Data

Research Methods for Human-Centered Computing

Data

Today's goal:

Explain how to collect and analyze data (the DVs of your study)

Outline:

– Measuring data
– Uses of data
– Exploring data (if there is time)

Methods outline

Give an overview of your research methods

Methods outline

- Write down your methods using the Guzdial Chart approach
  - https://computinged.wordpress.com/2016/10/03/defining-a-proposal-in-one-table-how-to-write-a-guzdial-chart/
  - Normally, you do this for multiple studies
  - This is a table with a row for each study, and multiple columns describing the methods per study
- For this proposal, you will do one “row” of a Guzdial Chart
  - You don’t have to put it in a table/row for me

Methods outline

What's the research question?

What data are you going to collect?

- What is your target population? How are you going to recruit study participants?
- What measures are you going to use (e.g., survey, interaction logs, GPS location)? How will you collect them?

Methods outline

How much data are you going to collect?

- How many participants?
- How often are you going to use these measures with these participants (within/between/repeated)?

How are you going to analyze these data?

- What are the manipulations and conditions?
- Which outcomes are you going to compare them on?
- Do you expect any interactions between manipulations?
- Are there any important covariates?

Methods outline

What do you expect to find?

- What are your hypotheses for what is going to happen?
- Note: Your research question likely has multiple hypotheses!

Submit as a PDF before class on 9/25

Measuring data

What types of data are there, and how can we collect them?

Measuring data

Levels of measurement

Categorical vs. continuous

Categorical: Nominal - you can do counts

Examples: gender, country of origin, religion

Categorical: Ordinal - you can order them

Examples: 5-point scale items (e.g.: strongly disagree, disagree, neutral, agree, strongly agree)

Note: the differences between levels are not equal!

Measuring data

Continuous: Interval - you can do addition, averaging

- Examples: temperature, most test scores
- Note: there is no meaningful zero point; hence you cannot say “A is twice as large as B”

Continuous: Ratio - you can multiply

- Examples: distance, time, dollar value, number of clicks

Note: sometimes continuous variables are measured on a discrete scale

Measuring data

Are the following data nominal, ordinal, interval, or ratio?

- Favorite snack
- Age in years
- Birth year
- Satisfaction on a 7-point scale
- Yearly income
- Yearly income, measured in categories: below $40k, $40-60k, $60-80k, $80-100k, above $100k

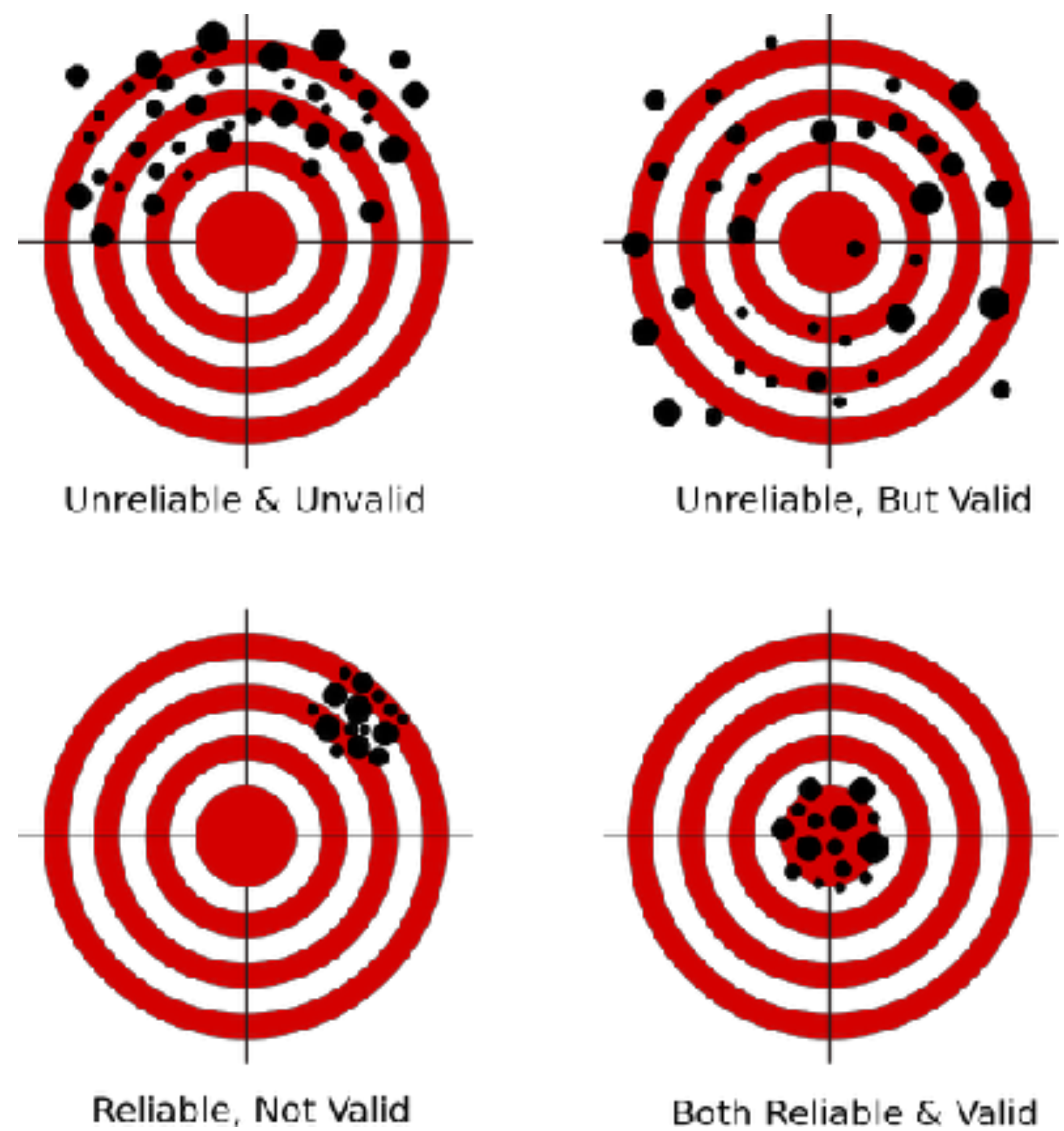
Validity and error

Validity: does it measure what you intend to measure?

- Is "number of clips clicked" a valid measure of satisfaction?

Error (opposite of reliability)

- Will you get the same value on repeated measurement?



Error

Note: Error = random variation due to…

- Environment
- Participants
- Measurements

What has more **environment** error? A study conducted on:

- participants' home computers
- participants' smart phones
- a lab computer

Error

What has more **participant** error?

- a study conducted on the general population
- a study conducted on businesspeople
- a study conducted on CEOs of tech companies

What has more **measurement** error?

- direct observation (e.g. height)
- indirect observation (e.g. pH test strip)
- self-report (e.g. number of vacations in past 3 years)

Validity

Content validity (face validity)

Criterion validity

- Predictive validity
- Concurrent validity

Construct validity

- Discriminant validity
- Convergent validity

Validity

Content validity is assessed by specialists in the concept to be measured

Do the items cover the breath of the content area? (not too wide, not too narrow?)

Are they in an appropriate format?

Bad:

- A attitude scale that also has behavioral items
- A usability scale that only asks about learnability
- A relative measure of risk, trying to measure absolute risk

Validity

Predictive validity

Test how well a measure predicts a future outcome (e.g. behavioral intention —> future behavior)

Concurrent validity

Compare the measure with some other measure that is known to correlate with the concept (e.g. correlate a new scale for altruism with an existing scale for compassion)

Or, compare the measure between groups that are known to differ on the concept (e.g. compare altruism of nuns and homicidal maniacs)

Validity

Discriminant validity

Are two scales really measuring different things? (e.g. attitude and satisfaction may be too highly correlated)

Convergent validity

Is the scale really measuring a single thing? (e.g. a usability scale may actually consist of several sub-scales: learnability, effectiveness, efficiency, satisfaction, etc.)

These can be confirmed statistically

Other types you have to confirm yourself!

Validity

Convergent and discriminant validity are precursors for reliability

Reliability and validity are difficult to establish with a single metric, but with multiple metrics it is possible

- This is the idea behind multi-item measurement: each item uses the others as a yardstick
- Unreliable/invalid items don't correlate well with other items

Validity in context

Note: validity is always assessed in **context**! It depends on:

- the specific **population** to be measured
- the **purpose** of the measure

E.g.: Questions with football metaphors work in the US but not abroad

Likewise, you cannot use a children's IQ test to measure adult intelligence

Tip: Use validated measurement instruments

But always re-validate them for your specific study

Uses of data

A brief intro to how data is used in statistical models

Uses of data

Describe the data

Model the data

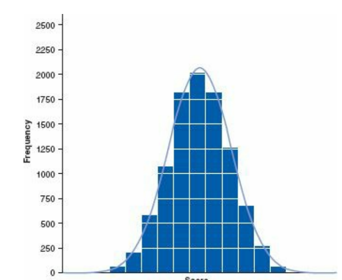
Describing data

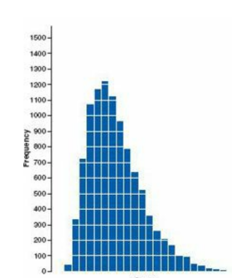
Frequency distribution

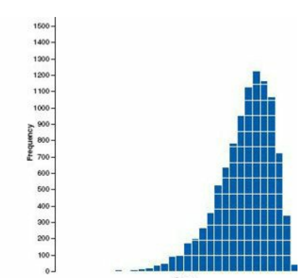
Plot a graph of how many times each score occurs

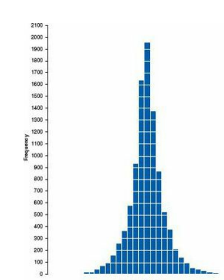
Distributions:

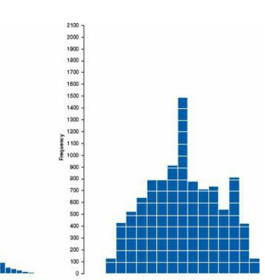
- Normal
- Positive skew
- Negative skew
- Leptokurtic (+ kurtosis)
- Platykurtic (- kurtosis)











Why normal?

Statisticians like normal distributions

- Because they have been studied extensively

We know the probability of a certain event occurring

- e.g. what is the probability that a man is 7ft tall (or taller)?

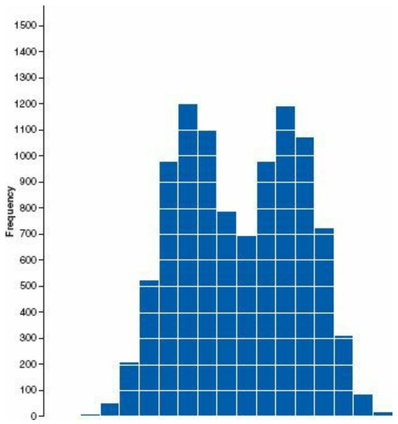
Using the mean and standard deviation, we can turn this question into a Z score:

- $z = (82 - 70) / 4 = 3$, which has a probability of .0013 (0.13%)

Describing data

Center of the distribution

- Mode (most common)
- Median (middle value)
- Mean (average)



Dispersion

- Range
- Interquartile range (IQR)
- Variance and standard deviation

Modeling data

A model is a way to explain or summarize the data

- The mean is a model

The quality of the model depends on how well it fits the data

- We can measure the deviance between the model and the data

Modeling data

$error_i = x_i - mean$

$SS = \sum error_i^2$

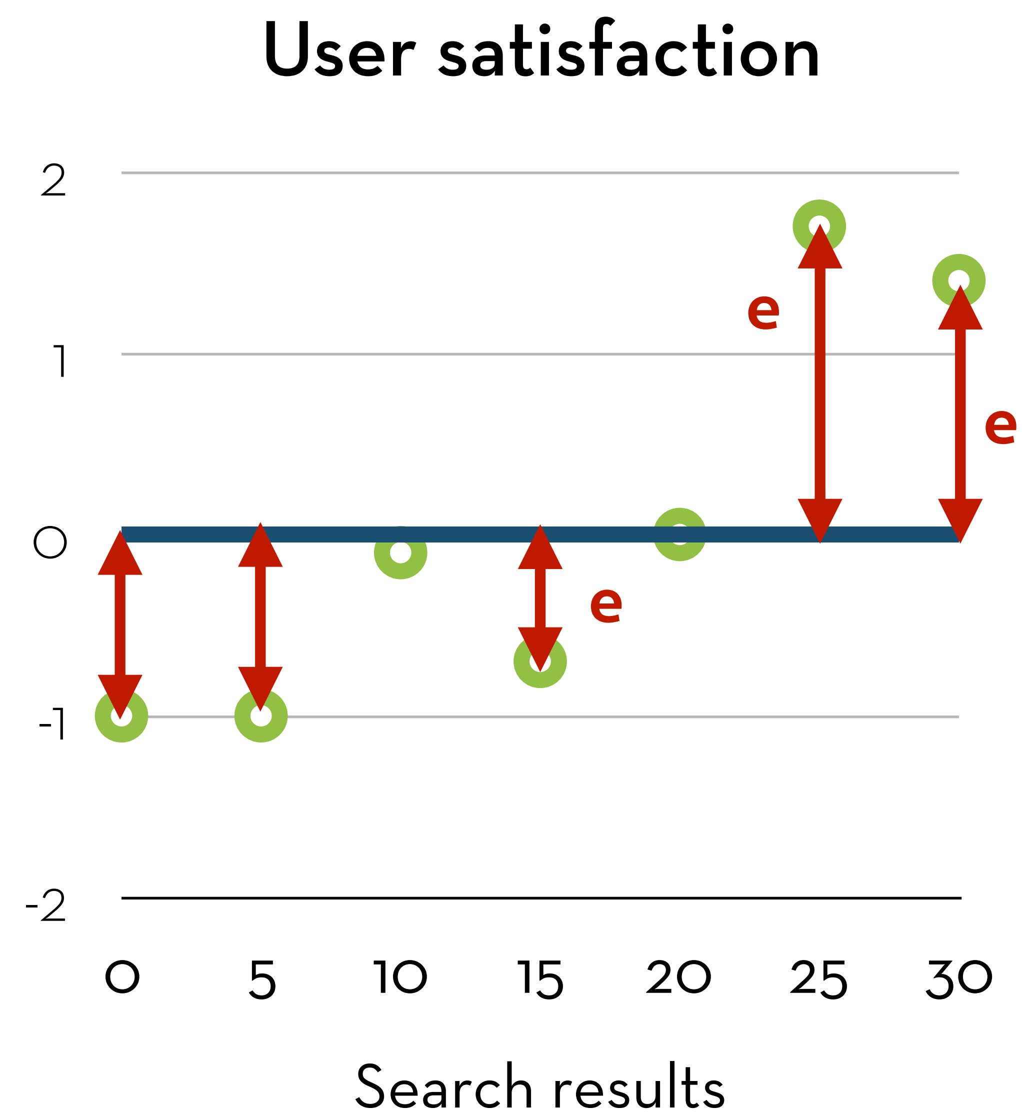
SS = sum of squared errors

$s^2 = SS/(N-1)$

s^2 = variance

s = standard deviation

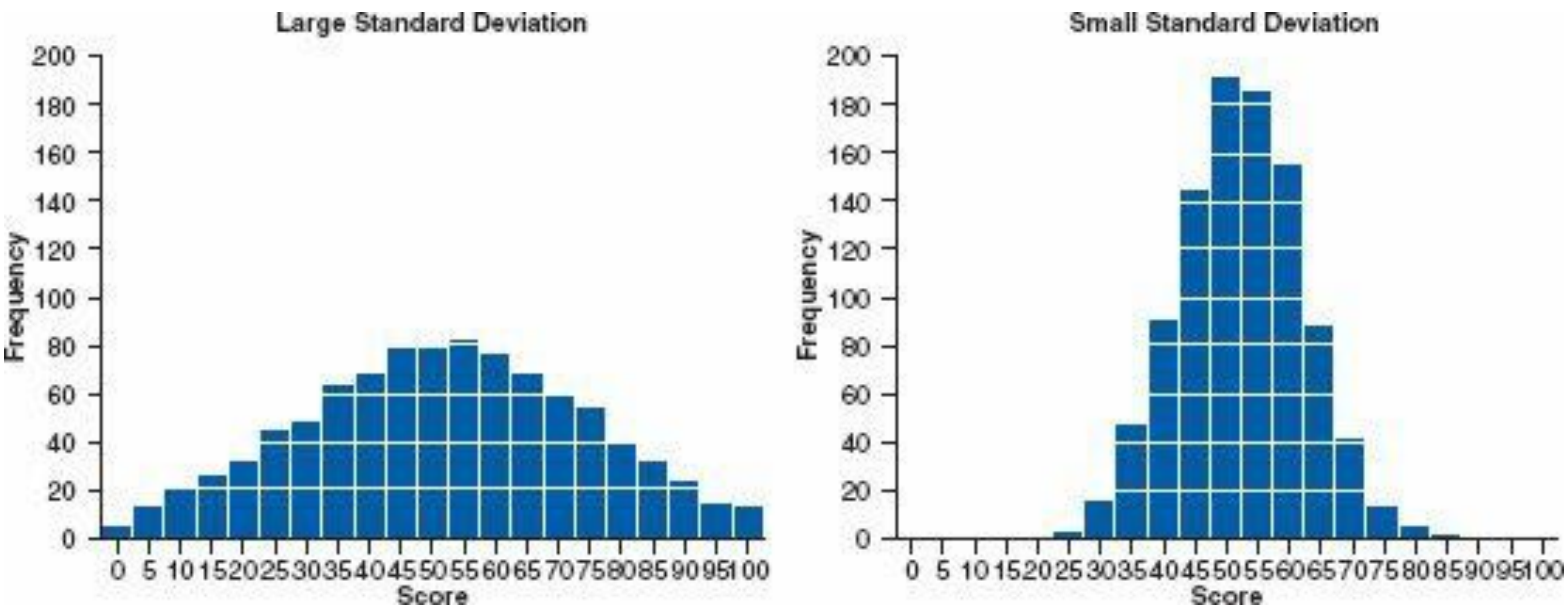
N-1 = degrees of freedom



Effect of deviation

High deviation = more spread

Not the same as kurtosis!



Beyond the sample

(Standard) deviation tells us how well the mean represents the **sample**

But how well does the sample mean represent the **population** mean?

Answer: we can calculate the **standard error**

Beyond the sample

What is the standard error?

Standard deviation = variability of a sample

e.g. variability of age or height of people in this class

Standard error = variability of the mean of a sample

e.g. if I taught this class several times, how much would the average age and the average height differ between classes?

Standard error = standard deviation / √(sample size)

Hypotheses

Compared to text, comic-based policies increase privacy knowledge

- Experimental hypothesis: H1: Mcomic > Mtext
- Null hypothesis: H0: Mcomic = Mtext

To test H1, we try to **reject H0**

How? By comparing the difference in means to the **standard error**

- If the SE is small, we expect small differences under H0
- If the SE is large, large differences are more likely

Hypotheses

If the difference is larger than expected based on the SE:

- We may still have found a difference by chance (no real effect), or...
- There is a real difference in means (H0 is incorrect).

The larger the difference, the more confident we are that H0 is incorrect. Then, H1 is **supported**

But never **proven**, because the first option may still apply!

We calculate the chance; this is the **p-value**

Generally, if $p < 0.05$, we reject H0

Hypotheses

What if $p > 0.05$?

- Does that mean that there is no difference between the two means?

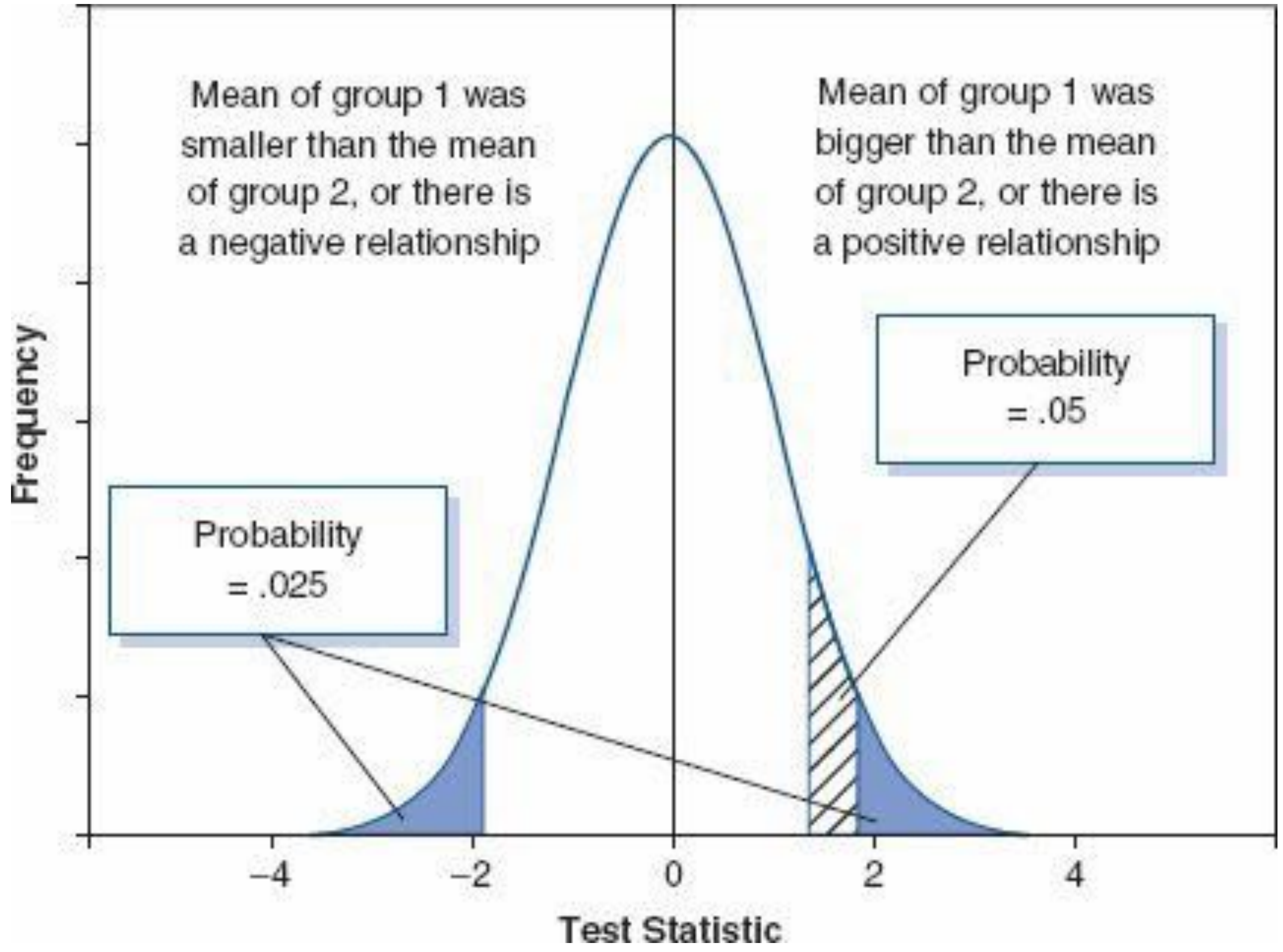
Remember: absence of evidence $\neq$ evidence of absence!

Hypotheses

Where does the 0.05 go?

If H1: Mb > Ma —> one-tailed

If H1: Mb ≠ Ma —> two-tailed



Getting it wrong

But what about the 5% of the times that we reject the null hypothesis, but we got it wrong?

- This is a Type I error
- 5% is the alpha-level

And what about the cases where there is a real effect but we didn't find it?

- This is a Type II error
- We want this error to be smaller than 20%... the beta-level

Getting it wrong

	There is a real effect	There is no real effect
Found an effect	**Power**	alpha (false positive)
Found no effect	beta (false negative)	1–alpha (true negative)

Power analysis

Power depends on alpha, sample size, and the effect size

If your sample is small, even substantially-sized effects may be non-significant (p > alpha)!

- If this happens for > 20% of effects of a substantial size, then the test is under-powered!
- More on this on next week!

Exploring data

Graphs! Graphs! Graphs!

Exploring data

Demonstration of ggplot2 in R

ggplot: a plot object

myGraph <- ggplot(myData); creates a plot

geom: a layer on the plot

myGraph + geom_histogram(); adds a histogram layer

aes: aesthetics of the graph or a layer

myGraph <- ggplot(myData, aes(xvar, yvar, color = cvar)); specifies the variables for the x-axis, y-axis, and color

Exploring data

Other things:

theme()

adds options, such as a title

labels(x = “Text”, y = “Text”)

adds x and y labels

stats:

things that make the geoms magically do what you want (e.g. generates counts when you run geom_histogram)

Exploring data

position

command used to avoid overlap

facet_grid(x ~ y) and facet_wrap(~ y, nrow, ncol)

split your plot into smaller plots

Examples

Scatterplot

Histogram

Boxplot

Bar charts

Line graphs

Scatterplot

Dataset:

Effect of exam stress on exam performance

Variables:

Code: participant id

Revise: hours spent revising

Exam: performance (%)

Anxiety: anxiety level (questionnaire score)

Gender: male/female

Scatterplot

Download the datasets from the course website

Read the data (easy in RStudio)

- Click on “import dataset” in the top-right panel
- Find the file “Exam Anxiety.dat”, click open
- Change the Name to exam, make sure Heading is set to Yes, click Import

Enable ggplot2 using the checkbox under “packages”

- (tab on bottom-right panel)

Scatterplot

Make a plot object; x = Anxiety, y = Exam:

```
scatter <- ggplot(exam, aes(Anxiety, Exam))
```

Create a dot plot:

```
scatter + geom_point()
```

Add labels:

```
scatter + geom_point() + labs(x = "Exam Anxiety, y =
"Exam Performance %)
```

Scatterplot

Add smoother:

scatter + geom_point() + geom_smooth() + labs(x = “Exam Anxiety, y = “Exam Performance %)

Make the smoother a red straight line, without CI:

scatter + geom_point() + geom_smooth(method=“lm”, color=“red”, se = F) + labs(x = “Exam Anxiety”, y = “Exam Performance %”)

Scatterplot

Grouped scatterplot:

groupscatter <- ggplot(exam, aes(Anxiety, Exam, color = Gender)

groupscatter + geom_point() + geom_smooth(method = "lm", aes(fill = Gender), alpha = 0.1) + labs(x = "Exam Anxiety", y = "Exam Performance %", color = "Gender")

Histogram

Read the data

- File: DownloadFestival.dat, set Name to festival
- Dataset: festival-goer hygiene (repeated measures)

Variables:

- ticknumb: participant id
- gender: male/female
- day1, day2, day3: hygiene level at days 1-3 (0-4 scale)

Histogram

Make a plot object with day3 data:

```
histo <- ggplot(festival, aes(day3))
```

Create a histogram:

```
histo + geom_histogram(binwidth = 0.4) + labs(x = 
"Hygiene at day 3", y = "Frequency")
```

Is this a normal distribution?

Boxplot

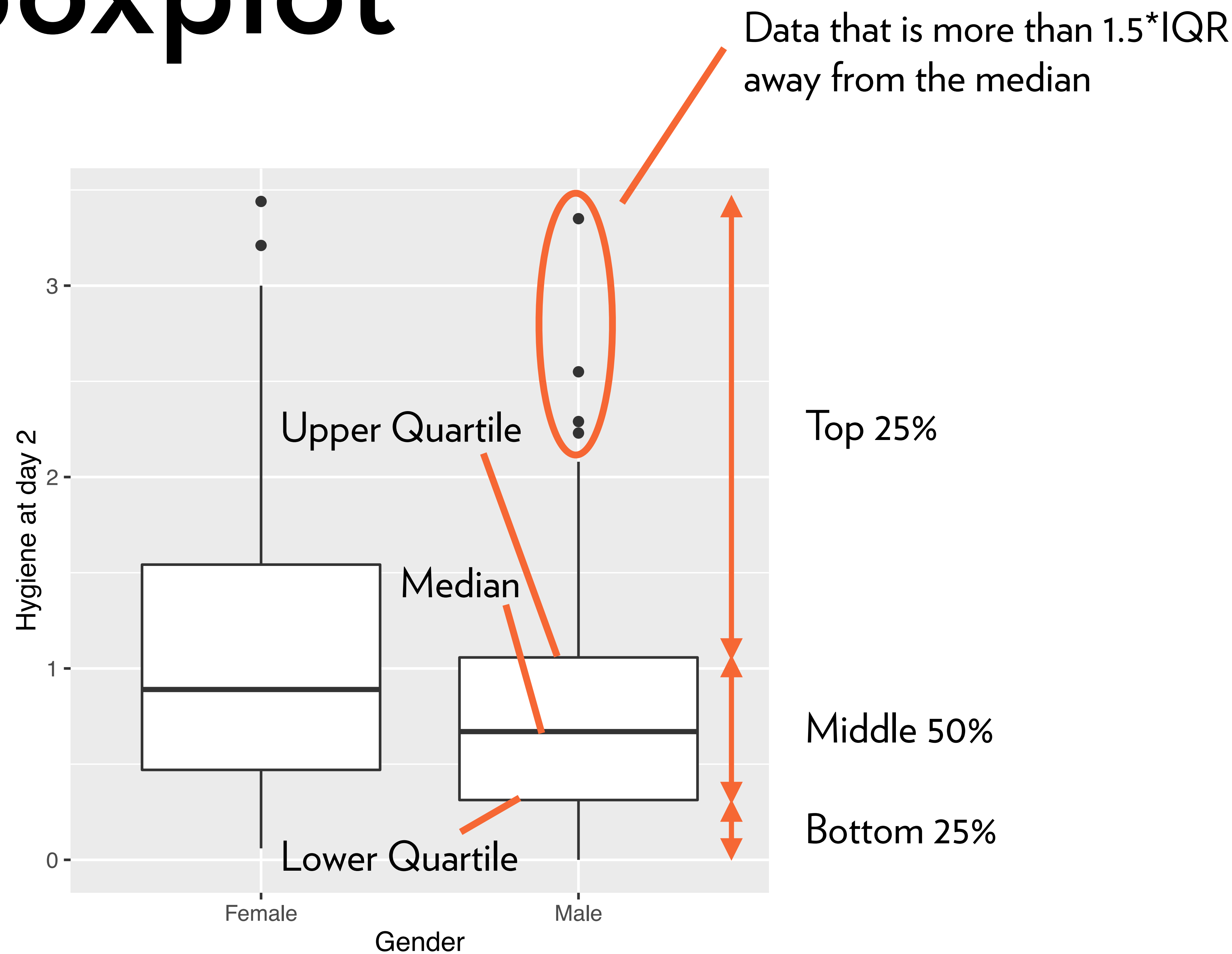
Make a plot object, x = gender, y = day2:

box <- ggplot(festival, aes(gender,day2))

Create a boxplot:

box + geom_boxplot()+ labs(x = "Gender", y = "Hygiene at day 2")

Boxplot



Data that is more than 1.5*IQR away from the median
Upper Quartile
Median
Lower Quartile
Top 25%
Middle 50%
Bottom 25%
Hygiene at day 2
0
1
2
3
Female
Male
Gender

Bar chart

Read the data

- File: ChickFlick.dat, set Name to flick
- Dataset: enjoyment of movies by gender

Variables:

- gender: male/female
- film: the movie (Bridget Jones' Diary, Memento)
- arousal: physiological arousal score (indicator of enjoyment)

Bar chart

Make a plot object with x = film and y = arousal:

bar <- ggplot(flick, aes(film, arousal))

Create a bar chart:

bar + stat_summary(fun.y = mean, geom = "bar", fill = "white", color = "black)

Add error bars of a 95% confidence interval

bar + stat_summary(fun.y = mean, geom = "bar", fill = "white", color = "black) + stat_summary(fun.data = mean_cl_normal, geom = "pointrange")

Line graph

Read the data

File: text.dat

Dataset: effects of text messaging on grammar (repeated measures)

Variables:

Id: participant ID number

Group: text message or control group

Time: at the start (baseline), six_months

Score: grammar scores

Line graph

Make a plot object:

```
line <- ggplot(text,aes(Time,Score,color=Group))
```

Create dots, lines connecting them, and some error bars

```
line + stat_summary(fun.y = mean, geom = "point") +
stat_summary(fun.y = mean, geom = "line",
aes(group=Group)) + stat_summary(fun.data =
mean_cl_boot, geom="errorbar", width = 0.2)+ylim(0,100)
```

ylim scales the plot area