Search

Fork Sign in

👋 Welcome. This is live code! Click the left margin to view or edit. ×

Kris Sankaran

♡ 1

Published Sep 26, 2019

Introduction to Inference

IFT6758, Fall 2019

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

Optional reading: MSMB 6.1 - 6.6 and Statistics for Hackers

?

Inference

- Meta-Algorithms: Evaluation of processes people use to learn from data

- Science: How to go from the finite to the general?

"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

Plan

- Inference for coin clips
  - 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 large a discrepancy would you need to see?
- How many flips would convince you?

.. only need to reme

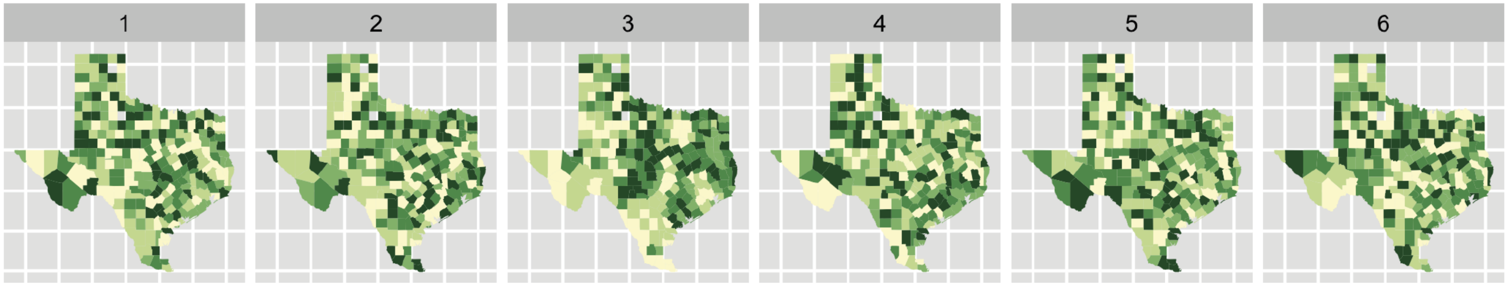
Is your coin fair?

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

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

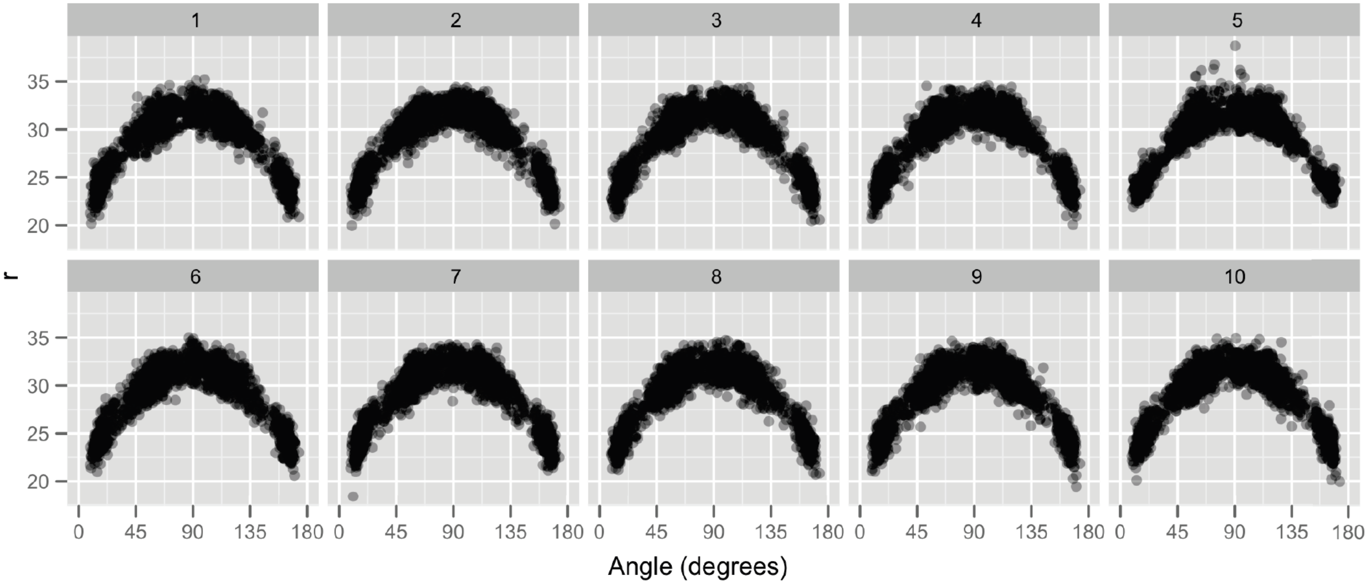


Counts of cancer deaths. 5 come from null model, where there is no spatial correlation.

Model of Realit

Model of Reality

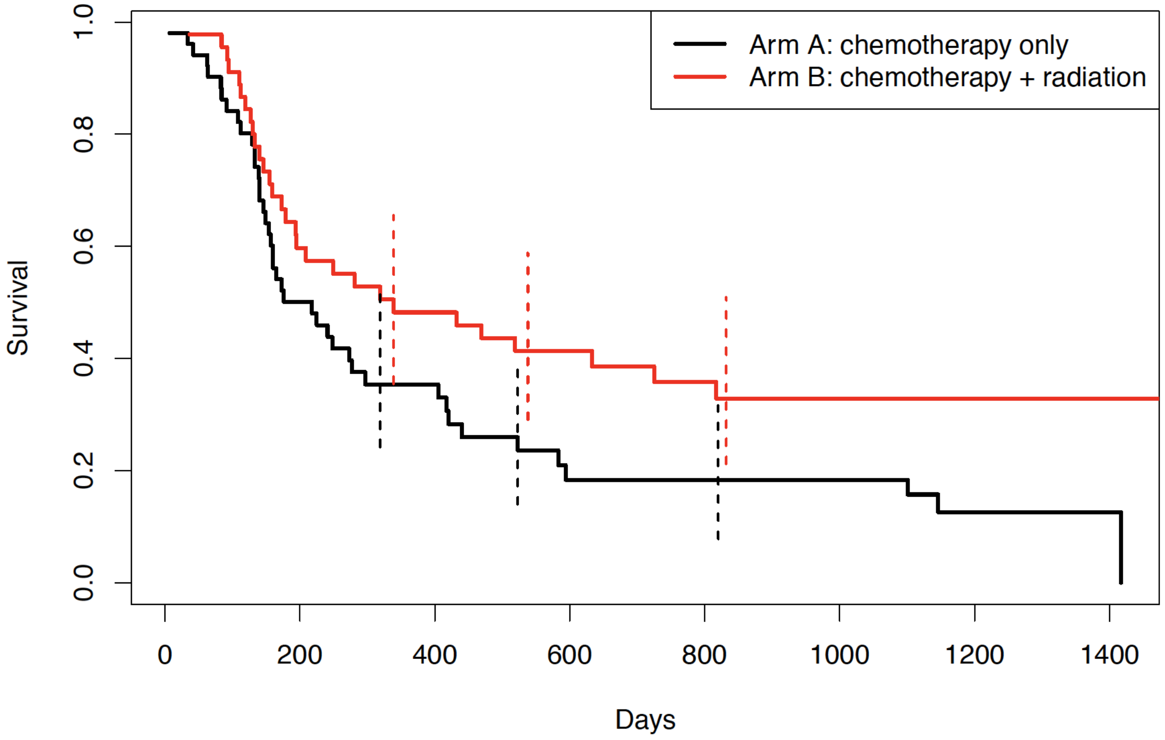
- A null model defines a default simulator of reality
- Our goal is to determine whether an observed dataset is consistent with that reality



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

Measuring Discrepancies

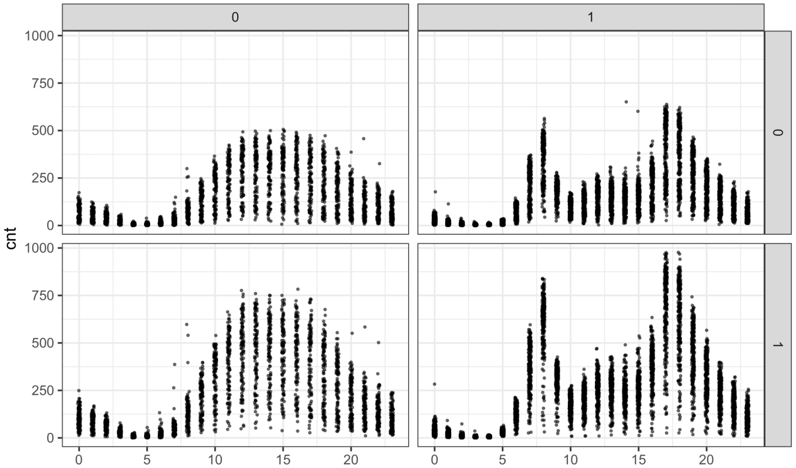
- We'll look in the data for potential discrepancies, and try to judge whether they are meaningfully large



If you didn't have the error bars, you might think the two treatments are quite different.

Challenges

- All theories are approximations to reality
- If your approximation isn't that believable, you will obviously reject the null (and no one will care)
- Most critical assumption: Independence
    - Thinking you have more evidence than you do is usually much worse than small distributional differences



hr

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

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

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

and sound an alarm if p_H is very far from 0.5

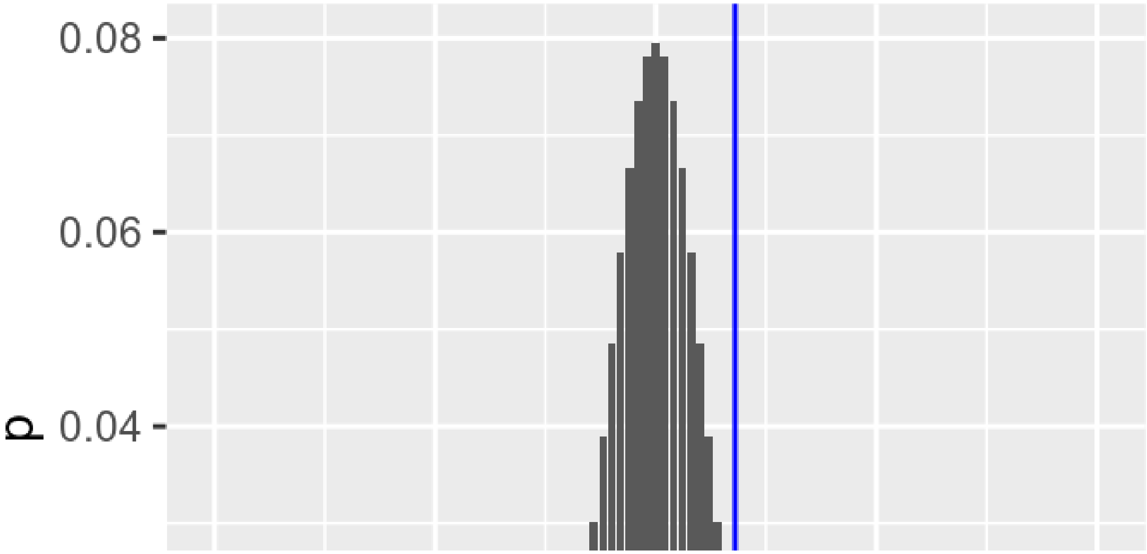
Not always obvious

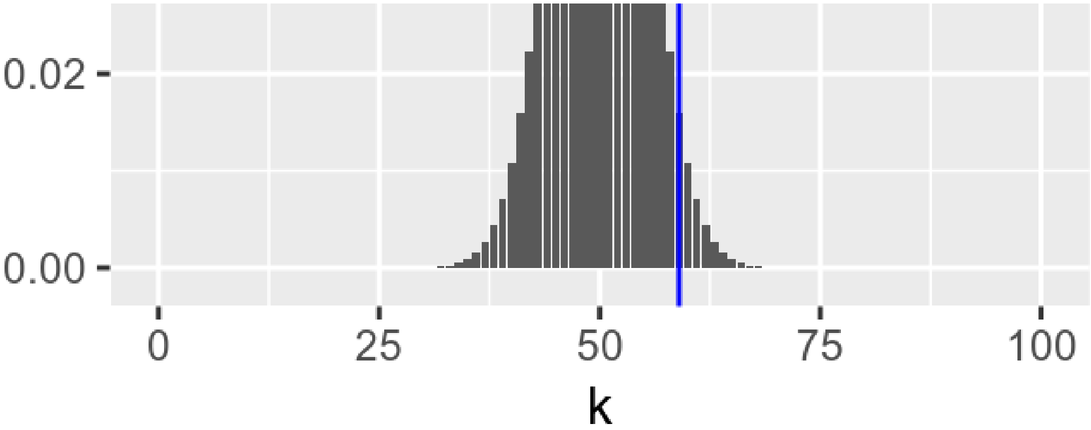
- What if you saw HHHHHHHHHHTTTTTTTTTT?
- Seems fair according to our metric....
- Different choices of test statistic are sensitive to different departures from your theory

Reference Distribution: Theory

- 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

$$\mathbb{P}\left[\hat{p}_n = \frac{k}{n}\right] = \binom{n}{k} p^k \left(1 - p\right)^{n-k}$$



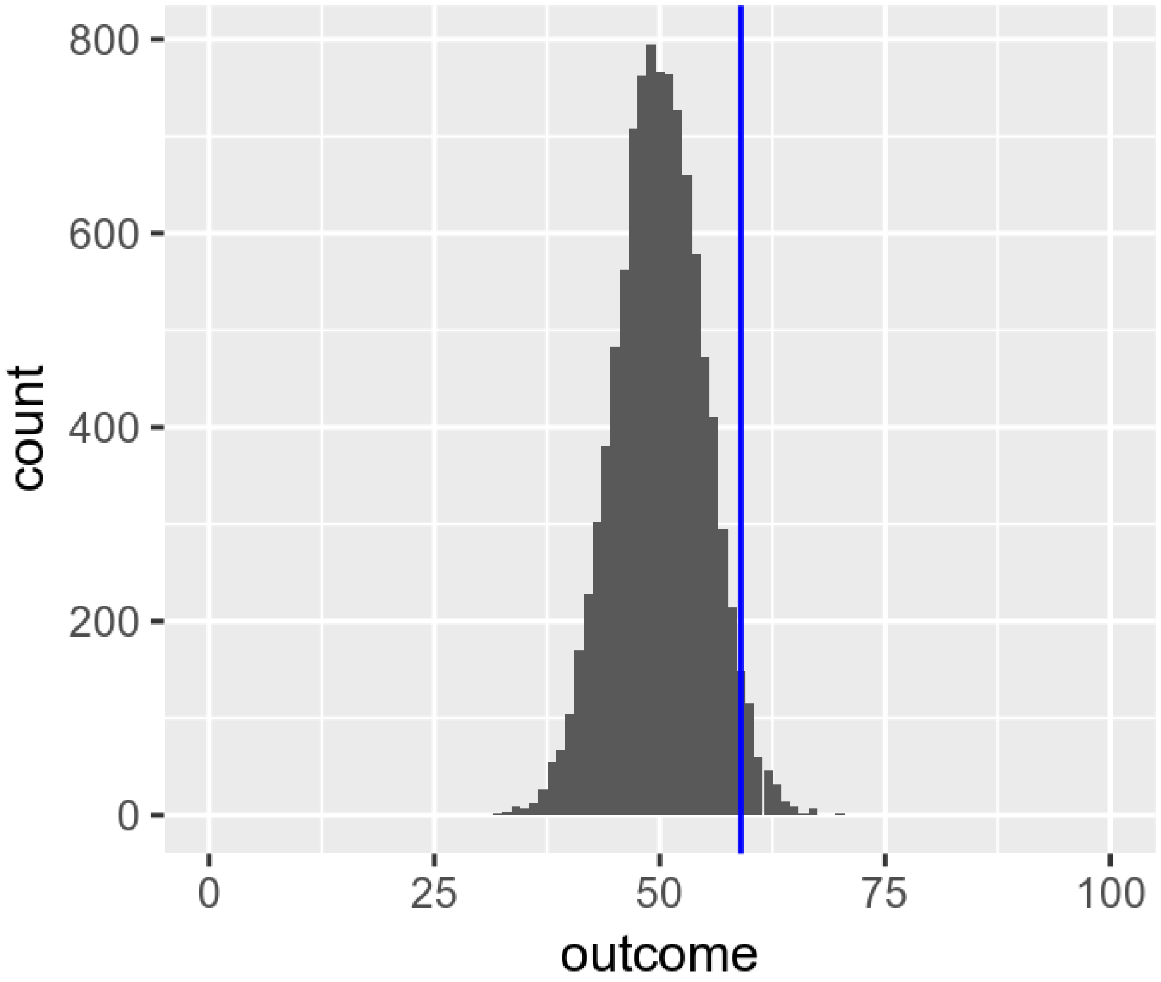


This image (and others) from MSMB.

Reference Distribution: Simulation

- 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

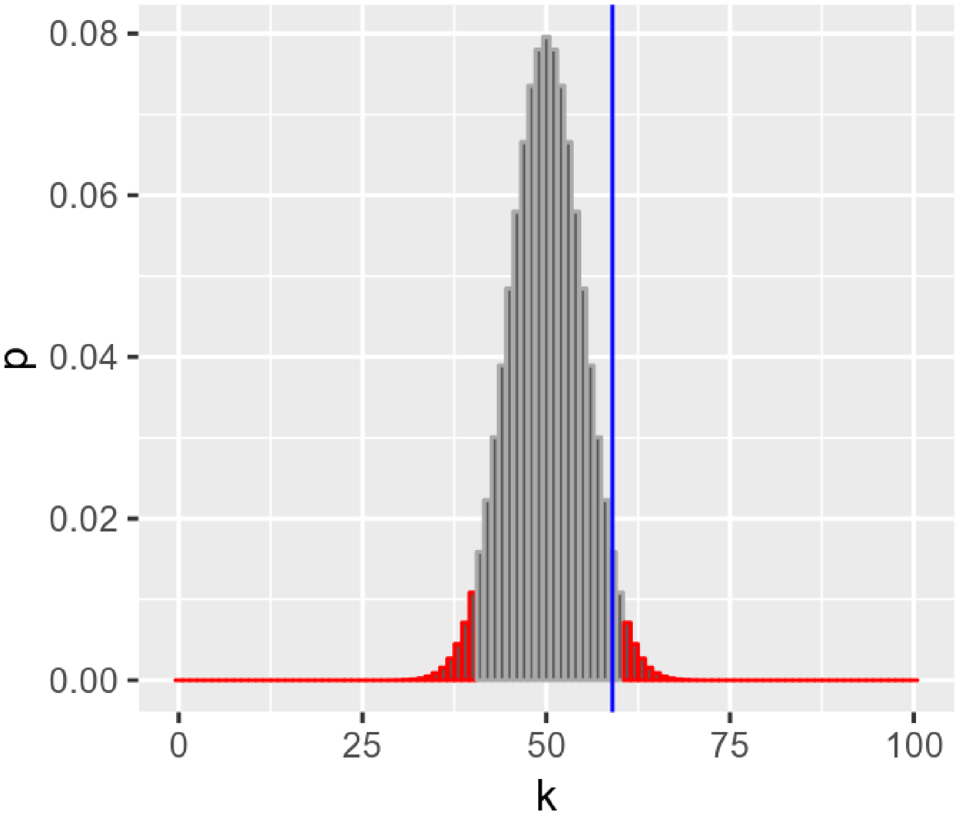
- ○ Especially useful when no closed form analysis



Same reference distribution, from 10^4 simulations

Rejection Region

- Signifance level: Want to keep false alarm rate below α%
- Power: Want to be sensitive to small discrepancies

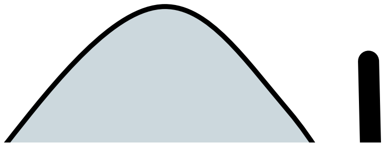


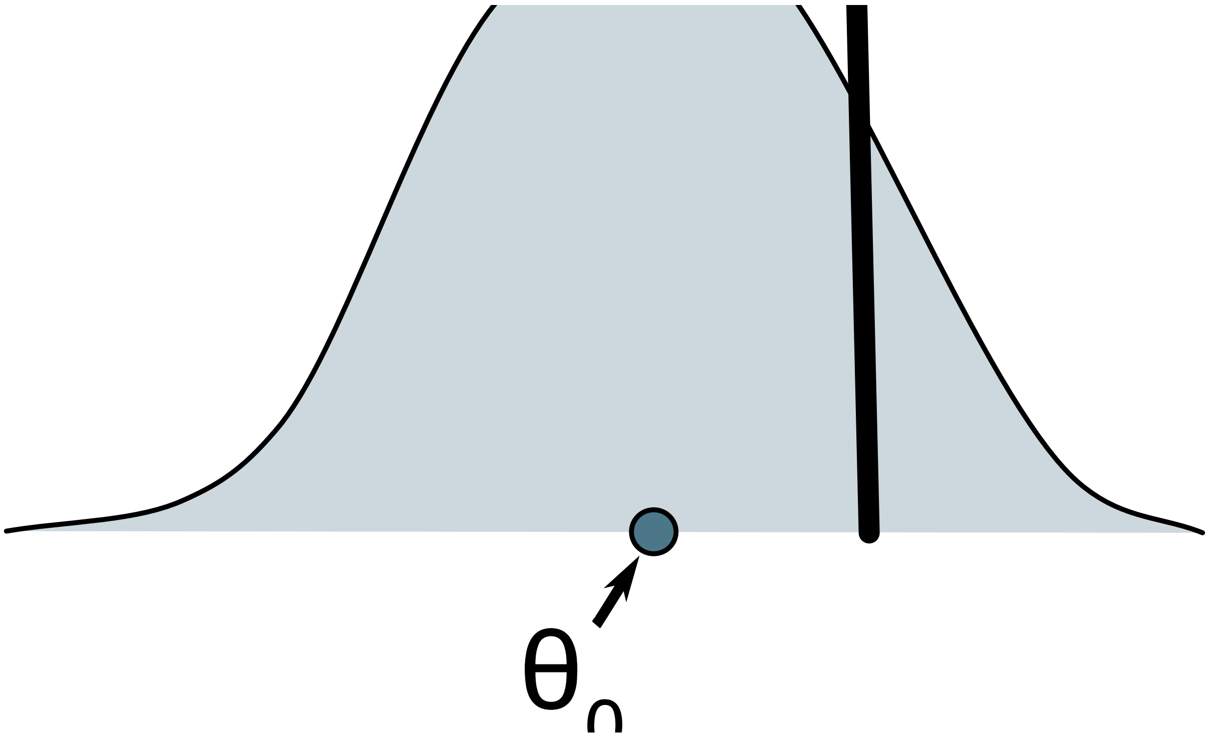
By rejecting in the tails, we can reject as many types of outcomes as possible while making sure we rarely reject when the null ($p = 0.5$) is actually true.

Definition: The p-value gives the area in the reference distribution of discrepancies that are "more extreme" than the one you've observed.

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

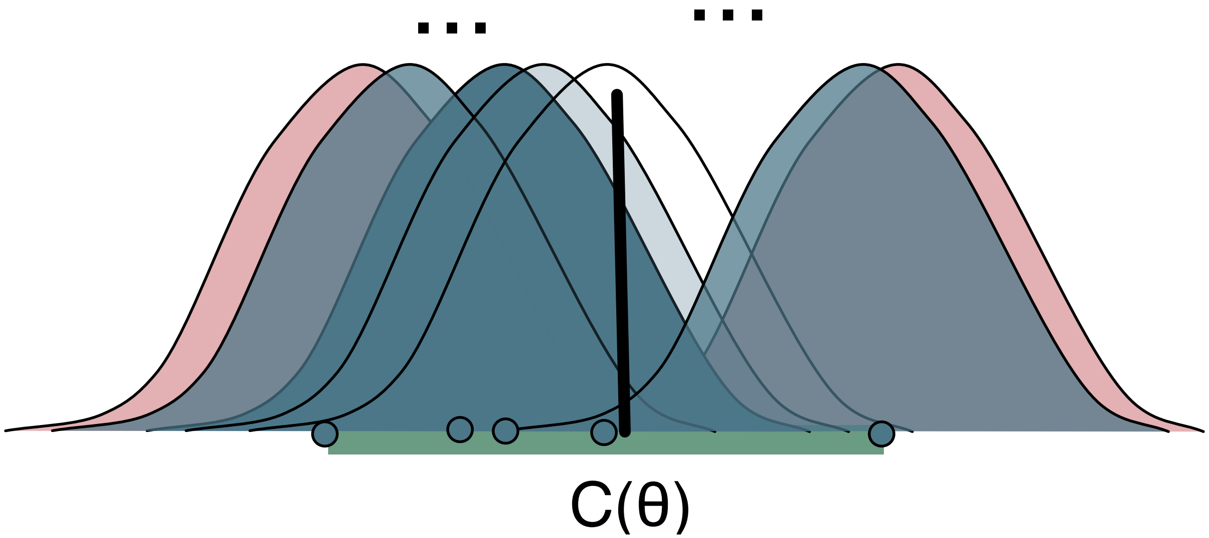




We wouln't have rejected this θ_0 because the test statistic lies in the bulk of the reference distribution.

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



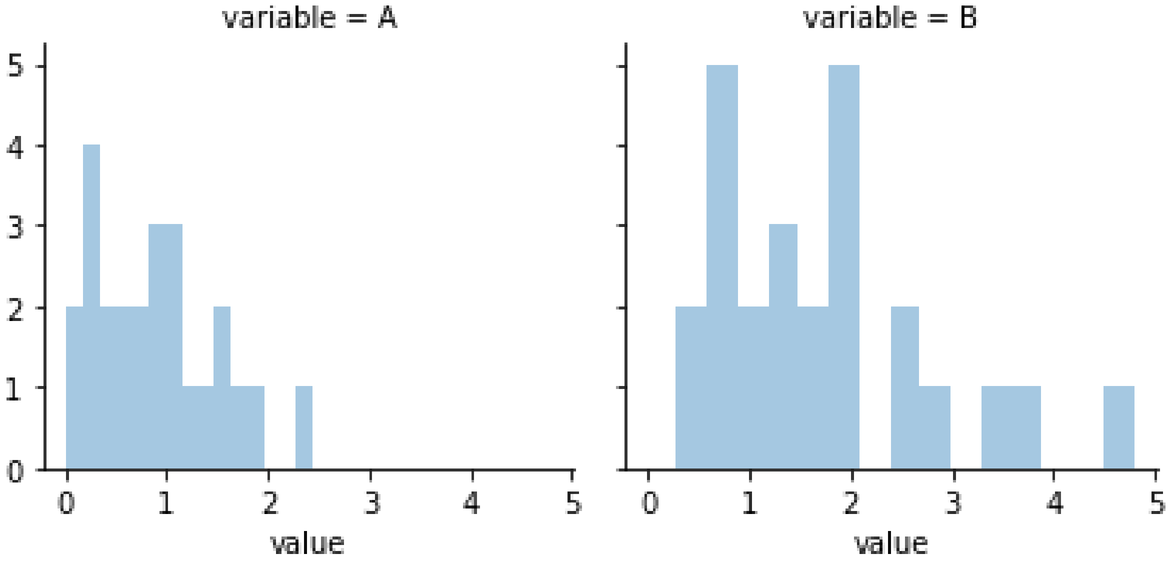
We start rejecting once we get past the red reference distributions. All the θ's between these belong to our confidence interval.

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



Follow along.

Measuring Discrepancy

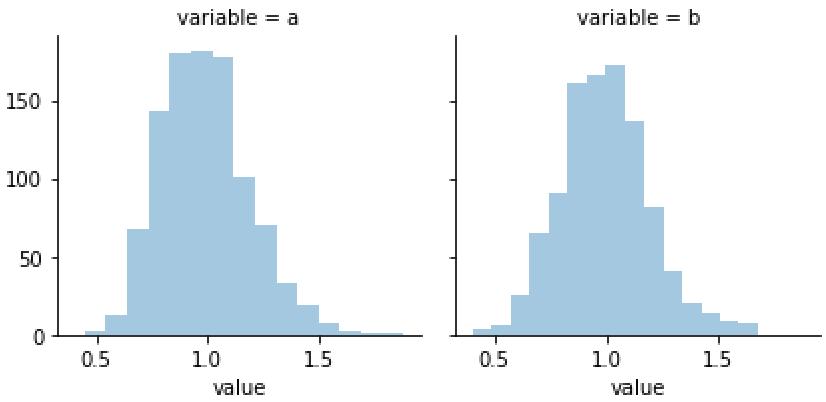
- Common to use a normalized difference, e.g.

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

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

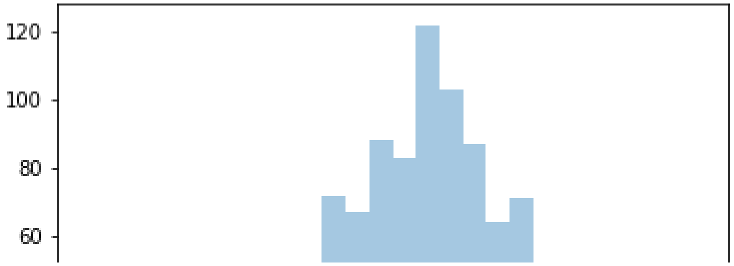
Measuring Discrepancy

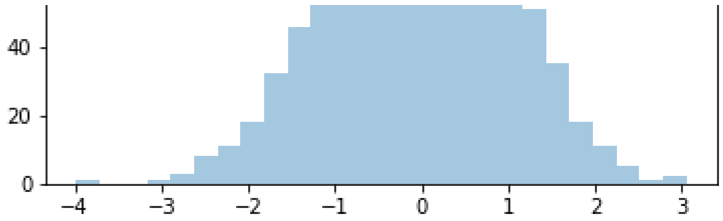
These are the distributions for $\bar{x}_1$ and $\bar{x}_2$. Even though the original data are relatively far from gaussian, the central limit theorem kicks in.



Measuring Discrepancy

The resulting normalized difference is about t-distributed.

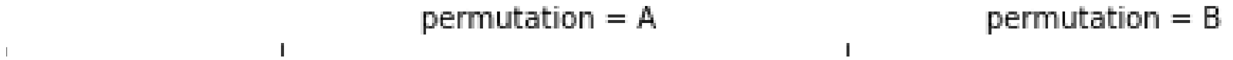


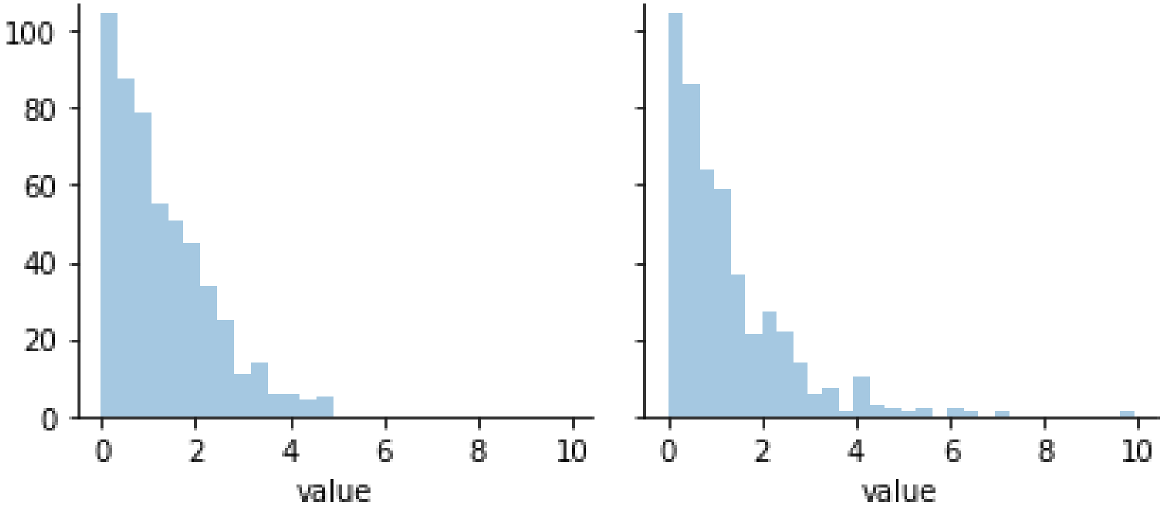


Evaluating Discrepancy

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

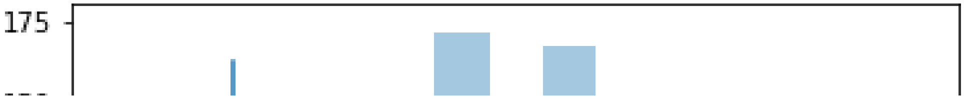
The separate histograms (one per group) become indistinguishable -- it's a simulation from the null for the raw data. Computing difference in means gives a reference distribution.

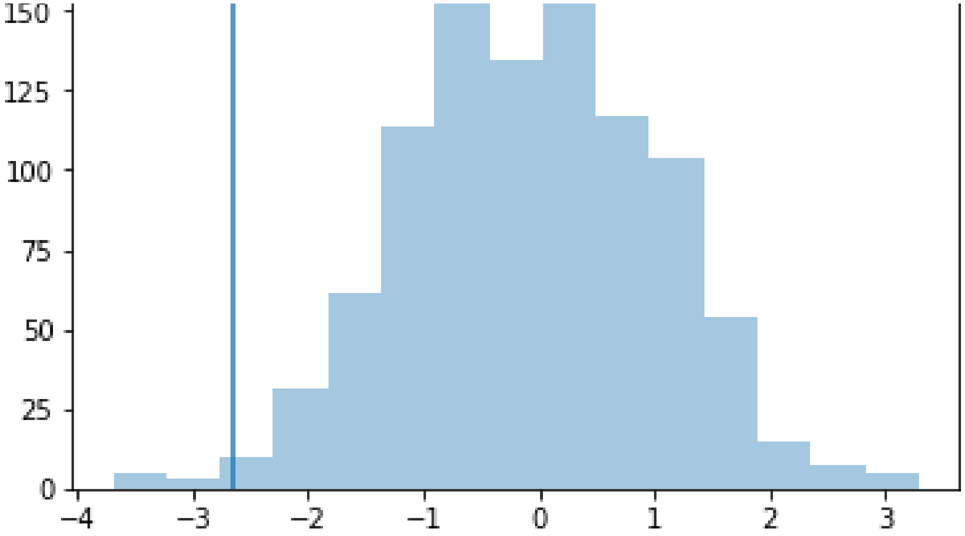




Evaluating Discrepancy

- 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

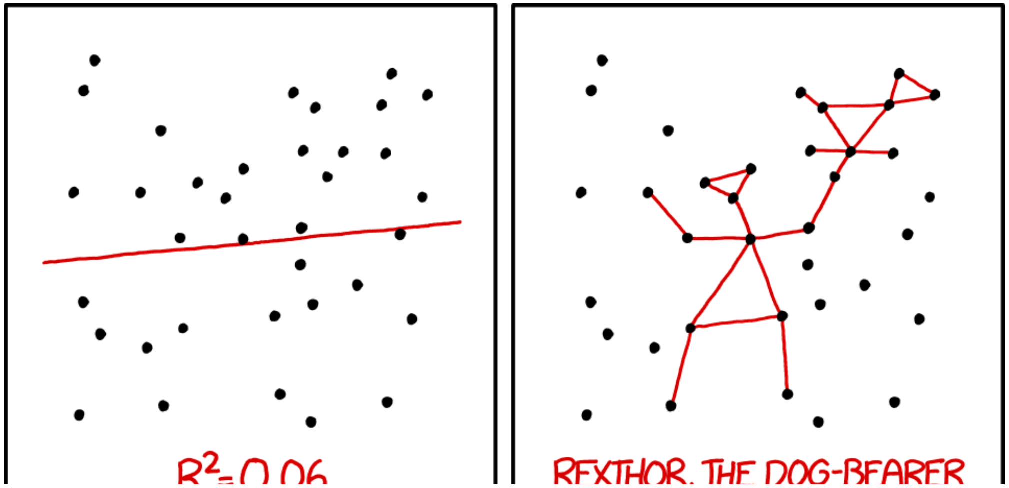
- Association: Is there are relationship between x and y?
- Model comparison: Is a model using additional features actually better?

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?



I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

Discrepancy

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

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

Evaluating: Also has a t-distribution if you assume i.i.d. gaussian errors $\epsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$. We'll see some simulation-based alternatives in later lectures.

+

Comment: Experimental Design

The formula for the standard error is illuminating,

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

- 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,

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

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,

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

- Submodel's error is,

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

  where $\hat{\beta}^0$ is fit assuming the last $p - j$ coordinates are 0.
- Small model always has larger error rate, but is it really *that* much worse?

Evaluating Discrepancy

- Compare the error rates between the models,

$$\hat{F} = \frac{\frac{1}{p-j}\left(RSS_0 - RSS\right)}{\frac{1}{n-p-1}RSS}$$

- Under same regression assumptions, statistic follows an F-distribution

Nuances

- A model comparison with one coefficient extra looks for an association with x_j, after controlling for other features
  - 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?
- If you reject when there are many additional coefficients, you won't be able to pinpoint *which* of the β_j's are responsible

Bayesian inference

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

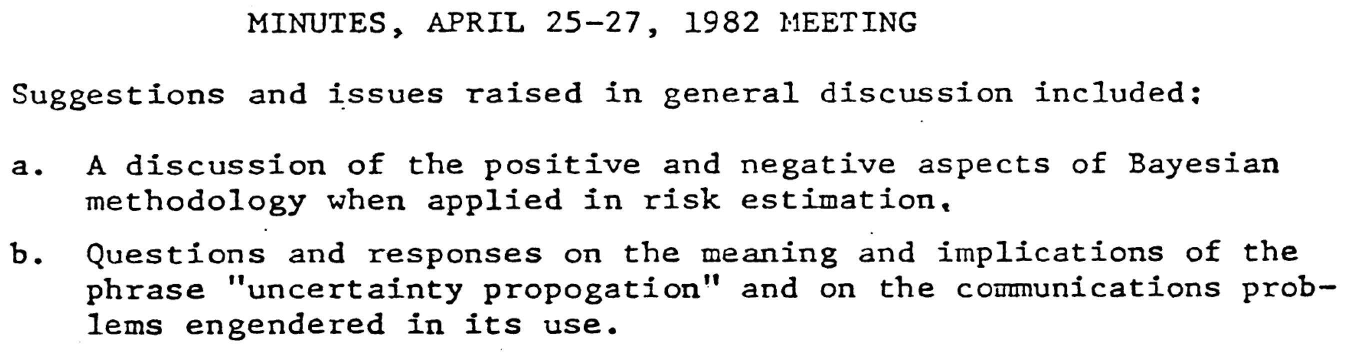
Posterior Updating

- Key logical device is Bayes rule,

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

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

Some healthy perspective...



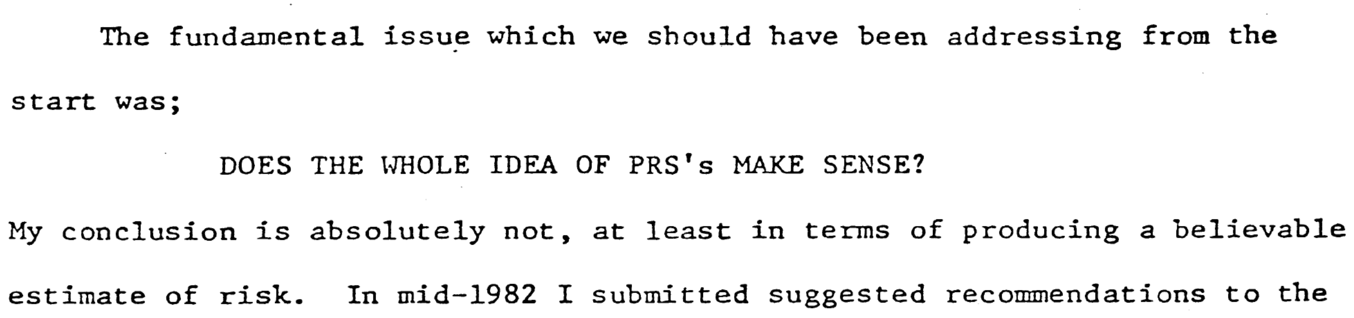
MINUTES, APRIL 25-27, 1982 MEETING

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.

...



The fundamental issue which we should have been addressing from the start was;

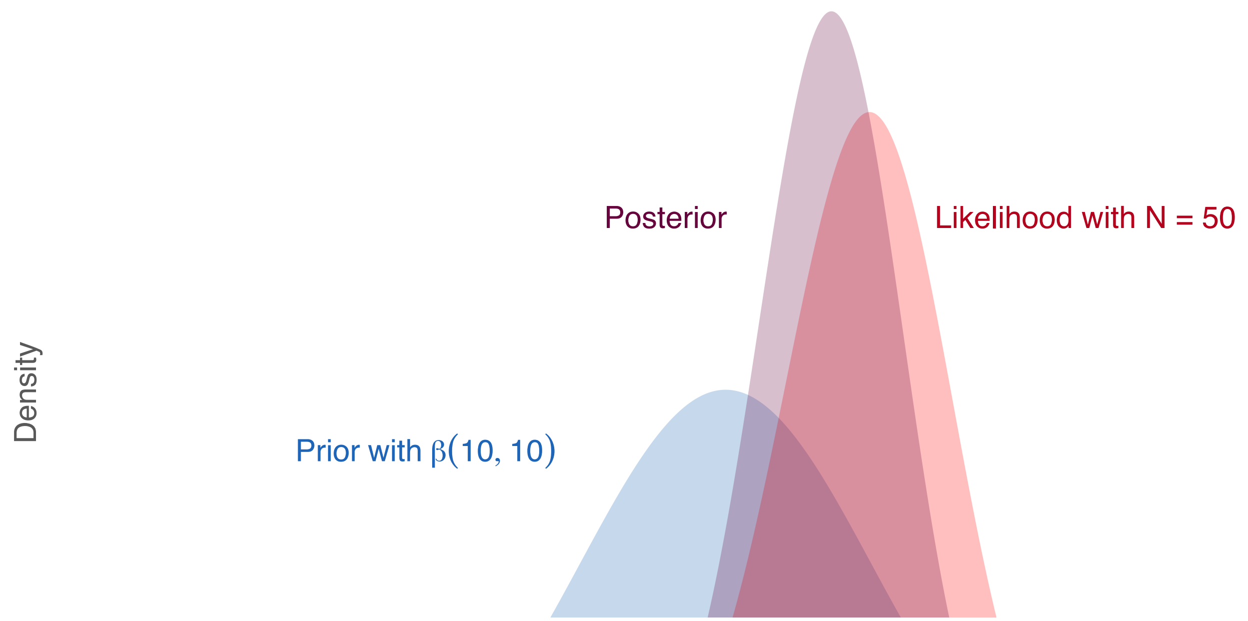
DOES THE WHOLE IDEA OF PRS's MAKE SENSE?

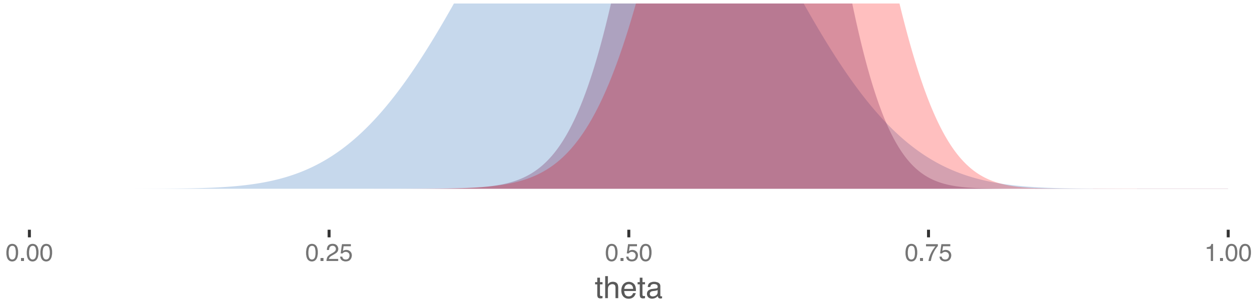
My conclusion is absolutely not, at least in terms of producing a believable estimate of risk. In mid-1982 I submitted suggested recommendations to the

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

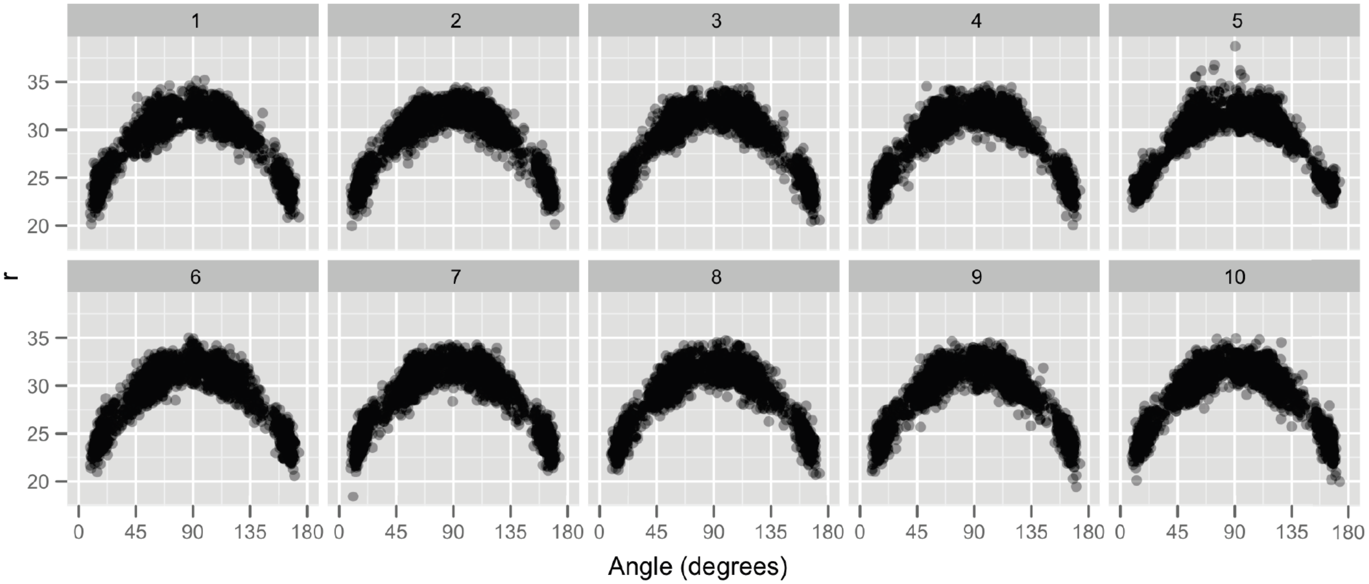
- Go to example here

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()
```