HCI and Design

Topics for today

Statistical significance

Simple statistical tests in HCI

Useful tools to know

Controlled experiment terminology

factor

- An independent variable, e.g., *input device*.

levels

- The possible values of a factor, e.g., *touchpad* and *trackball* are two levels of the factor *input device*.

between-subjects factor

- A factor for which each subject performs with *one level*, e.g., each subject uses the *touchpad* or the *trackball* but not both.

within-subjects factor

- A factor for which each subject performs with *all levels*, e.g., each subject uses the *touchpad* and the *trackball*.

Controlled experiment terminology

population

- All the people in the world who might be relevant to the research question asked, e.g., all potential touchpad and trackball users.

sample

- A representative portion of the whole population used in an experiment, e.g., some subset of touchpad and trackball users.

independent variable

- The variable encapsulating the conditions being tested in an experiment, e.g., *input device*.

dependent variable

- The outcome measure being used to assess differences in the independent variable, e.g., *throughput, speed, accuracy*.

Controlled experiment terminology

counterbalance

- Ordering the levels of a factor so as to avoid confounding the results, e.g., making sure half of the subjects do *touchpad* first, and half do *trackball* first in a within-subjects design.

ANOVA

- Abbreviation for "analysis of variance," which is a common statistical method used to determine if there are differences between levels of different factors.

t-test

- A simple statistical test to compare the means and distributions of two groups, that is, of two levels of a single factor, e.g., *touchpad* vs. *trackball* throughput.

p-value

- The result of a statistical test. By convention, a *p*-value less than 0.05 is deemed "statistically significant."

Significance tests

Why do we need significance tests?

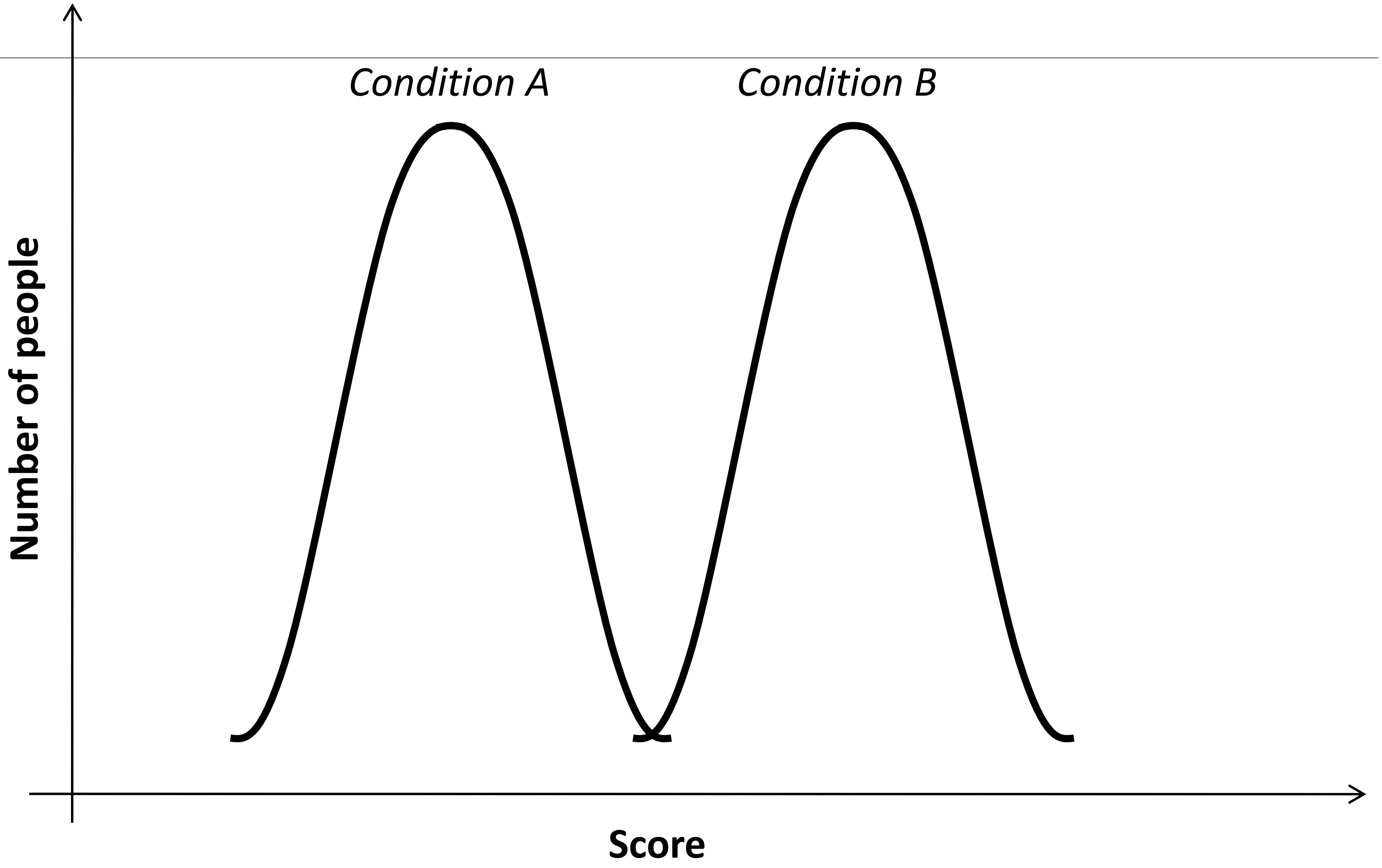
- When the values of the members of the comparison groups are all known, you can directly compare them and draw a conclusion. No significance test is needed since there is no uncertainty involved.
- When the population is large, we can only sample a sub-group of people from the entire population.
- Significance tests allow us to determine how confident we are that the results observed from the sampling population can be generalized to the entire population.

Example

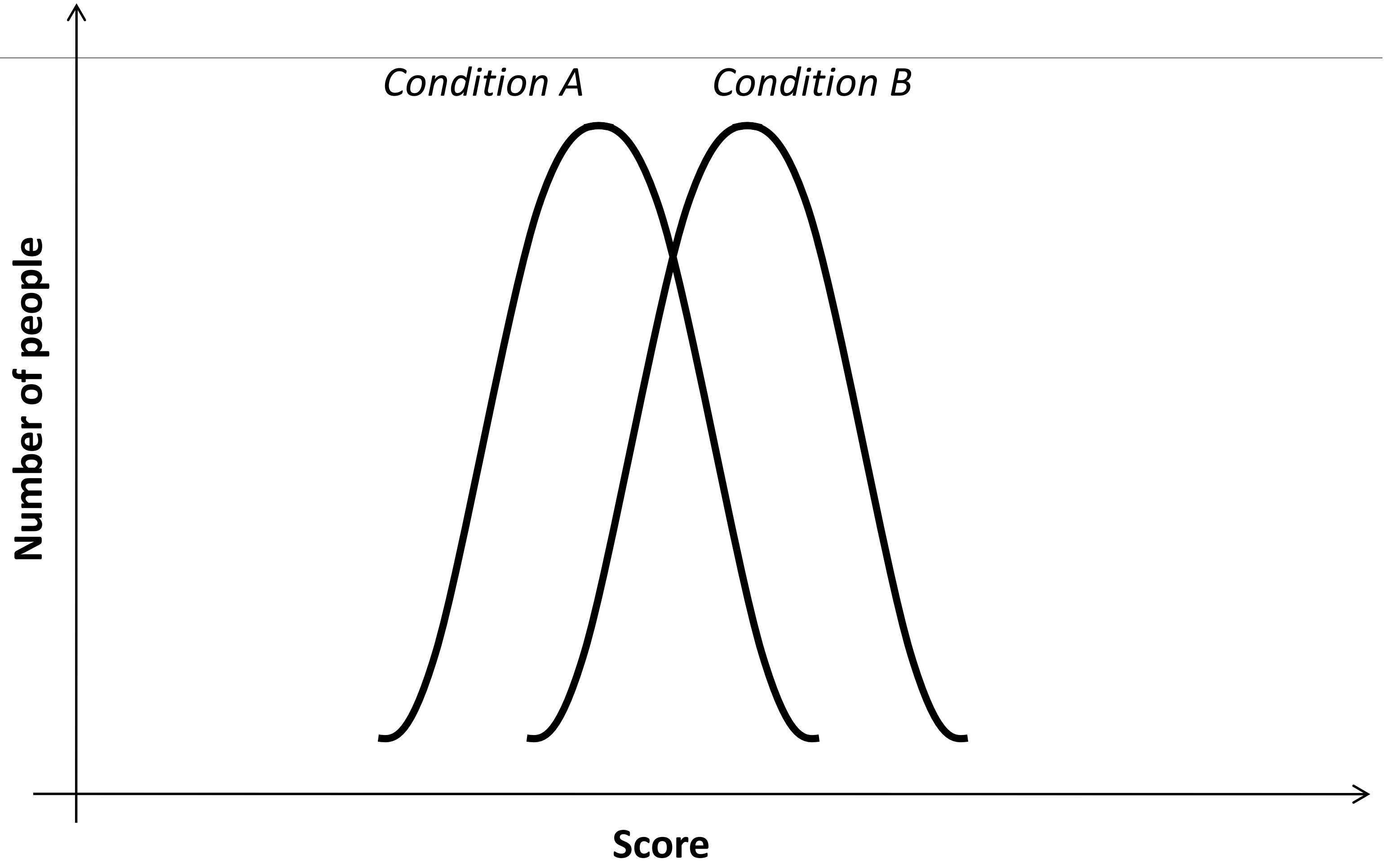
You recruit 30 people, 15 of which do a test using a touchpad, and 15 of which do the same test using a trackball. You end up with 30 measures of *throughput*. The mean for the touchpad is 4.30 clicks/s. The mean for the trackball is 5.08 clicks/s.

Can you conclude the trackball is better than the touchpad?

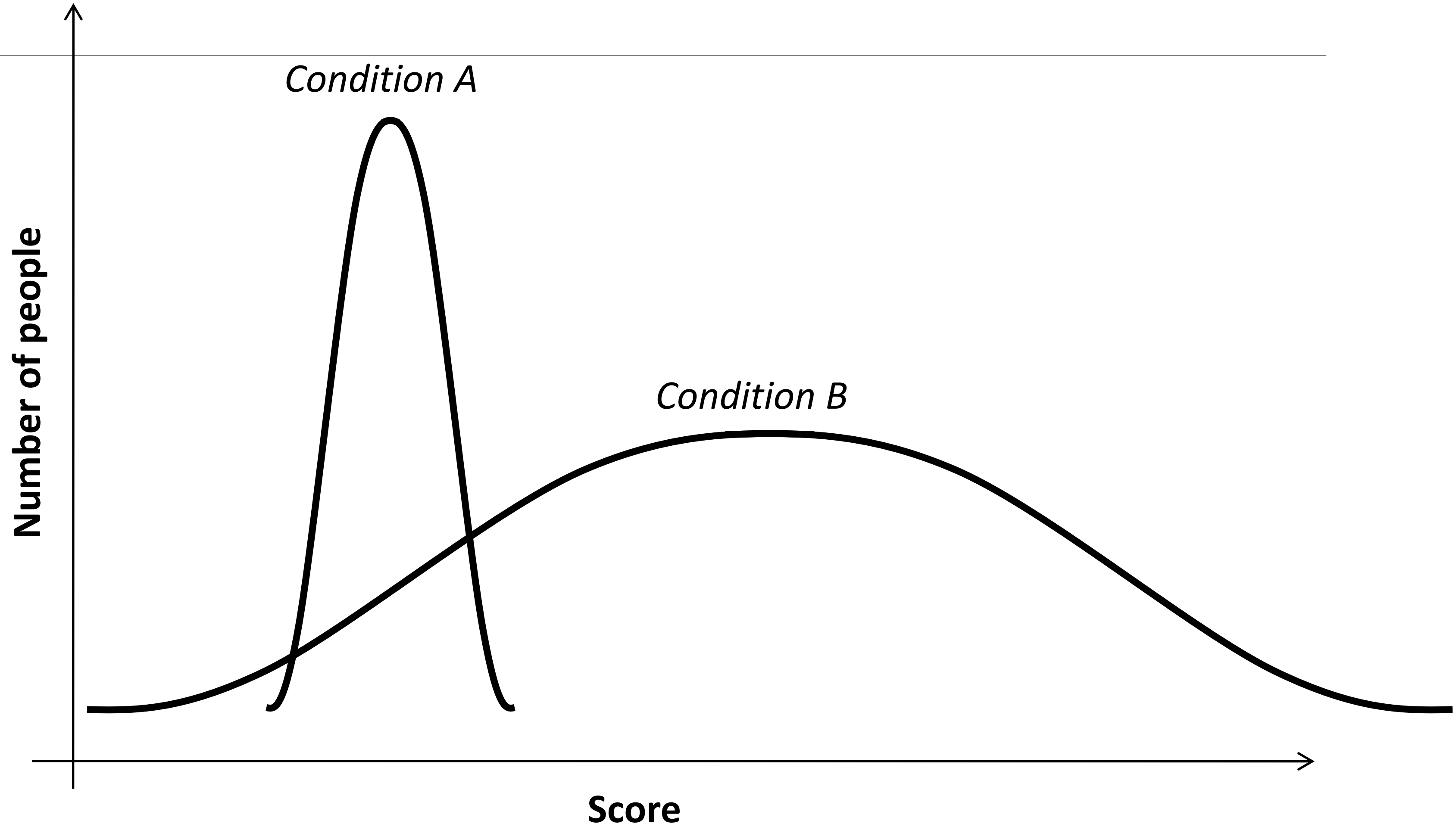
Are they different?



Are they different?



Are they different?



Bottom line

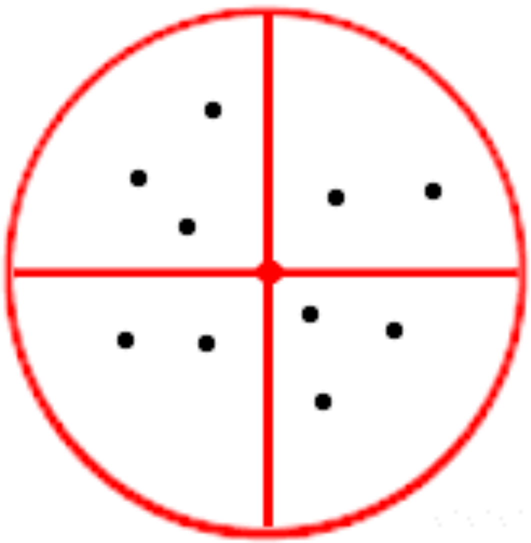
You can’t just compare means.

You must take “spreads” into account.

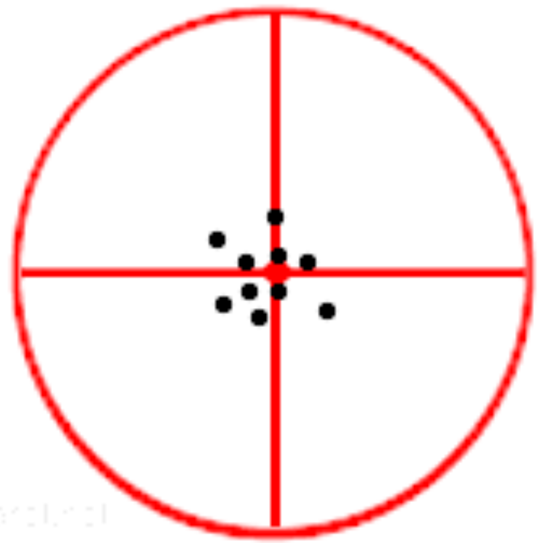
Statistics can perform *analyses of variance*, or the “amount of spread” around means to tell us how reliable/probable a real difference is.

- A real difference is a “statistically significant difference.”
- An unreliable difference is a “statistically non-significant difference.”
  - This does not prove two things are equal. Statistics cannot show equality, only difference.

Another view of *variance*

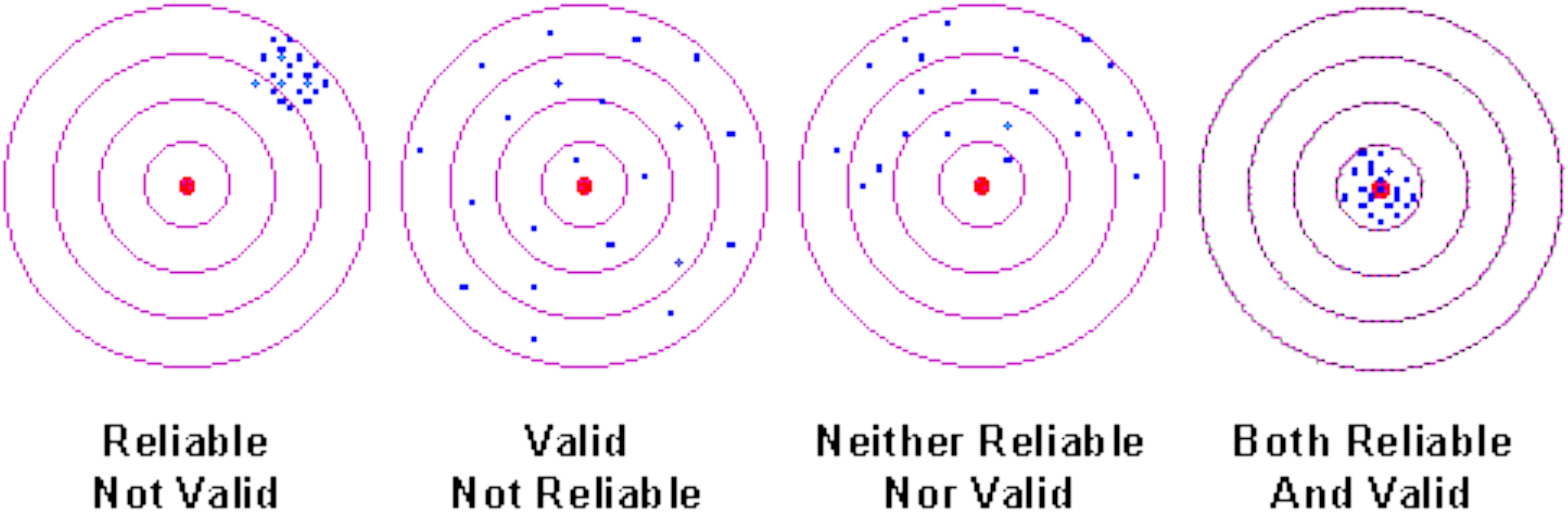


High variance



Low variance

Reliability and validity



Statistical significance

How do we determine if something is statistically significant?

Recall:

Experimental hypothesis: there is a difference between the levels

- e.g. the trackball is faster than the touchpad

Null hypothesis: there is no difference between the levels

- e.g. there is no difference between the trackball and the touchpad

Significance tests: p-values

We perform an analysis of variance and get a p-value.

The p-value comes from the sampling distribution of the sample mean.

The p-value is the probability of randomly getting a test statistic as (or more) extreme than what you observed if the null hypothesis was true.

i.e. the probability that your results occurred by chance

$p = 0.45$ means there is a 45% chance the data occurred by chance.

$p = 0.05$ means there is a 5% chance the data occurred by chance.

Significance tests

We now need to use the p-value to choose a course of action . . .

Either reject the null hypothesis, or fail to reject the null hypothesis

We need to decide if our sample result is unlikely enough to have occurred by chance.

Standard cutoff is $p < .05$ .i.e. we're at least 95% confident that our results did not occur by chance.

Errors

All significance tests are subject to the risk of Type I and II errors.

Type I errors (alpha)

- When there really is no significant difference but you say there is.
- More formally: When you incorrectly fail to accept the null hypothesis.

Type II errors (beta):

- When there really is a significant difference but you say there isn’t.
- More formally: When you incorrectly fail to reject the null hypothesis.

Type I and Type II errors

		Jury decision	
		Not guilty	Guilty
Reality	Not guilty	√	Type I error
	Guilty	Type II error	√

Table 2.3 Type I and Type II errors in the judicial case.

Type I and Type II errors

It is generally believed that Type I errors are worse than Type II errors.

Statisticians call Type I errors a mistake that involves “gullibility”.

- A Type I error may result in a condition worse than the current state.

Type II errors are mistakes that involve “blindness”

- A Type II error can cost the opportunity to improve the current state.

Controlling risks of errors

In statistics, the probability of making a Type I error is called alpha (or significance level, p value).

The probability of making a Type II error is called beta.

Alpha and beta are interrelated. Under the same conditions, decreasing alpha reduces the chance of making Type I errors but increases the chance of making Type II errors.

The statistical power of a test refers to the probability of successfully rejecting a null hypothesis when it is false and should be rejected

Simple statistical tests for HCI

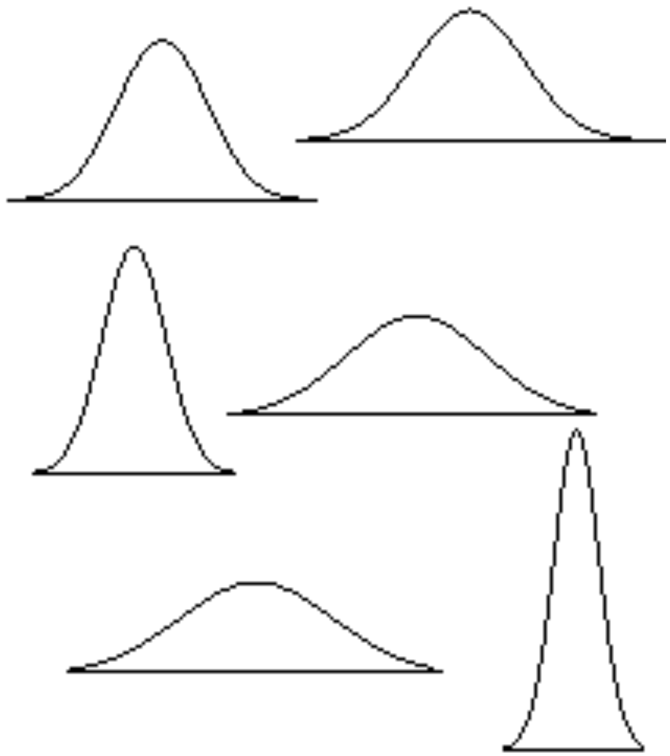
ANOVA - analysis of variance

- t-test
- F-test

ANalysis Of VAriance

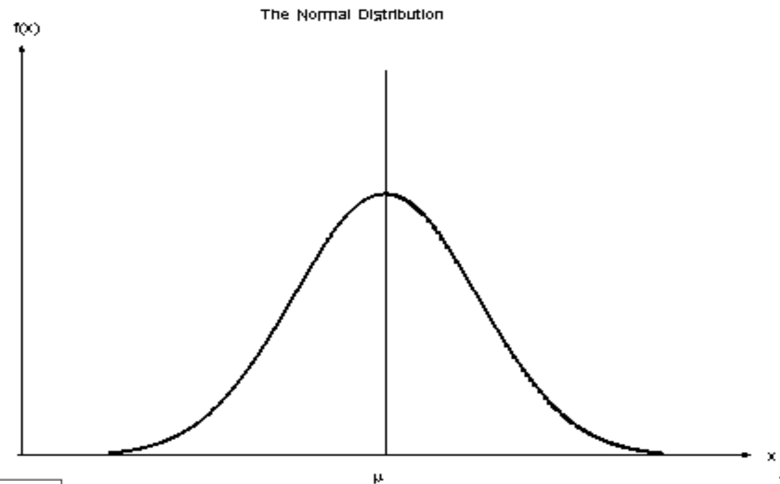
Statistical Workhorse

- Supports moderately complex experimental designs and statistical analysis
- Lets you examine differences between multiple independent variables at the same time
- Assumes a normal distribution (Bell curve)

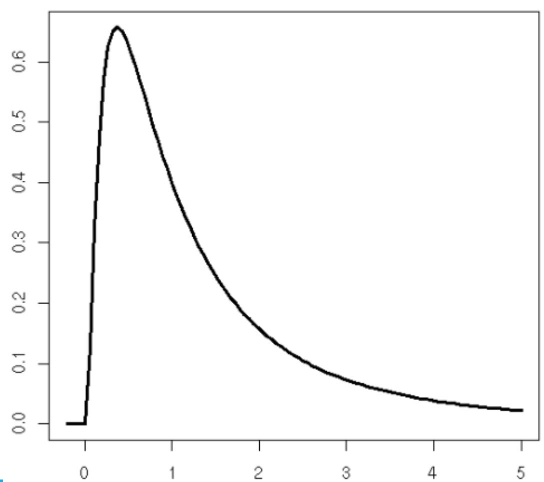


Common distributions

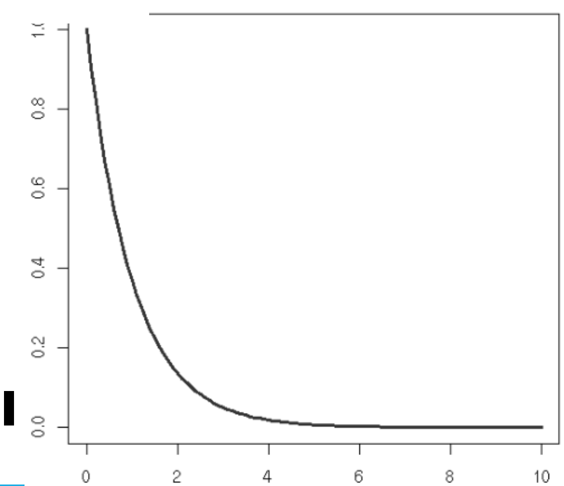
normal



lognormal

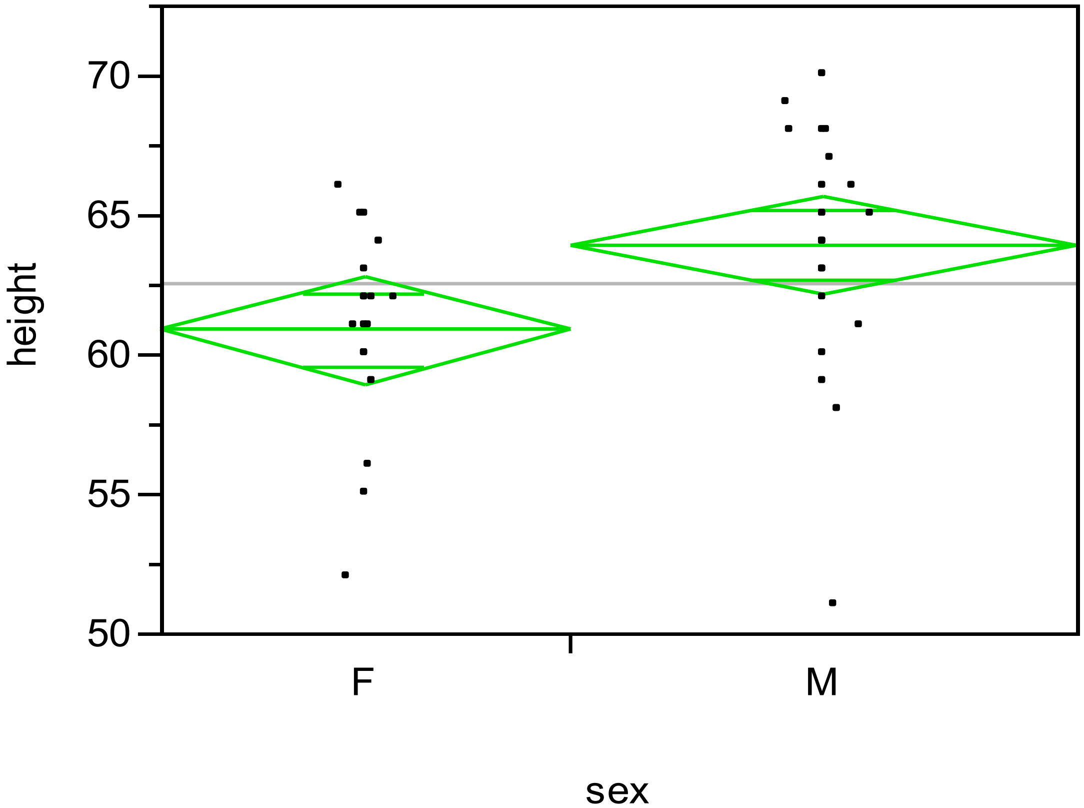


exponential



The *t* test

Simple test for differences between means on one independent variable.



Reporting a *t* test

t Ratio	3.820674
DF	14
Prob > |t|	0.0019*

"Gender had a significant effect on hours of game-play ($t(14)=3.82$, $p<.01$)."

Tests where $p>.05$ are "nonsignificant." They are not "insignificant."

- Or "not detectably different"
  - ($t(14)=1.23$, n.s.)
- Does ***not*** show equality!

Usually report p-values for only...

- $p<.05$
- $p<.01$
- $p<.001$
- $p<.0001$

“Marginal result” or “trend”

What if it’s almost significant? ($.05 < p < .10$)

Often this is called a “marginal result” or a “trend”.

Example

- “Our results indicate a nonsignificant effect of *Gender* on hours played ($t(14)=1.75$, $p=.06$), although the trend suggests that males may play more. Further experimentation is necessary to confirm this.”

More complex experiments

What if we have more than 2 levels of our factor?

What if we have multiple independent variables?

t test won’t work

The F-test

Compares relationships between many factors

In reality, we must look at multiple variables to understand what is going on

Provides more informed results

- considers the *interactions* between factors

The F-test

For one factor, same p-value as a *t* test.

But can handle >1 factors.

- Let's add *Posture* as a factor
- Levels: seated, standing

	Gender	Posture	Hours Played
1	Male	Seated	32
2	Male	Seated	39
3	Male	Standing	41
4	Male	Standing	47
5	Male	Standing	66
6	Male	Seated	21
7	Male	Seated	37
8	Male	Standing	44
9	Female	Seated	21
10	Female	Standing	19
11	Female	Seated	37
12	Female	Standing	15
13	Female	Standing	8
14	Female	Standing	18
15	Female	Seated	19
16	Female	Seated	24

Main Effect

You have a main effect when there are significant differences among levels of any one factor.

Tests of Between-Subjects Effects

Dependent Variable:HoursPlayed

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	2527.500[a]	3	842.500	11.943	.001
Intercept	14884.000	1	14884.000	210.996	.000
Gender	1722.250	1	1722.250	24.415	.000
Posture	49.000	1	49.000	.695	.421
Gender * Posture	756.250	1	756.250	10.721	.007
Error	846.500	12	70.542		
Total	18258.000	16			
Corrected Total	3374.000	15			

a. R Squared = .749 (Adjusted R Squared = .686)

Reporting main effects

There was a significant effect of *Gender* on hours played ($F(1,12)=24.41$, $p<.001$).

The effect of *Posture* on hours played was non-significant ($F(1,12)=0.69$, n.s.).

Dependent Variable:HoursPlayed

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	2527.500[a]	3	842.500	11.943	.001
Intercept	14884.000	1	14884.000	210.996	.000
Gender	1722.250	1	1722.250	24.415	.000
Posture	49.000	1	49.000	.695	.421
Gender * Posture	756.250	1	756.250	10.721	.007
Error	846.500	12	70.542		
Total	18258.000	16			
Corrected Total	3374.000	15			

a. R Squared = .749 (Adjusted R Squared = .686)

Reporting main effects

What about the interaction of Gender*Posture?

Dependent Variable:HoursPlayed

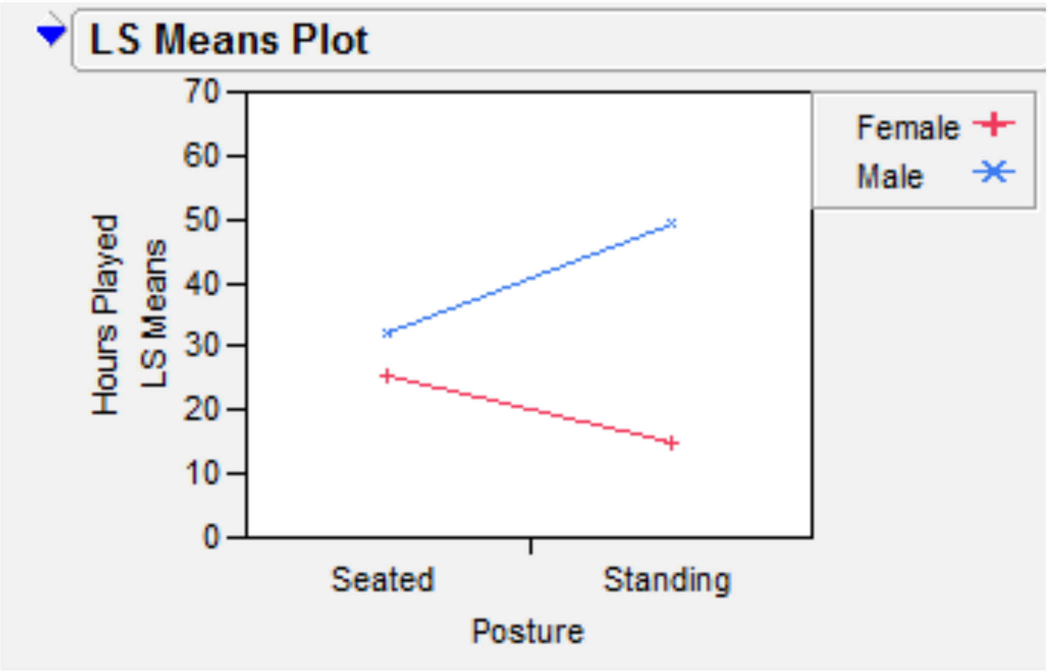
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	2527.500[a]	3	842.500	11.943	.001
Intercept	14884.000	1	14884.000	210.996	.000
Gender	1722.250	1	1722.250	24.415	.000
Posture	49.000	1	49.000	.695	.421
Gender * Posture	756.250	1	756.250	10.721	.007
Error	846.500	12	70.542		
Total	18258.000	16			
Corrected Total	3374.000	15			

a. R Squared = .749 (Adjusted R Squared = .686)

Interaction Effects

You have an interaction effect when the levels of one factor cause significant changes in the dependent variable for the levels of another factor.

i.e. levels change but in different ways



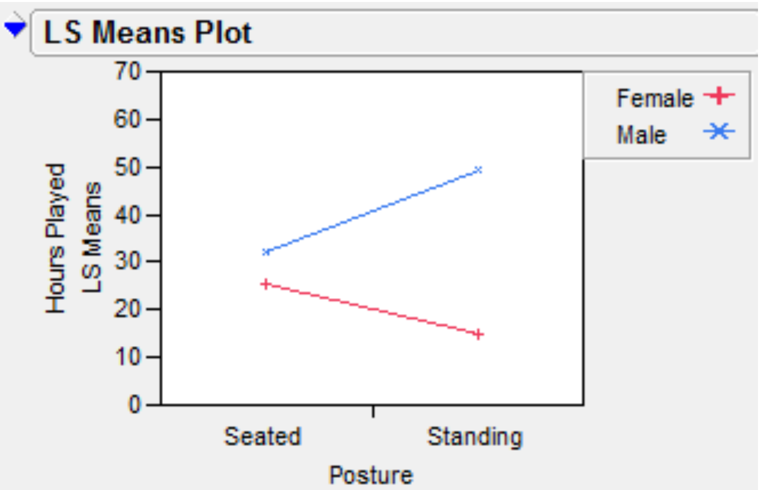
Reporting interactions

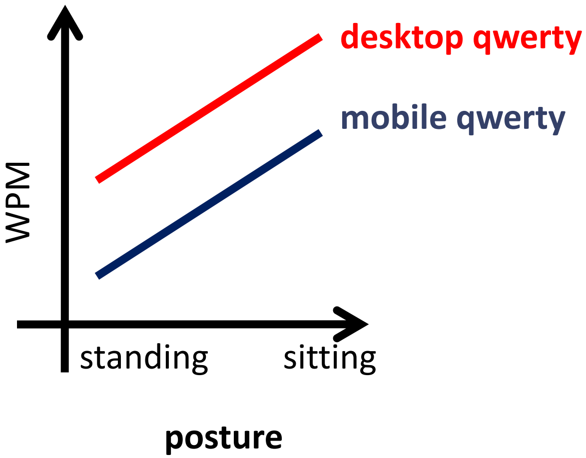
There was a significant *Gender*Posture* interaction ($F(1,12)=10.72$, $p<.01$).

Effect Tests

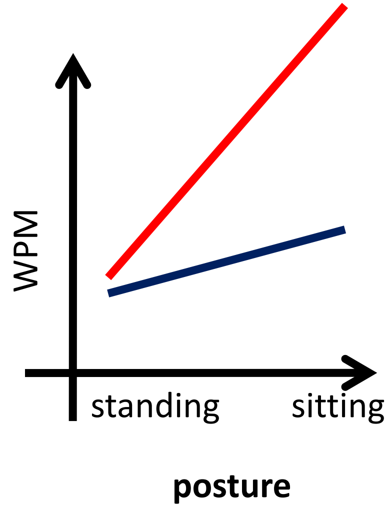
Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Gender	1	1	1722.2500	24.4146	0.0003*
Posture	1	1	49.0000	0.6946	0.4209
Gender*Posture	1	1	756.2500	10.7206	0.0067*

"An examination of our data reveals that females played less while standing than sitting, but males played more."

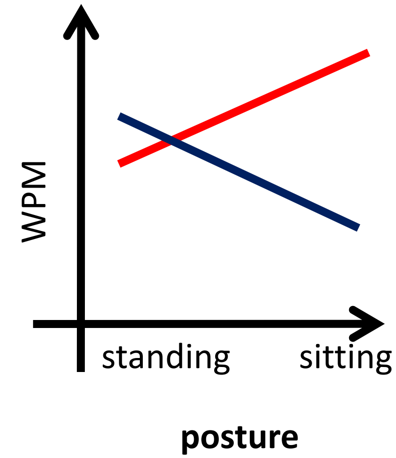




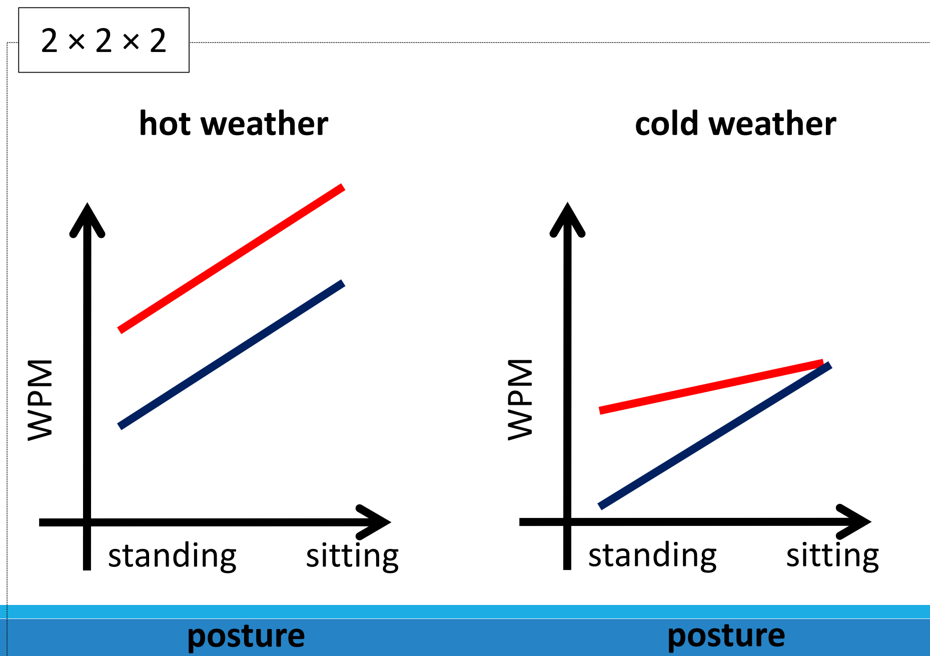
Main effect of *keyboard type*.
Main effect of *posture*.
No interaction between
keyboard type and *posture*.



Main effect of *keyboard type*.
Main effect of *posture*.
Interaction between
keyboard type and *posture*.



Main effect of *keyboard type*.
No main effect of *posture*.
Interaction between
keyboard type and *posture*.



Main effect of posture.
Main effect of keyboard type.
Main effect of weather.
Posture*keyboard type (probably)
Keyboard type*weather (probably)
Posture*weather (probably not)
Posture*keyboard*weather (definitely)

Tools for statistical analysis

JMP

- Stats package from SAS Institute
- Trial version available: http://www.jmp.com/software/

SPSS

- Stats package owned by IBM
- Trial version available: http://www.spss.com/downloads/

R

- Open source command-line statistics package
- Very powerful, extensible, and **FREE**
- Difficult to learn, but powerful in the end
- Lots of online resources to help: http://www.r-project.org/