PLSC 502: "Statistical Methods for Political Research"

Hypothesis Testing
October 18, 2016

Introduction: Elements of Statistical Testing

Statistical testing has its origins in decision theory, since – unsurprisingly – what statistics are designed to do is to help us make decisions on the basis of data. Gill (1999) goes into more detail on the history of this (as do many others), but the standard Neyman-Pearson approach to statistical testing brings together four components:

1. A *null hypothesis*, usually denoted H_0,
2. an *alternative* (or *research*) *hypothesis*, usually denoted H_a or H_1,
3. a *test statistic*, which we might denote generically by θ, which is a function of the sample measurements of the data $\mathbf{X}$, and
4. a *rejection region* in the space of the sample statistic; this is usually chosen with reference to the sampling distribution of the test statistic.

Consider this example: The most recent Quinnipiac poll in Pennsylvania has Hillary Clinton receiving 47 percent of the vote, and Donald Trump receiving 41 percent, out of a random sample of 660 likely voters. Of course, what we everyone really would like to know is who will win the state on Nov. 8; that is, we care(d) about (say) Pr(Clinton's Vote Share $> 50\%$), or – equivalently – Pr(Vote For Clinton) $\equiv \pi > 0.5$. So, we might think of the relevant alternative hypothesis as:

$$H_a : \pi > 0.5$$

and the corresponding null hypothesis as

$$H_0 : \pi = 0.5$$

(For the reason(s) that we don't typically specify (say) $H_0 : \pi < 0.5$, see Wackerly et al. (2008, pp. 518-20)). Our test statistic is essentially the estimate of π from the poll (that is, $\hat{\pi} = 0.47$, with $N = 660$).

The rejection region is a bit more complicated, but basically boils down to answering the question "How far from 0.5 does $\hat{\pi}$ need to be before we reject the null hypothesis that $\pi = 0.5$?" That answer is best determined probabilistically, in a way similar to what we did for confidence intervals. That is, we can answer the question by first answering a different one: "How confident do we want to be in our decision regarding rejection?" That is, are we willing to be wrong one time in 20? One in 100? One in 1000?

Types of Errors

In making a decision regarding whether to reject the null hypothesis or not, we can make two types of mistakes.

- A **Type I error** occurs when we reject the null hypothesis and, in fact, that null hypothesis is true. Think of this as a "false positive."
- A **Type II error** occurs when we fail to reject the null hypothesis when it is not, in fact, true. Think of this as a "false negative."

Either of these two things might happen, for example, if we (by pure chance) drew a sample that was wildly unrepresentative of the population. In the case of a Type I error, imagine if, for example, we were curious to see if Geminis were more athletic than people with other star signs, and – completely by chance – the Geminis in our sample were Hope Solo, Ronaldo, J.J. Watt, Venus Williams, Picabo Street, and Steph Curry. Conversely, if we wanted to see whether African-Americans were (on average) more politically liberal in their views than whites, and our sample of African-Americans (again, completely by chance) consisted of Ward Connerly, Condoleeza Rice, Thomas Sowell, Clarence Thomas, and J.C. Watts, we might commit a Type II error. The standard presentation of this is in terms of a 2×2 table:

	Reailty / Population	
Test Statistic / Sample	H_a	H_0
H_a	True	Type I error
H_0	Type II Error	Correct

By convention, we denote:

$$\Pr(\text{Type I Error}) = \alpha$$

and

$$\Pr(\text{Type II Error}) = \beta.$$

Thus,

1. the false positive rate α is the same as the "significance level;"
2. $1 - \alpha$ is known as the "specificity" of the test in question;
3. the false negative rate is β, and
4. $\mathbf{1} - \beta$ is usually termed the "sensitivity" (or "power") of the test.

This allows us to reformulate our 2×2 table a bit:

	Reailty / Population		
Test Statistic / Sample	Positive	Negative	Frequency
Positive	True Positive (N_{TP})	Type I error (False Positive) (N_{FP})	$N_P = N_{TP} + N_{FP}$
Negative	Type II Error (False Negative) (N_{FN})	True Negative (N_{TN})	$N_N = N_{TN} + N_{FN}$
Frequency	$N_{(+)} = N_{TP} + N_{FN}$	$N_{(-)} = N_{TN} + N_{FP}$	N

If we think of the numbers of cases in our data that fall into each of these cells, we can calculate a few other useful quantities:

- Looking down the columns, the false positive / significance level α can be written as $N_{FP}/N_{(-)}$, and the false negative level β is $N_{FN}/N_{(+)}$,
- Conversely, if we look across the rows, we can calculate the *false discovery rate* as N_{FP}/N_P (that is, the fraction of all positive test results that are actually negative), and the *false omission rate* as N_{FN}/N_N (that is, the fraction of all negative test results that are actually positive).
- Finally, the *accuracy* of the test is the proportion that it "gets right;" that is, $(N_{TP} + N_{FP})/N$,

There are other quantities that are of interest in this table, but that's enough for now.

Statistical testing always involves a tradeoff between α and β; any rejection region that prevents larger numbers of false positives will necessarily result in higher frequencies of false negatives, and vice-versa.[1] In practical terms, there is almost always a greater interest in minimizing "false positives" (Type I errors), because such results are often costlier than the reverse (but note that this isn't always the case: good counterexamples are areas like drug safety testing, where the cost of incorrectly finding a safe (and effective) drug unsafe – a Type II error – can at times be much greater than the cost of finding an unsafe drug safe).

Hypothesis Testing Explained

Our knowledge of the sampling distribution of our statistics allow us to say something directly about the probability of observing what we observe (that is, the data) given a hypothesized population parameter of interest. So, in the Clinton/poll example, we know that the sampling distribution of the proportion of Clinton voters is distributed in repeated sampling as

[1]Note, however, that $\alpha \neq 1 - \beta$.

$$\hat{\pi} \sim \mathcal{N}(\pi, \sigma^2_{\hat{\pi}}).$$

It is standard practice to think of our null hypothesis as a "guess" as to what π is. We suggested above that such a guess might be $\pi = 0.5$ – an even split of PA voters between the two candidates. From this, we can develop a rejection region for the null hypothesis. Since we know that $\sigma^2_{\hat{\pi}} = \frac{\sigma^2}{N}$, and (in the Bernoulli case) $\sigma^2 = \pi(1-\pi)$, we can calculate

$$\begin{aligned} \hat{\sigma}^2 &= 0.470(1-0.470) \\ &= 0.249 \end{aligned}$$

and

$$\begin{aligned} \hat{\sigma}^2_{\hat{\pi}} &= \frac{0.249}{660} \\ &= 0.00038. \end{aligned}$$

Thus, the sampling distribution of our test statistic $\hat{\pi}$ under the null hypothesis $\pi = 0.5$ is

$$\hat{\pi} \sim \mathcal{N}(0.5, 0.00038).$$

By converting our test statistic $\hat{\pi}$ to a z-score, we can then use a standard normal distribution in place of this one:

$$\frac{\hat{\pi} - \pi}{\sigma_{\hat{\pi}}} = Z \sim \mathcal{N}(0, 1).$$

We can then use the standard normal distribution to calculate the probability of rejection. Call z_α the value of Z that corresponds to confidence level α. Formally, our decision rule is

$$\text{Reject } H_0 \text{ if } Z \geq z_\alpha.$$

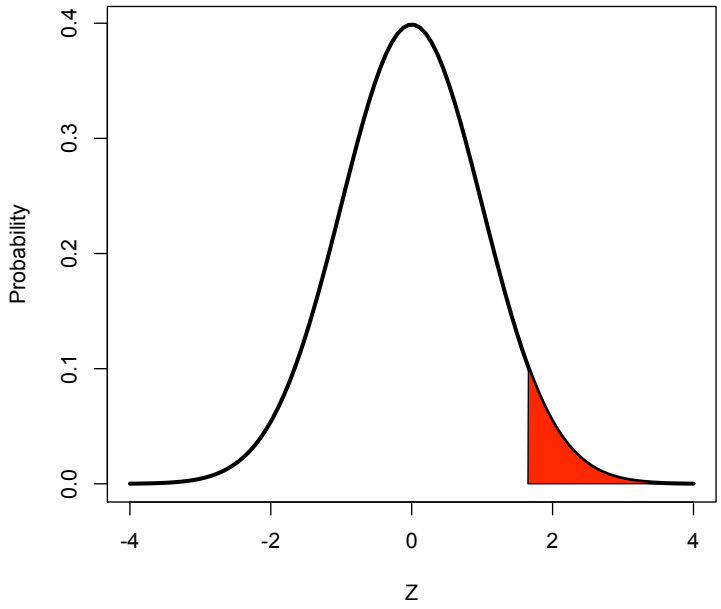
So, suppose we want to be 95 percent confident in our inference – that is, we set $\alpha = 0.05$. For the test $\Pr(Z > 0)$ (that is, for a test with a "directional" component – more on this below), we know that 95 percent of the area of the standard normal density lies below the value $z = 1.65$. Thus, if $Z \geq 1.65$, we would reject the null hypothesis that $\pi = 0.5$ at the 95 percent confidence level. Conversely, if $Z < 1.65$, we would fail to reject the null hypothesis.

Here, $Z = \frac{0.470 - 0.50}{0.0195} = -1.54$, indicating that we cannot reject the null hypothesis $\pi = 0.5$ at the 95 percent level of confidence. We could do a similar test for a different H_0, for example, $H_0 : \pi = 0.40$. That would yield:

$$\begin{aligned} Z &= \frac{0.47 - 0.40}{0.0195} \\ &= 3.59 \end{aligned}$$

This value of Z would therefore lead us to reject the null hypothesis $\pi = 0.40$ at the 95 percent level of significance.

Figure 1: 95% Confidence Rejection Region: One-Tailed Test ($Z > 0$)



"Tailedness"

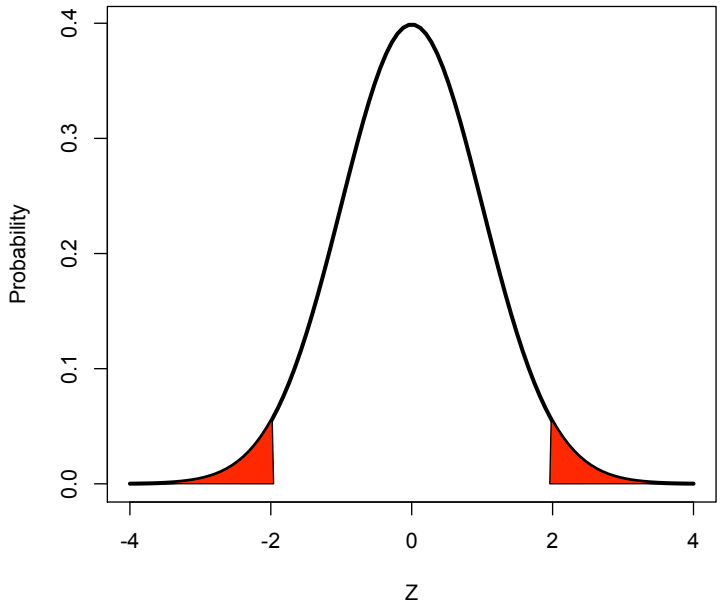
In our Clinton example, we have an alternative hypothesis H_a that is "directional" – specifically, that Clinton will receive *more* than 50 percent of the vote. An different approach might be to state that you don't think Clinton will get 50 percent of the vote, but you don't know (or necessarily care) whether her share will be more or less than that. Formally, this amounts to stating:

$$H_a : \pi \neq 0.5.$$

We are often interested in hypotheses of this sort in cases where we might anticipate a difference, but our expectations about the direction of that difference are not clear. This happens quite commonly for questions where different theories suggest different expectations about relationships in our data. For example, neorealist IR theory suggests that asymmetric trade between nations will lead to a higher incidence of conflict, while neoliberal theories predict that such asymmetric trade (and, in fact, *any* trade) has a pacifying effect. In both cases, the relevant "null" hypothesis is that there is no relationship between trade and conflict, but the direction of the "alternative" hypothesis is indeterminate.

"Tailedness" – the directionality or lack thereof in H_a – is important for hypothesis testing because it is part of how one determines the rejection region for one's test statistic. Formally,

Figure 2: 95% Confidence Rejection Region: Two-Tailed Test ($Z \neq 0$)



we use one "tail" of the sampling distribution if our H_a is "directional," as it was (for example) in the Clinton case above, and we refer to this as a "one-tailed test." We use both tails of the distribution if our hypothesis is not directional, and refer to the resulting hypothesis test as "two-tailed." Formally, a two-tailed test rejects H_0 if

$$\text{Reject } H_0 \text{ if } |Z| \geq z_{\alpha/2}$$

A one-tailed test will always be "easier" to reject the null hypothesis with, provided that the "sign" of the test statistic is correct. That's because, when we are looking at only one tail of the distribution, the location at which $1 - \alpha$ percent of the density is greater (or less) than z is closer to zero than when we have to consider both tails. Put somewhat differently, for a two-tailed test, we use $z_{\alpha/2}$, rather than z_α, to calculate the cutoff point for our test statistic.

So (and as we noted above), for example, the 95 percent confidence rejection value for Z using a one-tailed test is 1.65; that is, 5 percent of the density of a standard normal variate Z has a value greater than or equal to 1.65. For the corresponding two-tailed test, the equivalent Z value is 1.96; formally, that means that 2.5 percent of the density of a standard normal variate has values greater than 1.96, and an additional 2.5 percent has values less than -1.96.

P-Values Versus Significance Tests

An alternative approach to tests of significance is to report the actual P-value (also referred to as the "attained significance level") of a given test statistic – that is, the smallest level of significance α for which the observed data indicate that the null hypothesis should be rejected. Every value of Z corresponds to a particular probability α; the smaller the P-value, the stronger the case that H_0 can be rejected.

As a general rule, P-values are better than significance tests, for at least two (related) reasons:

1. **P-values avoid arbitrary "cutoffs."** Rather than simply saying that one can or cannot reject the null at some (more-or-less arbitrary) level of significance, reporting a P-value lets the reader come to his or her own conclusions regarding the strength of the case against the null hypothesis.

2. **P-values provide more information.** This is related to (1), and underscores the idea that the reader can make his or her own decisions about the results of the study.

In the Pennsylvania survey described above, we could use a table or some statistical software to learn that the Z value of -1.54 corresponds to a P-value of 0.88; that means that, if we'd set α equal to 0.90, we would have rejected the null hypothesis that $\pi = 0.5$. In other words, the P-value of 0.88 means that a researcher could reject the null of $\pi = 0.5$ for any significance level up to and including 12 percent. (That's not very good). The corresponding P-value for the Z-score equal to 3.59 (testing $H_0 : \pi = 0.40$) is 0.00017.

Significance Tests and Confidence Intervals

There is also a straightforward relationship between significance testing and the confidence intervals we discussed last time. Recall that for a sample statistic that is normally distributed, we defined an $(1 - \alpha) \times 100$ - percent confidence interval as

$$\text{c.i.}_\alpha = \hat{\theta} \pm z_{\alpha/2}\sigma_{\hat{\theta}}$$

and we said above that, for a two-tailed test, we reject H_0 if

$$|Z| \equiv \left| \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} \right| \geq z_{\alpha/2}. \tag{1}$$

One way of reconceptualizing the rejection region for Z is as the complement of an "acceptance region,"[2] the region in which, if Z falls within it, we fail to reject the null hypothesis. Building on (1), this region can be thought of as

[2] I suppose I probably ought to call this a "fail-to-rejection region," but that's damned awkward...

$$-z_{\alpha/2} \leq \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} \leq z_{\alpha/2}. \tag{2}$$

Working a bit of algebra on this (in a fashion akin to what we did last time), we can reexpress this "acceptance region" as

$$\hat{\theta} - z_{\alpha/2}\sigma_{\hat{\theta}} \leq \theta \leq \hat{\theta} + z_{\alpha/2}\sigma_{\hat{\theta}}, \tag{3}$$

that is, as exactly the confidence interval we discussed before. Thus, the connection between c.i.s and hypothesis testing are made clear: one way to restate our decision rule for hypothesis testing is:

> "Do not reject H_0 at P $= \alpha$ if θ lies within a $(1 - \alpha) \times 100$-percent confidence interval around $\hat{\theta}$, and reject H_0 if it does not."

This connection also illustrates at least two other useful things. First, note that the "acceptance region" contains many values; this underscores why we do not typically "accept" a null hypothesis. In point of fact, *any* value(s) in the "acceptance region" are "acceptable" (or, at least, "not rejectable"). This in turn highlights the importance of not drawing conclusions about the value of θ from an "insignificant" estimate $\hat{\theta}$.

Second, the relationship points out the value of confidence intervals as a means of inference. Confidence intervals have a number of nice, intuitive qualities – such as getting smaller as the sample size grows – that make them excellent for conveying inferential information. In contrast, hypothesis testing of the conventional sort (as we'll see below) raises all kinds of possibilities for inaccuracies and mistakes.

Important Things To Remember

Hypothesis testing is a dangerous business; it is very, very easy to make incorrect statements. Below are some examples, statements in red are *wrong* (and should not be made), while those in green are acceptable.

1. ***P*-values are *not* "the probability that the null hypothesis is false."** This is the most common mistake people make, and the most egregious.[3] So, for example, do not say:

   > "The test statistic allows us to reject the null hypothesis at $P < 0.01$, indicating that there is a less than one in 100 chance that the null hypothesis is true."

   Instead, say:

   > "The test statistic allows us to reject the null hypothesis at $P < 0.01$, which is strong evidence that the observed result is not due to chance."

2. **One does not, in general, "accept" the null hypothesis.** More generally, it is a serious – but very common – mistake to conclude much of *anything* from a test statistic that is not in the rejection region. It's quite common to say in a regression context, for example, to make a statement like:

   > "The *P*-value for the regression coefficient on *Female* is 0.56, indicating that there is no relationship between gender and support for immigrants' rights."

   A better version of this statement is:

   > "The *P*-value for the regression coefficient on *Female* is 0.56, indicating the data do not support the hypothesized relationship between gender and support for immigrants' rights."

3. ***P*-values are *not* the long-run frequency of a "statistically significant" test statistic.** Gill discusses this as the "replication fallacy," noting that it is incorrect to say (e.g.):

   > "The *P*-value of 0.01 means that 99 out of 100 hypothetical replications would reject the null hypothesis."

[3] Formally – as Gill notes – this amounts to conflating $\Pr(\text{Data}|H_0)$, which we *do* have, with $\Pr(H_0|\text{Data})$, which we do not.

4. **Statistical significance does not equate to substantive significance.** Particularly in very large samples, one can have P-values that are quite small even for estimates and relationships that are substantively trivial (e.g., astrological sign differences in voting, etc.). The most common error in this regard is stating that two different estimates (regression slopes, t-tests, etc.) on the same data are of differential (substantive) importance because one is more "statistically significant" than the other.

   Consider what happens, for example, if (using the data from Exercise Five) we compare Scorpios in Centre County to the population as a whole, in terms of whether (=1) or not (=0) they are currently listed as "active" on the voter rolls:

```
> popmean <- with(data, prop.table(table(1-Active)))[1]
> popmean
     0
0.8956

> with(data[data$Sign=="Scorpio",],
+      prop.test(table(1-Active),p=popmean,
+                correct=FALSE))

1-sample proportions test without continuity correction

data:  table(1 - Active), null probability popmean
X-squared = 2, df = 1, p-value = 0.2
alternative hypothesis: true p is not equal to 0.8956
95 percent confidence interval:
 0.8832 0.8975
sample estimates:
     p
0.8905
```

   We'd like to think that Scorpios are more or less a random sample of registered voters in Centre County. But, with more than 7,000 of them in the "sample," the 95% confidence interval for `Active` just barely includes the population value. The actual (substantive) difference between Scorpios and others is very small (89.56% active overall, versus 89.05% for Scorpios, or a difference of about 0.5%), but because the "sample" of Scorpios is large, the difference is (nearly) enough to cause us to reject the true population value.

5. **A statistic can never be "significant in the wrong direction."** If you have a directional hypothesis, then you have a one-tailed test; if the sign of the test statistic is the opposite of what you are expecting/hypothesizing, then it cannot be "statistically significant," *no matter how precisely estimated it is.* That is, it is badly incorrect to say:

"Our estimate of the effect of trade liberalization on the probability of a civil war – which we expected to be negative – is in fact positive, and statistically significant at $P = 0.02$."

6. **Identical P-values are not "better" or "more reliable" if they are based on a larger sample.** In fact, quite the opposite: it is "easier" to reject the null hypothesis at a specified confidence level as the sample size grows. However...

7. **At the same time, *failing* to reject the null hypothesis in a larger sample is a bigger deal than failing to do so in a small one.** This can be hard to get one's head around, but if we think about the asymptotic case (where $N \to \mathfrak{N}$), it becomes clearer why.