

TEST ABOUT PROPORTIONS*

Ewa Paszek

This work is produced by OpenStax-CNX and licensed under the
Creative Commons Attribution License 2.0[†]

Abstract

This course is a short series of lectures on Introductory Statistics. Topics covered are listed in the Table of Contents. The notes were prepared by Ewa Paszek and Marek Kimmel. The development of this course has been supported by NSF 0203396 grant.

1 TEST ABOUT PROPORTIONS

Tests of statistical hypotheses are a very important topic, let introduce it through an illustration.

1.1

Suppose a manufacturer of a certain printed circuit observes that about $p=0.05$ of the circuits fails. An engineer and statistician working together suggest some changes that might improve the design of the product. To test this new procedure, it was agreed that $n=100$ circuits would be produced using the proposed method and the checked. Let Y equal the number of these 200 circuits that fail. Clearly, if the number of failures, Y , is such that $Y/200$ is about to 0.05, then it seems that the new procedure has not resulted in an improvement. On the other hand, If Y is small so that $Y/200$ is about 0.01 or 0.02, we might believe that the new method is better than the old one. On the other hand, if $Y/200$ is 0.08 or 0.09, the proposed method has perhaps caused a greater proportion of failures. What is needed is to establish a formal rule that tells when to accept the new procedure as an improvement. For example, we could accept the new procedure as an improvement if $Y \leq 5$ or $Y/n \leq 0.025$. We do note, however, that the probability of the failure could still be about $p=0.05$ even with the new procedure, and yet we could observe 5 or fewer failures in $n=200$ trials.

That is, we would accept the new method as being an improvement when, in fact, it was not. This decision is a mistake which we call a **Type I error**. On the other hand, the new procedure might actually improve the product so that p is much smaller, say $p=0.02$, and yet we could observe $y=7$ failures so that $y/200=0.035$. Thus we would not accept the new method as resulting in an improvement when in fact it had. This decision would also be a mistake which we call a **Type II error**.

If it we believe these trials, using the new procedure, are independent and have about the same probability of failure on each trial, then Y is binomial $b(200, p)$. We wish to make a statistical inference about p using the unbiased $\hat{p} = Y/200$. We could also construct a confidence interval, say one that has 95% confidence, obtaining

$$\hat{p} \pm 1.96 \sqrt{\frac{\hat{p}(1-\hat{p})}{200}}.$$

*Version 1.2: Oct 8, 2007 4:14 pm -0500

[†]<http://creativecommons.org/licenses/by/2.0/>

This inference is very appropriate and many statisticians simply do this. If the limits of this confidence interval contain 0.05, they would not say the new procedure is necessarily better, at least until more data are taken. If, on the other hand, the upper limit of this confidence interval is less than 0.05, then they feel 95% confident that the true \mathbf{p} is now less than 0.05. Here, in this illustration, we are testing whether or not the probability of failure has or has not decreased from 0.05 when the new manufacturing procedure is used.

The **no change** hypothesis, $H_0 : p = 0.05$, is called **the null hypothesis**. Since $H_0 : p = 0.05$ completely specifies the distribution it is called **a simple hypothesis**; thus $H_0 : p = 0.05$ is **a simple null hypothesis**.

The research worker's hypothesis $H_1 : p < 0.05$ is called **the alternative hypothesis**. Since $H_1 : p < 0.05$ does not completely specify the distribution, it is a composite hypothesis because it is composed of many simple hypotheses.

The rule of rejecting H_0 and accepting H_1 if $Y \leq 5$, and otherwise accepting H_0 is called **a test of a statistical hypothesis**.

It is clearly seen that two types of errors can be recorded

- **Type I error:** Rejecting H_0 and accepting H_1 , when H_0 is true;
- **Type II error:** Accepting H_0 when H_1 is true, that is, when H_0 is false.

Since, in the example above, we make a Type I error if $Y \leq 5$ when in fact $\mathbf{p}=0.05$. we can calculate the probability of this error, which we denote by α and call **the significance level of the test**. Under an assumption, it is

$$\alpha = P(Y \leq 5; p = 0.05) = \sum_{y=0}^5 \binom{200}{y} (0.05)^y (0.95)^{200-y}.$$

Since \mathbf{n} is rather large and \mathbf{p} is small, these binomial probabilities can be approximated extremely well by Poisson probabilities with $\lambda = 200(0.05) = 10$. That is, from the Poisson table, the probability of the Type I error is

$$\alpha \approx \sum_{y=0}^5 \frac{10^y e^{-10}}{y!} = 0.067.$$

Thus, the approximate significance level of this test is $\alpha = 0.067$. This value is reasonably small. However, what about the probability of Type II error in case \mathbf{p} has been improved to 0.02, say? This error occurs if $Y > 5$ when, in fact, $\mathbf{p}=0.02$; hence its probability, denoted by β , is

$$\beta = P(Y > 5; p = 0.02) = \sum_{y=6}^{200} \binom{200}{y} (0.02)^y (0.98)^{200-y}.$$

Again we use the Poisson approximation, here $\lambda=200(0.02)=4$, to obtain

$$\beta \approx 1 - \sum_{y=0}^5 \frac{4^y e^{-4}}{y!} = 1 - 0.785 = 0.215.$$

The engineers and the statisticians who created this new procedure probably are not too pleased with this answer. That is, they note that if their new procedure of manufacturing circuits has actually decreased the probability of failure to 0.02 from 0.05 (a big improvement), there is still a good chance, 0.215, that $H_0: \mathbf{p}=0.05$ is accepted and their improvement rejected. Thus, this test of $H_0: \mathbf{p}=0.05$ against $H_1: \mathbf{p}=0.02$ is unsatisfactory. Without worrying more about the probability of the Type II error, here, above was presented a frequently used procedure for testing $H_0: \mathbf{p}=\mathbf{p}_0$, where \mathbf{p}_0 is some specified probability of success. This test is based upon the fact that the number of successes, \mathbf{Y} , in \mathbf{n} independent Bernoulli trials is such that Y/\mathbf{n} has an approximate normal distribution, $N[\mathbf{p}_0, \mathbf{p}_0(1-\mathbf{p}_0)/\mathbf{n}]$, provided $H_0: \mathbf{p}=\mathbf{p}_0$ is true and \mathbf{n} is large. Suppose the alternative hypothesis is $H_0: \mathbf{p}>\mathbf{p}_0$; that is, it has been hypothesized by a research worker

that something has been done to increase the probability of success. Consider the test of $H_0: p=p_0$ against $H_1: p > p_0$ that rejects H_0 and accepts H_1 if and only if

$$Z = \frac{Y/n - p_0}{\sqrt{p_0(1-p_0)/n}} \geq z_\alpha.$$

That is, if Y/n exceeds p_0 by standard deviations of Y/n , we reject H_0 and accept the hypothesis $H_1: p > p_0$. Since, under H_0 Z is approximately $N(0,1)$, the approximate probability of this occurring when $H_0: p=p_0$ is true is α . That is the significance level of that test is approximately α . If the alternative is $H_1: p < p_0$ instead of $H_1: p > p_0$, then the appropriate α -level test is given by $Z \leq -z_\alpha$. That is, if Y/n is smaller than p_0 by standard deviations of Y/n , we accept $H_1: p < p_0$.

In general, without changing the sample size or the type of the test of the hypothesis, a decrease in α causes an increase in β , and a decrease in β causes an increase in α . Both probabilities α and β of the two types of errors can be decreased only by increasing the sample size or, in some way, constructing a better test of the hypothesis.

1.1.1 EXAMPLE

If $n=100$ and we desire a test with significance level $\alpha=0.05$, then $\alpha = P(\bar{X} \geq c; \mu = 60) = 0.05$ means, since \bar{X} is $N(\mu, 100/100=1)$,

$$P\left(\frac{\bar{X} - 60}{1} \geq \frac{c - 60}{1}; \mu = 60\right) = 0.05$$

and $c - 60 = 1.645$. Thus $c=61.645$. The power function is

$$K(\mu) = P(\bar{X} \geq 61.645; \mu) = P\left(\frac{\bar{X} - \mu}{1} \geq \frac{61.645 - \mu}{1}; \mu\right) = 1 - \Phi(61.645 - \mu).$$

In particular, this means that β at $\mu=65$ is

$$= 1 - K(\mu) = \Phi(61.645 - 65) = \Phi(-3.355) \approx 0;$$

so, with $n=100$, both α and β have decreased from their respective original values of 0.1587 and 0.0668 when $n=25$. Rather than guess at the value of n , an ideal power function determines the sample size. Let us use a critical region of the form $\bar{x} \geq c$. Further, suppose that we want $\alpha=0.025$ and, when $\mu=65$, $\beta=0.05$. Thus, since \bar{X} is $N(\mu, 100/n)$,

$$0.025 = P(\bar{X} \geq c; \mu = 60) = 1 - \Phi\left(\frac{c - 60}{10/\sqrt{n}}\right)$$

and

$$0.05 = 1 - P(\bar{X} \geq c; \mu = 65) = \Phi\left(\frac{c - 65}{10/\sqrt{n}}\right).$$

That is, $\frac{c-60}{10/\sqrt{n}} = 1.96$ and $\frac{c-65}{10/\sqrt{n}} = -1.645$.

Solving these equations simultaneously for c and $10/\sqrt{n}$, we obtain

$$c = 60 + 1.96 \frac{5}{3.605} = 62.718;$$

$$\frac{10}{\sqrt{n}} = \frac{5}{3.605}.$$

Thus, $\sqrt{n} = 7.21$ and $n = 51.98$. Since n must be an integer, we would use $n=52$ and obtain $\alpha=0.025$ and $\beta=0.05$, approximately.

1.1.2

For a number of years there has been another value associated with a statistical test, and most statistical computer programs automatically print this out; it is called **the probability value** or, for brevity, **p-value**. The **p-value** associated with a test is the probability that we obtain the observed value of the test statistic or a value that is more extreme in the direction of the alternative hypothesis, calculated when H_0 is true. Rather than select the critical region ahead of time, the **p-value** of a test can be reported and the reader then makes a decision.

1.1.2.1

Say we are testing $H_0: \mu=60$ against $H_1: \mu>60$ with a sample mean \bar{X} based on $n=52$ observations. Suppose that we obtain the observed sample mean of $\bar{x} = 62.75$. If we compute the probability of obtaining an \bar{x} of that value of 62.75 or greater when $\mu=60$, then we obtain the **p-value** associated with $\bar{x} = 62.75$. That is,

$$\begin{aligned} p - \text{value} &= P(\bar{X} \geq 62.75; \mu = 60) = P\left(\frac{\bar{X}-60}{10/\sqrt{52}} \geq \frac{62.75-60}{10/\sqrt{52}}; \mu = 60\right) \\ &= 1 - \Phi\left(\frac{62.75-60}{10/\sqrt{52}}\right) = 1 - \Phi(1.983) = 0.0237. \end{aligned}$$

If this **p-value** is small, we tend to reject the hypothesis $H_0: \mu=60$. For example, rejection of $H_0: \mu=60$ if the **p-value** is less than or equal to 0.025 is exactly the same as rejection if $\bar{x} = 62.718$. That is, $\bar{x} = 62.718$ has a **p-value** of 0.025. To help keep the definition of **p-value** in mind, we note that it can be thought of as that **tail-end probability**, under H_0 , of the distribution of the statistic, here \bar{X} , beyond the observed value of the statistic. See Figure 1 (Figure 1) for the **p-value** associated with $\bar{x} = 62.75$.

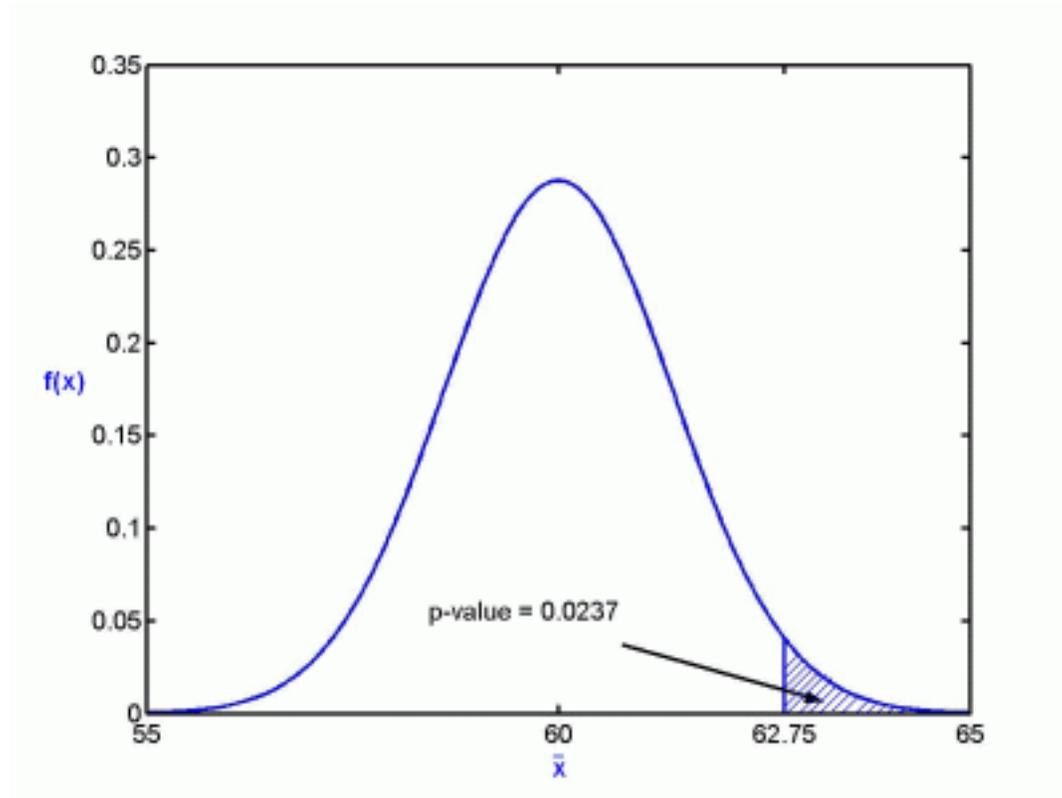


Figure 1: The p -value associated with $\bar{x} = 62.75$.

Example 1

Suppose that in the past, a golfer's scores have been (approximately) normally distributed with mean $\mu=90$ and $\sigma^2=9$. After taking some lessons, the golfer has reason to believe that the mean μ has decreased. (We assume that σ^2 is still about 9.) To test the null hypothesis $H_0: \mu=90$ against the alternative hypothesis $H_1: \mu < 90$, the golfer plays 16 games, computing the sample mean \bar{x} . If \bar{x} is small, say $\bar{x} \leq c$, then H_0 is rejected and H_1 accepted; that is, it seems as if the mean μ has actually decreased after the lessons. If $c=88.5$, then the power function of the test is

$$K(\mu) = P(\bar{X} \leq 88.5; \mu) = P\left(\frac{\bar{X} - \mu}{3/4} \leq \frac{88.5 - \mu}{3/4}; \mu\right) = \Phi\left(\frac{88.5 - \mu}{3/4}\right).$$

Because $9/16$ is the variance of \bar{X} . In particular,

$$\alpha = K(90) = \Phi(-2) = 1 - 0.9772 = 0.0228.$$

If, in fact, the true mean is equal to $\mu=88$ after the lessons, the power is $K(88) = \Phi(2/3) = 0.7475$. If $\mu=87$, then $K(87) = \Phi(2) = 0.9772$. An observed sample mean of $\bar{x} = 88.25$ has a

$$p\text{-value} = P(\bar{X} \leq 88.25; \mu = 90) = \Phi\left(\frac{88.25 - 90}{3/4}\right) = \Phi\left(-\frac{7}{3}\right) = 0.0098,$$

and this would lead to a rejection at $\alpha=0.0228$ (or even $\alpha=0.01$).