Understanding p Values

James H. Steiger
Vanderbilt University

Introduction

In this module, we introduce the notion of a p value, a concept widely used (and abused) in statistics.

We'll learn what a p value is, what it isn't, and how it is employed in standard hypothesis testing situations.

We'll discover how to compute p values from several distributions.

Quantifying the Unusual

In standard hypothesis testing, we employ a strategy that is based on falsifying the opposite of what we are trying to show. For example, suppose we believe that giving a drug to a group of people will make them unusual. We employ the following logic.

We begin by stating a hypothesis that is the opposite of what we believe. That is, we state as our "statistical null hypothesis" that the group is *not* unusual.

What we then seek to do is show that, in fact, the group *is unusual*, thereby falsifying the opposite of what we wish to show, and demonstrating the truth of what we wish to show.

So we need a way of demonstrating that an observation (our data) is unusual enough to be unlikely to have occurred by chance.

Quantifying the Unusual

Suppose a population of scores has a probability distribution.

How can we decide if an observation is *unusual*, or *extreme*?

For example, suppose IQ scores are normally distributed with a mean of 100 and a standard deviation of 15. Is an IQ of 145 unusual, or extreme? How would you decide?

Quantifying the Unusual

One way of quantifying how unusual a result is involves a probability calculation.

A result X is unusual if the probability of a result more extreme than X is small.

For example, how unusual is an IQ of 145?

We can see from the tiny red sliver on the far right of the plot on the next slide, that the probability is really small.

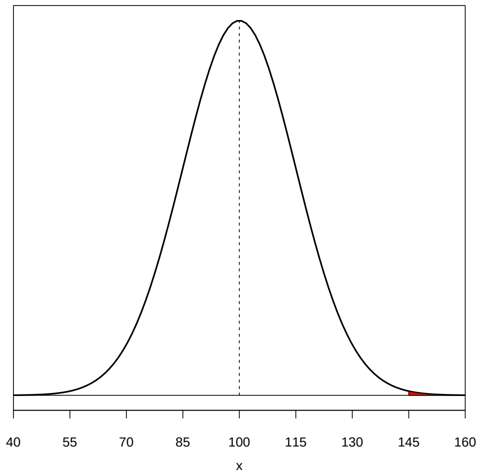
We can compute this probability as

```
> 1 - pnorm(145, 100, 15)

[1] 0.00135
```

In fact, statisticians call this quantity the "one-sided p value."

Quantifying the Unusual



How Extreme? Two-Sided vs. One-Sided

There is a catch. Suppose you consider an IQ of 100 to be normal.

Wouldn't it be correct to say that an IQ of 55 is just as unusual as an IQ of 145? The probability of being as extreme as or more extreme than 145 in the positive direction is exactly the same as the probability of being as extreme as or more extreme than 55 in the negative direction.

Here we show the calculations:

```
> ## probability X >= 145
> 1 - pnorm(145, 100, 15)

[1] 0.00135

> ## probability X <= 55
> pnorm(55, 100, 15)

[1] 0.00135
```

How Extreme? Two-Sided vs. One-Sided

This creates a problem. If you want to characterize how unusual a result is, you may want to add up these two probabilities. If you do, you have computed a "two-sided p value." In this case, then, we would say that an IQ of 145 has a "two-sided p value of .00270."

Calculating p Values

The preceding considerations have led to some standard conventions for calculating p values (or "significance levels") in the continuous case.

In statistical hypothesis testing, it is common to test a hypothesis about a parameter with a "test statistic" that is some function of an estimate of that parameter.

For example, we test a hypothesis that a population mean μ is equal to 100 by examining a test statistic t that is a function of an estimate of how much μ deviates from 100.

If the statistical null hypothesis is true, then the test statistic has a known distribution, the "null sampling distribution."

If a value occurs that is unlikely to occur in that distribution, we consider the result as evidence against the null.

In *one-sided* (or one-tailed) testing situations, evidence to reject the null can come at only one end of the number line, while in *two-sided* (or two-tailed) situations, the evidence can come at either end.

Calculating p Values

Connected with this idea is a value called the *significance level* or p-value.

With these ideas in mind, we can produce a rule for computing p-values that works well for continuous sampling distributions, but is more controversial for discrete distributions.

The p-value is the probability of getting a result as extreme or more extreme as the observed value:

1. If the test is one-sided, the calculation is performed in the direction that would result in rejection, and the p value is the probability of obtaining a result more extreme than the observed statistic in that rejection direction.
2. If the test is two-sided, the calculation is performed toward the nearest rejection side, then the resulting probability is doubled.

In general, the lower the p-value, the "more significant" the result.

In particular, if the p-value is less than α, we say that the result is "significant at the α level."

Calculating p Values

Some people have been led to believe that this means that the lower the p-value, the more powerful the effect shown in the result, but this is most definitely not the case.

Calculating *p* Values

To see how this works, suppose you are testing the hypothesis that $\mu_1 = \mu_2$ with a *t*-statistic with 24 degrees of freedom. The value of the statistic is 2.54. What is the *p*-value? The probability of getting a result more extreme on the positive side is

```
> 1 - pt(2.54, 24)

[1] 0.008987
```

However, since this is a hypothesis about equality, it could be rejected with either a very low or a very high value of the test statistic. It is a two-sided test. So we must double the above probability to get a correct 2-sided *p*-value

```
> 2 * (1 - pt(2.54, 24))

[1] 0.01797
```

Since the value is lower than 0.05, we reject the null hypothesis at the 0.05 level, but not at the 0.01 level, since the value is larger than 0.01.

Calculating p Values

We can automate this process with R functions that we can then use in subsequent statistical functions.

We'll load the functions in the file `t1.r` on the course website.

```
> source("http://www.statpower.net/R/t1.R")
```

Calculating *p* Values

Examining the first function, we see that it simply implements the procedure we discussed above.

You enter whether the null hypothesis is of the form $X = a$, $X \leq a$, or $X \geq a$, and give the value of the statistic, and the routine does the rest.

```
> z.test.p.value

function (z, null.hypothesis = "equals")
{
    if (null.hypothesis == "equals")
        p.value <- 2 * (1 - pnorm(abs(z)))
    if (null.hypothesis == "less")
        p.value <- 1 - pnorm(z)
    if (null.hypothesis == "greater")
        p.value <- pnorm(z)
    return(p.value)
}
```

Calculating *p* Values

Suppose we observe a Z-statistic of 2.08, with a 2-sided hypothesis of the form $\mu = 50$.

```
> z.test.p.value(2.08)

[1] 0.03753
```

Is this result significant at the 0.05 level?

Review Questions

Question 1

Suppose you are using a normally distributed statistic Z statistic to test the hypothesis that $\mu = 100$, i.e., that a population has a mean of exactly 100, against the alternative that $\mu \neq 100$. Is this a 2-sided or a 1-sided test?

Review Questions

Question 1

Answer. It is a 2-sided test, because we need to consider two ways in which the hypothesis might be false. The population mean might be larger than 100, or smaller than 100, and our test statistic may reflect either possibility. Consequently, one must compute a “2-sided *p*-value.”

Review Questions

Question 2

In the situation described in Question 1, suppose you obtain a Z statistic value of -2.45. What is the p-value?

Review Questions

Question 2

Answer. We need to calculate a the probability of obtaining a value more extreme than -2.45 in a standard normal distribution. But we also need to double this probability to obtain the 2-sided p value. Thus, we compute

```
> print(p.obtained <- 2 * pnorm(-2.45))

[1] 0.01429
```

Review Questions

Question 3

In the preceding question, we obtained a two-sided p-value of 0.0143. Is this result "statistically significant" at the 0.05 level? Is it significant at the 0.01 level?

Review Questions

Question 3

Answer. The obtained p value is less than 0.05, so the result is significant at the 0.05 level. However, it is not less than 0.01, so the result is not significant at the 0.01 level.

Review Questions

Question 4

If you are running a two-sided Z-test, what is the highest possible p-value you can obtain, and what observed value of Z will yield that value?

Review Questions

Question 4

Answer. Remember that, with a two-sided test, you compute the probability of obtaining a more extreme result than the one observed, *and then you double that probability*. As a value of Z gets closer and closer to zero, the area beyond that value gets closer and closer to 0.50. When Z is exactly 0, the area beyond it on either side is 0.50, and so the 2-sided p value is 1.00.

Review Questions

Question 4

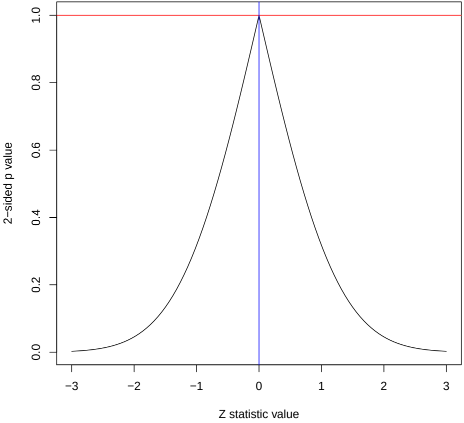
Incidentally, this is just one more example of how using R functions an plotting might have helped give you an insight about the correct answer. Specifically, we have a `z.test.p.value` function online, which we already loaded in this lecture.

Suppose we now simply plot the result over a wide range of Z values. For clarity, I drew a horizontal line at a probability value of 1.0, and a vertical line at 0.0 The plot clearly shows that the p value hits its maximum of 1.0 at a value of 0.

Review Questions

Question 4

```
> curve(z.test.p.value(x), -3, 3, xlab = "Z statistic value", ylab = "2-sided p value")
> abline(h = 1, col = "red")
> abline(v = 0, col = "blue")
```



Review Questions

Question 5

If you are running a one-sided Z-test of the hypothesis that $\mu \leq 20$, and you obtain a Z-statistic of 1.99? What is the p value?

Review Questions

Question 5

Answer. In this case the statistic is “in the rejection direction”, because the null hypothesis is on the lower side of 20. So the p value is the probability of getting a result larger than 1.99, which is

```
> 1 - pnorm(1.99)

[1] 0.0233
```

Note that you could have used our utility function as well. We need to tell the function what kind of null hypothesis is being tested.

```
> z.test.p.value(1.99, null.hypothesis = "less")

[1] 0.0233
```

Review Questions

Question 6

Suppose that, in the preceding question, we had obtained a Z value of -1.99. What would the p value have been?

Review Questions

Question 6

Answer. The *p*-value, in the one-sided case, is the area probability of a more extreme value in the rejection direction. In this case, the rejection direction is positive, so we compute the probability of obtaining a result of -1.99.

```
> 1 - pnorm(-1.99)

[1] 0.9767
```

Of course, we could have used the utility function

```
> z.test.p.value(-1.99, null.hypothesis = "less")

[1] 0.9767
```