The Basics of Hypothesis Testing

36-600

Week 2 Thursday – Fall 2021

Hypothesis Testing: Basic Idea

When we conduct an experiment, we (almost always) do not know the true underlying probability distribution from which the data are sampled. For instance: what is the true distribution of adult human heights? (Probably kinda like a normal…but also probably not exactly a normal. The truth is messy.)

When we perform a hypothesis test, what we are doing is, e.g.,

1. selecting a property of a particular distribution (e.g., the mean of a normal distribution);
2. specifying a value for that parameter (e.g., $\mu_o = 6$), which we'll dub the *null hypothesis*;
3. specifying a alternative to the null (e.g., $\mu > \mu_o$);
4. specifying a *test statistic*, i.e., an informative function of the data (e.g., the sample mean $\bar{x}$), whose distribution *given the null* we know;
5. determining how likely it is that we would observe the value of our test statistic *or something more extreme*, if the null is correct; and
6. deciding whether we can reject the null (or whether we fail to reject it).

Note that some hypothesis tests revolve around distributions themselves, and not just particular distribution properties like the mean. We'll return to this point later.

Also note that in a hypothesis test, **we are not proving anything!** *A hypothesis test is simply a mechanism by which to make a decision* (e.g., "I fail to reject the null: my data are consistent with the hypothesized value $\mu_o = 6$" versus "I reject the null: my data are not consistent with that value"). It is always possible that the decision you make could be the wrong one!

Hypothesis Test: Extended Illustrative Example

I am about to conduct an experiment in which I will collect n measurements, where n is my *sample size*. *Before* starting this experiment, I set down my hypotheses:

- H_o (the null hypothesis): the mean of the distribution from which the data are to be sampled is $\mu_o = 10$.
- H_a (the alternative hypothesis): the mean of the distribution from which the data are to be sampled is $\mu > \mu_o$.

I will thus conduct an *upper-tail test* (because the alternative specifies "greater than"). This is to be constrasted against a *lower-tail test* (e.g., $\mu < \mu_o$) or a *two-sided test* (e.g., $\mu \neq \mu_o$).

Also before starting the experiment, I will set down my test statistic and decision rule:

- I will decide to reject the null if the probability of seeing the observed value of my test statistic $\bar{x}$ or something more extreme (here, something larger, since I am performing an upper-tail test), given that the null is correct, is less than $\alpha = 0.05$.

The computed probability of seeing the *observed* value of my test statistic, $\bar{x}$, or something more extreme, given that the null is correct, is dubbed the *p-value*.

¡MUY IMPORTANTE! A p-value is not the probability that the null hypothesis is correct!

(A hypothesis is an idea, not a random variable. There is no probability associated with it!) A p-value is simply the probability of observing a particular test statistic value, or a more extreme value, given the null. Full stop.

Hypothesis Test: Extended Illustrative Example

I now do my experiment. My sample size is n = 50.

```
##  [1]  6.331929  7.457654  7.385734 12.776347  7.768998  3.800809 12.986752  6.902247 17.230201  7.661783 10.010315
## [13] 15.526207 10.563105 18.822190 12.649759 15.520070 15.646156 14.798670 12.748517 12.216513 13.874132 13.049744
## [25] 11.988662 16.317378  7.056131  5.376409  7.451565  7.340971 12.762597  6.768779 21.296951 13.978057 10.495503
## [37] 12.610566 21.407899  6.371005  4.883713  7.195884 10.553674 23.949660 13.511982 14.573277 11.426570 11.657881
## [49] 10.502460 12.506495
```

```
## The sample mean is  11.53551
```

The value that I observe is 11.54. While this is indeed greater than 10, it is a value that could have arisen by chance, with 10 being what we'd actually see as an average if we kept repeating the experiment. What is the probability that I would actually observe a value of 11.54, or something greater (i.e., more extreme, in the context of an upper-tail test), if the true value is indeed 10? (This, again, is the p-value.) Is this value less than my pre-determined value $\alpha = 0.05$?

(About p-values and α: if the null hypothesis is correct, the distribution of the p value is *uniform*, meaning I am just as likely to observe p = 0.87 as p = 0.33 as p = 0.01, etc. Thus, when I say I will reject the null when $p < \alpha$, I am taking a risk: if the null is indeed true, there is a 5% chance I would (wrongly) reject it. Rejecting the null when it is actually true is a so-called *Type I error*.)

Hypothesis Test: Extended Illustrative Example

To determine the p-value, we have to know or assume the probability distribution from which $\bar{x}$ is sampled.

In this particular example, we can utilize the *central limit theorem*, which says that if our sample size is sufficiently large (rule of thumb: $n \geq 30$), then no matter what distribution the individual data are sampled from, the *sample mean* is distributed normally (at least approximately), with mean μ and standard deviation $\sigma/\sqrt{n}$. (The more data you observe, the better you can constrain μ!)

We don't know μ, but we are testing a particular value: $\mu = 10$.

We don't know σ, but we can simply substitute in the sample standard deviation s:

```
cat("The sample standard deviation is ",sd(x),"\n")
```

```
## The sample standard deviation is  4.440231
```

and so the standard deviation for the mean itself (or the *standard error*) is, approximately,

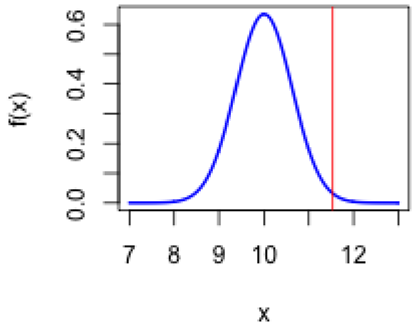
```
cat("The estimated standard error is ",sd(x)/sqrt(length(x)),"\n")
```

```
## The estimated standard error is  0.6279435
```

Hypothesis Test: Extended Illustrative Example

Let's draw a normal distribution with mean 10 and standard deviation 0.628:

```
fx = dnorm(seq(7,13,by=0.01),mean=10,sd=sd(x)/sqrt(length(x)))
plot(seq(7,13,by=0.01),fx,typ="l",col="blue",lwd=2,xlab="x",ylab="f(x)")
abline(v=mean(x),col="red")
```



The p-value is the area under the blue curve to the right of the red line (which denotes the observed sample mean). The function `pnorm()` tells us the area to the **left** of the line...so the p-value will be

```
1 - pnorm(mean(x),mean=10,sd=sd(x)/sqrt(length(x)))
```

```
## [1] 0.007236742
```

This is less than 0.05, so we reject the null hypothesis that $\mu = 10$. The end.

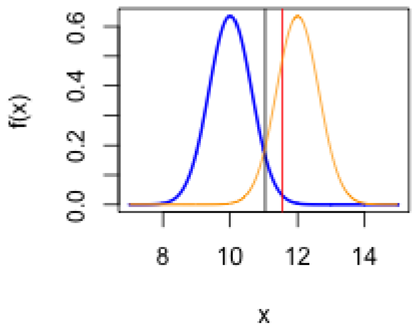
Hypothesis Test: Extended Illustrative Example

Well, almost the end.

One thing we might like to know is the *power* of our hypothesis test. The power is the probability that we would reject the null given a particular alternative hypothesis. For instance, let's say that before I started to collect data, I specified that I was interested in testing a particular alternative, $\mu_a = 12$.

That alternative helps define a second probability distribution, which we illustrate here:

```
fx = dnorm(seq(7,15,by=0.01),mean=10,sd=sd(x)/sqrt(length(x)))
plot(seq(7,15,by=0.01),fx,typ="l",col="blue",lwd=2,xlab="x",ylab="f(x)")
abline(v=mean(x),col="red")
abline(v=qnorm(0.95,mean=10,sd=sd(x)/sqrt(length(x))),col="black")
fx.alt = dnorm(seq(7,15,by=0.01),mean=12,sd=sd(x)/sqrt(length(x)))
lines(seq(7,15,by=0.01),fx.alt,col="orange")
```



Hypothesis Test: Extended Illustrative Example

Since this is an upper-tail test, the power is the area under the *orange* curve to the right of the *black* line. (This is the probability of rejecting the null; the black line bounds the rejection region, and the area under the blue curve to the right of the black line is α.)

```
power = 1 - pnorm(qnorm(0.95,mean=10,sd=sd(x)/sqrt(length(x))),mean=12,
                  sd=sd(x)/sqrt(length(x)))
cat("The test power for mu=12 is ",power,"\n")
```

```
## The test power for mu=12 is  0.9382376
```

Note that the power calculation did not bring in the observed test statistic $\bar{x}$ (although it does bring in the observed sample standard deviation).

How do you increase power?

- Increase the sample size, n. This makes the blue and orange curves "thinner," and moves the black line to the left (to keep the area under the blue curve to the right of the black line, which is α, the same). This increases the probability of rejecting the null.

- Increase the difference between μ_o and μ_a. This pulls the blue and orange curves farther apart, which increases the probability of rejecting the null.

Hypothesis Testing: Beware Multiple Comparisons!

Suppose I perform 100 independent hypothesis tests with Type I error α, and that for all 100 tests, the underlying truth is that H_o is correct. Let's assume $\alpha = 0.05$. Because the p-values output by these tests are distributed uniformly, we expect that we will reject the null approximately five times!

If you test enough (true!) null hypotheses, eventually you *will* find an apparently significant result.

Testing multiple hypotheses, without correcting for "multiple comparisons," in order to find that significant result, is *p-hacking*. Don't be a p-hacker.

Corrections for multiple comparisons (which we will not cover in the class) include the following.

- The *Bonferroni correction*: if you run k tests, change your allowable Type I error from α to α/k. This tends to be too conservative.
- The *False Discovery Rate*: this implements a correction that changes error rate to something between α and α/k. FDR tends to work well in practice.

Hypothesis Tests You Are Likely to See

1) The Large-Sample z Test

What is It? A test of a distribution mean μ. The distribution itself is unspecified. (If the distribution is normal, use a t test. See below.)

What is Assumed? The test utilizes the Central Limit Theorem, so the sample size is $n \geq 30$. And all data are sampled from the same distribution, i.e., all data are *independent and identically distributed*, or *iid*.

What is the Test Statistic? The standardized quantity

$$Z = \frac{\bar{x} - \mu_o}{s/\sqrt{n}} \, .$$

What is the p-Value?

- upper-tail: `1-pnorm(Z)`
- lower-tail: `pnorm(Z)`
- two-sided: `2*(1-pnorm(abs(Z)))`

The standardized quantity Z is standard-normal-distributed, so the mean is 0 and the standard deviation is 1…these are default values for `pnorm()` and thus need not be specified.

Note: so-called *population proportions* tests, based on binomial distributions, are really just large-sample z tests with different statistics used in place of $\bar{x}$ and $s/\sqrt{n}$.

Hypothesis Tests You Are Likely to See

2) The One-Sample t Test

What is It? A test of a *normal* distribution mean μ

What is Assumed? All data are iid samples from a normal distribution of unknown mean and standard deviation.

What is the Test Statistic? The standardized quantity

$$T = \frac{\bar{x} - \mu_o}{s/\sqrt{n}} \,.$$

What is the p-Value?

- upper-tail: `1-pt(T,df=n-1)`
- lower-tail: `pt(T,df=n-1)`
- two-sided: `2*(1-pt(abs(T),df=n-1))`

Note: while the formulae above are straightforward, you could just use the `t.test()` function in `R`. We'll return to this point in today's lab.

(As an aside: you'd think there would be a `z.test()` function in the base `R` function set...but no. However, several add-on packages have coded the `z.test()`. You can always install and use one of those.)

Hypothesis Tests You Are Likely to See

3) The Two-Sample t Test

What is It? A test of whether the data in two separate samples were drawn from normal distributions with the same mean μ.

What is Assumed? The data in each sample are normal and are independent and identically distributed. Note that the data may be paired...there may be a one-to-one correspondance between each datum in one sample and each in the other. (This would necessitate a paired t test.)

What is the Test Statistic? Now things start getting a bit complicated. For a classic unpaired two-sample t test, see the formulae on, e.g., this web page.

What is the p-Value? Here, you are definitely falling back on using the `t.test()` function in `R`. (Life's too short.) See today's lab for more details.

Hypothesis Tests You Are Likely to See

4) One-Way Analysis of Variance (ANOVA)

What is It? A test of whether the data in k separate samples were drawn from normal distributions with the same mean μ. (It is essentially a k-sample t test.)

What is Assumed? The data in each sample are normal and are independent and identically distributed.

What is the Test Statistic? See this web page.

What is the p-Value? It is output by the `anova()` function in `R`. See today's lab for more details.

(Wait...is this really useful? It can tell me that one of the means is different...but which one?)

We are glad you asked. You are correct. A standard analysis "protocol" is first to apply ANOVA to see if there is at least one mean that is statistically significantly different from the others (i.e., if $p < \alpha$), then to apply a *post-hoc comparisons test* (like the so-called Tukey test) to see which pairs of means are different. As life is short (as already stated), we will not cover post-hoc tests here.

Hypothesis Tests You Are Likely to See

There are a few more often-used tests, but we'll stop for now and introduce others as they are needed.

For instance, there is the *chi-square goodness-of-fit test*, which we'll talk about when we talk about general curve fitting in a future lecture.

Also, we mentioned earlier that some hypothesis tests revolve around the testing of distributions rather than distribution properties like the mean. Examples of such tests include

- The *one- and two-sample Kolmogorov-Smirnov (KS) tests*: was a sample of data drawn from a hypothesized distribution…or, were two samples drawn from the same underlying distribution?
- The *Shapiro-Wilk test*: was a sample of data drawn from a normal distribution?

We will start looking into some of these in today's lab.