Statistical Evaluation

Research Methods for Human-Centered Computing

Statistical Evaluation

Today's goal:

A (very) brief overview of statistical analysis

Outline:

- An overview of some stats
- An example

Statistics

A very brief overview

Statistics

Covariance and correlation

Linear regression

t-tests, ANOVA, factorial ANOVA

Non-normal data

Multi-level data

Subjective data

Mediation analysis

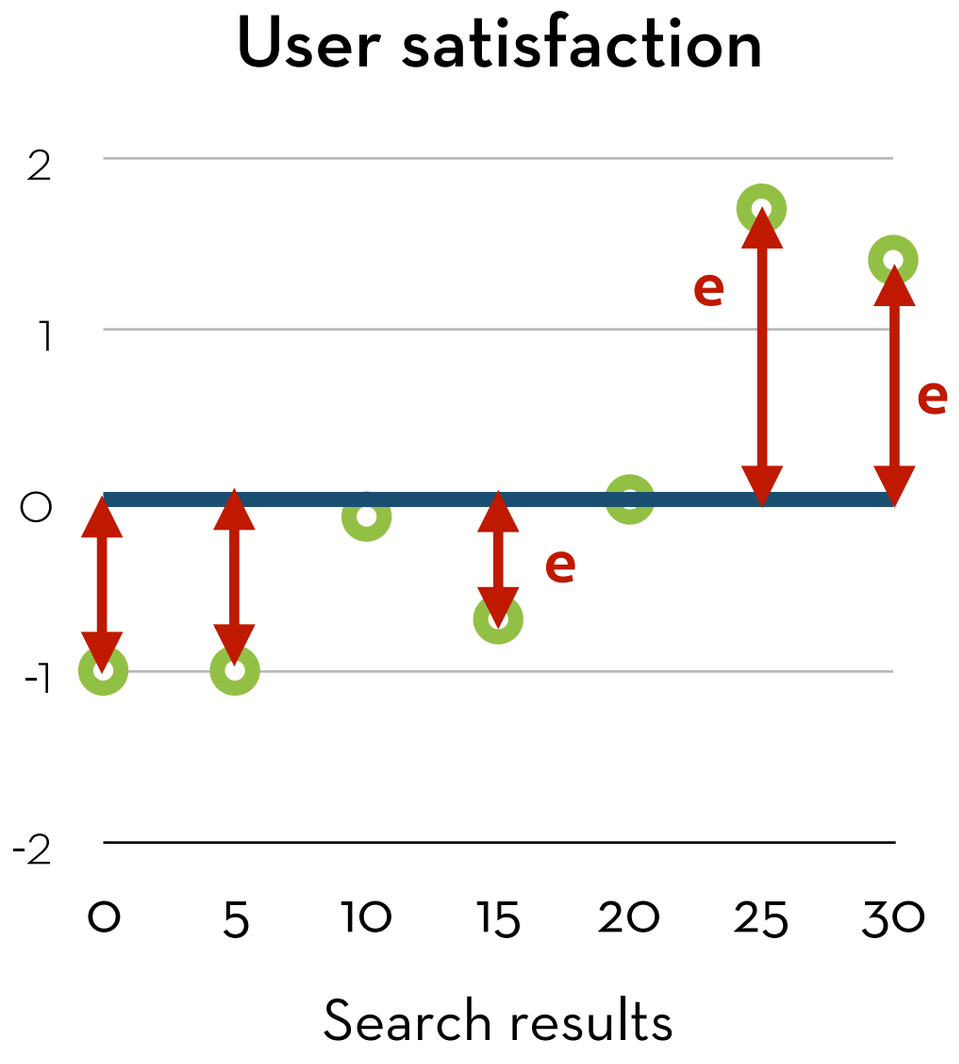
Modeling data

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

$error_i = x_i - mean$

$SS = \sum error_i^2$

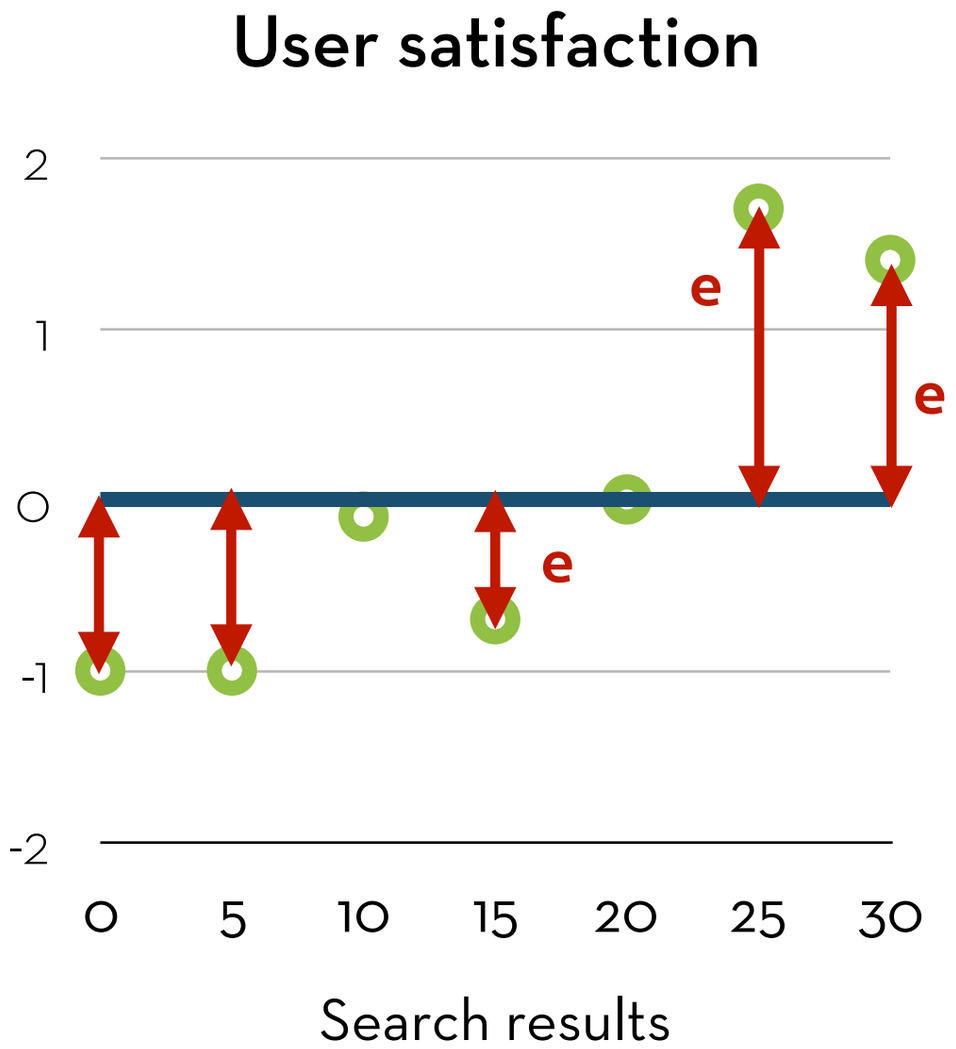
SS = sum of squared errors

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

s^2 = variance

s = standard deviation

N-1 = degrees of freedom



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!

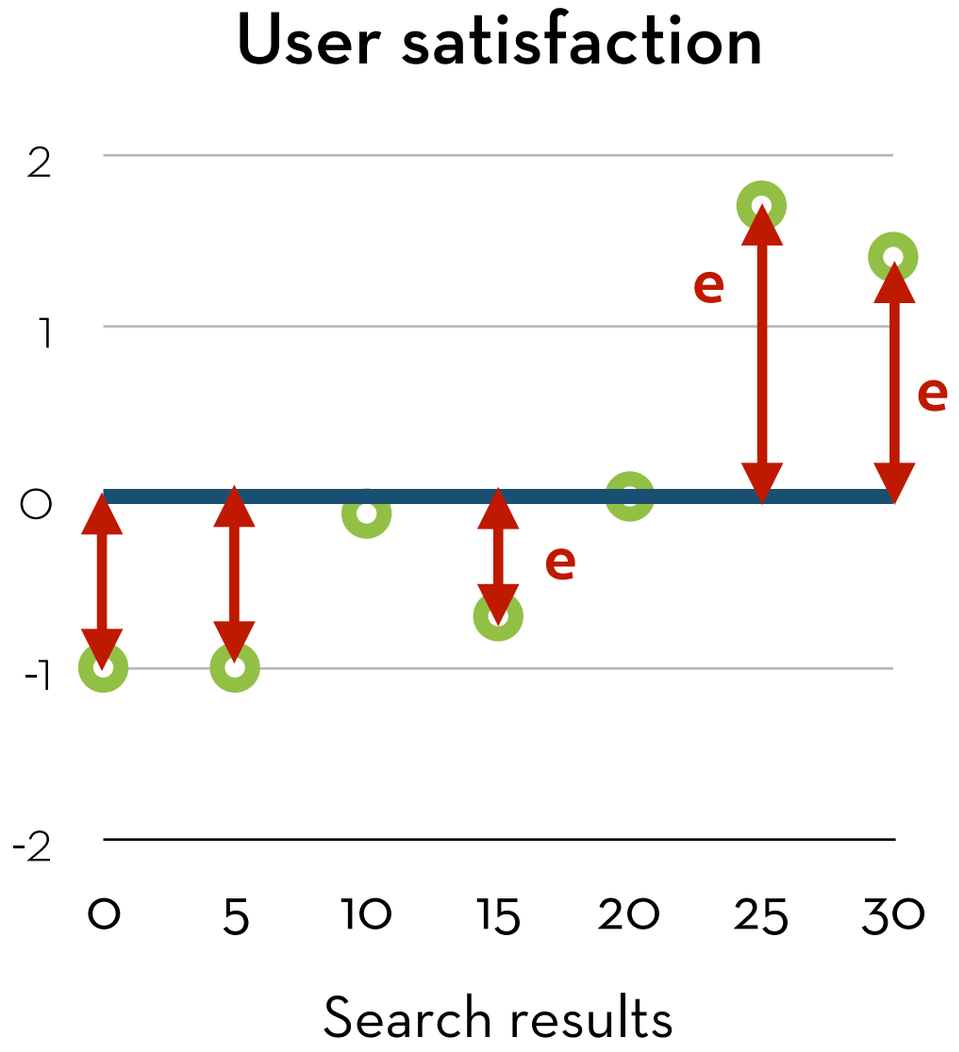
(1+3+4+x)/4 = 4 —> x has to be 8!

Variance

Variance is the variation of the data around a model (e.g. the mean)

$$s^2 = \sum(x_i - mean_x)^2/(N-1)$$

It is the sum of the **error in x times the error in x**, divided by the degrees of freedom



Covariance

Covariance measures the relationship between the variations of two variables, x and y

$$cov(x,y) = \sum(x_i - mean_x)(y_i - mean_y)/(N-1)$$

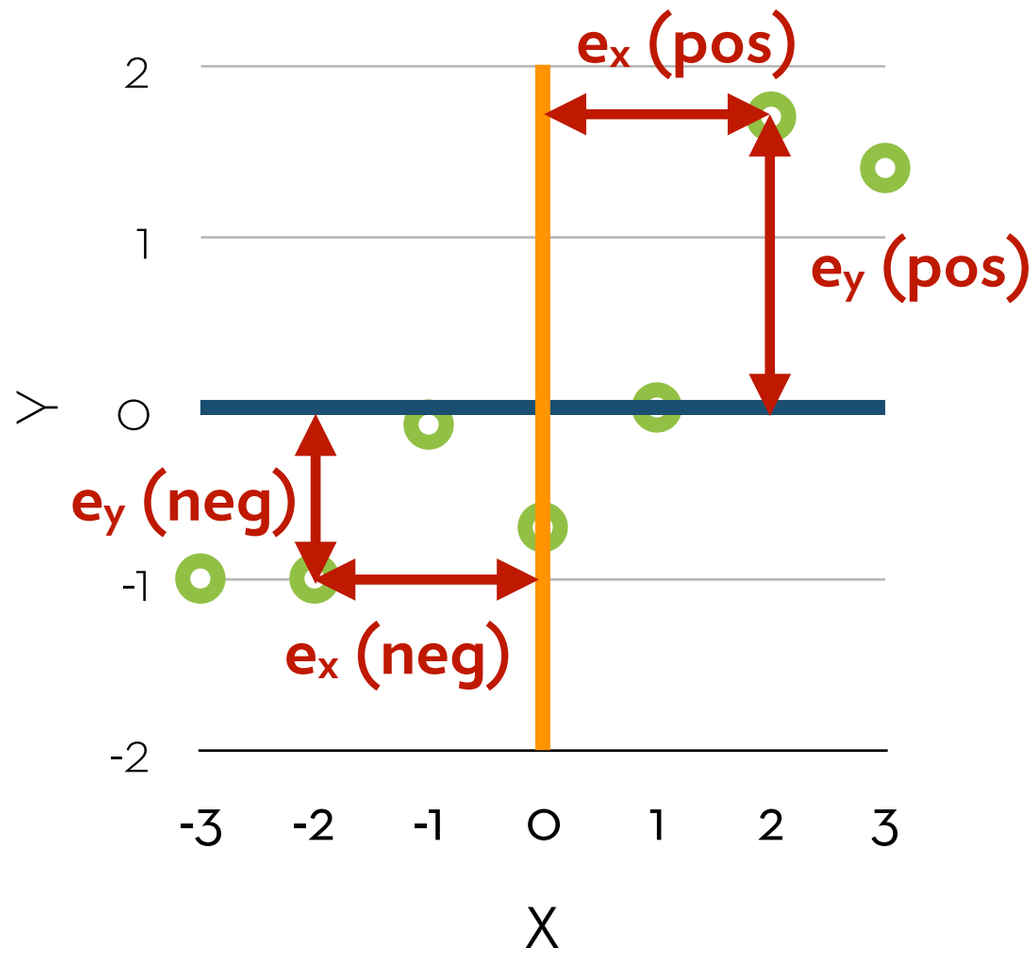
It is the sum of the **error in x times the error in y**, divided by the degrees of freedom

Covariance

Covariance measures the relationship between the variations of two variables, x and y

$$cov(x,y) = \sum(x_i - mean_x) * (y_i - mean_y)/(N-1)$$

It is the sum of the **error in x times the error in y**, divided by the degrees of freedom



Correlation

Standardization:

We can standardize any deviation by dividing it by the **standard deviation** of the measure ($\sqrt{}$variance)

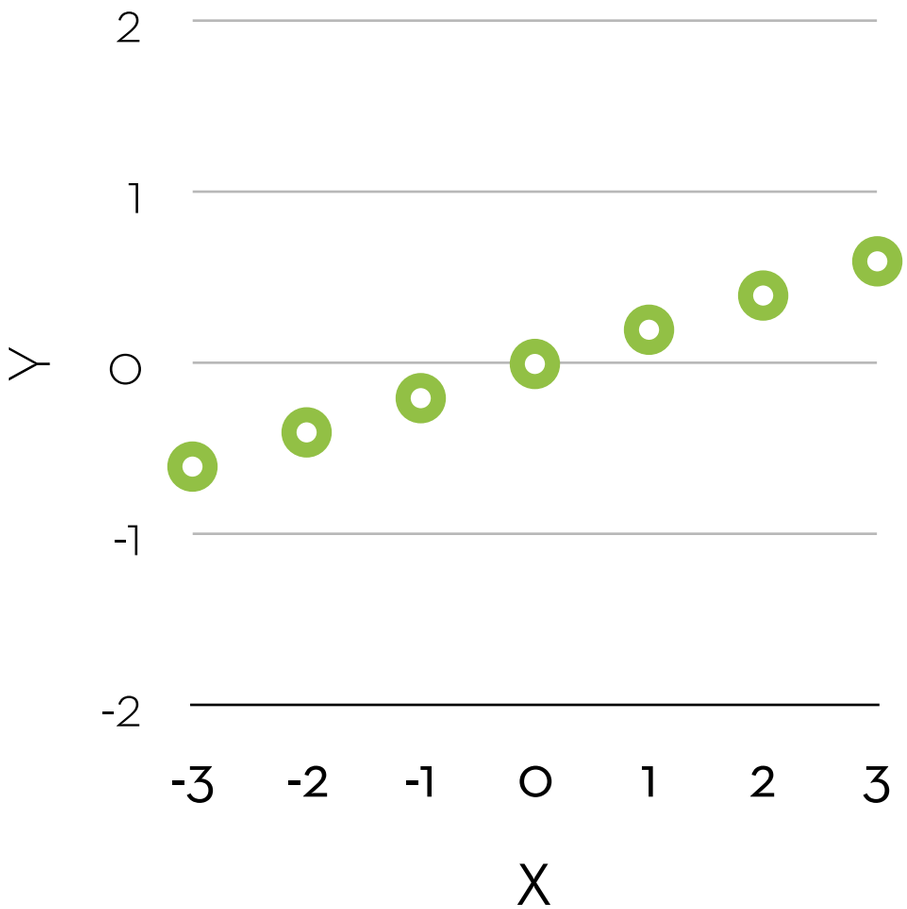
If we want to standardize the covariance, we divide by **both** the standard deviation of x and the standard deviation of y.

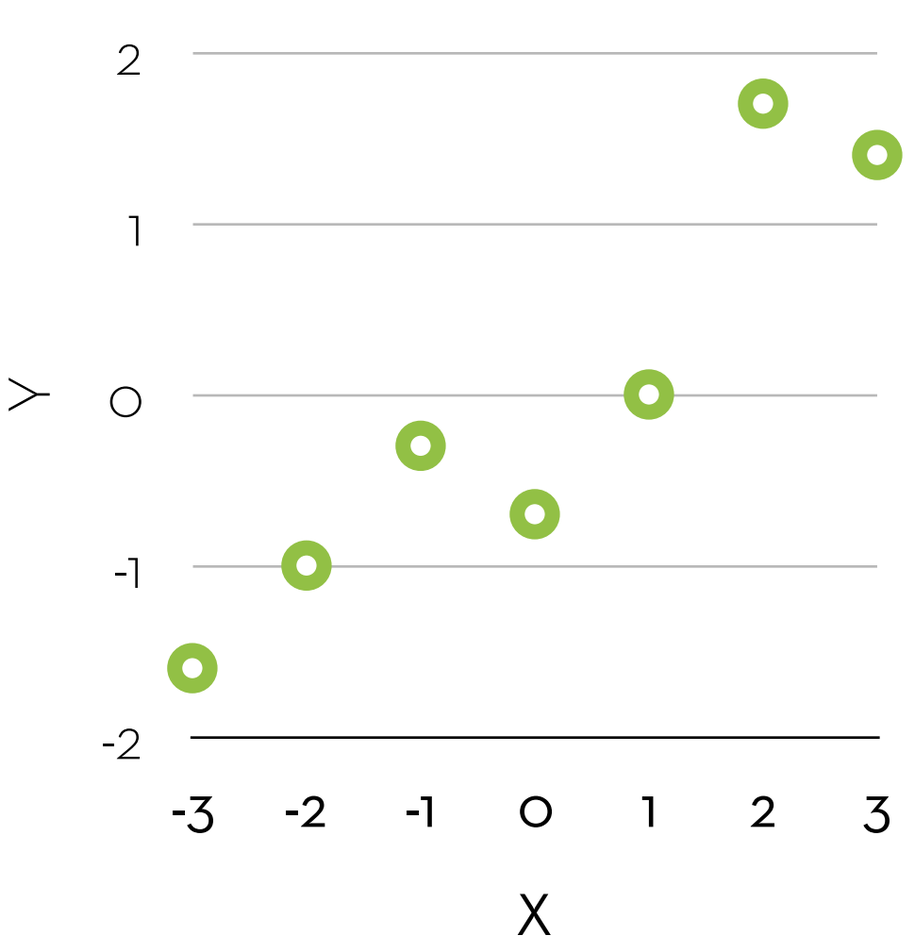
The resulting metric is the **correlation coefficient**:

$$r = cov(x,y)/s_x s_y = \sum(x_i - mean_x)(y_i - mean_y)/(N-1)s_x s_y$$

Correlation

Which of these two graphs shows the strongest correlation?





Variables

Independent variables (X): things that are manipulated (experiment) or innate (survey)

- Low vs. high diversity
- Number of search results
- Gender
- Age

They are outside the participants' control (in the experiment)

Variables

Dependent variables (Y): things that are measured as an outcome of X

- Number of clicks
- Interaction time
- Facial expression
- Satisfaction*

Variables

Random variables (also X): variables that are not of interest, but they may influence Y, so we measure them just in case.

Control variables (not X): variables that are not of interest, but they may influence Y, so we try to keep them stable

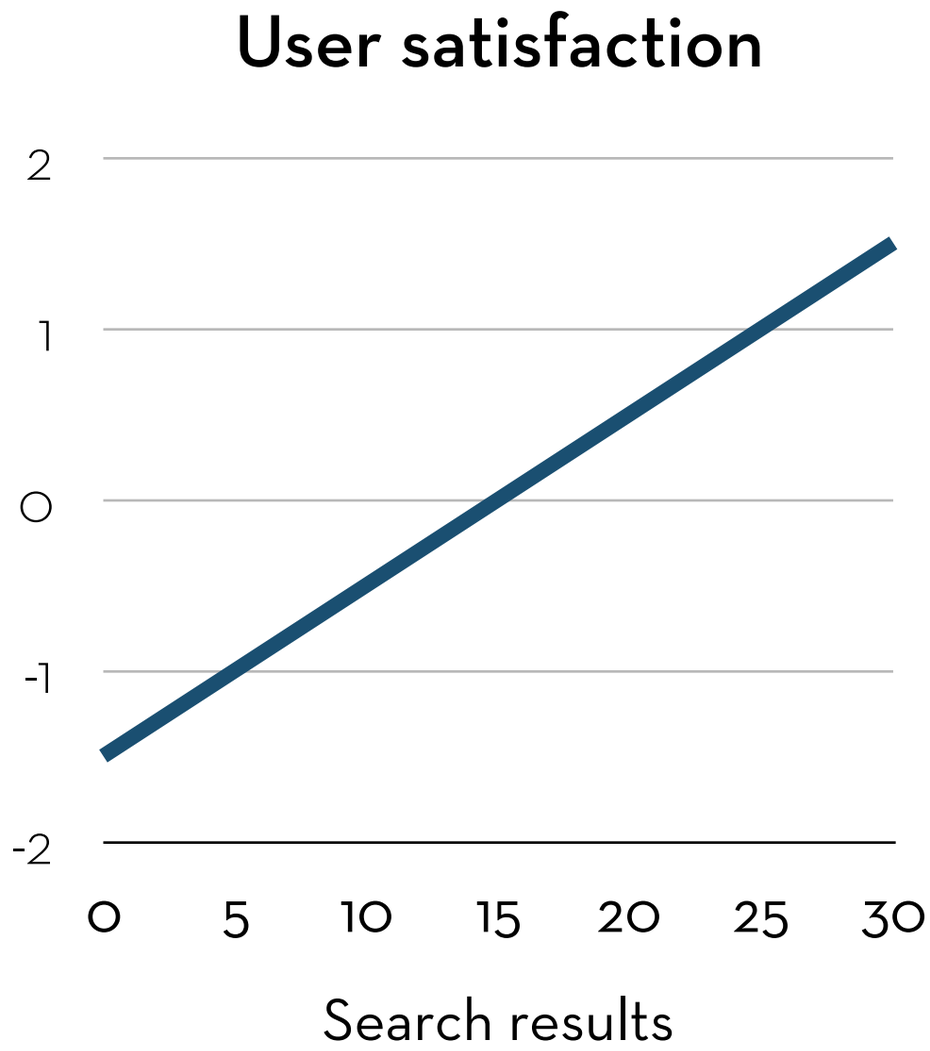
Linear regression

More of X -> more of Y:

Does user satisfaction increase with the number of search results?

More of X -> less of Y:

Does Facebook usage satisfaction decrease with age?



Linear regression

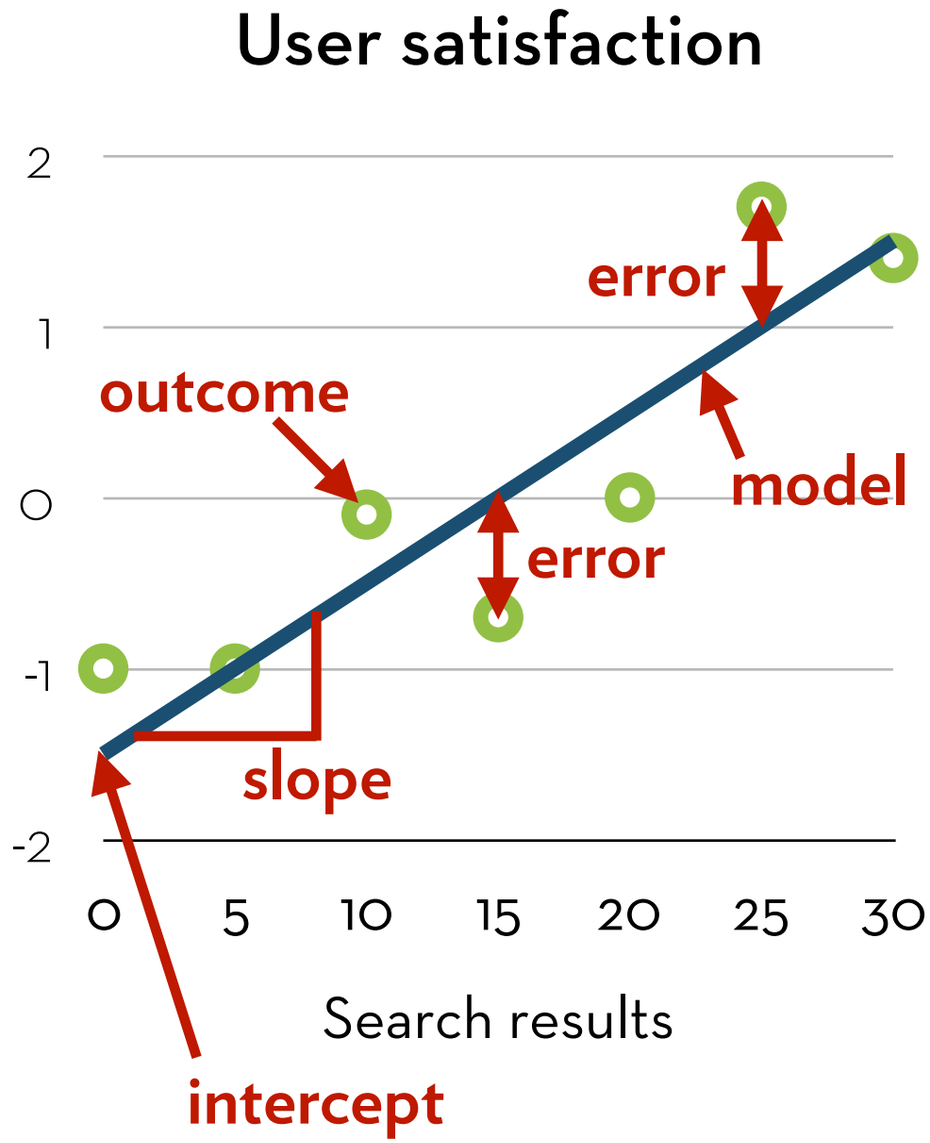
Any type of model:

$$outcome_i = model + error_i$$

Linear regression:

The model is a line with an intercept (a) and a slope (b)

$$Y_i = a + bX_i + e_i$$



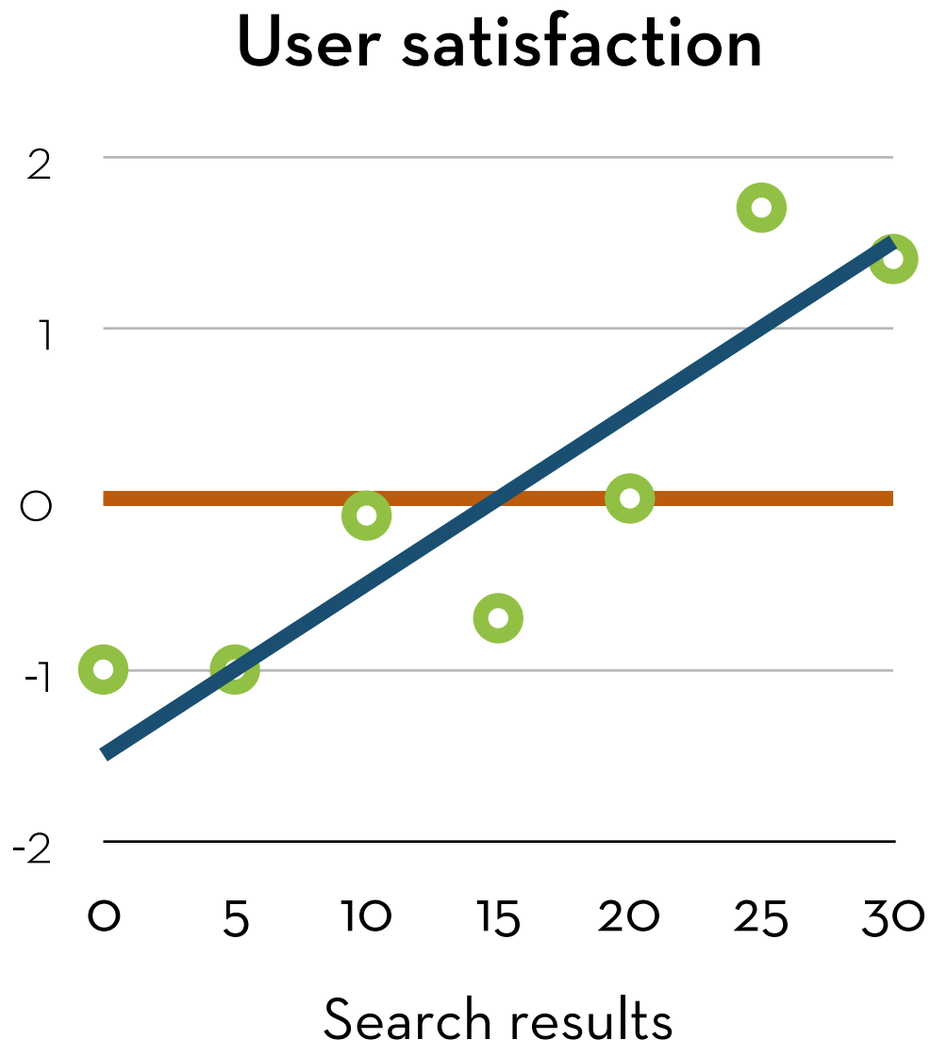
Linear regression

How good is the **model**?

We can use deviation for this as well!

Compare against the deviation of the simplest model

In this case: the mean



Multiple Regression

$outcome_i = model + error_i$

Multiple regression:

The model is a line with an intercept (a) and **several** slopes ($b_1 ... b_n$)

$$Y_i = a + b_1X_{1i} + b_2X_{2i} + \ldots + b_nX_{ni} + e_i$$

This means you can predict satisfaction using usability **and** gender, in each case controlling for the other variable

Note: bs are partial correlations (not the same as r!)

Multiple Regression

E.g.: $satisfaction_i = 1.00 + 2.00^*usability_i + 1.50^*gender_i + e_i$

For every 1 point increase in usability, satisfaction is expected to increase by 2 points, controlling for gender

Controlling for usability, the satisfaction for males (1) is expected to be 1.5 points higher than for females (0)

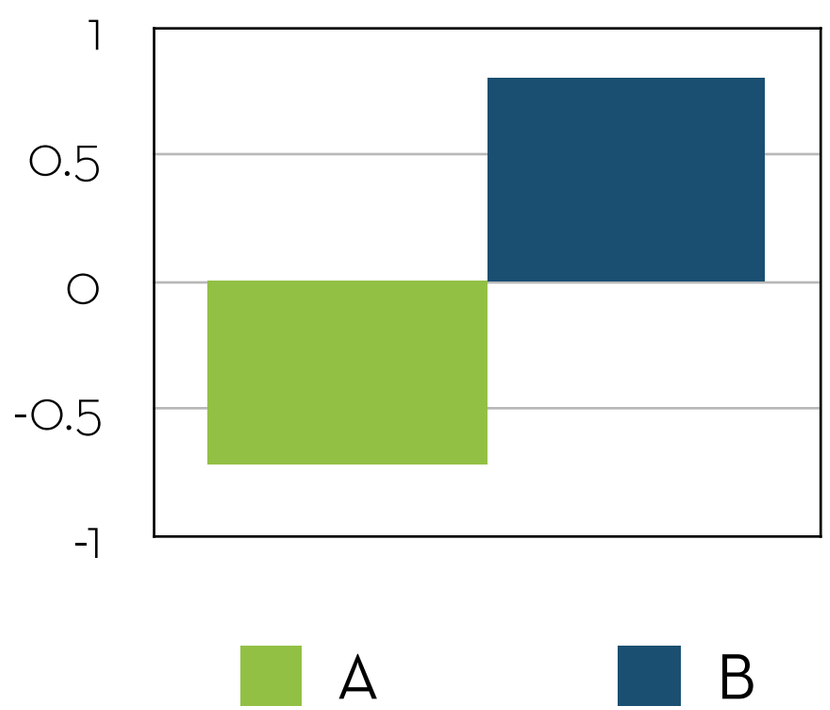
T-test

Difference between two systems:

Do these two UIs (A and B) lead to a different level of usability?

Differences between two groups of people:

Do men (A) and women (B) perceive different levels of usability?



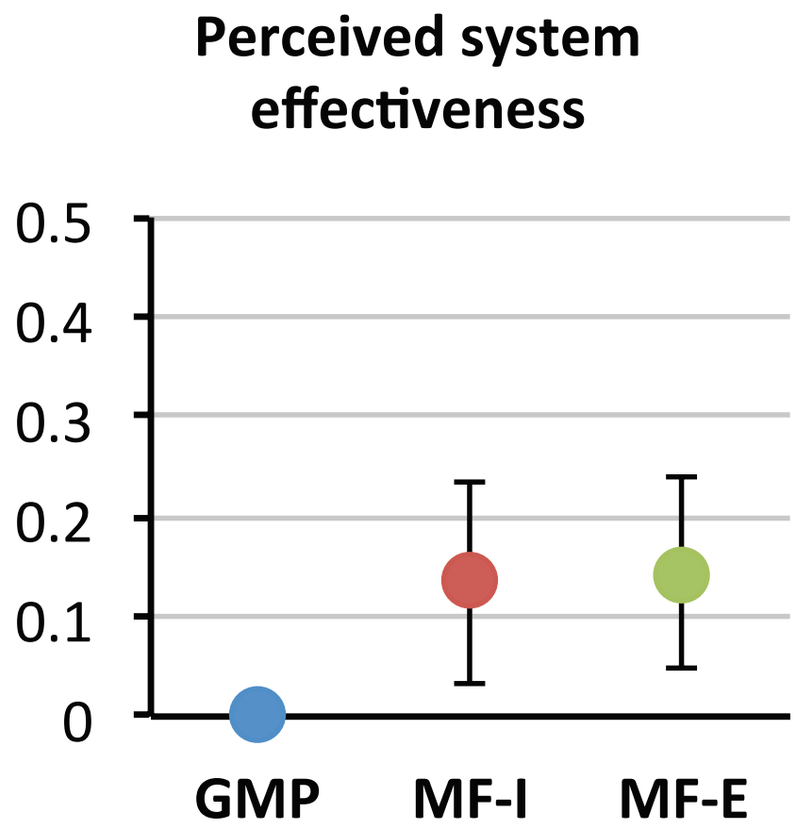
ANOVA

Differences between >2 systems / groups:

Are there differences in perceived system effectiveness between these 3 algorithms?

First do an omnibus test, then post-hoc tests or planned contrasts

Family-wise error!

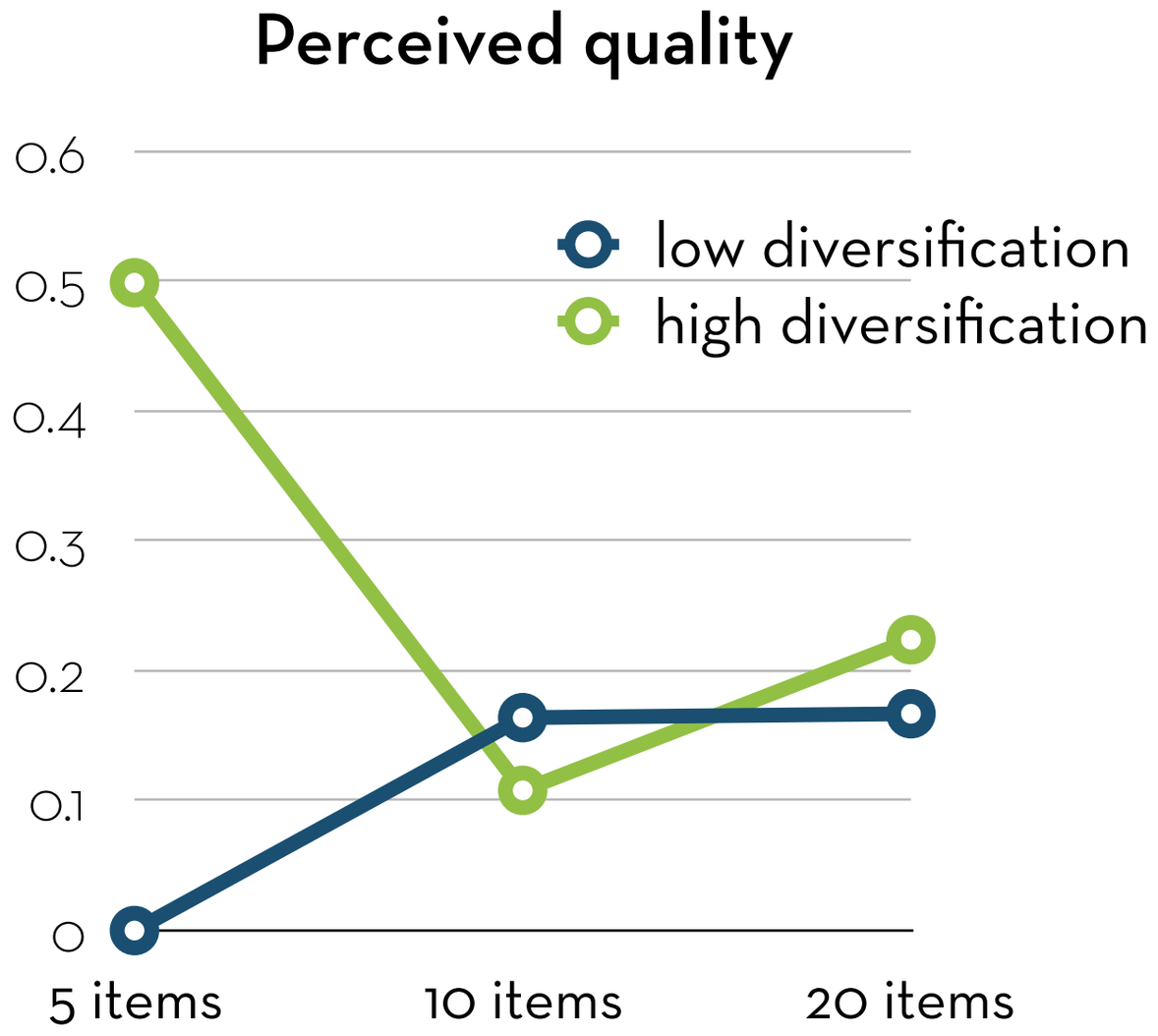


Factorial ANOVA

Two manipulations at the same time:

What is the combined effect of list diversity and list length on perceived recommendation quality?

Test for the interaction effect!



Willemsen et al.: "Understanding the Role of Latent Feature Diversification on Choice Difficulty and Satisfaction", UMUAI

Y is not normal

Standard tests assume that the dependent variable (Y) is an continuous, unbounded, normally distributed interval variable

Continuous: variable can take on any value, e.g. 4.5 or 3.23 (not just whole numbers)

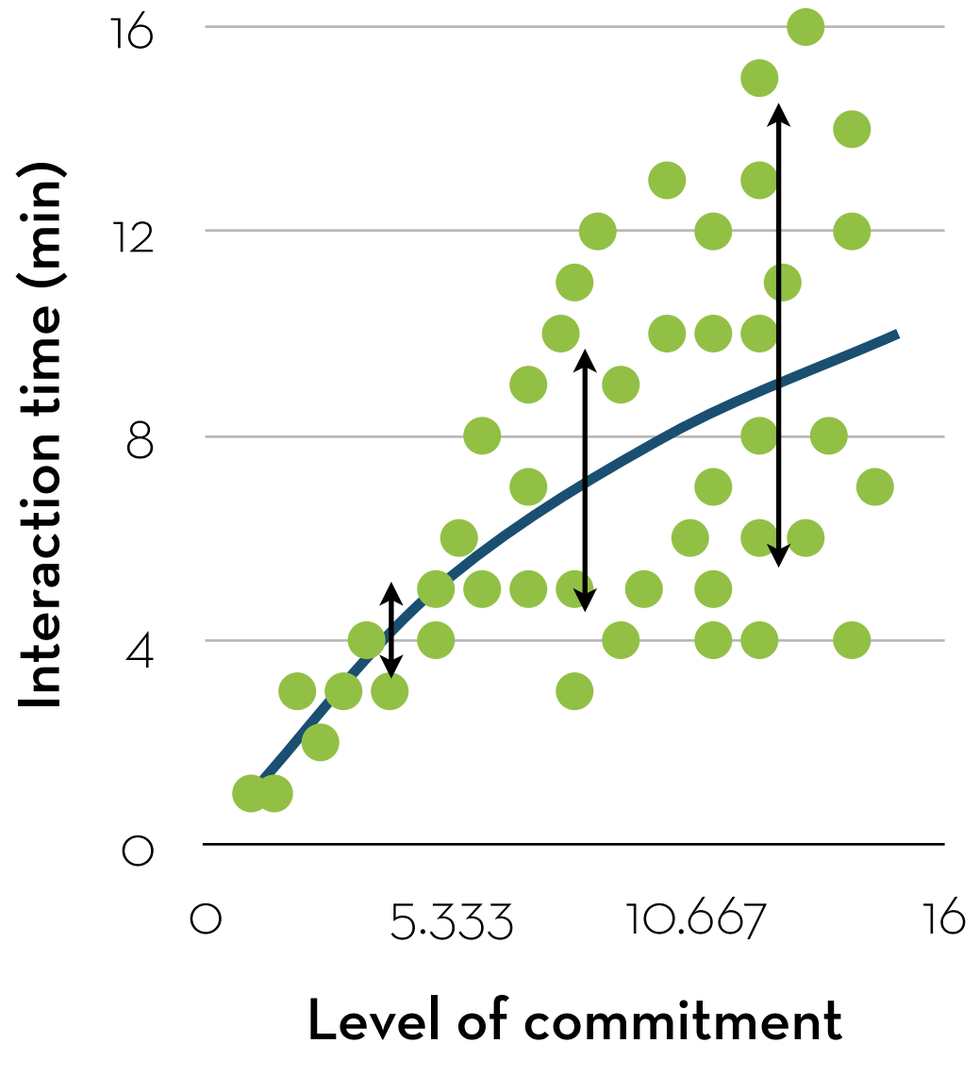
Unbounded: range of values is unlimited (or at least does not stop abruptly)

Interval: differences between values are comparable; is the difference between 1 and 2 the same as the difference between 3 and 4?

Y is not normal

Not true for most behaviors!

- Number of clicks
- Time, money
- 1-5 ratings
- Decisions



Logistic regression

Linear regression:

$$Y_i = a + b_1X_{1i} + b_2X_{2i} + \ldots + b_kX_{ki} + e_i$$

What if Y is binary (0 or 1)?

We can try to predict the **probability** of Y=1 — P(Y)

However, this probability is a number between 0 and 1

For linear regression, we want an unbounded linear Y!

Can we find some transformation that allows us to do this?

Yes: $P(Y) = 1 / (1+e^{-U})$

Logistic regression

$P(Y) = 1 / (1+e^{-U})$

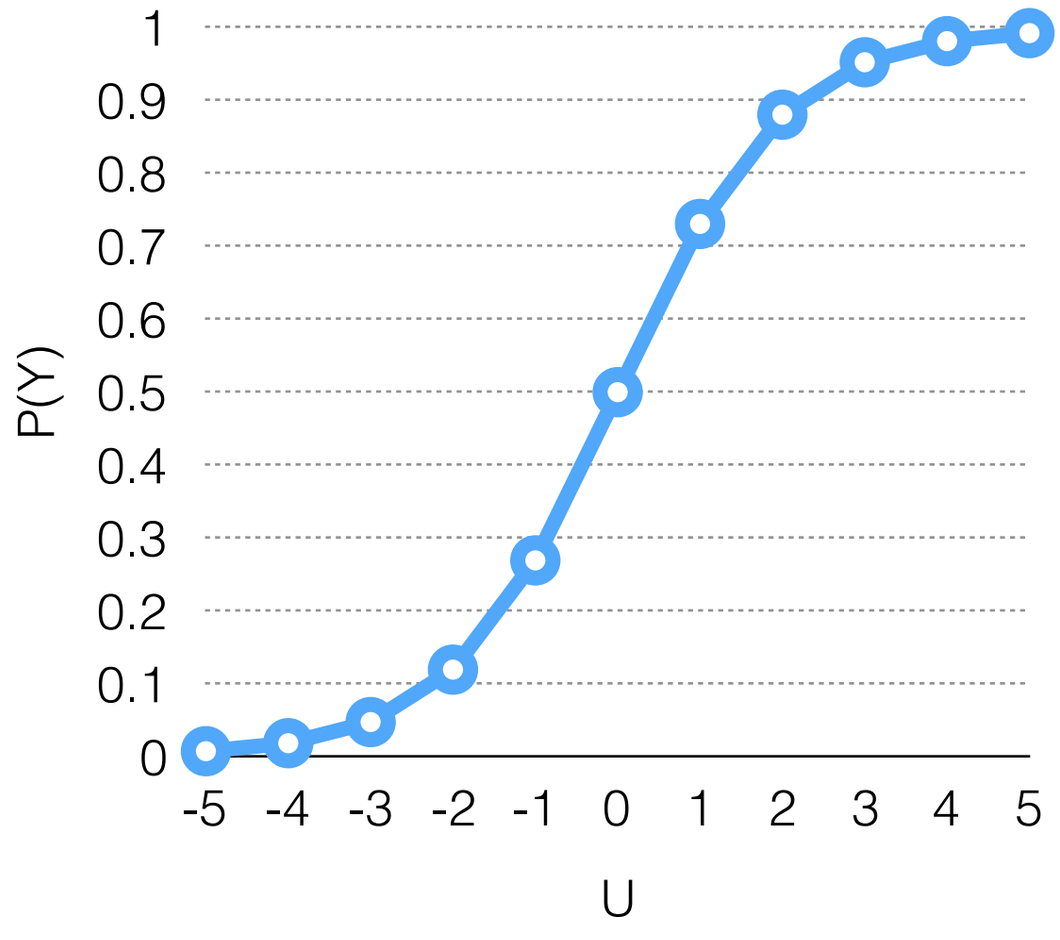
Conversely:

$U = \ln(P(Y)/(1-P(Y)))$

Interpretation:

$P(Y)/(1-P(Y))$ is the **odds** of Y

Therefore, U is the log odds, or **logit** of Y



Correlated errors

Standard regression requires **uncorrelated errors**

This is not the case when...

...you have repeated measurements of the same participant (e.g. you measured 5 task performance times per participant, for 60 participants)

...participants are somehow related (e.g. you measured the performance of 5 group members, for 60 groups)

Poisson regression

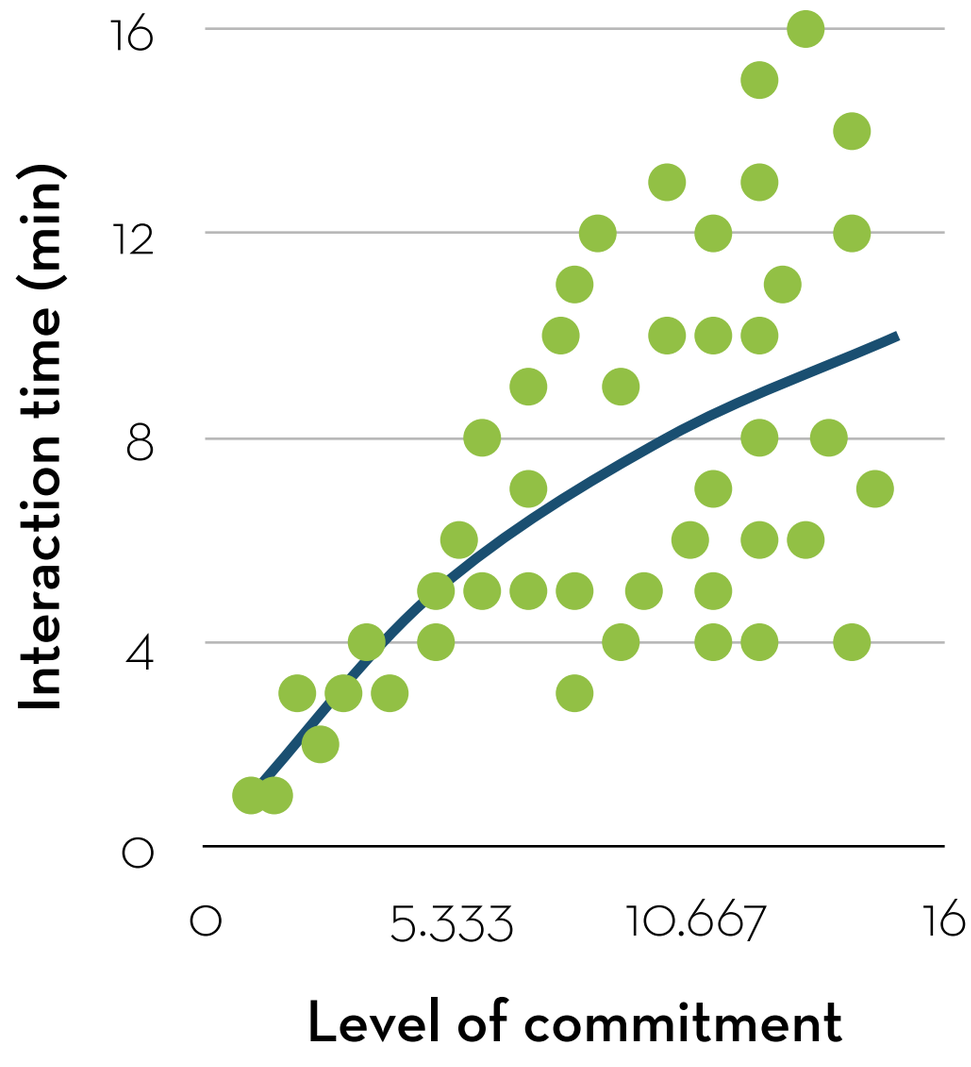
Count variables often look like this

Examples: # of purchases, # of clicks, time*, price*

Not normal, heteroscedastic!

Can we find some transformation that makes this work?

Yes: $Y = e^U$



Coefficients

How to interpret the b coefficients?

b is the increase in U for each increase of X

b is the increase in the **log rate** of Y for each increase in X

e^b is the ratio of rate Y for each increase in X

e^b is the **rate ratio**

Why the ratio?

$b = \log(rate_{X+1}) - \log(rate_X) = \log(rate_{X+1} / rate_X)$

therefore, $e^b = rate_{X+1} / rate_X$

Ordered logistic

Question: “I only act to satisfy immediate concerns, figuring the future will take care of itself.”

Answer categories:

1=extremely uncharacteristic

2=somewhat uncharacteristic

3=uncertain

4=somewhat characteristic

5=extremely characteristic

A problem...

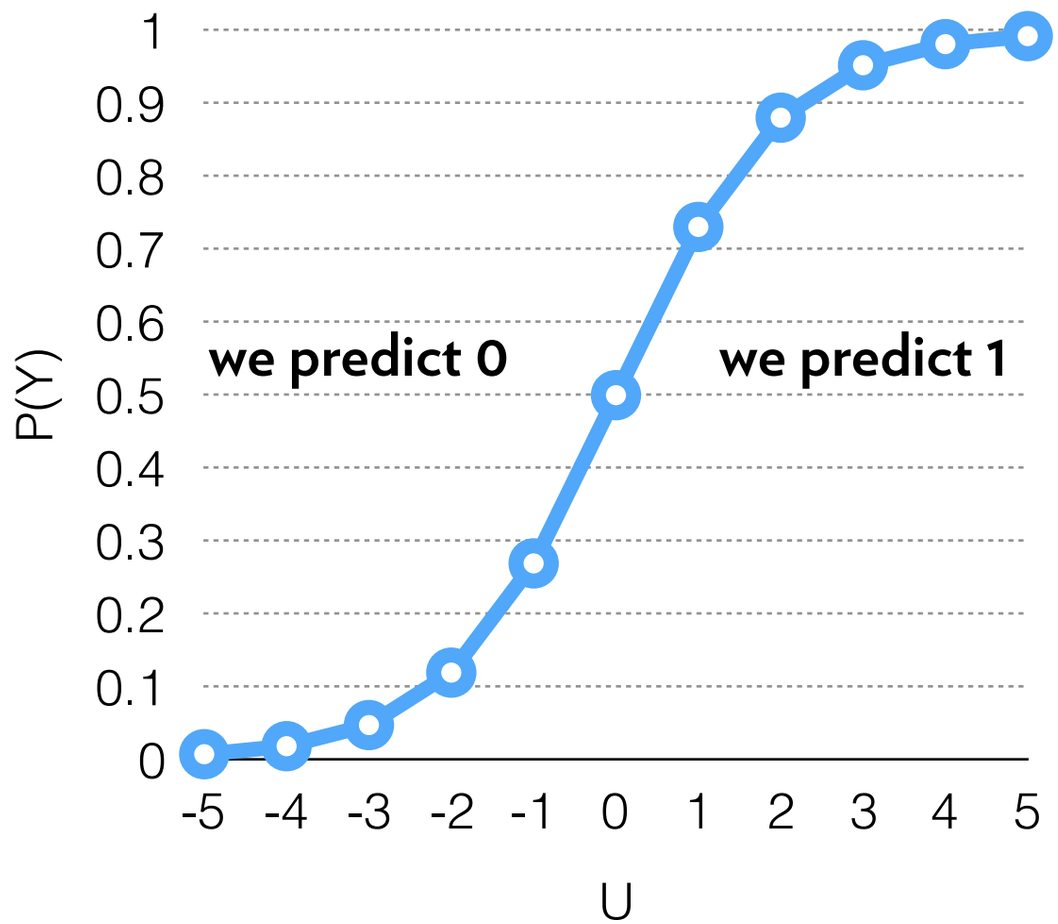
This is ordinal, not interval!

Is the difference between “extremely uncharacteristic” and “somewhat uncharacteristic” the same as the difference between “uncertain” and “somewhat characteristic”?

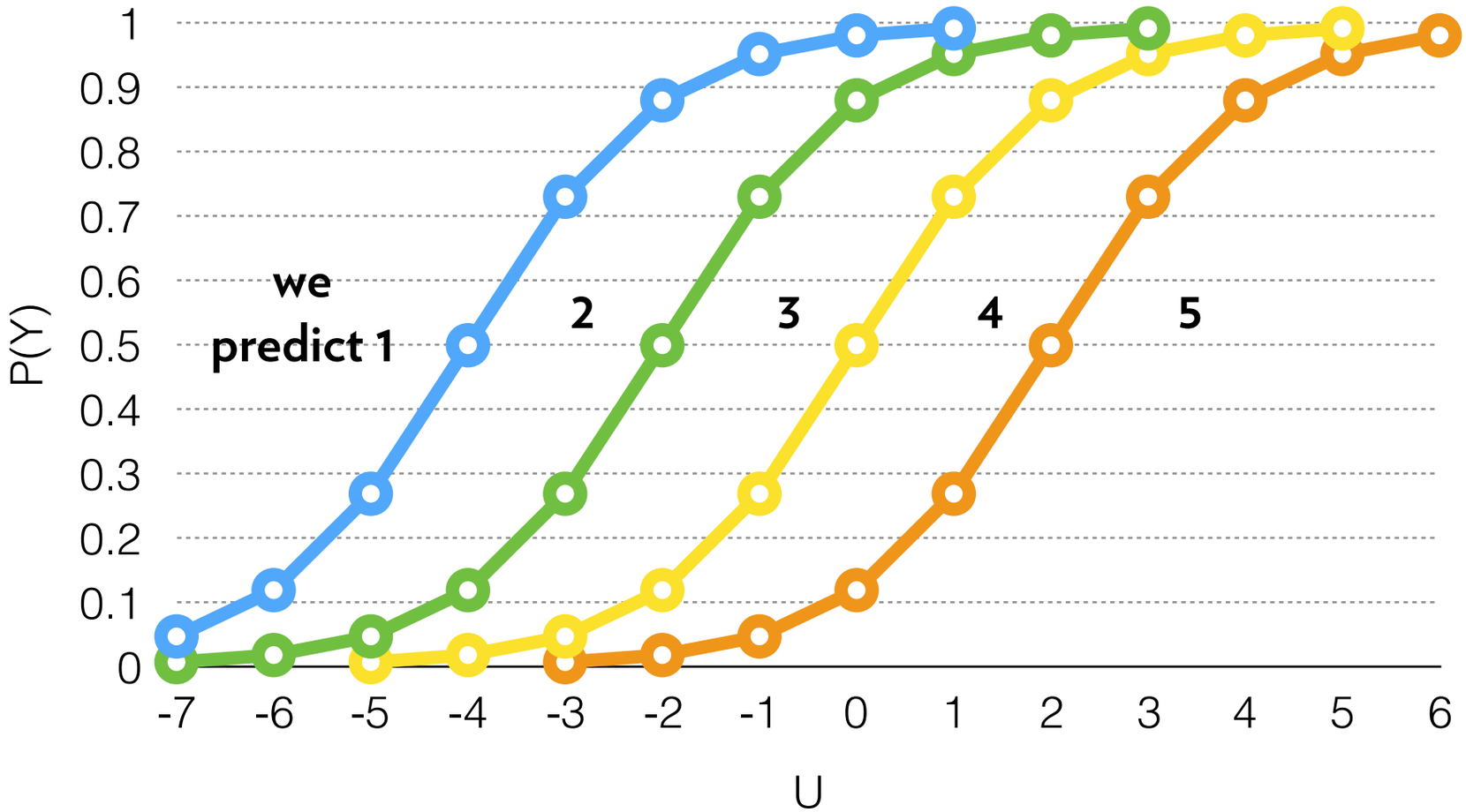
Also, likely not very normally distributed!

How can we solve these problems?

Logistic regression



Ordered logistic



Coefficients

The model estimates intercepts for each threshold

1|2, 2|3, 3|4, 4|5

These thresholds are the **log odds** of any person having **at least** this value

How to interpret the b coefficients?

e^b is the **odds ratio** for a 1pt increase in X

e.g. if the odds ratio is 1.40, then the odds of a higher value increase by 40% if X is 1 higher

Multi-level models

Repeated measurements

e.g. participants make 30 decisions

(Partially) within-subjects design

e.g. participants are randomly assigned to 1 of 3 games, and tested once with sound on and once with sound off

Grouped data

e.g. participants perform tasks in groups of 5

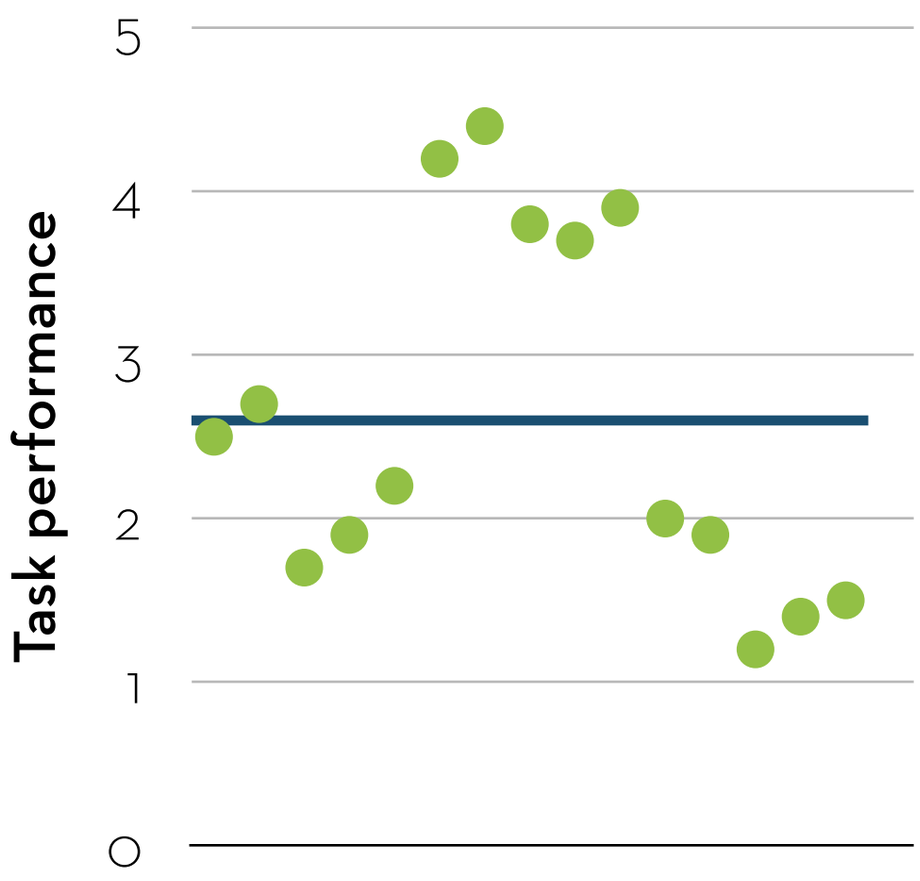
A combination of the above

Multi-level models

Consequence: errors are correlated

There will be a user-bias (and maybe an task-bias)

Golden rule: data-points should be **independent**

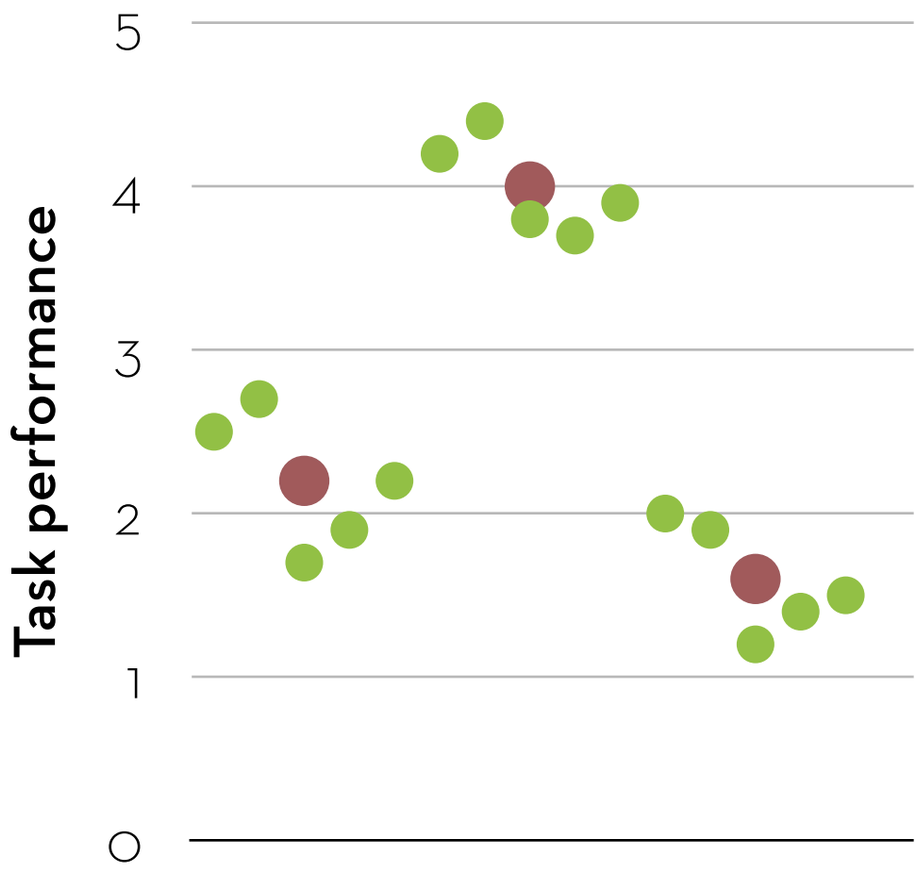


Multi-level models

Take the average of the repeated measurements

Reduces the number of observations

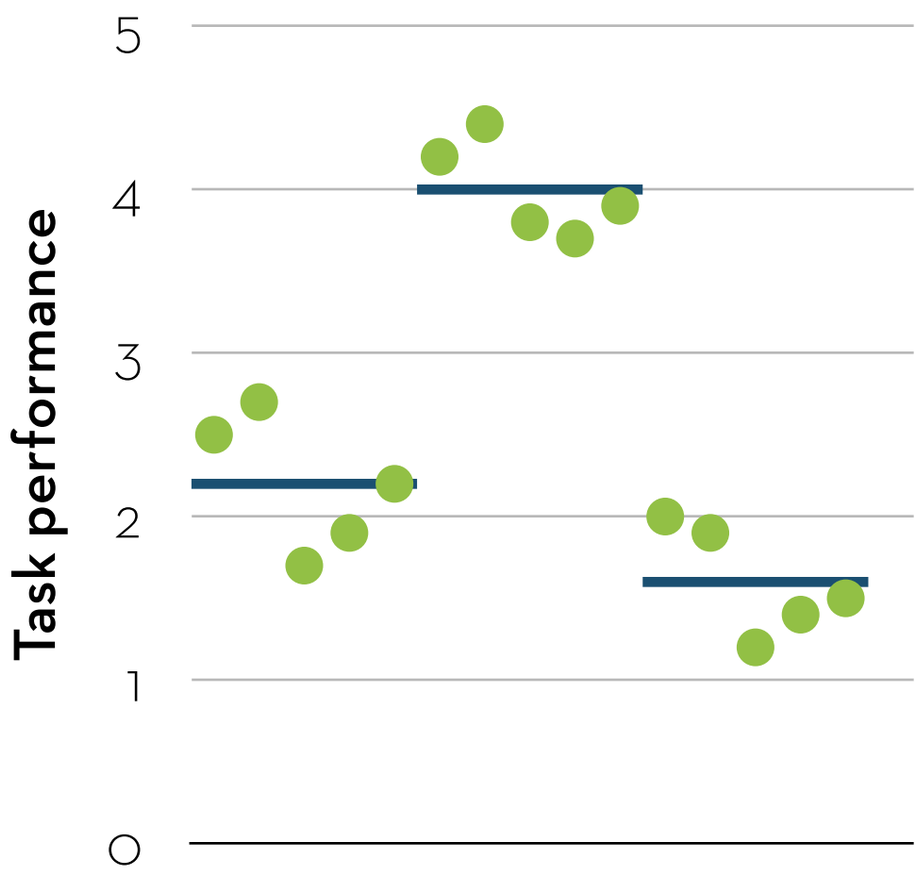
It becomes impossible to make inferences about individual tasks/users/etc.



Multi-level models

In regression:

- define a random intercept for each user
- impose an error covariance structure



Y is unobserved

Behavior is an “observed” variable

- Relatively easy to quantify
  - E.g. time, money spent, click count, yes/no decision

Perceptions, attitudes, and intentions (subjective valuations) are “unobserved” variables

- They happen in the user’s mind

How can we quantify them?

How to quantify?

Psychometrics:

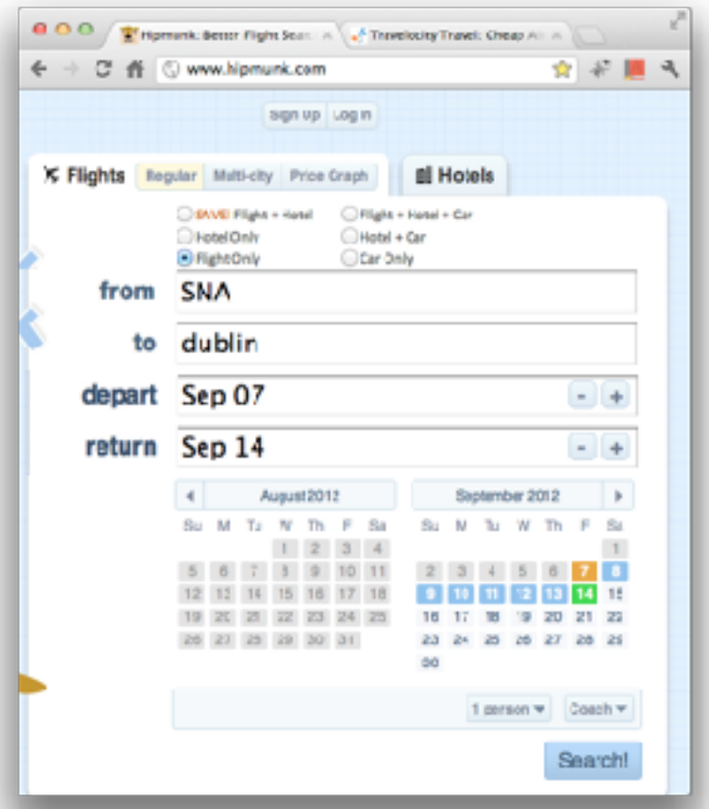
Ask multiple questions on a 5- or 7-point scale

E.g. perceived system effectiveness:

- "Using the system is annoying"
- "The system is useful"
- "Using the system makes me happy"
- "Overall, I am satisfied with the system"
- "I would recommend the system to others"
- "I would quickly abandon using this system"

Use **factor analysis** to validate the scales

Theory behind x->y



Why would the new system (X) have a higher usability (Y)?

Theory behind x->y

To learn something from a study, we need a **theory** behind the effect

- This makes the work generalizable
- This may suggest future work

Measure **mediating** variables

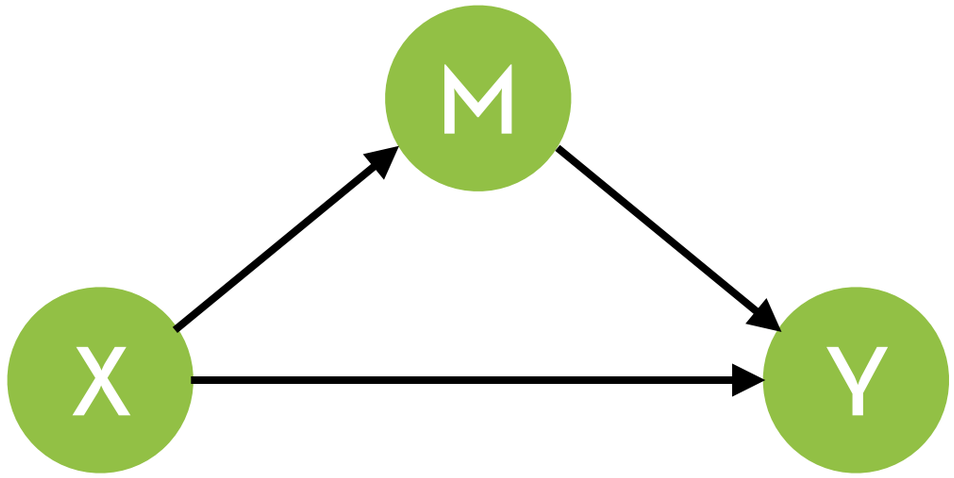
- Measure subjective system aspects
- Find out how they mediate the effect on user experience

Statistical method: **structural equation modeling** (SEM)

Mediation analysis

Manipulation -> perception
-> experience

Does the system
influence usability
via understandability?



Types of mediation

Partial mediation

Full mediation

Negative mediation

Mediation Analysis

Manipulation -> perception -> experience

Does the system influence usability via understandability?

Types of mediation

Partial mediation

Full mediation

Negative mediation

Mediation Analysis

Manipulation -> perception -> experience

Does the system influence usability via understandability?

Types of mediation

Partial mediation

Full mediation

Negative mediation

Mediation Analysis

Manipulation -> perception -> experience

Does the system influence usability via understandability?

Types of mediation

Partial mediation

Full mediation

Negative mediation

Mediation Analysis

Manipulation -> perception -> experience

Does the system influence usability via understandability?

Types of mediation

Partial mediation

Full mediation

Negative mediation

Mediation Analysis

Manipulation -> perception -> experience

Does the system influence usability via understandability?

Types of mediation

Partial mediation

Full mediation

Negative mediation

Example

6 jams

Less attractive

30% sales

Higher choice satisfaction

ıst

24 jams

More attractive

3% sales

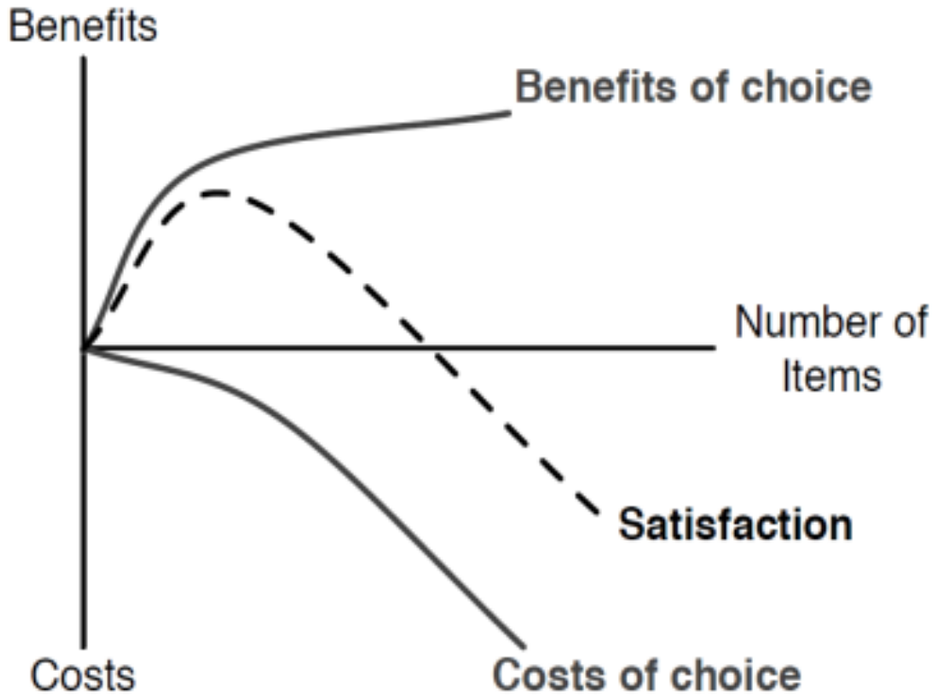
Lower choice satisfaction

Example

Satisfaction = benefit – cost

Benefit of more options: easier to find the right option

Cost of more options: more comparisons, higher potential regret



Is this also true for **recommendations**?

Example

Example from Bollen et al.: "Choice Overload"

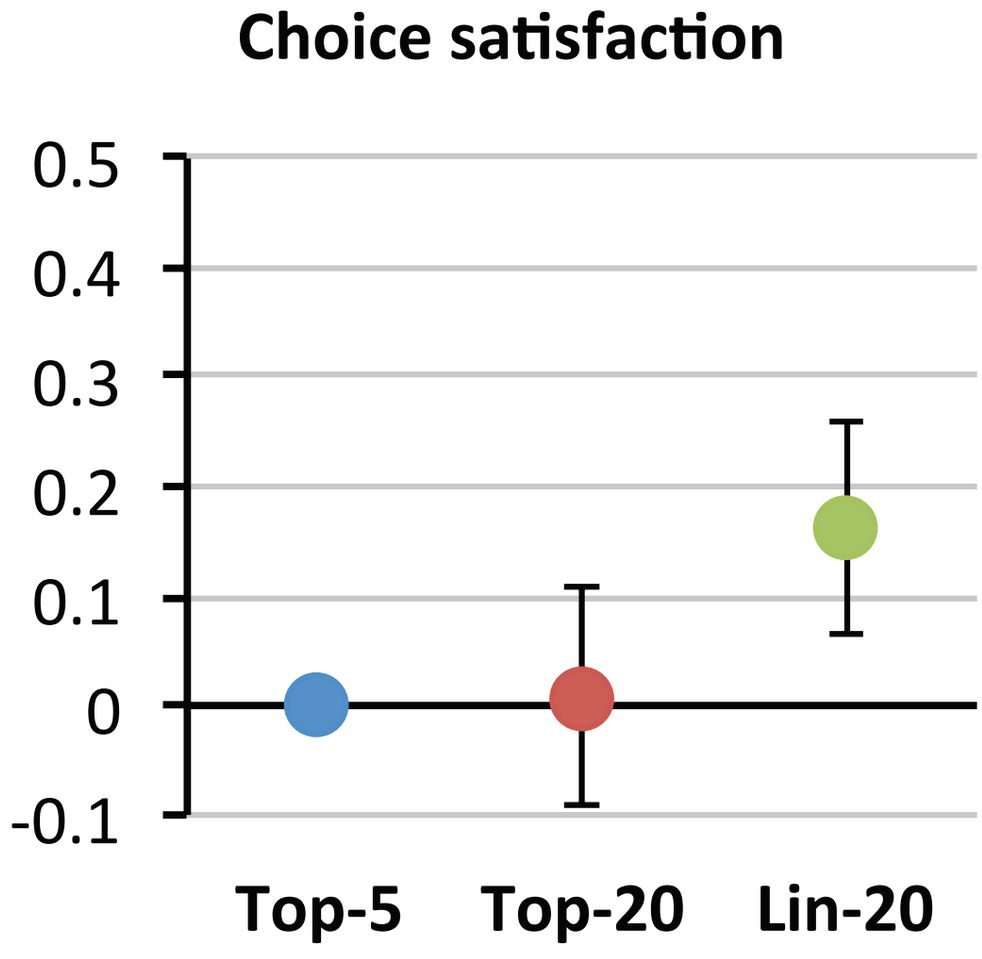
What is the effect of the number of recommendations?

What about the composition of the recommendation list?

Tested with **3 conditions**:

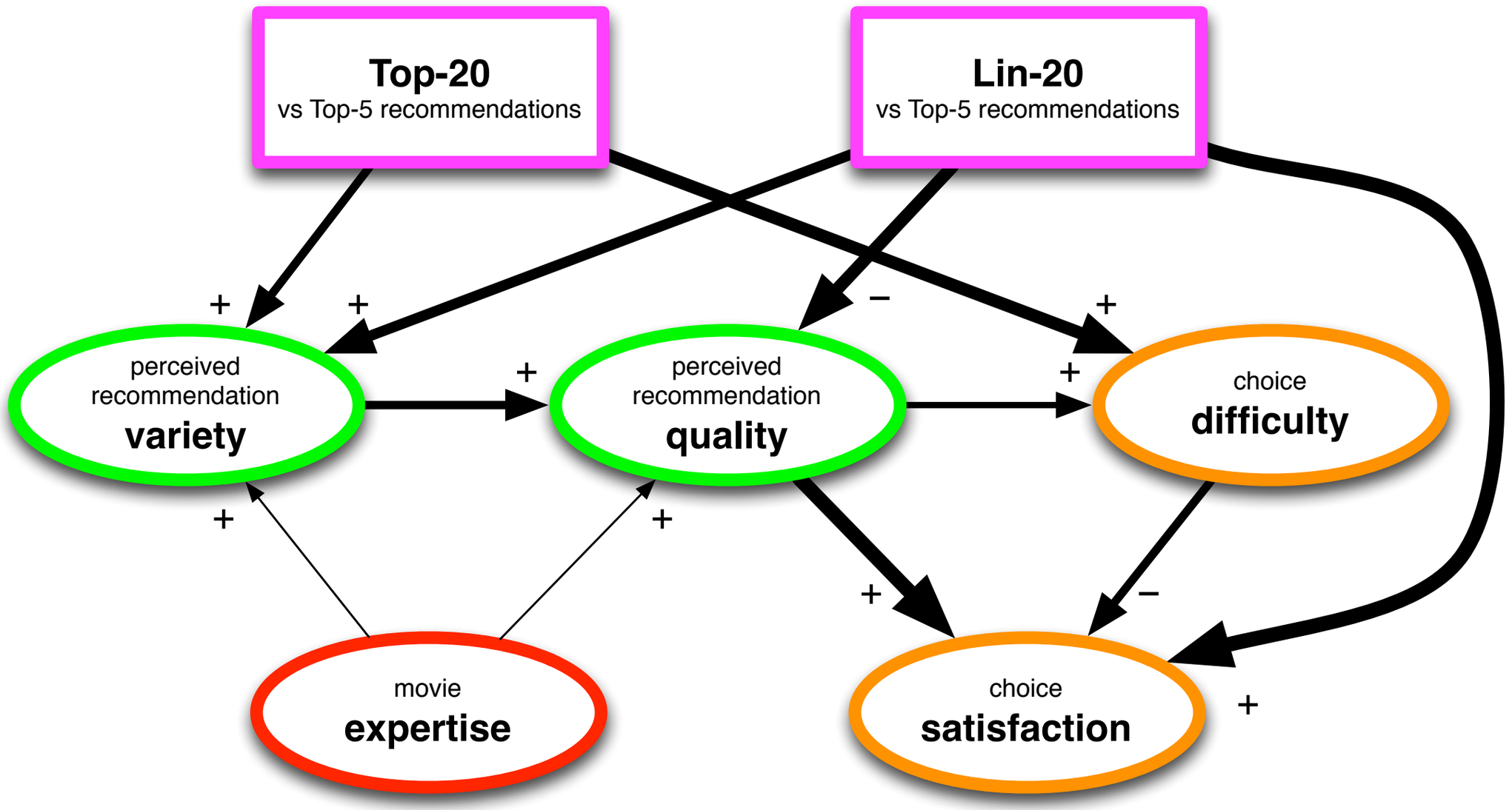
- Top 5:
  - recs: 1 2 3 4 5
- Top 20:
  - recs: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
- Lin 20:
  - recs: 1 2 3 4 5 99 199 299 399 499 599 699 799 899 999 1099 1199 1299 1399 1499

Example



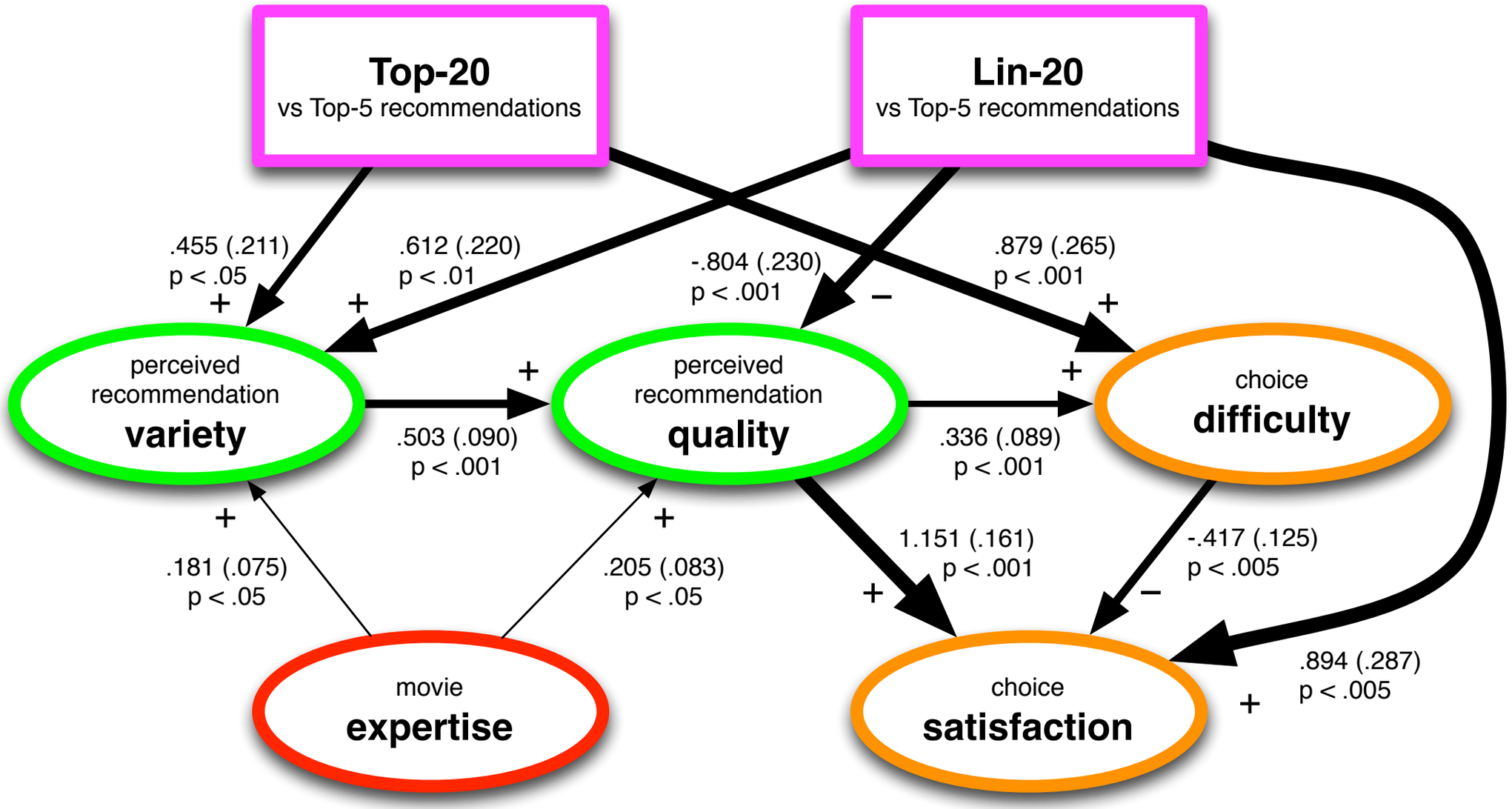
Bollen et al.: “Understanding Choice Overload in Recommender Systems”, RecSys 2010

Example



Bollen et al.: “Understanding Choice Overload in Recommender Systems”, RecSys 2010

Example



Bollen et al.: "Understanding Choice Overload in Recommender Systems", RecSys 2010