Human-Computer Interaction

Statistics II

Inferential Statistics (Cont'd)

Professor Bilge Mutlu

Recap: *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?

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.

Recap: *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.

Recap: *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.
- 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

Recap: *Which methods will we cover in this class?*

» X^2 ✅ Last week + recap today

» Student's t 👈 Today

» ANOVA 👈 Today

» Regression ❎ *

---

* Although we won't explicitly cover regression in this class, ANOVA is mathematically equivalent to a linear regression model with categorical predictors, and most software computes ANOVA using regression-based methods.

Recap: *Contingency analysis*

Pearson Chi-Square — Compares observed vs. expected counts using squared differences. Approximation that follows a chi-square distribution when sample size is large.

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

Works well when all expected counts ≥ 5.

Likelihood Ratio (G^2) — Also an approximation, but derived from maximum likelihood estimation (MLE). Compares how likely the observed data are under the null vs. full (saturated) model.

$$G^2 = 2 \sum O \ln\left(\frac{O}{E}\right)$$

Converges to χ^2 as sample size increases.

Fisher's Exact Test — Not an approximation. Computes the exact probability of observing the data under fixed row and column totals (hypergeometric).

$$P(\text{observed}) = \frac{\#\text{ ways to get observed table}}{\#\text{ possible tables}}$$

Sums probabilities of tables as or more extreme than observed. Used when sample sizes or expected counts are small.

How do we use the test outputs?

- Once χ^2 is computed, we compare it to a chi-square distribution with the appropriate degrees of freedom.
- The **p-value** is the probability of obtaining a χ^2 value *as large or larger* than the observed one *if the null hypothesis were true*.
- A small p-value means the observed deviations are unlikely under the null → evidence against the null hypothesis.
- When the deviations are *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 $p > .05$ to establish *significance*.

How do we conduct a χ^2 test?

Data is summarized in a **contingency table** that cross-tabulates multivariate frequency distributions of variables in a matrix format.[4]

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
```

[4] Data from "Mutlu et al. (2009). Nonverbal leakage in robots: communication of intentions through seemingly unintentional behavior. *HRI 2009.*"

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

Student's t-test

How do we conduct a t-test?

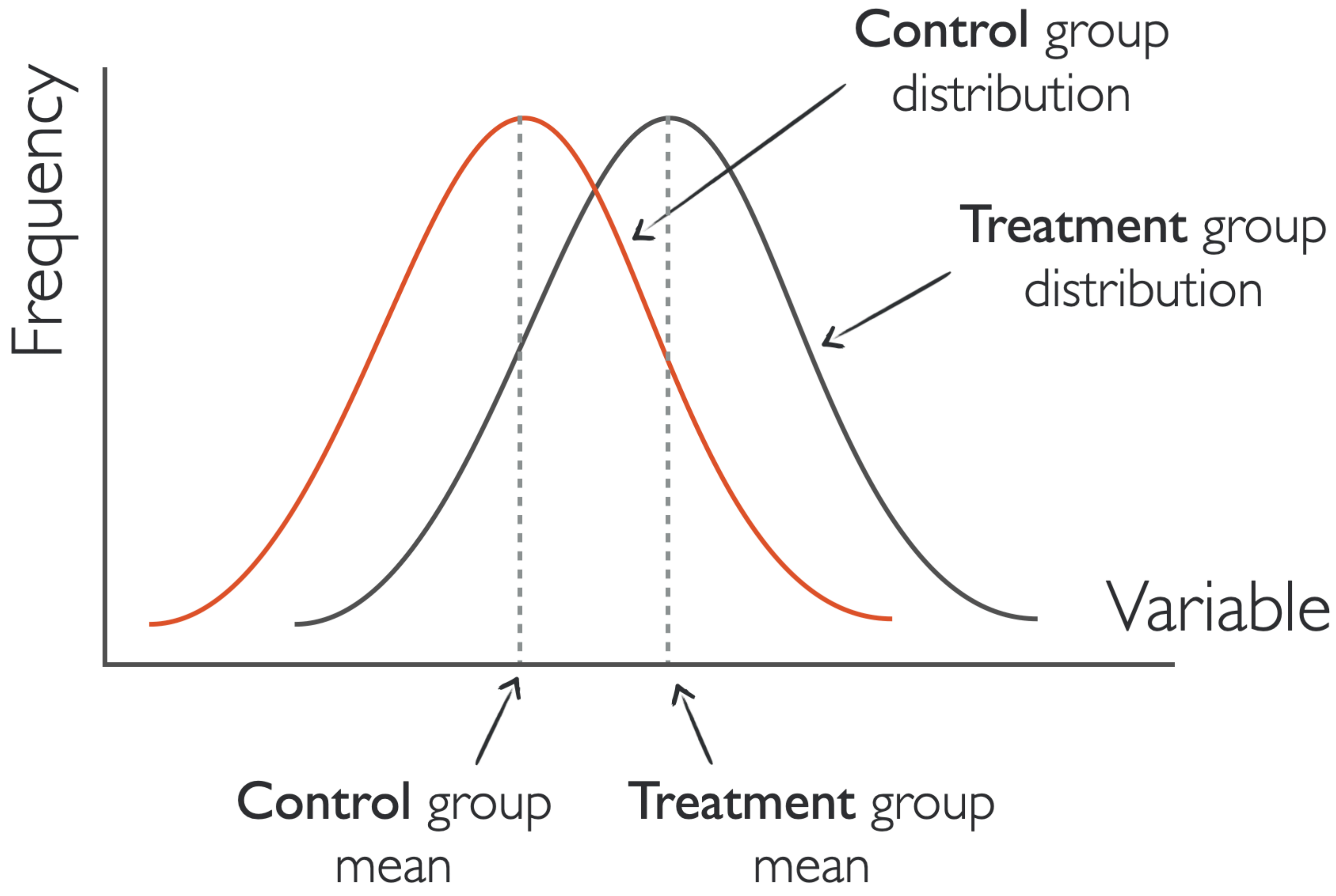
The *Student's t-test* assesses whether the means of two groups are **statistically different**.

Similar to the χ^2 test, when a different is *statistically significant*, the likelihood of it occurring by change is low, determined by a margin, called α level.

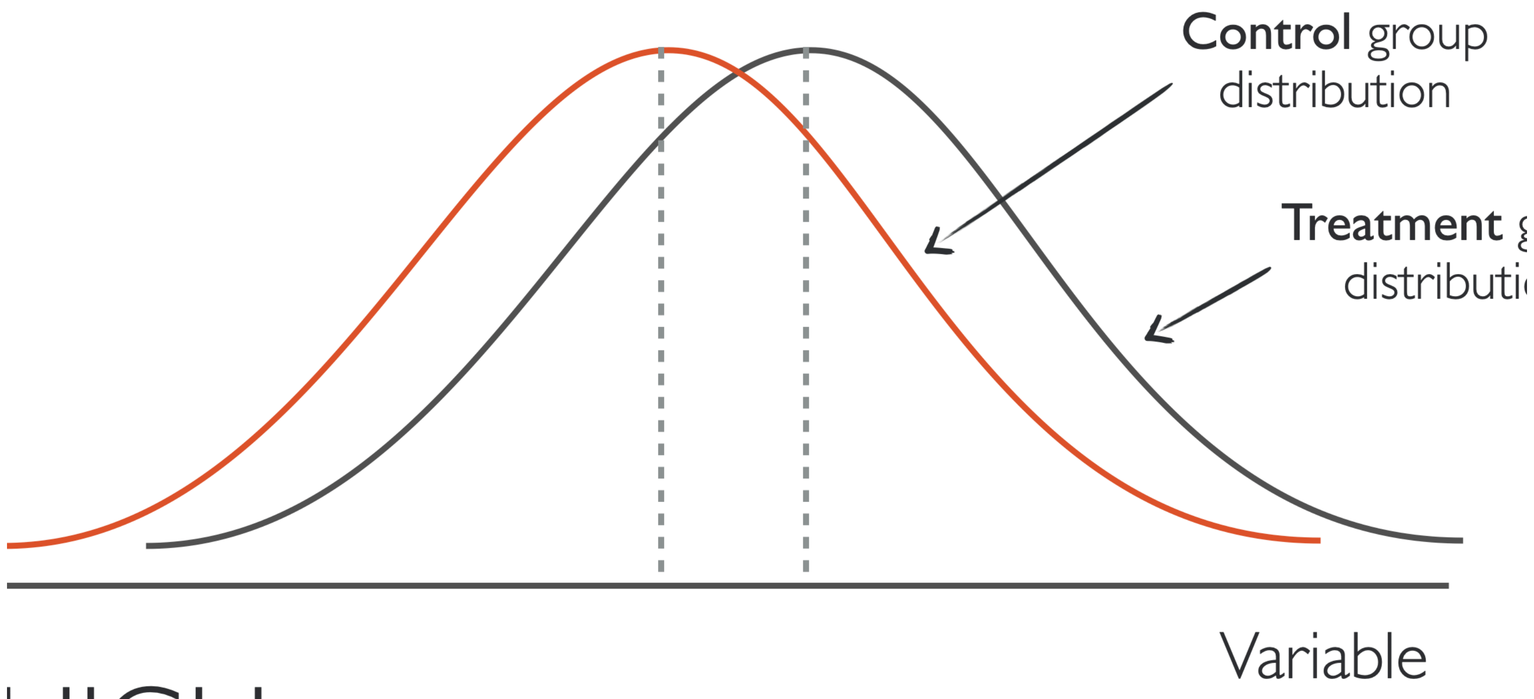
Again, in HCI research, $\alpha = .05$ is used, thus the probability, p, that the difference is occurring by change has to be $p > .05$ to establish *significance*.

So, how do we conduct a t-test?

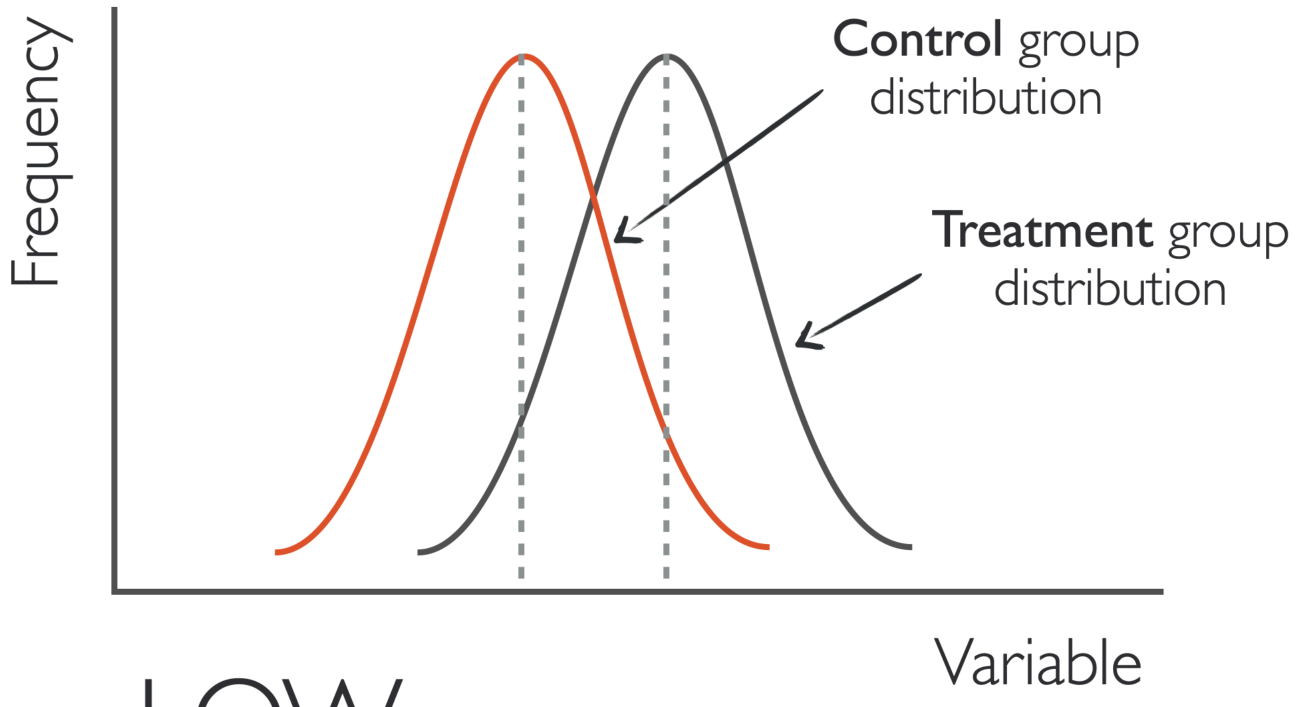
We look at two things: *difference in means* and *variability*.



Which two distributions are more likely to be statistically significant?



HIGH



LOW

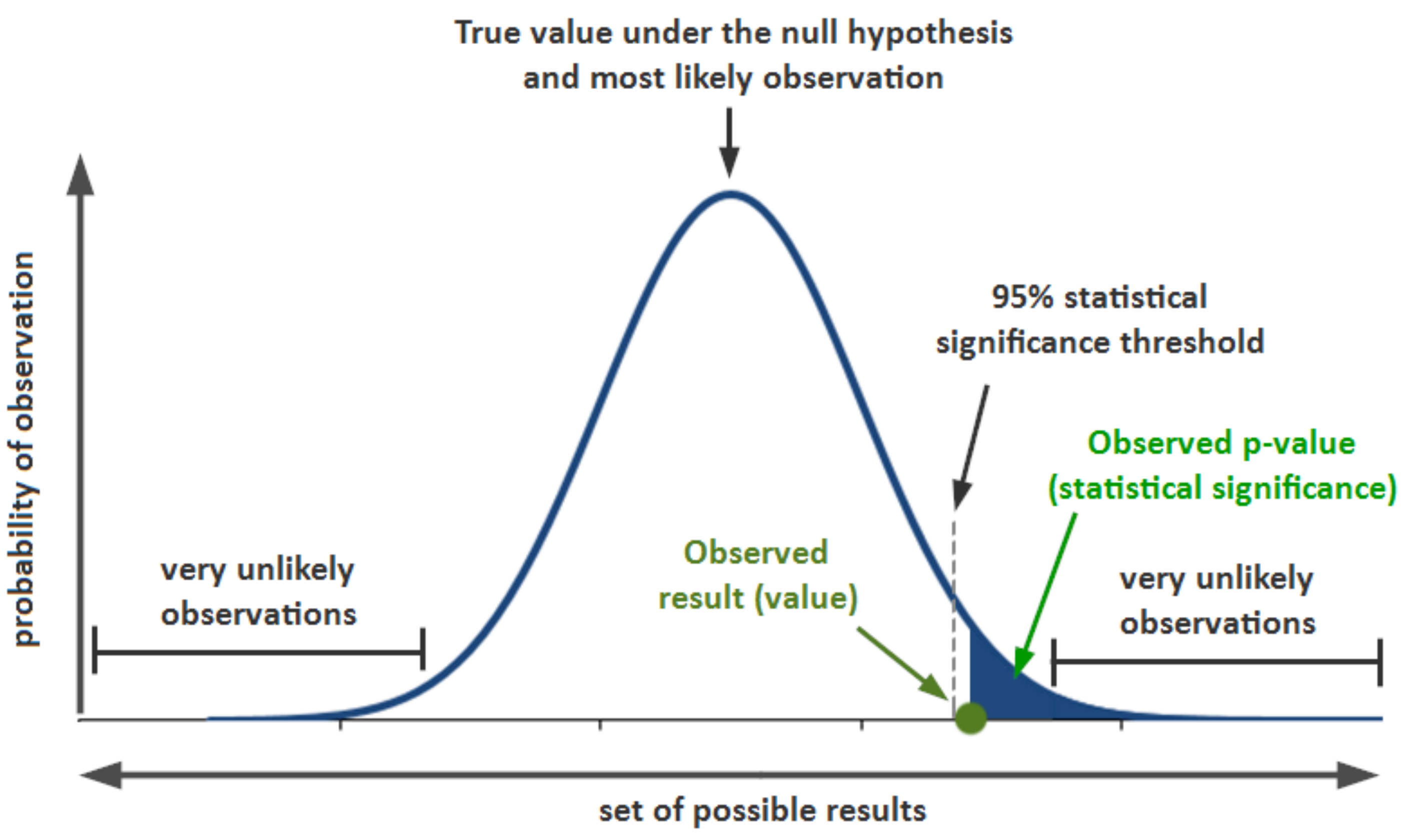
We need to calculate the t-statistic:

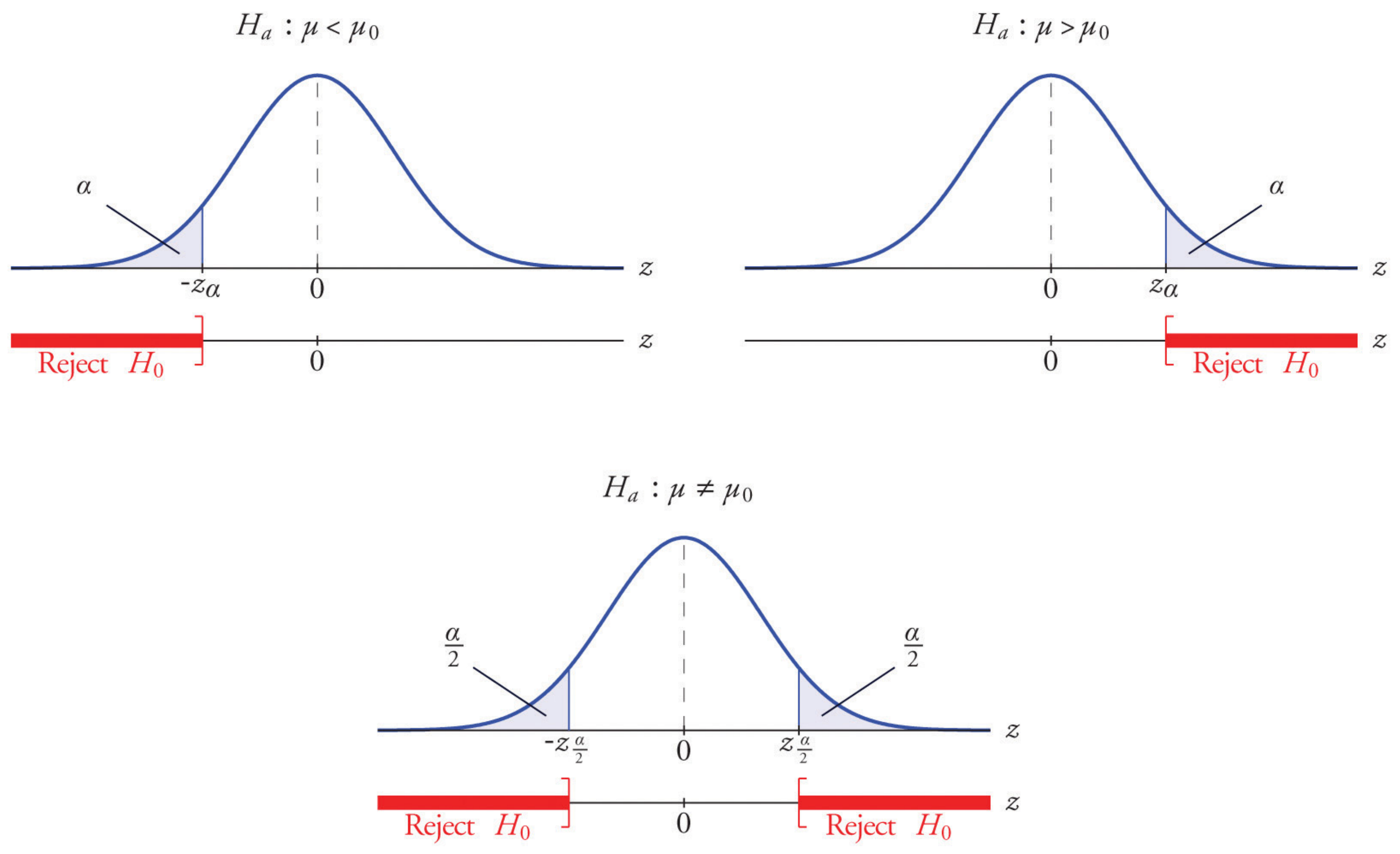
$$t = \frac{signal}{noise} = \frac{difference}{variability} = \frac{\mu_t - \mu_c}{\sqrt{\frac{\sigma_t}{n_t} + \frac{\sigma_c}{n_c}}}$$

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.

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

- $H_0: h_p = h_n$
- $H_1: h_p > h_n \lor h_p < h_n$

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

```
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

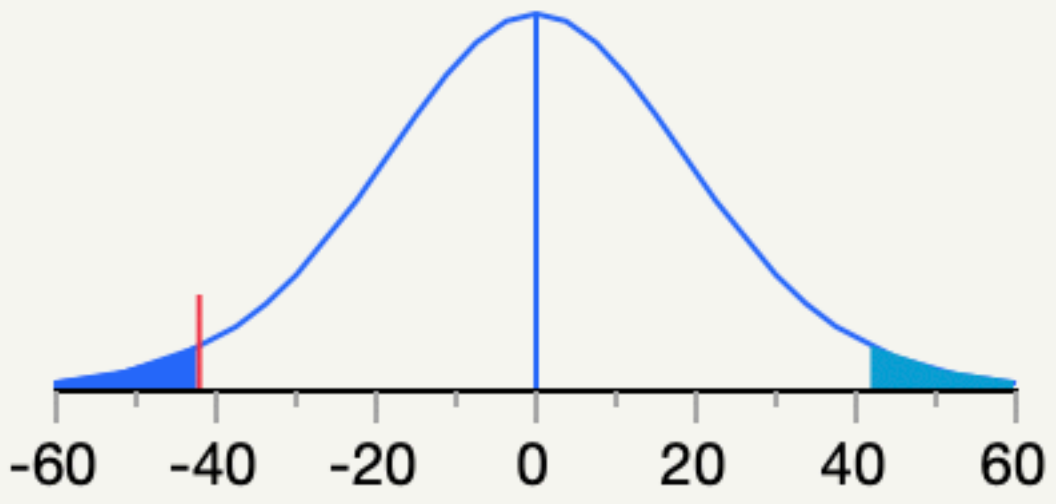
Analyze > Fit X by Y

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

- $H_0: h_p = h_n$
- $H_1: h_p > h_n \vee h_p < h_n$

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		

Consider this dataset. Can we use multiple t-tests?

Participant ID	Group	Time	Coding
Participant 01	Standard	245	0
Participant 02	Standard	236	0
Participant 03	Standard	321	0
Participant 04	Standard	212	0
Participant 05	Standard	267	0
Participant 06	Standard	334	0
Participant 07	Standard	287	0
Participant 08	Standard	259	0
Participant 09	Prediction	246	1
Participant 10	Prediction	213	1
Participant 11	Prediction	265	1
Participant 12	Prediction	189	1
Participant 13	Prediction	201	1
Participant 14	Prediction	197	1
Participant 15	Prediction	289	1
Participant 16	Prediction	224	1
Participant 17	Speech-based dictation	178	2
Participant 18	Speech-based dictation	289	2
Participant 19	Speech-based dictation	222	2
Participant 20	Speech-based dictation	189	2
Participant 21	Speech-based dictation	245	2
Participant 22	Speech-based dictation	311	2
Participant 23	Speech-based dictation	267	2
Participant 24	Speech-based dictation	197	2

$H_0 : \mu_1 = \mu_2 = \mu_3 \, , \, \alpha = .05$

3 pairwise tests: $(1 - \alpha)^3 = 0.86$

Reject H_0 when $p < 0.14$ instead of $p < 0.05$

➜ **Type I error** (reject H_0 when it is true)

Participant ID	Group	Time	Coding
Participant 01	Standard	245	0
Participant 02	Standard	236	0
Participant 03	Standard	321	0
Participant 04	Standard	212	0
Participant 05	Standard	267	0
Participant 06	Standard	334	0
Participant 07	Standard	287	0
Participant 08	Standard	259	0
Participant 09	Prediction	246	1
Participant 10	Prediction	213	1
Participant 11	Prediction	265	1
Participant 12	Prediction	189	1
Participant 13	Prediction	201	1
Participant 14	Prediction	197	1
Participant 15	Prediction	289	1
Participant 16	Prediction	224	1
Participant 17	Speech-based dictation	178	2
Participant 18	Speech-based dictation	289	2
Participant 19	Speech-based dictation	222	2
Participant 20	Speech-based dictation	189	2
Participant 21	Speech-based dictation	245	2
Participant 22	Speech-based dictation	311	2
Participant 23	Speech-based dictation	267	2
Participant 24	Speech-based dictation	197	2

What are errors in hypothesis testing?

Type I error: Rejecting H_0 when it is true

Type II error: Accepting H_0 when it is false

Type III error: Correctly rejecting H_0 for the wrong reason

	Null Hypothesis is true	**Alternative Hypothesis is true**
Fail to reject H_0	*Right decision*	*Wrong decision* **Type II error** (False negative)
Reject H_0	Wrong decision **Type I error** (False positive)	*Right decision*

Analysis of Variance (ANOVA)

Definition: Analysis of variance (ANOVA) is a collection of statistical models and their associated estimation procedures (such as the "variation" among and between groups) used to analyze the differences among group means in a sample.[1]

Procedures:

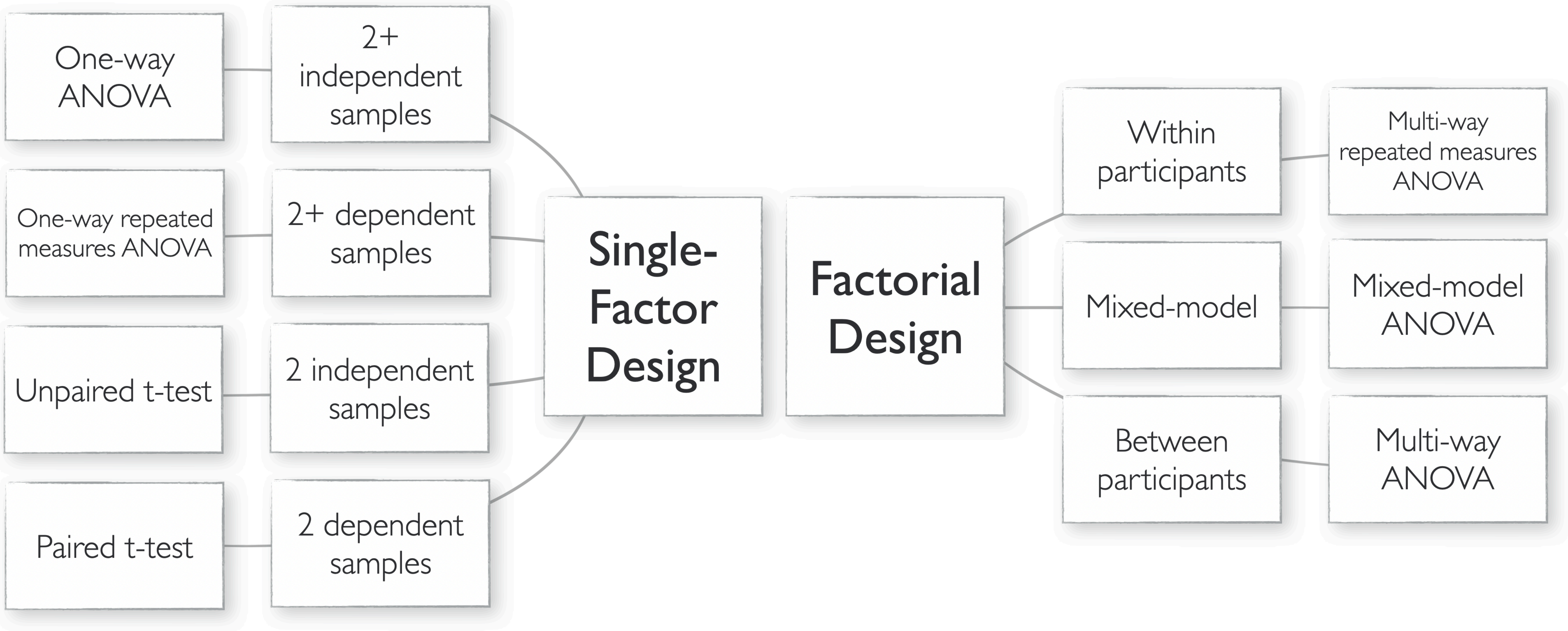
1. One-way (single factor)
2. Two-way (two factors)
3. Multi-way (multiple factors)

Models:

1. Fixed effects (between)
2. Random effects (within)
3. Mixed effects (mixed)

[1] Wikipedia: ANOVA

How do we choose among these procedures?



How do we conduct ANOVA?

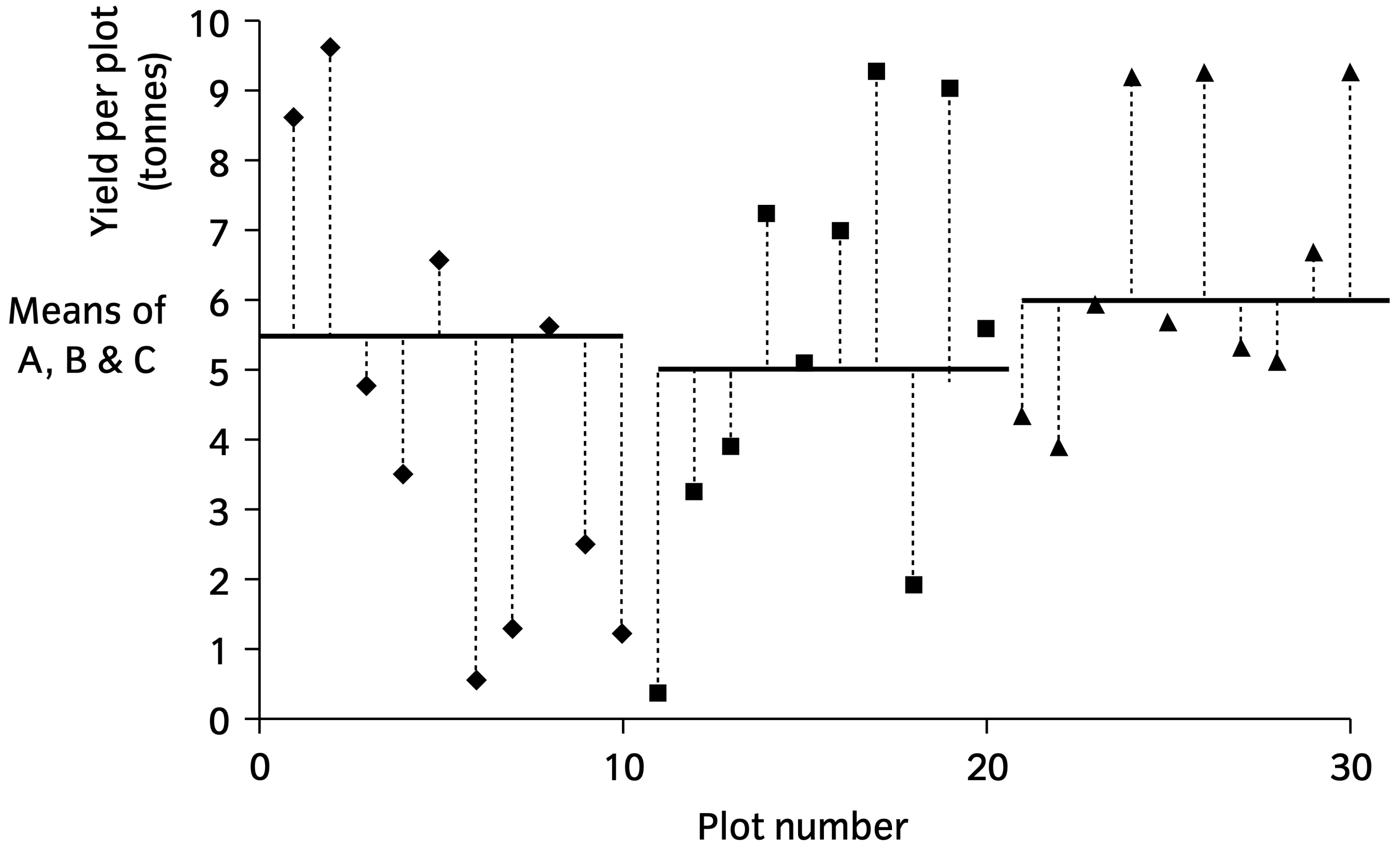
We calculate the *F*-statistic.

$$F = \frac{\sigma_{explained}}{\sigma_{unexplained}} = \frac{SS_{treatment}/(k-1)}{SS_{error}/(n-k)}$$

$$F = \frac{\sum n_i (M_i - \sum (Mi/k))^2/(k-1)}{\sum\sum (X_{it} - M_i)^2/(n-k)}$$

k: number of populations

n: sample size



One-way ANOVA in R

```
model = aov(Time~Group,data=data)
summary(model)
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
Group        2   7842    3921   2.174  0.139
Residuals   21  37880    1804
```

Participant ID	Group	Time	Coding
Participant 01	Standard	245	0
Participant 02	Standard	236	0
Participant 03	Standard	321	0
Participant 04	Standard	212	0
Participant 05	Standard	267	0
Participant 06	Standard	334	0
Participant 07	Standard	287	0
Participant 08	Standard	259	0
Participant 09	Prediction	246	1
Participant 10	Prediction	213	1
Participant 11	Prediction	265	1
Participant 12	Prediction	189	1
Participant 13	Prediction	201	1
Participant 14	Prediction	197	1
Participant 15	Prediction	289	1
Participant 16	Prediction	224	1
Participant 17	Speech-based dictation	178	2
Participant 18	Speech-based dictation	289	2
Participant 19	Speech-based dictation	222	2
Participant 20	Speech-based dictation	189	2
Participant 21	Speech-based dictation	245	2
Participant 22	Speech-based dictation	311	2
Participant 23	Speech-based dictation	267	2
Participant 24	Speech-based dictation	197	2

One-way ANOVA in JMP

Analyze > Fit X by Y

Oneway Anova

Summary of Fit

Rsquare	0.171518
Adj Rsquare	0.092615
Root Mean Square Error	42.47149
Mean of Response	245.125
Observations (or Sum Wgts)	24

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Group	2	7842.250	3921.13	2.1738	0.1387
Error	21	37880.375	1803.83		
C. Total	23	45722.625			

Means for Oneway Anova

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
Prediction	8	228.000	15.016	196.77	259.23
Speech-based dictation	8	237.250	15.016	206.02	268.48
Standard	8	270.125	15.016	238.90	301.35

Std Error uses a pooled estimate of error variance

Are we done?

The ANOVA analysis only told us whether the *methods* had a significant effect on *time*, not which method is more effective.

We can make two types of *pairwise* comparisons:

1. *A priori* comparisons (planned contrasts)

   $H_0: \mu_1 = \mu_2; H_1: \mu_1 > \mu_2$

2. *Post hoc* comparisons (exploratory pairwise tests)

   Test μ_1 vs μ_2, μ_1 vs μ_3, μ_2 vs μ_3

A priori comparisons in R

```
levels(data$Group)
comparison = c(1,-1,0)
mat = cbind(comparison)
contrasts(data$Group) <- mat
model = aov(Time~Group, data= data)
summary.aov(model, split = list(Group=list("mu1 vs mu2"=1)))
```

```
                     Df Sum Sq Mean Sq F value Pr(>F)
Group                 2   7842    3921   2.174  0.139
  Group: mu1 vs mu2  1     342     342   0.190  0.668
Residuals            21  37880    1804
```

A priori comparisons in JMP

Compare Means > Each pair, Student's t

Means Comparisons

Comparisons for each pair using Student's t

Confidence Quantile

t	Alpha
2.07961	0.05

LSD Threshold Matrix

Abs(Dif)-LSD

	Standard	Speech-based dictation	Prediction
Standard	-44.162	-11.287	-2.037
Speech-based dictation	-11.287	-44.162	-34.912
Prediction	-2.037	-34.912	-44.162

Positive values show pairs of means that are significantly different.

Connecting Letters Report

Level		Mean
Standard	A	270.12500
Speech-based dictation	A	237.25000
Prediction	A	228.00000

Levels not connected by same letter are significantly different.

Ordered Differences Report

Level	- Level	Difference	Std Err Dif	Lower CL	Upper CL	p-Value
Standard	Prediction	42.12500	21.23574	-2.0371	86.28715	0.0605
Standard	Speech-based dictation	32.87500	21.23574	-11.2871	77.03715	0.1365
Speech-based dictation	Prediction	9.25000	21.23574	-34.9121	53.41215	0.6676

Post hoc comparison in R

`TukeyHSD(model)`

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

Fit: aov(formula = Time ~ Group, data = data)

$Group
                                     diff       lwr      upr     p adj
Speech-based dictation-Prediction  9.250 -44.27619 62.77619 0.9011856
Standard-Prediction               42.125 -11.40119 95.65119 0.1409733
Standard-Speech-based dictation   32.875 -20.65119 86.40119 0.2896872
```

Post hoc comparison in JMP

Compare Means > All Pairs, Tukey HSD

Comparisons for all pairs using Tukey-Kramer HSD

Confidence Quantile

q*	Alpha
2.52057	0.05

HSD Threshold Matrix

Abs(Dif)-HSD

	Standard	Speech-based dictation	Prediction
Standard	-53.526	-20.651	-11.401
Speech-based dictation	-20.651	-53.526	-44.276
Prediction	-11.401	-44.276	-53.526

Positive values show pairs of means that are significantly different.

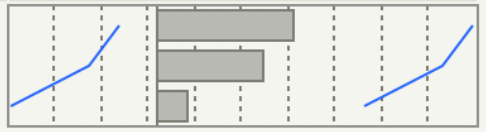
Connecting Letters Report

Level		Mean
Standard	A	270.12500
Speech-based dictation	A	237.25000
Prediction	A	228.00000

Levels not connected by same letter are significantly different.

Ordered Differences Report

Level	- Level	Difference	Std Err Dif	Lower CL	Upper CL	p-Value
Standard	Prediction	42.12500	21.23574	-11.4012	95.65119	0.1410
Standard	Speech-based dictation	32.87500	21.23574	-20.6512	86.40119	0.2897
Speech-based dictation	Prediction	9.25000	21.23574	-44.2762	62.77619	0.9012



What if we had a within-participants design?

We conduct a *repeated-measures* or *random-effects* one-way ANOVA.

Participant ID	Group	Time	Coding
Participant 01	Standard	245	0
Participant 01	Prediction	246	1
Participant 01	Speech-based dictation	178	2
Participant 02	Standard	236	0
Participant 02	Prediction	213	1
Participant 02	Speech-based dictation	289	2
Participant 03	Standard	321	0
Participant 03	Prediction	265	1
Participant 03	Speech-based dictation	222	2
Participant 04	Standard	212	0
Participant 04	Prediction	189	1
Participant 04	Speech-based dictation	189	2
Participant 05	Standard	267	0
Participant 05	Prediction	201	1
Participant 05	Speech-based dictation	245	2
Participant 06	Standard	334	0
Participant 06	Prediction	197	1
Participant 06	Speech-based dictation	311	2
Participant 07	Standard	287	0
Participant 07	Prediction	289	1
Participant 07	Speech-based dictation	267	2
Participant 08	Standard	259	0
Participant 08	Prediction	224	1
Participant 08	Speech-based dictation	197	2

Within-participants one-way ANOVA in R

```
model = aov(Time~Group+Error(Participant.ID/Group), data= data)
summary(model)

Error: Participant.ID
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals  7  19113    2730

Error: Participant.ID:Group
          Df Sum Sq Mean Sq F value Pr(>F)
Group      2   7842    3921   2.925 0.0868 .
Residuals 14  18767    1341
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Within-participants one-way ANOVA in JMP

Using the Full Factorial Repeated Measures ANOVA Add-In:

Add-ins > Repeated Measures > Full-Factorial Design (Mixed Effects)

For additional options (e.g., comparisons):

Launch Dialog > Emphasis: Effect Leverage

Response Time

Effect Summary

Summary of Fit

RSquare	0.48879
RSquare Adj	0.440103
Root Mean Square Error	36.61292
Mean of Response	245.125
Observations (or Sum Wgts)	24

Parameter Estimates

Random Effect Predictions

REML Variance Component Estimates

Random Effect	Var Ratio	Var Component	Std Error	95% Lower	95% Upper	Wald p-Value	Pct of Total
Participant ID	0.3456318	463.32143	514.98022	-546.0213	1472.6641	0.3683	25.685
Participant ID*Group		1340.506	506.66363	718.52371	3334.1618	<.0001*	74.315
Total		1803.8274	592.26174	1037.3604	3890.013		100.000

-2 LogLikelihood = 224.22780502

Note: Total is the sum of the positive variance components.

Total including negative estimates = 1803.8274

Covariance Matrix of Variance Component Estimates

Residual is confounded with Participant ID*Group and has been removed.

Iterations

Fixed Effect Tests

Source	Nparm	DF	DFDen	F Ratio	Prob > F
Group	2	2	14	2.9251	0.0868

Effect Details

Between-participants two-way ANOVA in R

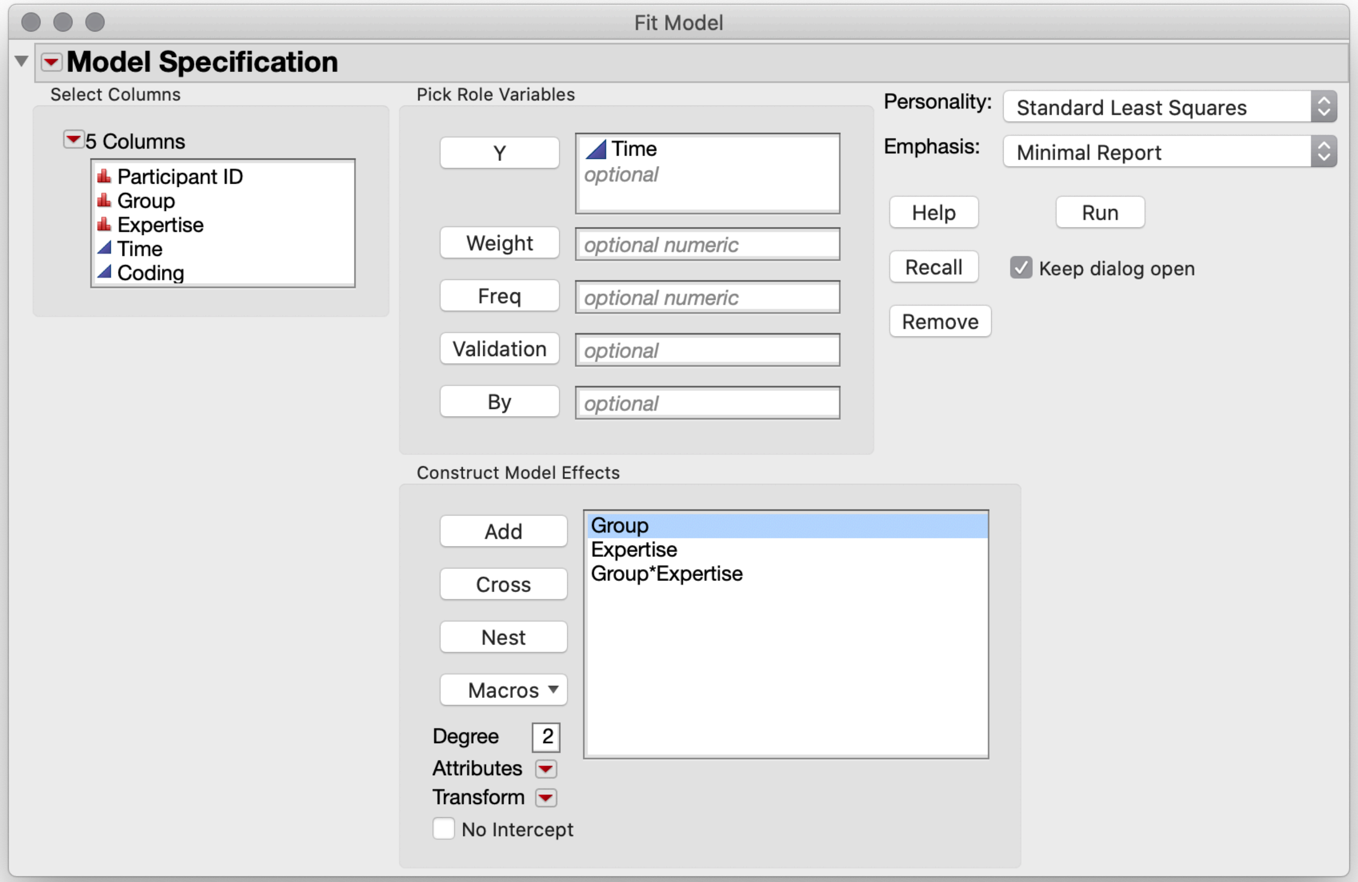
Task type	Entry method	Participant Number	Task time	Task Type coding	Entry Method coding
Transcription	Standard	Participant 1	245	0	0
Transcription	Standard	Participant 2	236	0	0
…	…	…	…	…	…
Transcription	Prediction	Participant 9	246	0	1
Transcription	Prediction	Participant 10	213	0	1
…	…	…	…	…	…
Transcription	Speech-based dictation	Participant 17	178	0	2
Transcription	Speech-based dictation	Participant 18	289	0	2
…	…	…	…	…	…
Composition	Standard	Participant 25	256	1	0
Composition	Standard	Participant 26	269	1	0
…	…	…	…	…	…
Composition	Prediction	Participant 33	265	1	1
Composition	Prediction	Participant 34	232	1	1
…	…	…	…	…	…
Composition	Speech-based dictation	Participant 41	189	1	2
Composition	Speech-based dictation	Participant 42	321	1	2
…	…	…	…	…	…
Composition	Speech-based dictation	Participant 48	202	1	2

```
model = aov(Time~Group*Expertise, data=data)
summary(model)
```

```
                Df Sum Sq Mean Sq F value Pr(>F)
Group            2   7842    3921   2.175  0.143
Expertise        1   1395    1395   0.774  0.391
Group:Expertise  2   4030    2015   1.117  0.349
Residuals       18  32455    1803
```

Between-participants two-way ANOVA in JMP

Analyze > Fit Model



Summary of Fit

RSquare	0.290171
RSquare Adj	0.092996
Root Mean Square Error	42.46257
Mean of Response	245.125
Observations (or Sum Wgts)	24

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	13267.375	2653.48	1.4716
Error	18	32455.250	1803.07	**Prob > F**
C. Total	23	45722.625		0.2477

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	245.125	8.667635	28.28	<.0001*
Group[Prediction]	-17.125	12.25789	-1.40	0.1794
Group[Speech-based dictation]	-7.875	12.25789	-0.64	0.5287
Expertise[Expert]	-7.625	8.667635	-0.88	0.3906
Group[Prediction]*Expertise[Expert]	-14.625	12.25789	-1.19	0.2483
Group[Speech-based dictation]*Expertise[Expert]	16.875	12.25789	1.38	0.1855

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Group	2	2	7842.2500	2.1747	0.1426
Expertise	1	1	1395.3750	0.7739	0.3906
Group*Expertise	2	2	4029.7500	1.1175	0.3488

Within-participants two-way ANOVA in R

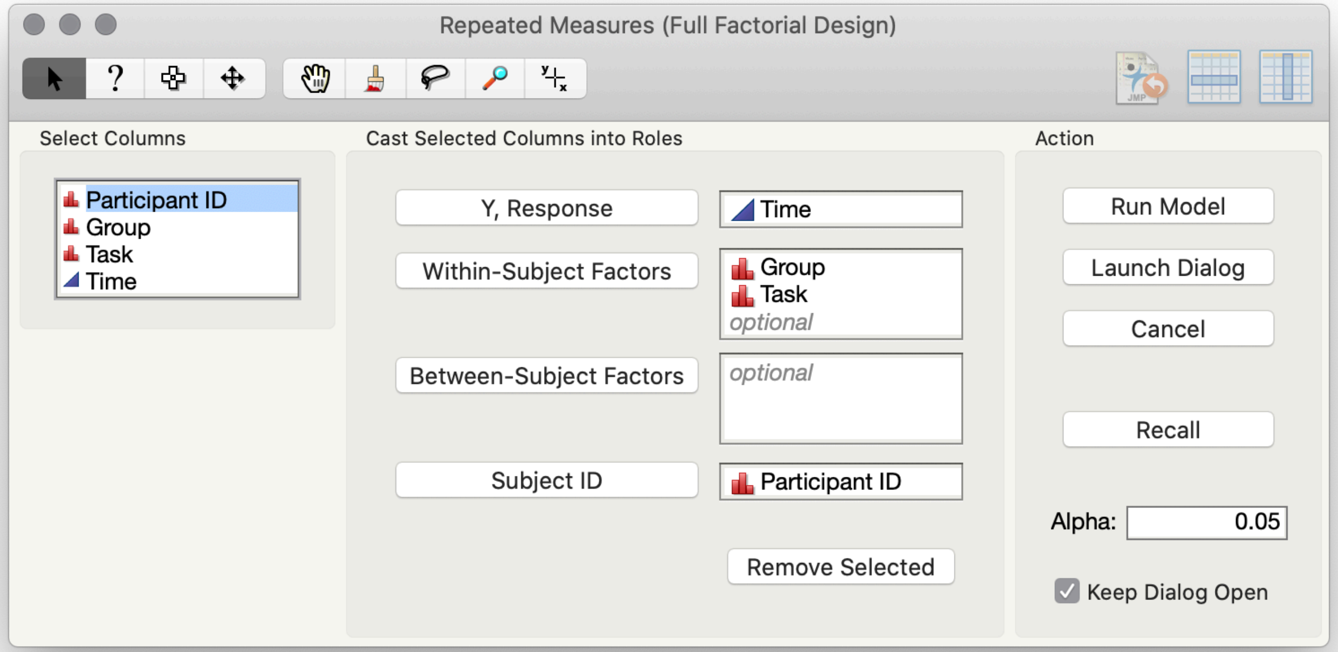
```
model = aov(Time~(Group*Task)+Error(Participant.ID/(Group*Task)), data= data)
summary(model)
```

Participant ID	Group	Task	Time
Participant 01	Standard	Complex	285
Participant 01	Prediction	Complex	160
Participant 01	Speech-based dictation	Complex	201
Participant 01	Standard	Simple	272
Participant 01	Prediction	Simple	191
Participant 01	Speech-based dictation	Simple	161
Participant 02	Standard	Complex	189
Participant 02	Prediction	Complex	250
Participant 02	Speech-based dictation	Complex	178
Participant 02	Standard	Simple	247
Participant 02	Prediction	Simple	288
Participant 02	Speech-based dictation	Simple	180
Participant 03	Standard	Complex	233
Participant 03	Prediction	Complex	285
Participant 03	Speech-based dictation	Complex	225
Participant 03	Standard	Simple	200
Participant 03	Prediction	Simple	202
Participant 03	Speech-based dictation	Simple	162

```
Error: Participant.ID
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals  7   7224    1032

Error: Participant.ID:Group
          Df Sum Sq Mean Sq F value Pr(>F)
Group      2   1650   825.2   0.345  0.714
Residuals 14  33441  2388.6

Error: Participant.ID:Task
          Df Sum Sq Mean Sq F value Pr(>F)
Task       1    341   341.3   0.119   0.74
Residuals  7  20055  2865.0

Error: Participant.ID:Group:Task
           Df Sum Sq Mean Sq F value Pr(>F)
Group:Task  2   1845   922.5   0.644   0.54
Residuals  14  20053  1432.3
```

Within-participants two-way ANOVA in JMP

Add-ins > Repeated Measures > Full-Factorial Design (Mixed Effects)



Summary of Fit

RSquare	0.397171
RSquare Adj	0.325405
Root Mean Square Error	37.84614
Mean of Response	216.625
Observations (or Sum Wgts)	48

Parameter Estimates

Term	Estimate	Std Error	DFDen	t Ratio	Prob>|t|
Intercept	216.625	4.636889	7	46.72	<.0001*
Group[Prediction]	1.6875	9.976255	14	0.17	0.8681
Group[Speech-based dictation]	-7.875	9.976255	14	-0.79	0.4431
Task[Complex]	2.6666667	7.725769	7	0.35	0.7401
Group[Prediction]*Task[Complex]	-2.229167	7.725311	14	-0.29	0.7772
Group[Speech-based dictation]*Task[Complex]	8.4583333	7.725311	14	1.09	0.2920

Random Effect Predictions

REML Variance Component Estimates

Random Effect	Var Ratio	Var Component	Std Error	95% Lower	95% Upper	Wald p-Value	Pct of Total
Participant ID	-0.324559	-464.875	323.07858	-1098.097	168.34739	0.1502	0.000
Participant ID*Group	0.3338216	478.14286	526.34376	-553.4719	1509.7577	0.3637	20.022
Participant ID*Task	0.3334123	477.55655	541.42698	-583.6208	1538.7339	0.3778	19.998
Participant ID*Group*Task		1432.3304	541.36999	767.74244	3562.5512	<.0001*	59.980
Total		2388.0298	687.37698	1457.7831	4611.3193		100.000

-2 LogLikelihood = 455.15548778

Note: Total is the sum of the positive variance components.

Total including negative estimates = 1923.1548

Covariance Matrix of Variance Component Estimates

Residual is confounded with Participant ID*Group*Task and has been removed.

Iterations

Fixed Effect Tests

Source	Nparm	DF	DFDen	F Ratio	Prob > F
Group	2	2	14	0.3455	0.7138
Task	1	1	7	0.1191	0.7401
Group*Task	2	2	14	0.6441	0.5400

Two-way mixed-effects ANOVA in R

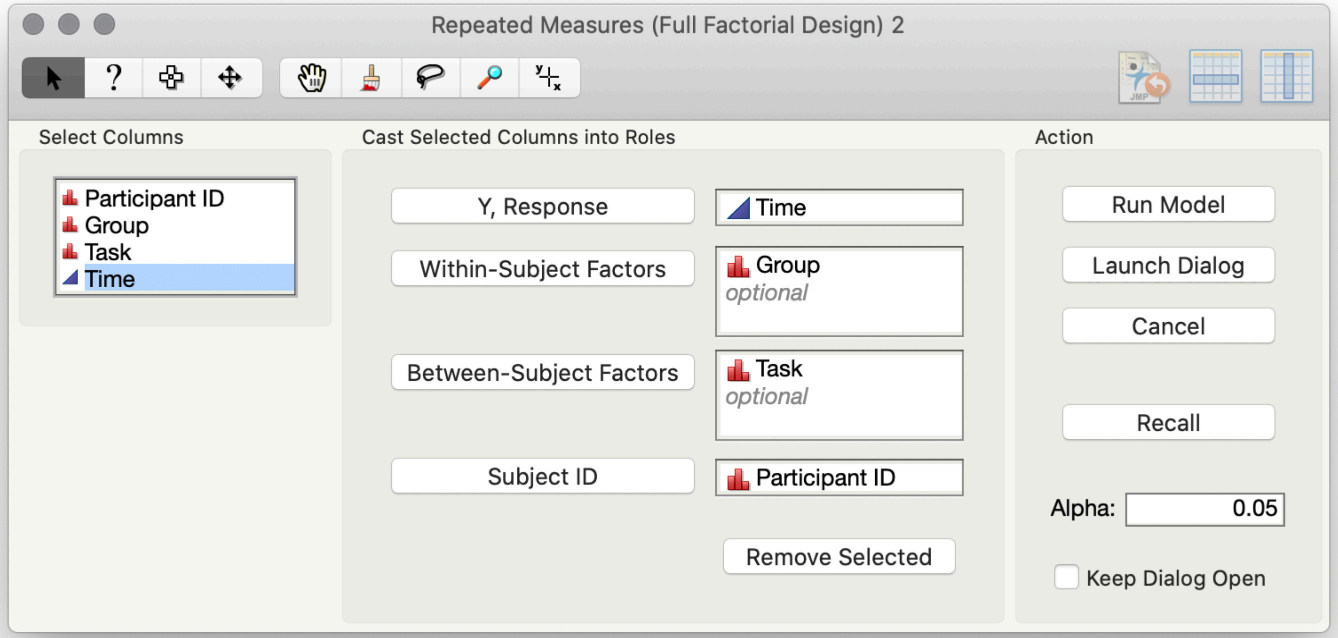
```
model = aov(Time~(Group*Task)+Error(Participant.ID/Group)+Task,data=data)
summary(model)
```

Participant ID	Group	Task	Time
Participant 01	Standard	Complex	285
Participant 01	Prediction	Complex	160
Participant 01	Speech-based dictation	Complex	201
Participant 02	Standard	Simple	272
Participant 02	Prediction	Simple	191
Participant 02	Speech-based dictation	Simple	161
Participant 03	Standard	Complex	189
Participant 03	Prediction	Complex	250
Participant 03	Speech-based dictation	Complex	178
Participant 04	Standard	Simple	247
Participant 04	Prediction	Simple	288
Participant 04	Speech-based dictation	Simple	180
Participant 05	Standard	Complex	233
Participant 05	Prediction	Complex	285
Participant 05	Speech-based dictation	Complex	225
Participant 06	Standard	Simple	200
Participant 06	Prediction	Simple	202
Participant 06	Speech-based dictation	Simple	162

```
Error: Participant.ID
          Df Sum Sq Mean Sq F value Pr(>F)
Task       1    341   341.3   0.175  0.682
Residuals 14  27279  1948.5

Error: Participant.ID:Group
           Df Sum Sq Mean Sq F value Pr(>F)
Group       2   1650   825.2   0.432  0.654
Group:Task  2   1845   922.5   0.483  0.622
Residuals  28  53493  1910.5
```

Two-way mixed-effects ANOVA in JMP

Add-ins > Repeated Measures > Full-Factorial Design (Mixed Effects)



Summary of Fit

RSquare	0.057814
RSquare Adj	-0.05435
Root Mean Square Error	43.70896
Mean of Response	216.625
Observations (or Sum Wgts)	48

Parameter Estimates

Term	Estimate	Std Error	DFDen	t Ratio	Prob>|t|
Intercept	216.625	6.371352	14	34.00	<.0001*
Task[Complex]	2.6666667	6.371352	14	0.42	0.6819
Group[Prediction]	1.6875	8.922054	28	0.19	0.8513
Group[Speech-based dictation]	-7.875	8.922054	28	-0.88	0.3849
Task[Complex]*Group[Prediction]	-2.229167	8.922054	28	-0.25	0.8045
Task[Complex]*Group[Speech-based dictation]	8.4583333	8.922054	28	0.95	0.3512

Random Effect Predictions

REML Variance Component Estimates

Random Effect	Var Ratio	Var Component	Std Error	95% Lower	95% Upper	Wald p-Value	Pct of Total
Participant ID[Task]	0.0066379	12.681548	298.71885	-572.7966	598.15973	0.9661	0.659
Participant ID*Group[Task]		1910.4732	510.59544	1203.1556	3494.4955	<.0001*	99.341
Total		1923.1548	419.68502	1307.4704	3106.8671		100.000

-2 LogLikelihood = 457.81133323

Note: Total is the sum of the positive variance components.

Total including negative estimates = 1923.1548

Covariance Matrix of Variance Component Estimates

Residual is confounded with Participant ID*Group[Task] and has been removed.

Iterations

Fixed Effect Tests

Source	Nparm	DF	DFDen	F Ratio	Prob > F
Task	1	1	14	0.1752	0.6819
Group	2	2	28	0.4319	0.6535
Task*Group	2	2	28	0.4829	0.6221

What if I would like to include a covariate?

Participant ID	Group	Time	Years
Participant 01	Standard	245	12
Participant 02	Standard	236	5
Participant 03	Standard	321	6
Participant 04	Standard	212	13
Participant 05	Standard	267	19
Participant 06	Standard	334	18
Participant 07	Standard	287	18
Participant 08	Standard	259	18
Participant 09	Prediction	246	14
Participant 10	Prediction	213	3
Participant 11	Prediction	265	19
Participant 12	Prediction	189	13
Participant 13	Prediction	201	24
Participant 14	Prediction	197	21
Participant 15	Prediction	289	5
Participant 16	Prediction	224	18
Participant 17	Speech-based dictation	178	21
Participant 18	Speech-based dictation	289	18
Participant 19	Speech-based dictation	222	23
Participant 20	Speech-based dictation	189	16
Participant 21	Speech-based dictation	245	12
Participant 22	Speech-based dictation	311	15
Participant 23	Speech-based dictation	267	16
Participant 24	Speech-based dictation	197	9

Consider the one-way between-subjects analysis and also measuring the *years of experience* the user had in the task to control for that factor.

We conduct what is called an analysis of co-variance (ANCOVA).

One-way between-participants ANCOVA in R

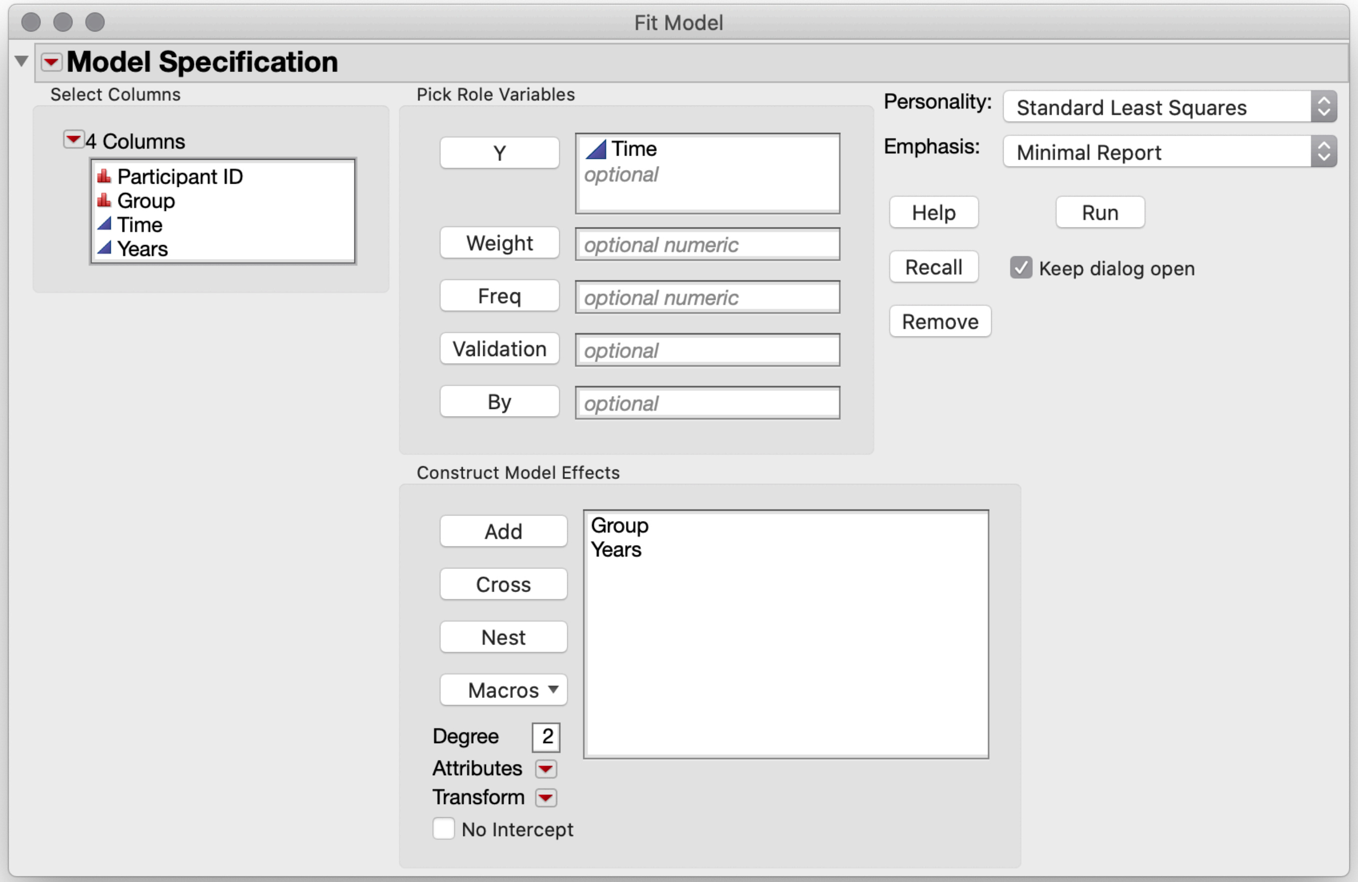
```
model = aov(Time~Group+Years, data=data)
summary(model)
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
Group         2   7842    3921   2.090   0.15
Years         1    350     350   0.187   0.67
Residuals    20  37530    1877
```

Because `Years` has no effect, we would remove it from our model (called *model simplification*) and rerun our analysis as an ANOVA.

One-way between-participants ANCOVA in JMP

Analyze > Fit Model



Summary of Fit

RSquare	0.179172
RSquare Adj	0.056048
Root Mean Square Error	43.31881
Mean of Response	245.125
Observations (or Sum Wgts)	24

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	8192.233	2730.74	1.4552
Error	20	37530.392	1876.52	Prob > F
C. Total	23	45722.625		0.2568

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	255.22257	24.99753	10.21	<.0001*
Group[Prediction]	-17.26682	12.50938	-1.38	0.1827
Group[Speech-based dictation]	-6.910626	12.70288	-0.54	0.5924
Years	-0.680735	1.576272	-0.43	0.6705

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Group	2	2	7341.1619	1.9561	0.1675
Years	1	1	349.9828	0.1865	0.6705

Data files used in Statistics I & II

Assignment: Quantitative Data Analysis

- Identify correct *statistical test* — (review survey design, justify)
- Prepare a *wide-format item-level table* — (1 row per participant × condition; columns = items)
- Compute *scale variables* — (reverse-score if needed; average items per construct)
- Conduct *statistical test* — comparing your two survey conditions
- Write *Participants*, *Analysis*, *Results* — following APA reporting conventions
- Conclude with a brief *reflection* — (what challenged you + what you learned)