Jhelum Chakravorty

Published Sep 23
Fork of Introduction to Inference by Kris Sankaran

Introduction to Inference

IFT6758, Fall 2020

Reading: ISLR 3.1.1, 3.1.2 and Bayesian Basics (intro - regression models)

Optional reading: MSMB 6.1 - 6.6, Statistics for Hackers and CASI Chapter 3

Inference

- Meta-Algorithms: Evaluation of processes people use to learn from data
- Science: How to go from particular to general cases?

"It is easy to lie with statistics, but hard to tell the truth without them."

Hypothetical Reasoning

- We've mostly been thinking about recognizing patterns occuring in observed data
- But *hypothetical reasoning* is important. How could the world look intead?
- Statistical inference provides quantitative machinery to help hypothetical reasoning

Hypothesis is our assumption in the real world. we do the obersvation and you can either accept or reject based on your findings.

Topics to be covered

- Inference for coin flips
  - Both mathematical and computational approaches
- Inference in linear regression
- Revisit problems from Bayesian perspective

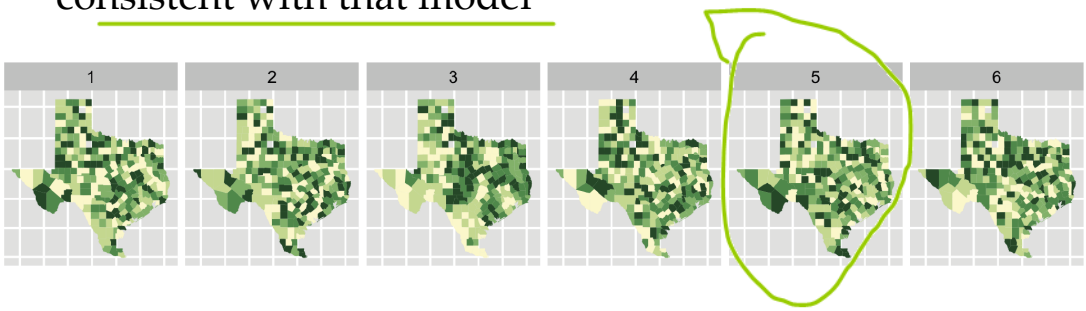
Is your coin fair?

Hypothesis testing gives a formal framework for answering questions like,

- How will you measure discrepancies?
- How will you tell if it's meaningfully large?

Model of Reality aka. hypothesis

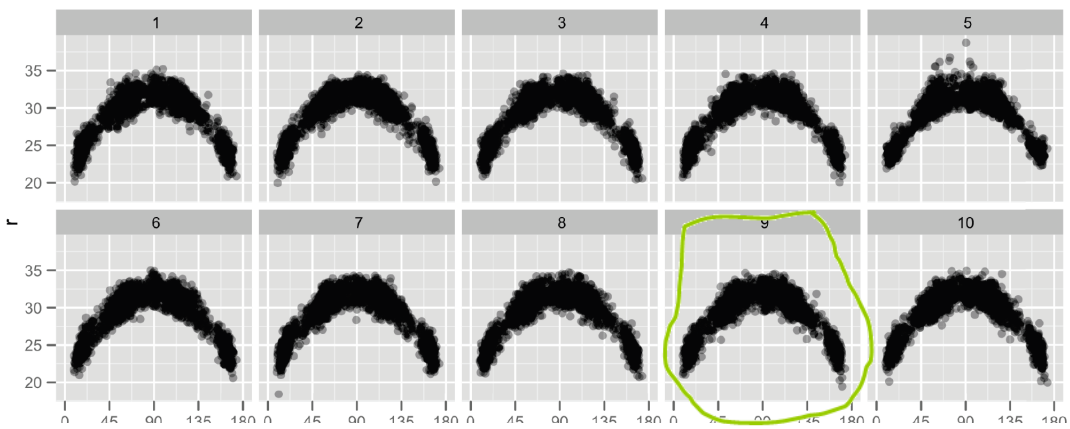
- A null model defines a default simulator of reality
- Our goal is to determine whether an observed dataset is consistent with that model



Counts of cancer deaths. 5 come from null model that there is no spatial correlation.

Model of Reality

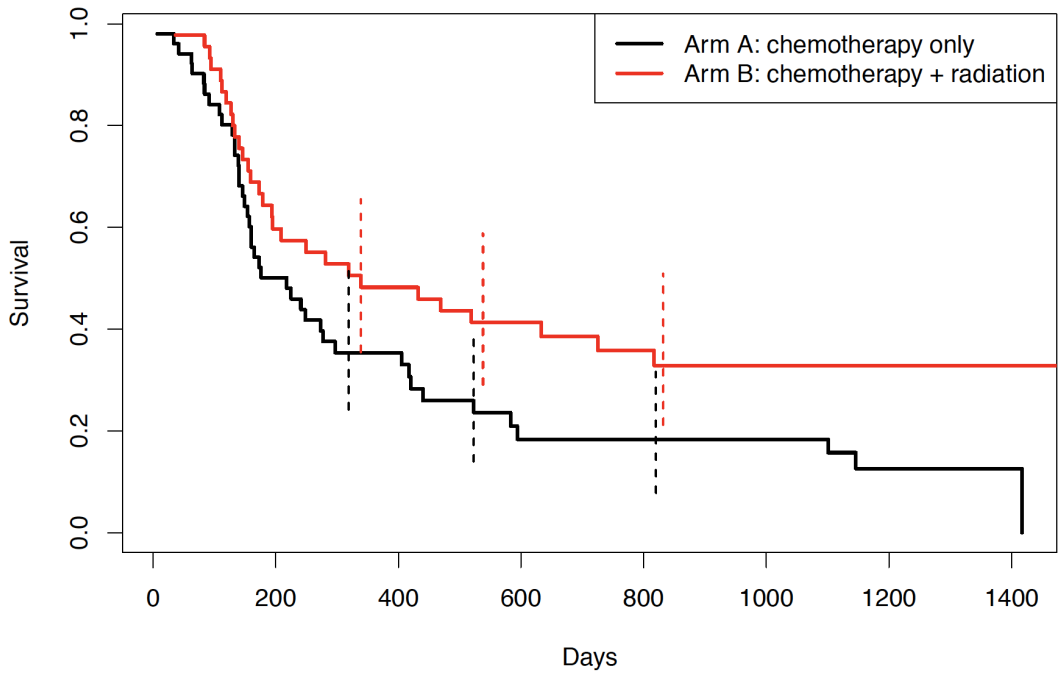
- A null model defines a default simulator of reality
- Our goal is to determine whether an observed dataset is consistent with that model



Distance vs. angle in three pointers from LA Lakers. 9 are from null model that there is a quadratic relationship.

Measuring Discrepancies

- We'll look in the data for potential discrepancies, and try to judge whether they are meaningfully large



Are these two treatments quite different?

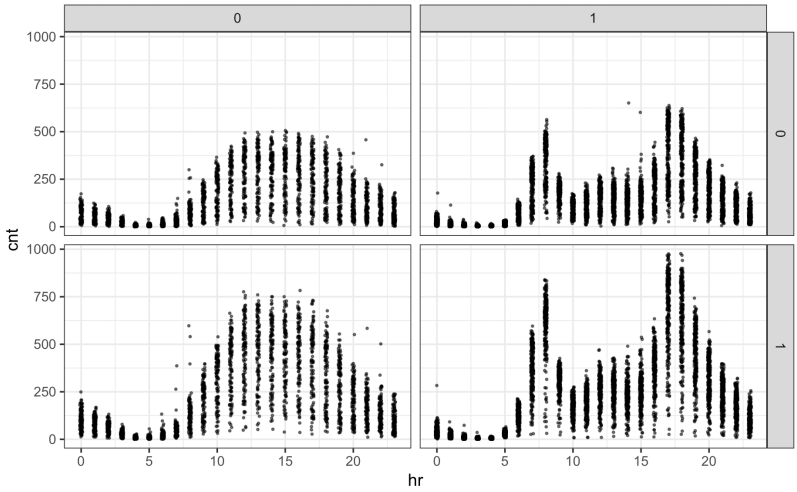
If you didn't have the error bars, you might think *yes*.

Challenges

- All theories are approximations to reality (null

hypothesis)

- If your approximation isn't that believable, you will reject the null.
- One should be careful *not to make any error in rejecting or accepting such approximation*.
- Most critical assumption: Independence
  - Thinking you have more evidence than you do is usually much worse than small distributional differences



If you had sampled twice as frequently, will you really have doubled your sample size?

Measuring Discrepancies

- The goal is to identify a statistic (any function of the data) that detects problems in your simulation of reality
- For coin tossing, it makes sense to use

$$\hat{p}_n(X_1, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$$
$$= \text{fraction heads}$$

and sound an alarm if $\hat{p}_n$ is very far from 0.5

then the coin is not fair

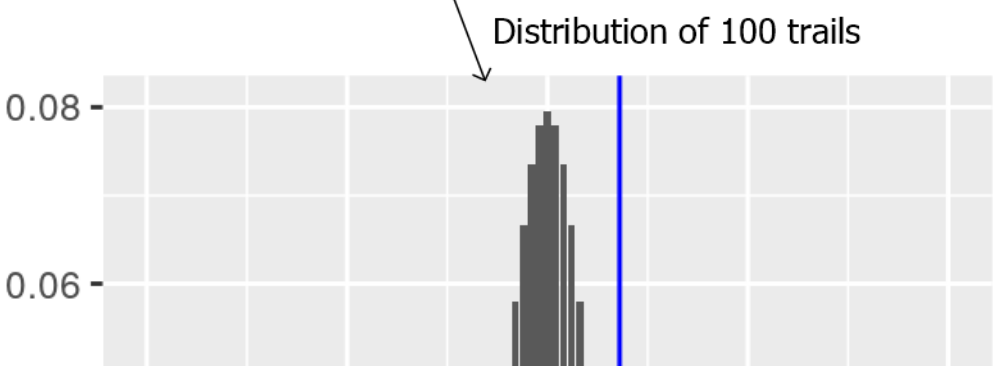
- What if you saw HHHHHHHHHHTTTTTTTTTT?
- Seems fair according to our metric....
- Different choices of test statistic are sensitive to different departures from your theory

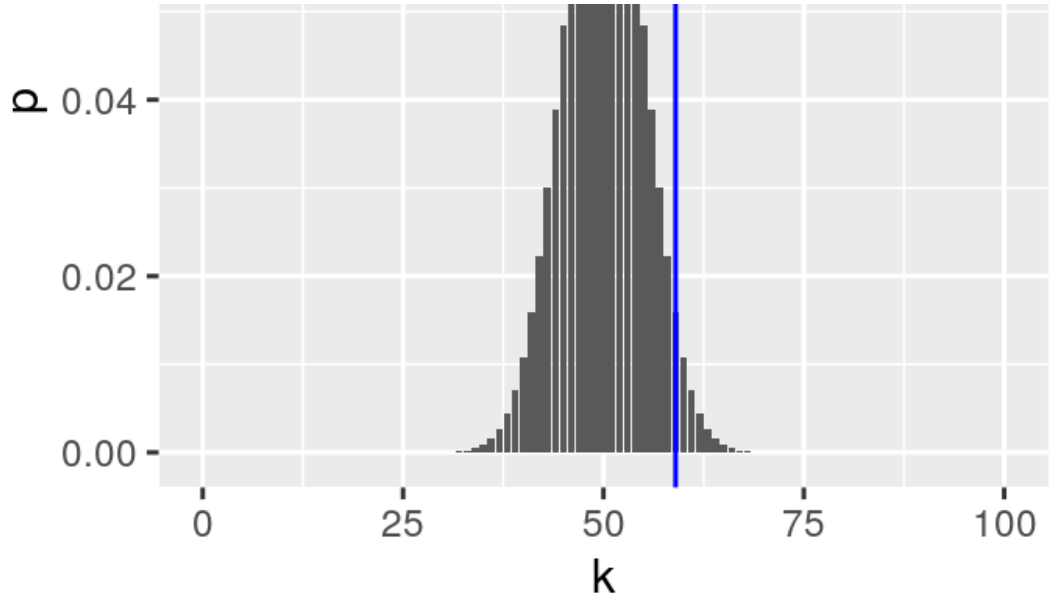
Reference Distribution: Theory

EG: what is the prob. that in 100 trails(n) we get 70 heads, can be calculated

- Say you observed 59 heads in 100 flips.
- How will you tell whether the measured discrepancy is meaningfully large?
- For coin tossing, supposing a binomial simulator

$$\mathrm{P}\left[\hat{p}_n = \frac{k}{n}\right] = \binom{n}{k} p^k (1-p)^{n-k}, \quad n = 100,\ p = 0.5$$





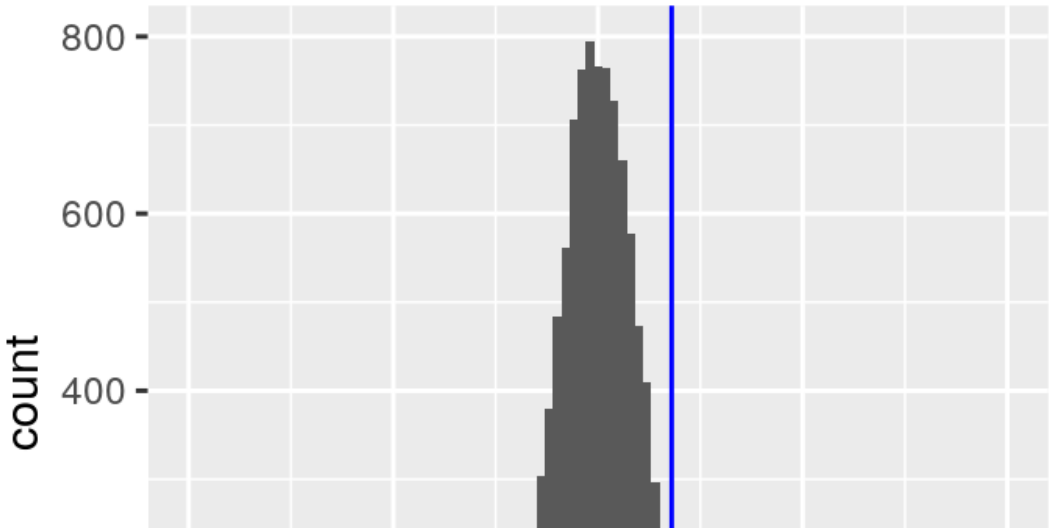
Blue line: number of heads apearing in 100 flips.

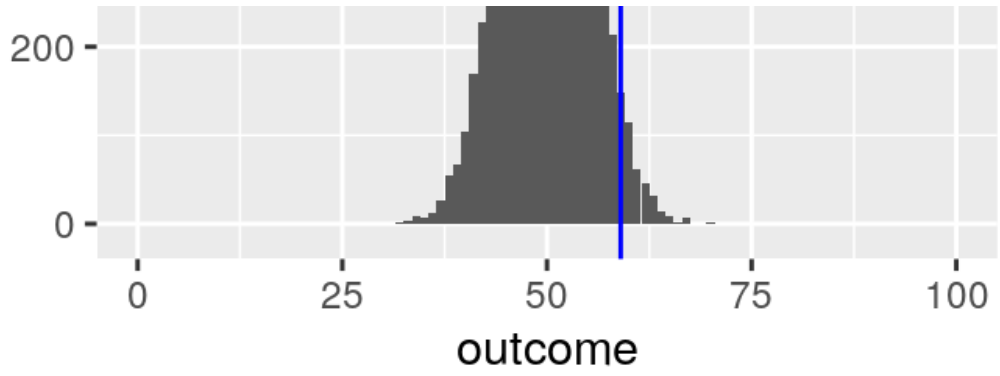
This image (and others) from MSMB.

Reference Distribution: Simulation

Rather than using the closed form equ(above), we can simulate it. The results are similar

- Alternatively, we can directly simulate from our model of the world
- Computing test statistic on simulated data provides a reference distribution
  - Can *always* be done when there's a generative model
  - Especially useful when no closed form analysis





Same reference distribution, from 10^4 simulations

Sample with replacement 100 times and repeat 10^4 times

Rejection Region

Some definitions

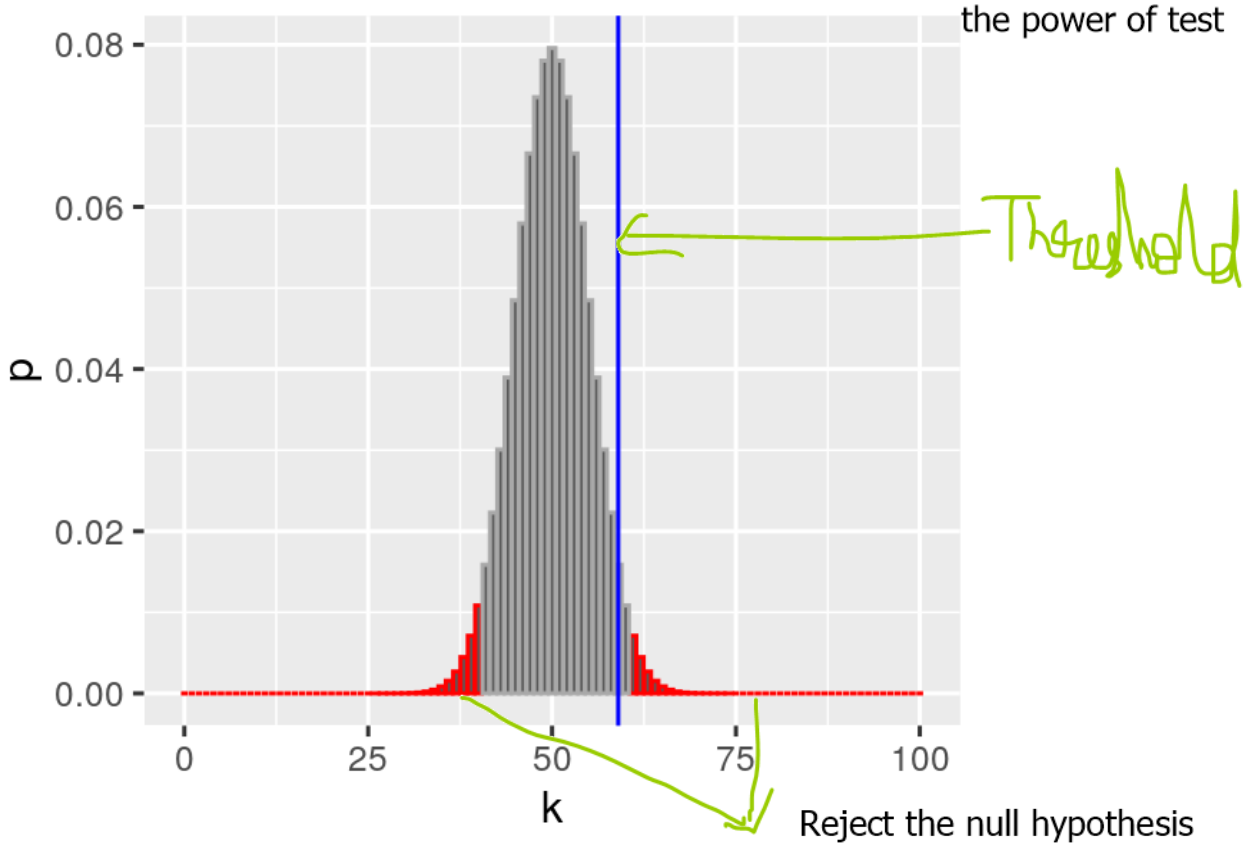
- *p*-value: the probability (i.e., the area in the reference distribution) of obtaining results at least as extreme as the observed results of a statistical hypothesis test, assuming that the null hypothesis is correct.
    - A smaller *p*-value means that there is stronger evidence in favor of the alternative hypothesis.
- Signifance level α: it defines the strength of evidence (i.e., significance) in probabilistic terms. α represents the probability that tests will produce statistically significant results when the null hypothesis is correct.
    - e.g.: $\alpha = 0.05$ means your analysis has a 5% chance of producing a significant result when the null hypothesis is correct.
- Type-I error: Rejecting a true null hypothesis. α equals the Type-I error rate.
    - Interpretation: You can think of this error rate as the *probability of a false positive*. The test results lead you to

believe that an effect exists when it actually does not exist.

Type 2 error is costly

- Type-II error: Failing to reject a false null hypothesis (i.e., when the alternative hypothesis is true).
    - Interpretation: You can think of this error rate as the *probability of a false negative*.
- Power (of a test): it is the probability that we make the right decision when the null is not correct (i.e. we correctly reject it). Power = 1- Prob(Type-II error).

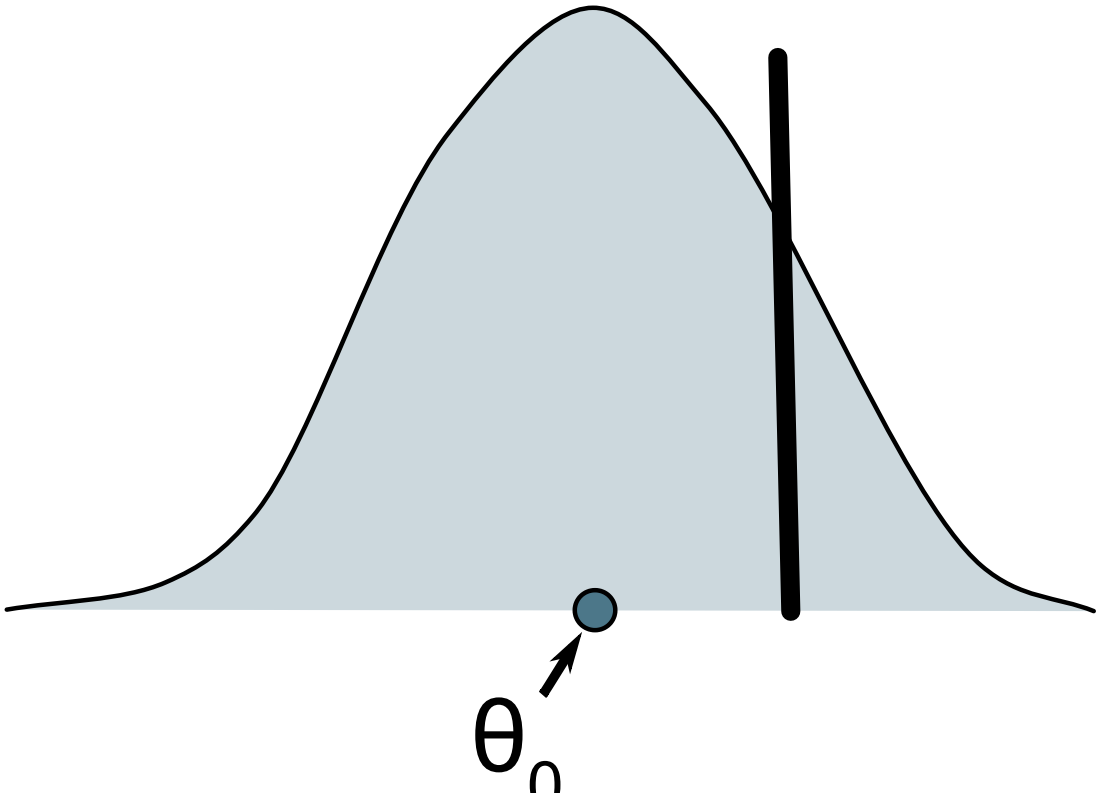
Lower the type 2 error, then higher the power of test



By rejecting in the tails (red regions, each area equals to α), we can reject as many types of outcomes as possible while making sure we rarely reject when the null ($p = 0.5$) is actually true.

Confidence Intervals

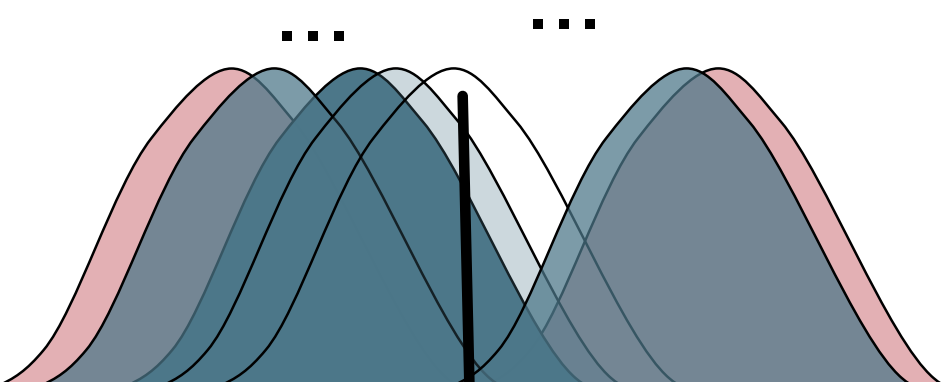
- For your dataset, what are all the possible null hypotheses that you wouldn't have been able to reject?
- This is richer information than just the test outcome

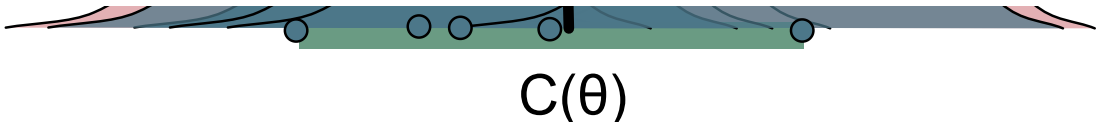


We wouln't have rejected this θ_0 because the test statistic lies in the bulk of the reference distribution.

Confidence Intervals

- For your dataset, what are all the possible null hypotheses that you wouldn't have been able to reject?
- This is richer information than just the test outcome





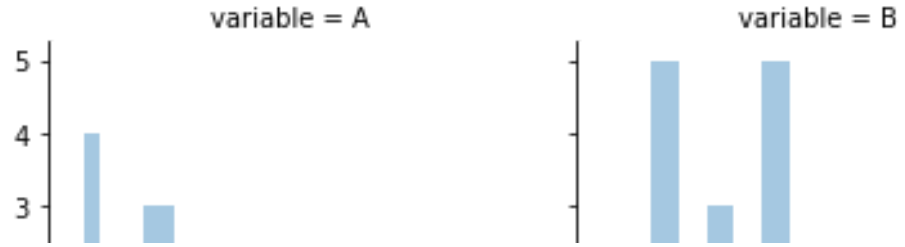
We start rejecting once we get past the red reference distributions. All the θ's between these belong to our confidence interval.

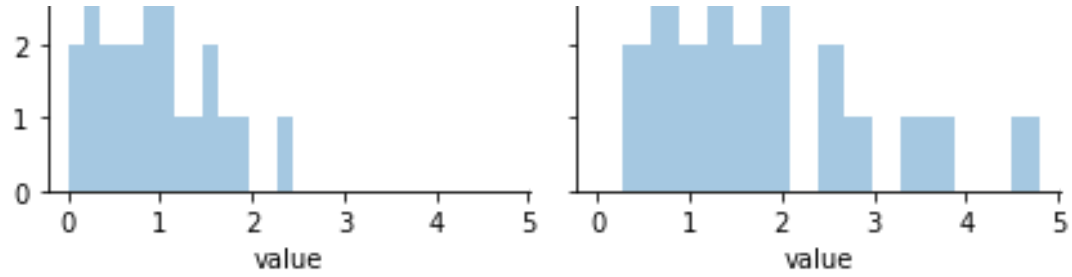
Recap: Overall Approach

- Determine the effect of interest, and design a suitable test statistic
- Define a null hypothesis, and come up with a null distribution for that statistic
- Define the rejection region
- Do the experiment and draw conclusions

Difference in Means

- Common situation: Have two distributions, and we want to see whether they have the same mean
- We don't have to assume they are normally distributed





An example.

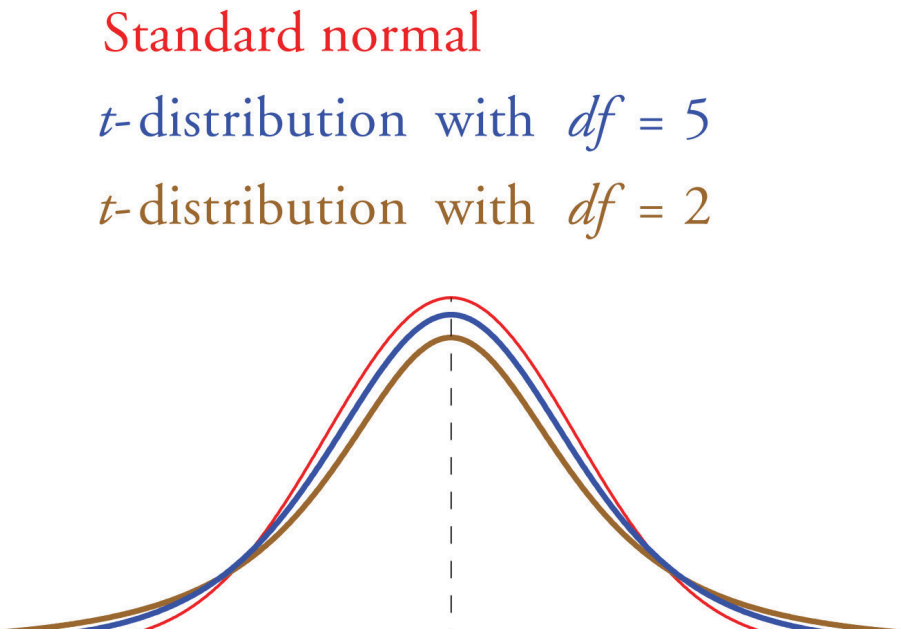
Measuring Discrepancy

- Common to use a normalized difference, e.g.

$$\hat{t}_n(X_1, \ldots, X_n) = \frac{\bar{x}_1 - \bar{x}_2}{\widehat{\text{s.e.}}\,(\bar{x}_1 - \bar{x}_2)}$$

  where $\widehat{\text{s.e.}}$ is an estimate of the standard error (standard deviation of sample mean)
- If the means in the groups are the same, this is approximately *t*-distributed

***t*-distribution**: similar to the normal distribution with its bell shape but has heavier tails. The shape depends on *degrees of freedom*, $df = n - 1$.



0

Check out mathematical expressions.

Measuring Discrepancy

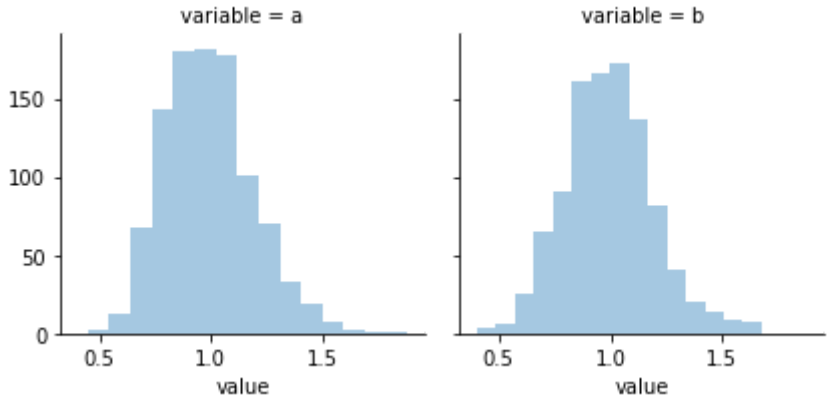
These are the distributions for $\bar{x}_1$ and $\bar{x}_2$. Even though the original data are relatively far from gaussian, the central limit theorem kicks in.

Central limit theorem

It states that sampling distribution of the sample means approaches a normal distribution as the sample size gets larger — *no matter what the shape of the population distribution*

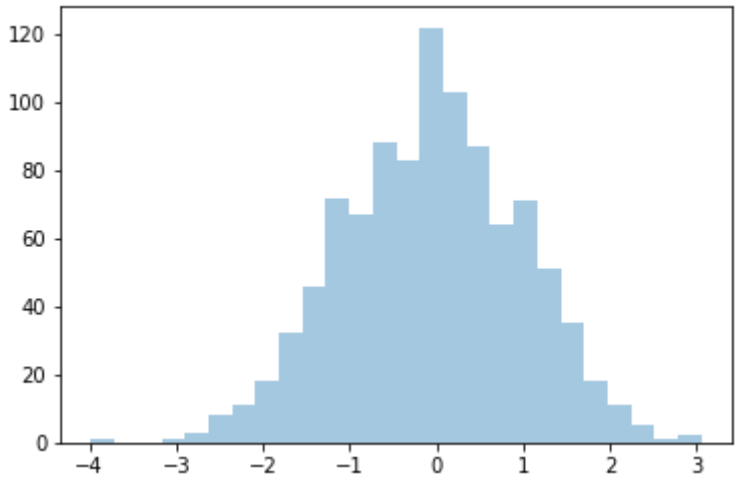
Mathematical statement

Let $\bar{X}_n$ be the *mean* of n random samples, each of size n (i.e., $\bar{X}_n$ is the *sample mean*). The population has mean μ and finite variance σ^2. Then the variable $Z = (\bar{X}_n - \mu)/(\sigma/\sqrt{n})$ follows *standard normal distribution* as $n \to \infty$.



Measuring Discrepancy

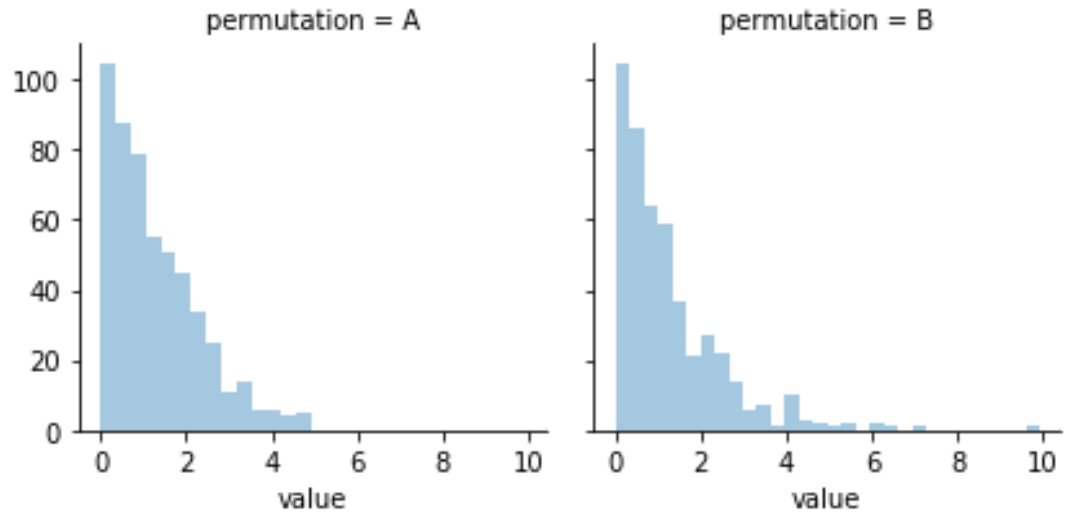
The resulting normalized difference is about *t*-distributed.



Evaluating Discrepancy

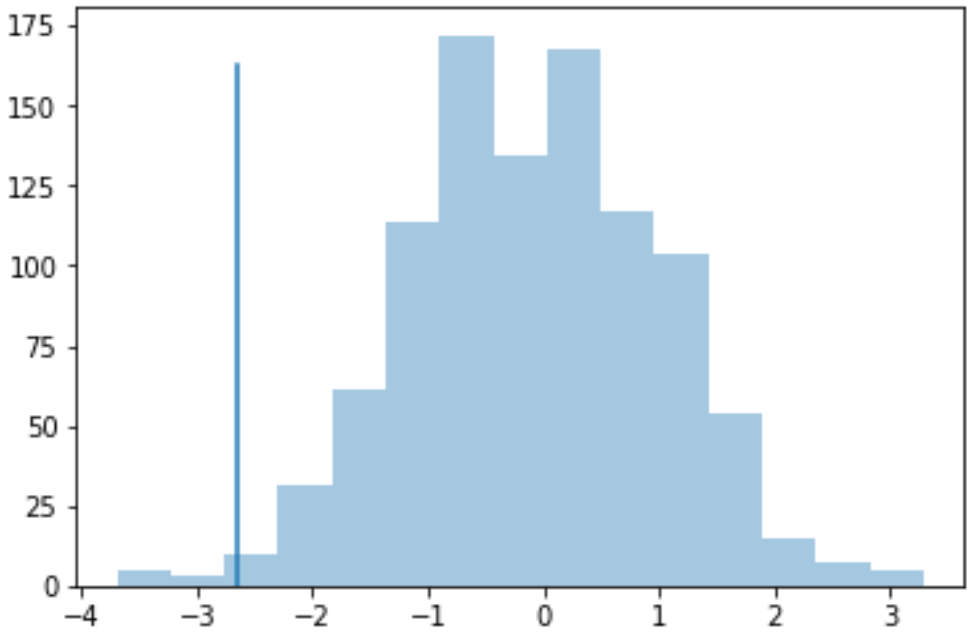
- You don't have to use theoretical reference distribution
- Permutation: Randomly reassign group labels and recompute the statistic

The separate histograms (one per group) become indistinguishable -- it's a simulation from the null for the raw data. Computing difference in means gives a reference distribution.



Evaluating Discrepancy

- You don't have to use theoretical reference distribution
- Simulation: Randomly reassign group labels and recompute the statistic



Linear Regression

Analogies

- $p \longleftrightarrow$ use $\bar{x}$ in binomial model
- $\mu_1 - \mu_2 \longleftrightarrow$ use $\hat{t}$ and central limit theorem
- $\beta \longleftrightarrow$ use $\hat{\beta}$ in linear regression with iid gaussian errors

Generally: Parameter of interest $\longleftrightarrow$ reference distribution of a statistic under a model

Types of Questions

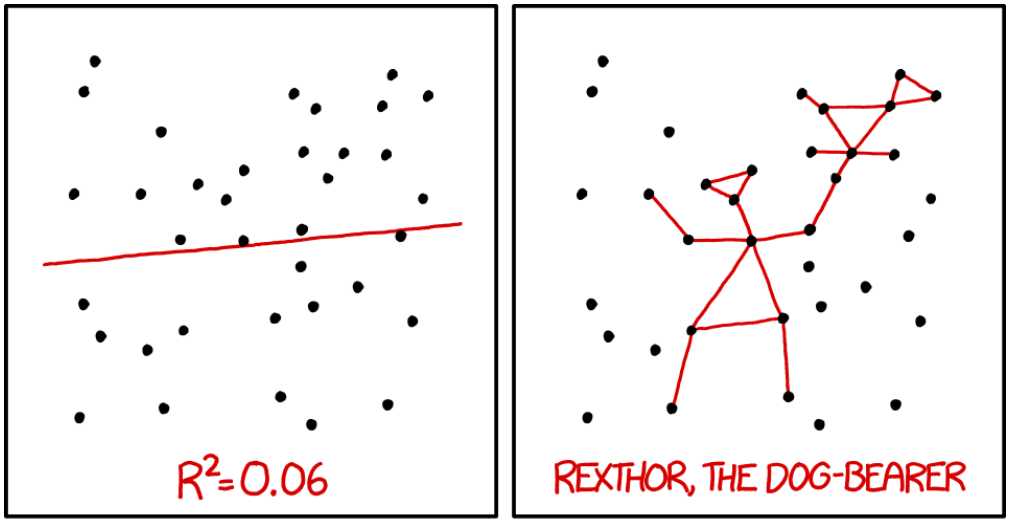
- Association: Is there are relationship between x and y?
- Model comparison: Is a model using additional features actually better?

Testing association [one predictor]

Default to there being no association between x and y. Formally,

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

How should we measure the discrepancy?



Discrepancy

Measuring: Continuous version of the earlier two-groups statistic,

$$\hat{t} = \frac{\hat{\beta}_1}{\widehat{\text{s.e}}(\hat{\beta}_1)}$$

Evaluation: $\hat{t}$ has a t-distribution if you assume i.i.d.

gaussian errors $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$. We'll see some simulation-based alternatives in later lectures.

Comment: Experimental Design

The formula for the standard error is illuminating,

$$\widehat{\text{s.e.}}(\hat{\beta}_1) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}$$

- Makes you want to spread out x_i
  - ... but make sure you can still check linearity!

Model comparison

- Suppose you want to compare a model that only uses first j features with the one that uses all p.
- The relevant test is,

$H_0 : \beta_{j+1} = \cdots = \beta_p = 0$
$H_1 : \beta_{j+1}, \ldots, \beta_p$ arbitrary
For example,

- Are the higher-order terms in a polynomial regression actually helping?
- Accounting for differences in backgrounds, does taking a data science course increase your income?

Measuring Discrepancy

- Full model's error,

$$RSS = \sum_i (y_i - x_i^T \hat{\beta})^2$$

  - Another relevant metric: MSE = RSS/v, where v is degree of freedom of residuals, $v = n -$ # fitted parameters)
  - $MSE = RSS/(n - p - 1)$

- Submodel's error is,

$$RSS_0 = \sum_i (y_i - x_i^T \hat{\beta}^0)^2$$

  where $\hat{\beta}^0$ is fit assuming the last $p - j$ coordinates are 0.

  - $MSE_0 = RSS_0/(n - j - 1)$

- Small model always has larger error rate, but is it really *that* much worse?

Evaluating Discrepancy

- Compare the error rates between the models,

$$\hat{F} = \frac{(RSS_0 - RSS)/(v \text{ of difference in residuals})}{MSE}$$

$$= \frac{\frac{1}{p-j}(RSS_0 - RSS)}{\frac{1}{}RSS}$$

$n-p-1$ RSS

- Under same regression assumptions, statistic follows an F-distribution
    - the random variable X of the F-distribution can be written as

$$X = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$$

where $s_i^2 = RSS_i/d_i$; RSS_i is sum of squares of d_i random variables drawn from $\mathrm{N}(0, \sigma_i^2)$.

Nuances

- Regression coefficients represent the mean change in the response variable for one unit of change in the predictor variable while holding other predictors in the model constant.
    - Question: I have a p-value for β_j. I introduce a new variable to my regression. How does the p-value for the original variable change?
- *t-test* can assess only one regression coefficient at a time.
    - Use this to pinpoint *which* of the β_j's are not significant in the predictor-response relationship.
- *F-test* can assess multiple coefficients simultaneously.
    - Use this to assess the significance of all coefficients of the overall model.
    - If you don't have any significant p-values for the individual coefficients in your model, the overall F-test won't be significant either.

Bayesian inference

Main Idea

- Instead of accepting or rejecting states of the world H, place probabilities over them
- Instead of arguing whether coin is fair, define a distribution over plausible probabilities

Posterior Updating

- Key logical device is Bayes rule,

$$p(H|\text{data}) \propto p(\text{data}|H)\,p(H)$$

- New belief about world = new data × old belief
- "Update the prior to the posterior"
- Stronger priors are harder to overcome

Does Bayesian approach make sense?

"If all you have is a hammer, then every problem will look like a nail."

As applied to statisticians, this refers to absorption with the technique rather than the problem; to the failure to see the problem whole; to ask, "Does it all make sense?"

...

MINUTES, APRIL 25-27, 1982 MEETING

Suggestions and issues raised in general discussion included:

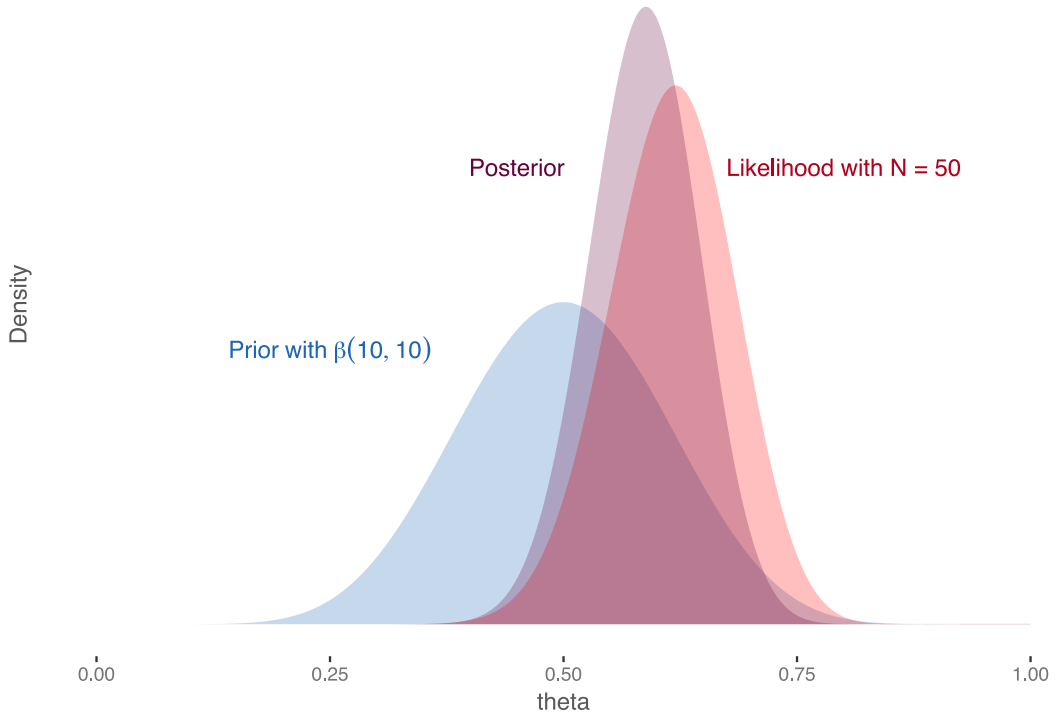
a. A discussion of the positive and negative aspects of Bayesian methodology when applied in risk estimation.

b. Questions and responses on the meaning and implications of the phrase "uncertainty propogation" and on the communications problems engendered in its use.

From Leo Breiman in Nail finders, edifices, and Oz

Revisiting Coin Flips

- Prior assigns probabilities to $p \in [0, 1]$
- We leave the update to a black box inference engine



Revisiting Coin Flips

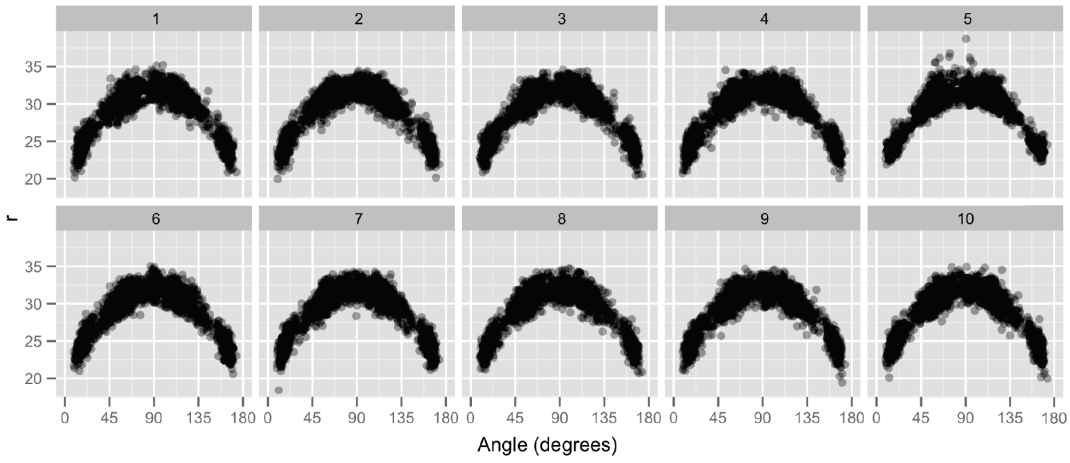
- Go to example here

Revisiting Linear Regression

- Almost exact same mechanics translates to linear regression
- Go through example here

Posterior Predictive Checks

- Once you've fitted a model, try generating data from it
- Do the simulated data look like your real data?
- Do they match specific properties of your real data?
  - (means, quantiles, correlations, ...)
- This idea is useful in frequentist settings too



Practical considerations

- Modular, flexible ways of building models and drawing inference
  - E.g. hierarchical modeling
- Prior choice is arbitrary, but you can measure sensitivity of inferences to that choice
- Freedom in design can outweigh complexity of computation

```
import {slide} from @mbostock/slide

import {coin_flipping} from @krisrs1128/sufficiency-illustrated

<style>

mtex_block = ƒ()

mtex = ƒ()
```