Power
&
Bayesian Inference

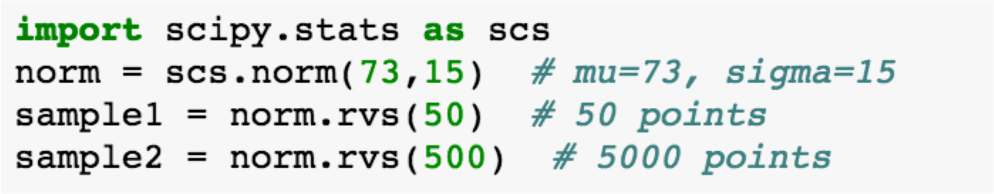
Overview

- **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
  - Coin flip example

Some Hypothesis Tests

- Two samples:



```
import scipy.stats as scs
norm = scs.norm(73,15)  # mu=73, sigma=15
sample1 = norm.rvs(50)  # 50 points
sample2 = norm.rvs(500)  # 5000 points
```

- ➢ Both samples come from the same underlying distribution
- ➢ We can perform various hypothesis tests on the two data sets---with varying outcomes

First Test

- **H0**: The mean of the population is 72
- **H1**: The mean of the population is different from 72 (two-sided)
- **Significance**: alpha = .05
- **Data** = Sample1 (50 points)

RESULTS

Sample mean = 73.06

Sample Variance = 127, Standard Error = 1.60

95% Confidence Interval = (69.9, 76.2)

FAIL TO REJECT NULL

Second Test

- **H0**: The mean of the population is 72
- **H1**: The mean of the population is different from 72 (two-sided)
- **Significance**: alpha = .05
- **Data** = *Sample2* (500 points)

RESULTS

Sample mean = 73.4

Sample Variance = 227, Standard Error = 0.67

95% Confidence Interval = (72.1, 74.8)

REJECT NULL

Third Test

- **H0**: The mean of the population is 72
- **H1**: The mean of the population is different from 72 (two-sided)
- **Significance**: *alpha = .01*
- **Data** = Sample2 (500 points)

RESULTS

Sample mean = 73.5

Sample Variance = 227, Standard Error = 0.67

99% Confidence Interval = (71.7, 75.2)

FAIL TO REJECT NULL

Fourth Test

- ***H0****: The mean of the population is 70*
- ***H1****: The mean of the population is different from 70 (two-sided)*
- **Significance**: alpha = .01
- **Data** = Sample2 (5000 points)

RESULTS

Sample mean = 73.5

Sample Variance = 227, Standard Error = 0.67

99% Confidence Interval = (71.7, 75.2)

REJECT NULL

Hypothesis Testing Results

- Four different tests with varying results
- Even though there is only one *ground truth* (Underlying population distribution)
- Factors that influence the outcome:
  - Sample Size
  - Desired Confidence level
  - Effect size

Powerful Test

- Which test do we like better?

TEST 1

$\alpha = 0.05$

"powerfulness" = 0.8

TEST 2

$\alpha = 0.05$

"powerfulness" = 0.3

Statistical Power

- We define the power of a hypothesis test as the probability that the test correctly rejects the null hypothesis (H0) when the alternative hypothesis (H1) is true

- Power = P(reject H0 | H1 is true)

  = P(accept H1 | H1 is true)

- How much chance do we have to reject the null hypothesis when the alternative is in fact true? (what's the probability of detecting a real effect?)

Find the Power

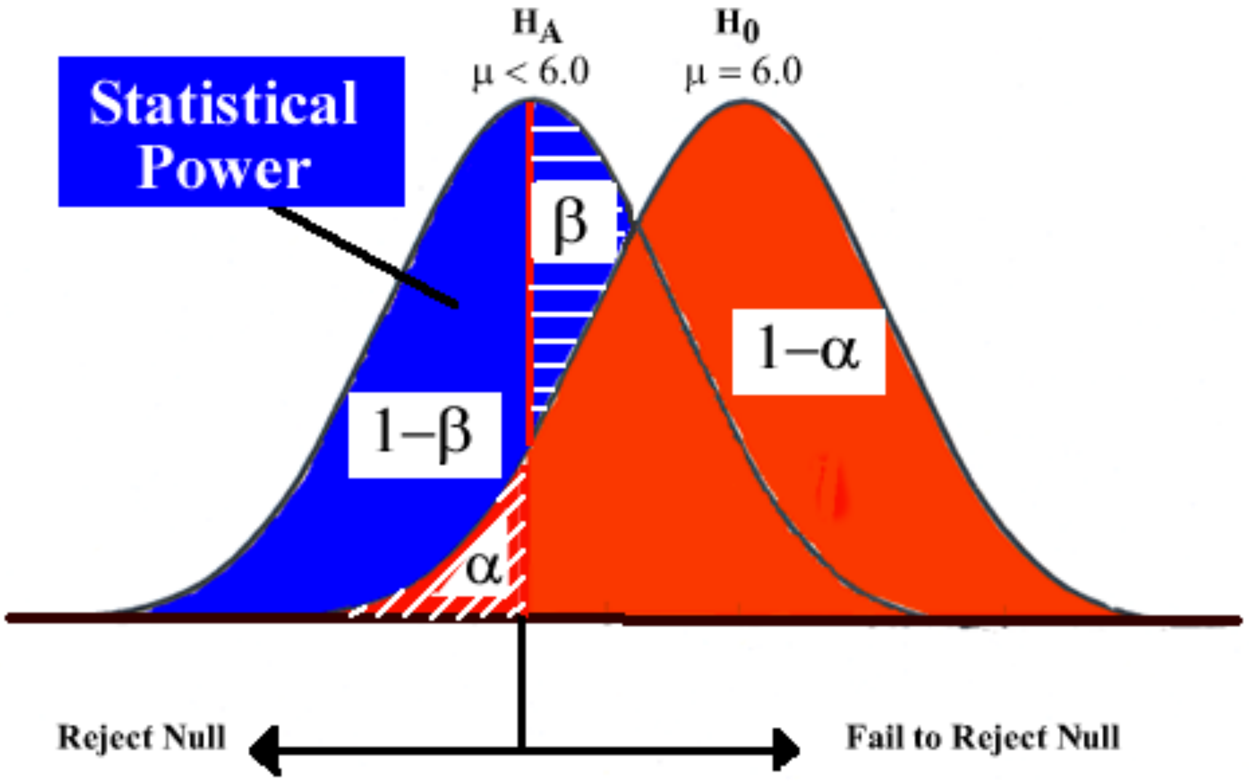
	Fail to reject the Null Hypothesis	Reject the Null Hypothesis
Null Hypothesis is true	✓	Type I Error
Alternative Hypothesis is true	Type II Error	✓

Type I Error – We see an effect that isn't really there.
Type II Error – The effect is there, but we don't see it.

Power = 1 – P(Type II Error)

Factors Influencing Power

1. Size of the effect ↑
2. Standard deviation of the characteristic ↓
3. Bigger sample size ↑
4. Significance level desired ↓



Calculating the Power

- β = P(fail to reject H0 | H0 is false)
  = P(fail to reject H0 | H1 is true)
  = P(Type II error)
- Power = 1 - β
- First, we want to find the value, under the null distribution, beyond which we would reject the null (H0)

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

- Then we find the corresponding tail probability of this value under the alternative distribution

$$power = P(X_1 > X^*) = P(Z > \frac{X^* - \mu_1}{s/\sqrt{n}})$$

- Note: we will replace “>” with “<“ in the power calculation above if the alternative distribution is to the left of the null distribution

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 fixed significance level (α)
  - An estimate of the population mean
  - An estimate of the population standard deviation
  - A desired power
- 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 the same sign

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.
  1. Conduct pilot experiment
  2. This gives you estimates of effect size and sample standard deviation, which you can plug into the sample size calculator
  3. Run the experiment with the computed sample size.

Recap: Power Calculation

- Decide the critical value for the test statistic, in general,
  - Z* = ± 1.96 for two-sided test
  - Z* = + 1.64 or – 1.64 for one-sided test
- Calculate the corresponding value under the null distribution
- 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

Two distinctive features of Bayesian Statistics

1. Bayesians incorporate *prior knowledge* into predictions using Bayes' rule:

$$p(\theta|x) \propto P(x|\theta) \cdot P(\theta)$$

2. Bayesians use the word probability to mean *degree of belief* rather than *long-run limiting frequency.*

Frequentist vs. Bayesian:

Incorporating prior evidence

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:

Incorporating prior evidence

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:

Probability as degree of belief

Bayesian: “I’d say the probability is about .8 that there is life on other planets”

Frequentist: “The probability is either 1 or 0. There either is or there isn’t”

Frequentist vs. Bayesian:

Probability as degree of belief

On reporting the value of an unknown parameter…

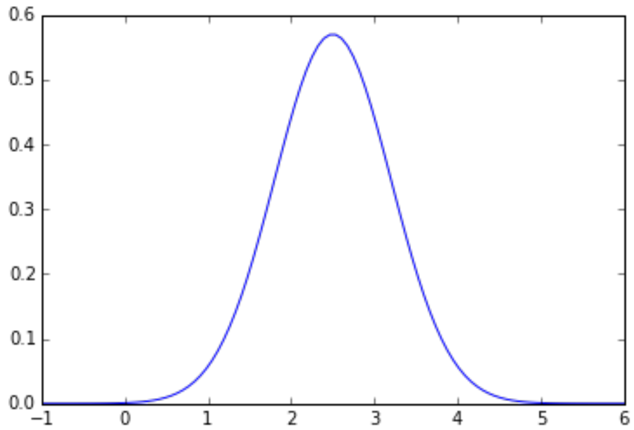
Frequentist:

A 95% confidence interval for the mean of the population is: [2.1, 2.9].

That is, 95 out of 100 times, when a sample is taken with these observed properties, the true population mean will be in the above interval.

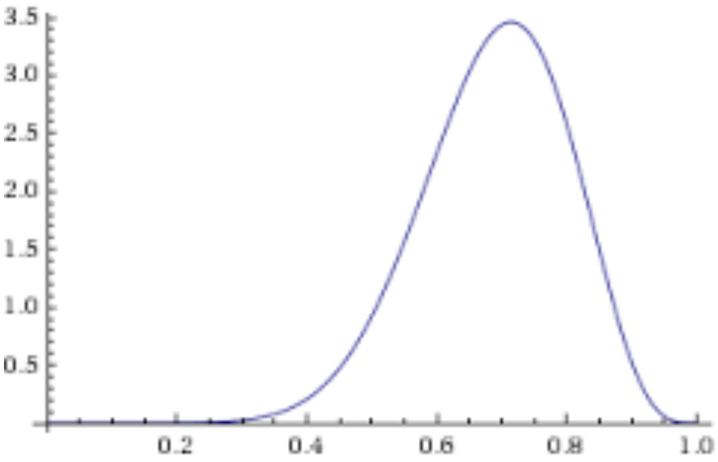
Bayesian:

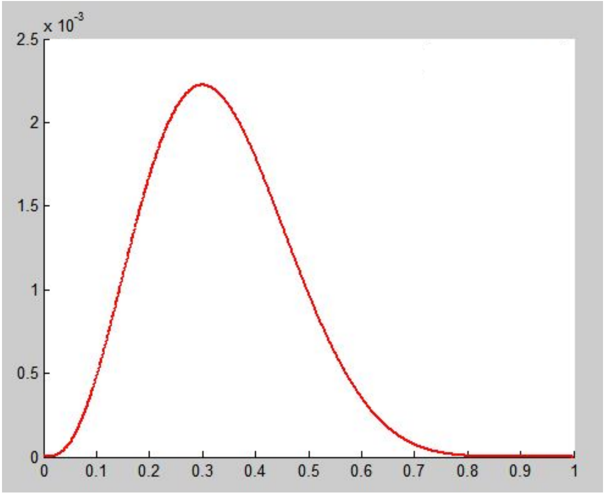
The population mean is:



Bayesian Inference

- Bayesian inference takes both distinctions into account. Report outcome as a probability, and update belief as new evidence comes in
- 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:

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

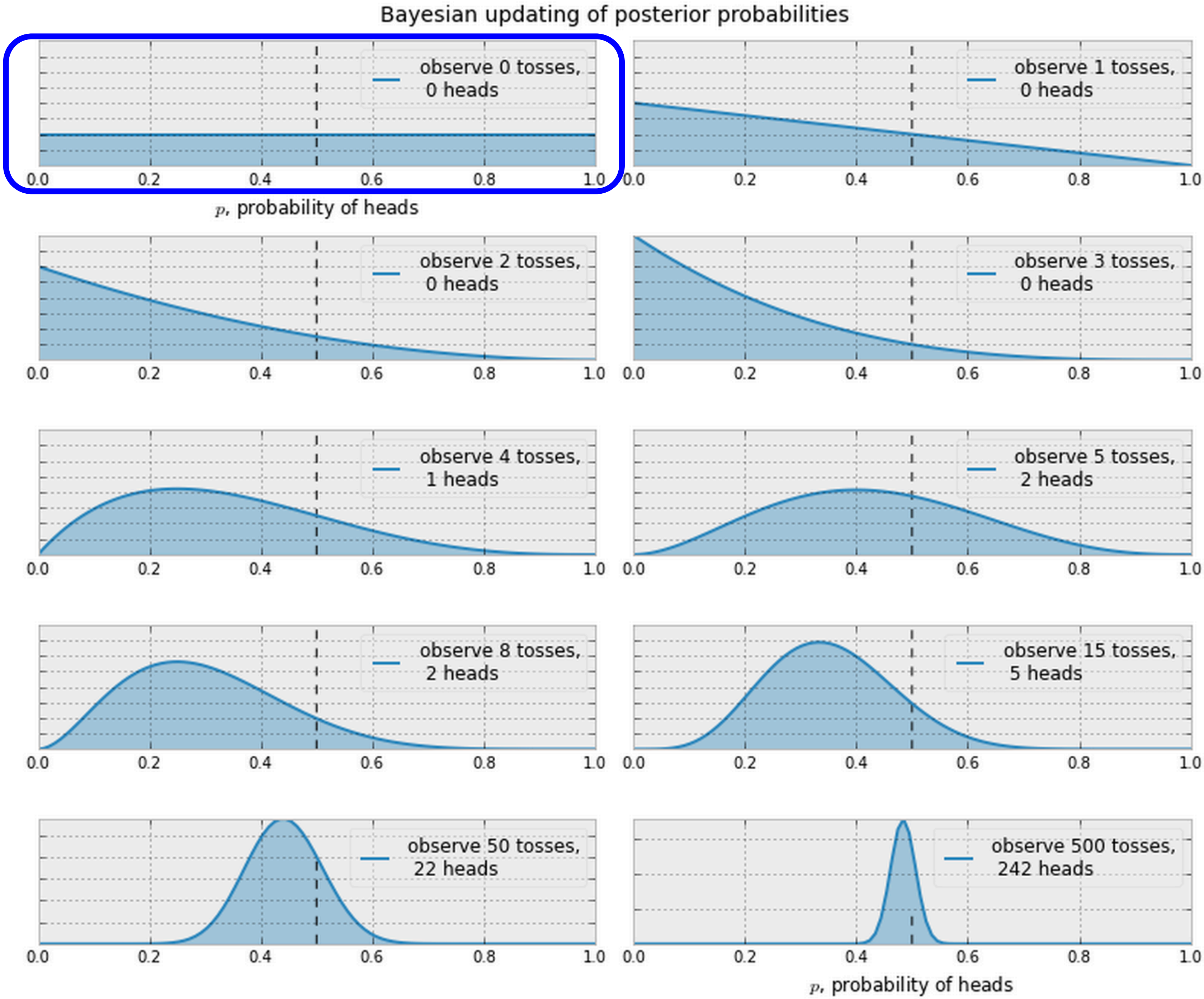
Posterior Distribution:

$$\pi(\theta|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\theta)\pi(\theta)}{\int_{\theta\in\Theta} f(\boldsymbol{x}|\theta)\pi(\theta)d\theta}$$

- prior $\pi(\theta)$ describes our current knowledge about θ
- likelihood $f(\boldsymbol{x}|\theta)$ is the distribution of the data for a given θ
- posterior $\pi(\theta|\boldsymbol{x})$ is our updated knowledge about θ after seeing the data

Coin Flip Example

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
0.0
0.2
0.4
0.6
0.8
1.0
p, probability of heads
p, probability of heads

Bayesian Inference

1. Choose a distribution for θ that incorporates prior knowledge. *i.e. Start with your beliefs*.
2. Choose a statistical model to calculate the likelihood of the data, given θ.
3. Calculate the posterior distribution for θ, given the data. *i.e. update your beliefs, based on the evidence.*
4. Take the posterior to be the prior in the next iteration*. i.e. start over with your new belief*