

Classical (Formal) Hypothesis Testing

Definitions

- 1) The **Null Hypothesis**, H_0 : A statement about the value of a population parameter (e.g. the mean μ or population proportion p) that contains the condition of **equality**. For example: $\mu = 7.3$ or $p \geq 0.36$.
The null hypothesis is tested directly by assuming that equality is true and then reaching the conclusion to either reject H_0 or to fail to reject H_0 . That is, even though we may be claiming that $\mu \geq 7.3$ the test is conducted by assuming that $\mu = 7.3$ is true.
- 2) The **Alternative Hypothesis**, H_A or H_1 : The statement that must be true (or accepted) if H_0 is rejected.
- 3) **Type I error**: The mistake of rejecting H_0 when it is in fact true. It is the actual error (as opposed to a simple calculational error) that can be made when a rare event happens by chance. The probability of committing a type 1 error is denoted by α : $P(\text{type 1 error}) = \alpha$ = the **significance level** of the test. Alpha is chosen by the tester in consideration of the seriousness of a type 1 error. Common values are $\alpha = 0.05$ or $\alpha = 0.01$.
- 3) **Type II error**: The mistake of failing to reject H_0 when it is false.
 $P(\text{Type II error}) = \beta$.
- 4) **Test Statistic**: The sample statistic based on sample data that will be used to make the decision about rejecting, or failing to reject, H_0 . It is chosen according to the relevant sampling distribution.
- 5) **Critical (Rejection) Region**: The set of values of the test statistic that would cause us to reject H_0 .
- 6) **Critical Value**: The value(s) that separate the rejection region from the values of the test statistic that would not lead to rejection of H_0 .

Controlling Alpha and Beta

For any fixed sample size n , decreasing α will increase β and vice versa. If α is fixed and the sample size is increased then β will decrease. The only way to decrease **both** α and β is to increase the sample size.

The “Tails” of the Test

The *tails* in a distribution are the extreme regions bounded by the critical values. They are the rejection region(s) of the test.

If you are testing a claim such as
$$\begin{array}{l} H_0 : m \geq 7 \\ H_A : m < 7 \end{array}$$
 then you use a **left-tailed** test. That is, the rejection region lies on the **left** of the distribution, corresponding to the fact that values of the test statistic “sufficiently” to the **left** of $m = 7$ will cause you to reject H_0 (and hence support the alternate hypothesis).

If you are testing a claim such as
$$\begin{array}{l} H_0 : m \leq 7 \\ H_A : m > 7 \end{array}$$
 then you will use a **right-tailed** test. That is, the rejection region lies on the **right** of the distribution, corresponding to the fact that values of the test statistic “sufficiently” to the **right** of $m = 7$ will cause you to reject H_0 (and hence support the alternate hypothesis).

If you are testing a claim such as
$$\begin{array}{l} H_0 : m = 7 \\ H_A : m \neq 7 \end{array}$$
 then you will use a **two-tailed** test and the rejection region lies on both sides of the distribution corresponding to the fact that you will reject H_0 if the test statistic is sufficiently far away from $m = 7$ in *either* direction.

The Significance of The Two Types of Errors.

Example: Suppose a paint manufacturer has heard of a new manufacturing process which will (supposedly) *significantly* reduce the drying time of his paint (from the old value of 4 hours) and he is considering switching to the new manufacturing process. The catch is that the retooling of his plant will also be a significant expense. He decides to base his decision on a statistical test designed to determine whether or not the mean drying time of paint really is reduced. He will spend the extra dollars on retooling and advertising if the test determines that $m < 4$. That is he will commission a sample to

test the hypotheses:
$$\begin{array}{l} H_0 : m \geq 4 \\ H_A : m < 4 \end{array}$$

Solution:

(a) **Type I error:** Reject H_0 when it is true. In this case, the boss will believe that the drying time is significantly reduced when it is NOT. He will buy new equipment and spend money on advertising and his “new and improved” product will prove no better than the old. This could have *disastrous* business consequences as he will lose his “good name” along with new and old customers.

(b) **Type II error:** Fail to reject H_0 when it is false. In this case, the boss believes the new process is no better than the old one when it actually is (i.e. he believes the new mean drying time is not less than the old one). He will not spend money to retool and advertise when he should have – he will miss out on potential benefits but he keeps his good name and his regular customers (as long as his competitor buys the new process).

Three Examples

A two-tailed test about a mean: Large Sample.

In April 1994, a columnist for the *Toronto Star* hinted at a “conspiracy” by baseball managers in which they “juiced” the baseballs (to encourage “good hitting”) to increase excitement in the game. Suppose that tests of old balls showed that when dropped 24 ft onto a concrete surface they bounced an average of 92.84 inches. In a test of a random sample of 40 new balls, the bounce heights had a mean of 92.67 inches and a standard deviation of 1.79 inches. Test the claim that the new balls have a bounce height significantly different from the old with a level of significance of 0.05.

Solution: The sample data are $\bar{Y} = 92.67$, $s = 1.79$ and $n = 40$. We will test the hypotheses

$$H_0 : \mathbf{m} = 92.84$$

$$H_A : \mathbf{m} \neq 92.84$$

Since we have a large sample ($n > 30$) involving the sample mean we will use the test statistic $z_T = \frac{\bar{Y} - \mathbf{m}}{\frac{s}{\sqrt{n}}} = \frac{92.67 - 92.84}{\frac{1.79}{\sqrt{40}}} = -0.6001$ which has the standard normal

distribution (where we have replaced σ by the sample standard deviation which is justified since n is big).

The critical values for the test are $z_C = \pm z_{\alpha/2} = \pm z_{0.025} = \pm 1.96$. Since $-z_C < z_T < z_C$ the test statistic lies in the “fail to reject” region and we conclude that “there is not enough evidence to support the claim that the balls have a bounce height different from before.” That is, the evidence does not support the belief that the balls are juiced.

A One-Tailed test about a mean: Small Sample.

Listed below are the total electric energy consumption amounts (in kWh) for the home of a statistics professor during 7 different years.

11300	11243	10789	9907	9012	9942	11053
-------	-------	-------	------	------	------	-------

The utility company claims that the mean annual consumption is 11000 kWh and offers a budget payment plan based on that amount. Test your claim that the real mean consumption is less than 11000 with a significance level of 0.10.

Solution: The sample data are: $\bar{Y} = 10464$, $s = 861.13$, $n = 7$. We will use the hypotheses

$$H_0 : \mathbf{m} \geq 11000$$

$$H_A : \mathbf{m} < 11000$$

Since the sample is small ($n < 30$) we will use the t distribution (hoping that the assumption that energy consumption measurements are “essentially” normal is justified) with $7 - 1 = 6$ degrees of freedom so that our test statistic is:

$$t_T = \frac{\bar{Y} - \mathbf{m}}{\frac{s}{\sqrt{n}}} = \frac{10464 - 11000}{\frac{861.13}{\sqrt{7}}} = -1.6468. \text{ The critical value is } -t_{6,0.10} = -1.440. \text{ Since}$$

$t_T < t_C$ the test statistic falls into the rejection region and we are justified in concluding that “the evidence supports the claim that the mean consumption is actually less than 11000 kWh.

However, it should be pointed out that P(type I error) is quite large (10%) and that a more realistic critical value (such as $-t_{0.05,6} = -1.943$) would NOT allow this conclusion!

A One-Tailed Test about a Proportion

In a study of store checkout scanners, 1234 items were checked and 20 of them were found to be overcharges (based on data from “UPC Scanner Pricing Systems: Are they accurate?” by Goodstein, *Journal of Marketing*, Vol. 58). Before scanners were used, the overcharge rate was estimated to be about 1%. Use a 0.05 level of significance to test the claim that with the use of scanners, the overcharge rate is more than 1%.

Solution: The sample data are:

$$\hat{p} = \frac{20}{1234} = 0.0162, \hat{q} = \frac{1214}{1234} = 0.9838, n = 1234.$$

The hypotheses for the test are $H_0 : p \leq 0.01$ and we will use the fact that $z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$ has

$$H_A : p > 0.01$$

approximately the standard normal distribution to get the test statistic

$$z_T = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} = 2.192. \text{ The rejection region is in the right tail of the S.N.D. and the critical}$$

value is $z_C = z_{0.05} = 1.645$. Since $z_T > z_C$ the null hypothesis is rejected and we conclude that “there is sufficient evidence to support the claim that the true proportion of overcharges is greater than 1% with the use of scanners.”