A Tale of Two Statistics

A hypothesis, **H**, is a model for how the world might work.

In practice, evidence is rarely conclusive.

- We would like to know, $P(H|D)$. That is, the probability that our hypothesis is true, given the data that we observe from the world.
- But there is a fundamental dilemma: We will never know whether our hypothesis is true.
- The world isn't a perfect laboratory, and although we can collect evidence and information, this evidence cannot identify a single, *unique* model from the set of all possible models.
- We cannot make a probability statement about the models either!

Example 1: Coin Flips

Suppose that you flip a coin once and it lands heads.
What is the probability that it is a double-headed coin?

- Is there enough information to answer this question?
- How did the coin get there? The context is missing?
- Even with more information, we still cannot ever possess the *entire* context.

Example 2: Gravity

- Both motions are constant with a gravitation attraction that is proportional to the square of the distance between two objects.
- What is the probability that Newton’s theory of gravity is “correct”?

Problem: Newton’s theory seemed to work well up to the precision of 17th-century instruments.

- But, it is the present now, and we have developed instruments that pretty clearly say that Newton’s laws are incorrect.
- How could Newton decide how likely was his model compared to general relativity (which hadn’t even been imagined yet!)?

Example 2: Take-away

We can never write down the infinite number of models that are possible consistent with our observations. And, even if we could, we could not assign a probability distribution to these models.

Example 3: Giant Squid

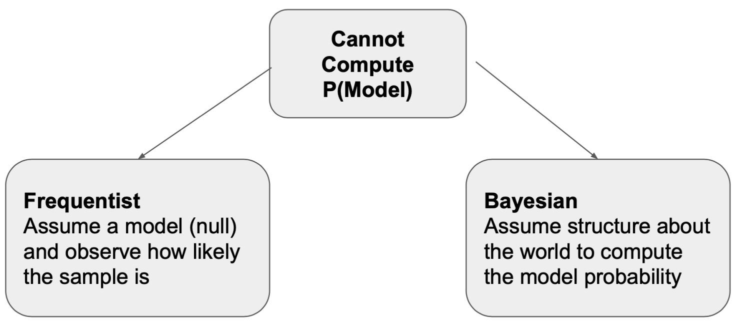
Suppose that you discover three new specimens of a new squid species that measure 3.2, 3.3, and 4.0 feet long.

- What is the probability that the average length among the entire species is 3.5 feet?
- Probability is zero for a single point (continuous probability measures).
- Probability that the average length is between 3 and 4 feet?

What we know and don't know:

- We know about our *sample*.
- We do not know about our *population*.

Two Branches of Statistics



Cannot Compute P(Model)

Frequentist
Assume a model (null) and observe how likely the sample is

Bayesian
Assume structure about the world to compute the model probability

The Frequentist Approach

The Birth of Modern Statistics

Before the 1930s

- Many “statistical” procedures; but, no coherent account of how to choose one.
- Neyman and Pearson published articles that added a formal mathematical treatment, laying the foundations for frequentist statistics.

The Central Dilemma

Despite the “modern” development of statistics, there is a problem.

- We observe data, D.
- Given the data that we observe, we *would like to know* the probability that our hypothesis is true.
- That is, $P(H|D)$.

But, frequentist statistics:

- Not only cannot compute this probability.
- Doesn’t think it even makes sense to assign a probability to a hypothesis.

Objective Probability

A frequentist defines probability as a matter of long-run frequencies.

- Frequentists specify a collective of elements (e.g. experiments, or throws of dice).
- As the number of observations approaches infinity, the proportion of the throws that show a 3 is $\frac{1}{6}$.
- For frequentists, **objective probability** is this long-run frequency of the event.

Objective Probability and Hypotheses

- If one views probability as objective, then they cannot talk about the probability of a hypothesis.
- A hypothesis is just a statement that is either *TRUE*, or *FALSE*.
- This mixes language a little, but when one talks about the probability of a hypothesis, they are talking about **subjective probability**, which is not the purview of data science.

What Probabilities Can We Study?

We can study events that have (or could have) a long-run collective.

$$P(D|H)$$

where H is a hypothesis and D is data.

- Assume that H is true, and call it the null hypothesis.
- Has to be *very* specific.
- Is the basis for predictions we need to make about the data that “should” come out of the experiment (if this H were actually true).

An Imprecise Statement

Vitamin W kicks the bad toxins right smack outta your system. (Read this like a TV salesperson.)

A Precise Statement

Vitamin W reduces blood pressure by 12 mmHg on average.

- With a precise statement, we also have a statement about the data that *should* be observed; and so,
- We have a collective to produce a long-run frequency, i.e. a probability.

Decision Rules

Hypothesis Test Example

Mad data science

Suppose that your lab has synthesized a new compound, *Vitamin W*.

Let random variable B represent the change in blood pressure that results from taking *Vitamin W*.

Let $\mu = \mathrm{E}[B]$.

You need to make a decision, to invest resources in Vitamin W or not.

Two Possible States of the World

Goal: Begin with a reasonable default supposition; leave this supposition behind if data provides compelling evidence

Null hypothesis

- Default assumption, status quo, statement that data might overturn
- $H_{\varnothing}$: Usually $\mu = 0$
- No effect

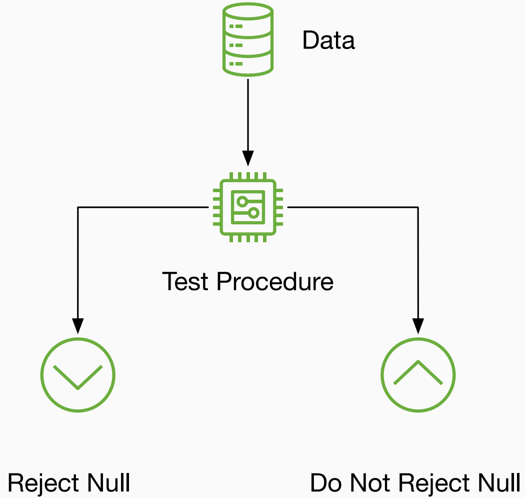
Alternative hypothesis

- Idea or alternative to status quo
- H_a : Usually $\mu \neq 0$
- Some effect exists

With compelling evidence, we leave the specific null hypothesis ($H_{\varnothing}$) for the alternative (H_a)

A Hypothesis Test

A *hypothesis test* is a procedure.



False Positive and False Negative Errors

	True state of the world	
	The null is true	*The null is false*
Reject the null	False Positive (Type I Error)	
Do not reject the null		False Negative (Type II Error)

False Positive and False Negative Errors (cont.)

False Positive Errors

- Typically the most destructive
- Error rate, denoted α, is the probability of rejecting the null hypothesis when we should not; $P(\text{Reject } H_\varnothing | H_\varnothing)$
- Starting with Ronald Fisher: set $\alpha = 0.05$

A hypothesis test is a procedure for rejecting or not rejecting a null, such that the false positive error rate is controlled ($\alpha = 0.05$).

Breaking Down a Test Procedure

A test statistic

- A function of our sample
- Measures deviations from the null hypothesis
- Distribution must be completely determined by the null

A rejection region

- A set of values for which we will reject the null
- Chosen to be contrary to the null
- Total probability must be $\alpha = 0.05$

What a Hypothesis Test Doesn't Do

A hypothesis test does not prove the null hypothesis.

- We control Type 1 error rates
- We cannot control Type 2 error rates
- How can you be sure the real B is not 0.01? Or 0.00001?

Never accept the null hypothesis.

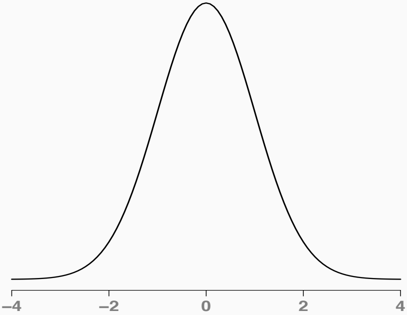
- The valid decisions are reject and fail to reject.

The One-Sample z-Test

Vitamin W Example

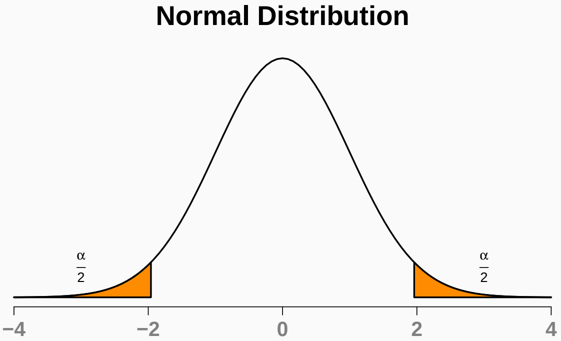
Suppose $(B_1, .., B_{100})$ are i.i.d. random variables with mean $\mu = \mathrm{E}[B]$, representing changes in blood pressure.

Assume $B \sim N(\mu, \sigma)$. Assume we know $\sigma[B] = 20$.



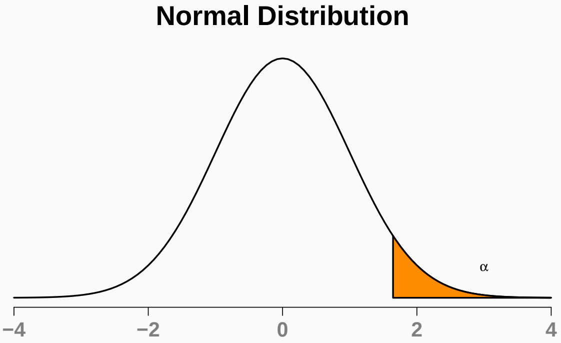
One- and Two-Tailed Tests

The Two-Tailed z-Test



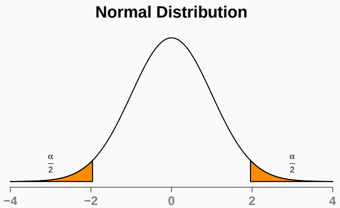
- **Null hypothesis**: $\mu = 0$
- **Alternative hypothesis**: $\mu \neq 0$

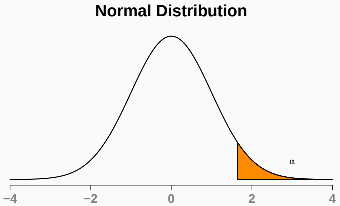
The One-Tailed z-Test



- **Null hypothesis**: $\mu = 0$
- **Alternative hypothesis 1**: $\mu > 0$
- **Alternative hypothesis 2**: $\mu < 0$

Choosing One or Two Tails





Switching your test after you see the statistic is cheating.

One-Tailed Test: Things to Consider

Before using a one-tailed test, ask yourself these questions:

1. Will the audience believe that I started with one tail before I saw the data?
2. Will the audience share my opinion of which tail is interesting?
3. Am I really 100% committed to only this tail?
    - What if the effect turns out to be huge, but in the other direction?
    - Would I be willing to call that a negative result?
    - Can I convince my audience I have this much commitment?

T-Test Assumptions

T-Test Assumptions, Part I

Assumptions of t-test

The textbook assumptions

- X is a metric variable.
- $\{X_1, X_2, \ldots, X_n\}$ is a random sample.
- X has a normal distribution.

Variables are almost never normal.

T-Test Assumptions, Part II

But, in the large sample case, this is more plausible.

Large sample t-test assumptions

If:

- X is a metric variable
- $\{X_1, X_2, \ldots, X_n\}$ is a random sample
- n is large enough that the CLT implies a normal distribution of mean

Then: The t-test is asymptotically valid

T-Test Assumptions, Part III

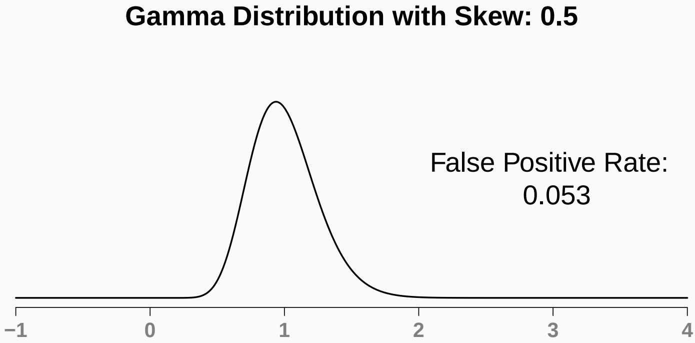
T-Test Assumptions, Part IV

The t-test is considered "reasonably robust," even when $n < 30$, as long as deviations from normality are moderate.

However, watch out for strong skewness, especially when $n < 30$.

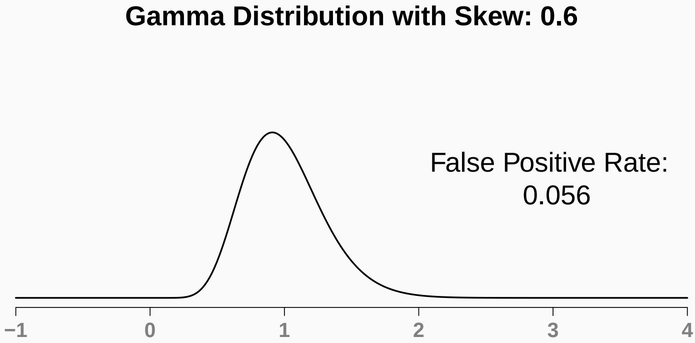
Gamma With Increasing Skew

Twenty draws from gamma distributions



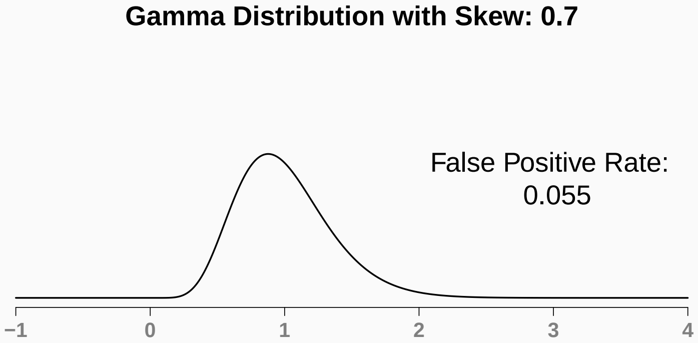
Gamma With Increasing Skew

Twenty draws from gamma distributions



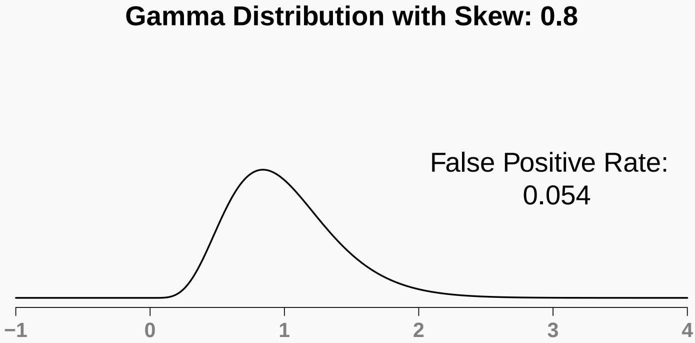
Gamma With Increasing Skew

Twenty draws from gamma distributions



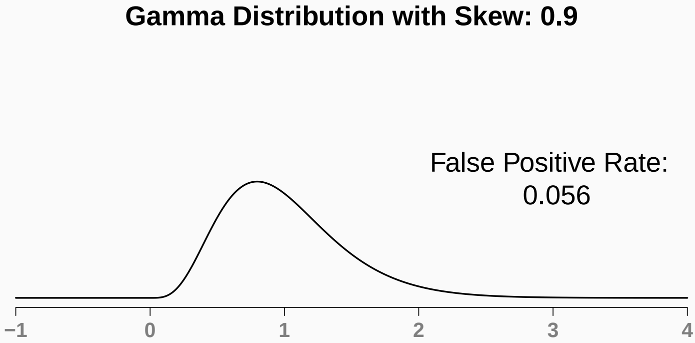
Gamma With Increasing Skew

Twenty draws from gamma distributions



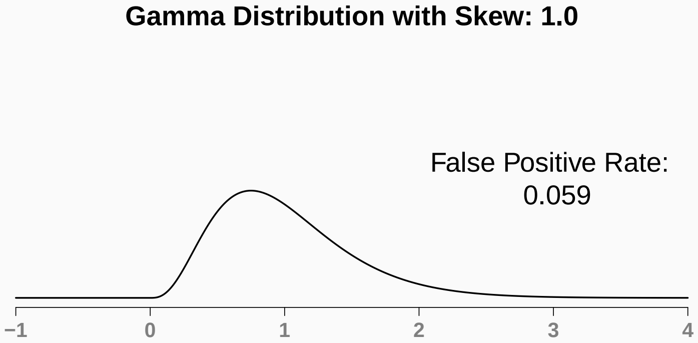
Gamma With Increasing Skew

Twenty draws from gamma distributions



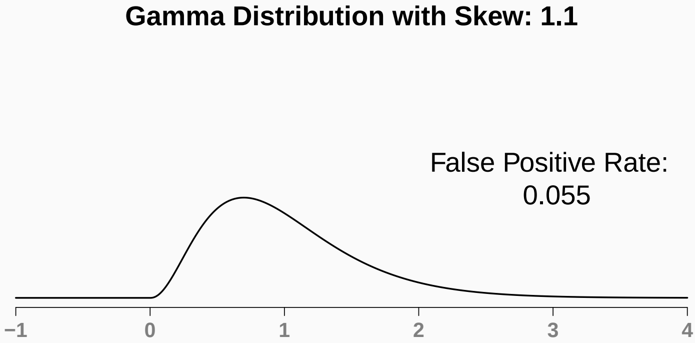
Gamma With Increasing Skew

Twenty draws from gamma distributions



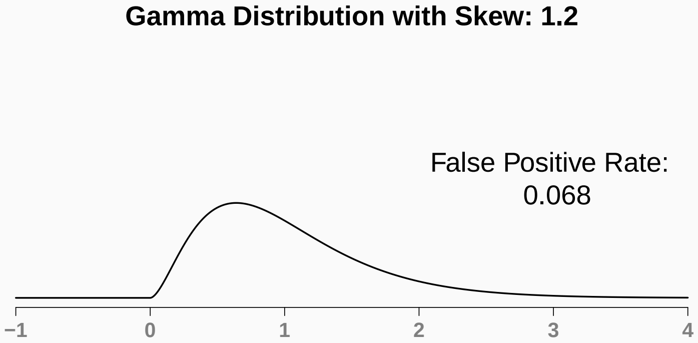
Gamma With Increasing Skew

Twenty draws from gamma distributions



Gamma With Increasing Skew

Twenty draws from gamma distributions

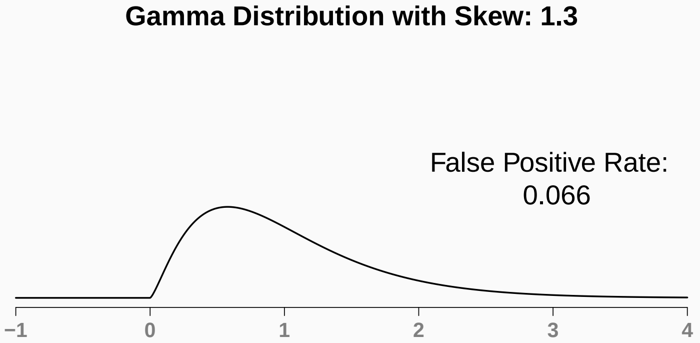


Gamma Distribution with Skew: 1.2

False Positive Rate:
0.068

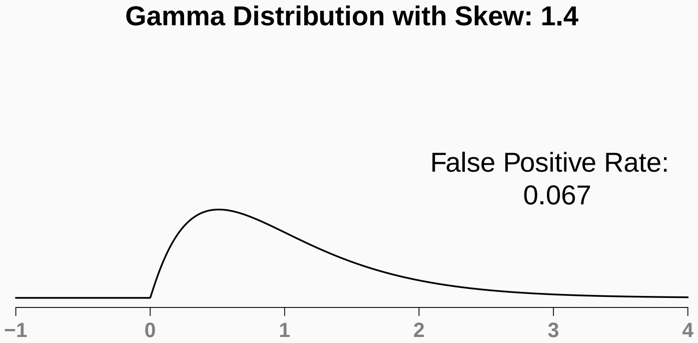
Gamma With Increasing Skew

Twenty draws from gamma distributions



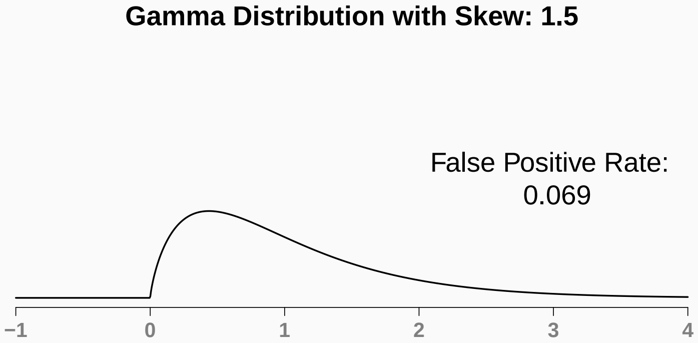
Gamma With Increasing Skew

Twenty draws from gamma distributions



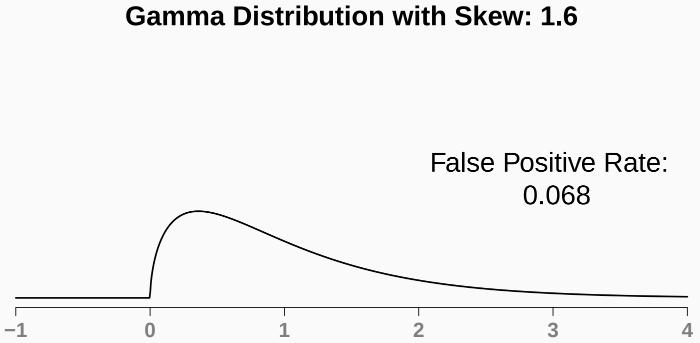
Gamma With Increasing Skew

Twenty draws from gamma distributions



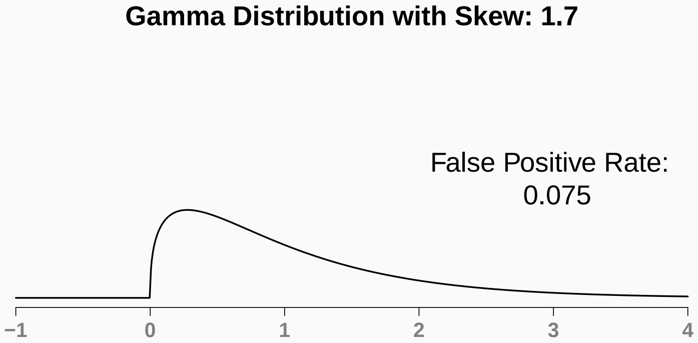
Gamma With Increasing Skew

Twenty draws from gamma distributions



Gamma With Increasing Skew

Twenty draws from gamma distributions

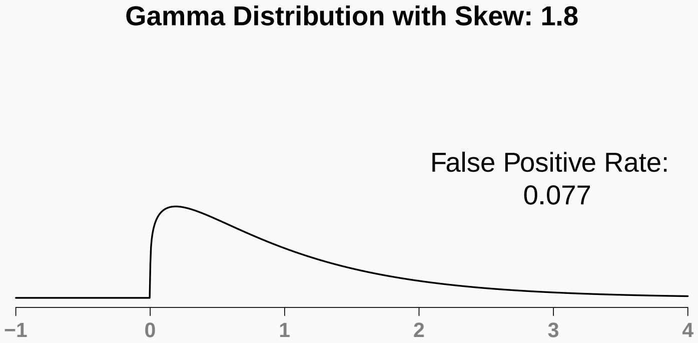


Gamma Distribution with Skew: 1.7

False Positive Rate:
0.075

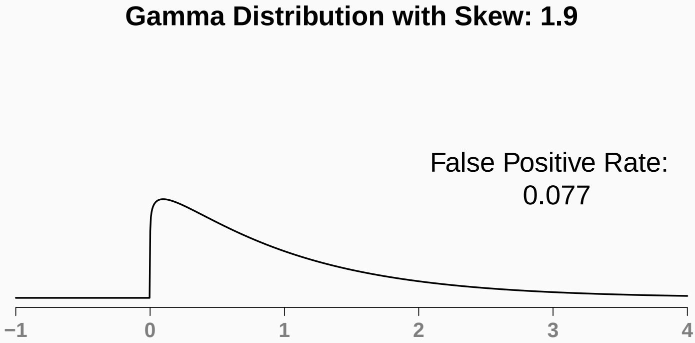
Gamma With Increasing Skew

Twenty draws from gamma distributions



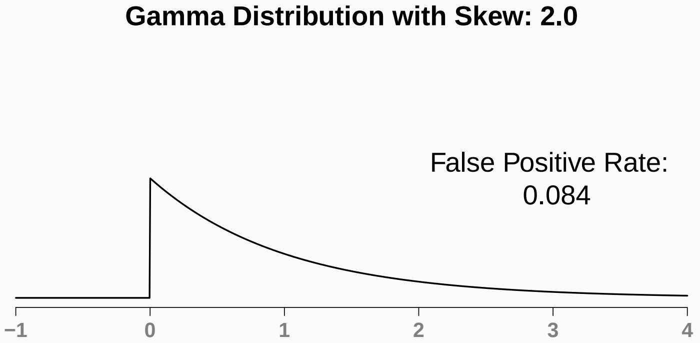
Gamma With Increasing Skew

Twenty draws from gamma distributions



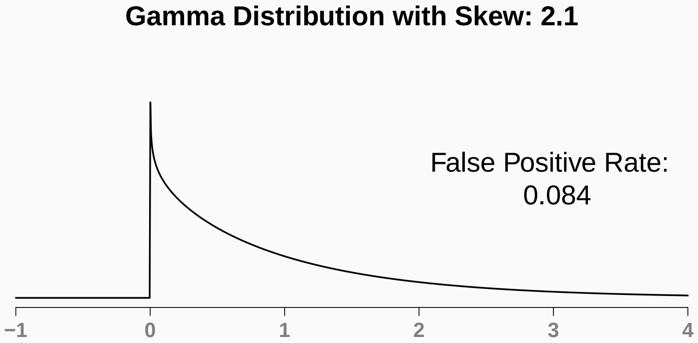
Gamma With Increasing Skew

Twenty draws from gamma distributions



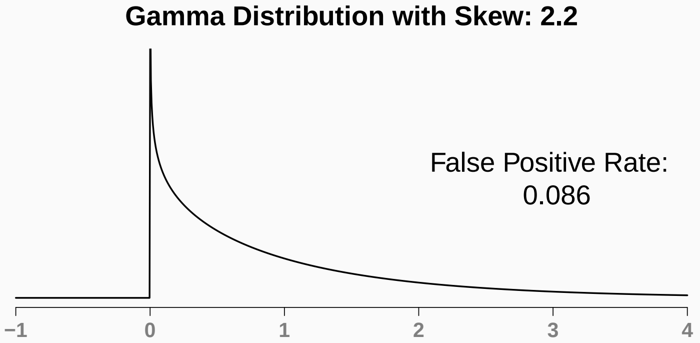
Gamma With Increasing Skew

Twenty draws from gamma distributions



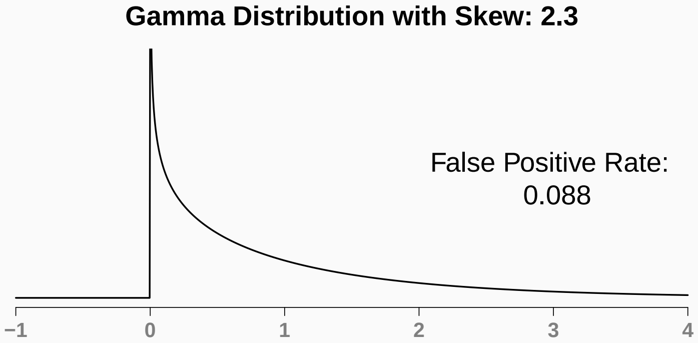
Gamma With Increasing Skew

Twenty draws from gamma distributions



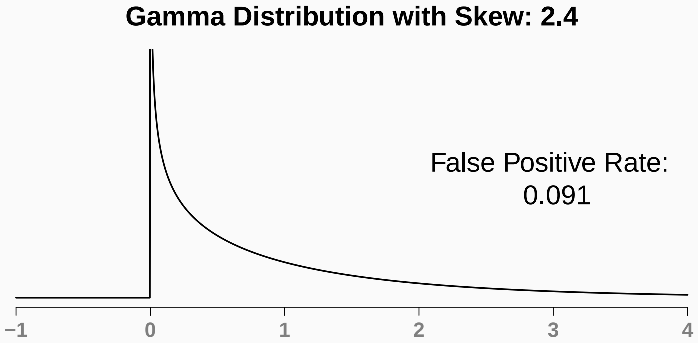
Gamma With Increasing Skew

Twenty draws from gamma distributions



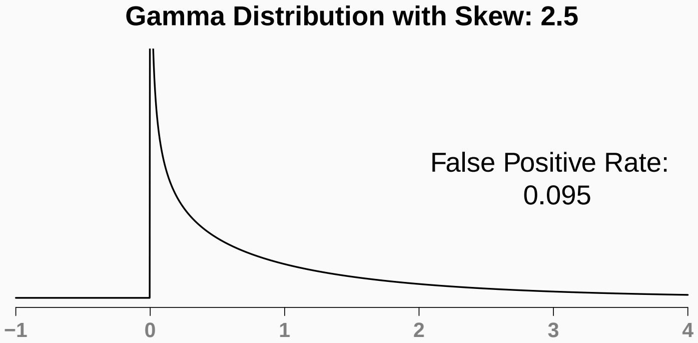
Gamma With Increasing Skew

Twenty draws from gamma distributions



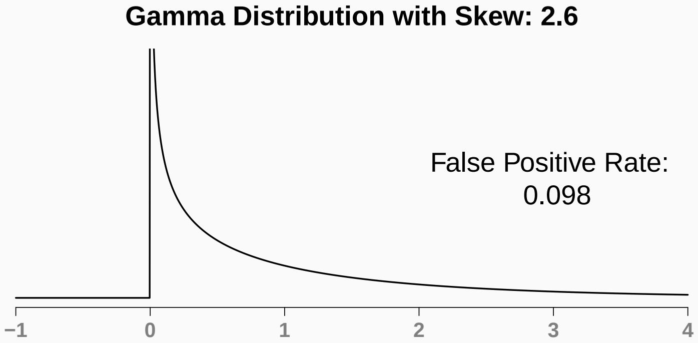
Gamma With Increasing Skew

Twenty draws from gamma distributions



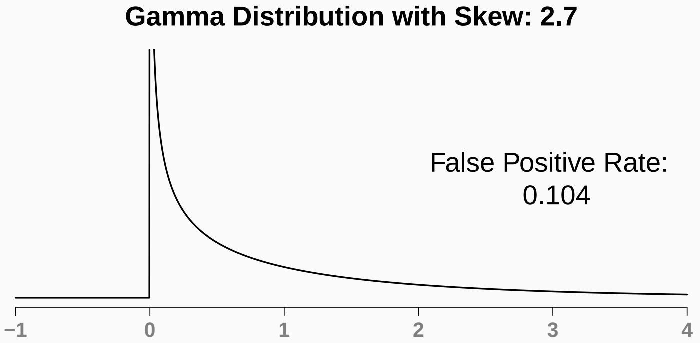
Gamma With Increasing Skew

Twenty draws from gamma distributions



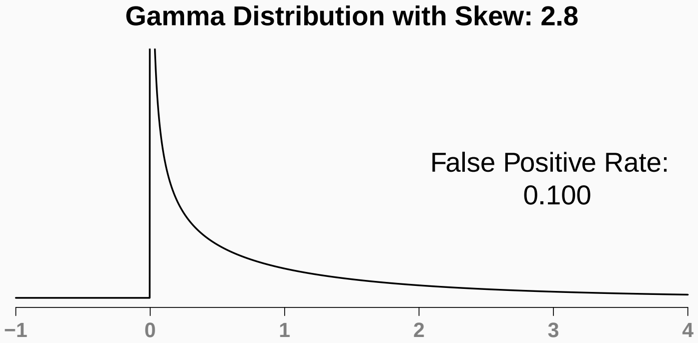
Gamma With Increasing Skew

Twenty draws from gamma distributions



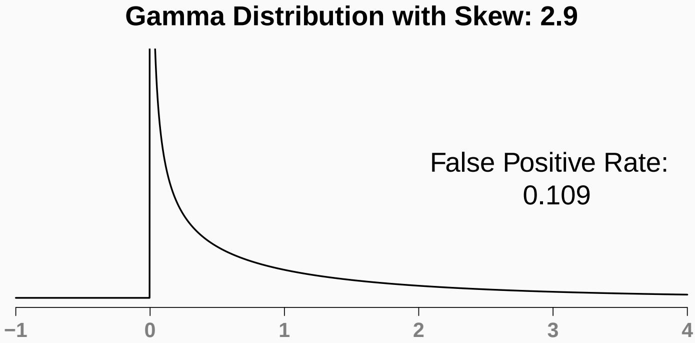
Gamma With Increasing Skew

Twenty draws from gamma distributions



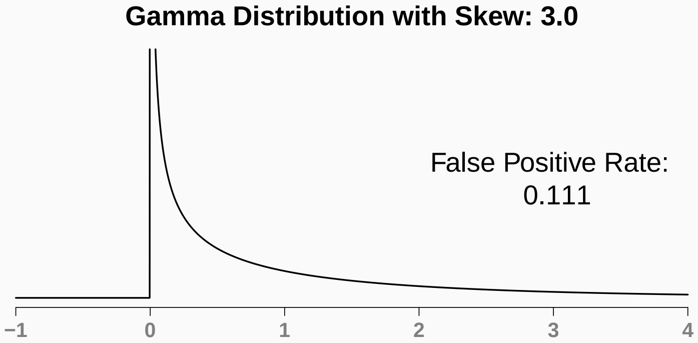
Gamma With Increasing Skew

Twenty draws from gamma distributions



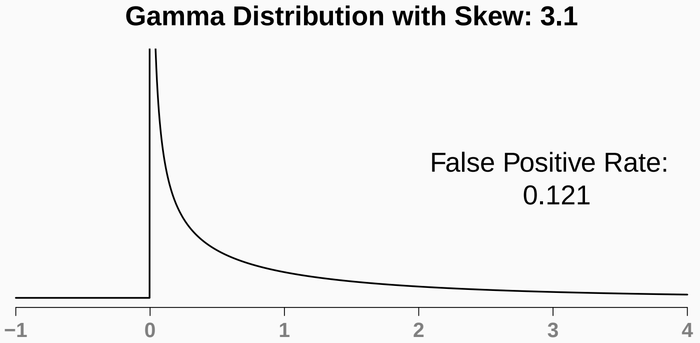
Gamma With Increasing Skew

Twenty draws from gamma distributions



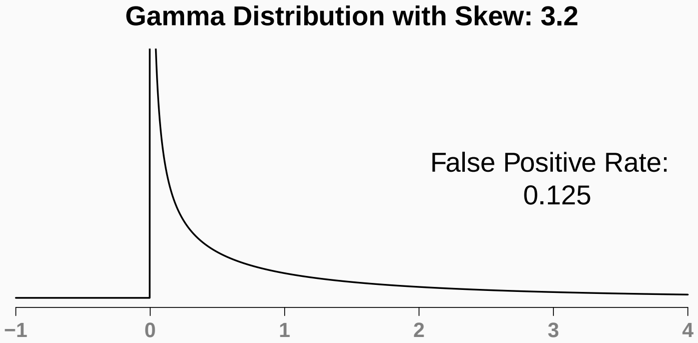
Gamma With Increasing Skew

Twenty draws from gamma distributions



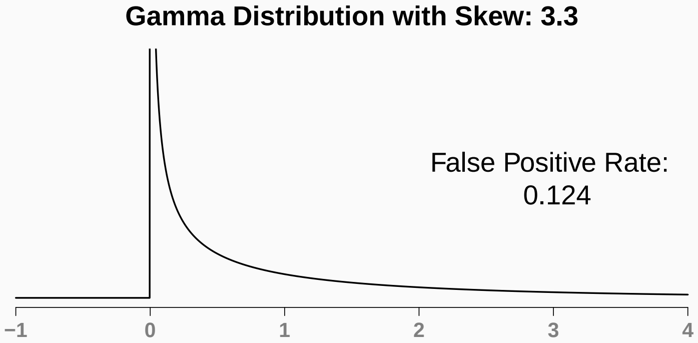
Gamma With Increasing Skew

Twenty draws from gamma distributions



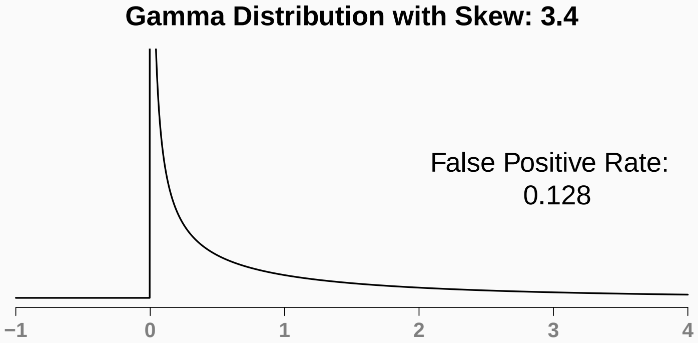
Gamma With Increasing Skew

Twenty draws from gamma distributions



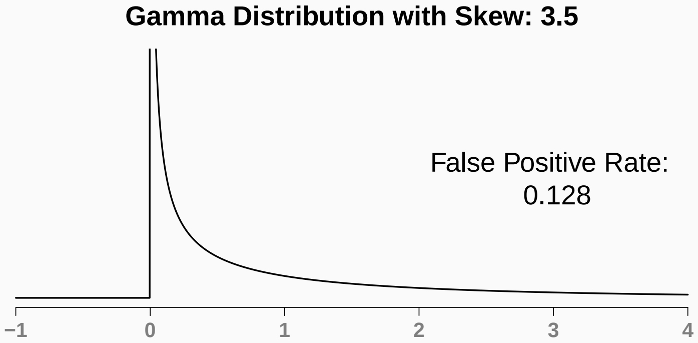
Gamma With Increasing Skew

Twenty draws from gamma distributions



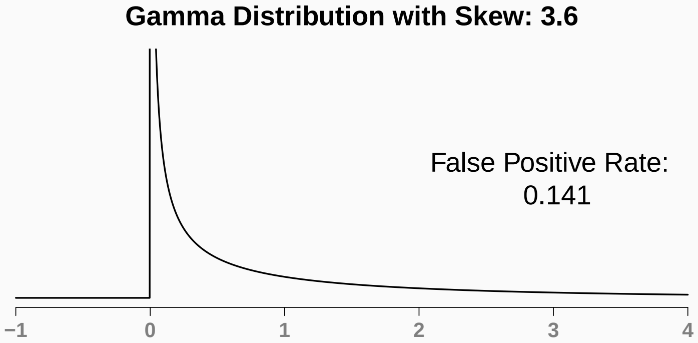
Gamma With Increasing Skew

Twenty draws from gamma distributions



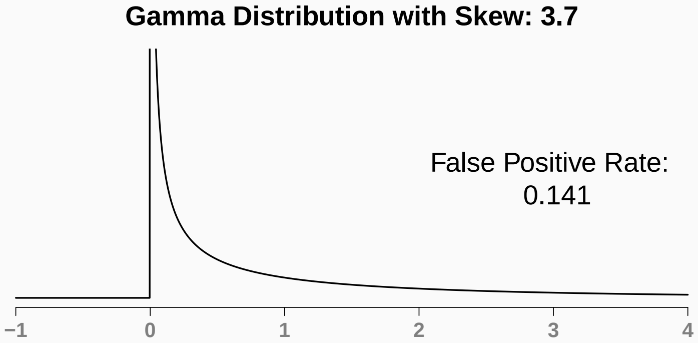
Gamma With Increasing Skew

Twenty draws from gamma distributions



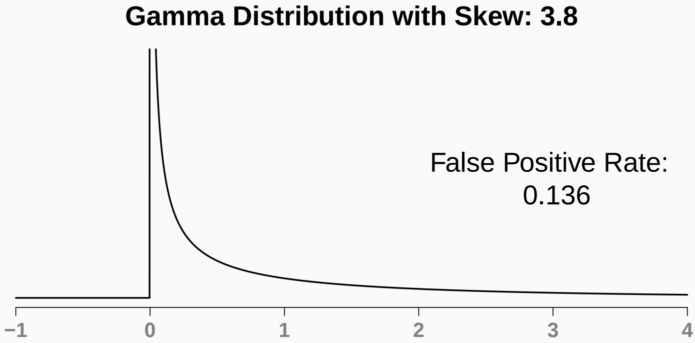
Gamma With Increasing Skew

Twenty draws from gamma distributions



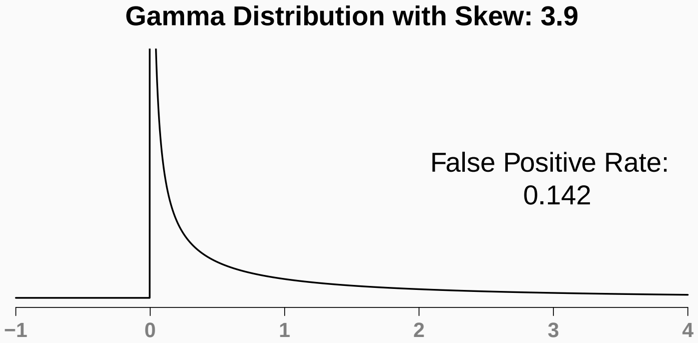
Gamma With Increasing Skew

Twenty draws from gamma distributions



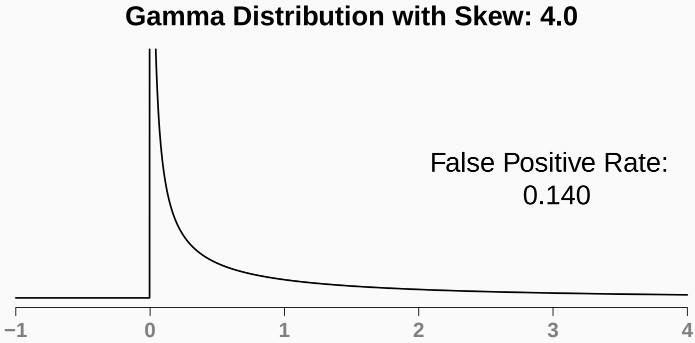
Gamma With Increasing Skew

Twenty draws from gamma distributions



Gamma With Increasing Skew

Twenty draws from gamma distributions



T-Test Assumptions

More practical guidance:

- X is a metric variable.
- $\{X_1, X_2, ..., X_n\}$ is a random sample.
- The distribution is not too non-normal, considering n.

When the t-test is not valid, consider using a non-parametric test instead.

Introduction to P-Values

Introducing P-Values

> *The p-value is the probability, calculated assuming that the null hypothesis is true, of obtaining a value of the test statistic at least as contradictory to H_0 as the value calculated from the available sample.*

Jay L. Devore (2015)

Z-Distribution

The P-Value for a Z-Test

Vitamin W

You measure the effects of Vitamin W on blood pressure (measured in *mmHg*) for 100 patients and get $\bar{X} = 3$.

Assume $X \sim N(\mu, 20)$.

- $H_0 : \mu = 0$
- $z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$

The P-Value and Decision Rules, Part I

Neyman-Pearson hypothesis testing: rules to make a decision and usually be right ($\alpha = 0.05$)

A classic z-test

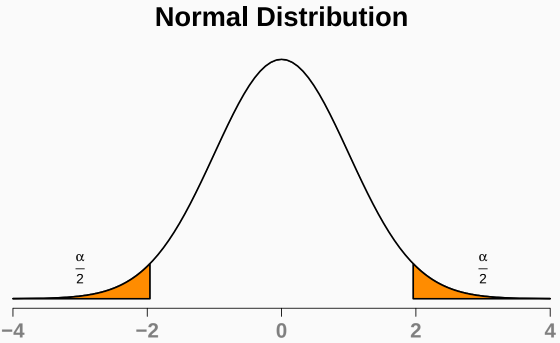
- z=1 $\rightarrow$ Do not reject null.
- z=2 $\rightarrow$ Reject null.
- z=10 $\rightarrow$ Reject null.

- Strict frequentist with a dichotomous decision rule: treat $z = 2$ and $z = 10$ identically.
- But is there value in knowing *how contrary* the data is to the null?

The P-Value and Decision Rules, Part II

$$|z| > \text{critical value} \Rightarrow \text{reject } H_0$$

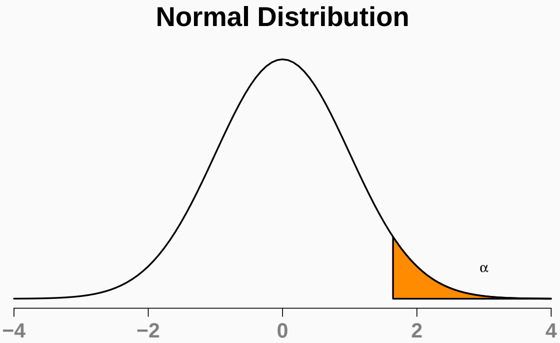
$$|z| < \text{critical value} \Rightarrow \text{fail to reject } H_0$$



The P-Value and Decision Rules, Part III

$$|z| > \text{critical value} \Rightarrow \text{reject } H_0$$

$$|z| < \text{critical value} \Rightarrow \text{fail to reject } H_0$$



An Equivalent Decision Procedure

Compute p-value.

- If $p < .05 \Rightarrow$ reject H_0
- If $p \geq .05 \Rightarrow$ do not reject H_0

But, can you justify making such a bright-line statement after reducing information so much?

1. Concept
2. Measurement
3. Statistic
4. Assumptions about distribution
5. **p-value**
6. Reject/fail to reject

t-Test and p-Values

P-Value Convention

p-value range	Convention	Symbol
$p > 0.10$	Non-significant	
$0.10 > p > 0.05$	Marginally-significant	.
$p < 0.05$	Significant	*
$p < 0.01$	Highly significant	**
$p < 0.001$	Very highly significant	***

Reporting Test Results

- A t-test for the effect of Vitamin W on blood pressure was highly significant ($t = 3.1$, $p = .008$).
- We found evidence that Vitamin W decreases blood pressure ($t = 2.3$, $p = .04$).
- The effect of Vitamin X on blood pressure was not statistically significant ($t = 1.2$, $p = .23$).

Vitamin W	Vitamin X
2.2 **	1.2
(0.6)	(0.8)

This is half the story; next, you'll need to describe practical significance.

Variable Importance and P-Values

Does a small p-value mean that a variable is "important"?

- Statistical significance
- Practical significance

A Warning

A very common mistake is to assume a p-value is the chance the null hypothesis is true.

Frequentist statistics cannot tell you the probability of a hypothesis!

A Warning (cont.)

Example

I test whether Vitamin X decreases blood pressure: $p = 0.03$.

However, you know that Vitamin X is secretly cornstarch because you created it yourself.

My test will not convince you that there is a 97% chance Vitamin X decreases blood pressure.

Statistical Power

False Positive and False Negative Errors

	The null is true	The null is false
Reject the null	False Positive (I)	
Do not reject the null		False Negative (II)

- False Positive (I) errors are jumping without cause
- False Negative (II) errors are failing to jump when you should
    - Failing to detect a real effect
    - Missed opportunity to create a product, publish a paper, or advance knowledge

Statistical Power, Part I

Much Vitamin W

Consider a *specific* alternate hypothesis:

- H_a : Vitamin W decreases blood pressure by 20 mmHg

- False Negative Error Rate: $\beta = P(\text{not rejecting } H_0 | H_a)$
- Statistical power: $1 - \beta$
- Statistical power is the probability of supporting the alternate hypothesis, assuming it is true

Statistical Power, Part II

Statistical Power, Part II

Statistical Power, Part III

How to increase power

- Increase sample size.
- Choose a powerful test (if you can justify its assumptions).

Practical Significance

Practical significance

Statistical significance

- How much does the data support the existence of an effect?

Practical significance

- Is the size of this effect important?
- What is the magnitude of the effect?
- Should we care about this effect?

Example

Productivity supplements

Vitamin W

$$n = 30$$
$$\mu_{treat} = 12.6$$
$$\mu_{control} = 6.1$$
$$p = 0.11$$

"The difference between groups was not statistically significant, ($t = 1.34$, $p = 0.11$)."

Vitamin Q

$$n = 30,000$$
$$\mu_{treat} = 6.25$$
$$\mu_{control} = 6.21$$
$$p = 0.0005$$

"The difference between the two groups was highly significant, ($t = 3.34$, $p < 0.001$)."

Practical Significance: Context

Primary goal: Provide context for your audience to reason about results.

- Who is your audience?
- What action might be taken based on these results?
- How does this result alter how you would run the business?
- What is the cost-benefit for implementing a change based on this result?
- How does this result “stack up” to other effects?

Practical Significance: Model Explainability

- Some tasks require *explainable* models.
- Finance, healthcare, insurance, and other regulated industries stipulate specific model forms .
- Humans reason in linear hypotheses—higher-dimensional and conditional hypotheses are too much to keep in mind.

Practical Significance: Effect Sizes

Effect sizes

- Single-number metrics that characterize the magnitude of an effect
- Population parameters that we estimate—*do not vary based on sample size*

Invalid effect size metrics

- t-stat
- p-value

Valid effect size metrics

- Mean values
- Difference in means between groups

Standard Effect Size Measures

Standardized effect sizes are designed to be flexible and apply in many scenarios:

- Cohen's *d*
- Correlation ρ
- Cramer's *V*

General metrics ignore the specific context around your research or business question.

Cohen's d

Sometimes, a mean (or difference in means) is hard to assess because the units are unfamiliar.

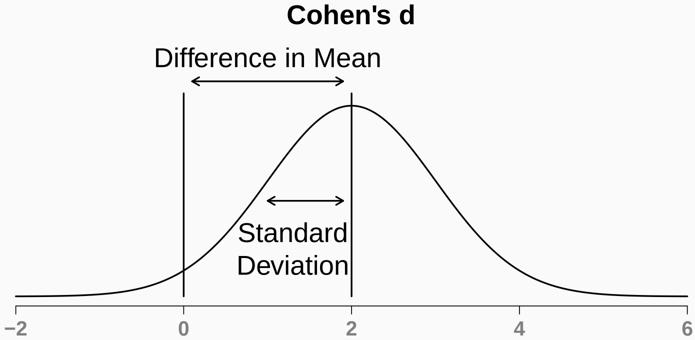
- **Example**: The effect of angled bristles on tooth decay is 5 millicaviparsecs per brushstroke

Cohen's d

Compare effect size relative to the underlying natural variation in the outcome.

$$\text{Cohen's } d = \frac{\text{mean difference}}{\text{standard deviation}}$$

Cohen's d (cont.)



Rules of thumb (according to Cohen)

Small effect	$d = 0.2$
Medium effect	$d = 0.5$
Large effect	$d = 0.8$

- Applicable across a huge number of contexts
- Ignores any important differences between context
- Saving dollars or saving lives are the same to Cohen’s d

Takeaways

- After a statistical test, it’s important to assess both statistical significance and practical significance.
- Standard effect size measures can help in a wide variety of situations.
- But don’t get carried away and reach for them automatically.
- The main objective is to clearly explain how important the magnitude of the effect is.

Guidelines for Statistical Reporting

Guidelines for Statistical Reporting

- Communicating results is a *key* part of statistical analysis.
- In this class, in other classes, and in your organization, you will be expected to submit your analysis as a written report.

Guidelines

1. A statistical analysis is a written argument.
2. If you don’t have something to say about some output, don’t display it.
3. Document every decision that you make.
4. Identify features of the population that should be reflected in statistical models.
5. Be clear about the difference between the sample and the population.
6. The code is a part of the argument.