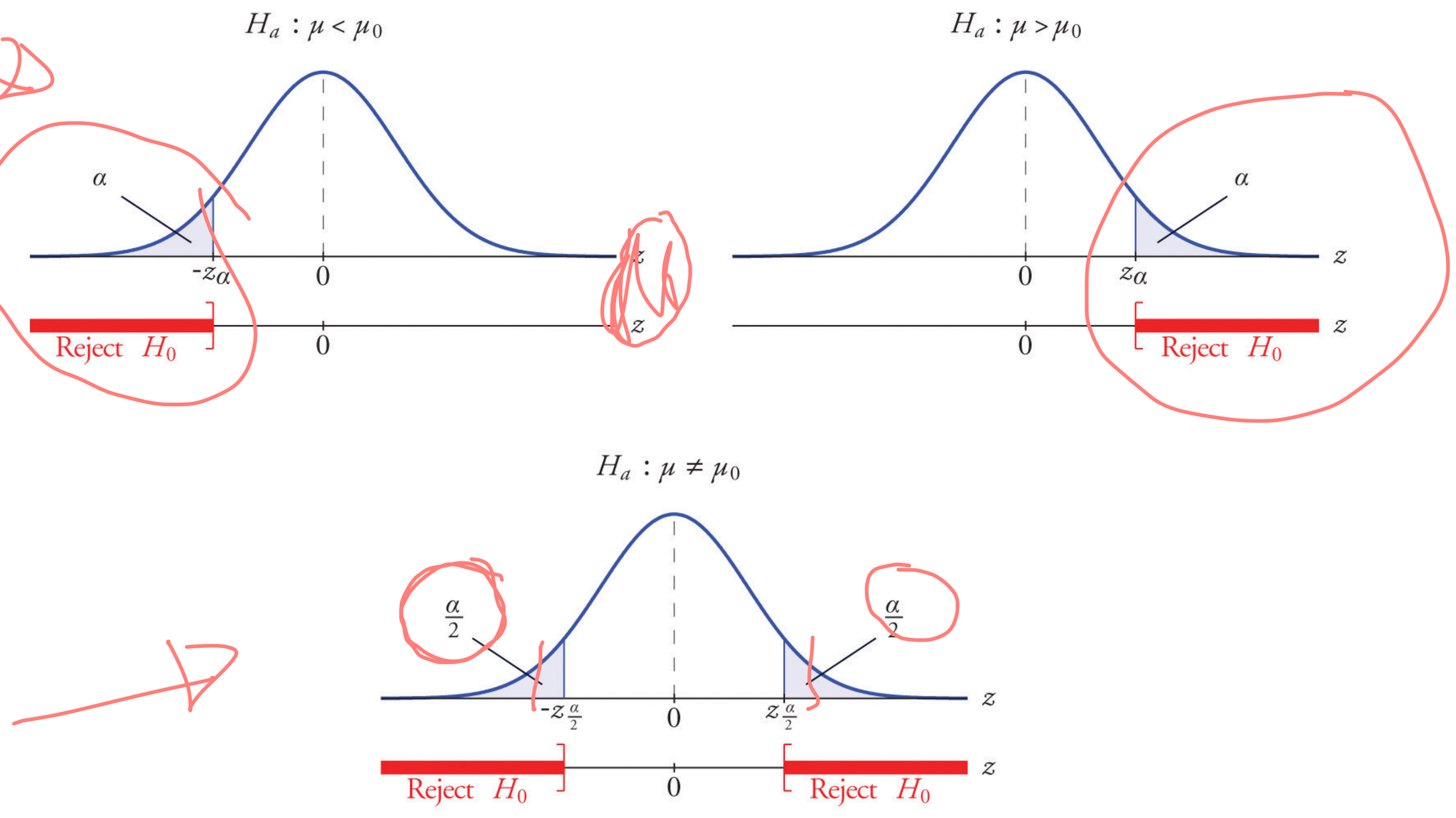


ANOVA = explained variance / unexpl. variance

The *t*-test will return the values of: (1) a **t-statistic** that will indicate signal/noise ratio, and (2) a **p-value** that indicates significance.

In *one*- and *two-tailed* tests, the p-value is interpreted differently.[9]





[9] Image sources: left, right

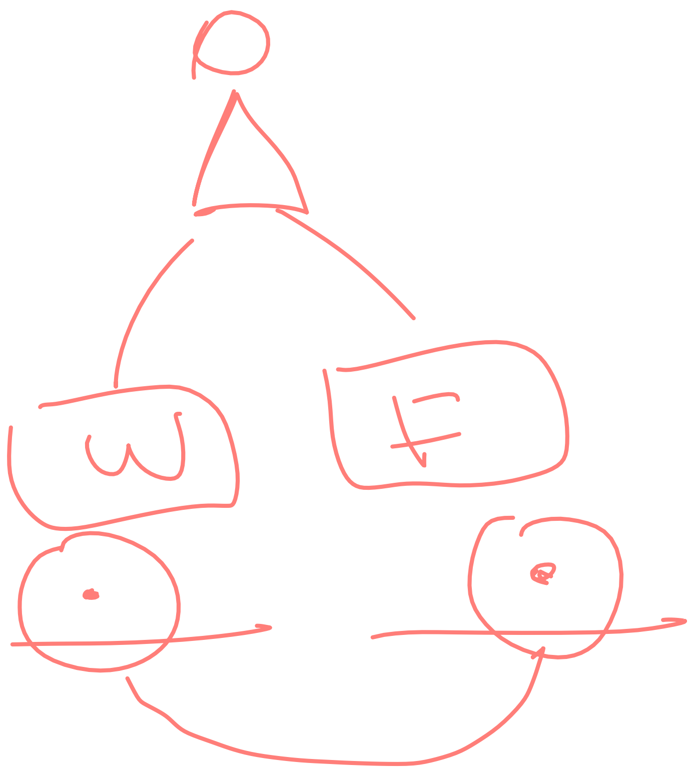
One-tailed and two-tailed tests are mathematically equivalent; they only differ in the application of the α level.

```
----------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+------------------------------------------------------------------
    male |      91    50.12088    1.080274    10.30516    47.97473    52.26703
  female |     109    54.99083    .7790686    8.133715    53.44658    56.53507
---------+------------------------------------------------------------------
combined |     200      52.775    .6702372    9.478586    51.45332    54.09668
---------+------------------------------------------------------------------
    diff |           -4.869947    1.304191               -7.441835   -2.298059
----------------------------------------------------------------------------
                         Degrees of freedom: 198

                  Ho: mean(male) - mean(female) = diff = 0

     Ha: diff < 0              Ha: diff != 0              Ha: diff > 0
       t =  -3.7341              t =  -3.7341              t =  -3.7341
   P < t =   0.0001        P > |t| =   0.0002          P > t =   0.9999
```

Does experimental design change how we perform the t-test?

Yes! There are two types of t-tests:

1. **Unpaired t-test**: When the data in the two distributions come from *different* populations.
2. **Paired t-test**: When the data in the two distributions come from the *same* population.



Unpaired t-test example

One-tailed

» $H_0: h_p = h_n$

» $H_1: h_p \neq h_n$

Two-tailed

» $H_0: h_p = h_n$

» $H_1: h_p \neq h_n$

Group	Participants	Task Completion Time	Coding
No prediction	Participant 1	245	0
No prediction	Participant 2	236	0
No prediction	Participant 3	321	0
No prediction	Participant 4	212	0
No prediction	Participant 5	267	0
No prediction	Participant 6	334	0
No prediction	Participant 7	287	0
No prediction	Participant 8	259	0
With prediction	Participant 9	246	1
With prediction	Participant 10	213	1
With prediction	Participant 11	265	1
With prediction	Participant 12	189	1
With prediction	Participant 13	201	1
With prediction	Participant 14	197	1
With prediction	Participant 15	289	1
With prediction	Participant 16	224	1

Unpaired t-test in R

```
data <- read.csv("t-test.csv")
t.test(data$Task.Completion.Time~data$Group)

Welch Two Sample t-test

data:  data$Task.Completion.Time by data$Group
t = 2.1688, df = 13.648, p-value = 0.04829
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  0.364964 83.885036
sample estimates:
  mean in group No prediction mean in group With prediction
                      270.125                        228.000
```

Unpaired t-test in JMP

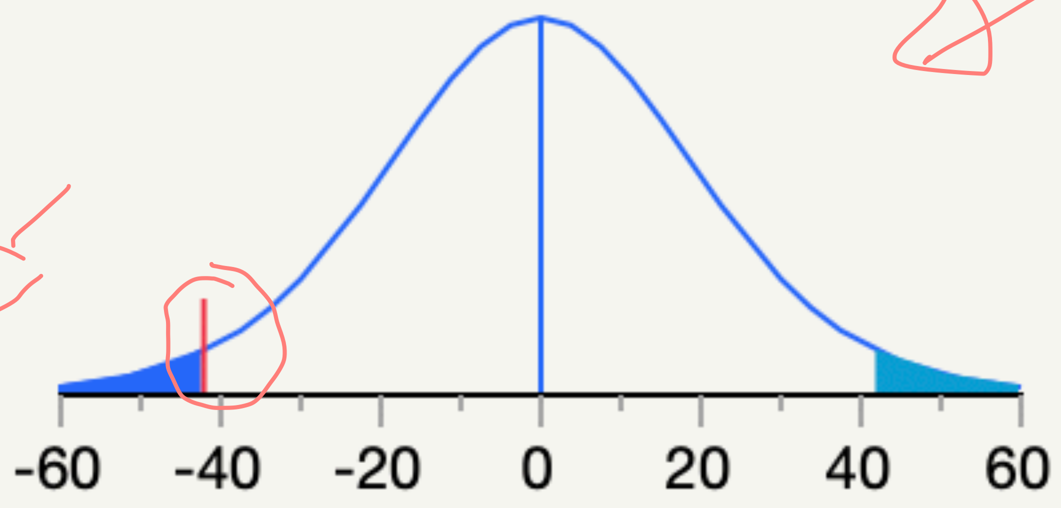
Analyze > Fit X by Y

t Test

With prediction-No prediction

Assuming unequal variances

Difference	-42.125	t Ratio	-2.16878
Std Err Dif	19.423	DF	13.6476
Upper CL Dif	-0.365	Prob > |t|	0.0483*
Lower CL Dif	-83.885	Prob > t	0.9759
Confidence	0.95	Prob < t	0.0241*



Paired t-test example

Participants	No Prediction	With Prediction
Participant 1	245	246
Participant 2	236	213
Participant 3	321	265
Participant 4	212	189
Participant 5	267	201
Participant 6	334	197
Participant 7	287	289
Participant 8	259	224

One-tailed

» $H_0: h_p = h_n$

» $H_1: h_p > h_n$

Two-tailed

» $H_0: h_p = h_n$

» $H_1: h_p \neq h_n$

Unpaired t-test in R

```
data <- read.csv("t-test-paired.csv")
t.test(data$No.Prediction,data$With.Prediction,paired=TRUE)
```

```
Paired t-test

data:  data$No.Prediction and data$With.Prediction
t = 2.6313, df = 7, p-value = 0.03385
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  4.268751 79.981249
sample estimates:
mean of the differences
                 42.125
```

Unpaired t-test in JMP

Analyze > Specialized Modeling > Matched Pairs

With Prediction	228	t-Ratio	-2.63126
No Prediction	270.125	DF	7
Mean Difference	-42.125	Prob > |t|	0.0339*
Std Error	16.0094	Prob > t	0.9831
Upper 95%	-4.2688	Prob < t	0.0169*
Lower 95%	-79.981		
N	8		
Correlation	0.32486		

What about when we have nominal output variables?

	Nominal	Categorical (2+)	Ordinal	Quantitative Discrete	Quantitative Non-Normal	Quantitative Normal
Nominal	Chi-squared, Fisher's	Chi-squared	Chi-squared Trend, Mann-Whitney	Mann-Whitney	Mann-Whitney, log-rank *	Student's *t*
Categorical (2+)	Chi-squared	Chi-squared	Kruskal-Wallis**	Kruskal-Wallis**	Kruskal-Wallis**	ANOVA***
Ordinal	Chi-squared Trend, Mann-Whitney	*****	Spearman rank	Spearman rank	Spearman rank	Spearman rank, linear regression
Quantitative Discrete	Logistic regression	*****	*****	Spearman rank	Spearman rank	Spearman rank, linear regression
Quantitative Non-Normal	Logistic regression	*****	*****	*****	Plot data-Pearson, Spearman rank	Plot data-Pearson, Spearman rank & linear regression
Quantitative Normal	Logistic regression	*****	*****	*****	Linear regression****	Pearson, linear regression

Contingency analysis

In contingency analysis, we calculate a chi-squared, X^2, statistic:

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

X^2 is the Pearson's test statistic, n is the number of observations, O_i is the observed frequency, and E_i is the expected frequency.

Data is summarized in a **contingency table** that cross-tabulates multivariate frequency distributions of variables in a matrix format.

Robot	Reported Gaze Cue
Robovie	Yes
Geminoid	Yes
Robovie	Yes
Geminoid	No
Robovie	Yes
Geminoid	No
Geminoid	No
Robovie	No
Robovie	Yes
Geminoid	No
Robovie	Yes
Geminoid	No
Robovie	No

```
          Reported.Gaze.Cue
Robot      No Yes
  Geminoid 10   3
  Robovie   3  10
```

Chi-squared test in R

```
gaze <- read.table('robot-gaze.csv', sep=",", header=TRUE)
chisq.test(table(gaze))
```

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

data:  table(gaze)
X-squared = 5.5385, df = 1, p-value = 0.0186
```

Chi-squared test in JMP

Analyze > Fit X by Y

N	DF	-LogLike	RSquare (U)
26	1	3.9765190	0.2207

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	7.953	0.0048*
Pearson	7.538	0.0060*

Fisher's Exact Test	Prob	Alternative Hypothesis
Left	0.9994	Prob(Robot=Robovie) is greater for Reported Gaze Cue=No than Yes
Right	0.0085*	Prob(Robot=Robovie) is greater for Reported Gaze Cue=Yes than No
2-Tail	0.0169*	Prob(Robot=Robovie) is different across Reported Gaze Cue

Hand-on activity

For your project, identify the input/output variable and appropriate statistical test for independent (unpaired) or dependent (paired) observations.