CSE 510: Advanced Topics in HCI

Experimental Design
and Statistical Analysis

James Fogarty
Daniel Epstein

Tuesday / Thursday
10:30 to 12:00

CSE 403

design:
use:
build:

University of Washington

Introduction

Experiments and statistics are not always "the right way" to do things in HCI or CS

- Hopefully we have established that by now

But you should come to understand effective experimental design and statistical analysis

- In designing, running, analyzing your own studies
- In reading / reviewing studies by others

Should be useful within and outside HCI

Introduction

Really good experiments are an art,
and can represent a breakthrough in a field

Why?

Introduction

Really good experiments are an art,
and can represent a breakthrough in a field

- Many things to account for in design
- Unexpected twists arise in analysis
- Small differences matter

And there are a ton of statistical tools out there,
more than you can learn in one day or course

- Remember your statistics course?

A Pragmatic Approach

So how do you get anything done?

A Pragmatic Approach

So how do you get anything done?

Beg: Learn who you can ask for help

Borrow: Learn and use effective patterns

Re-use designs you have used in the past

Look at papers published by good people

Steal: Do not get “caught” by your design

Learn how to recognize when over your head, when assumptions do not feel right

A Pragmatic Approach

Today is not about the many procedures you might learn in the abstract, but a handful that you are likely to repeatedly encounter in HCI

I strongly believe you learn statistics because you understand and apply them in your research, not because an instructor reviews them

Also keywords for how you can learn more

Design and Statistics

Even a seemingly simple experiment can be difficult or impossible to correctly analyze

Why?

Design and Statistics

Even a seemingly simple experiment can be difficult or impossible to correctly analyze

Design and analysis are inseparable

Consider your experiment and analyses together, to avoid running an experiment you cannot analyze

Design isolates a difference, statistics test it

Causality and Correlation

We cannot prove causality

- We can only show strong evidence for it
- Always something outside the scope of an experiment that could be the true cause

We can show correlation

- Treatment changes, so does outcome
- Hold all things equal except for one
- Eliminate possible rival explanations

Causality and Correlation

A negative result means little or nothing

- A given experiment failed to find a correlation, but that does not mean there is not a correlation, nor the experimental conditions are "equal"

See power analysis

- probability of correctly rejecting the null hypothesis (H0) when the alternative hypothesis (H1) is true
- Conceptually important, but not common in HCI

Why?

Internal and External Validity

Internal Validity

Convincingly link treatments to effects and the experiment is said to have high internal validity, it shows an effect

External Validity

An experiment likely to generalize beyond the things directly tested is said to have high external validity

Often at odds with each other

Why?

Achieving Control

Avoiding other plausible explanations

- Often referred to as confounds

General Strategies

- Remove and/or exclude
- Measure and adjust (i.e., with pre-test)
- Spread effect equally over all groups
  - Randomization (i.e., assign randomly)
  - Blocking / Stratification (i.e., assign balanced)

Variable Terminology

Factors – Variables of interest

(i.e., one variable is a single-factor experiment)

Levels – Variation within a factor

(i.e., factors are not necessarily binary)

Independent Variables

Variables you control

Dependent Variables

Your outcome measures

(they depend on your independent variables)

Factorial Designs

May have more than one factor

Factors may have multiple levels

A 2x2x3 study has two factors of two levels each and a third factor with three levels

Text entry method {Multitap, T9} x
Number of hands {one, two} x
Posture {seating, standing, walking}

Some potential dependent variables?

Within and Between Subjects

Within-Subjects Designs

- Each participant experiences multiple levels
- Much more statistically powerful,
  but much harder to avoid confounds

Between-Subjects Designs

- Each participant experiences only one level
- Avoids possible confounds,
  easier to statistically analyze,
  requires more participants

Why more participants?

Carryover Effects

For example: learning effects, fatigue effects

Counterbalanced designs help mitigate

e.g., Latin square

A	B	C	D
C	D	A	B
D	C	B	A
B	A	D	C

“Uncommon” / Special Designs

Some areas of research features experimental designs that are otherwise “uncommon”

Why?

“Uncommon” / Special Designs

Some areas of research features experimental designs that are otherwise “uncommon”

- Often based in solutions to likely confounds

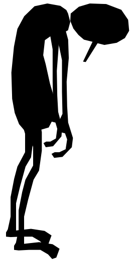
For example, “Wait List” interventions

- Self-selection effects
- Ethical dilemmas

Non-random cross-validation

- Sensor drift in physiological studies

Ethical Considerations



Testing is stressful, can be distressing

- People can leave in tears

You have a responsibility to alleviate

- Make voluntary with informed consent
- Avoid pressure to participate
- Let them know they can stop at any time
- Stress that you are testing the system, not them
- Make collected data as anonymous as possible

Human Subjects Approvals

Research requires human subjects review of process

This does not formally apply to your coursework

But understand why we do this and check yourself

Companies are judged in the eye of the public

Public Announcement

WE WILL PAY YOU $4.00 FOR ONE HOUR OF YOUR TIME

Persons Needed for a Study of Memory

*We will pay five hundred New Haven men to help us complete a scientific study of memory and learning. The study is being done at Yale University.

*Each person who participates will be paid $4.00 (plus 50c carfare) for approximately 1 hour's time. We need you for only one hour: there are no further obligations. You may choose the time you would like to come (evenings, weekdays, or weekends).

*No special training, education, or experience is needed. We want:

Factory workers	Businessmen	Construction workers
City employees	Clerks	Salespeople
Laborers	Professional people	White-collar workers
Barbers	Telephone workers	Others

All persons must be between the ages of 20 and 50. High school and college students cannot be used.

*If you meet these qualifications, fill out the coupon below and mail it now to Professor Stanley Milgram, Department of Psychology, Yale University, New Haven. You will be notified later of the specific time and place of the study. We reserve the right to decline any application.

*You will be paid $4.00 (plus 50c carfare) as soon as you arrive at the laboratory.

TO:
PROF. STANLEY MILGRAM, DEPARTMENT OF PSYCHOLOGY, YALE UNIVERSITY, NEW HAVEN, CONN. I want to take part in this study of memory and learning. I am between the ages of 20 and 50. I will be paid $4.00 (plus 50c carfare) if I participate.

NAME (Please Print). . . . .

ADDRESS . . . .

TELEPHONE NO. . . . . Best time to call you . . . .

AGE . . . . OCCUPATION . . . . SEX . . . . . .

CAN YOU COME:

WEEKDAYS . . . . EVENINGS . . . . . . WEEKENDS . . . .

Design and Statistics

Now that our design has allowed us to isolate what appears to be a difference,
we need to test whether it actually is

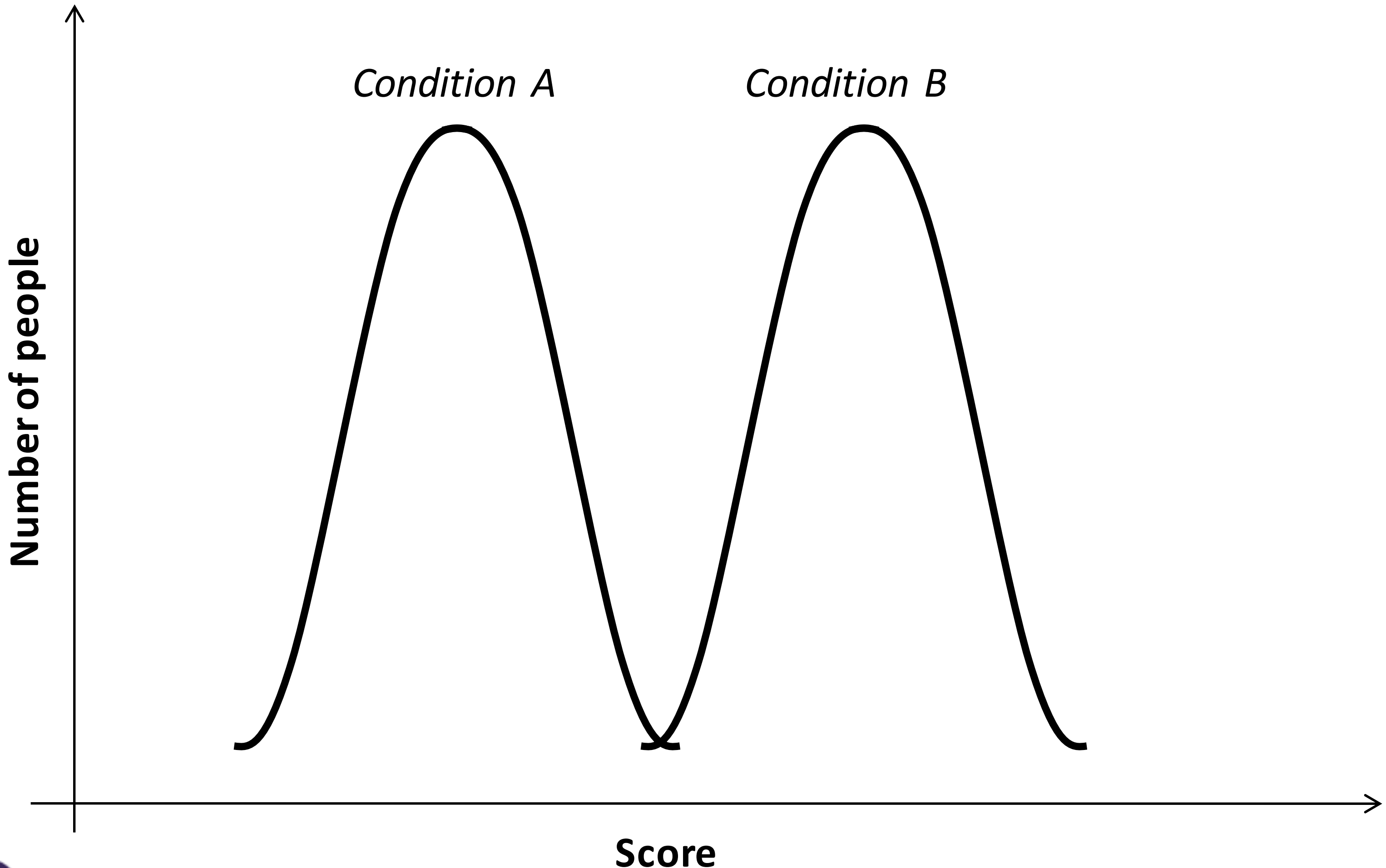
Test whether large enough,
in light of variance,
to indicate an actual difference

Simple Analysis

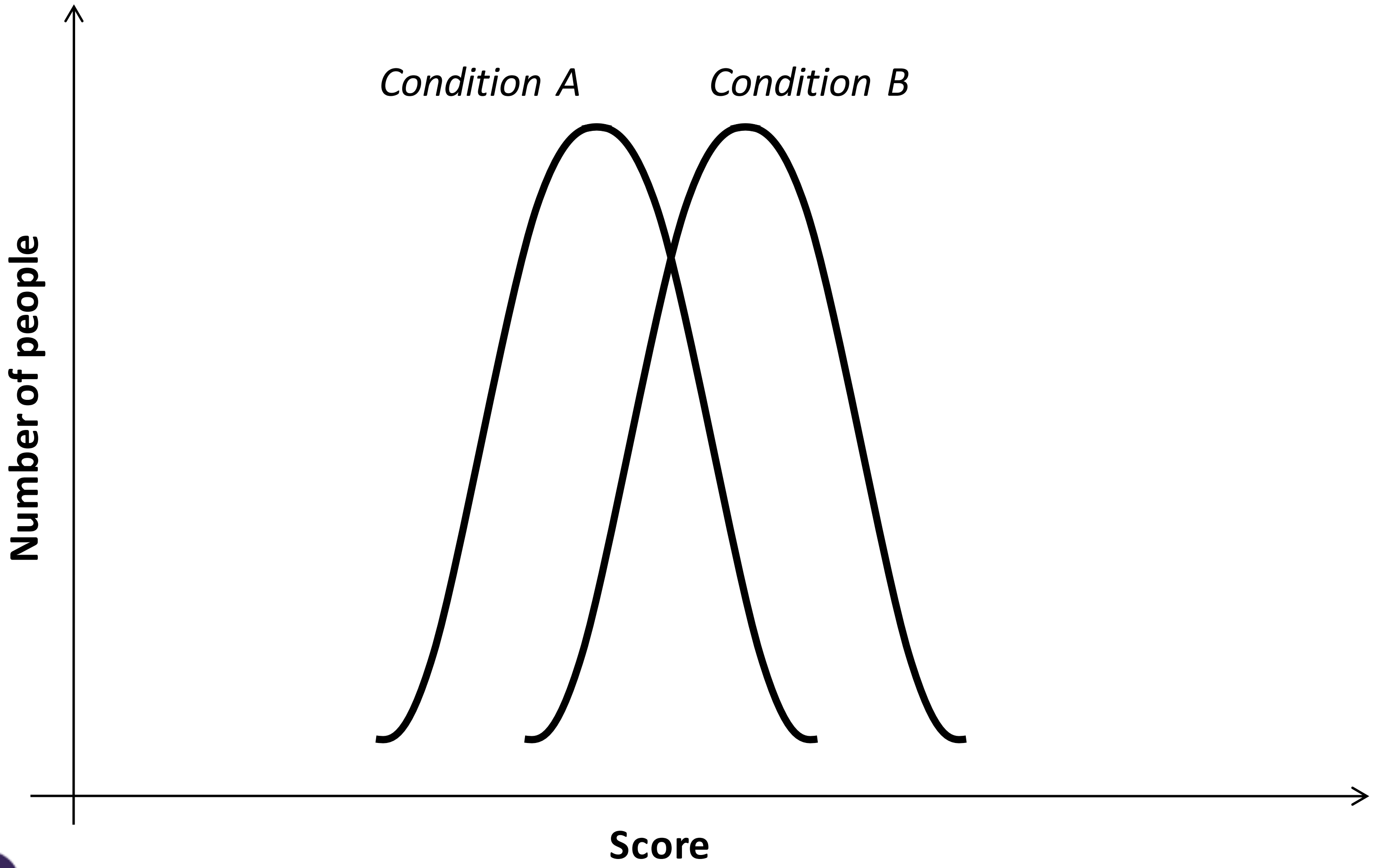
Two conditions, *Condition A* and *Condition B*

A common analysis we might conduct is to determine whether there is a significant difference between Condition A and Condition B

Difference?

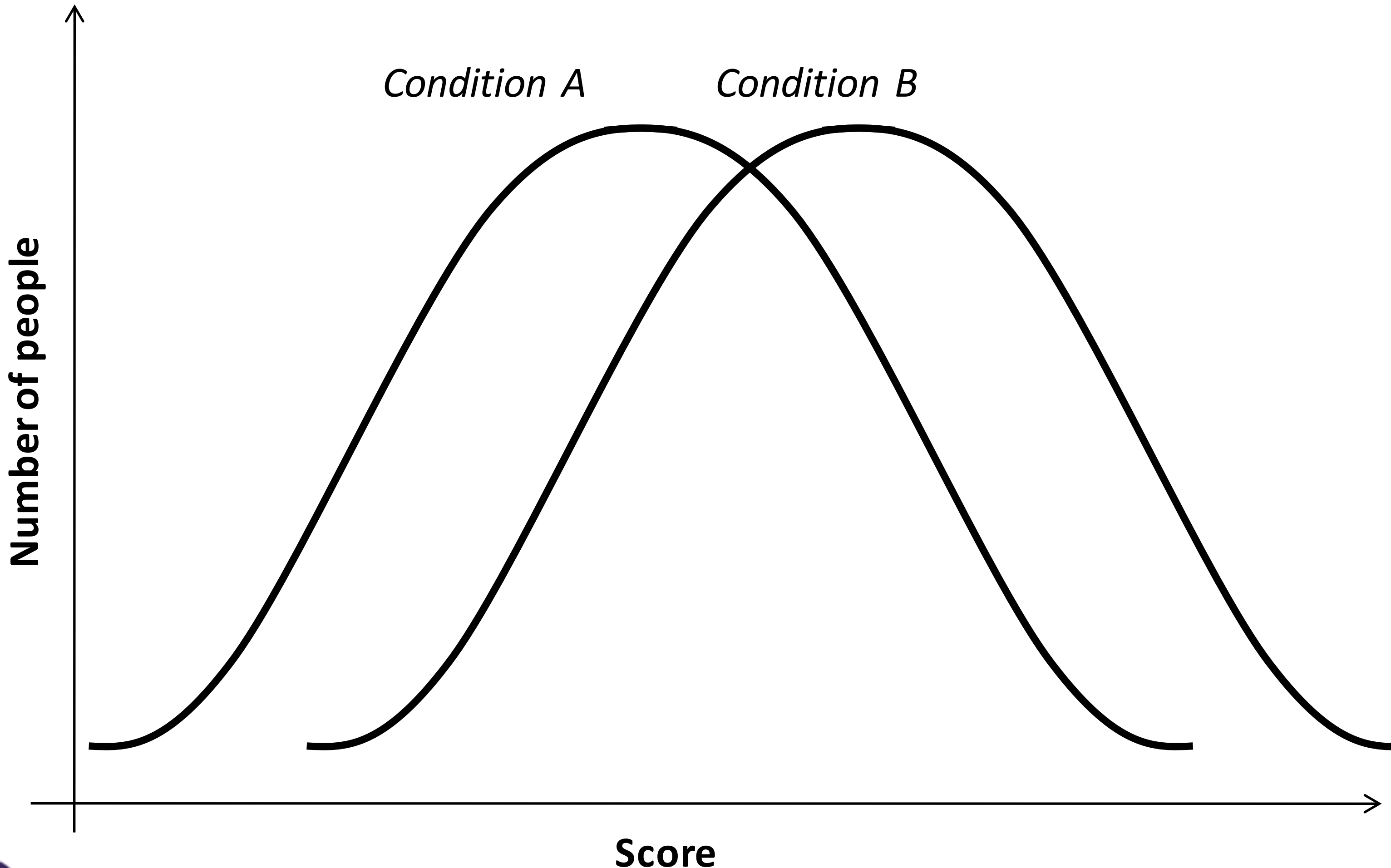


Difference?



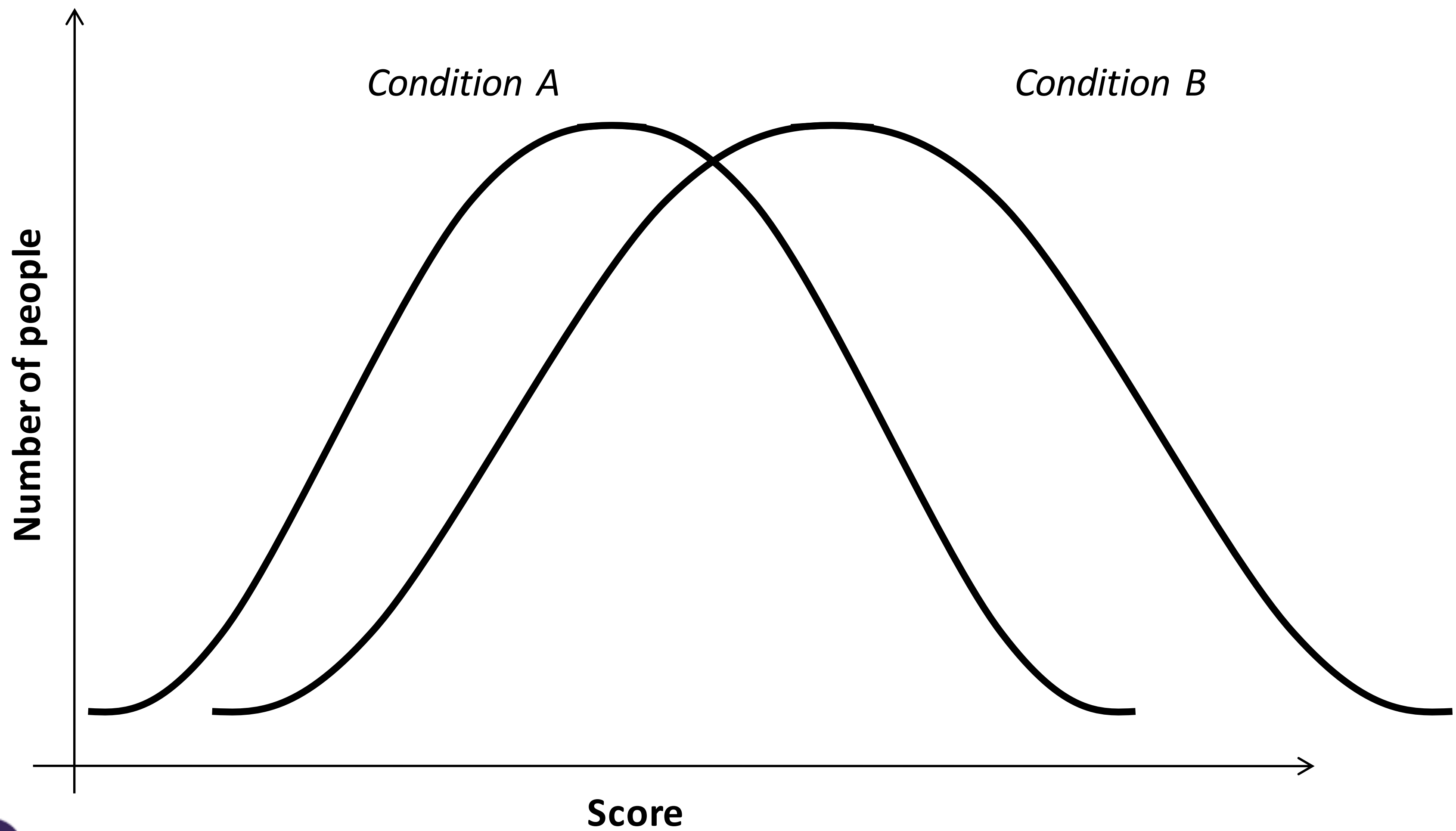
University of Washington

Difference?

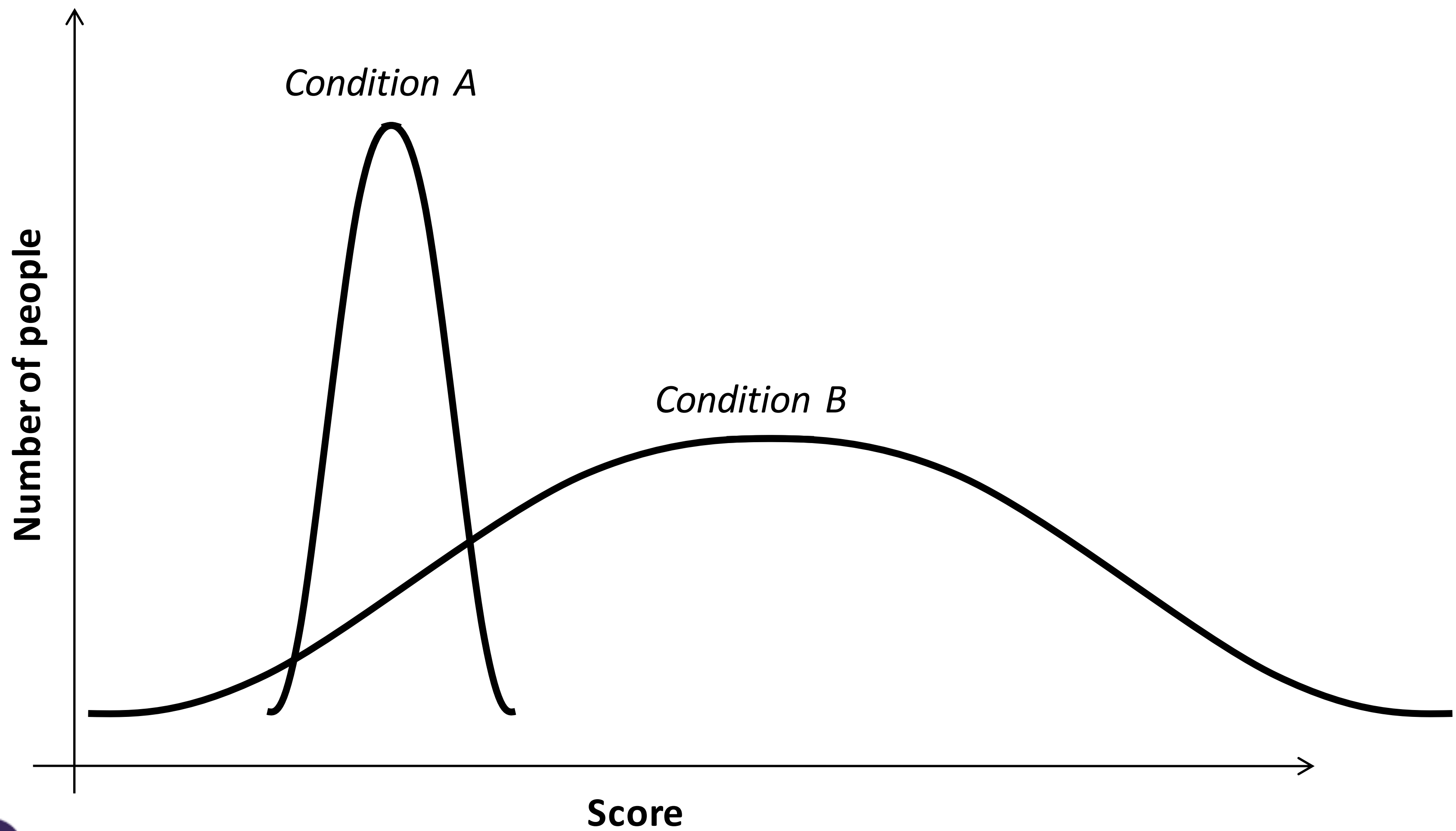


University of Washington

Difference?



Difference?



Difference

You cannot only compare means

You must take "spreads" into account

$$SD = \sqrt{\frac{\sum(X - \overline{X})^2}{n-1}}$$

Standard deviation (square root of variance), often preferred because it retains same units and magnitude

p values

The statistical significance of a result is often summarized as a p value

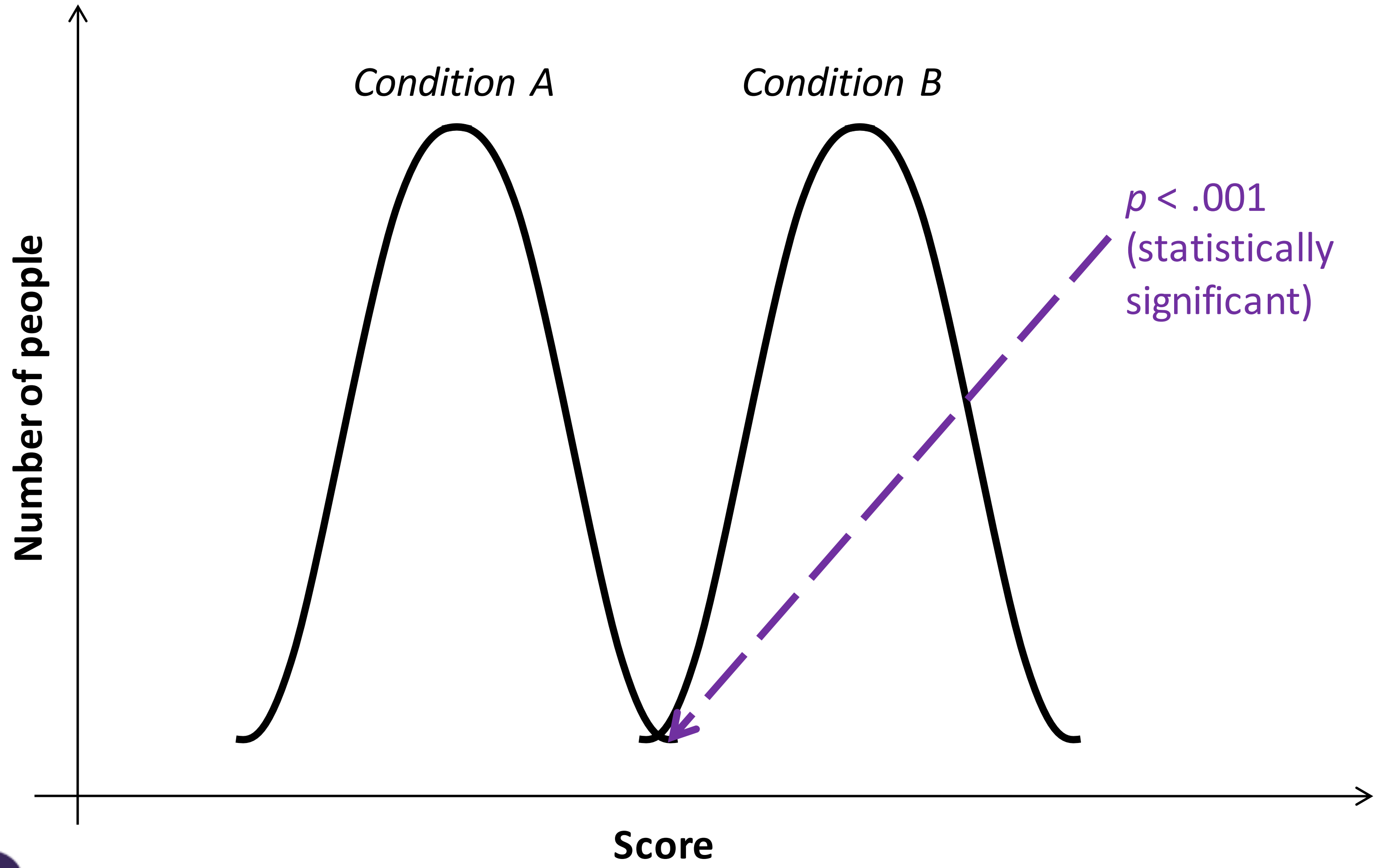
p is the probability the null hypothesis is true (there is no difference between conditions)

The same experiment, run 1 / *p* times, would generate this result by random chance

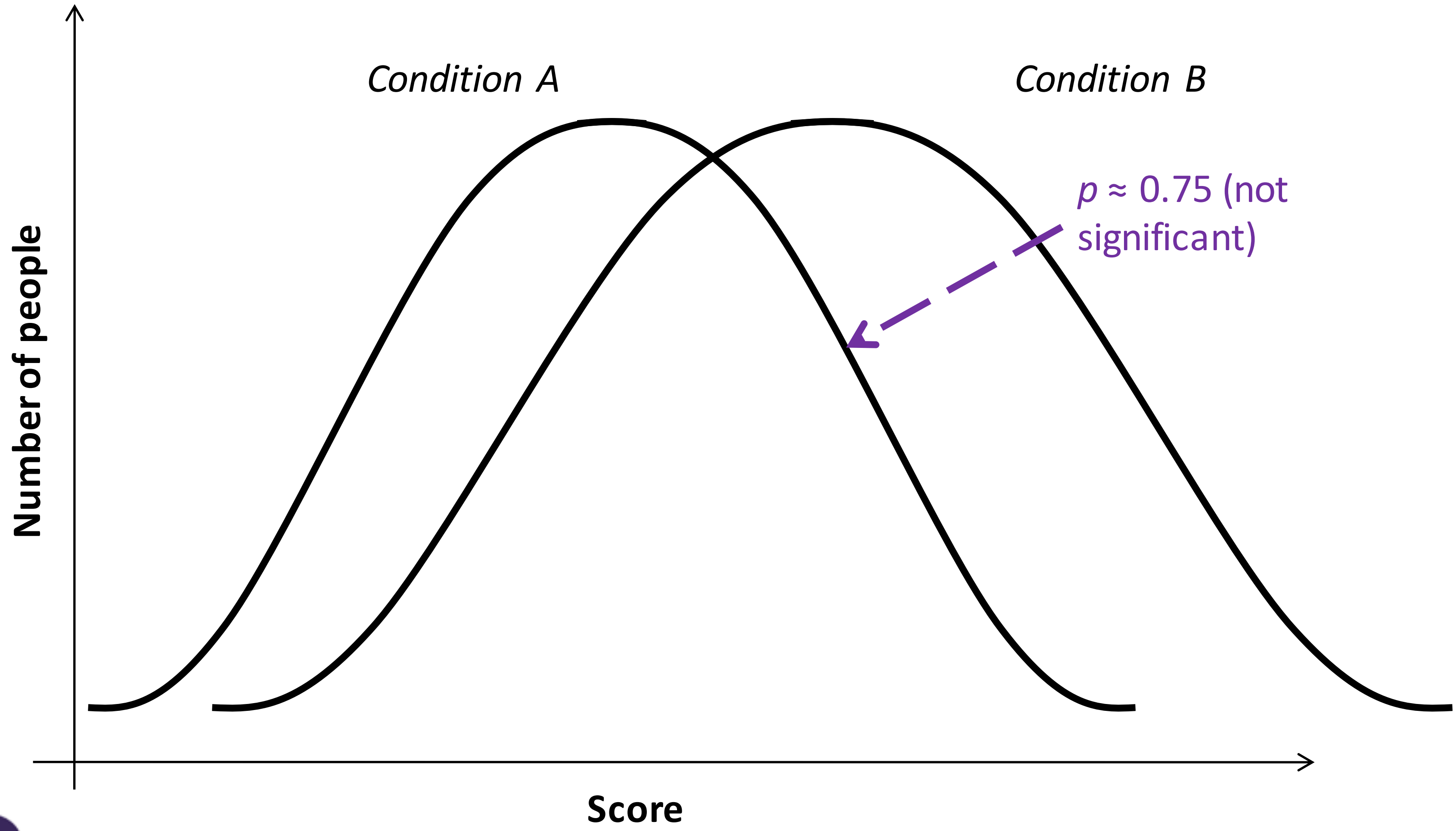
$p < .05$ is an arbitrary but widely used threshold of statistical significance

Report your *p*
Not just the comparison
And show your work

Difference?


Condition A
Condition B
p < .001
(statistically
significant)
Number of people
Score

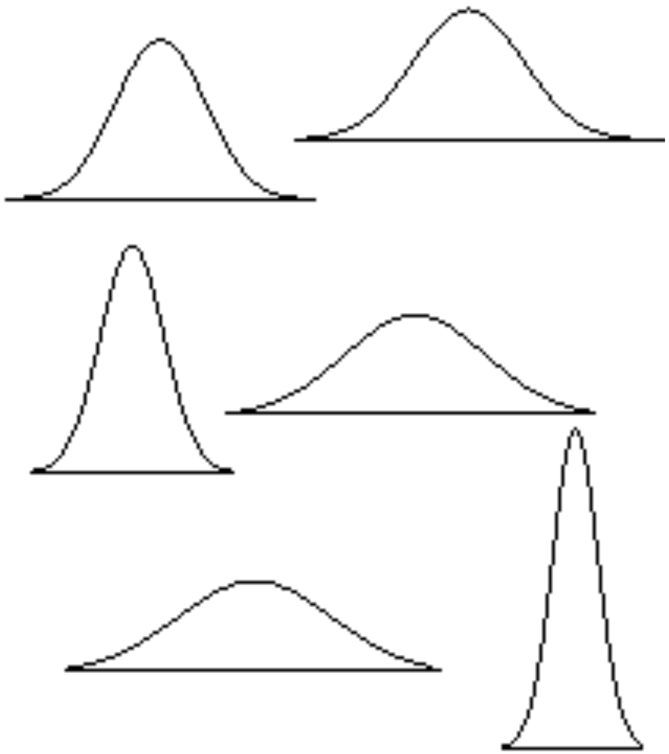
Difference?


Condition A
Condition B
p ≈ 0.75 (not significant)
Number of people
Score

p and Normal Distributions

Given a mean and a variance, assuming a Normal distribution allows estimating the likelihood of a value

Thus, parametric tests (most common tests) assume data is from normal distributions

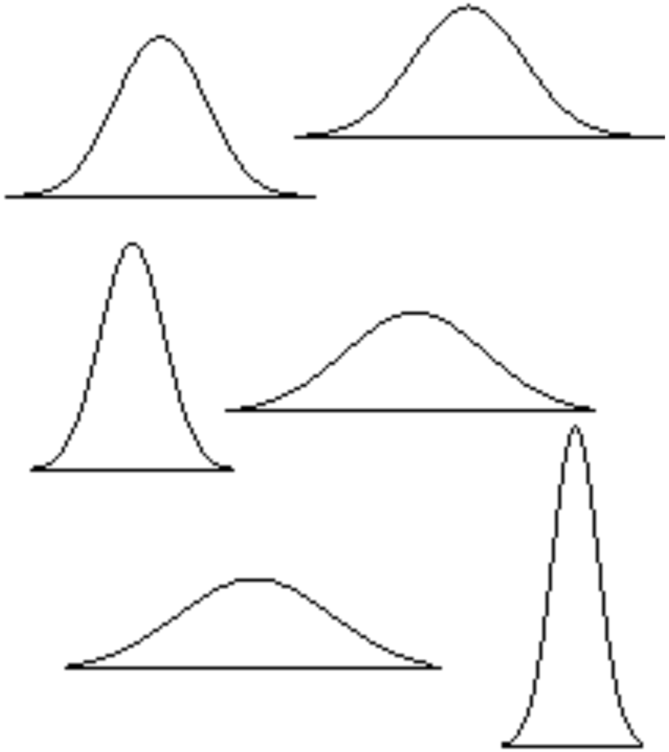


p and Normal Distributions

This is often a fair assumption

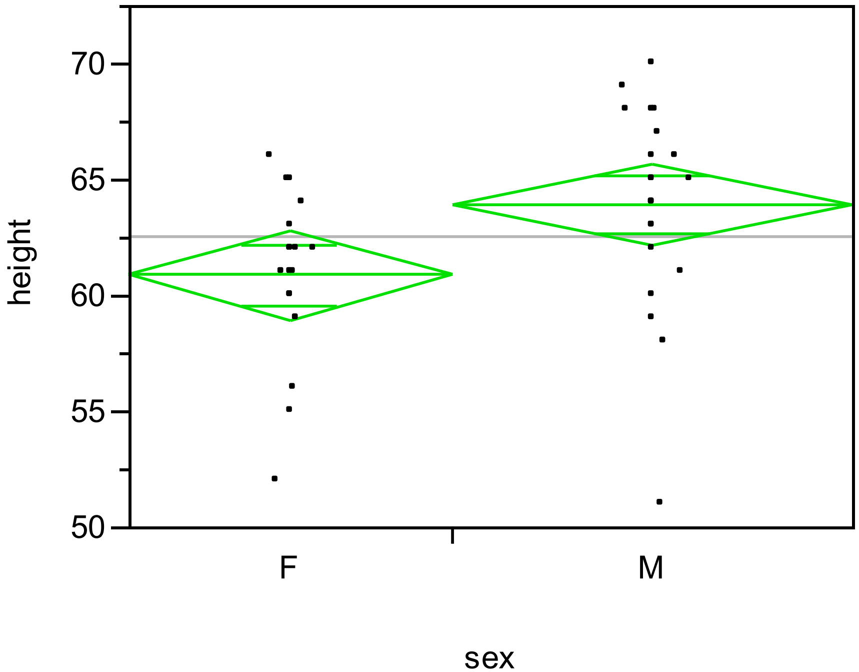
Central Limit Theorem:

Under certain conditions, the mean will be approximately normally distributed given a large enough sample



The t test

Simple test for differences between means on one independent variable



University of Washington

One-Way ANOVA

A t test is a "one-way" analysis of variance

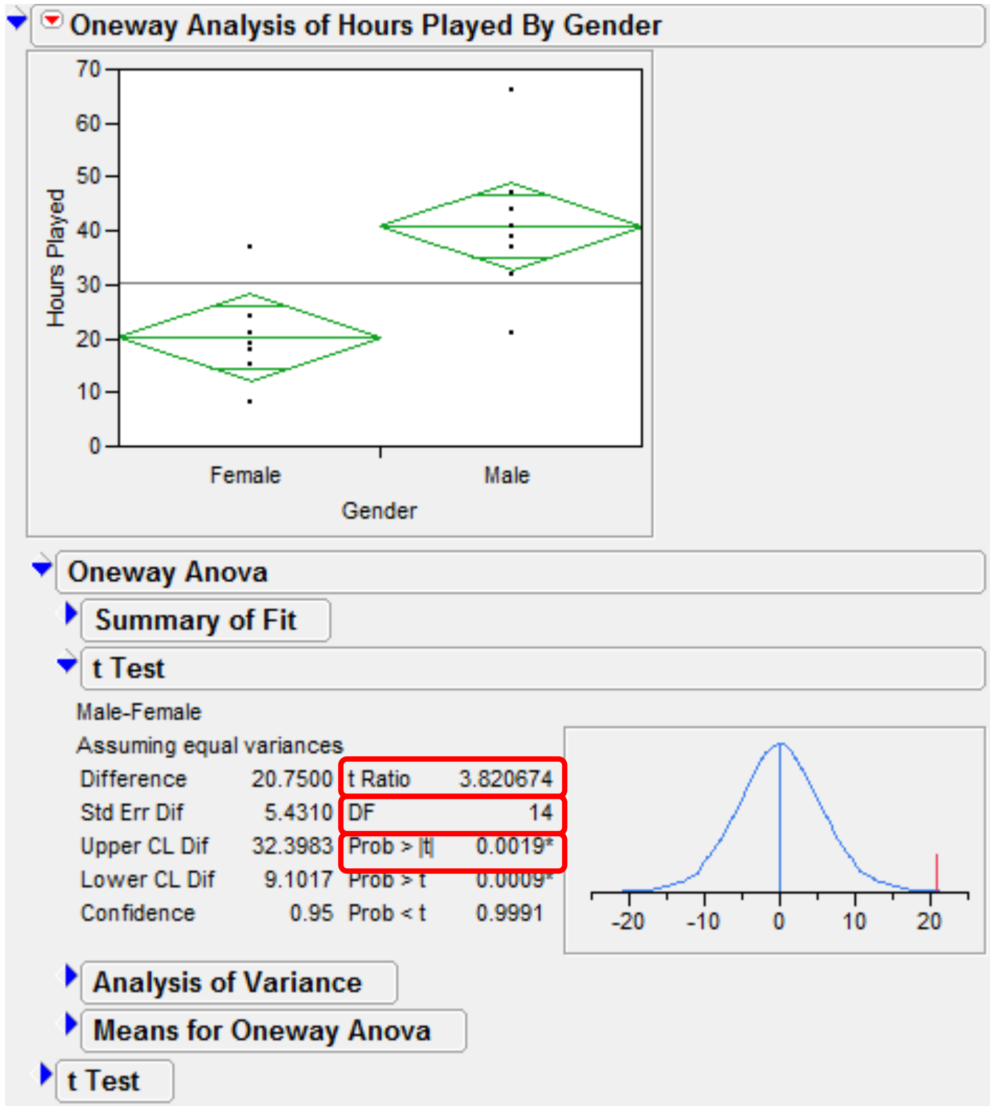
One independent variable, N > 1 levels

Example

Hours of game-play for 8 males and
8 females during the course of one week

Gender is a single factor with 2 levels (M/F)

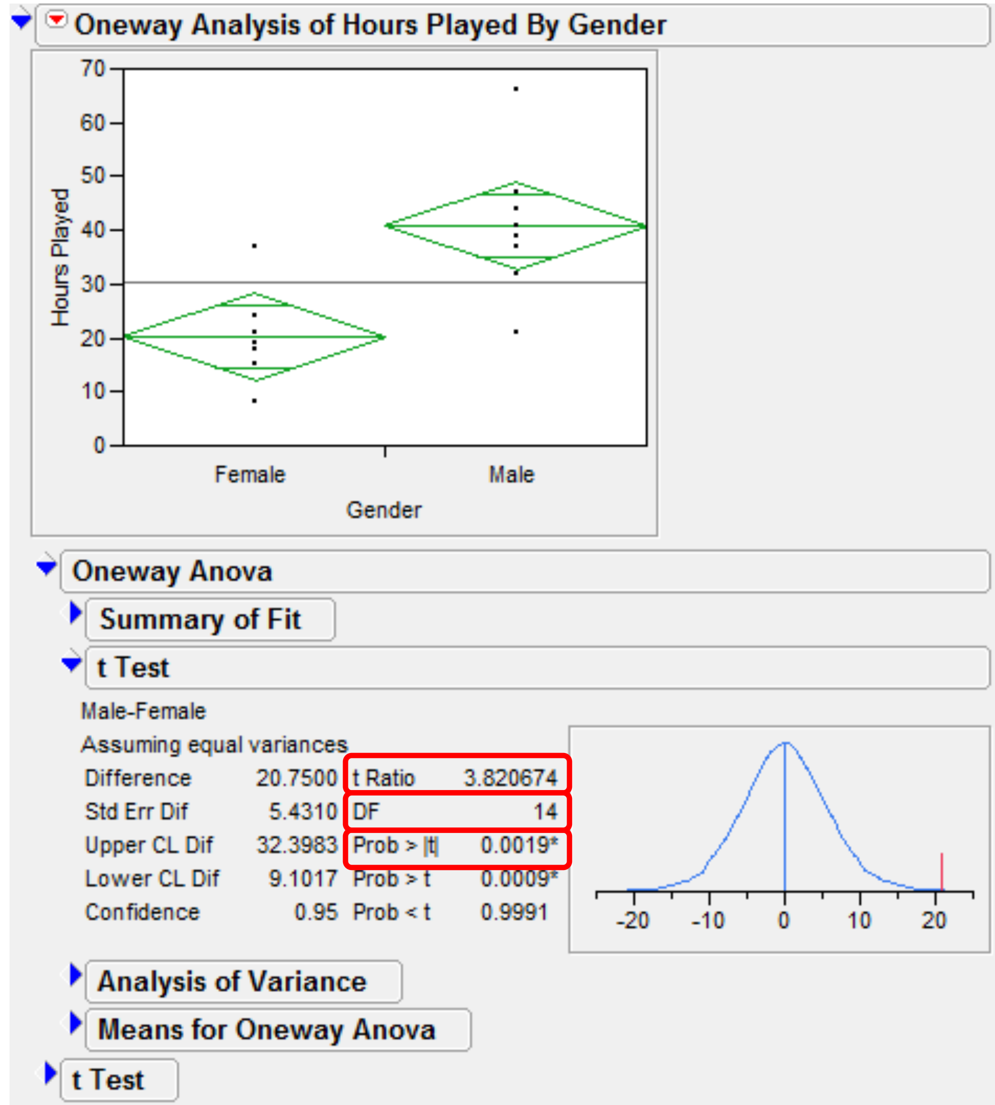
A t test Result



A t test Result

"Gender had a significant effect on hours of game-play ($t(14)=3.82$, $p \approx .002$)"

Show your work, resist the urge to report only *p*



University of Washington

The F-test

With one factor, gives the same *p* value as a t test

But can also handle multiple factors

We will add Posture

	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

The F-test

Based in a linear regression,
fitting an equation to the dependent variable

$v = ax + by + z$

$x = (0, 1)$, gender is “male”

$y = (0, 1)$, posture is “standing”

$a = ?$ $b = ?$ $z = ?$

ANOVA table

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	2527.5000	842.500	11.9433
Error	12	846.5000	70.542	Prob > F
C. Total	15	3374.0000		0.0006*

Parameter Estimates

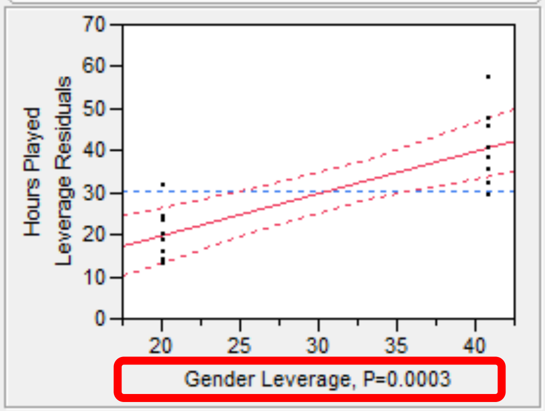
Effect Tests

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*

Main Effects

Gender

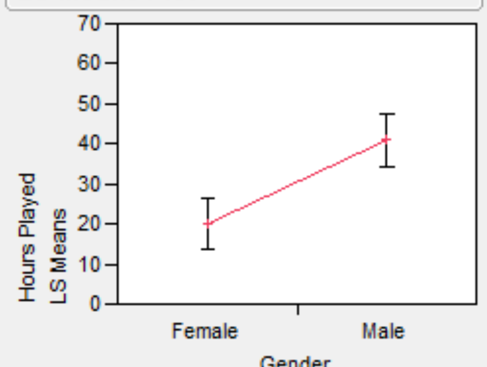
Leverage Plot



Least Squares Means Table

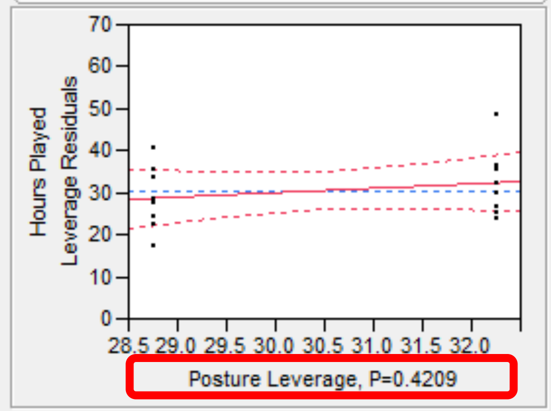
Level	Least Sq Mean	Std Error	Mean
Female	20.125000	2.9694626	20.1250
Male	40.875000	2.9694626	40.8750

LS Means Plot



Posture

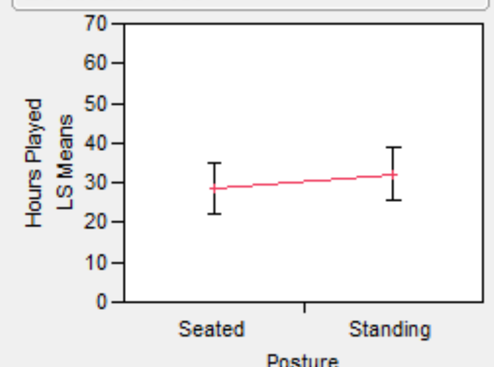
Leverage Plot



Least Squares Means Table

Level	Least Sq Mean	Std Error	Mean
Seated	28.750000	2.9694626	28.7500
Standing	32.250000	2.9694626	32.2500

LS Means Plot



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 not significant ($F(1,12)=0.69$, $p \approx .42$)

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	2527.5000	842.500	11.9433
Error	12	846.5000	70.542	Prob > F
C. Total	15	3374.0000		0.0006*

Parameter Estimates

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*

(this screenshot is a different presentation format than you will encounter in the analyses you perform in your assignment)

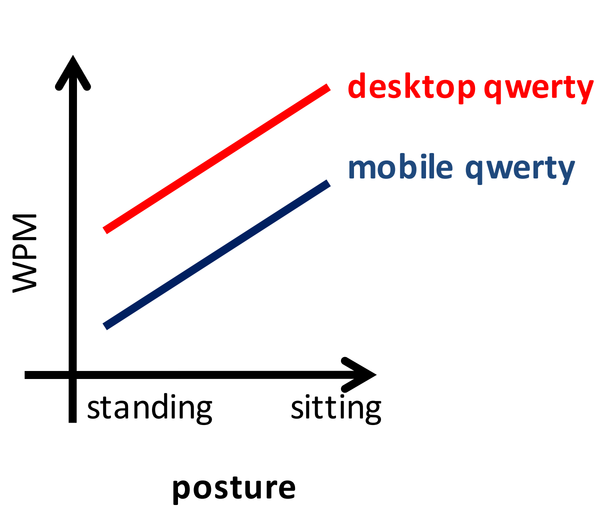
Interactions

Gender has a significant effect on hours played, and Posture does not

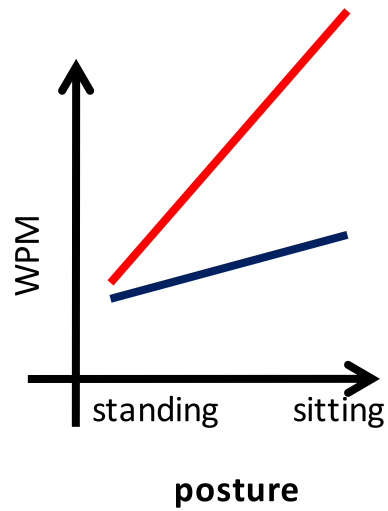
But these two effects are not independent, so we consider whether there is an *interaction effect*

	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

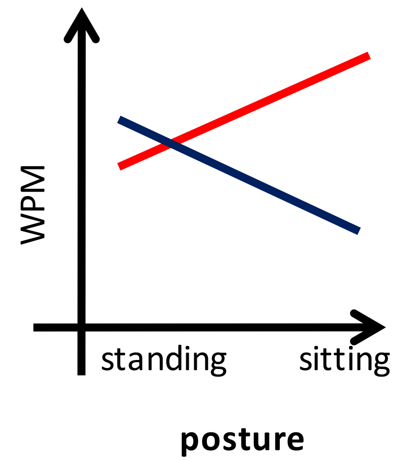
Interactions



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*.



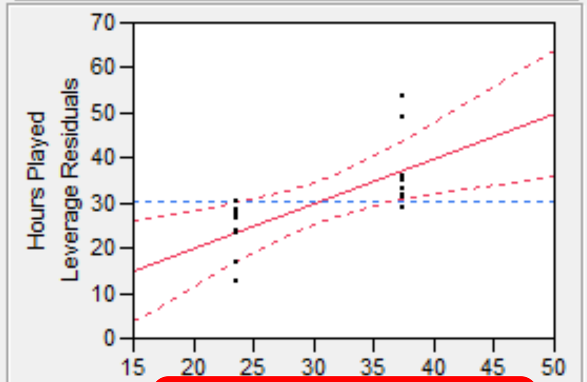
Main effect of *keyboard type*.
No main effect of *posture*.
Interaction between
keyboard type and *posture*.

University of Washington

Interactions

Gender*Posture

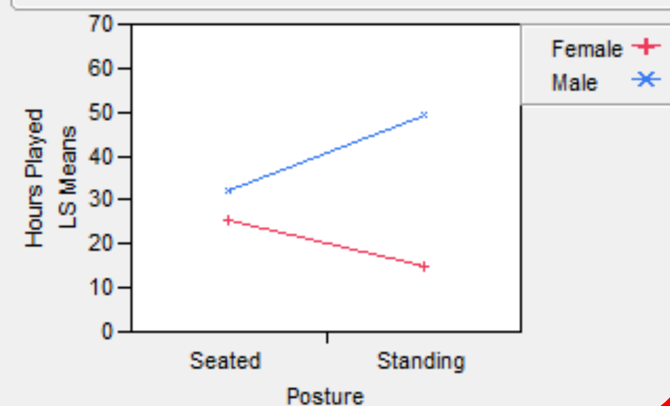
Leverage Plot



Least Squares Means Table

Level	Least Sq Mean	Std Error
Female,Seated	25.250000	4.1994543
Female,Standing	15.000000	4.1994543
Male,Seated	32.250000	4.1994543
Male,Standing	49.500000	4.1994543

LS Means Plot



University of Washington

Reporting Interactions

“However, there was a significant interaction of Gender with Posture (F(1,12)=10.72, p<.01).”

“An examination of our data reveals that females played less while standing, but males played more.”

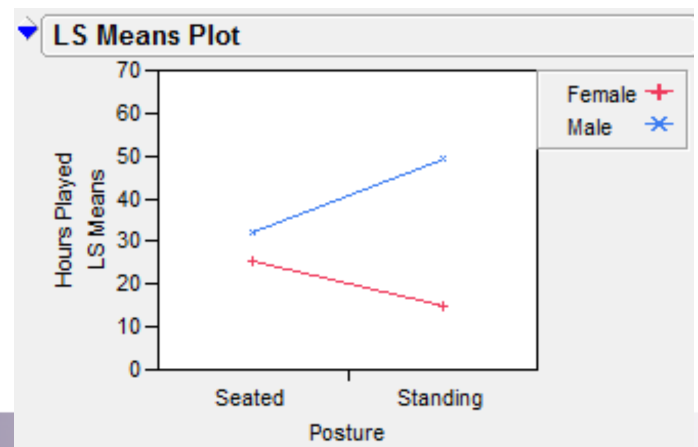
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	2527.5000	842.500	11.9433
Error	12	846.5000	70.542	Prob > F
C. Total	15	3374.0000		0.0006*

Parameter Estimates

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*



Scaling Regressions

Recall an F-test is based in linear regression

$v = ax + by + z$

$a = ?$ $b = ?$ $z = ?$

Can scale to more than two dimensions

$v = aw + bx + cy + dz + e$

$a = ?$ $b = ?$ $c = ?$ $d = ?$ $e = ?$

Concern for Fishing

It is bad form to simply test things until you find something significant, then to report that

Need a theoretical basis for
why you choose to make comparisons

Otherwise, you have gone fishing for results

Concern for Fishing

Recall the definition of p

Unprincipled comparisons
increase the risk of falsely identifying a result

Because if you test enough things,
something is bound to be significant

Unplanned Comparisons

If a multi-level factor is significant,
you need a principled approach
to comparing values of different levels

Tukey's Honestly Significant Difference (HSD)
is available in most statistical software

The sequential Bonferroni procedure
is quite easy to execute manually

Talk to somebody
who has used them

Non-Normal Data

If your data is not normally distributed:

Nominal (categorical) dependent variable:

Consider Chi Square Test

Otherwise:

Consider Non-Parametric Tests

Other Types of Regression

Logistic Regression

binary or ordered outcome

Poisson Regression

count data

Negative Bionomial Regression

"over-dispersed" count data (high stdev)

generalized Poisson

Zero-Inflated Regression

count data with excess zeros

Why are these more common than before?

Talk to somebody who has used them

Chi Square

Used for measuring differences
in proportions between two or more groups

- Number of participants prefer a given interface (out of multiple choices)
- Relative accuracy of binary predictions (perhaps between multiple statistical models or perhaps comparing human judgment, also see ROC curves)

Notation: $\chi 2(1, N=30)=3.28, p<.05$

- Degrees of freedom; report N

Non-Parametric Tests

Non-parametric tests do not assume data comes from normal or quasi-normal distributions

- Cannot use ANOVA (no t or F tests)
- Useful example: Likert scale data

A rank transformation makes data normal

- Wilcoxon signed-rank for matched pairs
- Wilcoxon rank-sum
- Mann-Whitney test
- Aligned Rank test

Talk to somebody who has used them

Bayesian Statistics

Statistics expressed in terms of *degrees of belief*

Start with "prior" beliefs, use data (e.g. an experiment) to create "posterior" beliefs

Report a probability distribution rather than a *p* value and an effect size/confidence interval

Useful for knowledge accrual/meta-analyses

Talk to somebody who has used them

CSE 510: Advanced Topics in HCI

Experimental Design
and Statistical Analysis

James Fogarty
Daniel Epstein

Tuesday / Thursday
10:30 to 12:00

CSE 403