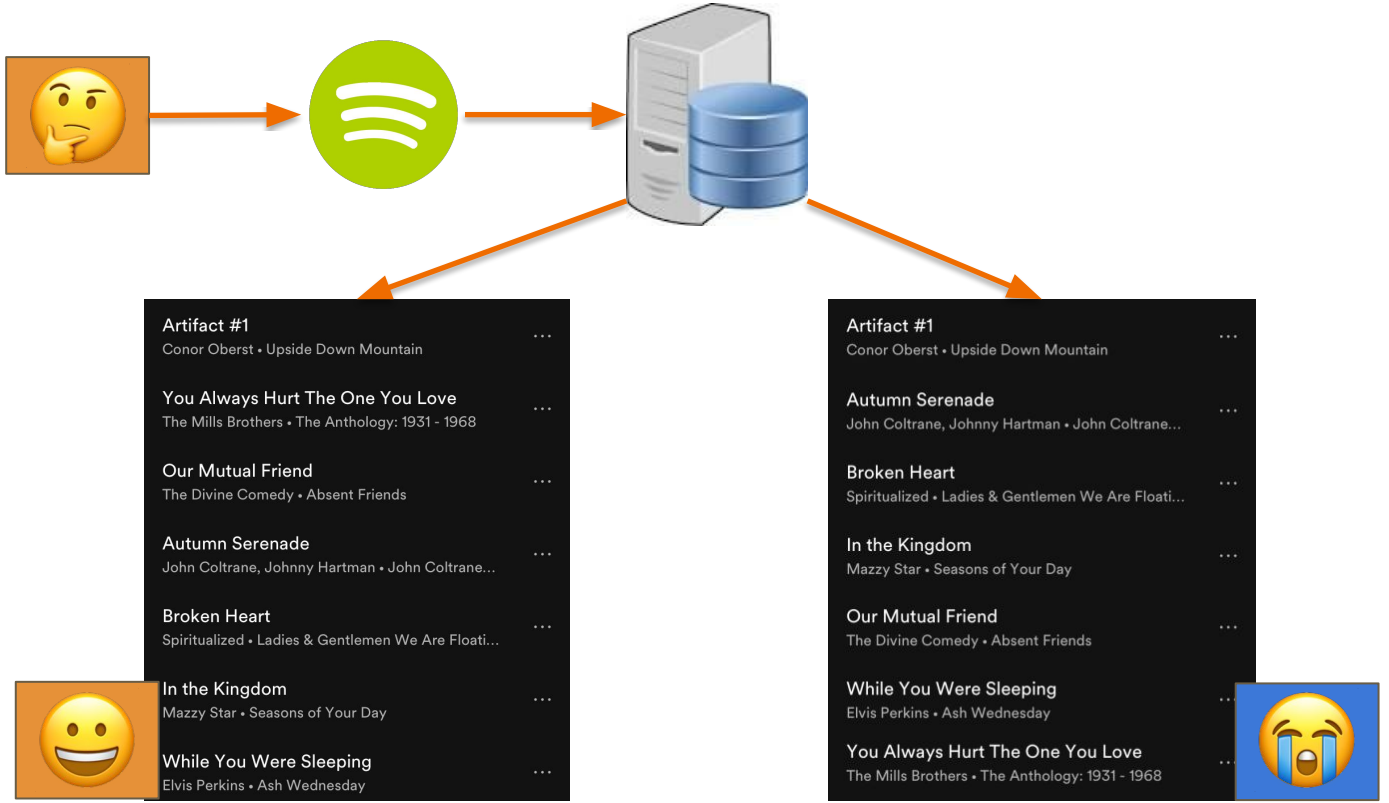
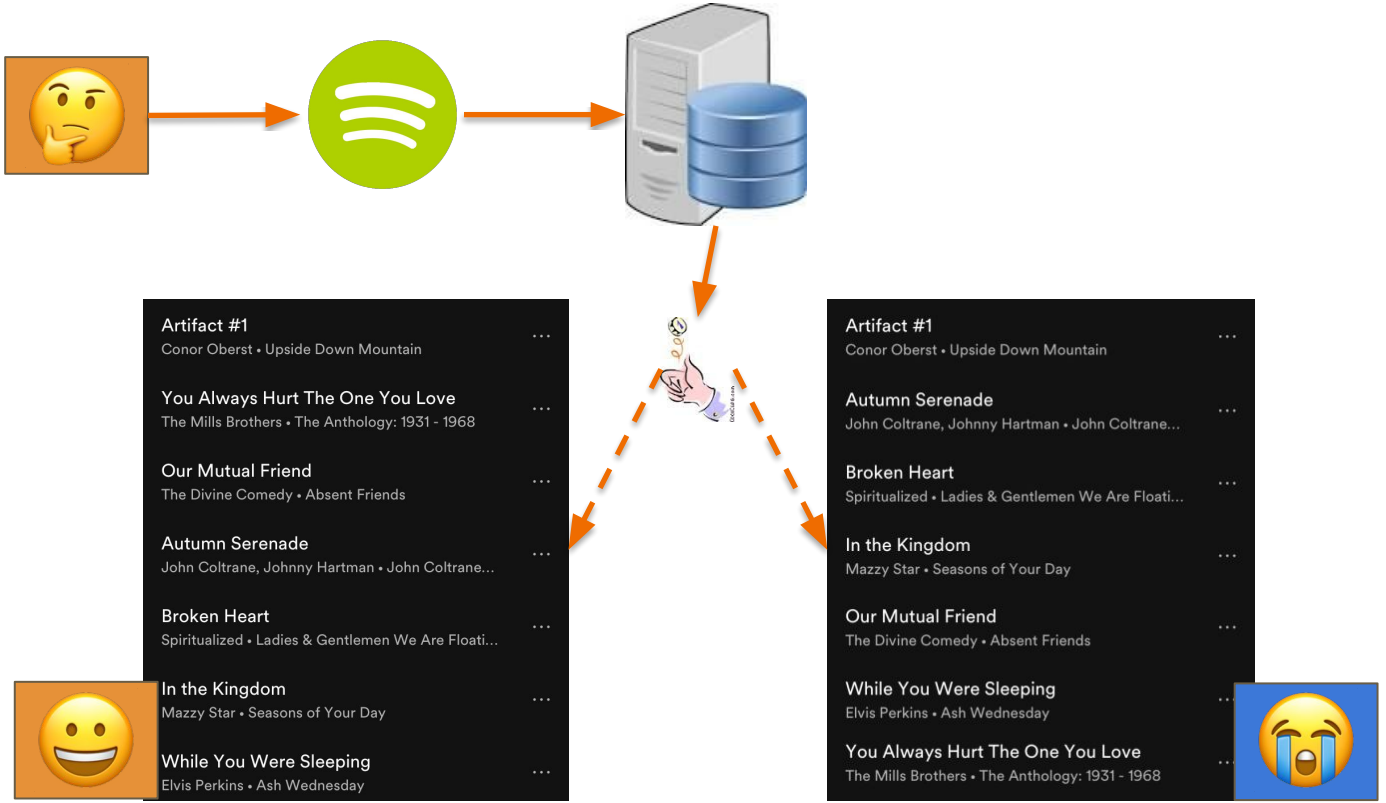
Hypothesis testing

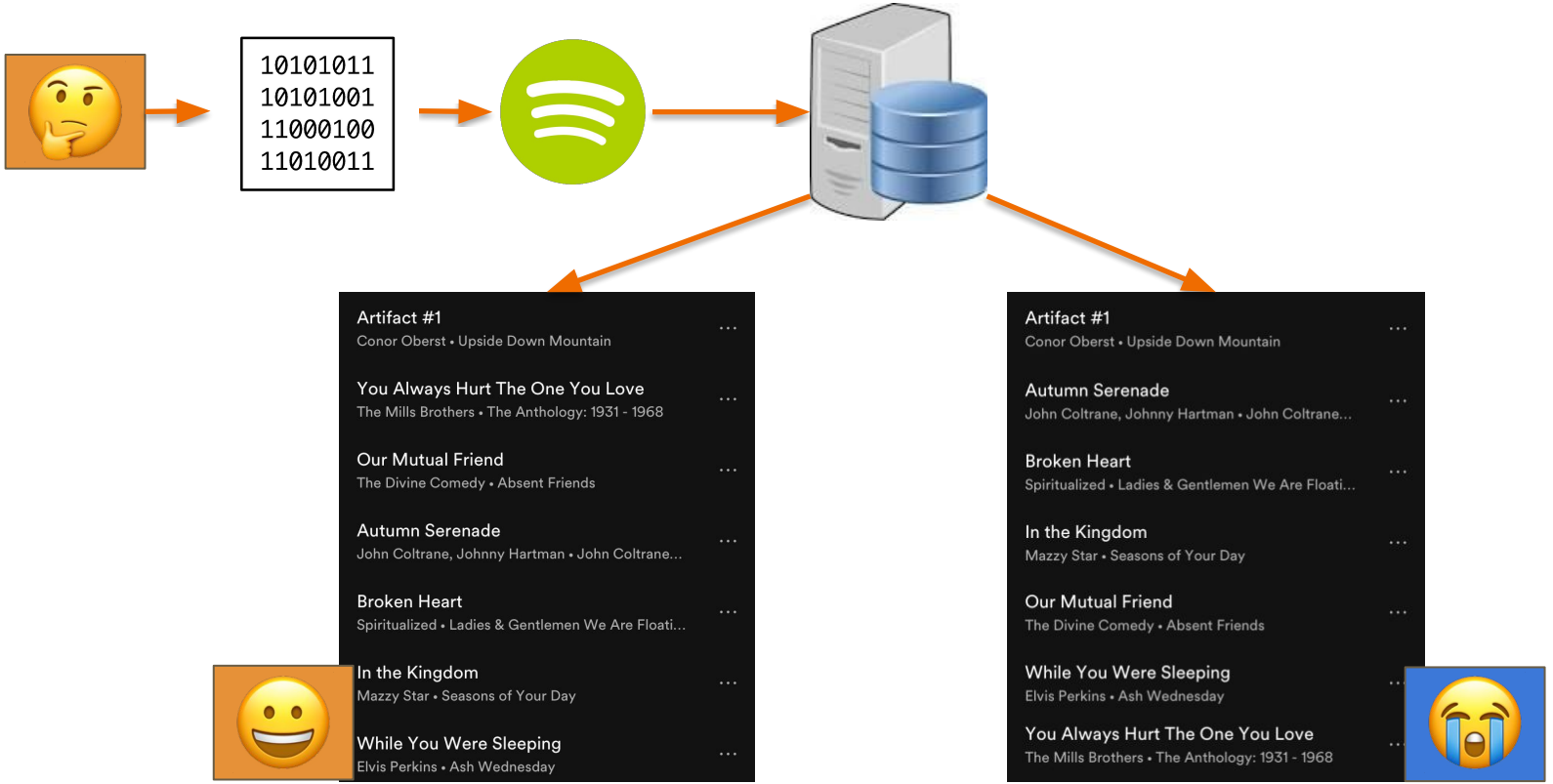
Recommender system experiments



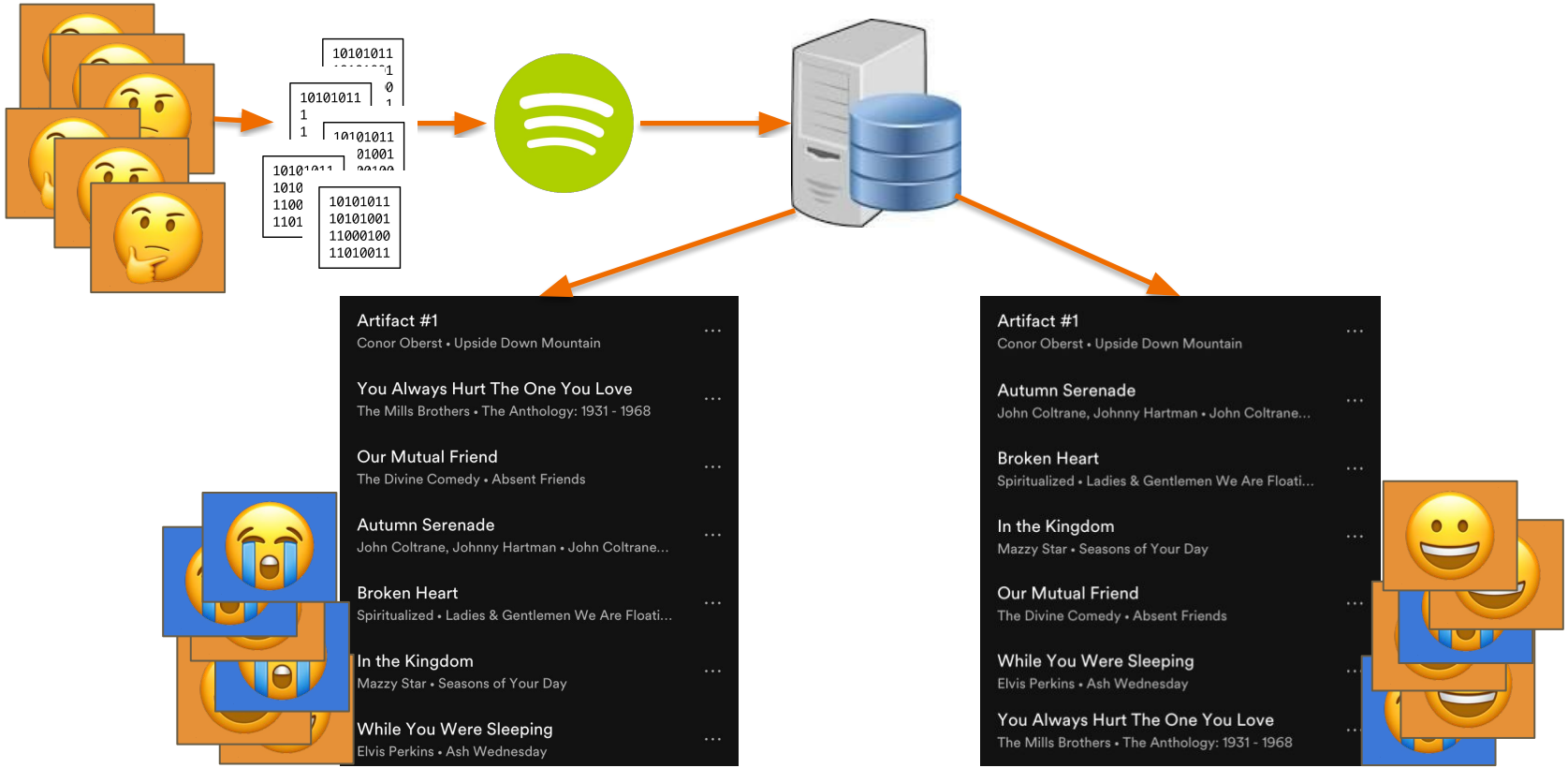
Rec sys experiments: A/B tests and user studies



Rec sys experiments: Offline experimentation



Experiments and measurement



Statistical Hypothesis Testing

- “Do these results support my hypothesis?”
- “Are these results meaningful?”
- “Is it possible that the differences are just random?”

Statistical hypothesis testing

- Significance tests lend rigor to our experimentation
  - Without them, the usual differences we see in experiments would be difficult to interpret
- But they are widely misunderstood
  - Test results can be incorrectly interpreted
  - Test results can be easily manipulated (even unintentionally)
- They are fundamentally no more rigorous than any AI/ML approach to classification
  - Though they may have a deeper theoretical basis

Outline of this presentation

Part 1: Testing statistical significance

Part 2: Fundamentals of significance testing

Part 3: Applications

Part 1

TESTING STATISTICAL SIGNIFICANCE

Six easy tests

- Non-parametric
  - Sign test / binomial test
  - Wilcoxon rank tests
- Parametric
  - T-test
  - ANOVA
- Distribution-free
  - Randomization test / Fisher exact test
  - Bootstrap test

Sign test (a.k.a. Binomial test)

User	A	B	B-A	sign(B-A)
1	.25	.35	+.10	+1
2	.43	.84	+.41	+1
3	.39	.15	-.24	-1
4	.75	.75	0	0
5	.43	.68	+.25	+1
6	.15	.85	+.70	+1
7	.20	.80	+.60	+1
8	.52	.50	-.02	-1
9	.49	.58	+.09	+1
10	.50	.75	+.25	+1

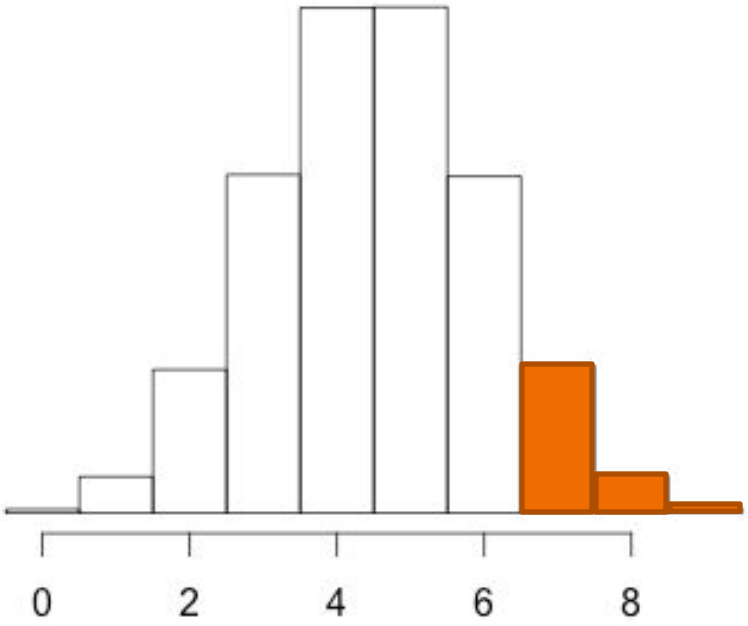
7 “successes” in 9 complete trials

What if each +1/-1 was just the result of flipping a fair coin?

What is the probability we would see 7 or more heads if the coin is fair?

Binomial distribution

What is the probability we would see 7 or more heads given a fair coin flipped 9 times?



P(7 heads | 9 trials, ½ probability)

+ P(8 heads | 9 trials, ½ probability)

+ P(9 heads | 9 trials, ½ probability)

= 0.09

p-value = 0.09

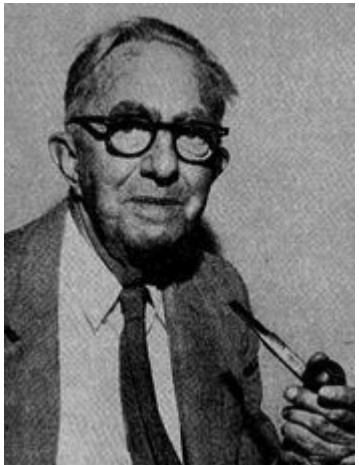
Wilcoxon signed-rank test

User	A	B	B-A
1	.25	.35	+.10
2	.43	.84	+.41
3	.39	.15	-.24
4	.75	.75	0
5	.43	.68	+.25
6	.15	.85	+.70
7	.20	.80	+.60
8	.52	.50	-.02
9	.49	.58	+.09
10	.50	.75	+.25

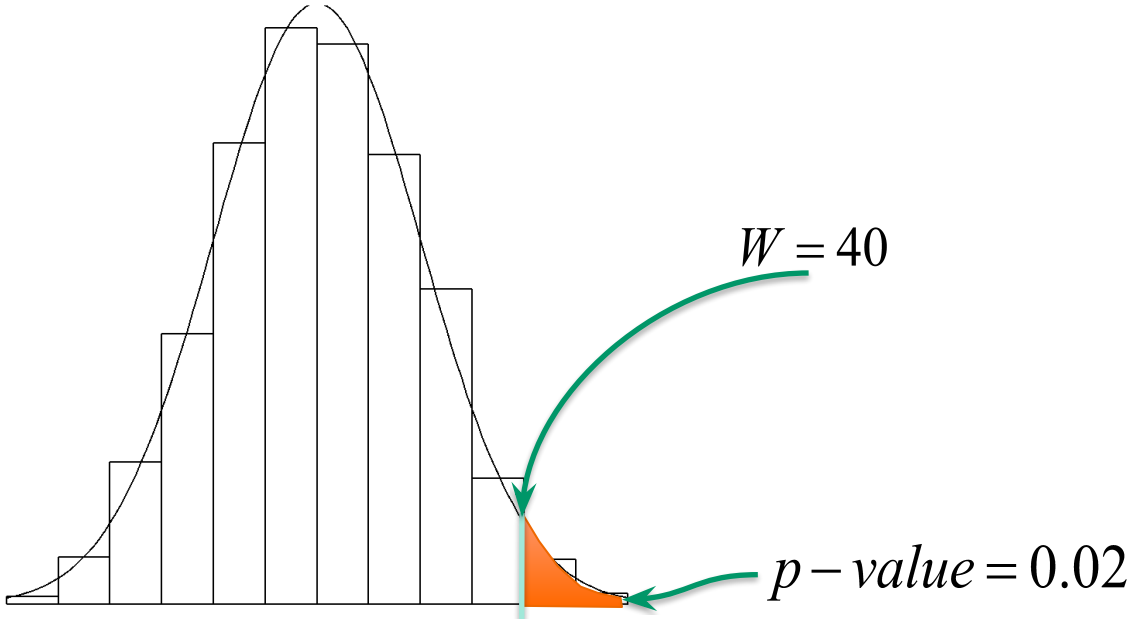
Rank	B-A
1	-.02
2	+.09
3	+.10
4	-.24
5.5	+.25
5.5	+.25
7	+.41
8	+.60
9	+.70

$$W = 2+3+5.5+5.5+7+8+9$$

$$W = 40$$



Wilcoxon signed-rank test



Matched pairs

- Offline tests can leverage *matched pairs*
  - Every user can be tested in every condition

- Online tests and user studies generally can’t

Wilcoxon rank-sum test

User	A	User	B
1	.25	11	.35
2	.43	12	.84
3	.39	13	.15
4	.75	14	.75
5	.43	15	.68
6	.15	16	.85
7	.20	17	.80
8	.52	18	.50
9	.49	19	.58
10	.50	20	.75

Rank	y	System
1.5	.15	A
1.5	.15	B
3	.20	A
4	.25	A
5	.35	B
6	.39	A
7.5	.43	A
7.5	.43	A
9	.49	A
10.5	.50	A
10.5	.50	B
12	.52	A
13	.58	B
14	.68	B
16	.75	A
16	.75	B
16	.75	B
18	.80	B
19	.84	B
20	.85	B

Sum up the ranks of the measurements produced by system B

a.k.a. Mann-Whitney U

Student's t-test for matched pairs

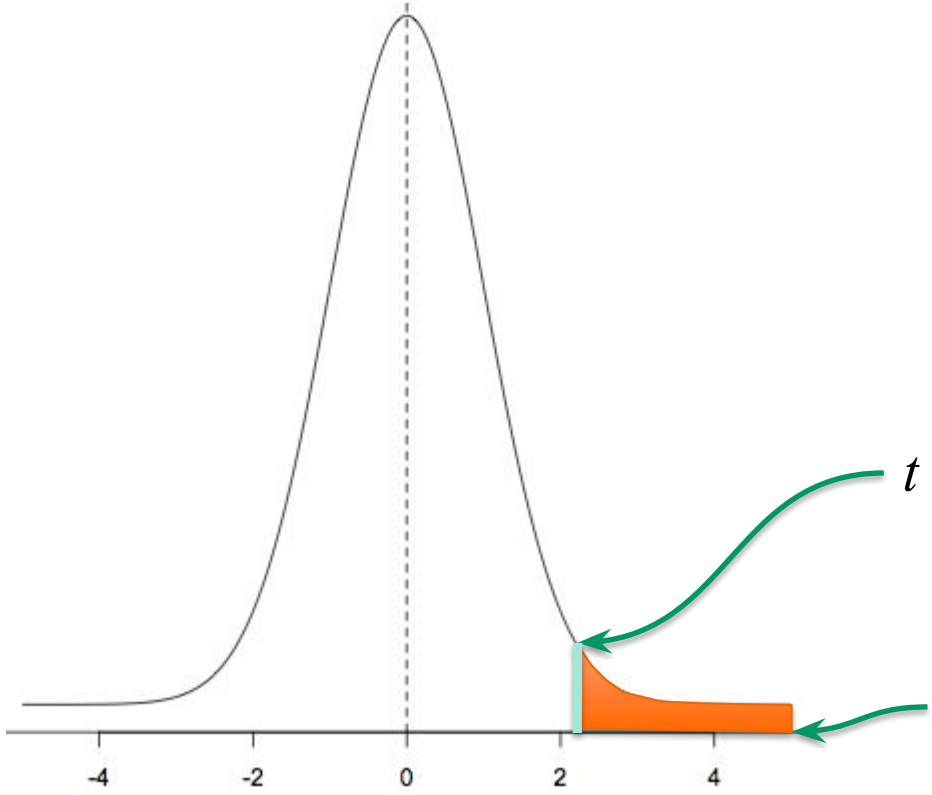
User	A	B	B-A
1	.25	.35	+.10
2	.43	.84	+.41
3	.39	.15	-.24
4	.75	.75	0
5	.43	.68	+.25
6	.15	.85	+.70
7	.20	.80	+.60
8	.52	.50	-.02
9	.49	.58	+.09
10	.50	.75	+.25

$$\hat{\mu} = \overline{B - A} = 0.214$$

$$\hat{\sigma}_{B-A} = 0.291$$

$$t = \frac{\hat{\mu}}{\hat{\sigma}_{B-A}}\sqrt{n} = 2.33$$

Student's t-test for matched pairs



$$\hat{\mu} = \overline{B - A} = 0.214$$

$$\hat{\sigma}_{B-A} = 0.291$$

$$t = \frac{\hat{\mu}}{\hat{\sigma}_{B-A}}\sqrt{n} = 2.33$$

$$p - value = 0.02$$

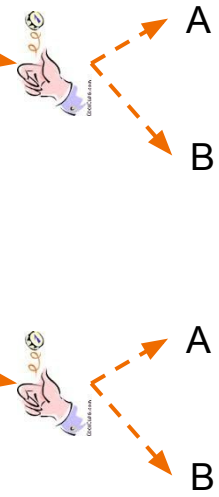
Student's t-test for independent samples

User	A	User	B
1	.25	11	.35
2	.43	12	.84
3	.39	13	.15
4	.75	14	.75
5	.43	15	.68
6	.15	16	.85
7	.20	17	.80
8	.52	18	.50
9	.49	19	.58
10	.50	20	.75

$$t = \frac{\overline{B} - \overline{A}}{\sqrt{\widehat{\sigma}^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Randomization test for independent samples

User	A	User	B
1	.25	11	.35
2	.43	12	.84
3	.39	13	.15
4	.75	14	.75
5	.43	15	.68
6	.15	16	.85
7	.20	17	.80
8	.52	18	.50
9	.49	19	.58
10	.50	20	.75



Randomization test for matched pairs

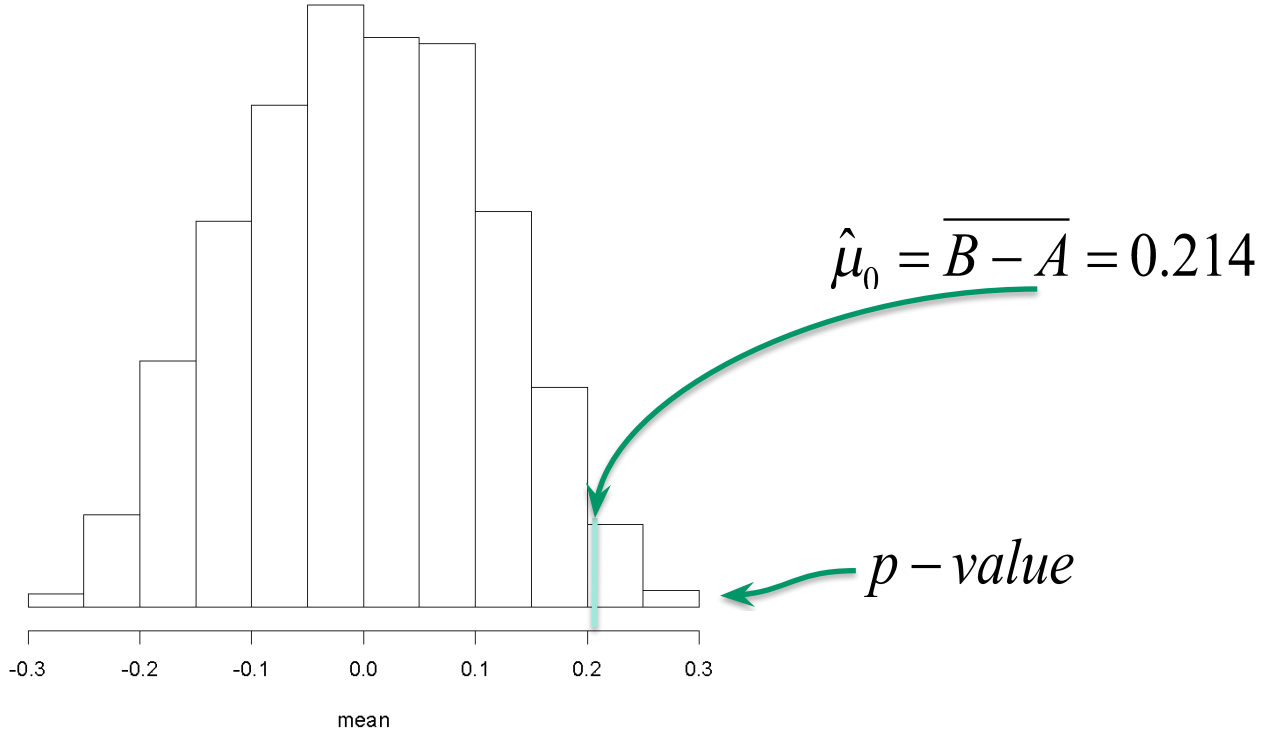
User	A	B	B-A
1	.25	.35	+.10
2	.84	.43	+.41
3	.39	.15	-.24
4	.75	.75	0
5	.68	.43	-.25
6	.15	.85	+.70
7	.80	.20	+.60
8	.50	.52	+.02
9	.58	.49	0.09
10	.75	.50	+.25

$$\hat{\mu}_0 = \overline{B - A} = 0.214$$

$$\hat{\mu}_1 = -0.008$$

$$\hat{\mu}_2 = -0.093$$

Randomization test for matched pairs

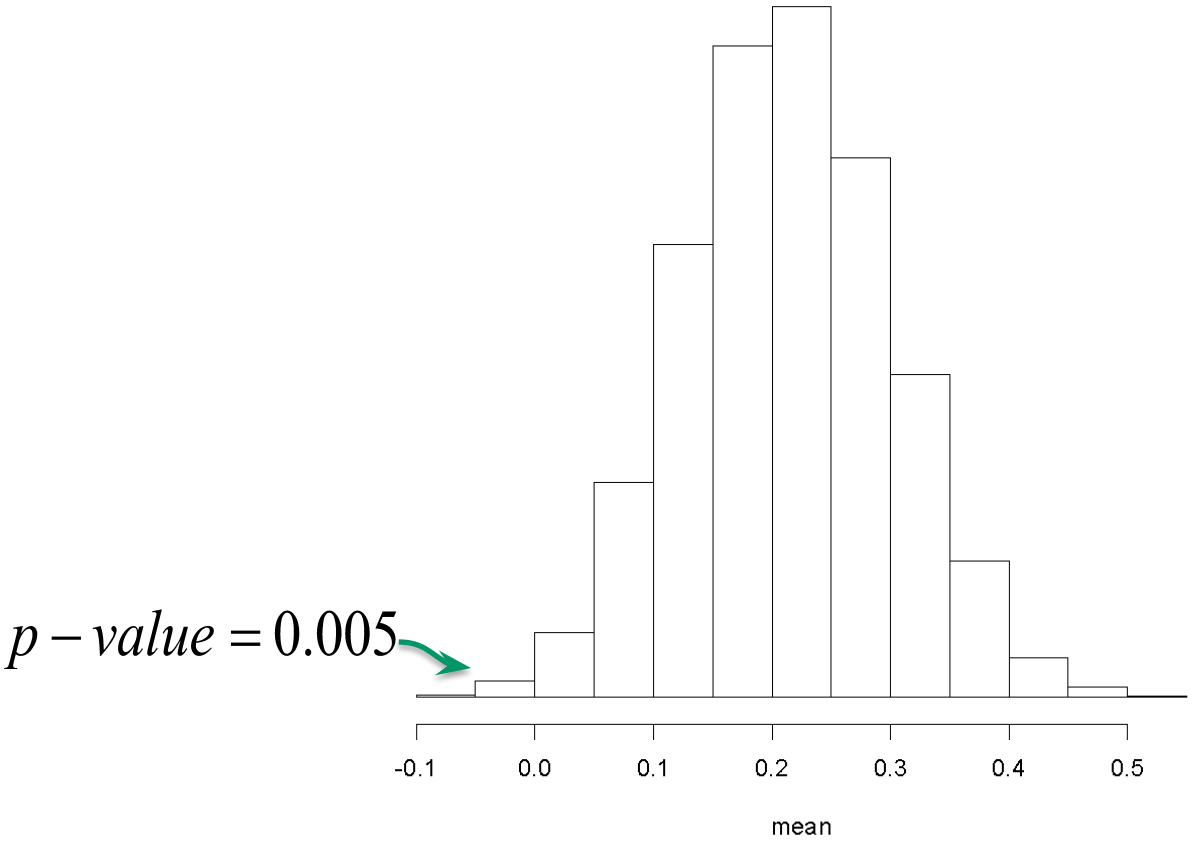


Bootstrap test for matched pairs

User	A	B	B-A
1	.25	.35	+.10
2	.43	.84	+.41
3	.39	.15	-.24
4	.75	.75	0
5	.43	.68	+.25
6	.15	.85	+.70
7	.20	.80	+.60
8	.52	.50	-.02
9	.49	.58	+.09
10	.50	.75	+.25

Sample 1	Sample 2	Sample 3
-.24	+.25	-.24
+.41	+.10	+.60
-.02	+.25	-.70
0	+.60	+.25
+.25	+.70	+.70
+.10	-.02	+.41
+.25	+.10	-.02
+.10	+.25	-.24
+.25	0	+.70
+.10	-.02	+.25

Bootstrap distribution



ANOVA

- Compare variance due to system to variance due to user

User	A	B	B-A
1	.25	.35	+.10
2	.43	.84	+.41
3	.39	.15	-.24
4	.75	.75	0
5	.43	.68	+.25
6	.15	.85	+.70
7	.20	.80	+.60
8	.52	.50	-.02
9	.49	.58	+.09
10	.50	.75	+.25

$$\hat{\sigma}^2 = MSE = 0.042$$

$$\hat{\sigma}_s^{\,2} = MST = 0.229$$

$$F = \frac{MST}{MSE} = 5.41$$

ANOVA

- ANOVA is a generalization of the t-test
- Allows comparison of more than just 2 systems
  - And across more factors than just system and topic
- Commonly used to analyze user studies

Summary

- These are six tests that are easy to apply to recommender system experimentation
  - Many others in the literature: Chi-squared, proportion test, ANCOVA/MANOVA/MANCOVA
- All have in common the use of some probability distribution used to compute a p-value from measurements
  - Non-parametric tests transform data to be modeled with a closed-form distribution
  - Parametric tests estimate hyperparameters from data
  - Empirical tests compute a null distribution from the data itself
- All produce *different* p-values that are still highly correlated
  - Though they don't always agree on which pairs are significantly different

Part 2

FUNDAMENTALS OF SIGNIFICANCE TESTING

What are tests really telling us?

- Formal set-up:

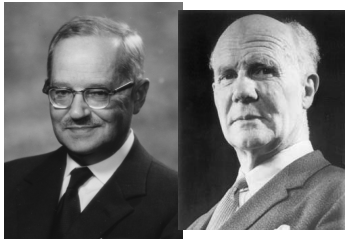
  H_0: $\mu = 0$ or H_0: $\mu \leq 0$

  H_1: $\mu \neq 0$ H_1: $\mu > 0$

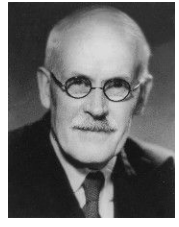
- The null hypothesis, along with the null distribution, is a model
  - The test summarizes evidence against the truth of the model
- A significance test is a procedure that takes data, a null hypothesis, and a procedure for computing a null distribution
  - It outputs a p-value
- The p-value is the probability that you would have seen the same or more extreme result if H_0 were true
  - If that probability is low, we typically “reject” H_0

What are tests really telling us?

- Fisher: p-value is the likelihood of the data under H_0
  - The p-value is a conclusion about this particular experiment only
  - Nothing more, nothing less—do not generalize from a single test



- Neyman-Pearson: $p < 0.05$ means we can reject H_0 as being unlikely to be true
  - p-values lead to inference about the population
  - The p-value itself is not interesting; the inference is
  - Note that we do *not* accept that H_1 is true!

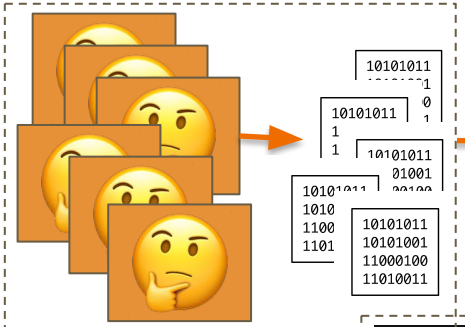


- Bayesian testing: posterior probability of H_0 being true can be compared to that of other models

What are tests NOT telling us?

- NOT the “probability that the results are due to chance”
- NOT whether the experiment is reliable
- NOT the probability that H_0 is true or false
- NOT that H_0 is false if and only if the p-value is sufficiently low
- NOT that our test is better than yours because it produced a lower p-value

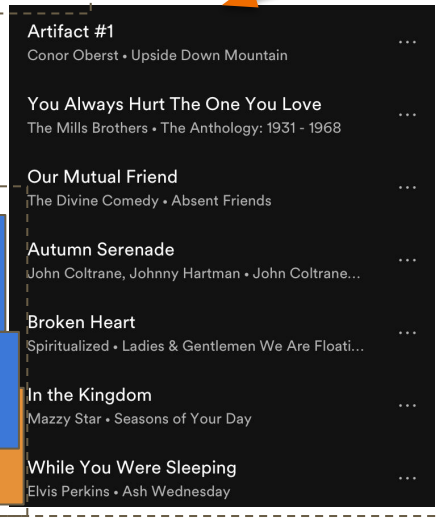
Some basic terms

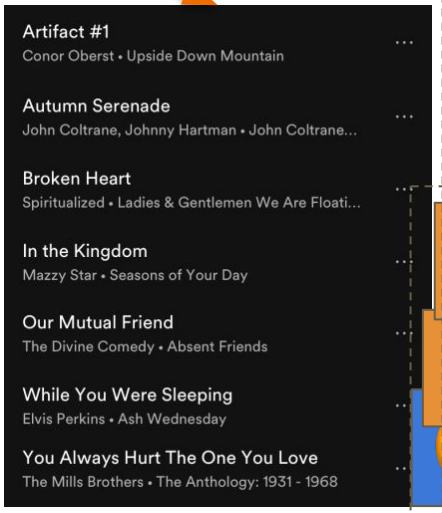


sample

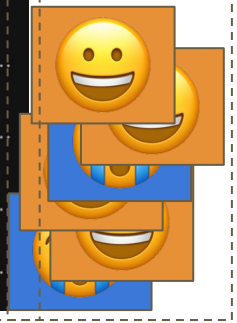
treatments (usually 1 control, 1 or more non-control)

measurements



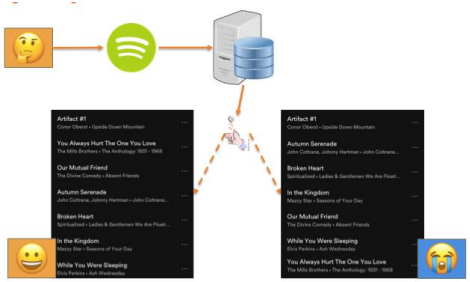


measurements



Terms and definitions

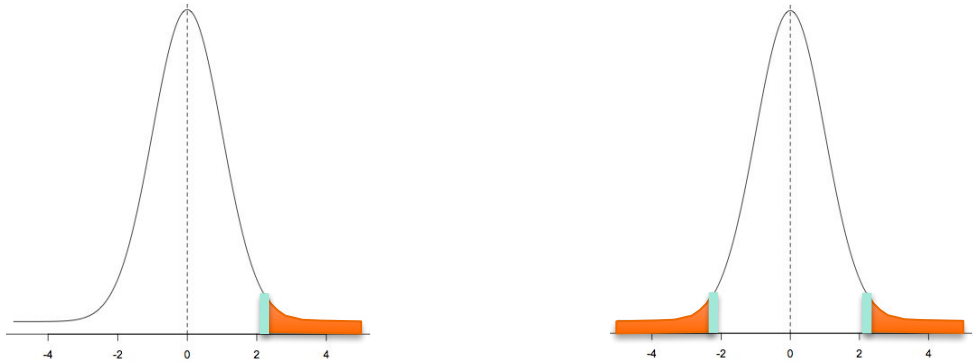
- Single-sample vs two-sample experiments



- Paired vs unpaired
  - Paired tests are a special case of single-sample tests
  - Unpaired tests can be single-sample too

Terms and definitions

- One-tailed vs two-tailed



- **Recommendation:** use one-tailed tests
    - More honest about what your hypothesis is, and a bit less susceptible to bad interpretation

Test statistics and distributions

- Test statistic
  - A summary of the data, usually designed to have specific distribution guarantees (asymptotically)
  - Examples: # of successes, Wilcoxon signed rank, t-statistic
- Parametric vs non-parametric
  - If the test statistic has any hyperparameters, the test is said to be “parametric”
- Confidence interval

Sizes and values

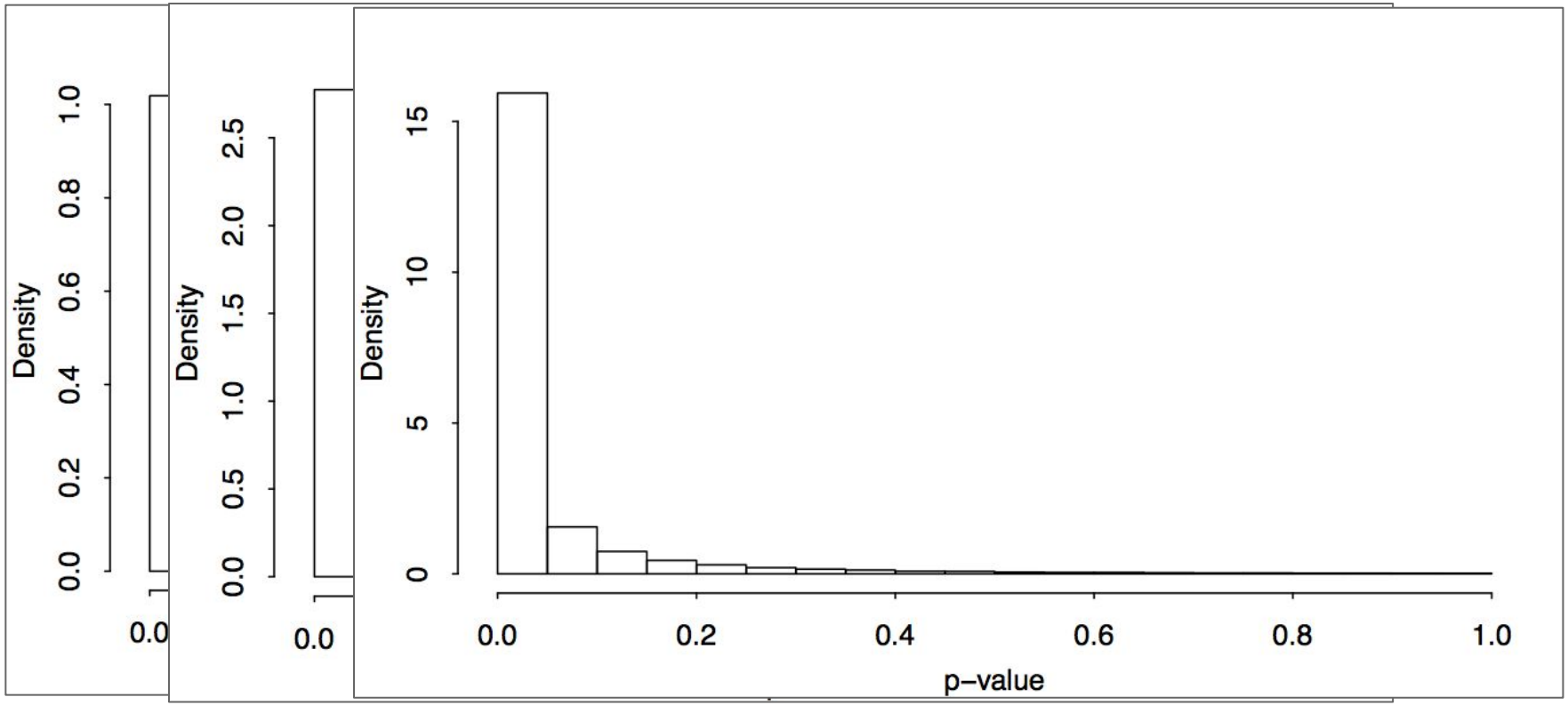
- Sample size
  - The number of users/examples sampled for the experiment
  - Often assumed to be sampled i.i.d. from a larger population
- Effect size
  - A measure of the difference between two treatments over the population
  - Independent of sample size
  - Null hypothesis is often equivalent to no effect
- P-value
  - The likelihood of observing the effect in the sample, assuming the null hypothesis is true
- Critical value
  - The minimum value of the test statistic needed to achieve $p < \alpha$

- **Recommendation**: report estimated effect size and confidence interval, not just the p-value or whether $p < \alpha$

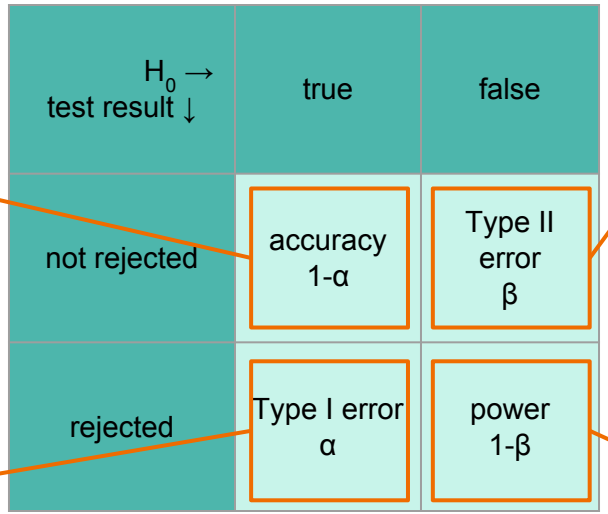
P-value distributions

- Variance due to subjects and treatments means variation in p-values over experiments
  - Therefore we can talk about the probability of observing a certain p-value conditional on an experiment
- When the null hypothesis is true, the p-value has a uniform distribution
  - All values equally likely
  - $P(p < 0.05 \mid H_0 \text{ true}) = 0.05$
- When the null is not true, the p-value distribution depends on the population effect size and sample size

Example p-value distributions



Accuracy and power



The probability of finding $p > \alpha$ when H_0 is true: true negative rate

H_0 is false but $p > \alpha$
False negative rate

H_0 is true but $p < \alpha$
False positive rate

The probability of finding $p < \alpha$ when H_0 is false: true positive rate

Most tests are designed so that the false positive rate is exactly equal to α.
Recommendation: think about power *before* your experiment

Statistical testing as classification

- Contingency tables? False positives and negatives? Looks familiar…
- **A statistical test with a threshold for significance is a binary classifier**
- Classifiers learn a model of the data: class modeled as a function of features
    - Of course, unlike classifiers, we cannot evaluate statistical tests directly—there is no ground truth
- Statistical tests implicitly model evaluation data as a function of features, then inference in that model

Modeling metrics for significance testing

- ANOVA is based on the linear regression model

$$y_{ij} = \mu + \alpha_j + \beta_i + \varepsilon_{ij}$$

- y_{ij} is the measurement of effectiveness of system i for user j
- μ is the intercept and represents baseline effectiveness
- α_j represents the “user effect”
- β_i represents the “system effect”
  - Different meaning from the α and β in Type I and Type II error rates
- ε_{ij} is random error
  - It represents every effect unspecified in the model

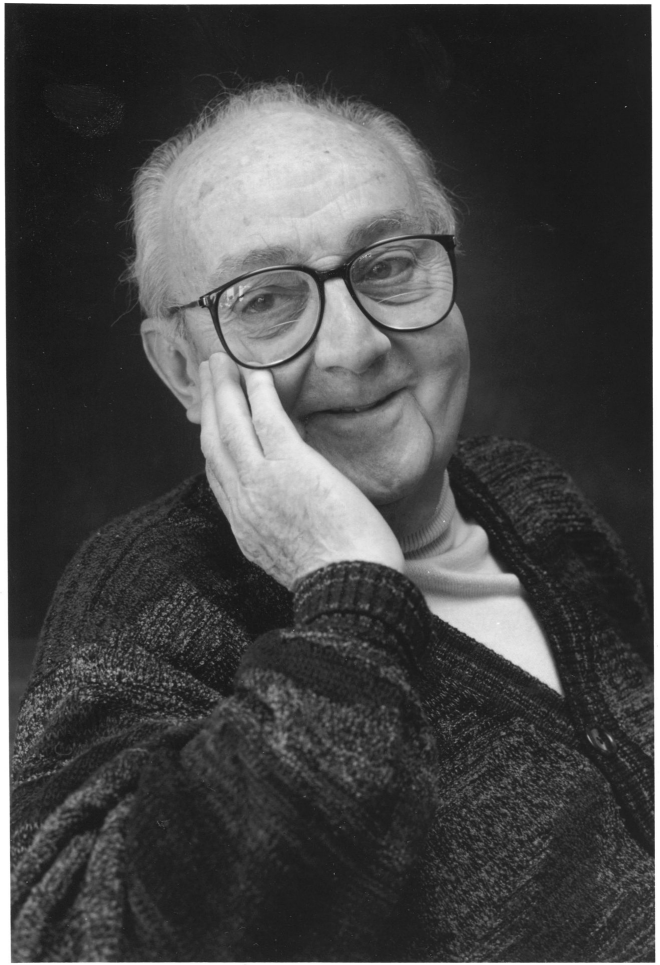
Modeling metrics for significance testing

$$y_{ij} = \mu + \alpha_j + \beta_i + \varepsilon_{ij}$$

- Fit the model using OLS
- Compare β values

- OLS estimator for β_i is mean effectiveness of system *i*
- ε_{ij} is assumed to have normal distribution with variance σ^2

- Inference about significance uses those two quantities, ignores everything else
  - But it's all still there, affecting the model



George E. P. Box

All models are wrong, but some are useful.

Summary

- There is disagreement about how to use and understand statistical significance!

- Recommendations:
  - Use one-tailed tests
  - Report effect sizes and confidence intervals, not just p-values or significance
  - Think about power *before* your experiment (coming up soon)
  - Remember that every test is wrong

Part 3

APPLICATIONS

What is a statistical significance test?

- A statistical test consists of four things:
  - A null hypothesis
  - A test statistic
  - A null distribution for the test statistic
  - A critical value in the null distribution
- You can invent any test you like!
  - ... as long as you can compute a test statistic and its null distribution

Example 1: Multiple comparisons

- Recommender system experimentation often happens like this:
  1. Modify a system, compare to baseline, run test
  2. Significant?
     - No: go back to step 1
     - Yes: deploy system / start writing a paper

- How many tests does it take to get to the endpoint?
  - P(m^{th} experiment gives significant result | m experiments lacking power to reject H_0)
  - P(at least one significant result | m experiments lacking power to reject H_0)

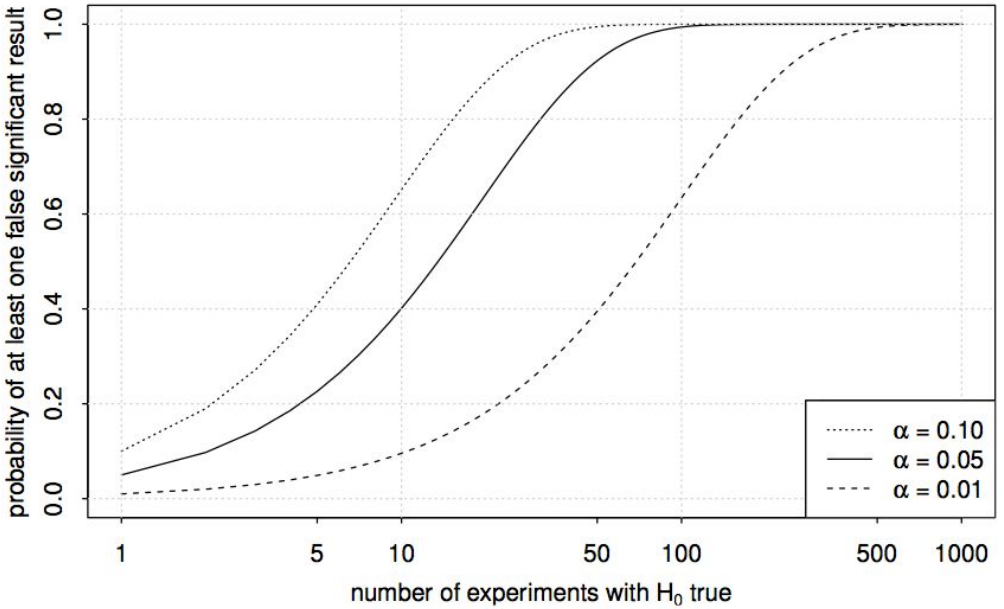
Multiple Comparisons Problem

- P(at least one significant result | m experiments lacking power to reject H_0)

  = P(one significant | m) + P(two significant | m) + …

  = 1 – P(none significant | m)

  = $1 - (1 - \alpha)^m$



Multiple comparisons correction

- Adjust our test results for the fact that we have made multiple comparisons
- Many different approaches in stats literature:
  - Bonferroni correction
  - Tukey's Honest Significant Differences
  - Multivariate t test
- Instead of picking one of those, let's reason from principles

Setting up a test

- Start by setting up a single null hypothesis that all systems are equal on average:
  - H_0: $S_1 = S_2 = S_3 = \ldots = S_m$
  - This is called the omnibus hypothesis
- Define a test statistic
  - Maximum difference between any two systems, assuming all are equal on average
- How do we compute a null distribution?
  - Even if all systems are equally effective, random variation will mean that we can order them by average effectiveness on a sample
  - What is the expected maximum difference between any two systems over a sample given that all are equally effective?
  - Compute the null distribution from there

Using randomization to find the null distribution

user	S1	S2	S3	S4	S5	S6
1	0.44	0.64	0.60	0.39	0.11	0.64
2	0.26	0.38	0.47	0.30	0.56	0.61
3	0.01	0.41	0.54	0.65	0.19	0.37
4	0.30	0.29	0.49	0.27	0.57	0.69
mean	0.25	0.43	0.53	0.40	0.36	0.58

Original data: maximum difference = 0.58 - 0.25 = 0.33

Using randomization to find the null distribution

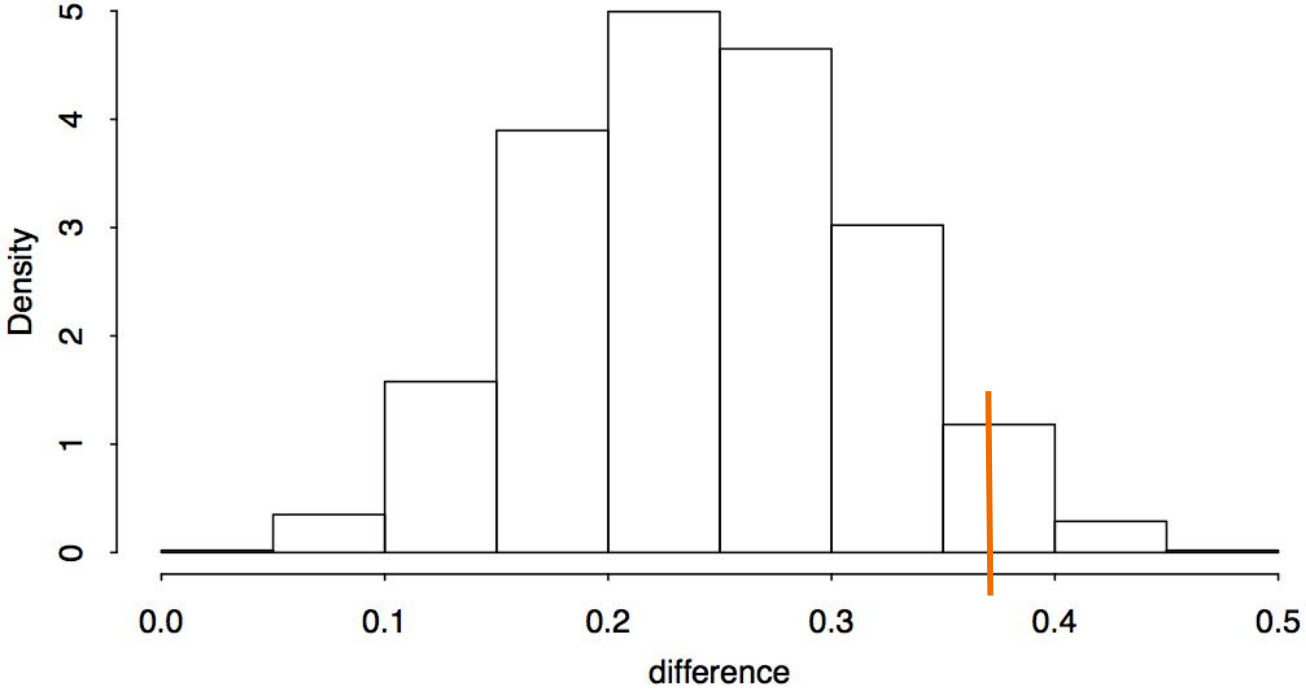
user	S1	S2	S3	S4	S5	S6
1	0.39	0.64	0.11	0.64	0.60	0.44
2	0.26	0.38	0.47	0.30	0.56	0.61
3	0.01	0.41	0.54	0.65	0.19	0.37
4	0.30	0.29	0.49	0.27	0.57	0.69
mean	--	--	--	--	--	--

Using randomization to find the null distribution

user	S1	S2	S3	S4	S5	S6
1	0.39	0.64	0.11	0.64	0.60	0.44
2	0.56	0.61	0.30	0.26	0.47	0.38
3	0.19	0.65	0.54	0.01	0.41	0.37
4	0.69	0.27	0.57	0.29	0.49	0.30
mean	0.41	0.54	0.31	0.39	0.44	0.51

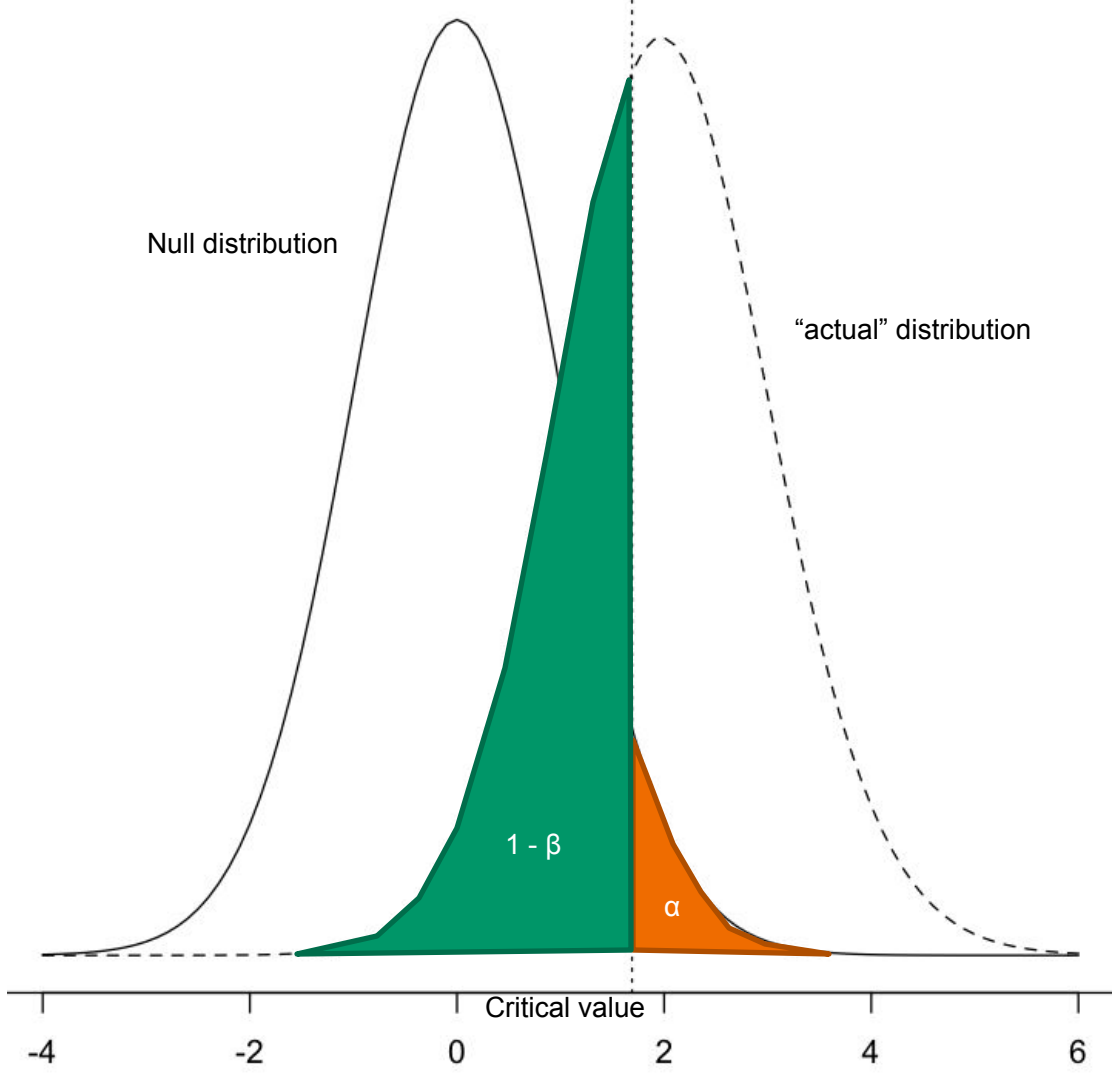
Trial 1: maximum difference = 0.54 - 0.31 = 0.23

Distribution of maximum difference



Example 2: Power Analysis

- High statistical power is desirable
  - $1-\beta \approx 0.8$ is generally considered good power
    - 80% chance of rejecting H_0 when it is false
- But you don't know statistical power *a priori*
  - Unlike accuracy, power cannot be guaranteed
- *Power analysis* is a technique to estimate the necessary sample size to achieve high power


Null distribution
“actual” distribution
1 - β
α
Critical value
-4
-2
0
2
4
6

Power Analysis Without Pain

- "I want to be able to detect an effect of size 0.1 or higher with 80% probability"
  - In other words, I want to be able to reject H_0 even when the effect is relatively small
- Steps (using t-test as example):
  - Pick a value of n
  - Let c_α be the critical value for sample size n
  - Let $t = 0.1 \times \sqrt{n}$
  - Compute $\beta = P(T \geq c_\alpha \mid n, t)$
- Search for the smallest value of n that results in $\beta \geq 0.8$

Power analysis without pain

- “I want to be able to detect an effect of size 0.1 or higher with 80% probability”
  - In other words, I want to be able to reject H_0 even when the effect is relatively small

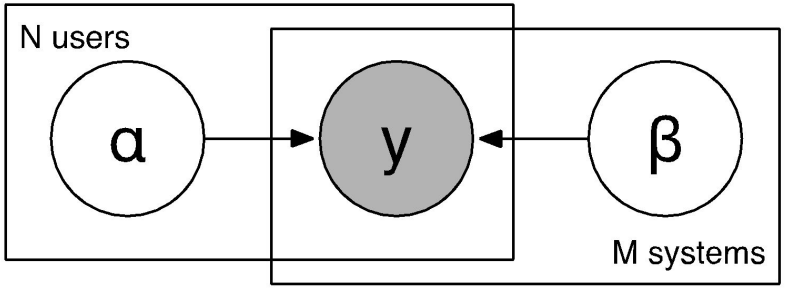
- Steps (using t-test as example):
  - Pick a value of n
  - Let c_α be the critical value for sample size n
  - Let $t = 0.1 \times \sqrt{n}$
  - Compute $\beta = P(T \geq c_\alpha \mid n, t)$

- Search for the smallest value of n that results in $\beta \geq 0.8$

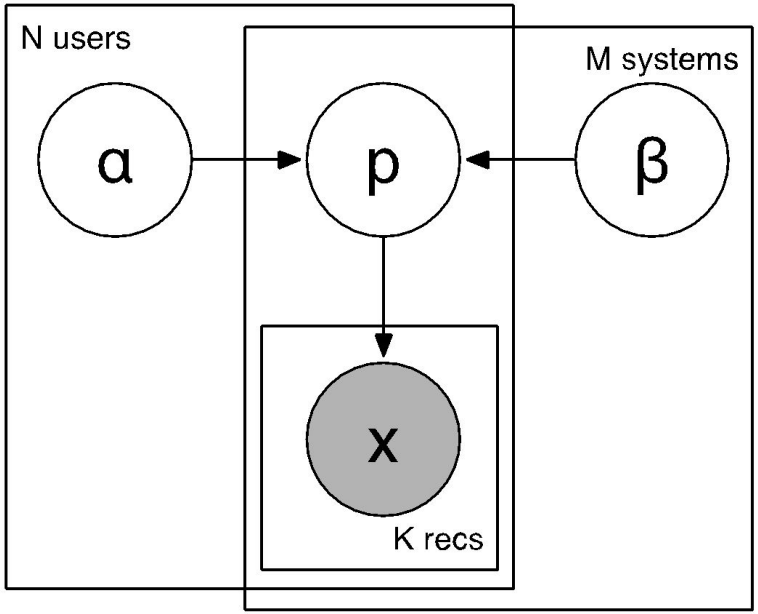
Example 3: Item-Level Testing

- Testing for differences in metrics averaged over a list/set of items means causality can only be attributed in aggregate
  - Cannot necessarily say why it failed for a specific user
- De-aggregate, test for differences at the item level instead

ANOVA model



Hierarchical model



Item-level model

$$x_{ijk} \sim Bernoulli(p_{ij})$$

$$\text{logit } p_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$$

$$\epsilon_{ij} \sim N(0, \sigma^2)$$

- Likelihood function: $L = \prod_{i=1}^{N} \prod_{j=1}^{M} \prod_{k=1}^{K} P(x_{ijk})$

$$L_{M_3} = \prod_{i=1}^{N} \prod_{j=1}^{M} \prod_{k=1}^{K} p_{ij}^{x_{ijk}} (1 - p_{ij})^{1 - x_{ijk}}$$

Takeaways

- Always do significance tests
  - But don’t worry too much about which tests to use
  - The t-test is always a good option
- Don’t just report p-values or * to indicate significance
  - Always report estimated effect sizes and confidence intervals
- Always take results of tests with a grain of salt
  - Especially when the effect size is low
  - Don’t expect them to generalize a priori
  - Build your intuition and use it
- Learn how to design tests to get at the questions that are really interesting