

Introduction to Hypothesis Testing

I. Terms, Concepts.

A. In general, we do not know the true value of population parameters - they must be estimated. However, we do have hypotheses about what the true values are.

B. The major purpose of hypothesis testing is to choose between two competing hypotheses about the value of a population parameter. For example, one hypothesis might claim that the wages of men and women are equal, while the alternative might claim that men make more than women.

C. The hypothesis actually to be tested is usually given the symbol H_0 , and is commonly referred to as the null hypothesis. As is explained more below, the null hypothesis is assumed to be true unless there is strong evidence to the contrary – similar to how a person is assumed to be innocent until proven guilty.

D. The other hypothesis, which is assumed to be true when the null hypothesis is false, is referred to as the alternative hypothesis, and is often symbolized by H_A or H_1 . *Both the null and alternative hypothesis should be stated before any statistical test of significance is conducted.* In other words, you technically are not supposed to do the data analysis first and then decide on the hypotheses afterwards.

E. In general, it is most convenient to always have the null hypothesis contain an equals sign, e.g.

$$\begin{aligned} H_0: \quad \mu &= 100 \\ H_A: \quad \mu &> 100 \end{aligned}$$

F. The true value of the population parameter should be included in the set specified by H_0 or in the set specified by H_A . Hence, in the above example, we are presumably sure μ is at least 100.

G. A statistical test in which the alternative hypothesis specifies that the population parameter lies entirely above or below the value specified in H_0 is a one-sided (or one-tailed) test, e.g.

$$\begin{aligned} H_0: \quad \mu &= 100 \\ H_A: \quad \mu &> 100 \end{aligned}$$

H. An alternative hypothesis that specified that the parameter can lie on either side of the value specified by H_0 is called a two-sided (or two-tailed) test, e.g.

$$\begin{aligned} H_0: \quad \mu &= 100 \\ H_A: \quad \mu &\neq 100 \end{aligned}$$

I. Whether you use a 1-tailed or 2-tailed test depends on the nature of the problem. Usually we use a 2-tailed test. A 1-tailed test typically requires a little more theory.

For example, suppose the null hypothesis is that the wages of men and women are equal. A two-tailed alternative would simply state that the wages are not equal – implying that men could make more than women, or they could make less. A one-tailed alternative would be that men make more than women. The latter is a stronger statement and requires more theory, in that not only are you claiming that there is a difference, you are stating what direction the difference is in.

J. In practice, a 1-tailed test such as

$$\begin{aligned} H_0: \mu &= 100 \\ H_A: \mu &> 100 \end{aligned}$$

is tested the same way as

$$\begin{aligned} H_0: \mu &\leq 100 \\ H_A: \mu &> 100 \end{aligned}$$

For example, if we conclude that $\mu > 100$, we must also conclude that $\mu > 90$, $\mu > 80$, etc.

II. The decision problem.

A. How do we choose between H_0 and H_A ? The standard procedure is to assume H_0 is true - just as we presume innocent until proven guilty. Using probability theory, we try to determine whether there is sufficient evidence to declare H_0 false.

B. We reject H_0 only when the chance is small that H_0 is true. Since our decisions are based on probability rather than certainty, we can make errors.

C. Type I error - We reject the null hypothesis when the null is true. The probability of Type I error = α . Put another way,

$$\alpha = \text{Probability of Type I error} = P(\text{rejecting } H_0 \mid H_0 \text{ is true})$$

Typical values chosen for α are .05 or .01. So, for example, if $\alpha = .05$, there is a 5% chance that, when the null hypothesis is true, we will erroneously reject it.

D. Type II error - we accept the null hypothesis when it is not true. Probability of Type II error = β . Put another way,

$$\beta = \text{Probability of Type II error} = P(\text{accepting } H_0 \mid H_0 \text{ is false})$$

E. EXAMPLES of type I and type II error:

$$\begin{aligned} H_0: \mu &= 100 \\ H_A: \mu &\neq 100 \end{aligned}$$

Suppose μ really does equal 100. But, suppose the researcher accepts H_A instead. A type I error has occurred.

Or, suppose $\mu = 105$ - but the researcher accepts H_0 . A type II error has occurred.

The following tables from Harnett help to illustrate the different types of error.

		The true situation is	
		H_0 is true	H_a is true
Action	Reject H_0 (Accept H_a)	Type I error	Correct decision
	Reject H_a (Accept H_0)	Correct decision	Type II error

		The true situation is	
		Not guilty (H_0)	Guilty (H_a)
Action	Jury finds guilty (Accept H_a)	Type I error	Correct decision
	Jury finds not guilty (Accept H_0)	Correct decision	Type II error

F. α and β are not independent of each other - as one increases, the other decreases. However, increases in N cause both to decrease, since sampling error is reduced.

G. In this class, we will primarily focus on Type I error. But, you should be aware that Type II error is also important. A small sample size, for example, might lead to frequent Type II errors, i.e. it could be that your (alternative) hypotheses are right, but because your sample is so small, you fail to reject the null even though you should.

III. Hypothesis testing procedures. The following 5 steps are followed when testing hypotheses.

1. Specify H_0 and H_A - the null and alternative hypotheses. Examples:

(a) $H_0: E(X) = 10$ $H_A: E(X) \neq 10$	(b) $H_0: E(X) = 10$ $H_A: E(X) < 10$	(c) $H_0: E(X) = 10$ $H_A: E(X) > 10$
---	--	--

Note that, in example (a), the alternative values for $E(X)$ can be either above or below the value specified in H_0 . Hence, a two-tailed test is called for - that is, values for H_A lie in both the upper and lower halves of the normal distribution. In example (b), the alternative values are below those specified in H_0 , while in example (c) the alternative values are above those specified in H_0 . Hence, for (b) and (c), a one-tailed test is called for.

When working with binomially distributed variables, it is often common to use the proportion of successes, p , in the hypotheses. So, for example, if X has a binomial distribution and $N = 20$, the above hypotheses are equivalent to:

(a) $H_0: p = .5$ $H_A: p \neq .5$	(b) $H_0: p = .5$ $H_A: p < .5$	(c) $H_0: p = .5$ $H_A: p > .5$
---------------------------------------	------------------------------------	------------------------------------

2. Determine the appropriate test statistic. A test statistic is a random variable used to determine how close a specific sample result falls to one of the hypotheses being tested. That is, the test statistic tells us, if H_0 is true, how likely it is that we would obtain the given sample result. Often, a Z score is used as the test statistic. For example, when using the normal approximation to the binomial distribution, an appropriate test statistic is

$$z = \frac{\# \text{ of successes} \pm .5 - Np_0}{\sqrt{Np_0 q_0}}$$

where p_0 and q_0 are the probabilities of success and failure as implied or stated in the null hypothesis. When the Null hypothesis is true, Z has a $N(0,1)$ distribution. Note that, since X is not actually continuous, it is sometimes argued that a *correction for continuity* should be applied. To do this, add .5 to x when $x < Np_0$, and subtract .5 from x when $x > Np_0$. Note that the correction for continuity reduces the magnitude of z . That is, failing to correct for continuity will result in a z-score that is too high. In practice, especially when N is large, the correction for continuity tends to get ignored, but for small N or borderline cases the correction can be important.

Warning (added September 2004): As was noted earlier, the correction for continuity can sometimes make things worse rather than better. Especially if it is a close decision, it is best to use a computer program that can make a more exact calculation, such as Stata can with its `btest` and `btesti` routines. We will discuss this more later.

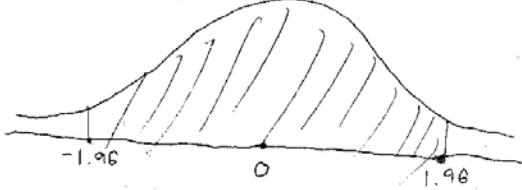
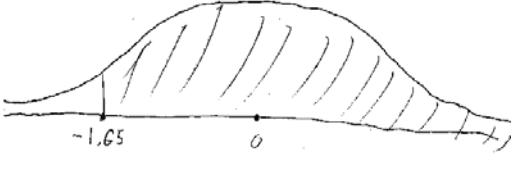
Intuitively, what we are doing is comparing what we actually observed with what the null hypotheses predicted would happen; that is, # of successes is the observed empirical result, i.e. what actually happened, while Np_0 is the result that was predicted by the null hypothesis. Now, we know that, because of sampling variability, these numbers will probably not be exactly equal; e.g. the null hypotheses might have predicted 15 successes and we actually got 17. But, if the difference between what was observed and what was predicted gets to be too great, we will conclude that the values specified in the null hypotheses are probably not correct and hence the null should be rejected.

If, instead, we work with the proportion p , the test statistic is

$$z = \frac{\hat{p} \pm .5/N - p_0}{\sqrt{\frac{p_0 q_0}{N}}} = \frac{\hat{p} \pm .5/N - p_0}{\sqrt{\frac{p_0 q_0}{N}}}$$

where \hat{p} = the observed value of p in the sample. Note that the only difference between this and the prior equation is that both numerator and denominator are divided by N . To correct for continuity, add $.5/N$ to \hat{p} when $\hat{p} < p_0$, and subtract $.5/N$ from \hat{p} when $\hat{p} > p_0$.

3. Determine the critical region (this is sometimes referred to as “designing a decision rule”). The following table summarizes the most crucial points.

Acceptance region: Choose “critical values” for a such that	When used	Example	Decision rule when α = .05
$P(-a \leq Z \leq a) = 1 - \alpha$ Or, equivalently, $F(-a) = \alpha/2$ $F(a) = 1 - \alpha/2$	for a two-tailed alternative hypothesis	$H_0: p = .5$ $H_A: p \neq .5$	Reject the null hypothesis if the computed test statistic is less than -1.96 or more than 1.96
			
$P(Z \leq a) = \alpha$, i.e., $F(a) = \alpha$	for a one-tailed alternative that involves a $<$ sign. Note that a is a negative number.	$H_0: p = .5$ $H_A: p < .5$	Reject the null hypothesis if the computed test statistic is less than -1.65
			

$P(Z \geq a) = \alpha$ i.e. $F(a) = 1 - \alpha$	for a one-tailed alternative that involves a $>$ sign. Note that a is a positive number.	$H_0: p = .5$ $H_A: p > .5$	Reject the null hypothesis if the computed test statistic is more than 1.65.

Values of the test statistic that do not fall within the specified range are said to be in the critical region - H_0 will be rejected for such values. Values of the test statistic that fall within the specified range are in the acceptance region (although a more precise, albeit much more awkward name would be the do not reject region).

Typically, α (often referred to as the “level of significance”) is set equal to .05 or .01. If, for example, $\alpha = .05$, we would erroneously reject H_0 when H_0 is true 5% of the time. Rejecting H_0 when H_0 is true is referred to as a Type I error, and α = probability of a Type I error. Accepting H_0 when H_0 is false is referred to as a Type II error, and β = probability of a Type II error.

Put another way - if $\alpha = .05$ and H_0 is true, there is only a 5% chance that we will falsely reject the null hypothesis.

4. Compute the value of the test statistic. A value of z calculated on the basis of a sample result is called a computed z -value, and is denoted by the symbols z_c or simply z .

5. Make decision. If the calculated value of the test statistic falls in the critical region, then H_0 is rejected. When the calculated value lies in the acceptance region, H_0 is not rejected.

ADDITIONAL COMMENTS:

1. For now, we will concentrate on variables that have a binomial distribution (see examples below). After the first exam, we will show how to compute test statistics for other sorts of variables.

2. Researchers often report the values of their test statistics (or provide enough information so that others can compute the test statistics). Hence, if someone else wants to use a different decision rule (that is, use a different value for α) they can do so.

SAMPLE PROBLEMS:

1. One researcher believes a coin is “fair,” the other believes the coin is biased toward heads. The coin is tossed 20 times, yielding 15 heads. Indicate whether or not the first researcher’s position is supported by the results. Use $\alpha = .05$.

Solution. If the coin is fair, $p = .5$, and $X \sim N(10, 5)$.

Step 1. The null and alternative hypotheses are

$$\begin{aligned} H_0: \quad E(X) &= 10 \text{ heads} & (\text{or, } p = .5) \\ H_A: \quad E(X) &> 10 \text{ heads} & (\text{or, } p > .5) \end{aligned}$$

(Note: a one-tailed test is appropriate, since the second researcher believes the coin is biased toward heads.)

Step 2. $Z = (\text{Number of heads} \pm .5 - 10) / \sqrt{5}$ is an appropriate test statistic. Recall that $Z \sim N(0, 1)$. To apply the correction for continuity, subtract .5 from X if the observed number of heads is greater than 10, add .5 to X if the observed number of heads is less than 10.

Alternatively, $Z = (\hat{p} \pm .5/20 - .5) / .1118$ (since the square root of $.5 * .5 / 20 = .1118$).

Step 3. Since $P(Z \leq 1.65) = .95 = 1 - \alpha$, reject H_0 if $Z > 1.65$. Equivalently, reject H_0 if $X > 14.18$ (since $X = (Z * \sqrt{5}) + .5 + 10$). Or, equivalently, reject H_0 if $p > .709$ (since $X = (Z * .1118) + .5/20 + .5$).

Step 4. $z = (15 -.5 - 10) / \sqrt{5} = 2.01$; or, $z = (.75 - .5/20 -.50) / .1118 = 2.01$

(NOTE: If we did not apply the correction for continuity, we would get $z = (15 - 10) / \sqrt{5} = 2.24$. This does not change our conclusion in Step 5. In practice, corrections for continuity tend not to be made, especially when N is large, but it is still a good idea to do it.)

Step 5. Since the computed z value is greater than 1.65, we reject H_0 .

Question: Suppose we used $\alpha = .01$. Would we still reject H_0 ?

2. Design a decision rule to test the hypothesis that a coin is fair if a sample of 64 tosses of the coin is taken and if a level of significance of (a) .05 and (b) .01 is used.

Step 1. The null and alternative hypotheses are:

$$\begin{aligned} H_0: \quad E(X) &= 32 & (\text{or, } p = .5) \\ H_A: \quad E(X) &\neq 32 & (\text{or, } p \neq .5) \end{aligned}$$

Note that a two-tailed test is appropriate, since a coin is not fair if it is biased toward either heads or tails.

Step 2. If H_0 is true, then $X \sim N(32, 4^2)$. Hence, an appropriate test statistic is

$$Z = (\text{number of heads} \pm .5 - 32) / 4$$

Step 3.

For $\alpha = .05$, do not reject H_0 if $-1.96 \leq Z \leq 1.96$. That is, do not reject H_0 if $23.66 \leq X \leq 40.34$.

For $\alpha = .01$, do not reject H_0 if $-2.58 \leq Z \leq 2.58$. That is, do not reject H_0 if $21.18 \leq X \leq 42.82$.

For example, if you tossed a coin 64 times and got 22 heads, you would reject H_0 if you used $\alpha = .05$, but you would not reject H_0 if $\alpha = .01$.

3. The mayor claims that blacks account for 25% of all city employees. A civil rights group disputes this claim, and argues that the city discriminates against blacks. A random sample of 120 city employees has 18 blacks. Test the mayor's claim at the .05 level of significance.

SOLUTION. If the mayor is correct, $X \sim N(30, 22.5)$

- a. $H_0: p = .25$ (or $E(X) = 30$)
- $H_A: p < .25$ (or $E(X) < 30$)

A one tailed test is called for - the civil rights group thinks the city discriminates, hence it believes there will be fewer blacks than the mayor claims.

- b. The appropriate test statistic is

$$Z = (\# \text{ blacks in sample} \pm .5 - Np_0) / \sqrt{(Np_0q_0)} =$$

$$(\# \text{ blacks in sample} \pm .5 - 30) / \sqrt{22.5}$$

- c. As we have previously proven, $P(Z \geq -1.65) = 1 - .05$. Hence, we will accept H_0 if $z_c \geq -1.65$, otherwise we will reject H_0 . Or, equivalently, accept H_0 if $X \geq 21.7$, reject H_0 otherwise.
- d. $z_c = (18 - 30) / \sqrt{22.5} = -2.53$. Or, to be more precise and apply the correction for continuity, $z_c = (18 + .5 - 30) / \sqrt{22.5} = -2.42$.
- e. Reject H_0 , because z_c does not lie in the acceptance region.

NOTE: If you just wanted the probability that the sample would contain 18 or fewer blacks, given that the mayor is correct, you would compute z_c as shown in step d. Then, $F(-2.53) = 1 - F(2.53) = .0057$ (or, with correction for continuity, $F(-2.42) = .0078$). Hence, there is less than a 1% chance that the mayor could be correct and the sample only contain 18 or fewer supporters.

IV. Additional comments on one-tailed tests.

Warning: In a one-tailed test, do not reject the null hypothesis unless the alternative hypothesis is better! It is not enough for the test statistic to be large in magnitude -- it must also have the correct sign!

Examples:

For each of the following, indicate whether you should reject or not reject the null hypothesis:

- a. A parts manufacturer claims that only 10% of its parts are defective (i.e. $p = .10$). Critics claim the defect rate is higher (i.e. $p > .10$). Testing of 500 randomly selected parts yields a Z statistic of -4.0.

Solution.

Do not reject! The null and alternative hypotheses are:

$$\begin{aligned} H_0: \quad p &= .10 \\ H_A: \quad p &> .10 \end{aligned}$$

The fact that the test statistic was negative means that the defect rate in the sample was *less than* .10 – that is, the product did better than the manufacturer claimed! (Remember, the formula for the test statistic involves taking the observed number of defects and subtracting the hypothesized number – so a negative test statistic means the hypothesized number was greater than the observed number.) The alternative hypothesis is clearly unreasonable in this case. In fact, you wouldn't even need to bother computing the test statistic once you saw that the sample p was less than .10.

In looking at this problem, it may be helpful to realize that a test of the above hypotheses is also a test of

$$\begin{aligned} H_0: \quad p &\leq .10 \\ H_A: \quad p &> .10 \end{aligned}$$

If $p > .10$, that also means it is greater than .09, .08, etc. Conversely, the null hypothesis is automatically a “winner” whenever the sample p is less than or equal to .10.

Incidentally, note that a one-tailed test makes a great deal of sense here. Nobody is going to be upset if the product performs better than claimed. By saying that the defect rate is .10, the manufacturer is really claiming that the defect rate is .10 or less.

b. A researcher hypothesized that women are more likely to be religious than are men. In her study, she found that

	Male	Female
Religious	70	50
Not religious	30	50

Solution.

Do not reject! The null and alternative hypotheses are

$$\begin{aligned} H_0: \quad p_F &= p_M \\ H_A: \quad p_F &> p_M \end{aligned}$$

You may think this is an unfair question, since I haven't told you how to work this kind of problem yet, but you already know everything you need to know. She hypothesized that women were more likely to be religious, when in reality her sample shows that men are (70% of the men are religious, as opposed to 50% of the women).

If instead she had hypothesized that men and women differed in how religious they were, without specifying which was more religious, a two-tailed test would be called for, and we would need to calculate the test statistic.