

Hypothesis Testing: What's the Alternative?

Danny Kaplan

Macalester College, Saint Paul, Minnesota

Stat Chat, Jan. 26, 2010

Outline

The role of the alternative hypothesis in introductory statistics.

- 1 A consensus view on the alternative hypothesis?
 - 1 Examples from textbooks: the “anything but” hypothesis.
 - 2 Claim: The alternative, in introductory statistics, is oriented mainly to discuss the distinction between one-sided and two-sided tests.
- 2 Opinion: It’s a bad idea to teach one-sided versus two-sided tests:
 - 1 It originates, I think, in a mathematical result that’s irrelevant to our students: the Neyman-Pearson lemma.
 - 2 It doesn’t really help, and people don’t stick to the rules anyways.
 - 3 It introduces prior views in somewhat arbitrary ways. If we’re going to be Bayesians, let’s be proper Bayesians.
- 3 Claim: A more important role for the alternative hypothesis relates to power.
- 4 Using a specific alternative hypothesis allows one to talk about power in concrete ways that are pretty easily understood.
 - 1 Abstract simulations: means, proportions.
 - 2 Simulations of mechanisms.
 - 3 Resampling: The population looks like our data.

Some Textbook Definitions I

Freedman, Pisani, and Purves, *Statistics* 3/ed.

The null hypothesis expresses the idea that an observed difference is due to chance. To make a test of significance, the null hypothesis has to be set up as a box model for the data. The alternative hypothesis is another statement about the box; it says that the difference is real. — p. 479 (in a box)

Utts and Heckard, *Mind on Statistics*

*The **alternative hypothesis**, represented by the symbol H_a , is a statement that something is happening. In most situations, this hypothesis is that what the researcher hopes to prove. It may be a statement that the assumed status quo is false, or that there is a relationship, or that there is a difference. — p. 315*

Some Textbook Definitions II

Agresti and Franklin, *Statistics: The Art and Science of Learning from Data*

The **null hypothesis** is a statement that the parameter takes a particular value.

The **alternative hypothesis** states that the parameter falls in some alternative range of values. — p. 369:

Some Textbook Definitions III

DeVeaux, Velleman, and Bock, *Intro Stats* 2nd ed.

The alternative hypothesis proposes what we should conclude if we find the null hypothesis to be unlikely. — p. 465

*The **alternative hypothesis**, H_A , contains the values [note the plural] of the parameter we accept if we reject the null. If we reject the null hypothesis, then we accept the alternative hypothesis. In the Coke vs. Pepsi example, our null hypothesis is that $p = 0.50$. What's the alternative? We would be interested in learning that either cola was preferred. In terms of the parameter, we can write $H_A : p \neq 0.50$. If the data convince us that we should reject the null hypothesis, we would accept the alternative. — p. 453*

The “Anything but” Hypothesis

The alternative introduced in this way is an “anything but” hypothesis — anything but the null.

This has problems:

- It encourages students to think too abstractly. If we believe that applied statistics requires some knowledge of the field of application, why encourage a statement of the alternative that has absolutely no contact with the field of application.
- We’re starting students off with a mathematically complicated form of a hypothesis: a “compound hypotheses,” not a simple statement of what the world is like.
- A consequence of the mathematical complexity of the “anything but” hypothesis is that it becomes difficult to estimate **power**.

Compound Hypotheses are Hard I

A quote from E.L. Lehmann, one of the authors of *Statistics: A Guide to the Unknown* 3/ed:

*Why these [tests] rather than some others? This is the question that Neyman and Pearson considered and which (after some preliminary work in Neyman and Pearson 1928) they later answered (Neyman and Pearson 1933a). Their solution involved not only the hypothesis but also a class of possible alternatives and the probabilities of two kinds of error: false rejection (Error I) and false acceptance (Error II). The “best” test was one that minimized P_A (Error II) subject to a bound on P_H (Error I), the latter being the significance level of the test. They completely solved this problem for the case of testing a simple (i.e., single distribution) hypothesis against a simple alternative by means of the Neyman-Pearson lemma. **For more complex situations, the theory required additional concepts, and working out the details of this program was an***

Compound Hypotheses are Hard II

important concern of mathematical statistics in the following decades. [?, p.1243]

Maybe a program that occupies decades of mathematical statistics is not one we want to emphasize in introductory statistics.

A Historical Speculation

Fisher and Pearson disagreed on the criteria that would lead to the rejection in a χ^2 test. Fisher wrote, “If P is between .1 and .9, there is no reason to suspect the hypothesis being tested.” (Fisher 1925, *Statistical Methods for Research Workers*, p. 71, quoted in [?, p. 9])

We typically do a one-tailed test on χ^2 , following Pearson’s view that we should not reject a “graduation curve” [a model of a distribution] because it is too close to the data. Fisher points out that if the hypothesized model is correct, then $p = 0.999$ is just as unlikely as $p = 0.001$ — suggesting a two-tailed test. The Pearson approach is justified if our alternative is one-tailed: another distribution that fits worse than the hypothesized one. A similar situation applies to F tests. We always do a one-tailed test, because our alternative is that the model term is more closely related to the response variable than would occur by chance. But very small F , suggesting close orthogonality of the model term to the response, are also very unlikely.

My speculation is that the one-sided versus two-sided dispute in χ^2 was resolved in theoreticians’ minds by the Neyman-Pearson lemma and so

What's the alternative for? I

In these books, the purpose of the alternative is to make a decision: Do we perform a one-sided or two-sided test?

Agresti and Franklin, *Statistics: The Art and Science of Learning from Data*

Their index entry:

Alternative hypothesis, 369

 one-sided, 374, 384-385, 392-393, 436

 two-sided, 374, 384-385

What's the alternative for? II

Utts and Heckard, *Mind on Statistics*

Example 11.1: Are Side Effects Experienced by Fewer than 20% of Patients?

Null: 20% (or more) of users will experience side effects.

Alternative: Fewer than 20% of users will experience side effects.

Example 11.2: Does a Majority Favor the Proposed Blood Alcohol Limit?

H_0 : $p \leq .5$ (not a majority)

H_a : $p > .5$ (a majority)

The Neyman-Pearson Lemma

This gives a scalar function, the ratio of likelihoods, to use to set the region for rejecting the null based on some data x :

$$\Lambda(x) = \frac{L(\theta_0|x)}{L(\theta_1|x)} \leq \eta \quad (1)$$

(Note: The convention seems to be $L(\theta|x)$ is the likelihood of x given θ , written in a way that's backwards from the standard probability notation.) Such a region defined by η will give the most powerful test of size α .

Trying to understand the reasoning

Convention is to fix α and accept the power that results.

Neyman-Pearson says that picking the rejection region based on Eq. 1 will give the most powerful result.

Since α is fixed, isn't the most powerful test the best one?

An (Emotional) Counter-Example I

You're studying a new drug that you believe might be more effective than the old one. If it's not, there's no use.

You have done a preliminary study with $n = 4$ which gives $\hat{\theta} = 1$ with a standard error of $SE_{\theta} = 1$. This gives $t = 1$ and therefore a p-value that is large (0.196 one sided, 0.391 two sided).

You're interested in showing that $\theta > 0$. To establish once and for all the utility of the treatment, you are going to do a study with $n = 100$.

Should the hypothesis test be done in a one-sided or two-sided way?

One-sided: All we're interested in is $\theta > 0$. By doing a one-sided test, we increase the power and make it easier for us to find significance.

Two-sided: The power is large anyways. The standard error will likely be $s.e._{\theta} = 1 \times \frac{\sqrt{4}}{\sqrt{100}} = 0.2$, so t will be around 5 — highly significant even for a two-sided test.

The reason to do a one-tailed test is to increase the power. No point in increasing it if it is already very large.

An (Emotional) Counter-Example II

We go ahead and do a one-sided test. Playing by the rules of the game, when we observe $\hat{\theta} = -1$ with $t = -5$ ($p < 0.0000006$, one sided). It's tiny, but we are unable to reject the null hypotheses, bound by our pledge to do a one-sided test.

But there are consequences to this. If not for our insistence on a one-sided test that we didn't need in the first place, we would have rejected the null and been able to publish that the treatment is deleterious. Now we can't. But if people continue to investigate this, not only will they be wasting effort, but they will be doing harm.

A Philosophical Aside: Falsifiability and Science I

Karl Popper, following others, has argued that a hypothesis must be falsifiable, and that one cannot regard a proposition or theory as scientific if it does not admit the possibility of being shown false. — Wikipedia

- Often this criterion for “scientific” is applied to areas where it’s hard to imagine what sort of evidence would be accepted as plausibly falsifying the hypothesis. Example: Creationism.
- But let’s apply it to statistical hypothesis testing.

In a statistical hypothesis test, the p-value is used to indicate the plausibility of evidence that the null hypothesis should be rejected.

- If a hypothesis is falsifiable, we should be able to imagine a set of observations that would lead to a p-value small enough to be compelling.

A Philosophical Aside: Falsifiability and Science II

- The anything-but form of the alternative allows the world to be just slightly different from the Null. That is, $H_a : \mu = \mu_0 + \epsilon$ and ϵ can be arbitrarily close to zero.
- In such a world, the required sample size n for non-trivial power is proportional to $1/\epsilon^2$. This can be infinite.
- So the anything-but alternative admits the possibility that the Null cannot be rejected with any amount of data. It's potentially non-falsifiable, hence not scientific.

It's the statistical process that dictates this, since the alternative plays a role in determining whether we believe the null is falsifiable.

Now imagine a statistical procedure that insists that a definite statement be made of the alternative, as opposed to the anything-but hypothesis. Now ϵ is finite. Consequently, the n required to reject the null with non-trivial power is finite.

Costs and Benefits to One-Sided Tests I

- Benefit: Increases the power (Neyman-Pearson Lemma)
- Cost (for teaching): Adds complexity.
- Cost (for research): Encourages fraud — ex post facto trimming of the p-value. Statistical circumcision. $p < 0.10$ is satisfactory, so long as you can justify a one-sided test.

Question: How much does it increase the power? Is it worthwhile. My calculations indicate that for a study with 80% power, giving up the added power of the one-sided test is compensated by an approximately 20% increase in sample size.

That's not nothing. But if we are going to be worrying about factors of 20% in sample size, we had better have already covered the matters that lead to order of magnitude estimates of sample size. This we typically have not done.

Costs and Benefits to One-Sided Tests II

Indeed, many statistics instructors that I talk to say that they don't cover power, and so how can a student understand the benefit of a one-sided test anyways.

First introduce the concept and give them the tools to get the first digit right, then worry about calculations on the order of 20%.

How Much Data does a Two-Sided Test Cost I

Assumptions:

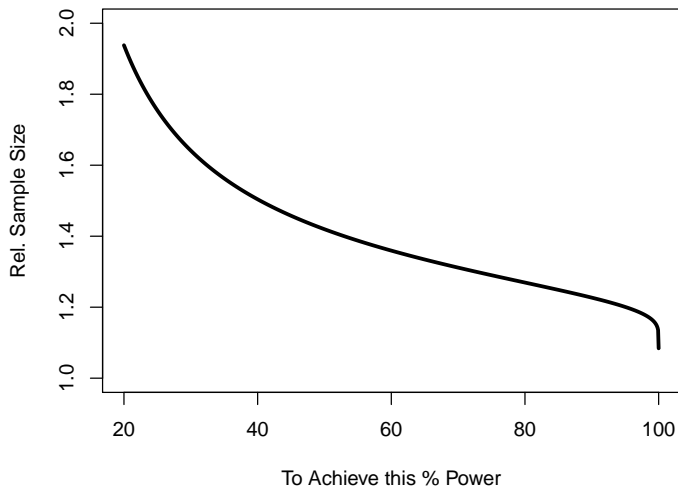
- Both Null and Alternative sampling distributions are normal with same SE.
- We reject when $x/SE > 1.645$ for one-sided and 1.960 for two-sided.
- One-sided test has sample size n , two-sided test will have sample size γn — larger to make up for the lost power due to the use of a two-sided test.
- Denote by $\tau_{1-\beta}$ the distance the rejection threshold needs to be from μ_A in order to achieve the desired power, e.g., $1 - \beta = 0.80$ power needs $\tau_{0.2} = 0.84$.

How Much Data does a Two-Sided Test Cost II

$$\gamma = \left(\frac{1.960 + \tau_{1-\beta}}{1.645 + \tau_{1-\beta}} \right)^2$$

How Much Data does a Two-Sided Test Cost III

Relative Sample Size for 2-Sided Test on 1-Sided Alternative



What to Teach about Power?

The motivation for the one- vs two-sided distinction is power.

What things affect power? In rough order of importance.

- Sample size.
- The hypothesized value for the alternative hypothesis. (If the effect is small ... don't bother.)
- Covariates.
- Appropriate experimental design, e.g. cross-over
- Precision of measurements.
- Multiple tests.
- One- vs two-sided.

We're teaching the last item on this list and largely ignoring the preceding items.

Example: Vitamin D and Blood Pressure, part I

Is there a link between Vitamin D deficiency and high blood pressure?

Approach: Collect vitamin D levels and BP from n people from a suitable population. See if we can reject the Null that D and BP are unrelated.

The Anything-But Alternative

We will do a one-sided test, since we believe that vitamin D deficiency causes **high** blood pressure.

Or we might not believe this. Wouldn't it be interesting to hear that vitamin D deficiency protects against high BP?

The one-sided test becomes a temptation, something we will fall back on if our data don't work out so well. If we're honest, we'll write something like, "We observe a trend toward higher BP with vitamin D deficiency ($p < 0.10$)."

And what should be our sample size?

Example: Vitamin D and Blood Pressure, part II

The Specific Alternative

Read the literature and figure out what would be a reasonable relationship between vitamin D and BP. Set this as our alternative.

Then figure out how often, in a world in which the alternative is true and with a proposed sample size n , we would reject the null.

Adjust n to get an acceptable power, keeping in mind the cost of additional subjects.

Objections:

- What if there is no literature?

A Response: There must be a reason that we have proposed the study. Perhaps we want to use as an alternative, a common level of vitamin D deficiency and a clinically meaningful increase in vitamin D.

Example: Vitamin D and Blood Pressure, part III

- We're teaching statistics, not public health or medicine.
Some Responses: We should be modeling for our students the idea that you should know something about the field of application if you are going to work in it.
If using “real data” is important in teaching statistics — as GAISE recommends — then the context in which that real data is embedded is also important.

The Calculations I

Supplies needed:

- A simulation of the alternative. This includes a specific parameter value for the alternative (`magic`, which can be set as desired), and a way to set the sample size.

```
> s = run.sim( vitaminD, magic=5, n=4 )
```

```
> s
```

```
  race  D systolic
1    W 62      125
2    W 73      146
3    W 78      129
4    W 75      170
```

- A means to see if the null can be rejected.

```
> summary( lm( systolic ~ D, data=s) )
```

The Calculations II

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	53.7329	136.0354	0.39	0.7310
D	1.2329	1.8828	0.65	0.5798

- A way to collect the relevant p-value

```
> p.value(summary(lm(systolic ~ D, data=s)), 2)
[1] 0.5798
```

- A way to carry out multiple random trials

```
> do(10)*p.value(summary(
+   lm(systolic~D,data=run.sim(vitaminD,magic=5,n=4))), 2)
[1] 0.5617 0.3348 0.6948 0.6705 0.3270 0.3543 0.7649
[8] 0.2257 0.8745 0.2137
```

NOTE in DRAFT: I should arrange the p.value syntax to look like this: p.value(2)@lm(...)) and define p.values to be p.values(all)

Computing the Power for $n = 100$, part I

- 1 Collect the p-values from many, say $k = 1000$ trials, each done with $n = 100$.

```
> k = 1000
```

```
> s = do(k)*p.value(summary(  
+   lm( systolic ~ D,data=run.sim(vitaminD,magic=5,n=100))
```

- 2 See how many of these trials resulted in a “small” p-value.

```
> table( s <= 0.05 )
```

```
FALSE  TRUE
```

```
  895   105
```

The power is about 10%. That’s terrible! We need a larger sample size.

Let’s try $n = 500$.

Computing the Power for $n = 100$, part II

```
> s = do(k)*p.value(summary(  
+   lm( systolic ~ D,data=run.sim(vitaminD,magic=5,n=500))), 2  
> table( s <= 0.05 )
```

```
FALSE  TRUE  
  766   234
```

Still horrible. We need more data! How about $n = 5000$?

```
> s = do(k)*p.value(summary(  
+   lm( systolic ~ D,data=run.sim(vitaminD,magic=5,n=5000))),  
> table( s <= 0.05 )
```

```
FALSE  TRUE  
   26   974
```

That's great for power, but it's going to be expensive to run the study. Let's back off a little bit and find something that might fit our budget.

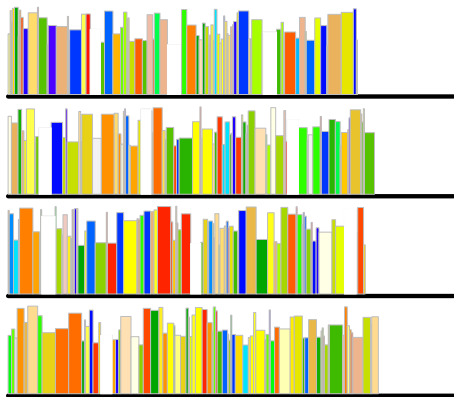
A Classroom Activity

To illustrate the importance of a formal approach to random sampling, I'm going to send students to the library to select books at "random." We want to get an estimate of the average page count of a book. Then we will compare their results to the results from a census of all the books.

Question: How many books should I have them select for their samples for there to be almost certain (say, 99%) that their confidence interval will **not** include the population parameter?

The Classroom Activity: A Simulation, I

```
> books = select.books(10)
```



The Classroom Activity (cont). I

Each student selects 10 books and gets something like this, which includes information on all the books as well as a flag saying which were selected.

```
> head(books)
  pages shelf selected
1    284     1         0
2    268     1         0
3    192     1         0
... and so on
304    92     4         0
305  1100     4         1
306   372     4         0
```

The students' job: Construct a confidence interval on the selected books, and see if this includes the population parameter.

The Classroom Activity (cont). II

```
> mean(books$pages)
[1] 366.1375
> mean(mybooks$pages)
[1] 636.4
> sd(mybooks$pages)
[1] 304.0845
> mean(mybooks$pages) +
  c(-1,1)*qt(.975,9)*sd(mybooks$pages)/sqrt(10)
[1] 418.8711 853.9289
```

The 95% confidence interval, 419 to 854, does not cover the population parameter, 366.

The Classroom Activity (cont). III

ASIDE: The same calculation done with modeling notation

```
> confint( lm( pages~1, data=mybooks))
              2.5 %  97.5 %
(Intercept) 418.8711 853.929
```

Or, done as a hypothesis test:

```
> summary(lm(pages-366.1375 ~ 1, data=mybooks))
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    270.26      96.16    2.811  0.0204
```

Finding the Power, part I

When I do this in class, I have each of the students run their own simulation of book selection, compute their own confidence interval, and compare to the population parameter.

- A show of hands, whose interval excludes the population parameter, indicates the power.
- The sampling process itself includes the human element — that's where the bias comes from. It might be difficult to simulate this mechanism on the computer.
- Instead, I'll adopt as the alternative hypothesis an easy but specific statement: *the population (under the alternative) looks just like my selection, but of unlimited size.*
- This can be implemented by **resampling**.

Finding the Power, part II

```
> p.value(summary(lm(pages~366.1375 ~ 1,
  data=mybooks)))
[1] 0.02035871
> p.value(summary(lm(pages~366.1375 ~ 1,
  data=resample(mybooks))))
[1] 0.003362427
> s = do(1000)*p.value(summary(
  lm(pages~366.1375 ~ 1, data=resample(mybooks))))
> table(s < .05)
FALSE  TRUE
  264    736
```

The power is about 74%. But I want 99% so that the large majority of students will be convinced of bias.

Let's try 20 books in the sample:

Finding the Power, part III

```
> s = do(1000)*p.value(summary(  
  lm(pages~366.1375 ~ 1, data=resample(mybooks,20))))  
> table(s < .05)  
FALSE  TRUE  
   20   980
```

Almost there!

In my class, I ask the students to do this calculation individually. They will get different sample sizes to achieve 99% power. That's because the calculation depends on the specific alternative hypothesis, and each of them has a different alternative, since each of them implements a somewhat different mechanism for random selection.

We talk about this and reach a compromise, such as using the largest sample size of any student in the class. It's also important to realize that in a class of 30 students, even the most extreme individual probably doesn't capture 99% of the actual population.

General Idea

The Alternative is like our preliminary study.

An Example on Sample Proportions I

We're going to take a poll in a population where we think support for a candidate is at 52% — that's the specific alternative. How large a poll is needed to have a 90% power of showing that the support is the majority. Contrast two different approaches:

An Example on Sample Proportions II

How often will our sample show a majority?

```
> alternative=.52
> sampsize=1000
> rbinom(1, size=sampsize, prob=alternative)

[1] 505

> table(
+ rbinom(1000,size=sampsize,prob=alternative) > sampsize/2 )

FALSE  TRUE
   106   894
```

A sample size of 1000 will, if the alternative is right, show a majority in about 90% of cases.

Of course, it's cleaner just to use the right operator:

```
> 1-pbinom(sampsize/2, size=sampsize, prob=alternative)

[1] 0.8914
```

Summary I

- The “anything-but” form of the alternative hypothesis addresses an issue of secondary importance — how to make our sample size 20% smaller.
- It makes it very difficult to calculate power. Without power, there is no meaningful calculation of sample size (except by rules of thumb, e.g., confidence interval should “clearly” exclude the null). So the 20% smaller doesn’t matter, since we don’t know what the sample size should be anyways.
- Power is straightforward to calculate — just count the number of trials where the p -values are < 0.05 .

Summary II

Doing this requires:

- A specific statement of the alternative hypothesis.
- A way to simulate sampling from the alternative:
 - A simulation of the system.
 - Resampling from a preliminary study's data.
 - Use of probability operators, e.g., `rbinom`
- A means to calculate p-values from a sample.
- A way to repeat this many times.
 - Many students doing it in parallel.
 - Computer iteration, e.g., `do(1000)`.