Power
&
Bayesian Inference

Overview

- **Power** (Frequentist Hypothesis Testing cont'd)
  - What is power?
  - Calculating Power
  - Calculating the sample size (n) for a given level of Power
  - Relation to A/B Testing

- Bayesian Inference
  - Frequentist vs. Bayesian
  - Prior, Likelihood, and Posterior Distributions
  - Revisiting MAP

TO BE **POWERFUL**!!!

Powerful Test

Which test do we like better?

TEST 1	TEST 2
$\alpha = 0.05$	$\alpha = 0.05$
"powerfulness" = 0.8	"powerfulness" = 0.3

Statistical Power

- The power of a hypothesis test is the probability that the test correctly rejects the null hypothesis (H_0) when the null hypothesis is false.

- Power = P(Reject H_0 | H_0 is false)
  = P(Reject H_0 | H_1 is true)

- What's our chance of rejecting the null when the null is in fact false?

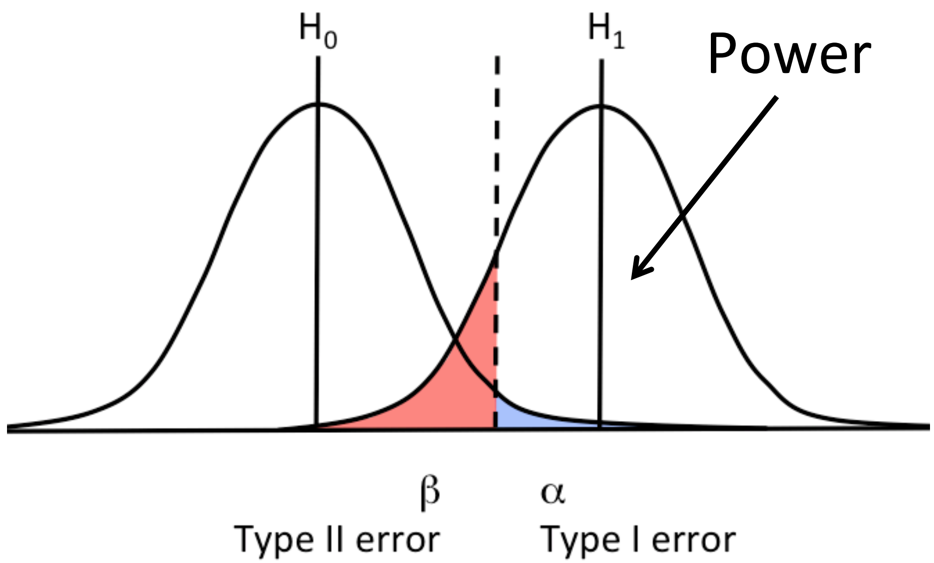
What's the probability of detecting a real effect?

Find the Power

	H_0 is true	H_0 is false
Accept H_0	Correct Decision (1-α)	Type II Error (β)
Reject H_0	Type I Error (α)	Correction Decision (1-β)

Power = 1 – Type II Error
= 1 - β

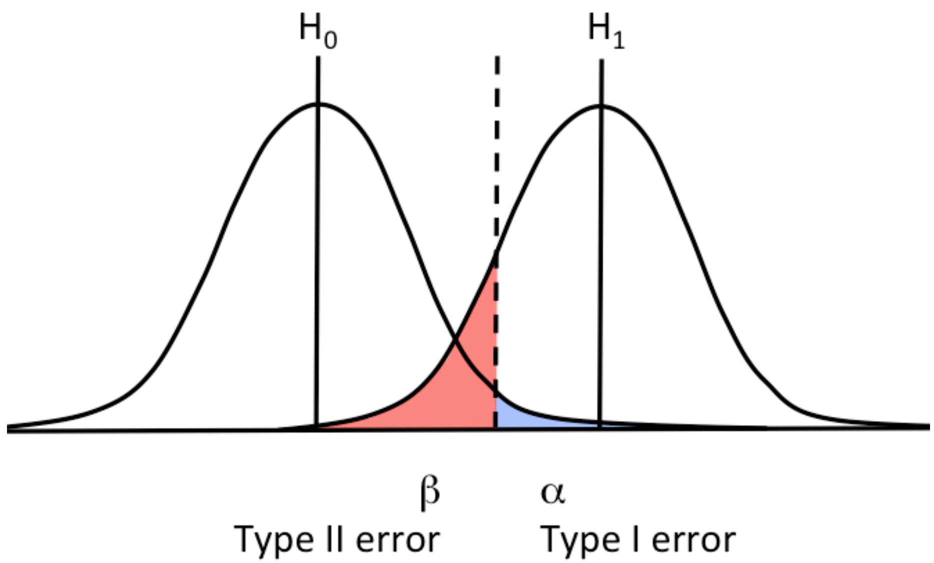
Type I and Type II Errors



	H_0 is true	H_0 is false
Accept H_0	Correct Decision (1-α)	Type II Error (β)
Reject H_0	Type I Error (α)	Correction Decision (1-β)

1. Draw two curves based on H_0 (get μ_0) and sample data ($\bar{x}$ and s)
2. Define rejection region (color region until you hit α)
3. Calculate β by computing area to left of decision boundary
4. Power is 1 - β

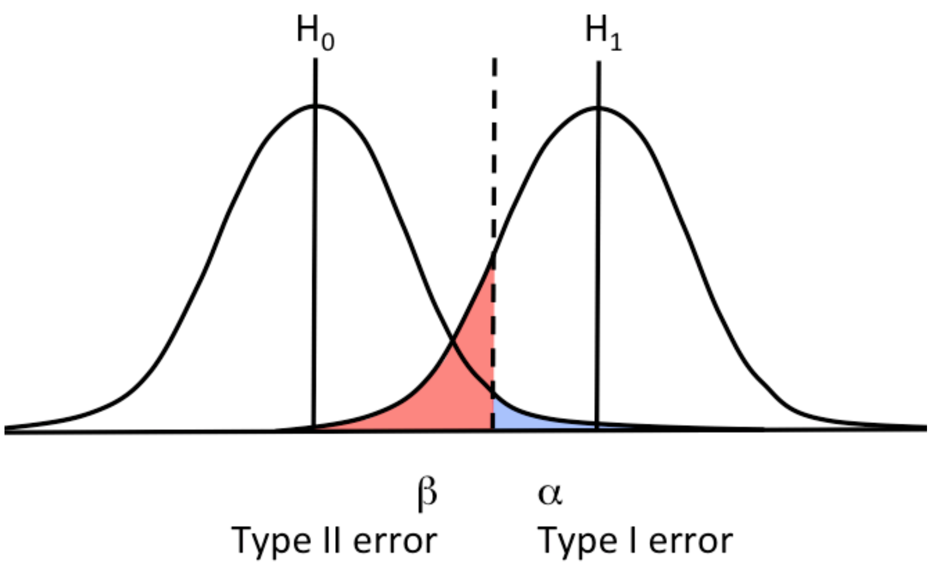
Type I and Type II Errors



	H_0 is true	H_0 is false
Accept H_0	Correct Decision (1-α)	Type II Error (β)
Reject H_0	Type I Error (α)	Correction Decision (1-β)

What happens when we increase the sample size?

Factors influencing Power



	H_0 is true	H_0 is false
Accept H_0	Correct Decision (1-α)	Type II Error (β)
Reject H_0	Type I Error (α)	Correction Decision (1-β)

*What happens to power as we **increase** each of the following?*

- α ↑
- effect size ↑
- standard deviation ↓
- sample size ↑

Design of Experiment

- Usually need to decide what sample size to collect
- Especially true if important to minimize number of samples because it may be painful (new surgical procedure) or costly (paying test subjects)

Power analysis can allow you to determine the sample size needed to detect a particular effect

Calculating Power

Example: One-sample Test of Mean

$$H_0 : \mu = \mu_0 \quad H_1 : \mu = \mu_1 (> \mu_0)$$

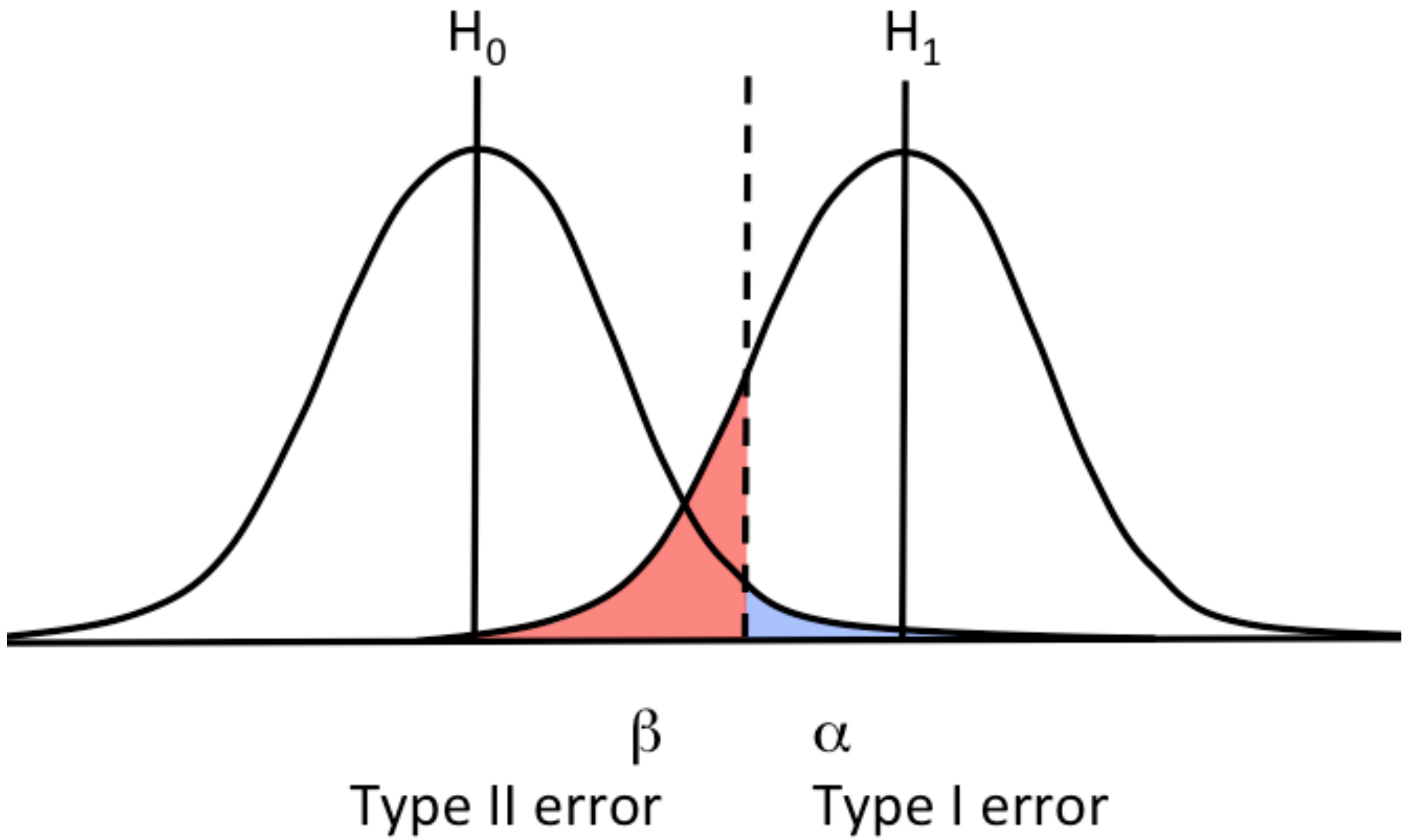
Recall: Power = P(Reject H_0|H_0 is false) = P(Reject H_0|H_1 is true)

- We want to find the critical value, under H_0, beyond which we would reject H_0

$$X^* = \mu_0 + Z^* \frac{s}{\sqrt{n}}$$

- Now find the corresponding probability under H_1

$$Power = P(X > X^* | H_1) = P(Z > \frac{X^* - \mu_1}{s/\sqrt{n}})$$



What happens when we increase the sample size?

Calculating Sample Size

- What if we do not know the true mean and want to collect a larger sample for the test?
- First, we need to have
  - A mean associated with the null hypothesis (μ_0)
  - An estimate of the population mean (μ_1)
  - An estimate of the population standard deviation
  - A fixed significance level (α, often 5%)
  - A desired power level ($1-\beta$, often 80%)
- Then we derive the value for n from the power calculation formula

$$\boxed{n = ((Z_{(1-power)} + Z^*)\frac{s}{\mu_1 - \mu_0})^2}$$

Both Zs should have same sign

A Small Interlude

Review - A/B Testing

- A/B testing is essentially two-sample hypothesis testing
- In practice, we often conduct a small pilot experiment to estimate the sample size for a given power

Recap: Power Calculation

- Decide the critical value for the test statistic, in general,
  - $Z^* = \pm 1.96$ for two-sided test
  - $Z^* = + 1.64$ or $- 1.64$ for one-sided test
- Calculate the corresponding value under the null distribution (X^*)
- Find the tail probability of the above value under the alternative distribution (power!)

Recap: Sample Size Calculation

- Obtain some sort of initial estimation of the parameter/effect we are trying to test
  - e.g. a pilot experiment
- Decide on the desired power of the test
  - e.g. power = 0.8
- Calculate the sample size using the initial estimation and desired power

Bayesian Inference

Frequentist vs. Bayesian

<u>Frequentist Probability</u>

"Long Run" frequency of an outcome

<u>Subjective Probability</u>

A measure of degree of belief

Bayesians consider both types

Frequentist vs. Bayesian

Adapted example from Jim Berger's book, the Likelihood Principle

Experiment 1:

A fine classical musician says he's able to distinguish Haydn from Mozart.
Small excerpts are selected at random and played for the musician.
Musician makes 10 correct guesses in exactly 10 trials.

Experiment 2:

Drunken man says he can correctly guess what face of the coin will fall down, mid air.
Coins are tossed and the drunken man shouts out guesses while the coins are mid air.
Drunken man correctly guesses the outcomes of the 10 throws.

Frequentist vs. Bayesian

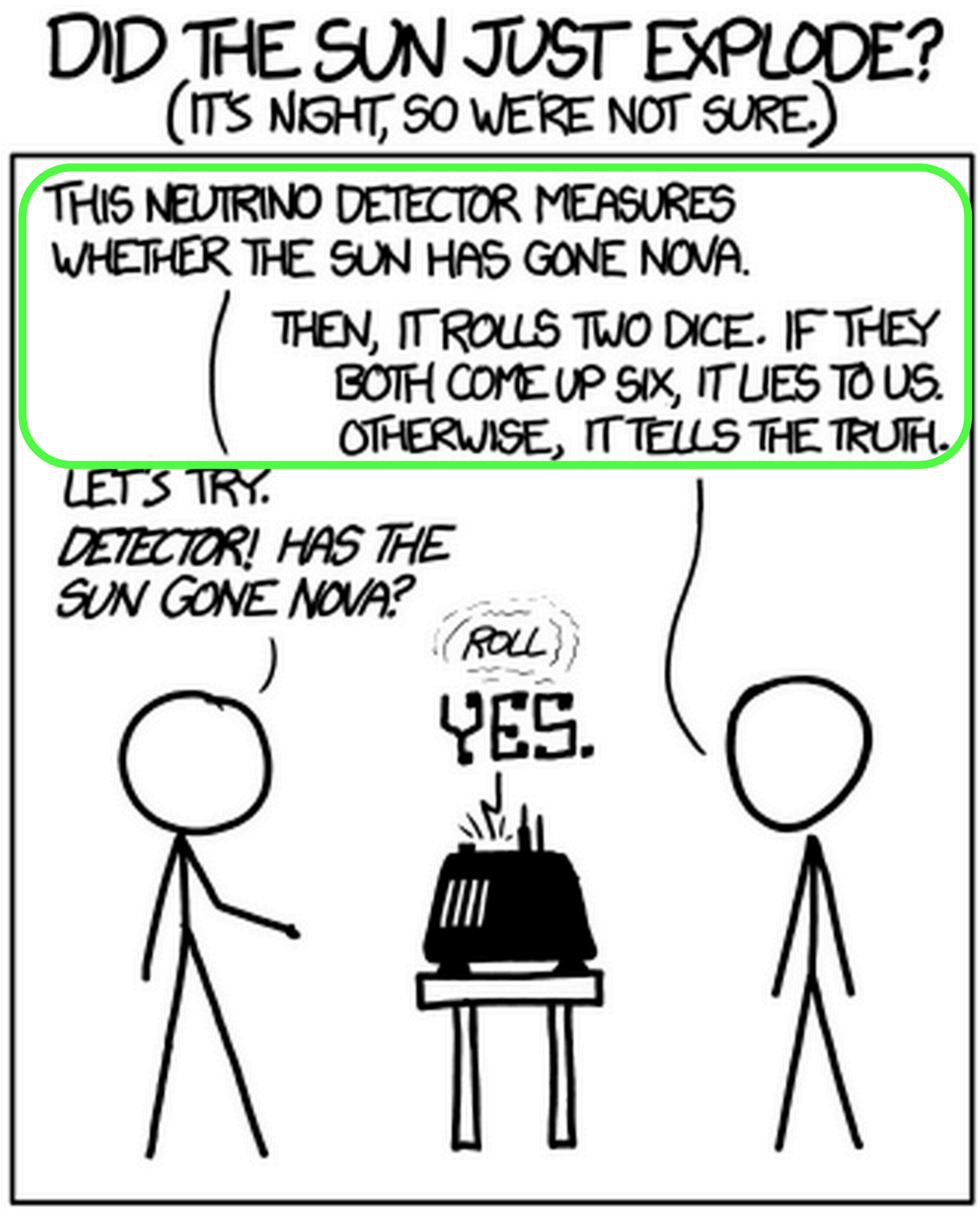
Frequentist: “They’re both so skilled! I have **as much confidence** in musician’s ability to distinguish Haydn and Mozart as I do the drunk’s to predict coin tosses”

Bayesian: “I don’t know man…”

- A Bayesian would incorporate some prior confidence about the musician’s ability and the drunk’s.

Frequentist vs. Bayesian

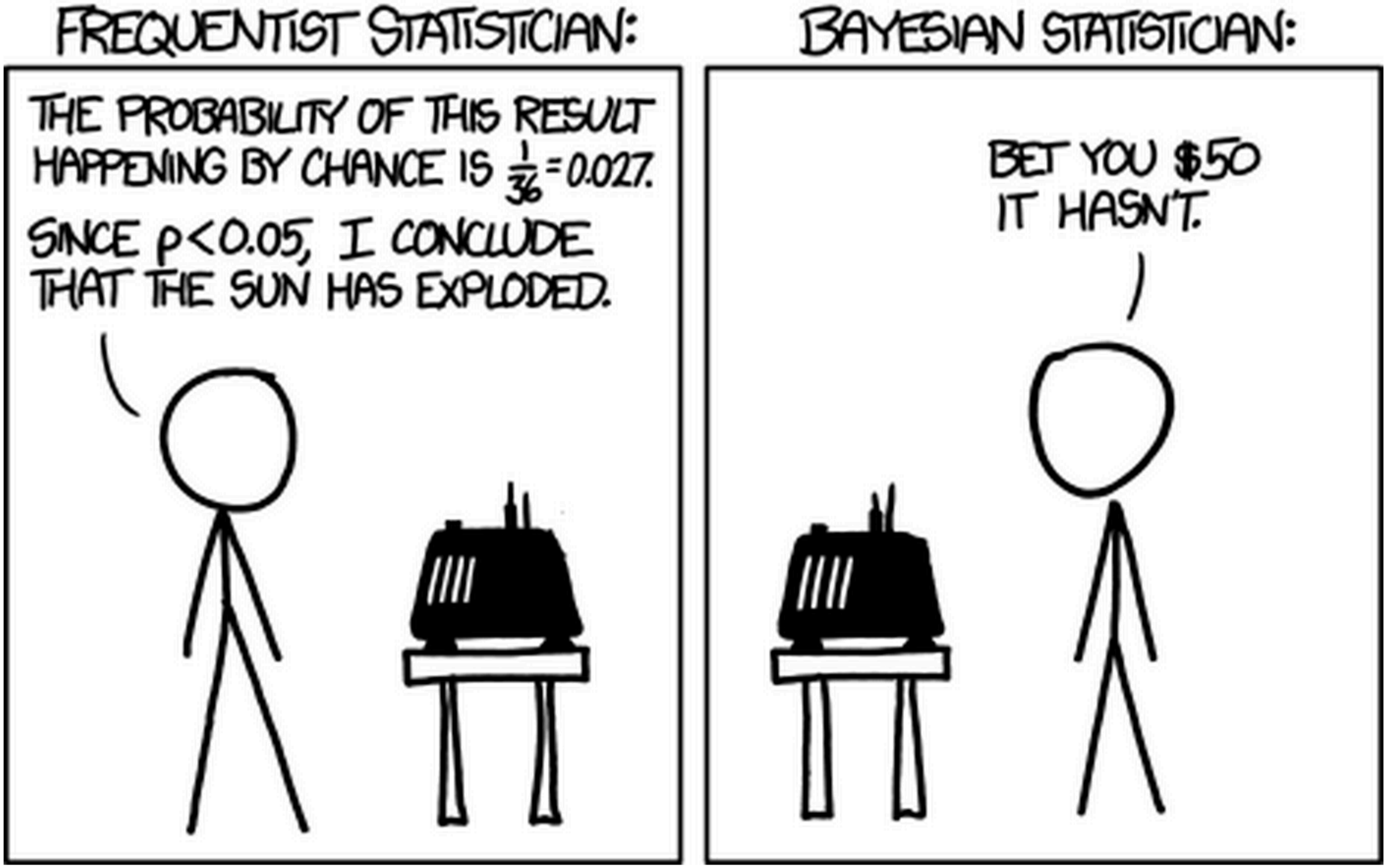
H_0: Sun has not exploded...



Evidence collecting process

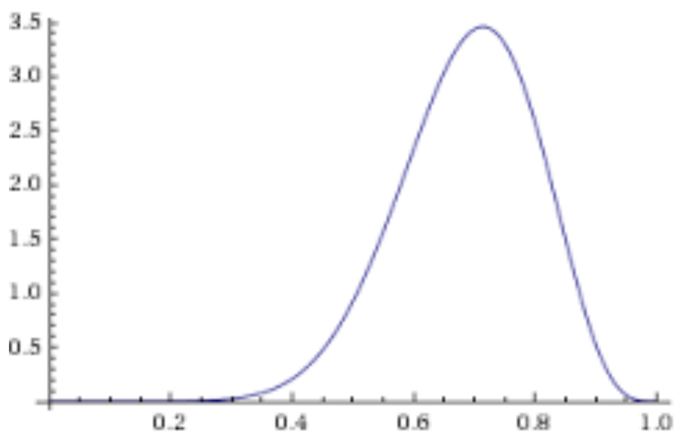
Evidence!

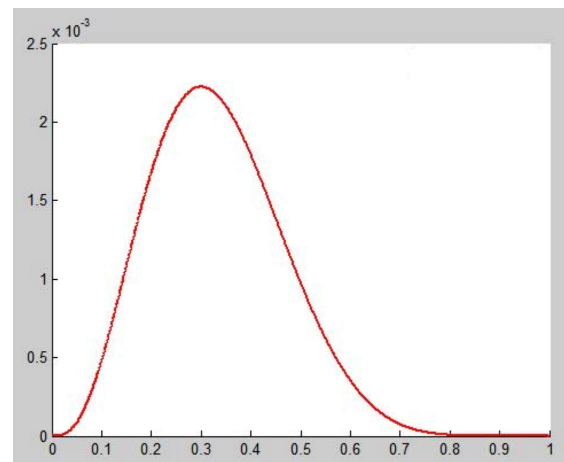
Frequentist vs. Bayesian



Bayesian Inference

- Simply updating beliefs after considering new evidence
- Probability as measure of believability in event
  - A priori, can just make something up…
  - For ex., musician’s ability to distinguish Haydn from Mozart





Bayesian Inference

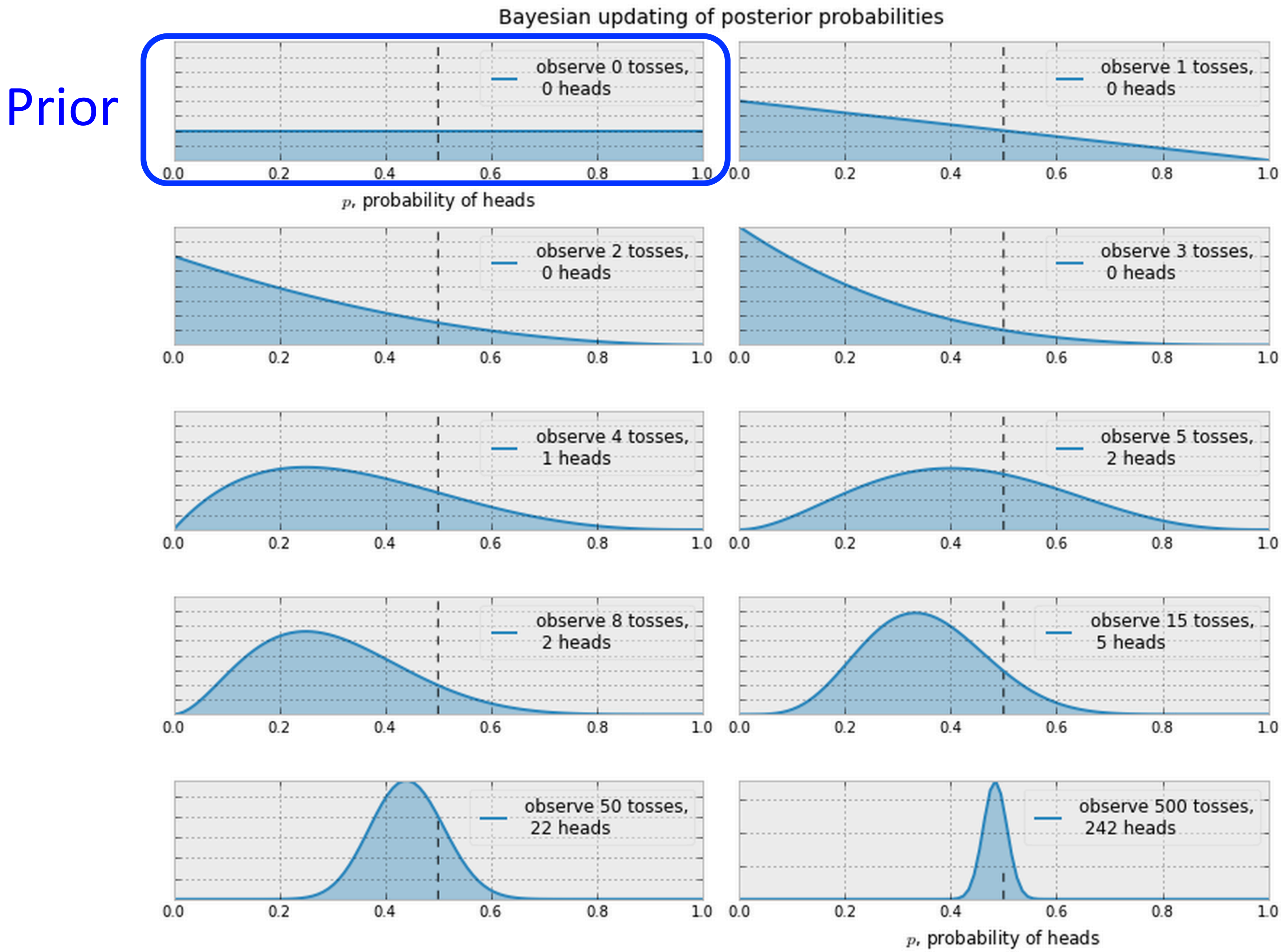
Bayes Rule

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}$$

Posterior distribution

$$\pi(\theta \mid \mathbf{x}) = \frac{f(\mathbf{x} \mid \theta)\pi(\theta)}{\int f(\mathbf{x} \mid \bar{\theta})\pi(\bar{\theta})d(\bar{\theta})}$$

- Prior distribution: Describes our current (prior) knowledge about θ (or A). Can be subjective.
- Likelihood: Distribution for the data (as a function of the parameter).
- Posterior distribution: Our updated knowledge about θ (or A) after seeing the data.



Prior
Bayesian updating of posterior probabilities
observe 0 tosses, 0 heads
observe 1 tosses, 0 heads
observe 2 tosses, 0 heads
observe 3 tosses, 0 heads
observe 4 tosses, 1 heads
observe 5 tosses, 2 heads
observe 8 tosses, 2 heads
observe 15 tosses, 5 heads
observe 50 tosses, 22 heads
observe 500 tosses, 242 heads
p, probability of heads
0.0 0.2 0.4 0.6 0.8 1.0

Inference – MAP

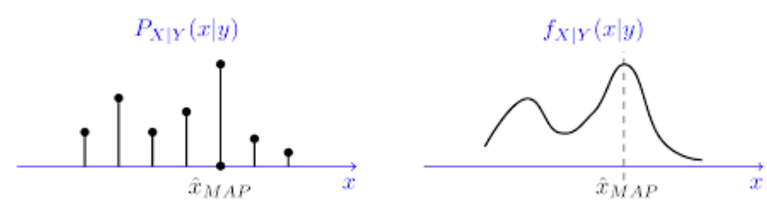
- Maximum a posteriori (MAP) – mode of the posterior distribution
  - For MLE, we have

$$\hat{\theta}_{mle} = \underset{\theta \in \Theta}{argmax}\ f(x|\theta) = \underset{\theta \in \Theta}{argmax}\ log\mathcal{L}(\theta|x_1, \ldots, x_n)$$

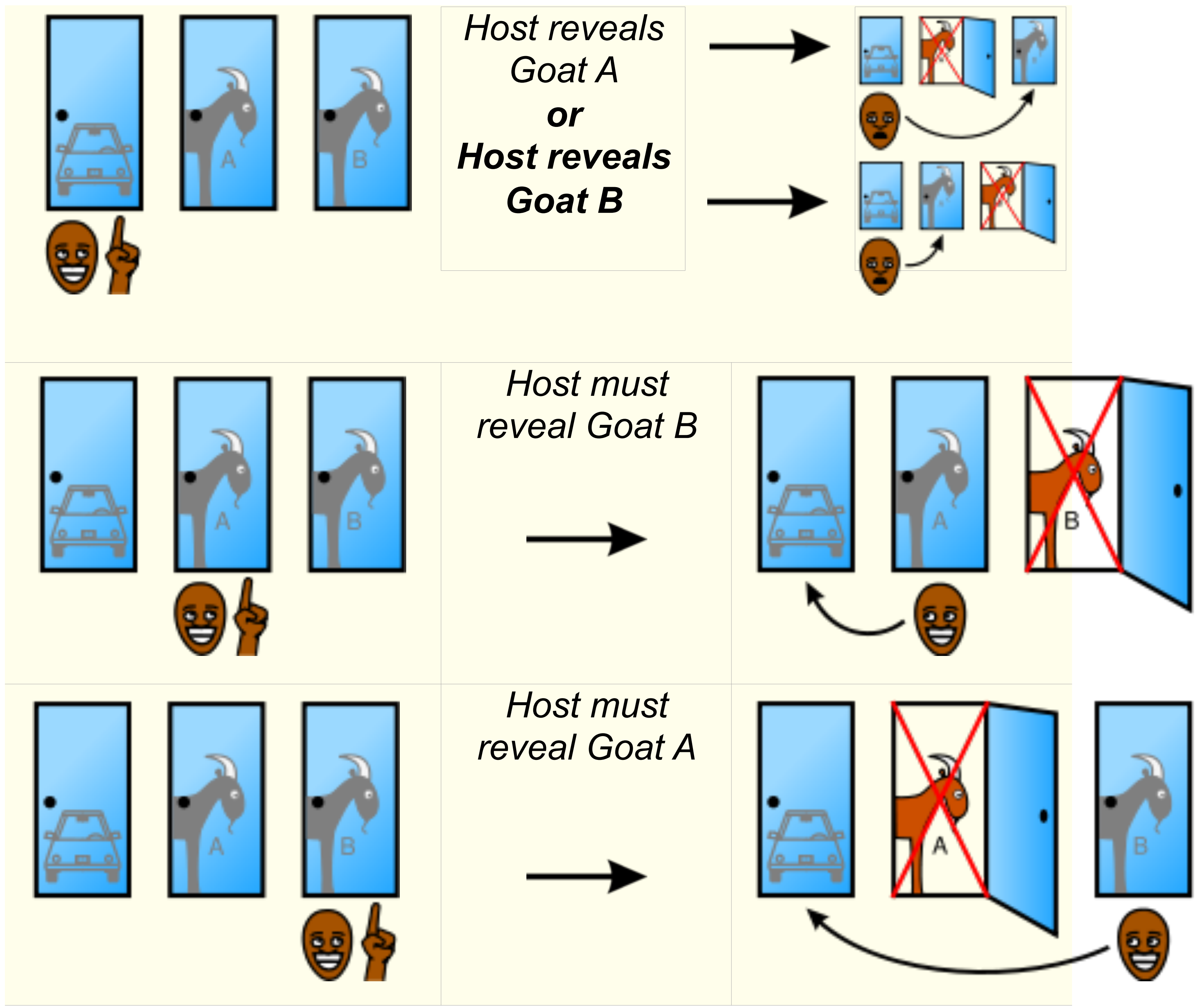
  - For MAP, we assume a prior *g* over Θ, and go one step further to get the posterior.

Simply get Posterior using Bayes

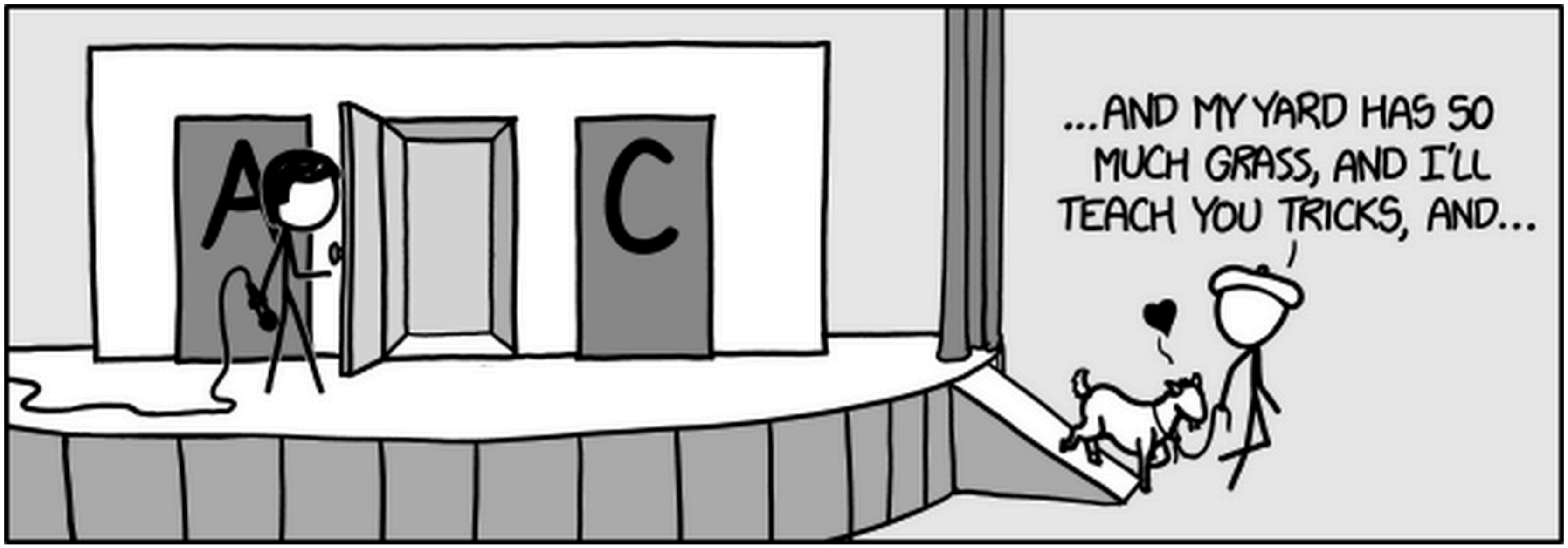
$$\theta \mapsto f(\theta|x) = \frac{f(x|\theta)\, g(\theta)}{\int_{\vartheta \in \Theta} f(x|\vartheta)\, g(\vartheta)\, d\vartheta}$$



$$\hat{\theta}_{map} = \underset{\theta \in \Theta}{argmax} \frac{f(x|\theta)\, g(\theta)}{\int_{\vartheta} f(x|\vartheta)\, g(\vartheta)\, d\vartheta} = \underset{\theta \in \Theta}{argmax} f(x|\theta)\, g(\theta).$$



Host reveals
Goat A
or
Host reveals
Goat B
Host must
reveal Goat B
Host must
reveal Goat A


...AND MY YARD HAS SO MUCH GRASS, AND I'LL TEACH YOU TRICKS, AND...

Questions

- What's power?
  - How does it relate to the Type II error?
  - What happens as n increases? How can we calculate n?
- Revisiting MOM and MLE
  - What do they solve for?
  - How does each approach tackle the problem?
- What's MAP?
  - How does it relate to the MLE?
- Frequentist vs. Bayesian?
- Bayesian: Write out equation for posterior distribution as function of prior and likelihood
  - In layman terms, what's going on here?

Questions

- What's power? Probability that the test correctly rejects the null when alt is true
  - How does it relate to the Type II error? 1-(Type II Error)
  - What happens as n increases? How can we calculate n? See Slide 11
- Revisiting MOM and MLE
  - What do they solve for? Parameter Estimation
  - How does each approach tackle the problem?
    - Both assume a specific distribution already.
    - MOM uses moment matching to get at parameters
    - MLE asks what parameter would maximize the the likelihood of the resulting data
- What's MAP?
  - How does it relate to the MLE? Similar to MLE, but need to account for Prior
- Frequentist vs. Bayesian? Long-term frequency vs. Subjective Belief
- Bayesian: Write out equation for posterior distribution as function of prior and likelihood See Slide 29
  - In layman terms, what's going on here? Updating belief (Prior) with data (Likelihood)