People and Data

Honors seminar

People and Data

Today's goal:

How to study people and data

Outline:

- User experiments
- Dealing with data

User experiments

Studying people

User experiments

A scientific method to investigate factors that influence how people interact with systems*

Systems can be anything:

- Software
- Hardware
- Other people
- Organizations
- Policies

What to ask?

“Is my new travel system good?”

Problem...

What does good mean?

- Learnability? (e.g. number of errors?)
- Efficiency? (e.g. time to task completion?)
- Usage satisfaction? (e.g. usability scale?)
- Outcome quality? (e.g. survey?)

We need to define **measures**

Better...

“Does the user interface of my travel system score high on this usability scale?”

However...

What does high mean?

Is 3.6 out of 5 on a 5-point scale “high”?

What are 1 and 5?

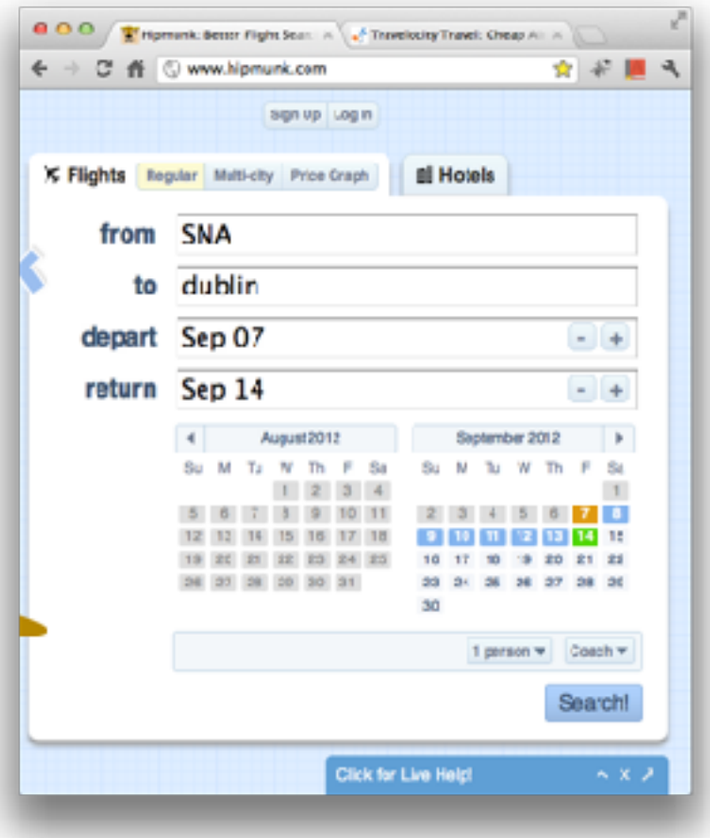
What is the difference between 3.6 and 3.7?

We need to compare the UI against something

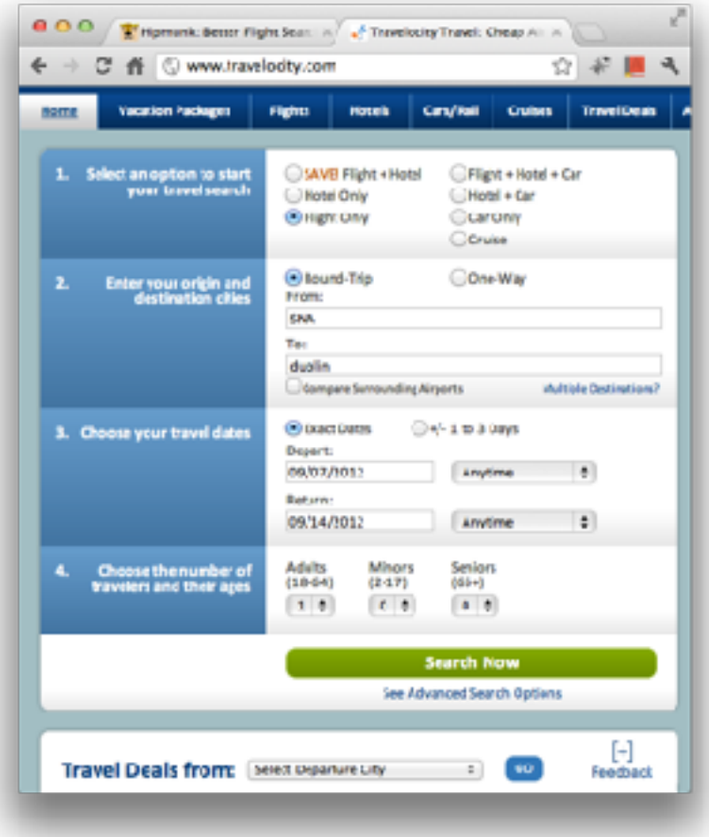
Even better...

“Does the UI of my system score high on this usability scale compared to this other system?”

Testing A vs. B



My new travel system



Travelocity

However...

Say we find that it scores higher on usability... why does it?

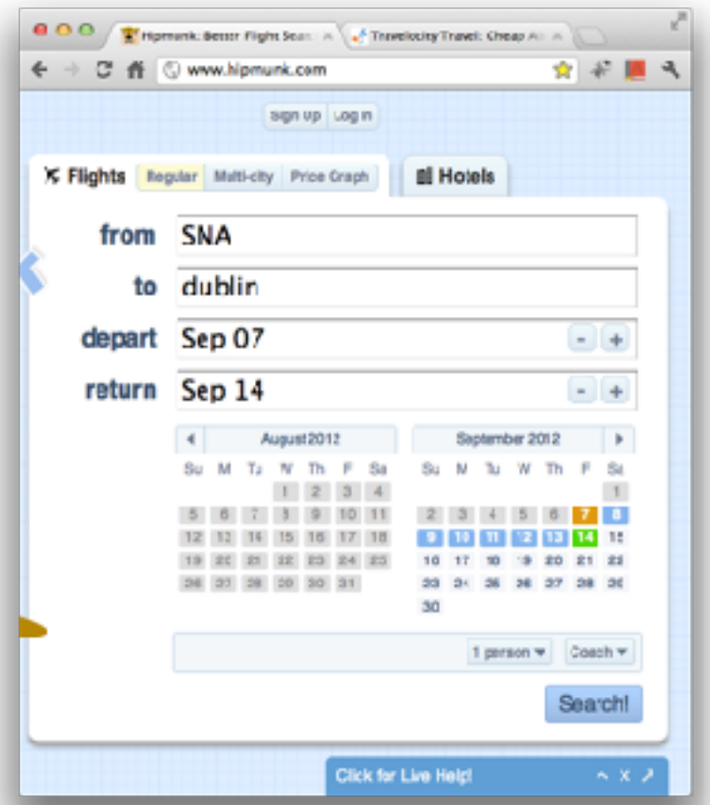
- different date-picker method
- different layout
- different number of options available

Apply the concept of ceteris paribus to get rid of confounding variables

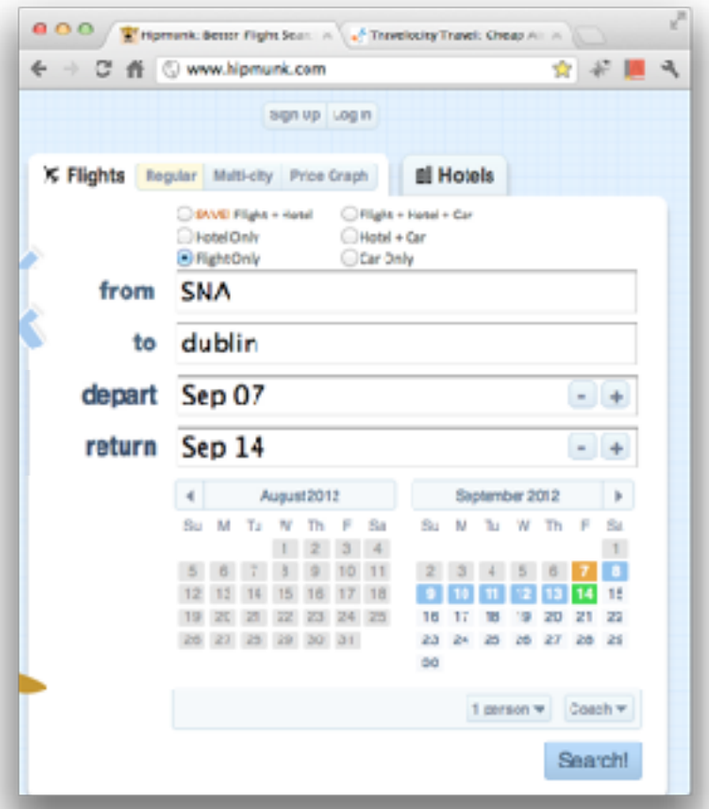
Keep everything the same, except for the thing you want to test (the manipulation)

Any difference can be attributed to the manipulation

Ceteris Paribus



My new travel system



Previous version
(too many options)

How it works

Create multiple versions of your system (intervention and control)

Recruit participants to take part in your study (a sample taken from a population)

You let people use one (or all) of these versions

You measure their behaviors and/or subjective evaluations (outcome)

You statistically evaluate the difference in outcome between intervention and control

Core components

“A user experiment systematically tests how different system aspects (manipulations) influence the users’ experience and behavior (observations).”

Manipulations

Manipulations are the things that you believe will "make a difference"

- Also called "independent variables"

In HCC research, our main manipulations are system aspects

- We are not testing entire systems, but system aspects

Each manipulation consists of multiple conditions (different versions of the system aspect)

- Simplest variant, two conditions: intervention and control

Manipulations

Examples of manipulations:

- Recommendations vs. random items
- Number of recommendations: 5, 10, or 20
- Comic-based privacy policy vs. text-based privacy policy
- Comic length: short, medium, long

Observations

Observations are the means by which you measure the differences between conditions

- Also called “dependent variables” or “outcomes”

In HCC research, our observations are either objective or subjective

- They are **always** quantitative (more on this in the next part)

Observations

Examples of observations:

Objective:

- Number of clicks
- Privacy knowledge (# correct answers on a privacy quiz)

Subjective

- Perceived privacy protection
- Perceived system effectiveness

Systematic

If you apply the concept of ceteris paribus...

...and you randomly assign participants to conditions...

...then any difference you find can be attributed to the manipulation!

i.e., the difference between the conditions!

The power of user experiments lies in the ability to make such causal inferences!

Hypotheses

Experiments can answer causal research questions

i.e., how one variable influences the other

Example: Does a comic-based privacy policy increase privacy awareness compared to a text-based policy?

Hypotheses are predictions regarding the influence of your independent variables (manipulations) on your dependent variables (outcomes)

Example: compared to text, comic-based policies increase privacy knowledge

Hypotheses

Compared to text, comic-based policies increase privacy knowledge

Experimental hypothesis: H1: Mcomic > Mtext

Question: what if they could just as well be worse?

Calculate the means. Do they differ a lot?

Given no effect, we expect the means to be roughly equal

H0: Mcomic = Mtext

To test H1, we try to reject H0

Hypotheses

If the difference is larger than expected:

- 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!

Survey/observation

What is the difference between men and women in Facebook usage satisfaction?

Downsides:

Purely correlational

- No manipulations!
- What causes what?
- Third variable problem

No ceteris paribus

- Hard to get rid of confounding variables

Dealing with data

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

Dealing with 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

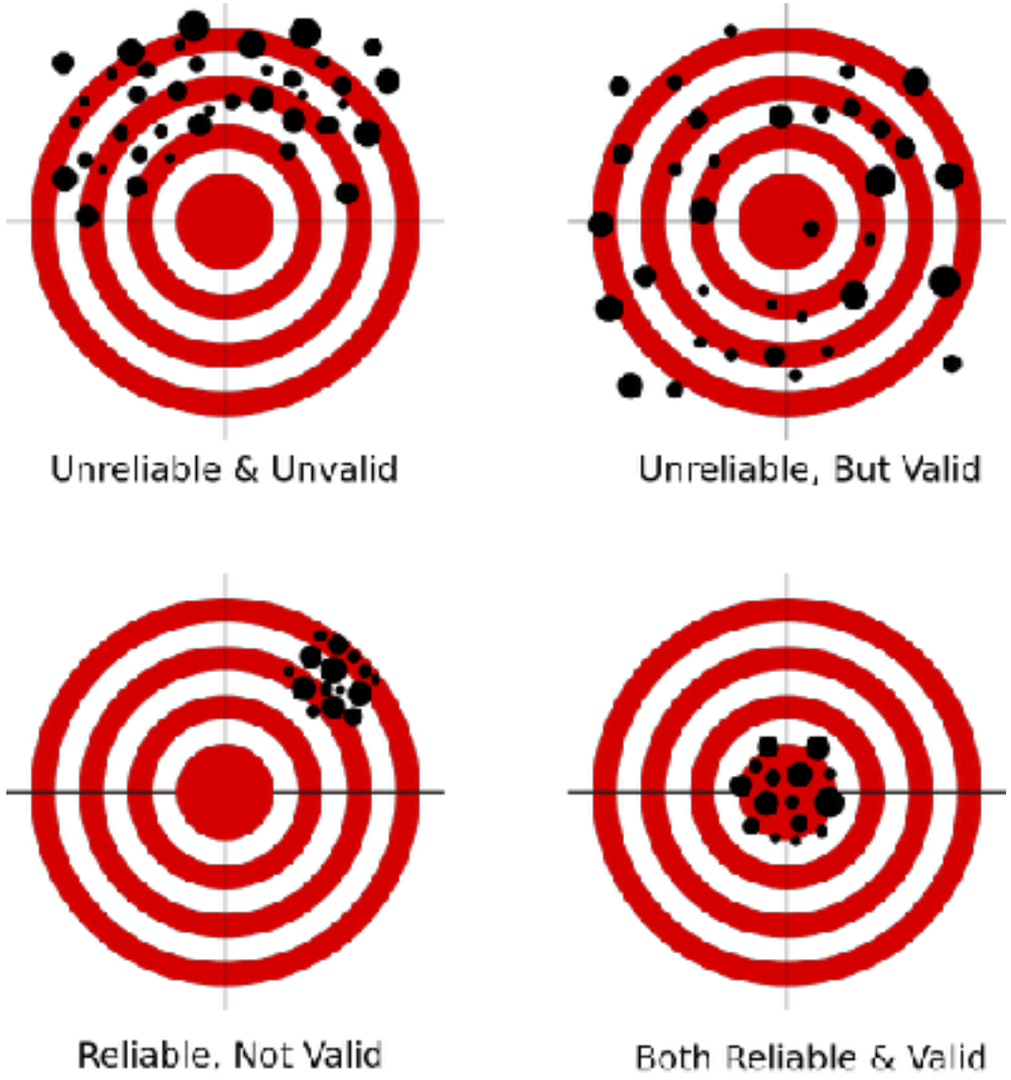
Validity and error

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 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

Describe the data

Model the data

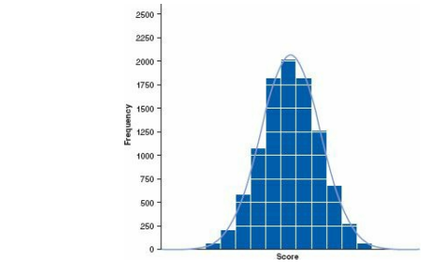
Describing data

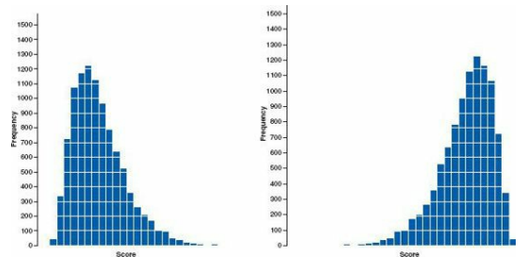
Frequency distribution

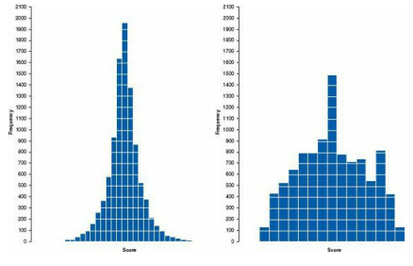
Plot a graph of how many times each score occurs

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)?

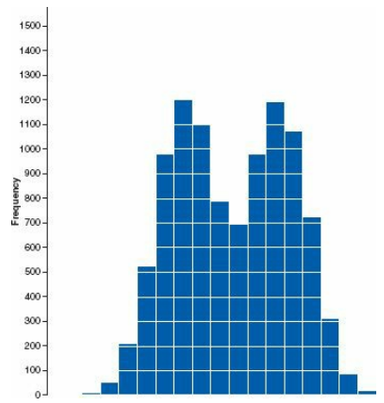
Using the mean and standard deviation, we can turn this question into a Z score:

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

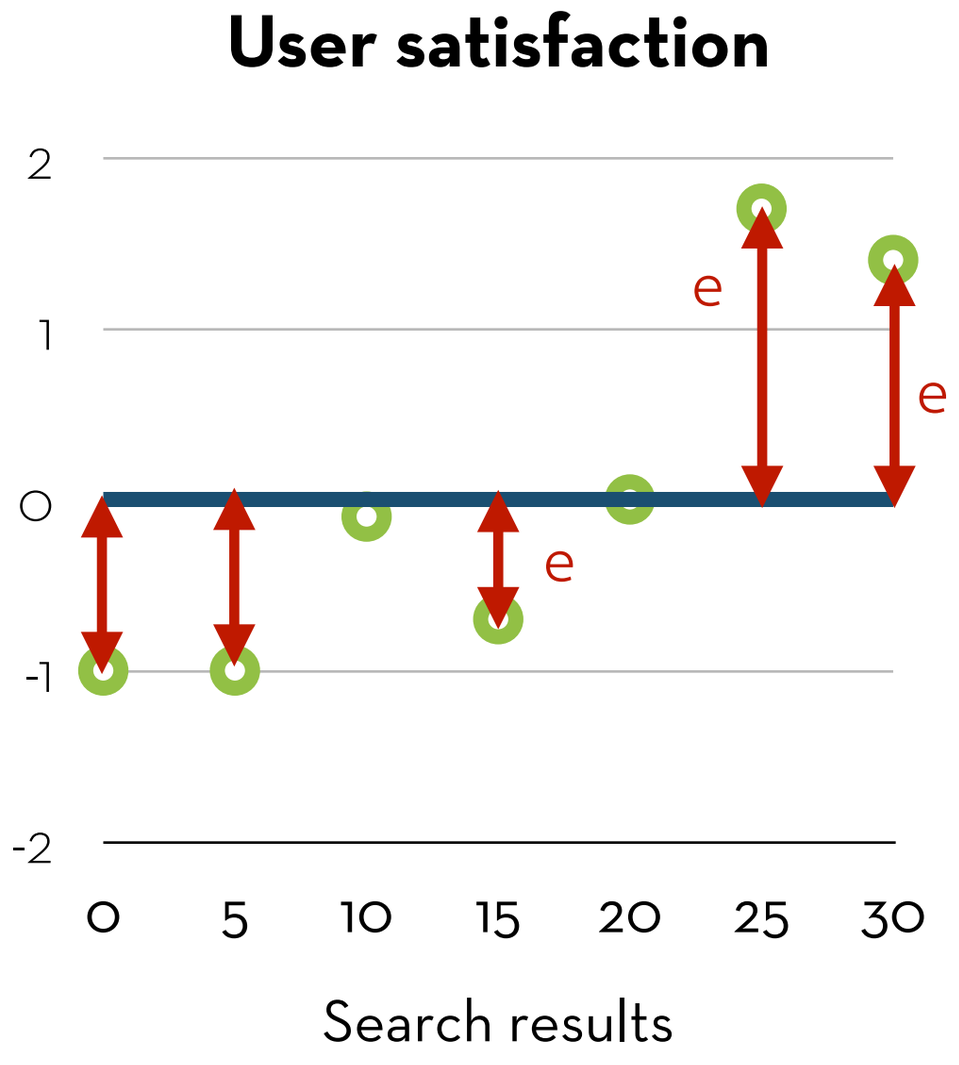
Modeling data

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

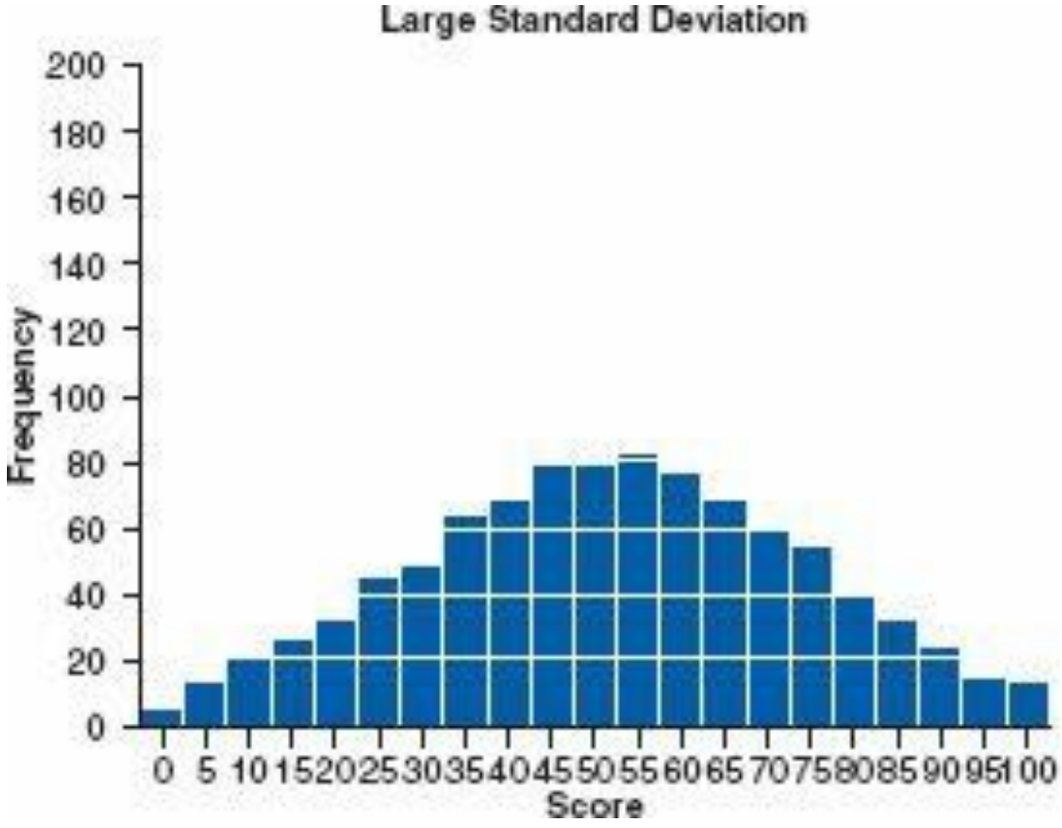
We can measure the deviance between the model and the data

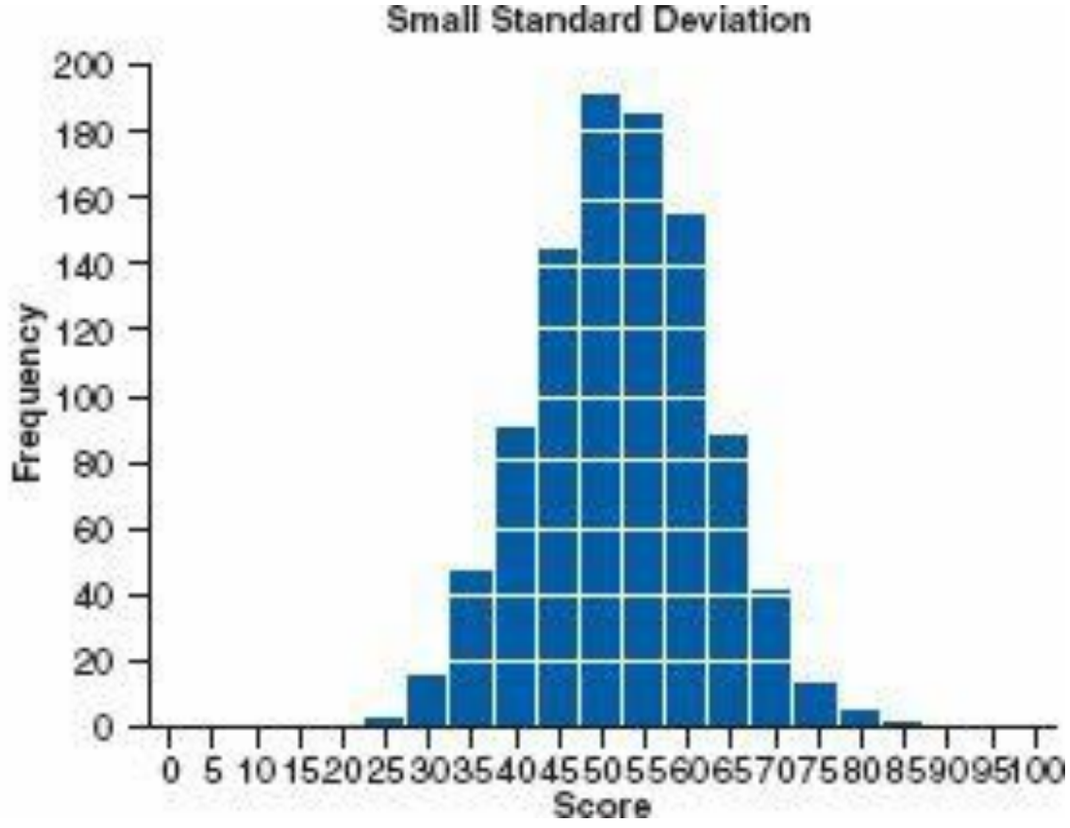


Effect of deviation

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≠ evidence of absence!