Power Calculation

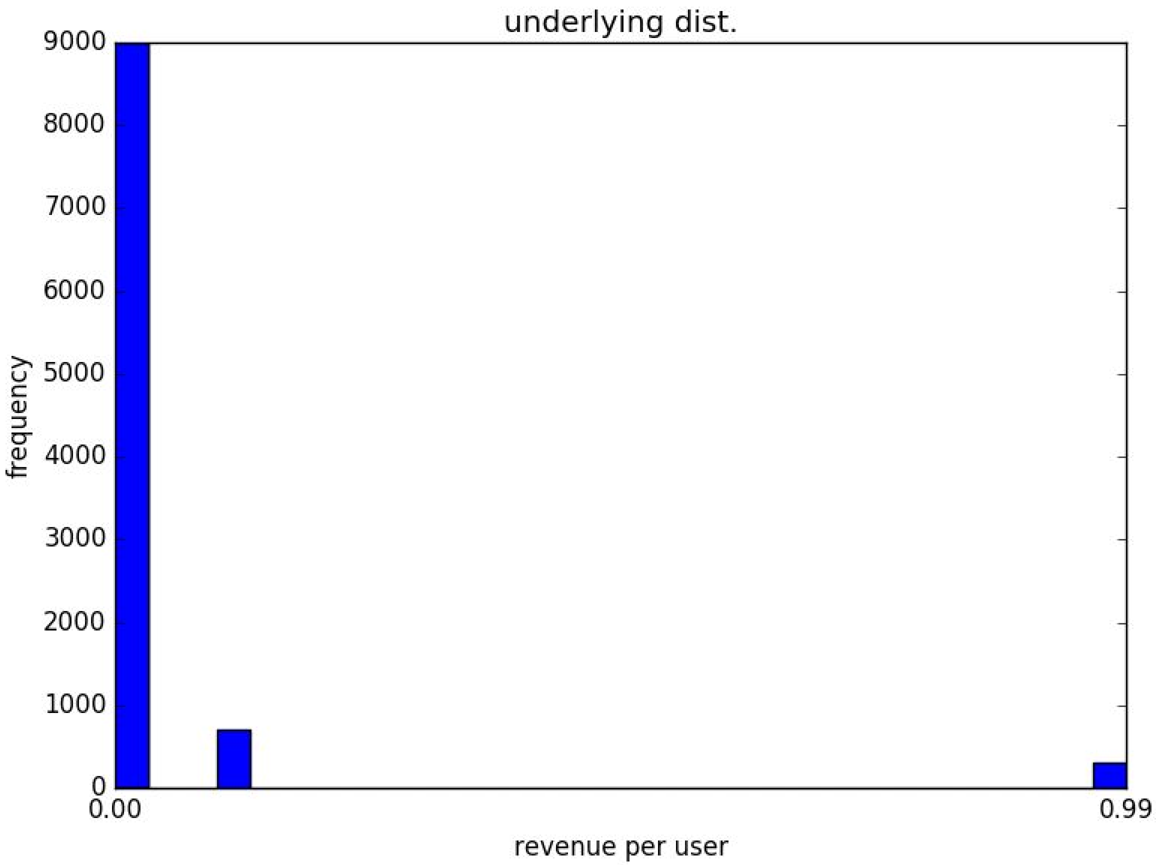
Ryan Henning

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

Distribution of website revenue per visitor

Underlying Distribution:

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



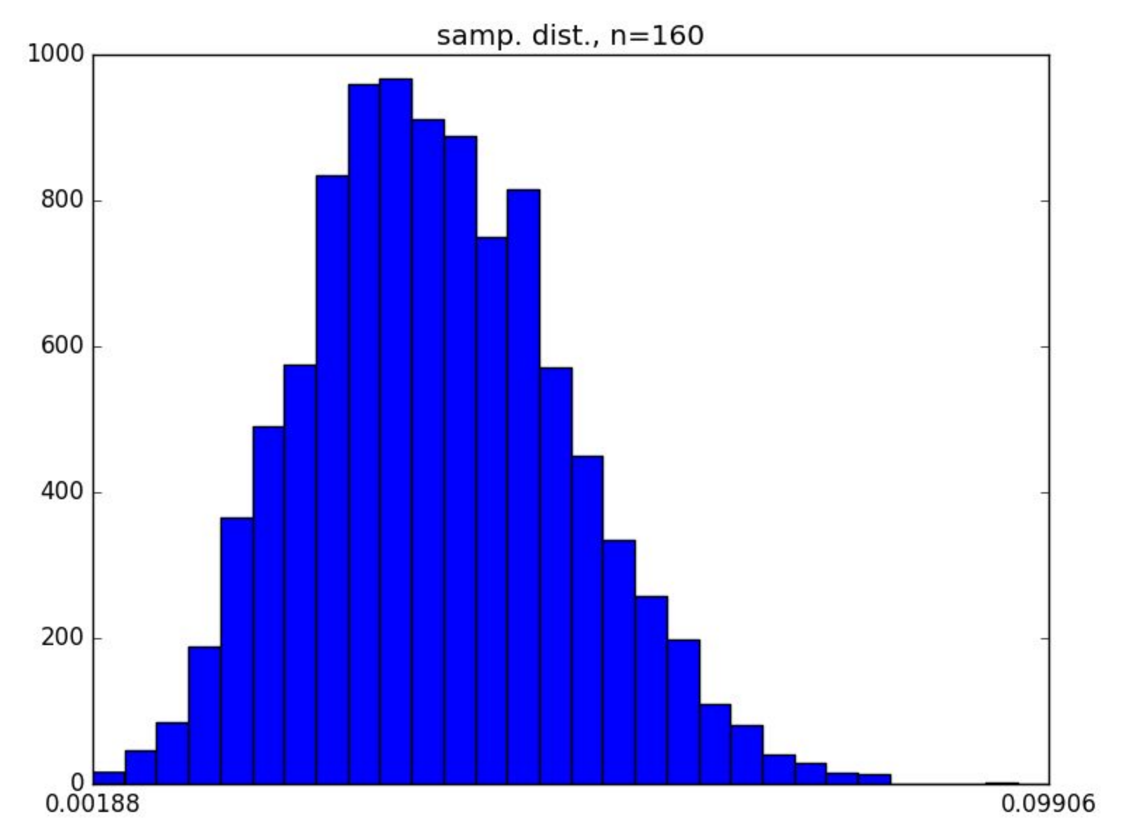
Distribution of sample means

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

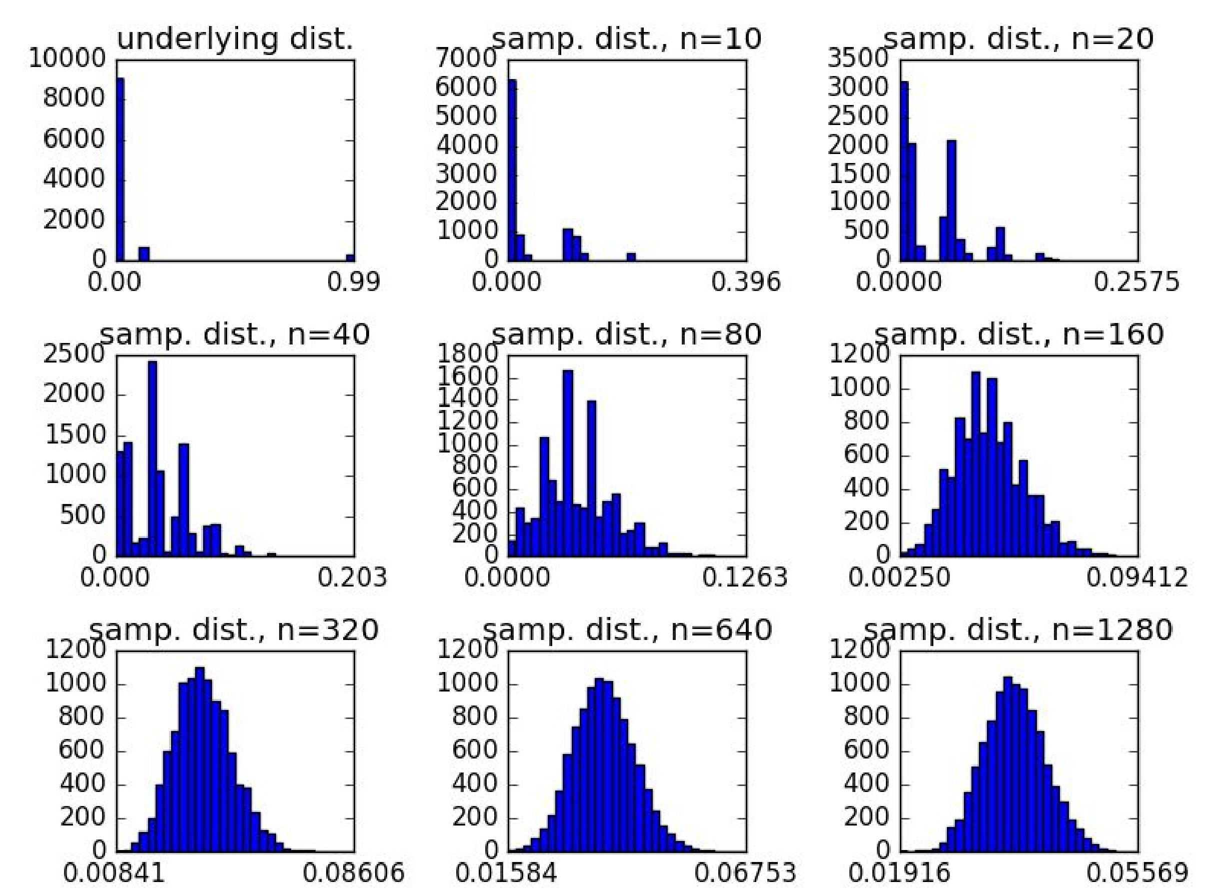
Repeat 10,000 times, we get:

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

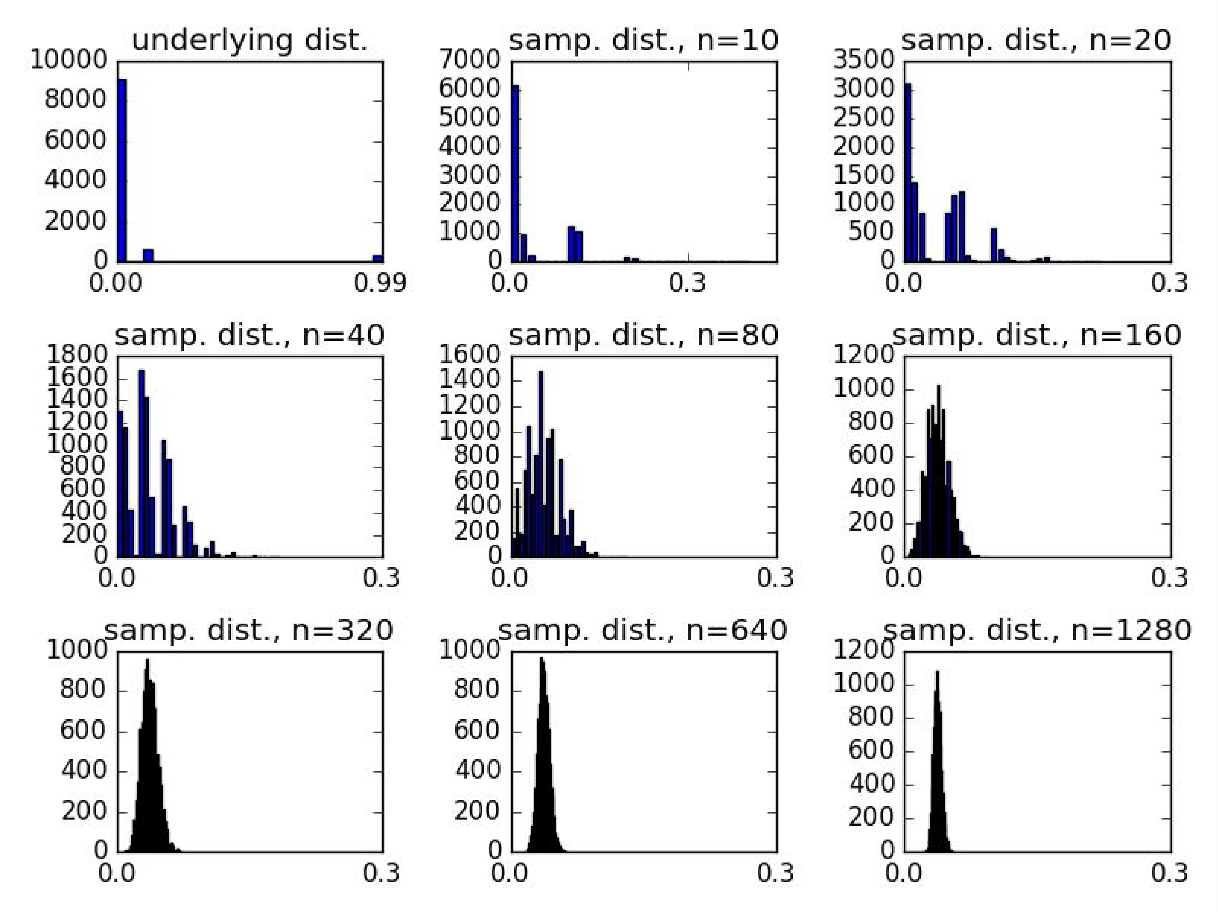
Plot all 10,000 sample means.



Central Limit Theorem



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



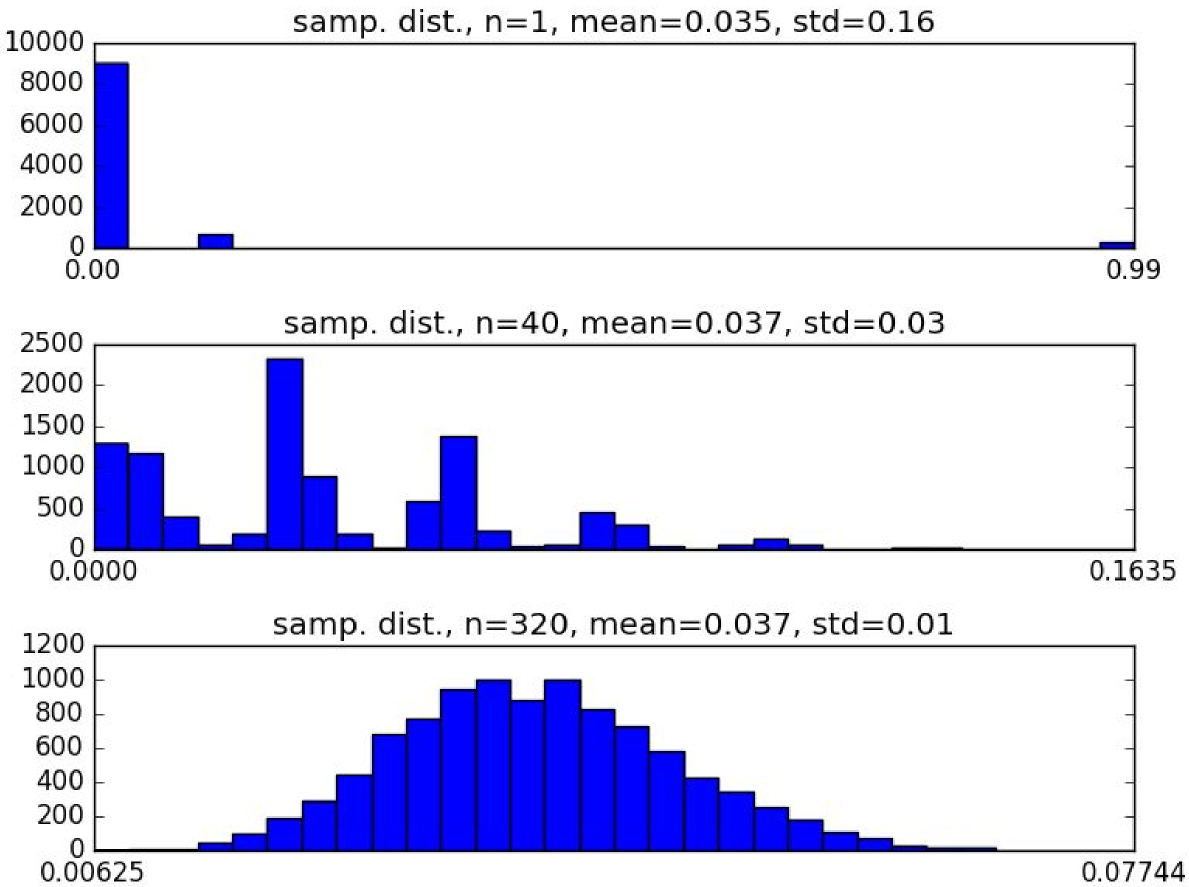
Central Limit Theorem: Std. Dev precise relationship to sample mean

Let the underlying distribution have mean and std. dev.

$$\mu \text{ and } \sigma$$

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

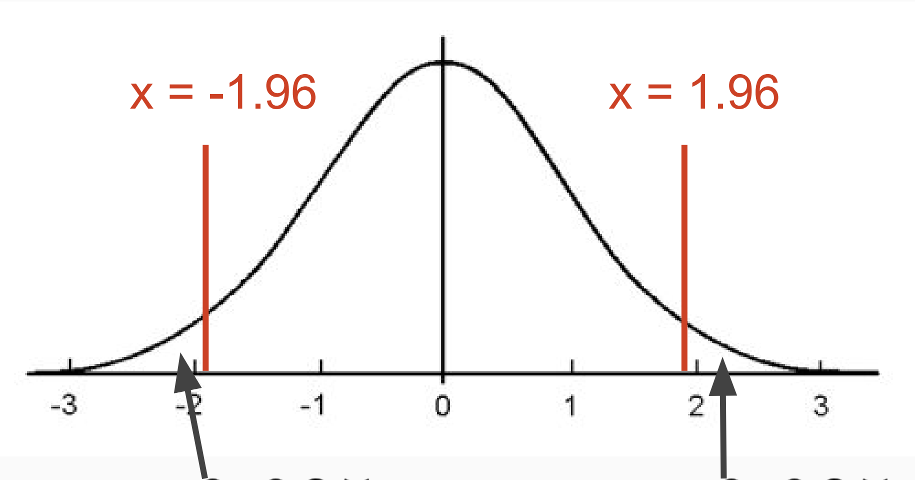
$$\mu' = \mu$$
$$\sigma' = \sigma / \sqrt{n}$$



Hypothesis Testing: Review

Two-sided test:

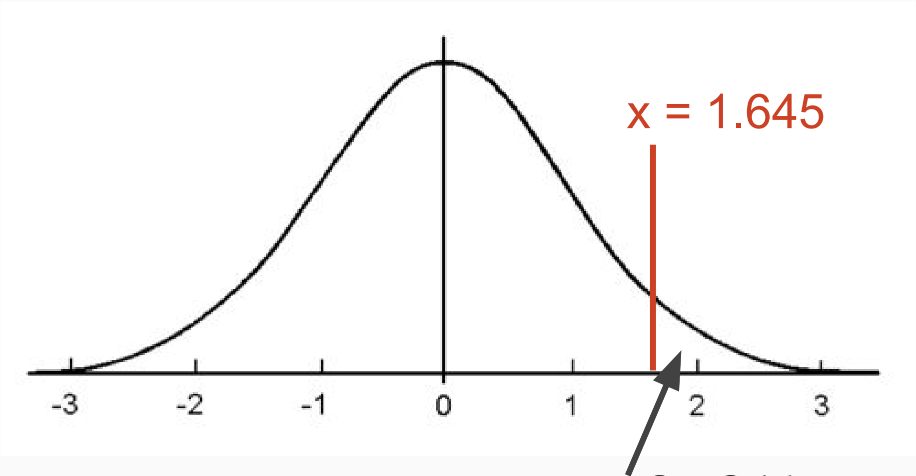
$$H_0 : \mu = 0$$
$$H_A : \mu \neq 0$$



$\alpha = 0.025$ $\alpha = 0.025$

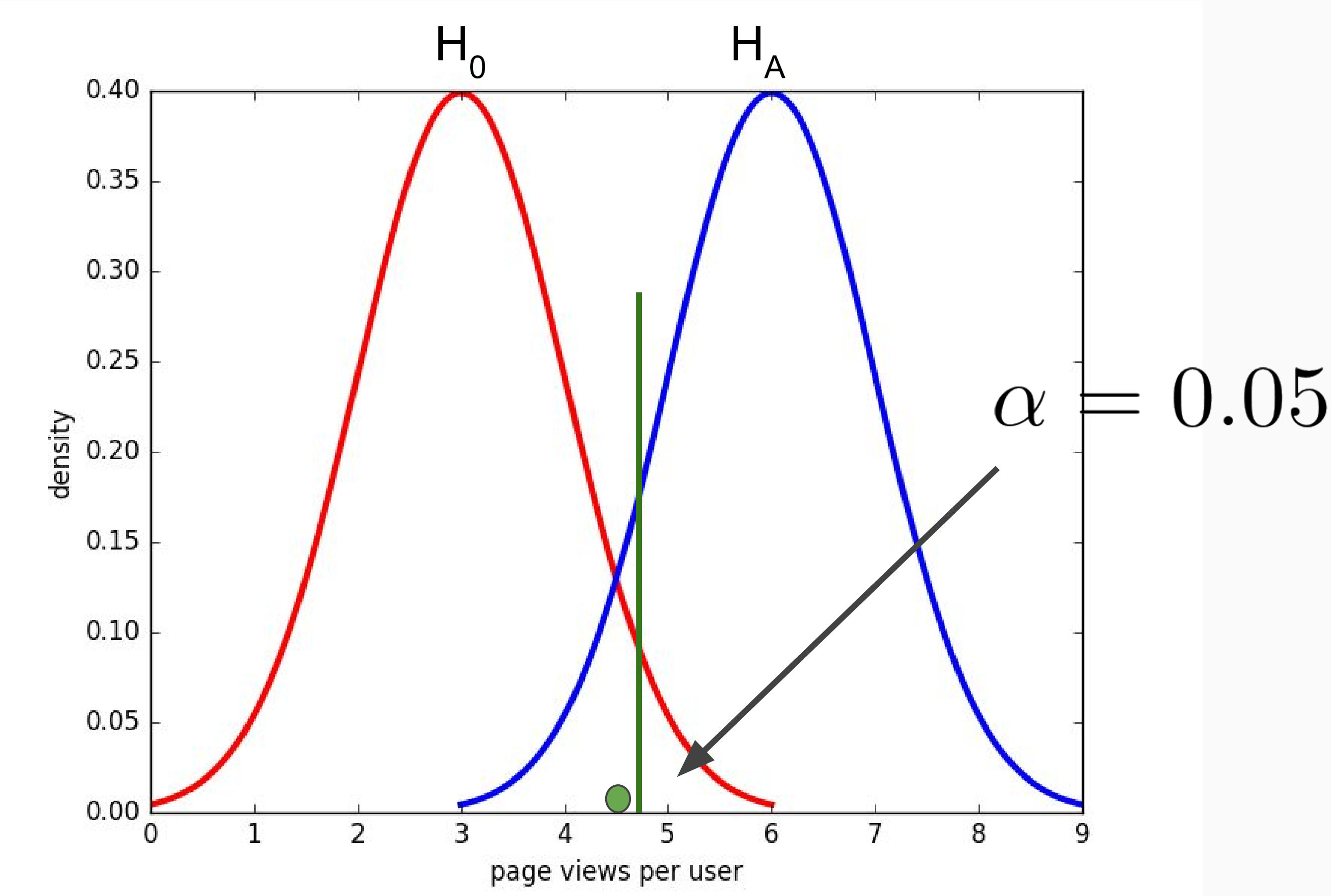
One-sided test:

$$H_0 : \mu = 0$$
$$H_A : \mu > 0$$



$\alpha = 0.05$

Guessing the unknown



Hypothesis Testing: Possible Outcomes

	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.

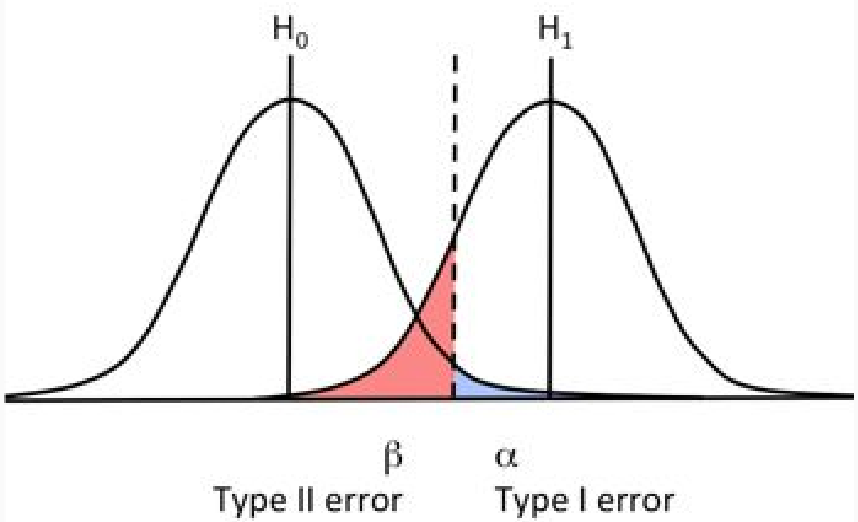
Hypothesis Testing: Possible Outcomes

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

false positive rate
(aka, 1 - specificity)

true positive rate
(aka, sensitivity)

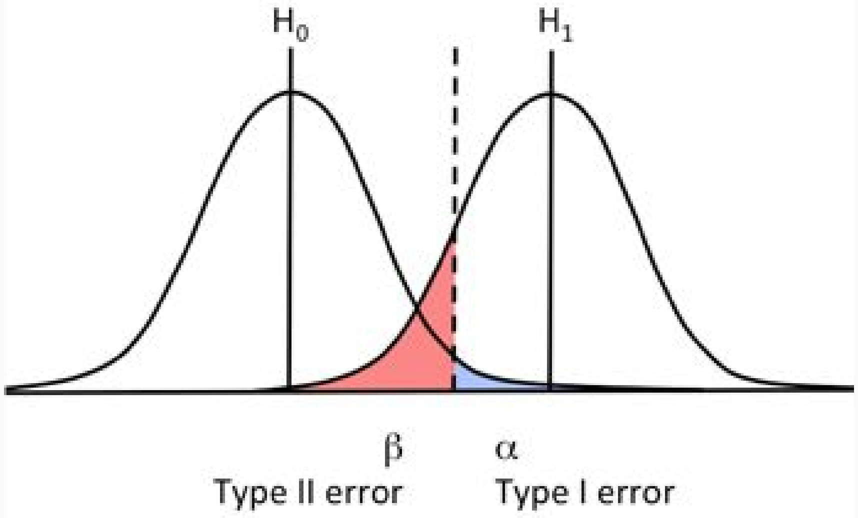
Hypothesis testing: the *power* region



	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 <u>specific</u> alternative hypothesis. Think of it as the *power* to detect a particular “effect size”.

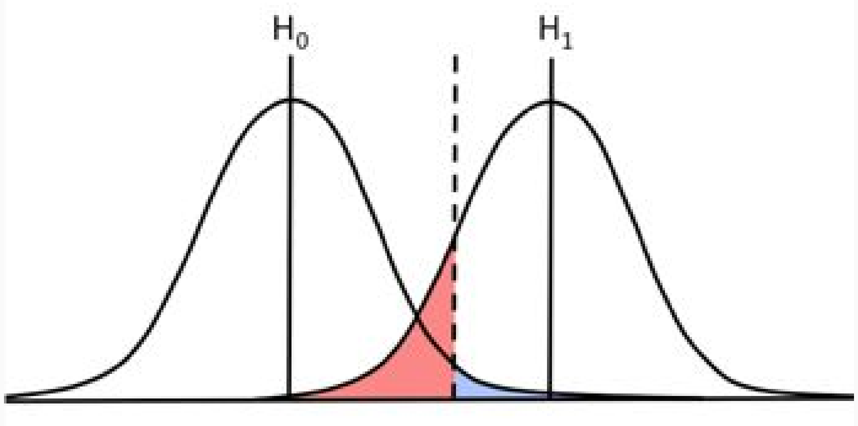
Hypothesis testing: the *power* region



	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?

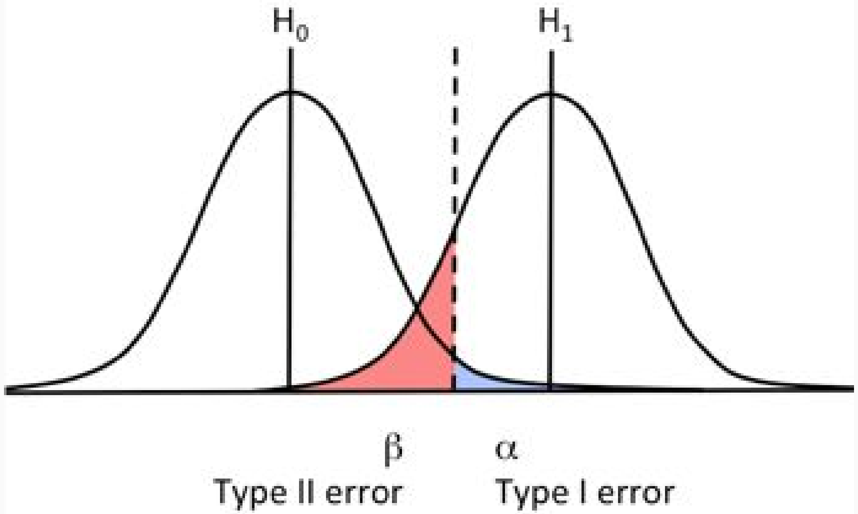
Hypothesis testing: the *power* region



	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* when we increase the effect size?

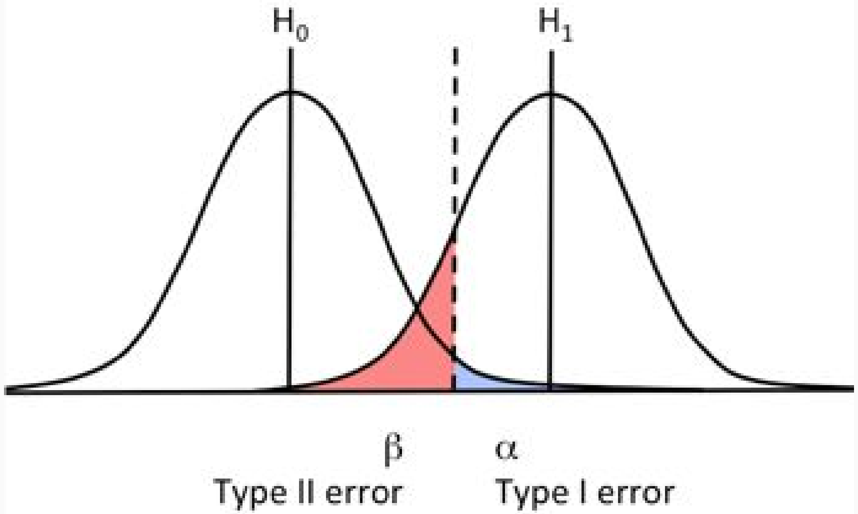
Hypothesis testing: the *power* region



	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* when we increase the sample std. deviation?

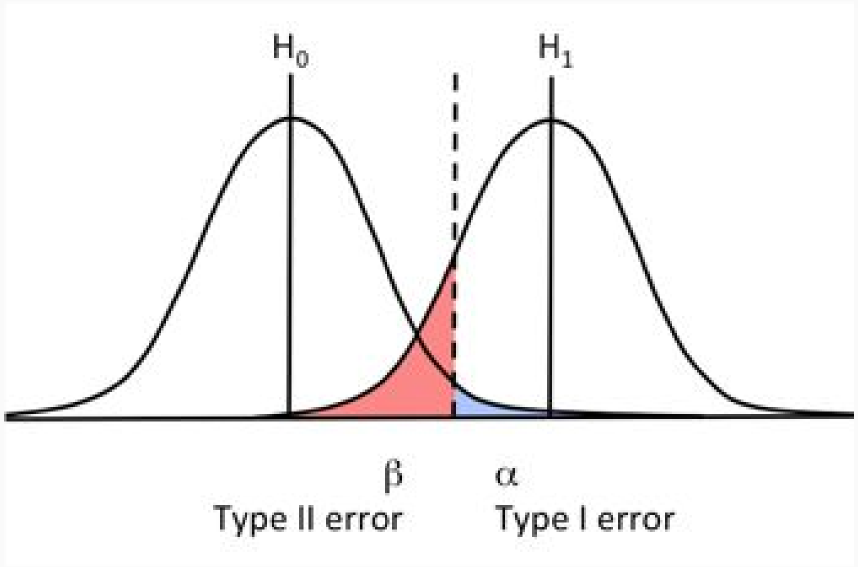
Hypothesis testing: the *power* region



	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* when we increase the sample size?

Hypothesis testing: the *power* region



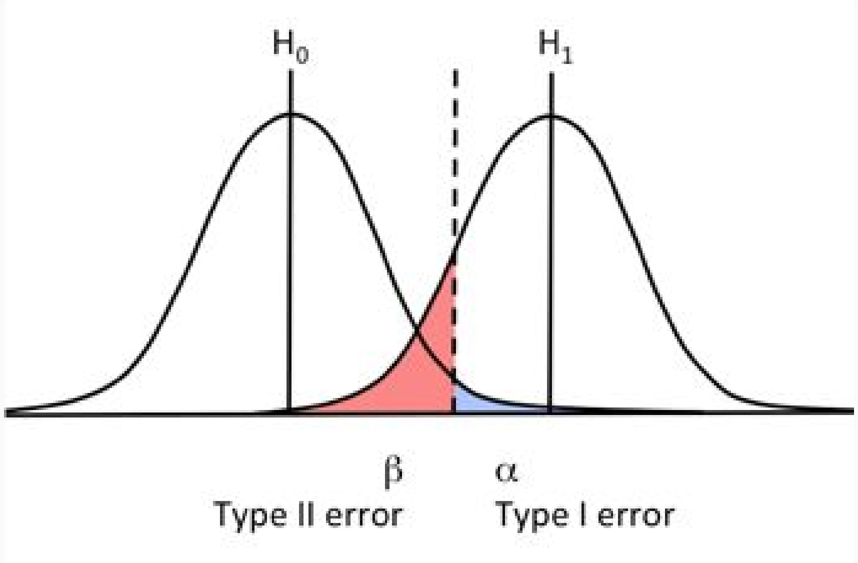
	H_0 is true	H_0 is false
Accept H_0	Correct Decision (1-α)	Type II Error (β)
Reject H_0	Type I Error (α)	Correction Decision (1-β)

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!

Hypothesis testing (revised with power calculation)



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

(new steps)

Calculating the required sample size

To the white board...

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

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

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

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

A/B Testing

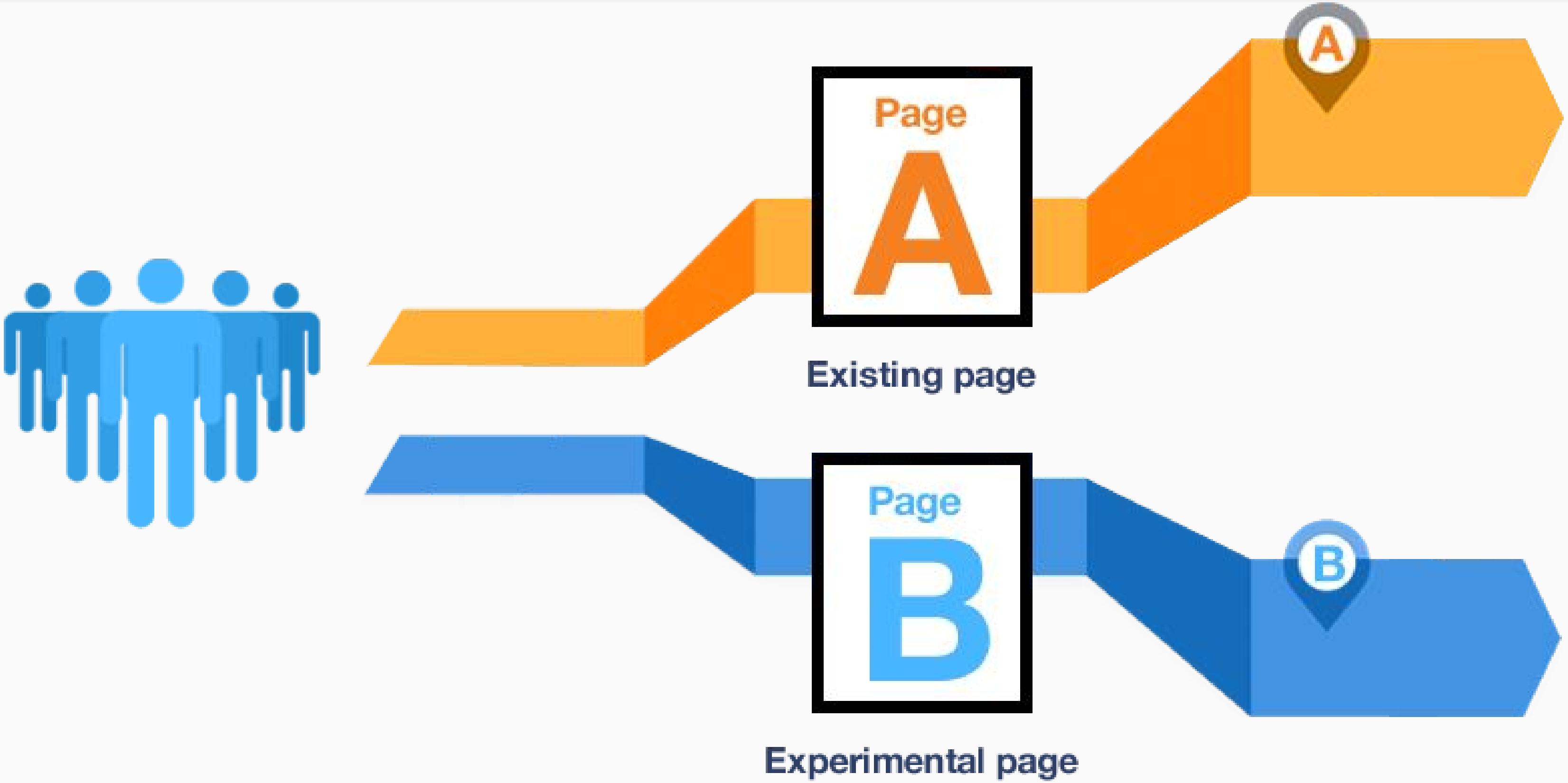


Image from: http://techcrunch.com/2014/06/29/ethics-in-a-data-driven-world/

A/B Testing

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 <u>7% signup rate</u>. We'll want an experiment where <u>alpha is 1%</u> and <u>power is 95%</u>.

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

$$n \geq 9,084$$

A/B Testing

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 <u>1.2% signup rate</u>. We'll want an experiment where <u>alpha is 1%</u> and <u>power is 95%</u>.

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

$$n \geq 39,427$$

A/B Testing

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

Ryan Henning

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

Intuition

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 <u>given</u> that I’m in Seattle?

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

Intuition

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 <u>given</u> that I live in Seattle <u>and</u> I see that the road is wet?

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

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

Experiment 1:

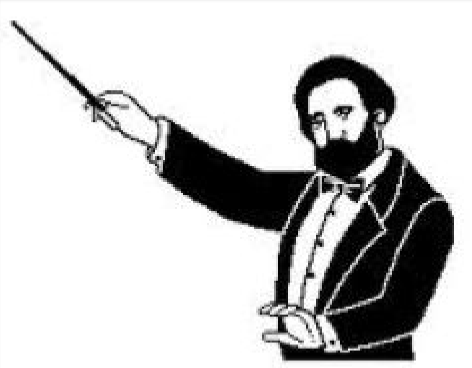
A fine classical musician says he's able to distinguish Haydn from Mozart.
Small excerpts are selected at random and played for the musician.
Musician makes 10 correct guesses in exactly 10 trials.

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.

Adapted example from Jim Berger's book, The Likelihood Principle. Also adapted from Tammy Lee's slides.

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'm not convinced by the drunken man…"

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

Bayes' Rule

posterior probability

likelihood

prior probability

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

marginal likelihood

Bayes' Rule: Example

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

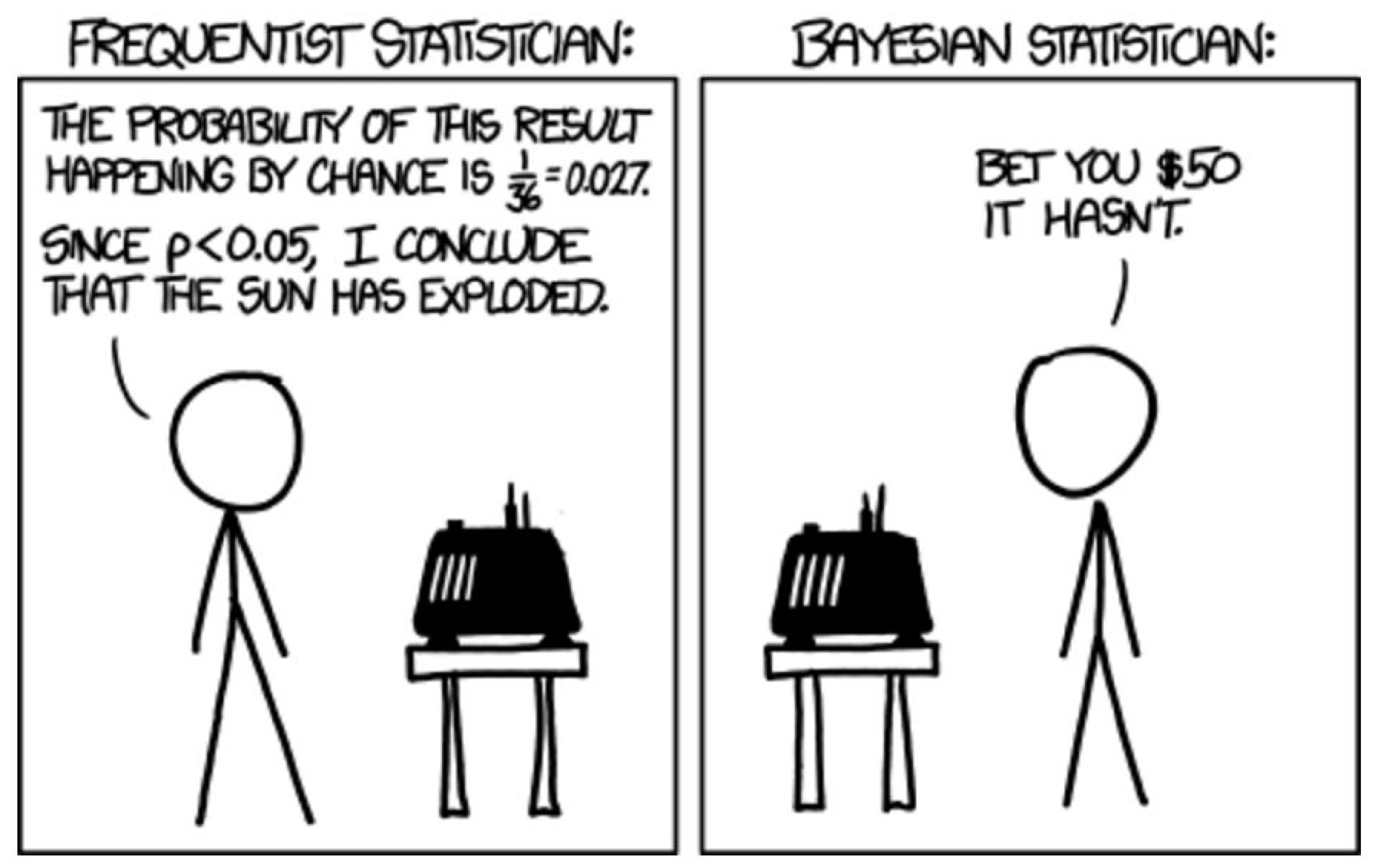
$$= \frac{1.0 * 0.0001}{0.5^{10}}$$

Very subjective!

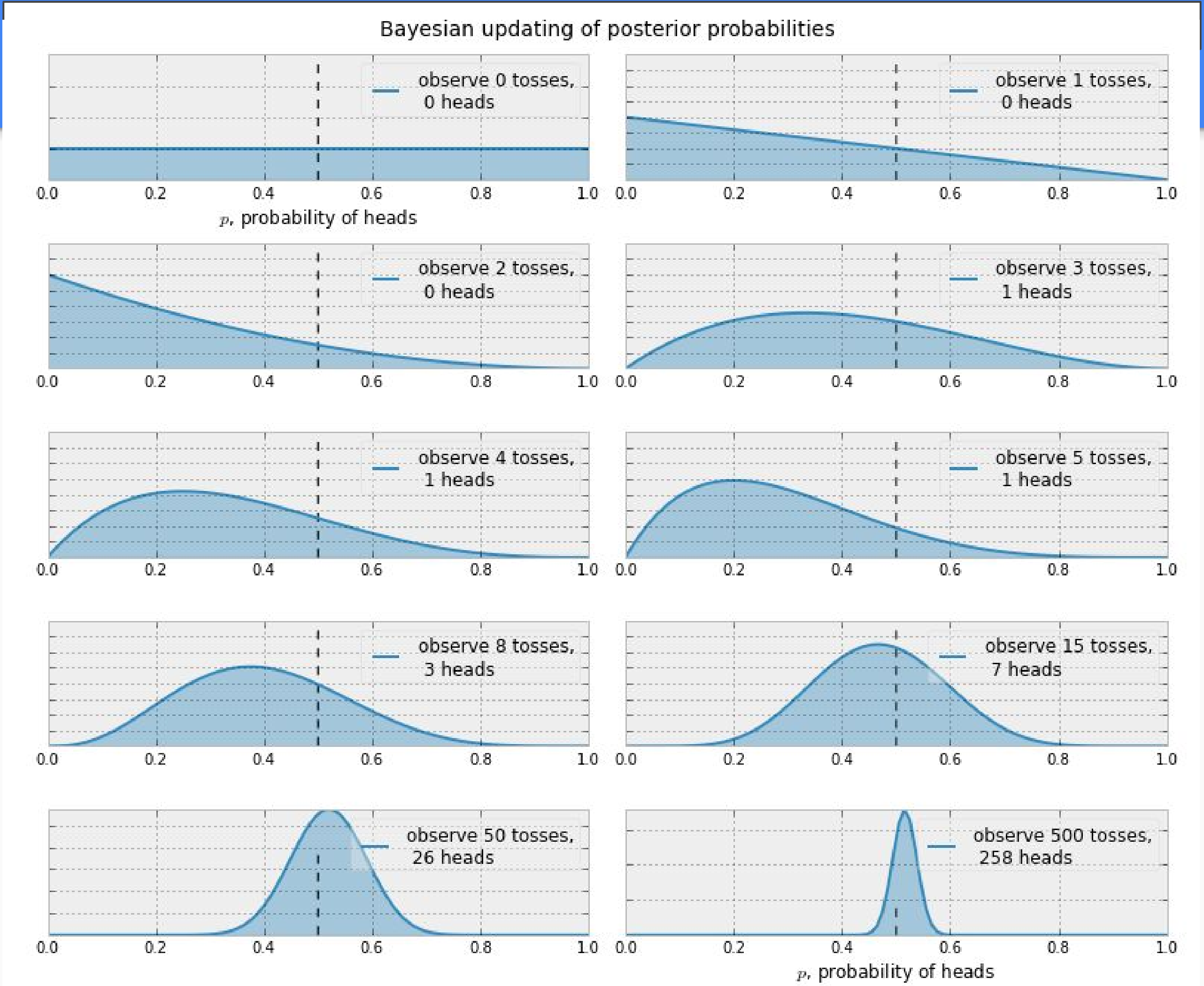
$$= 10.2\%$$

xkcd: Frequentists vs. Bayesians (#1132)

xkcd: Frequentists vs. Bayesians (#1132)



Bayesian Updates



Monty Hall Problem