Significance Tests

Paul E. Johnson[1] [2]

[1]Department of Political Science

[2]Center for Research Methods and Data Analysis, University of Kansas

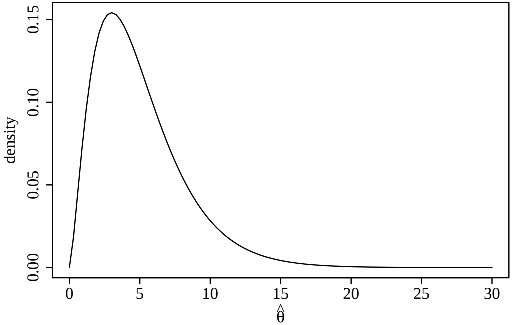
May 17, 2018

What is this Presentation?

1. Hypothesis Testing
2. Tails
3. p-value.
4. T Test Examples

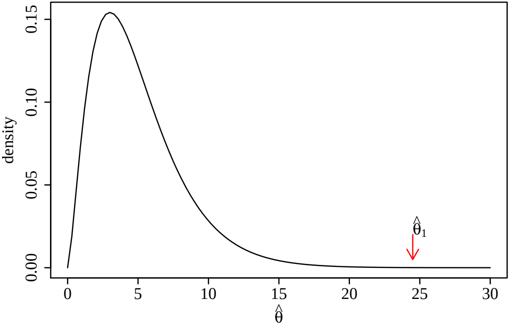
Here is the Big Idea

- Suppose I told you the sampling distribution of an estimator looks like this.
- The most common estimate is around 5, and estimates above 20 are almost never going to happen



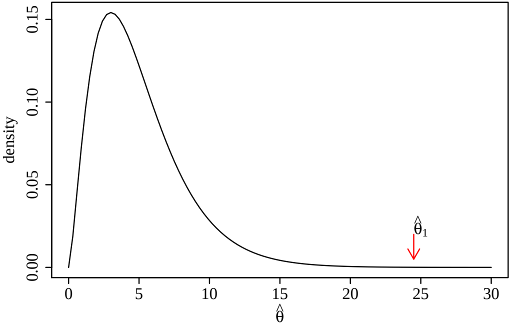
You draw a sample and calculate one estimate, $\hat{\theta}_1$

- $\hat{\theta}_1$ is *extreme! A VERY UNLIKELY THING HAS OCCURRED*



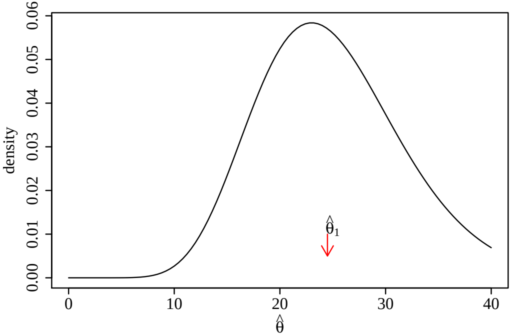
What should you conclude?

- Your research assistant does some calculations. The chance of an outcome as great or greater than $\hat{\theta}_1$ is 0.000001
- What should you conclude?
  - Something nearly impossible happened. (?)
  - The premise that "this is the sampling distribution of $\hat{\theta}$" was wrong from the start.



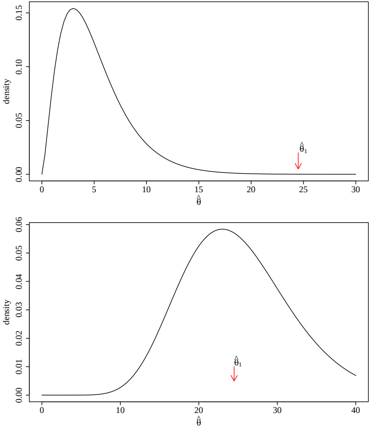
Here is the Big Idea

- Maybe the "true" sampling distribution is more like this
- In which case $\hat{\theta}_1$ is a completely ordinary, common occurrence.



The Big, Dramatic Question

- Should the evidence from one sample lead you to reject the premise on top?
- A "Hypothesis Tester" says "yes". A too-unlikely thing occurred. Reject the top model.
- Hypo tester does not tell you what you ought to do instead, but we know what we reject.



Hypo Testing Terminology

Parameter: θ is a "number" that with we estimate with $\hat{\theta}$.

Set Up A Decision Rule.

Null Hypothesis: A claim about the true value of θ,often called θ_0 ("Theta sub naught")

Alternative Hypothesis: Usually just a logical rejection of the Null Hypo.

More formal Terminology

Null Hypothesis: $H_0 : \theta = \theta_0$:

The presumed value of θ is θ_0.

Alternative Hypothesis: $H_A : \theta \neq \theta_0$

If we reject H_0 because $\hat{\theta}_1$was "unlikely", that means we think θ_0 was wrong. Then we are supposed to accept H_A.

Weakness in Hypo Testing Paradigm

The "alternative hypothesis" is not informative. Its Unsatisfying. We rejected θ_0, but then what? The "frequentist statistical paradigm" (which I'm teaching you now) does not help us too much there.

Define Statistically Significant

- If $\hat{\theta}$ is “really far” from the value supposed under the Null Hypothesis, then we reject the Null.
- Such a finding is “statistically significant”.

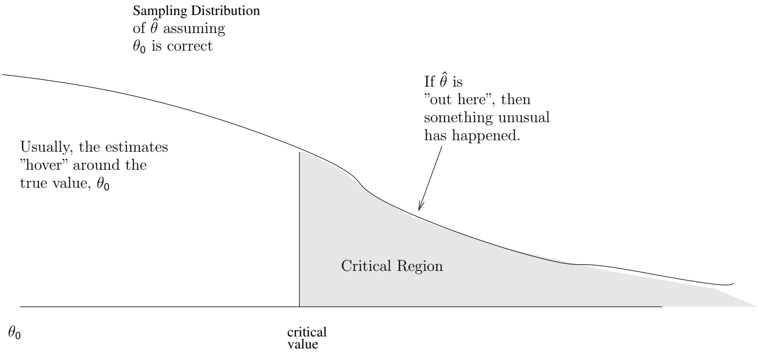
Example:

$H_0 : \mu = 7$

$H_A : \mu \neq 7$

- If the estimate $\hat{\mu}$ is ”far” from 7 (in a sense explained below), then the estimate is “statistically significantly different from 7.”
- Casual language often used
  - the difference between $\hat{\mu}$ and 7 is “statistically significant,”
  - $\hat{\mu}$ is “statistically significant.”

3 Steps of Hypothesis Testing



Step 1. Create A Sampling Distribution for $\hat{\theta}$, assuming θ_0 is correct

Step 2. Mark the "critical value" (and critical region) that holds, say, 5% of the area.

3 Steps of Hypothesis Testing ...

Step 3. Check if the sample estimate $\hat{\theta}$ is in “critical region”

Need to match estimators with probability density functions

- We have distributions like “Normal” “t” “χ^2” and “F”.
- We need a statistician to help us figure out which one matches up with our formula $\hat{\theta}$.
- You will develop intuition: comparison against the Normal or t will be “automatic”
- Comparison against the χ^2 will become crystal clear in the course on categorical data analysis.
- Comparison against the F will be needed in ANOVA and maximum likelihood.

Try this R code

- I've fiddled with R code to make these null hypo probability plots.
- There's a function in the first part that you can use to draw nice looking plots that highlite the extremes.
- The first part, by Joshua Wiley (psych grad student, UCLA), is pretty good, the other new versions of that function are just me trying to be cute.

`http://pj.freefaculty.org/R/WorkingExamples/plot-critical_regions-1.R`

Use T

Use T distribution to evaluate $\hat{\theta} - \theta_0$

- For many hypothesis tests, the test statistic is the difference between the estimate and the null:

$$\hat{\theta} - \theta_0$$

- We calculate a standard error of $(\hat{\theta} - \theta_0)$, which is the same as the standard error of $(\hat{\theta})$ (because θ_0 is a "constant", it does not affect variance).
- The ratio follows a T distribution

$$\hat{t} = \frac{\hat{\theta} - \theta}{\textit{standard error}\ (\hat{\theta})}$$

- T can be "one sided" or "two sided" (draw on blackboard)

T, F, and χ^2 are the most common sampling distributions

- The χ^2 distribution is used as a one tailed test. Usually, it is used to find out if a sum-of-squared "whatevers" is bigger than expected.
- Example.
  - The true variance is σ^2.
  - The unbiased sample variance is $\widehat{\sigma^2} = \frac{\sum(x_i - \bar{x})^2}{N-1}$.

$$\textit{The ratio } \frac{\sum(x_i - \bar{x})^2}{\sigma^2} \textit{ is distributed as } \chi^2_{(N-1)}$$

- if σ^2 were true, then the sum of squares would be approximately $N\sigma^2$.

The Alpha Level: Type I Error

Alpha level: α is the "risk" we are willing to take that H_0 is "true" and yet we (mistakenly) reject it.

If H_0 is correct, and we reject it, that is called a Type I error.

Statistical Significance: If the chance of making a mistake is smaller than α, and we reject H_0, a result is called "statistically significant".

Type 1 = alpha, Type II = beta

Type I error: H_0 is true, but our test leads us to reject it. The chance of this is α (alpha).

Type II error: H_0 is false, but our test does not reject it. The chance is β (beta).

Power: The "statistical power" of a test statistic refers to its ability to avoid Type II error. Power is not given enough emphasis in political science research, mainly because it is technically much more difficult to measure.

Type I error is called "α error".

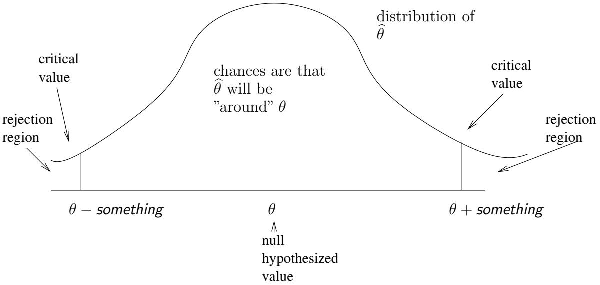
Type II error is called "β error".

In many statistical contexts, the α value is much easier to measure and evaluate, and many of us have been trained by people who do not put enough emphasis on β error.

Two-Tailed Test

Definition

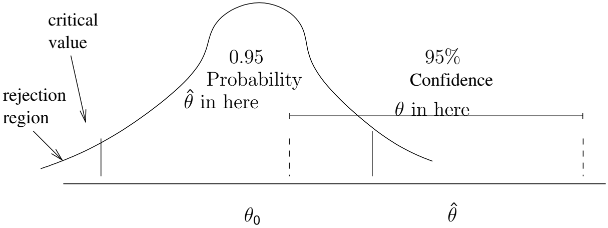
H_0 can be rejected if $\hat{\theta}$ is either “too high” or “too low”.


distribution of $\widehat{\theta}$
critical value
critical value
rejection region
rejection region
chances are that $\widehat{\theta}$ will be "around" θ
$\theta - \textit{something}$
θ
$\theta + \textit{something}$
null hypothesized value

Symmetric Two-Tailed Test

- Note area of “rejection region” on each side is $1/2 * \alpha$.
- Often, but not always, the null hypothesis is 0, symbolizing “no effect” or “no difference” for a variable.
  $H_0 : \theta = 0$
  $H_A : \theta \neq 0$

How is that different from a confidence interval?



- Important to note, the CI does not imply “shape”, it is just a flat interval.
- Width of “H_0 accepted region” is equal to the width of the Confidence Interval for some estimators (means, slope coefficients)

A Little Wrinkle

- CI and Hypo test: we might casually say “take an interval ’this wide’ and move it sideways.”
- That is not technically correct for all kinds of hypo tests. (Verzani, Intro Stat W/R, p. 218).
    - In hypo-testing, the standard error can be based on the null hypothesized value of the parameter.
    - For some models, we “know” the standard deviation of $\hat{\theta}$. E.g, for a proportion. The given value of θ determines uncertainty–there’s nothing to estimate.
    - (Recall, the test is always conducted “under the null” hypothesis)
    - In the creation of a confidence interval, the standard error was based on the estimated variance of the data.

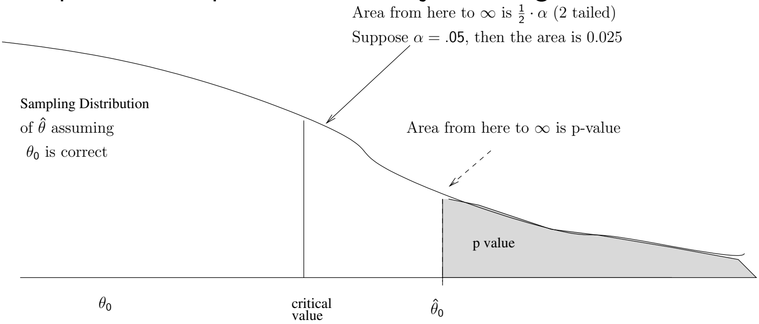
Definition of p-value

- 1. Calculate the test statistic $\hat{\theta}$.
- 2. Place $\hat{\theta}$ into the sampling distribution, assuming θ_0 is true.
- 3. Figure out how likely it might be that we would find a value of $\hat{\theta}$ more extreme than the observed value. This is usually done with a two-tailed test in mind, so you imagine you are calculating the chance that an estimate would be more extreme than $\hat{\theta}$ or $\hat{\theta}$.

The chance of observing a 'more extreme' value is known as the p-value.

Compare p and α

The p-value Compared to the α rejection region


Area from here to ∞ is $\frac{1}{2} \cdot \alpha$ (2 tailed)
Suppose $\alpha = .05$, then the area is 0.025
Sampling Distribution
of $\hat{\theta}$ assuming
θ_0 is correct
Area from here to ∞ is p-value
p value
θ_0
critical value
$\hat{\theta}_0$

I'm cautious about p-values because

- Emphasis on the p-value can lead you to some misinterpretations.

Compare p and α ...

- One common problem: scholars fish through data sets, scanning many estimated $\hat{\theta}$, looking for one with a small p-value.
    - Problem: H_0 is true, but 5% of the time we will observe and estimate with a p-value smaller than 0.05.

In a Perfectly Honest Setting, This Happens.

1. Construct the sampling distribution
   1. Select the alpha value
   2. Find the critical values of the test statistic (get the "rejection regions")
2. Calculate $\hat{\theta}$. If it exceeds the critical value, then reject the null hypothesis.
3. My opinion: Do not concern yourself with the magnitude of the test statistic, except to find out if it is in the rejection region.

NOTE: the observed "p-value" plays no role in this analysis

Some Silly People Do This

- If an estimate is to be 'statistically significant', it is of course necessary that $p < 0.05$.
- If the p value is smaller, say 0.0232, the proponents of the p-value want to emphasize the fact that they found a result that is "more unusual", somehow "more significant". *Don't be silly*.
- It would be correct to say, "the estimated value $\hat{\theta}$ would be statistically significantly different from θ_0 even if the standard error were higher" or "even if the sample size were smaller." But that doesn't mean it is "more significant". Saying "more significant" is one way to reveal yourself as a silly person. I think.
- I'd say "capable of rejecting the null hypothesis at a smaller value of α" if I had to discuss a p-value.

Statistical Significance and Substantive Significance.

- Effect Size: slang for measures of importance of a variable's effect
- If $\hat{\theta} - \theta_0$ is 0.000000001 (very small), it may still be the difference is "statistically significantly different from 0".
- But that doesn't mean it is important?
  - "important for public policy makers"
  - "useful for scientists"
  - "helpful to graduate students".
- "Substantive significance". Can we give a persuasive interpretation of the importance of the difference $(\theta - \hat{\theta})$?

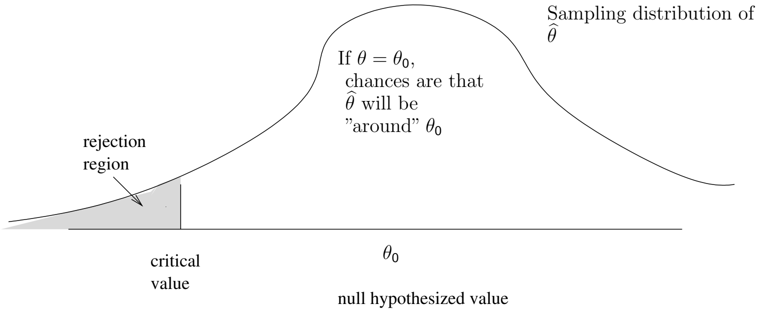
How Can Rejecting H_0 Be Not-Valuable?

Cases in which we reject θ_0 but there's no substantive satisfaction in it.

- Miniscule effects: If $\hat{\theta} - \theta_0$ is 0.000000001.
  - If that is "pounds lost per hour on treadmill", I don't give a hoot if the difference is bigger than 0.
  - Large sample sizes can make standard errors so small that any trivial effect $(\theta - \hat{\theta})$ is "statistically significantly different from 0"
- Ridiculous choice of θ_0. You can magnify $\hat{\theta} - \theta_0$ by choosing θ_0 in a ridiculous way.
  - Statistical significance may mean only that the null hypothesis was grossly wrong, not that $\hat{\theta}$ is especially meaningful.

One-Tailed Test

A one-tailed test is designed so that H_0 is rejected only if the estimated value is at one extreme.


Sampling distribution of $\widehat{\theta}$
If $\theta = \theta_0$, chances are that $\widehat{\theta}$ will be "around" θ_0
rejection region
critical value
θ_0
null hypothesized value

One-Tailed Test

- I'd say it should be stated

$$
\begin{aligned}
H_0 :&\theta \geq 0\\
H_A :&\theta < 0
\end{aligned}
$$

- Some texts want to put the null hypothesis as $H_0 : \theta > 0$.

Estimates of Means

- If we previously believed that the average Intelligence Quotient of American teenagers was 111, but our current sample (N=100) estimate is 123, is it likely that our previous belief is inconsistent with the process currently generating the data?
- $H_0 : \mu = 111$
- H_A: $\mu \neq 111$

Here's Where the t distribution comes into play

- The standard error (estimated standard deviation of the mean) is

$$std.err.(\hat{\mu}) = \frac{1}{\sqrt{N}} std.dev.(x)$$

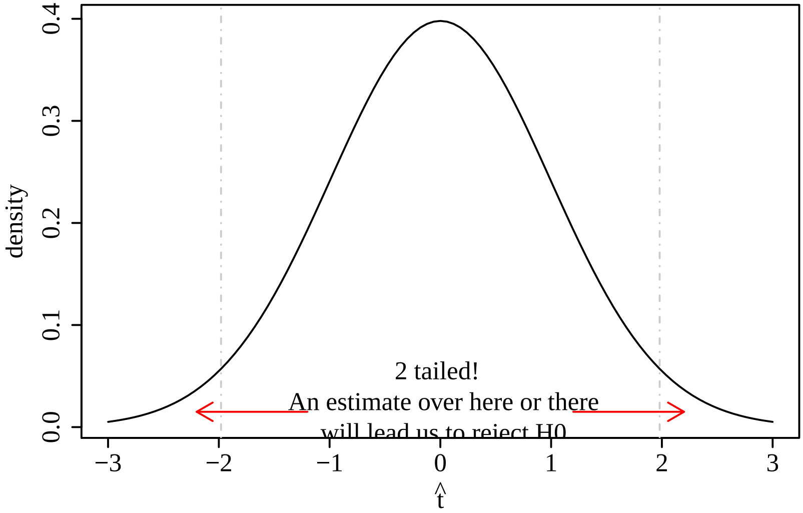
- The test statistic (which plays the role of $\hat{\theta}$ in this story) is the ratio:

$$\hat{t} = \frac{\hat{\mu} - \mu}{std.err.(\hat{\mu})}$$

T distribution

The t distribution is centered on 0

T distribution ...



Consider the IQ test:

H_0: $\mu = 111$ and the estimates are $\hat{\mu} = 123$ and $\widehat{\sigma}_{\hat{\mu}} = 7$

H_A: $\mu \neq 111$

$$t = \frac{123 - 111}{7} = \frac{12}{7} = 1.71$$

That is “almost statistically significant”, but not quite.

```
pt( 12/7, df = 95, lower = FALSE)
```

```
[1] 0.04486855
```

The chance of an estimate more extreme than 1.71 is 0.044, but it is only “statistically significantly different from 0” if it is smaller than 0.025 (this is a 2-tailed test).

Groveling for Statistical Significance

If you are desperate to find a "statistically significant result", you have 2 options.

1. Get a bigger sample. The larger N will make the denominator shrink. That is the sense in which, if you have a big enough sample, even very small, substantively unimportant differences can be statistically significant.
2. Switch to a 1 tailed test. The critical value of t in a two tailed test is 1.96, but in a one-tailed test it is 1.65.

Because of your temptation to follow route 2, one-tailed tests have gotten something of a bad reputation.

Difference of Means

- If you have two samples, you are interested to find out if the “true” mean in one is equal to the “true” mean in the other. $H_0 : \theta_1 = \theta_2$
- That is equivalent to $H_0 : \theta_1 - \theta_2 = 0$
- So the test statistic from the data is

$$\widehat{\theta_1 - \theta_2} = \hat{\theta}_1 - \hat{\theta}_2 \tag{1}$$

Difference of Means

- Hence in the test statistic, the
    - numerator is the difference between two estimates
    - denominator is an estimate of the standard error of that difference.

$$\hat{t} = \frac{(\widehat{\theta_1 - \theta_2}) - (\theta_1 - \theta_2)}{std.err.(\widehat{\theta_1 - \theta_2})} \tag{2}$$

- Of course, if the null is that the two parameters are the same, $\theta_1 - \theta_2 = 0$, then:

$$\hat{t} = \frac{\hat{\theta}_1 - \hat{\theta}_2}{std.err.(\widehat{\theta_1 - \theta_2})} \tag{3}$$

Standard error of the difference

Calculate $s.e.(\widehat{\theta_1 - \theta_2})$

- Recall

$$Var(\hat{\theta}_1 - \hat{\theta}_2) = Var(\hat{\theta}_1) + Var(\hat{\theta}_2) - 2Cov(\hat{\theta}_1, \hat{\theta}_2) \quad (4)$$

- The standard error in the denominator can be calculated under various suppositions (same variance within 2 groups, or not).
- The current advice in R is that we should assume that the two samples are drawn from populations that have different variances, so output from ?t.test will explain that Welch's method is used to calculate variance.

Estimated Regression Slope

- Suppose a model has been run and it says the effect of education (years) on starting salary (dollars) is $2843. That is to say, a linear model says that each additional year in school seems to predict a 2843 rise in salary.
- Ask, is 2843 a substantial effect? Perhaps it would be 0 if you drew another sample, or even -2843. We surveyed 1000 people, perhaps a different sample would be grossly different.
- Hypo testing approach. Push this into a t-test format. The df is 998 (don't worry why)
  - $H_0 : b = 0$
  - $\hat{t} = (\hat{b} - 0)/std.err.(\hat{b})$
  - Suppose $std.err.(\hat{b}) = 1000$
  - $\hat{t} = 2843/1000 = 2.843$