Dealing with data

Measurement & Evaluation of HCC Systems

Intro

Today's goal:

Teach you how to deal with data (fundamentals of stats)

Outline:

- Measuring data
- Uses of data
- Exploring data

Measuring data

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

Collecting data

Correlational (e.g. survey, observation)

- Measure both cause and effect
- High ecological validity

Experimental

- Manipulate cause, measure effect
- Able to establish causality

Measuring data

Levels of measurement

Categorical

- Nominal - you can do counts
- Ordinal (subjective) - no diffs, < >

Continuous

- Interval - distances equal, adding, averaging
- Ratio - 2 is twice as much as 1, multiply

Validity and error

Validity: does it measure what you intend to measure?

Is "purchase behavior" a valid measure of satisfaction?

Error (opposite of reliability)

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 in context

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

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

Types of validity

Content validity (face validity)

Criterion validity

- Predictive validity
- Concurrent validity

Construct validity

- Discriminant validity
- Convergent validity

Content 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

Criterion 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)

Construct 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.)
- Factor analysis helps you with construct validity
  - Other types you have to confirm yourself!

Uses of data

A brief intro to how data is used in statistical models

Uses of data

Describe the data

Model the data

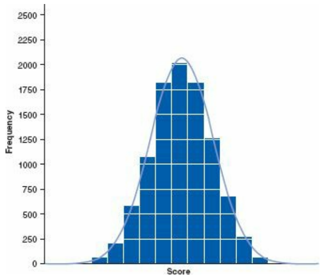
Describing data

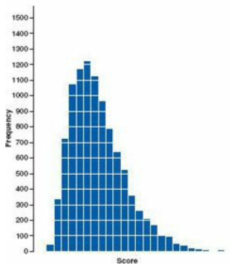
Frequency distribution

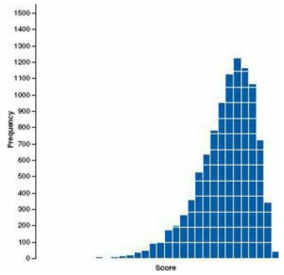
Plot a graph of how many times each score occurs

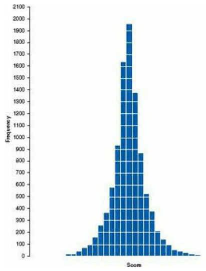
Distributions:

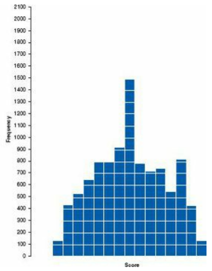
- Normal
- Positive skew
- Negative skew
- Leptokurtic (+ kurtosis)
- Platykurtic (- kurtosis)











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)?

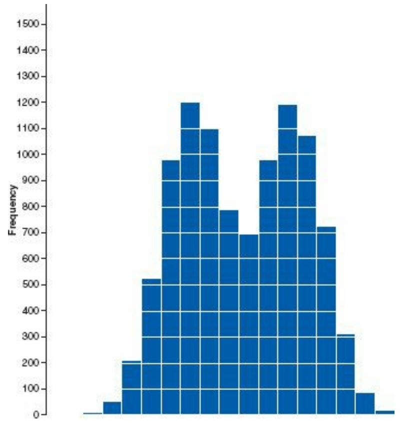
Using the mean and standard deviation, we can turn this question into a Z score:

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

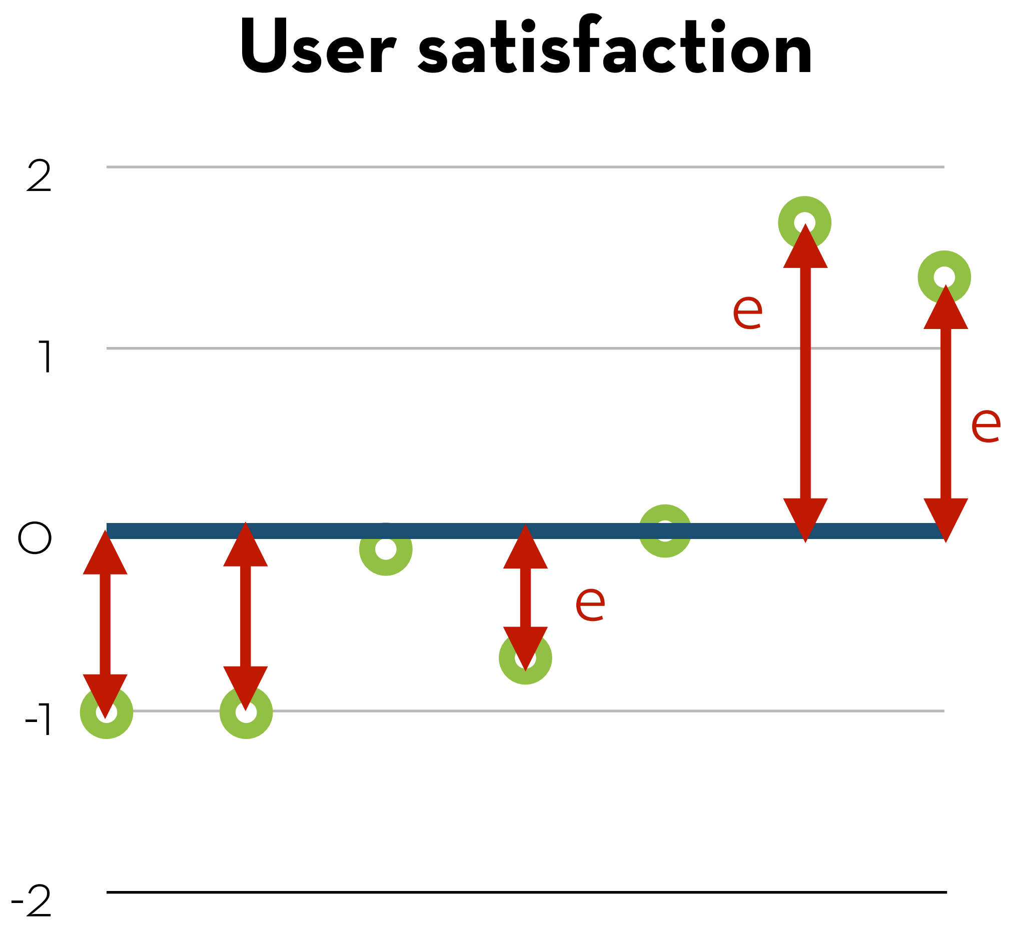
Modeling data

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

$\text{error}_i = x_i - \text{mean}$

$SS = \sum \text{error}_i^2$

SS = sum of squared errors

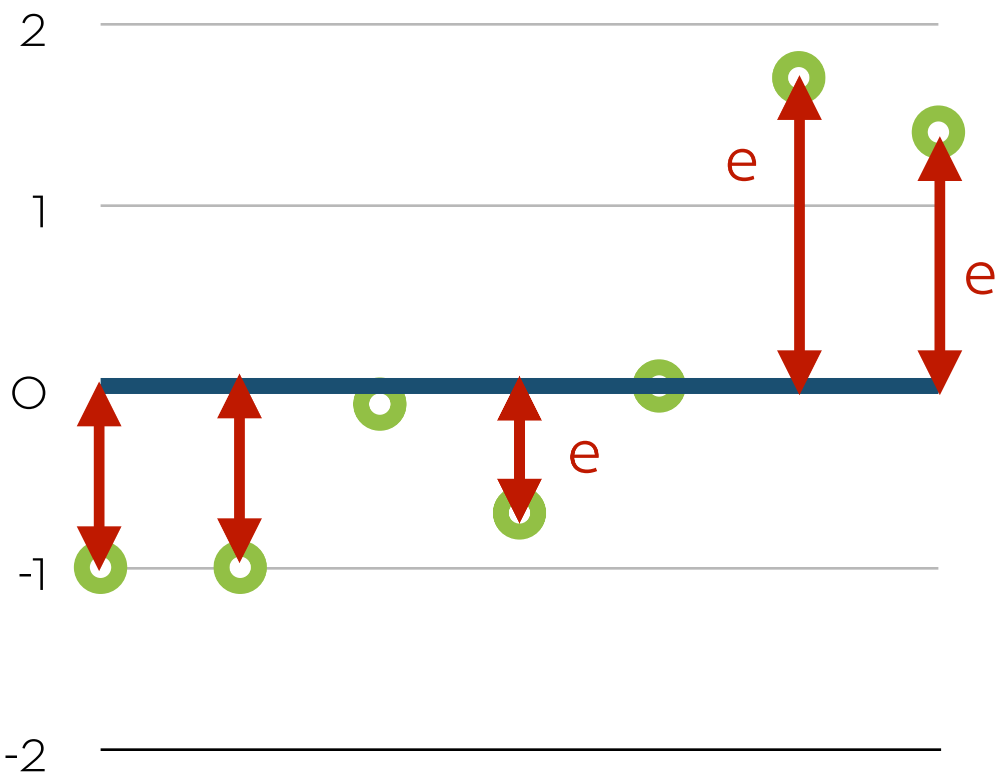
$s^2 = SS/(N-1)$

s^2 = variance

s = standard deviation

N-1 = degrees of freedom

User satisfaction



Why N-1?

Let's say you have 4 data points:

1, 3, 4, 8

Mean: 4

If you know the mean, how many data points are "free"?

Answer: Only three!

Once you know the first three, you will know the fourth one as well, because the mean needs to be 4!

(1+3+4+x)/4 = 4 —> x has to be 8!

Modeling data

Remember:

$\text{error}_i = x_i - \text{mean}$

$SS = \sum \text{error}_i^2$

More generally:

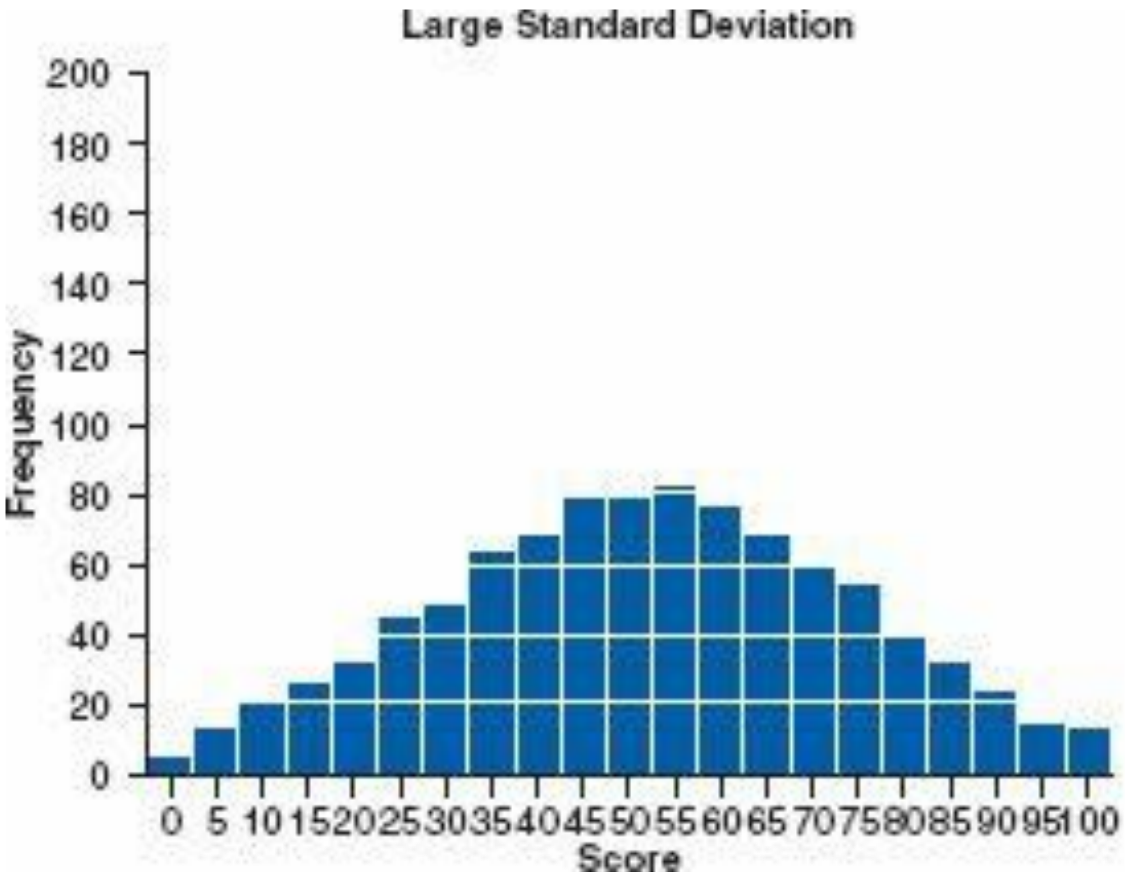
model: $\text{outcome}_i = \text{model} + \text{error}_i$

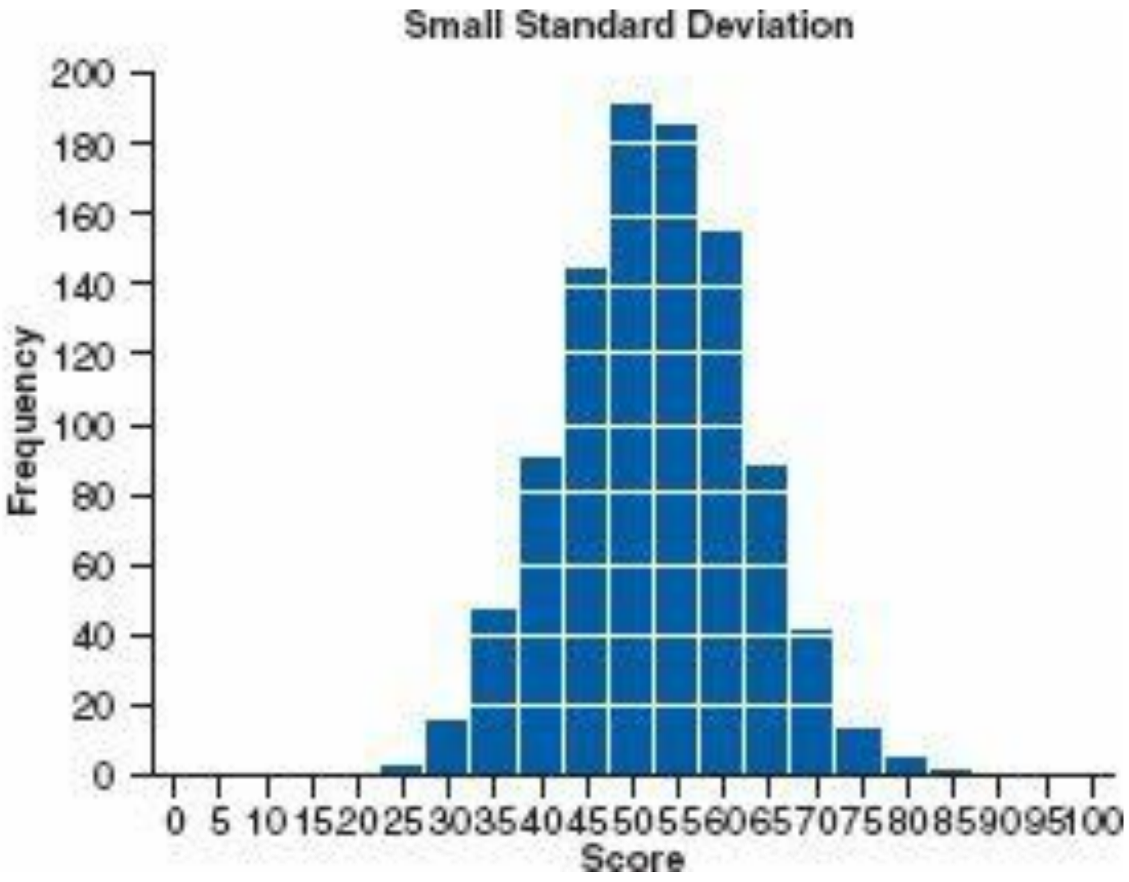
$\text{deviation} = \sum(\text{observation}_i - \text{model})^2$

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: 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)

Beyond the sample

The sample mean may deviate a bit from the population mean

- Can we say something about the population mean?

Answer: we can create a confidence interval:

- E.g. 95% CI: on repetition, we'd expect the true mean to be within the CI 95% of the time

Beyond the sample

Calculating the CI, using the z-score:

$z = (x - \text{mean})/s$

95% of the means fall within $z = -1.96$ and $z = +1.96$

upper x: mean + 1.96*SE

lower x: mean – 1.96*SE

General rule: to construct a x*100% confidence interval, use:

z-score of $p = (1-x)/2$ (you can look this up in a table)

for small samples: use t_{n-1} of $p = (1-x)/2$

Hypotheses

Research question:

- Is my new system (version B) better than version A?
- Experimental hypothesis: H1: Mb > Ma

Calculate the means. Do they differ a lot?

- Given no effect, we expect the means to be roughly equal
- H0: Mb = Ma

To test H1, we try to reject H0

Hypotheses

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 it 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!

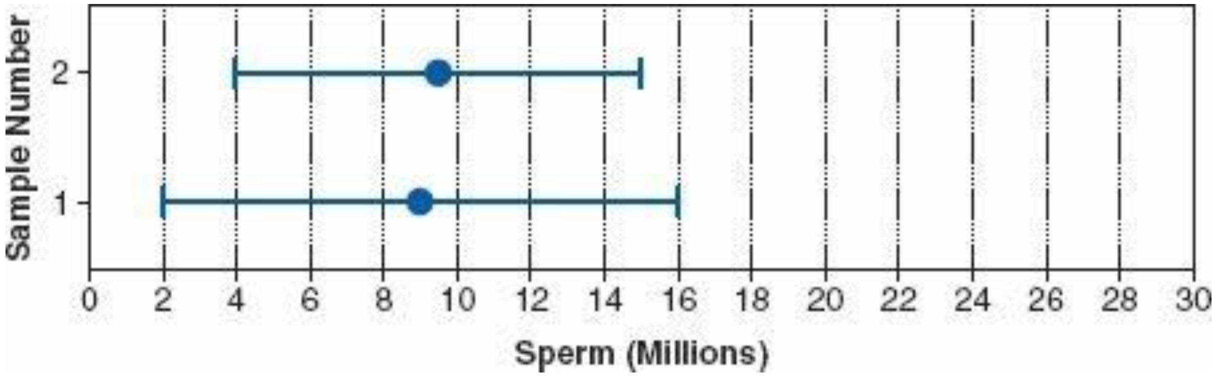
We calculate the chance; this is the **p-value**

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

Hypotheses

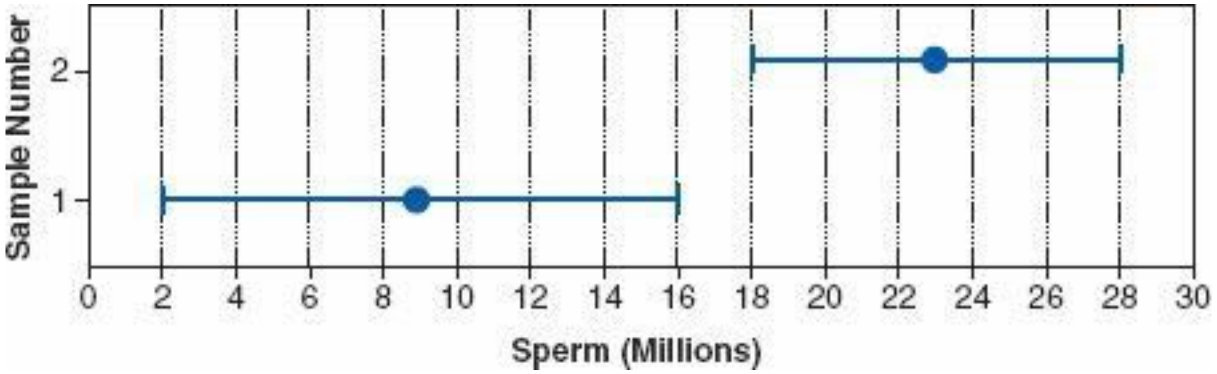
If CIs overlap, SE is large compared to the difference

> Means are likely to come from the same population



If they don't overlap, they are likely to come from different populations

> Because with 95% CIs, this happens only 5% of the time!



Hypotheses

Generally speaking:

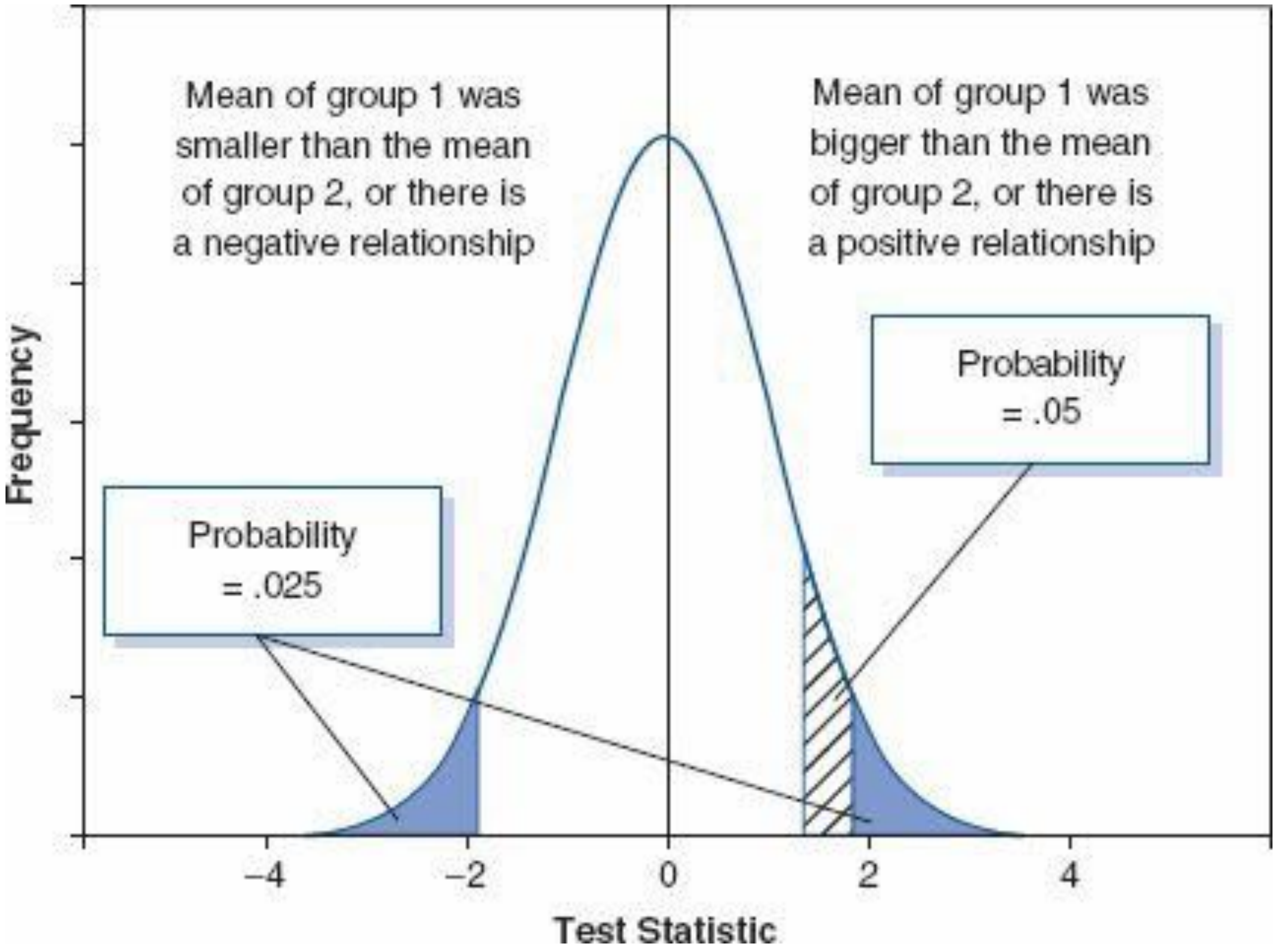
- Test statistic = the variance explained by the model / the variance not explained by the model
- For a good test statistic, we know the probability of finding a value at least this big
- The bigger the value, the smaller the chance
- If this $p < 0.05$, then we reject the null hypothesis that the test statistic = 0
- What if $p > 0.05$?

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

A calculation involving the following 4 parameters:

- Alpha (cut-off p-value, often .05)
- Power (probability of finding a true effect, often .80 or .85)
- N (sample size, usually the thing we are trying to calculate)
- Effect size (usually the expected effect size)

If N is small, true effects may be non-significant (p > alpha)!

If this happens for > 20% of effects of the expected size, then the test is under-powered!

More on this in the next lecture!

Exploring data

Graphs! Graphs! Graphs!

Exploring data

We will do this part in R

Exploring data

Plotting with ggplot2

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() note: Field uses the deprecated function “opts()”

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

Density plot

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 examData, 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(examData, 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(examData, 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 festivalData
- 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 day1 data:

```
histo <- ggplot(festivalData, aes(day1))
```

Create a histogram:

```
histo + geom_histogram(binwidth = 0.4) + labs(x =
“Hygiene at day 1”, y = “Frequency”)
```

Density plot

Fix the outlier

```
festivalData[festivalData$day1 == 20.02,]$day1 <- 2.02
```

Make a plot object with day1 data:

```
density <- ggplot(festivalData, aes(day1))
```

Create a density plot:

```
density + geom_density() + labs(x = “Hygiene at day 1”,
y = “Density Estimate”)
```

Boxplot

Make a plot object, x = gender, y = day1:

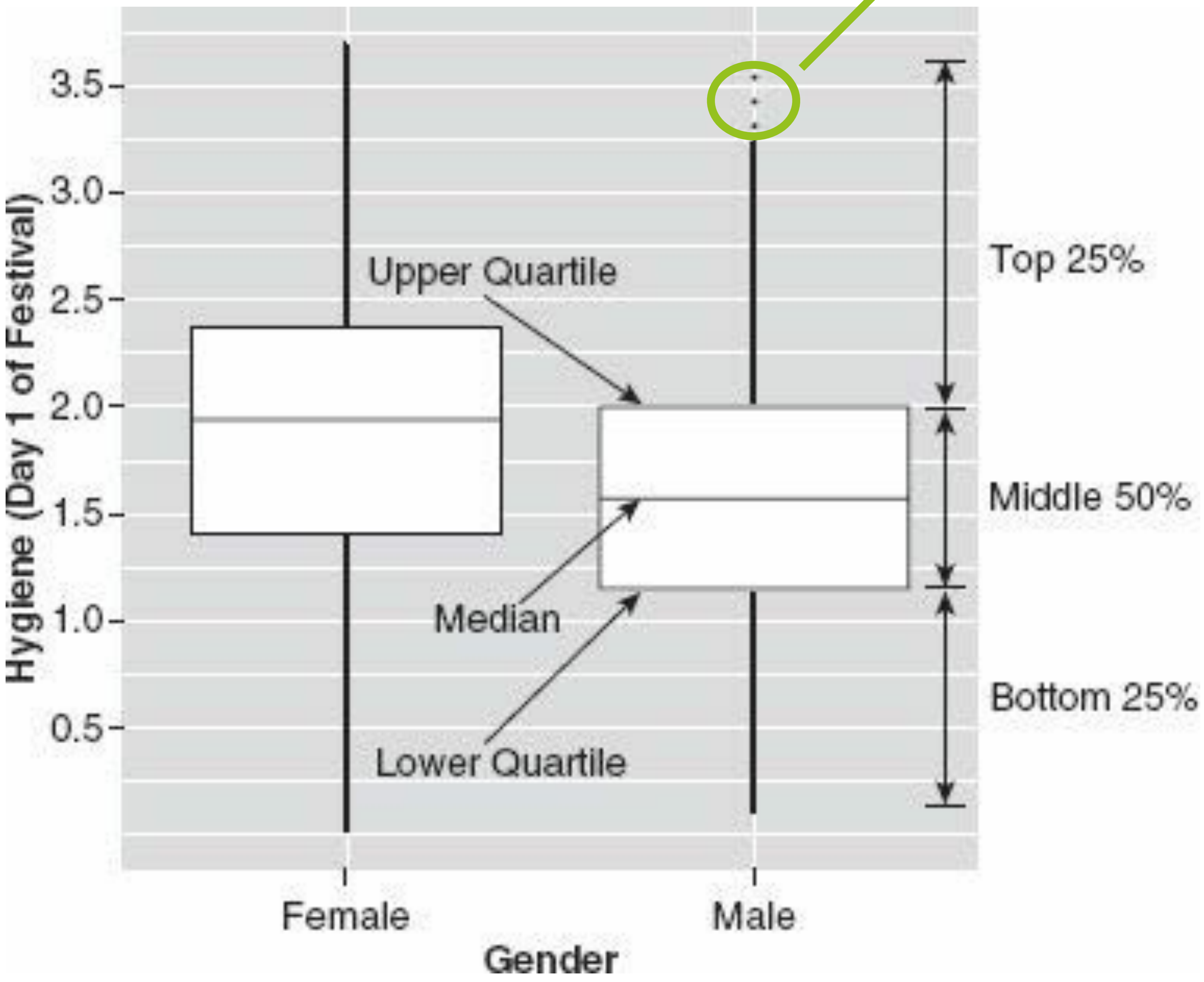
```
box <- ggplot(festivalData, aes(gender,day1))
```

Create a boxplot:

```
box + geom_boxplot()+ labs(x = “Gender”, y = “Hygiene at day 1”)
```

Boxplot

Data that is more than 1.5*IQR away from the median



Bar chart

Read the data

- File: ChickFlick.dat, set Name to chickFlick
- 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(chickFlick, aes(film, arousal))
```

Create a histogram:

```
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”)
```

Bar chart by gender

Make a plot object x = film, y = arousal, fill = gender:

```
genbar <- ggplot(chickFlick, aes(film, arousal, fill=gender))
```

Create a bar plot, genders side-by-side:

```
genbar + stat_summary(fun.y = mean, geom = “bar”,
position=“dodge”)
```

Add error bars of a 95% confidence interval

```
genbar + stat_summary(fun.y = mean, geom = “bar”,
position=“dodge”) + stat_summary(fun.data =
mean_cl_normal, geom = “errorbar”, position =
position_dodge(width=0.90), width = 0.2)
```

Bar chart by gender

Same thing, but now we are going to make separate plots for gender:

```
genbar2 <- ggplot(chickFlick, aes(film, arousal, fill=film))
```

Create a bar plot, genders in different “facets” (note: no dodge needed, remove the legend)

```
genbar2 + stat_summary(fun.y = mean, geom = “bar”) +
stat_summary(fun.data = mean_cl_normal, geom =
“errorbar”, width = 0.2) + facet_wrap(~gender) +
theme(legend.position = “none”)
```

Line graph

Read the data

- File: Hiccups.data, set Name to hiccupsData
- Dataset: cures for hiccups (repeated measures)

Variables:

- Baseline: hiccups at baseline
- Tongue: hiccups after tongue pulling
- Carotid: hiccups after carotid artery massage
- Rectum: you don't want to know

Line graph

Reshape the data

We want to go from:

Baseline	Tongue	Carotid	Rectum
15	9	7	2
13	18	7	4
9	17	5	4
7	15	10	5

to:

Hiccups	Intervention
15	Baseline
13	Baseline
9	Baseline
7	Baseline

Line graph

Reshape the data

```
hiccups <- stack(hiccupsData)
```

Give the correct column names

```
names(hiccups) <- c(“Hiccups”,”Intervention”)
```

Turn “Intervention” into a factor

```
hiccups$Intervention <- factor(hiccups$Intervention,
levels=c(“Baseline”,”Tongue”,”Carotid”,”Rectum”))
```

Line graph

Make a plot object:

```
line <- ggplot(hiccups,aes(Intervention,Hiccups))
```

Create dots, a blue dotted line connecting them, and some error bars

```
line + stat_summary(fun.y = mean, geom = “point”)
+ stat_summary(fun.y = mean, geom = “line”, aes(group=1),
color = “blue”, linetype = “dashed”)
+ stat_summary(fun.data = mean_cl_boot, geom =
“errorbar”, width = 0.2)
```

Double line graph

Read the data

- File: TextMessages.dat, set Name to textData
- Dataset: effects of text messaging on grammar (repeated measures)

Variables:

- Group: text message or control group
- Baseline: grammar scores before the experiment
- Six_months: grammar scores after the experiment

Double line graph

Install “reshape2” package (in the “packages” panel)

Reshape the data with “melt”

```
text <- melt(textData, id=c(“Group”),
measured=c(“Baseline”,Six_months”))
```

Give the correct column names

```
names(text) <- c(“Group”,”Time”,”Score”)
```

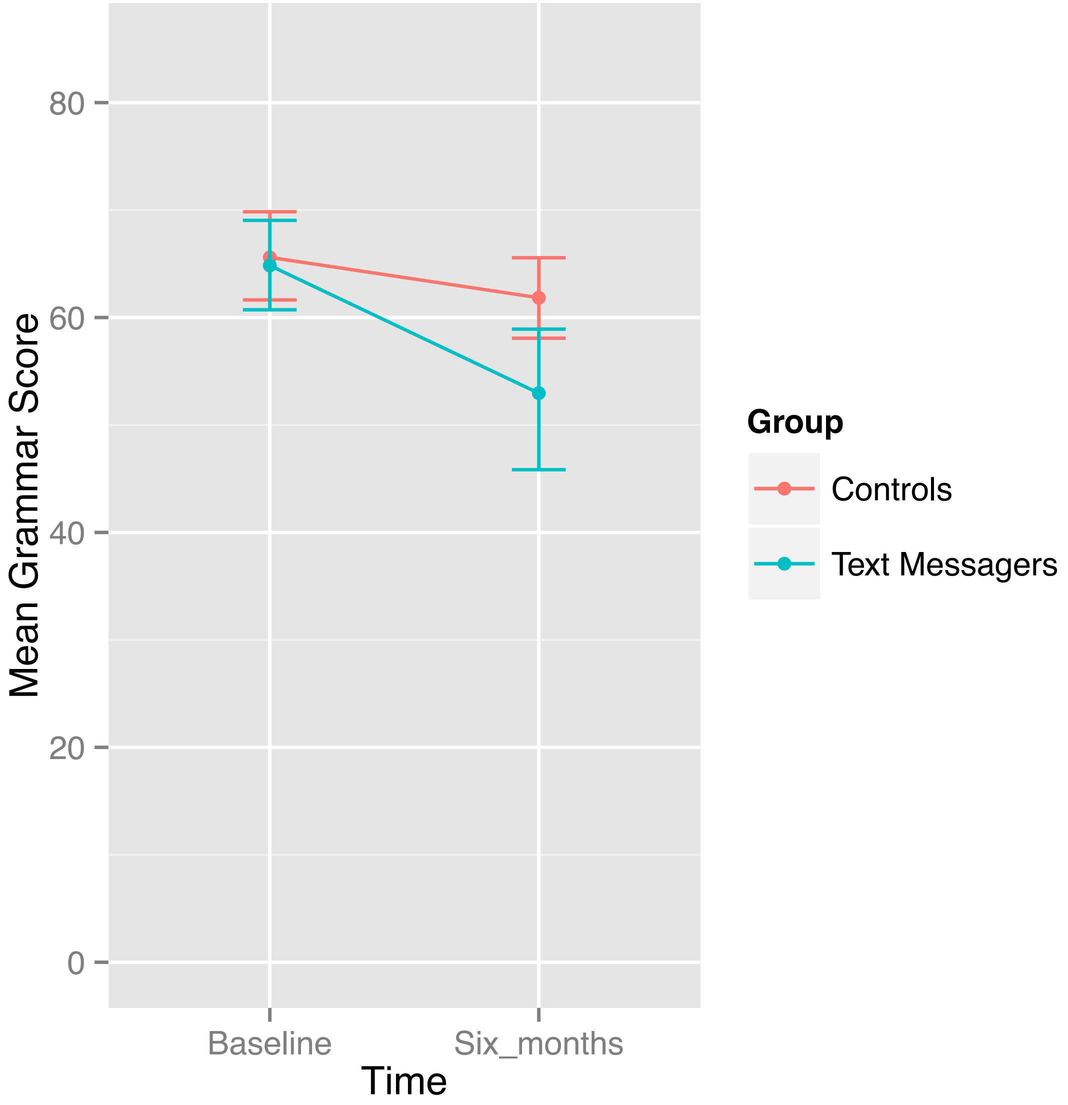
Turn “Time” into a factor

```
text$Time <- factor(text$Time, levels=c(“Baseline”,”
Six_months”))
```

Double line graph

Exercise: create the following plot

(hint: reshape the y-axis with the ggplot aesthetic “ymin” and “ymax”)



“It is the mark of a truly intelligent person to be moved by statistics.”

THANKS!

George Bernard Shaw