

Power Calculation

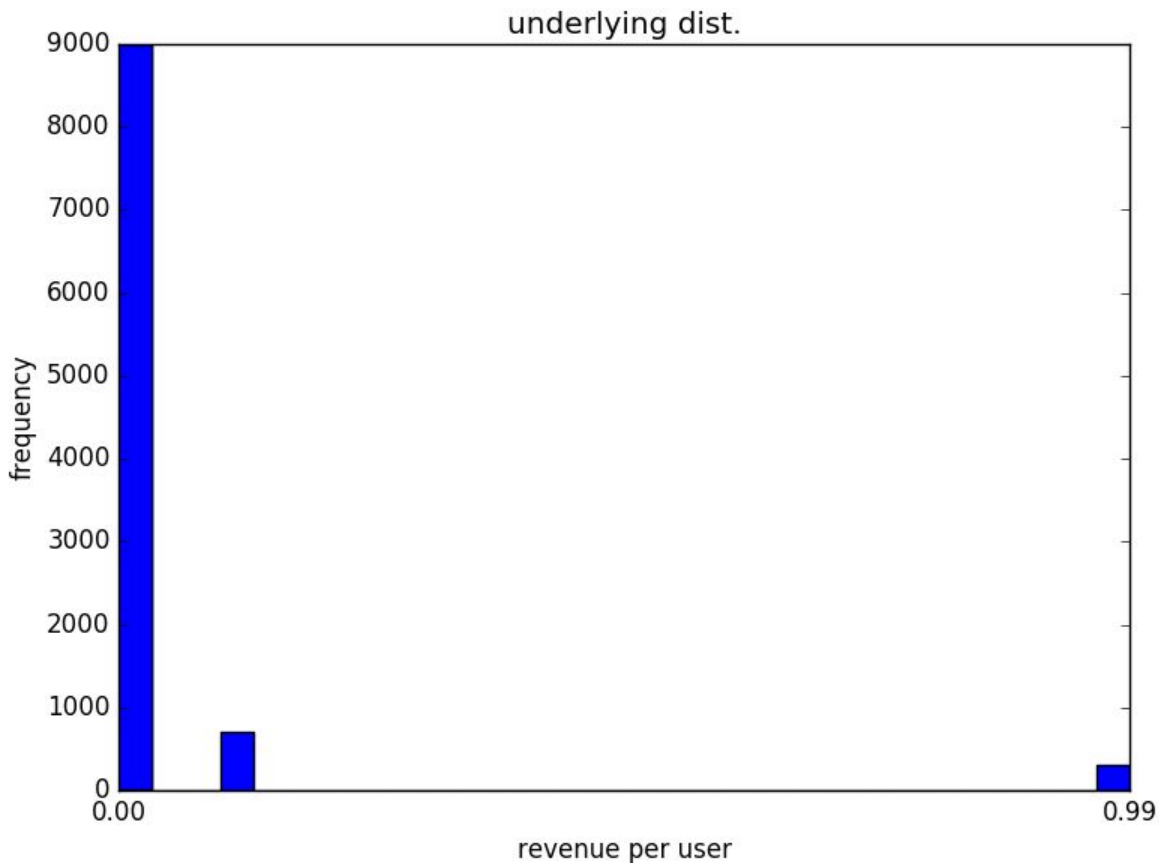
Natalie Hunt



1. Review:
 - a. Central Limit Theorem
 - b. Hypothesis Testing
2. Type I vs Type II errors
3. What is “Power”?
4. Calculating Power / Sample Size
5. A/B Testing w/ Power

Underlying Distribution:

Random variable: <i>X = revenue per visitor</i>	P(X):
X = \$0.00 (no revenue)	90%
X = \$0.10 (ad-click)	7%
X = \$0.99 (app purchase)	3%

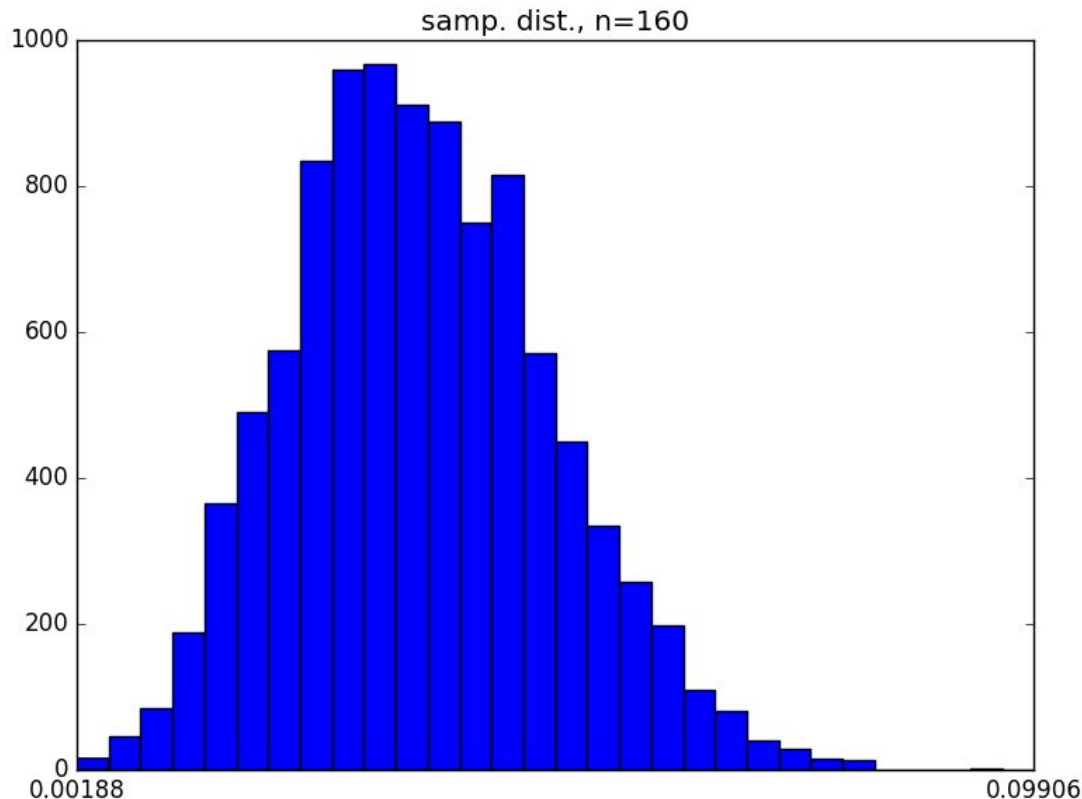


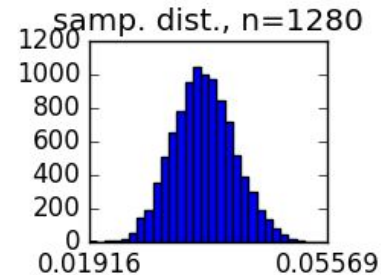
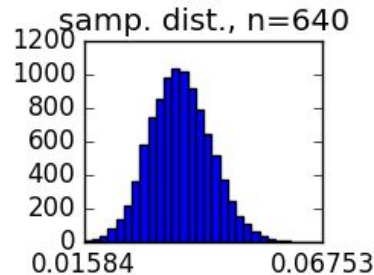
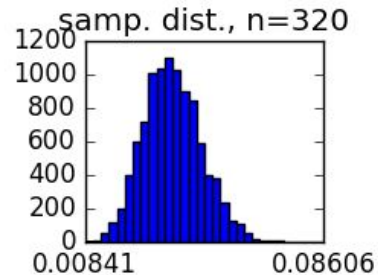
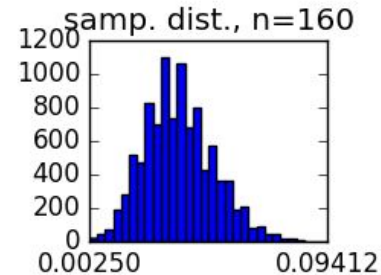
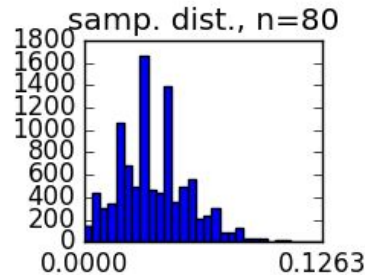
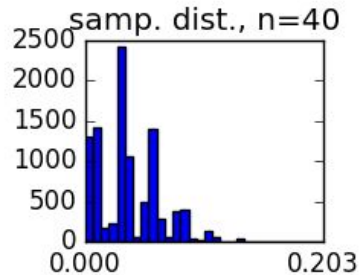
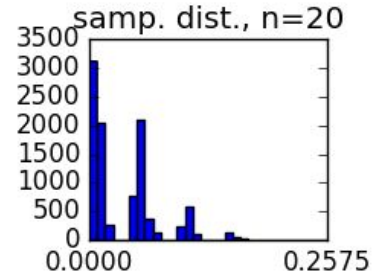
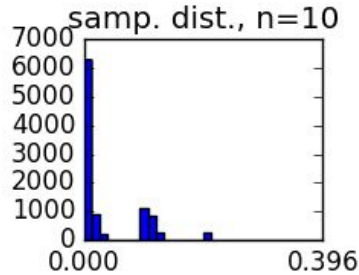
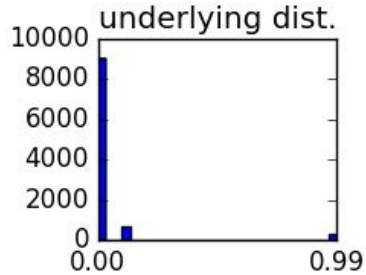
Collect n samples from the website revenue distribution, calculate the sample mean \bar{x}

Repeat 10,000 times, we get:

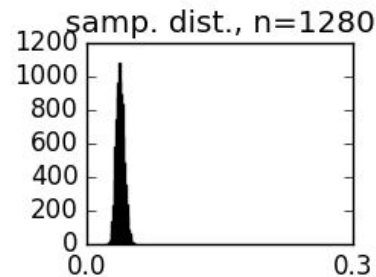
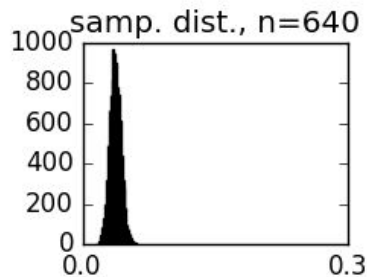
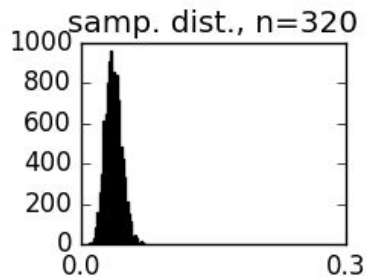
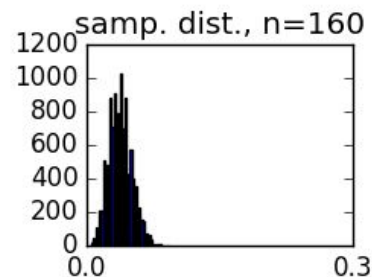
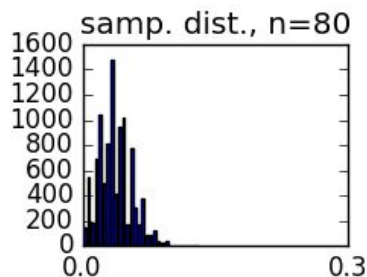
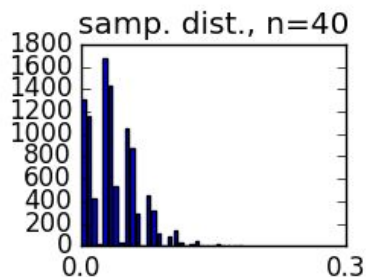
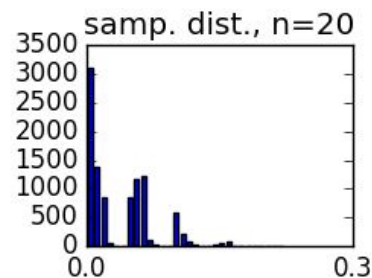
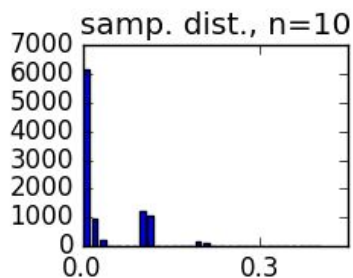
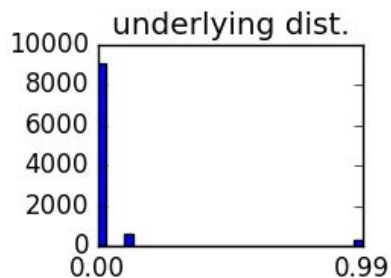
$\bar{x}_0, \bar{x}_1, \dots, \bar{x}_{9999}$

Plot all 10,000 sample means.





Central Limit Theorem: What happens when the sample size increases?



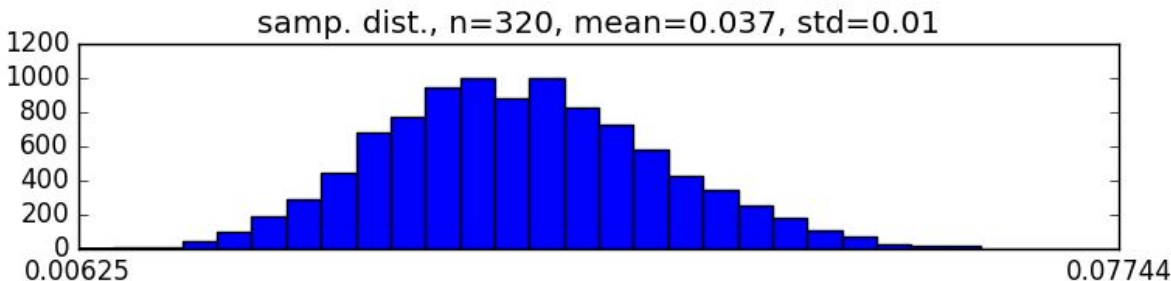
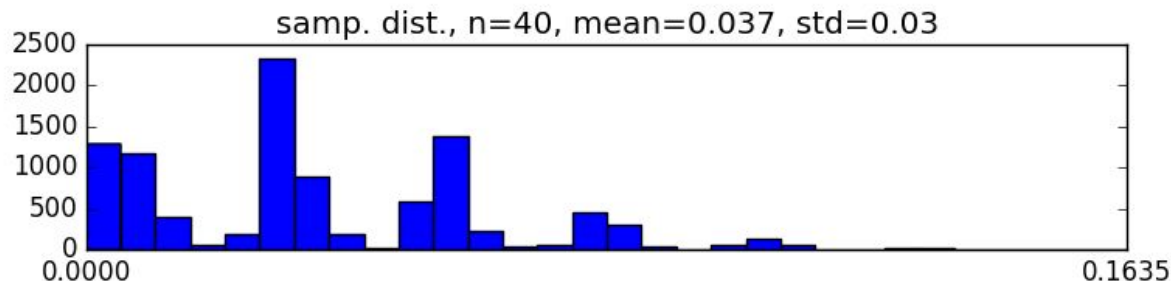
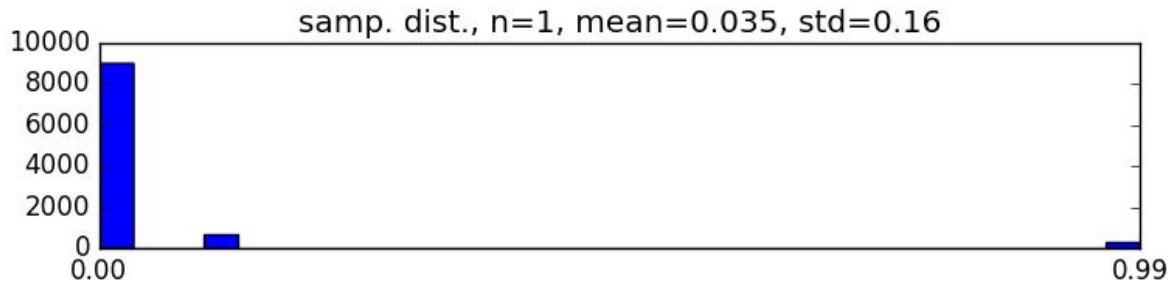
Let the underlying distribution have mean and std. dev.

μ and σ

The sampling distribution's mean and std. dev. will equal:

$$\mu' = \mu$$

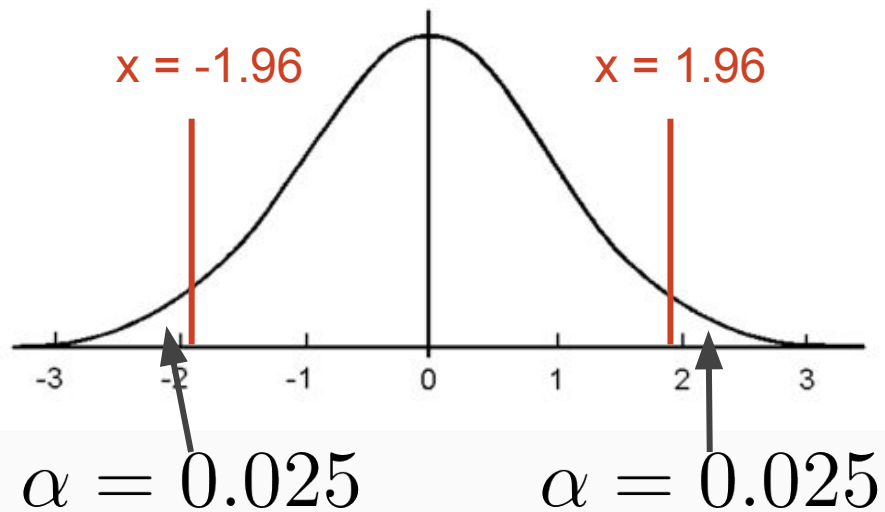
$$\sigma' = \sigma / \sqrt{n}$$



Two-sided test:

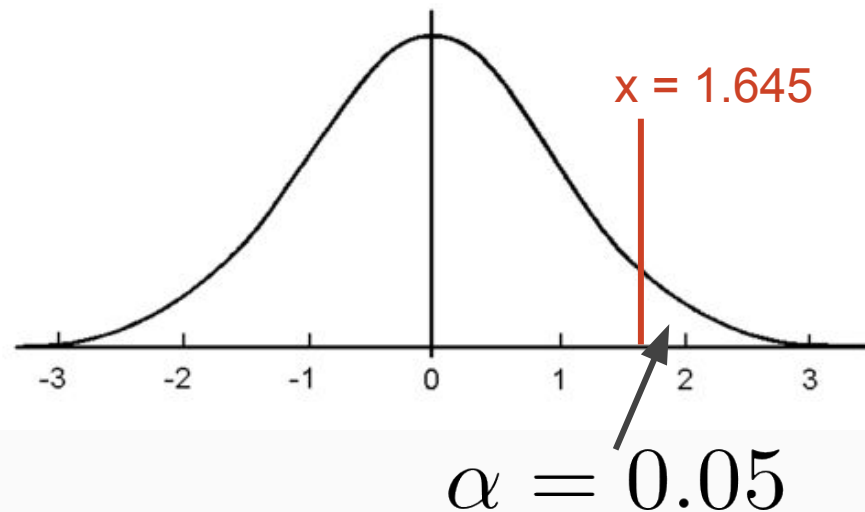
$$H_0 : \mu = 0$$

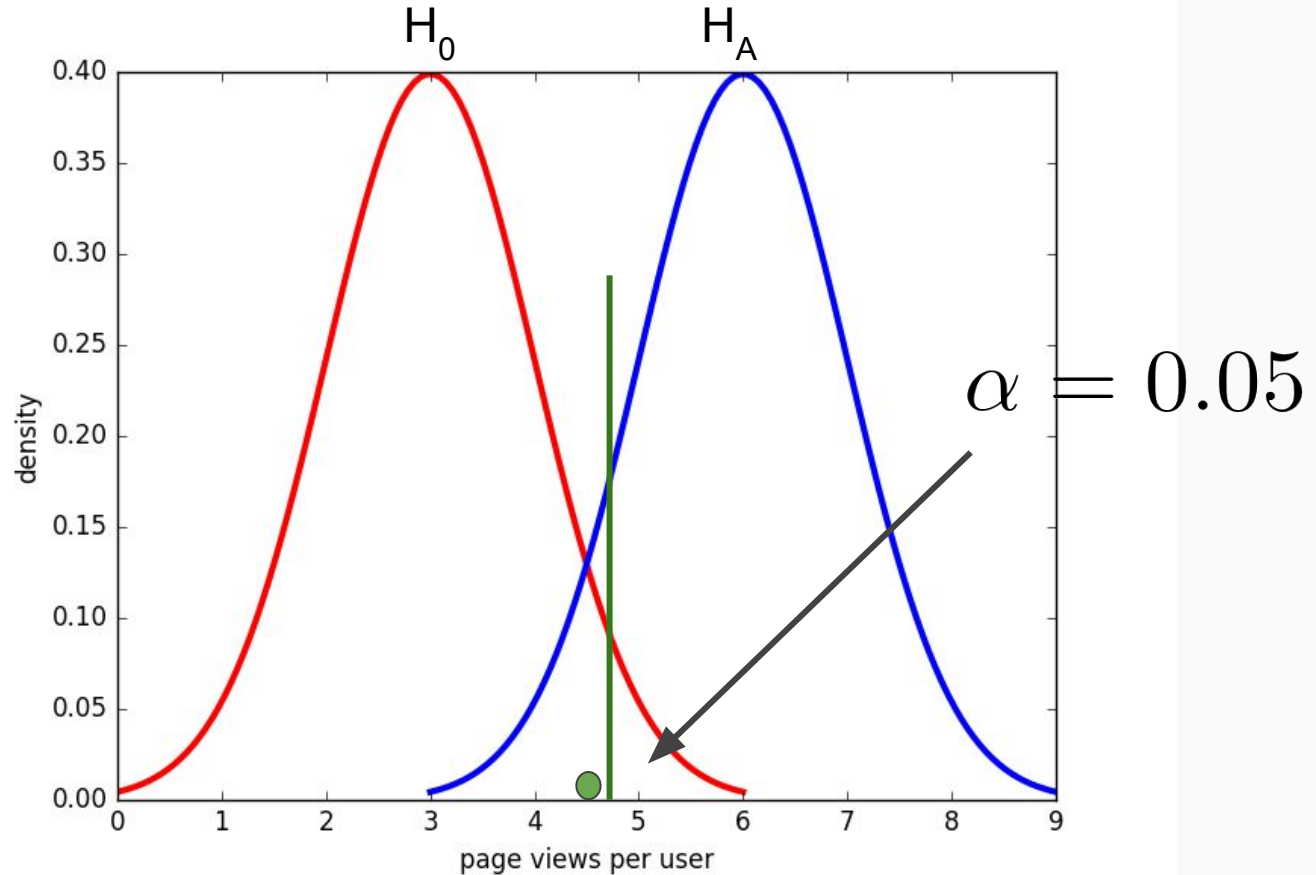
$$H_A : \mu \neq 0$$

**One-sided test:**


$$H_0 : \mu = 0$$

$$H_A : \mu > 0$$





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

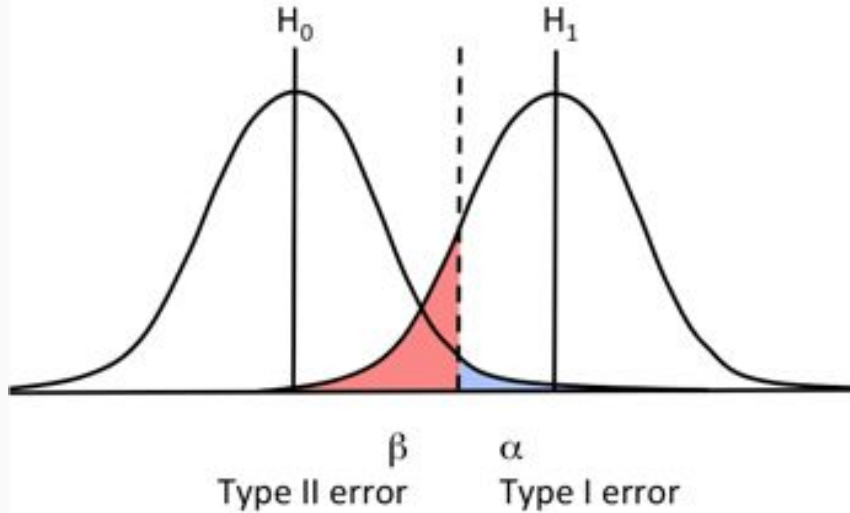


We call this the experiment's "Power". It is the probability that we **correctly reject H_0** when the null hypothesis is false.

	H_0 is true true -	H_0 is false true +
Accept H_0 predict -	Correct Decision ($1-\alpha$)	Type II Error (β)
Reject H_0 predict +	Type I Error (α)	Correction Decision ($1-\beta$)

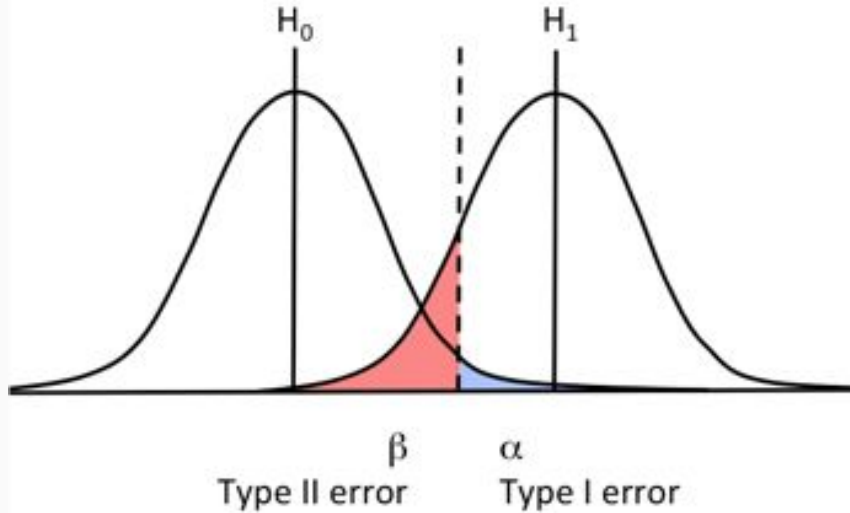
false positive rate
(aka, 1 - specificity)

true positive rate
(aka, sensitivity)



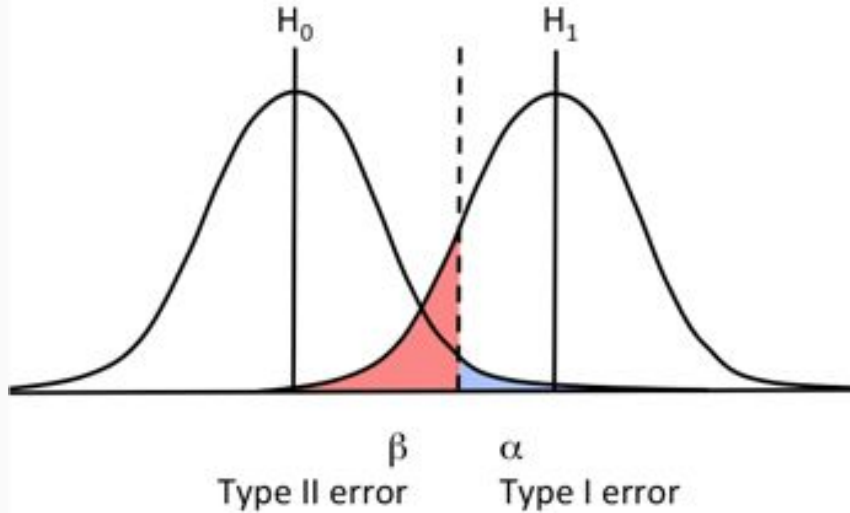
	H_0 is true	H_0 is false
Accept H_0	Correct Decision ($1-\alpha$)	Type II Error (β)
Reject H_0	Type I Error (α)	Correction Decision ($1-\beta$)

The *power* measurement is in relationship to a specific alternative hypothesis. Think of it as the *power* to detect a particular “effect size”.



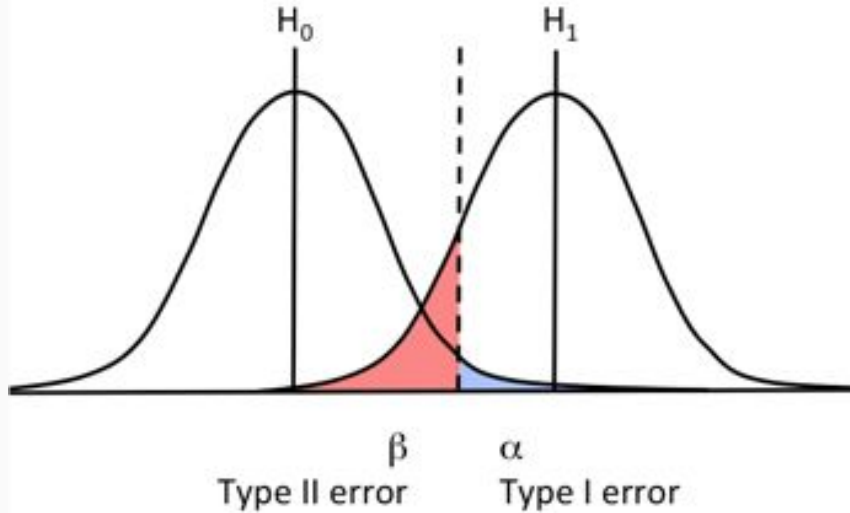
	H_0 is true	H_0 is false
Accept H_0	Correct Decision ($1-\alpha$)	Type II Error (β)
Reject H_0	Type I Error (α)	Correction Decision ($1-\beta$)

What happens to *power* when we increase alpha?



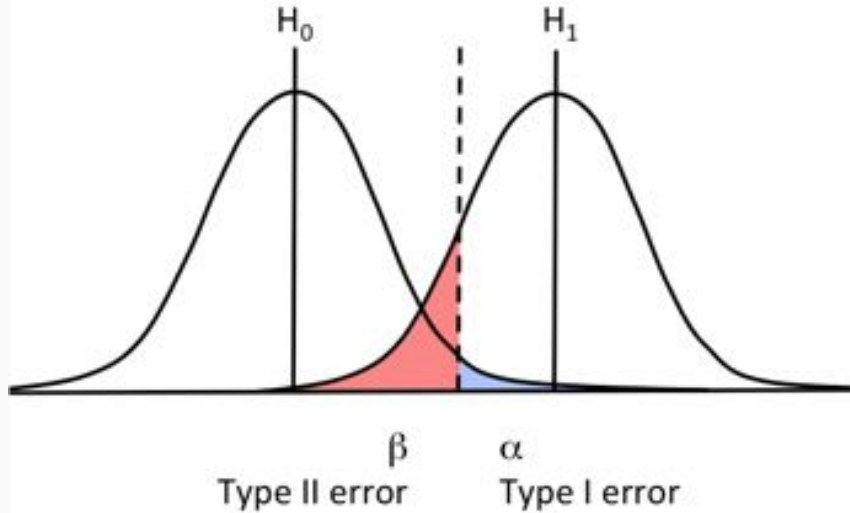
	H_0 is true	H_0 is false
Accept H_0	Correct Decision ($1-\alpha$)	Type II Error (β)
Reject H_0	Type I Error (α)	Correction Decision ($1-\beta$)

What happens to *power* when we increase the effect size?



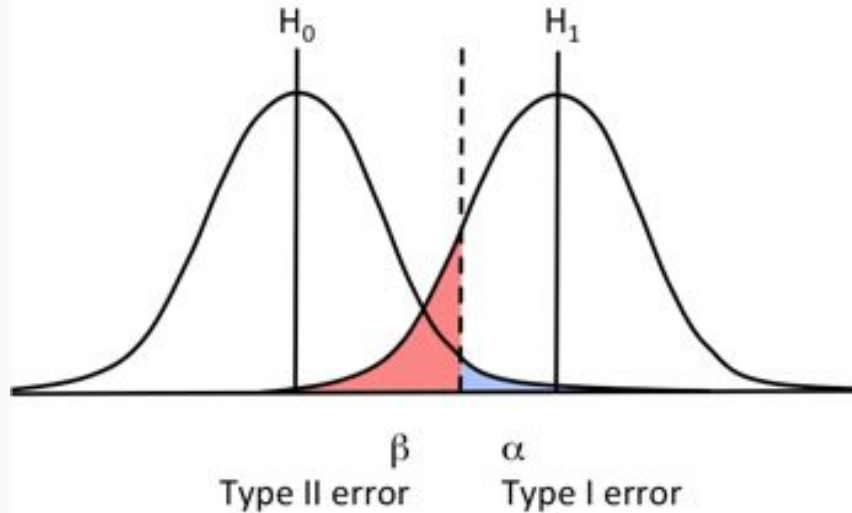
	H_0 is true	H_0 is false
Accept H_0	Correct Decision ($1-\alpha$)	Type II Error (β)
Reject H_0	Type I Error (α)	Correction Decision ($1-\beta$)

What happens to *power* when we increase the sample std. deviation?



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

What happens to *power* when we increase the sample size?

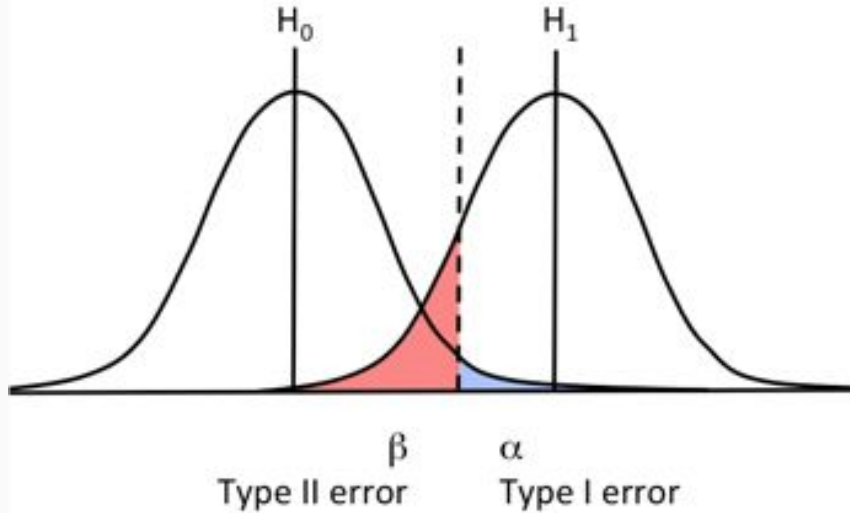


	H_0 is true	H_0 is false
Accept H_0	Correct Decision ($1-\alpha$)	Type II Error (β)
Reject H_0	Type I Error (α)	Correction Decision ($1-\beta$)

Often, we know:

1. The “effect size” that we want to detect, and
2. The *power* that we want to achieve.

We then calculate the *sample size* needed to get what we want!



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

1. Decide to run an experiment, choose α and $(1 - \beta)$
2. Calculate required sample size n
3. Take sample, obtain \bar{x} and s
4. Reject or “fail to reject” H_0

(new steps)

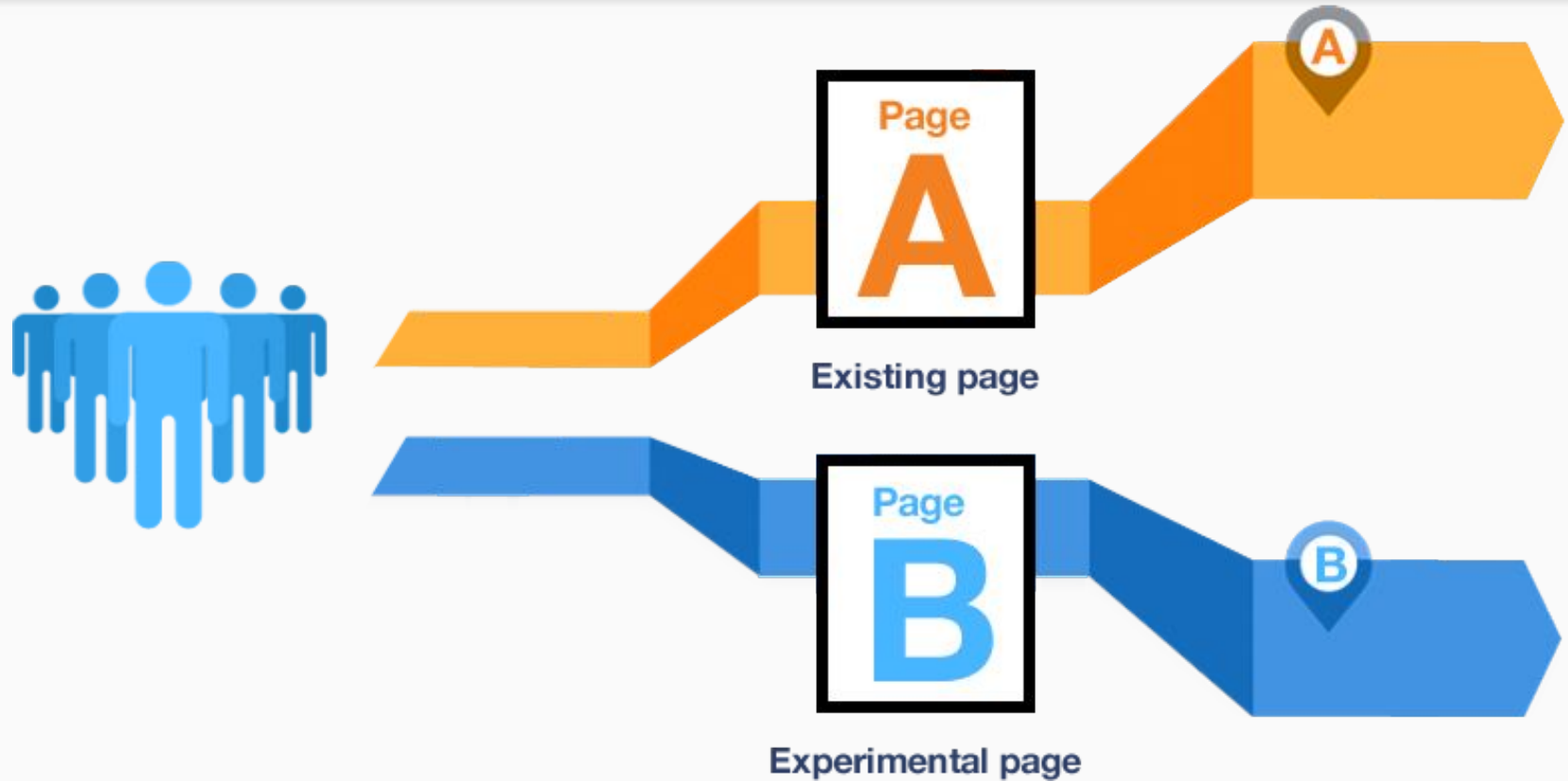
Calculating the required sample size

$$n > \left((Z_{(1-\beta)} - Z_{\alpha}) \frac{s}{\mu_b - \mu_a} \right)^2$$

```
import scipy.stats as st
```

```
st.norm.ppf(alpha)
```

```
st.norm.ppf(1 - beta)
```



Setup: A/B Test our website's homepage.

Our current homepage has a signup conversion rate of 6%. (The standard deviation would be 0.24.)

We want to test a new homepage design to see if we can get a 7% signup rate. We'll want an experiment where alpha is 1% and power is 95%.

How many visitors must visit the new homepage in order to fulfill the requirements of this experiment?

$$n \geq 9,084$$

Setup: A/B Test our website's homepage.

Our current homepage has a signup conversion rate of 1%. (The standard deviation would be 0.099.)

We want to test a new homepage design to see if we can get a 1.2% signup rate. We'll want an experiment where alpha is 1% and power is 95%.

How many visitors must visit the new homepage in order to fulfill the requirements of this experiment?

$$n \geq 39,427$$

Setup: A/B Test our website's homepage.

Our current homepage has a signup conversion rate of 20%. (The standard deviation would be 0.4.)

We want to test a new homepage design to see if we can get a 30% signup rate. We'll want an experiment where alpha is 1% and power is 95%.

How many visitors must visit the new homepage in order to fulfill the requirements of this experiment?

$$n \geq 253$$

Bayesian Inference

1. Frequentists vs. Bayesian
2. Bayes' Rule
3. Prior, likelihood, posterior distributions



What is the probability that it rained in my city last night?

(No info is given about which city I'm currently in.)

$$P(\text{rain}) = 0.1$$

What is the probability that it rained in my city last night given that I'm in Seattle?

$$P(\text{rain}|\text{Seattle}) = 0.65$$

What is the probability that it rained in my city last night?

(No info is given about which city I'm currently in.)

$$P(\text{rain}) = 0.1$$

What is the probability that it rained in my city last night given that I live in Seattle and I see that the road is wet?

$$P(\text{rain} | \text{Seattle, wet roads}) = 0.97$$

Frequentist vs. Bayesian

Frequentist Probability

“Long Run” frequency of an outcome

Subjective Probability

A measure of degree of belief

Bayesians consider both types

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.

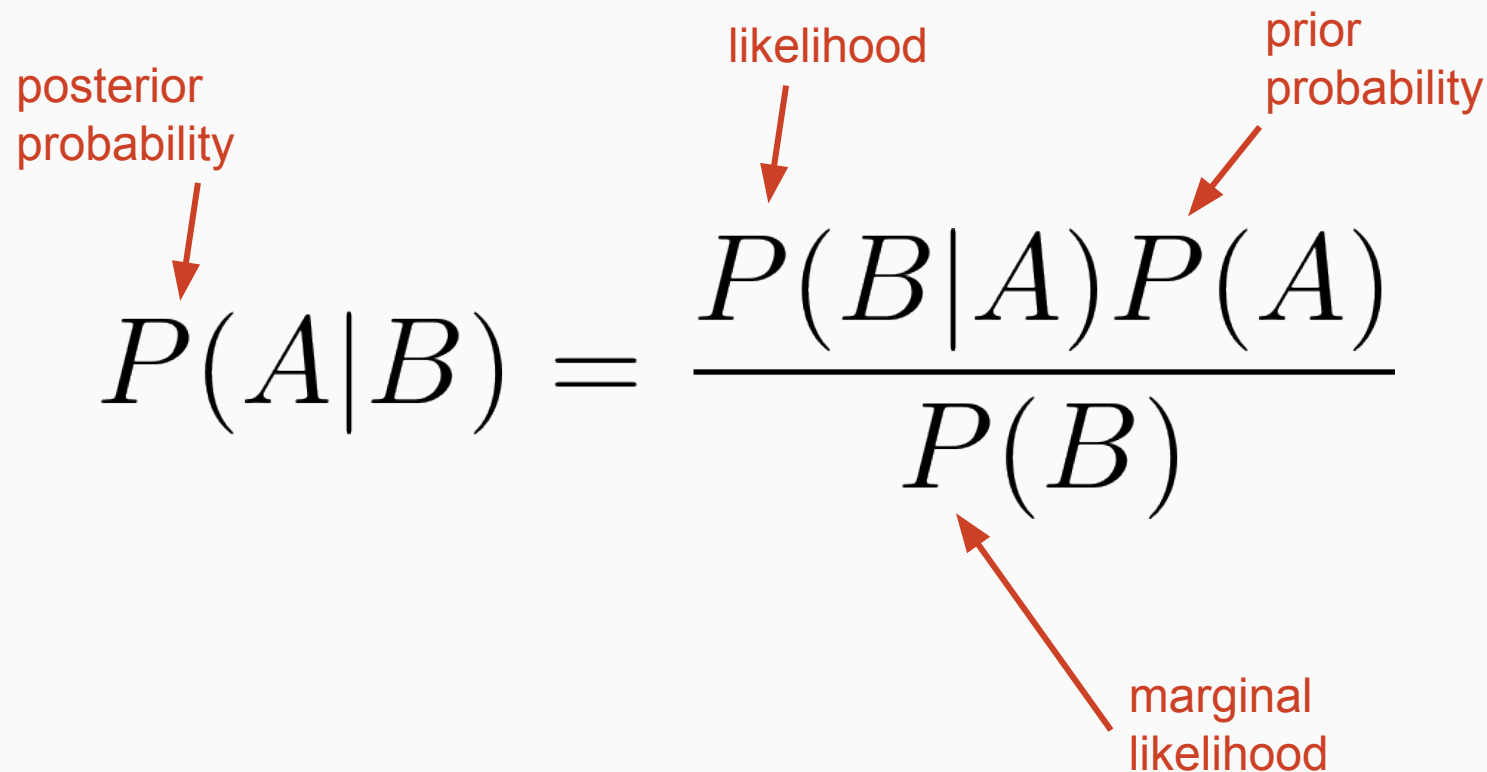




Frequentist: “They’re both so skilled! I have **as much confidence** in musician’s ability to distinguish Haydn and Mozart as I do the drunk’s to predict coin tosses”

Bayesian: “I’m not convinced by the drunken man...”

The Bayesian approach is to incorporate prior knowledge into the experimental results.



The diagram illustrates Bayes' Rule with the following components and annotations:

- posterior probability**: An arrow points to $P(A|B)$.
- likelihood**: An arrow points to $P(B|A)$.
- prior probability**: An arrow points to $P(A)$.
- marginal likelihood**: An arrow points to $P(B)$.

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

$$P(\text{psychic}|\text{correct}) = \frac{P(\text{correct}|\text{psychic})P(\text{psychic})}{P(\text{correct})}$$

$$= \frac{1.0 * 0.0001}{0.5^{10}}$$

$$= 10.2\%$$

Very subjective!



DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)

THIS NEUTRINO DETECTOR MEASURES
WHETHER THE SUN HAS GONE NOVA.

THEN, IT ROLLS TWO DICE. IF THEY
BOTH COME UP SIX, IT LIES TO US.
OTHERWISE, IT TELLS THE TRUTH.

LET'S TRY.
DETECTOR! HAS THE
SUN GONE NOVA?



FREQUENTIST STATISTICIAN:

THE PROBABILITY OF THIS RESULT
HAPPENING BY CHANCE IS $\frac{1}{36} = 0.027$.
SINCE $p < 0.05$, I CONCLUDE
THAT THE SUN HAS EXPLODED.



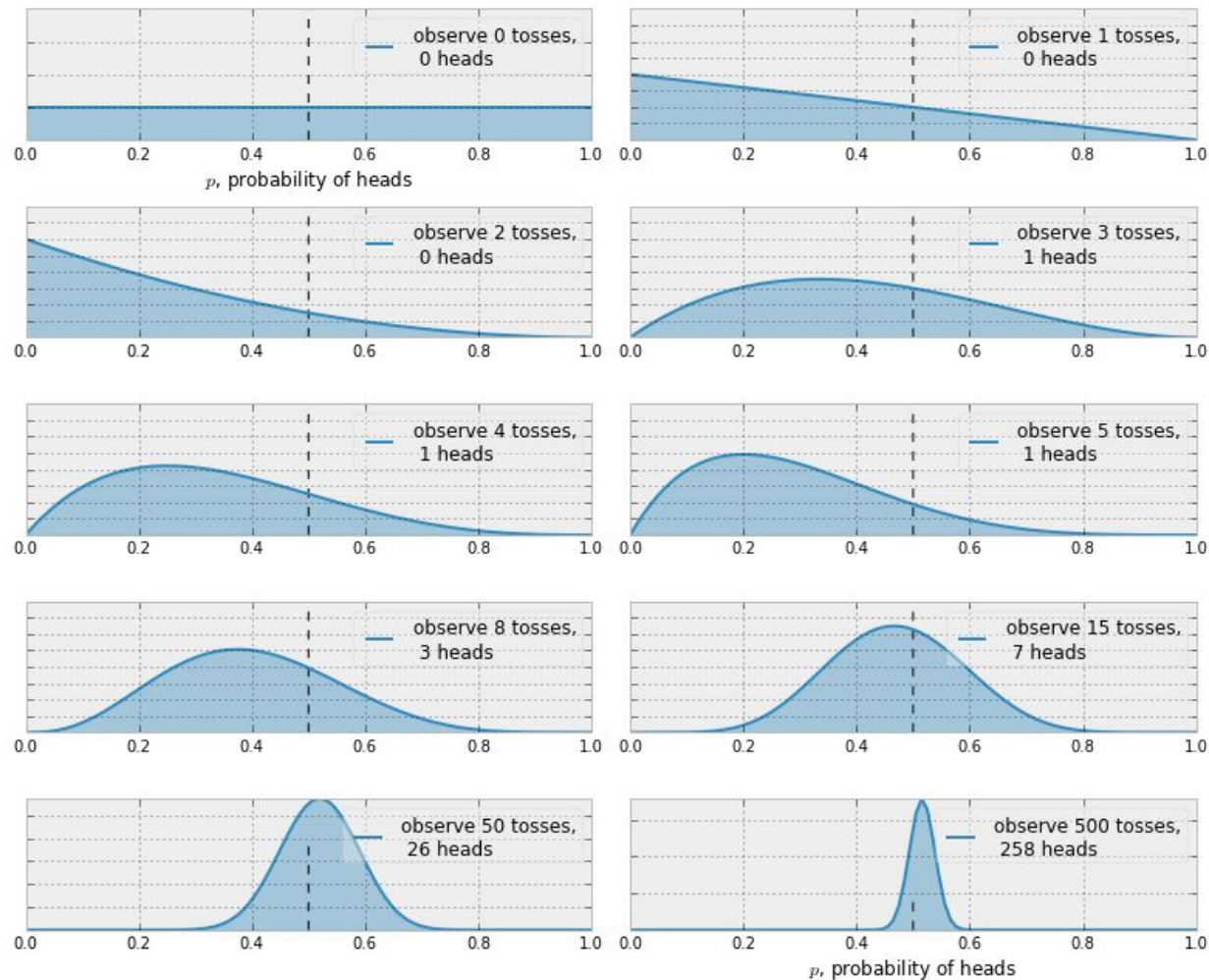
BAYESIAN STATISTICIAN:

BET YOU \$50
IT HASN'T.



Bayesian Updates

Bayesian updating of posterior probabilities



Monty Hall Problem

