

Introduction to Hypothesis Testing

In Consumer Reports, April, 1978, the results of a taste test were reported. Consumer Reports commented, "we don't consider this result to be statistically significant." At the time, Miller had just bought Lowenbrau and Consumer's Union wanted to know if people could tell the difference between the two beers. Twenty-four tasters were given three carefully disguised glasses, one of the three with a different beer. The tasters were attempting to correctly identify the one that was different.



Figure 5: Three glasses of beer

Here we have a straightforward binomial hypothesis, i.e. that the tasters cannot tell the difference. We test $H_0 : p = \frac{1}{3}$ against $H_a : p > \frac{1}{3}$, where p denotes the probability of a correct choice. Note that there is no consideration of $p < \frac{1}{3}$, since that would have no meaning with respect to the capabilities of tasters. There is a natural (and sufficient!) statistic, $Y = \text{number of successes by the tasters}$. If there is random guessing, or the experiment is modeled as random guessing, and H_0 is true, then

$$E(Y) = np = \frac{1}{3} \cdot 24 = 8$$

If we get 8 successes, that is consistent with random guessing; if we get 9, that's better than guessing, but that could happen by chance. In fact, the probability of guessing correctly exactly

9 times is $\binom{24}{9} \left(\frac{1}{3}\right)^9 \left(\frac{2}{3}\right)^{15} = 0.1517$.

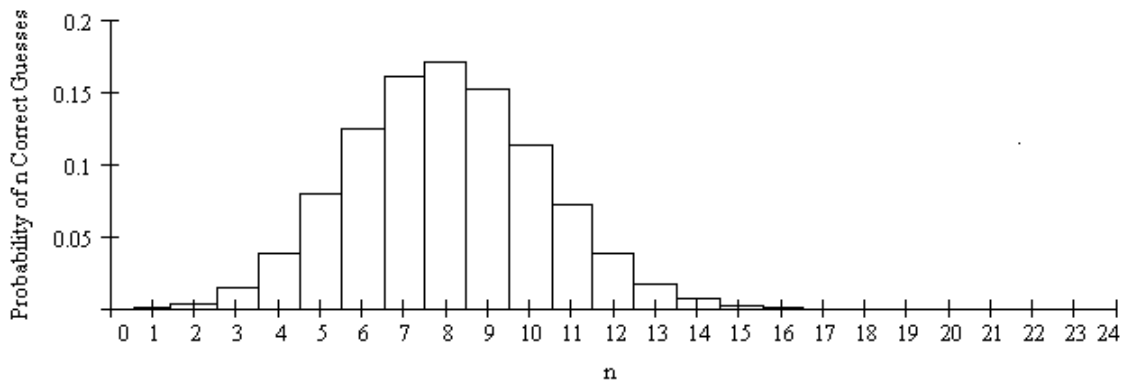
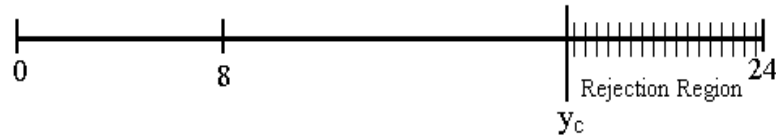


Figure 6: Distribution of Number of Correct Guesses with $p = \frac{1}{3}$

When should we reject H_0 ? Where do we draw the line to set off our rejection region? Somehow we need a *critical value*, y_c . That is, we need a value of y for which our decision will be to reject H_0 if $y \geq y_c$. Is having 11 or more correct sufficiently unusual to cause us to doubt the probability is one-third, or do we need stronger evidence?



Wherever we draw the line we could make a mistake! Here are the possibilities:

		Truth about Null Hypothesis	
		H_0 : True	H_0 :False
Decision Based on Data	Fail to Reject	Correct	Type II error
	Reject	Type I error	Correct

We don't treat the two types of errors equally. We discriminate on purpose against rejecting a true null hypothesis; we are conservative and want to make no claims of a false null hypothesis without good evidence. That is, we want $P(\text{type I error}) = P(\text{reject } H_0 | H_0 \text{ is true})$ to be small. $P(\text{type I error})$, denoted by α , can be determined once we know the form of the critical region. We decided previously to reject the null hypothesis if $y \geq y_c$. Therefore, we can find the probability of getting a set of outcomes in the critical region given that the null hypothesis is true:

$$\begin{aligned}
 P(\text{type I error}) &= P(\text{Rejecting } H_0 | H_0 \text{ is true}) \\
 &= P\left(y \geq y_c \mid p = \frac{1}{3}\right) \\
 &= \sum_{y=y_c}^{24} \binom{24}{y} \left(\frac{1}{3}\right)^y \left(\frac{2}{3}\right)^{24-y}
 \end{aligned}$$

We might, for instance, settle on a value of $\alpha = .05$ as a cutoff point for this probability. Then, we can calculate the following probabilities of type I error given possible cutoff values of y_c :

$$y_c = 12, P(\text{type I error}) = 0.0677 > 0.05$$

$$y_c = 13, P(\text{type I error}) = 0.0284 < 0.05$$

Our rejection region, based on these results, would be: $\{y : y \geq 13\}$ which tells us to reject the null hypothesis if y is greater than or equal to 13, and fail to reject if y is less than 13.

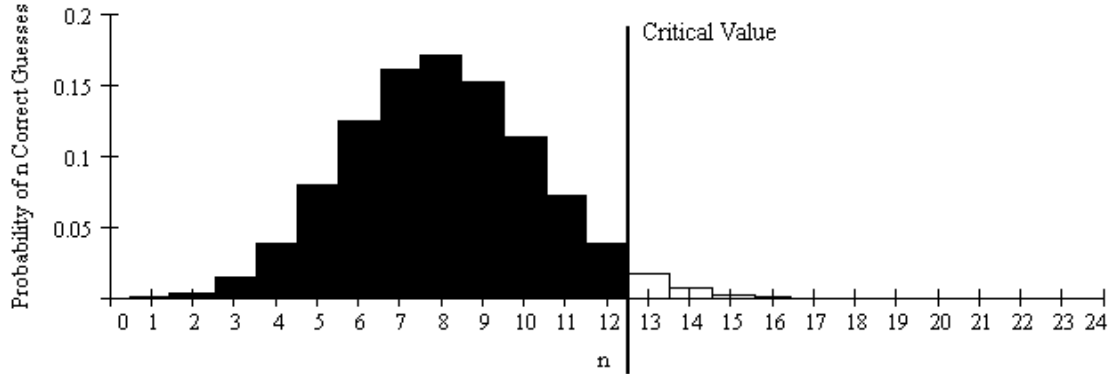


Figure 7: Distribution of Number of Correct Guesses with critical value at 13

Consumer Reports found eleven correct choices and concluded the results were not statistically significant. The p -value is the probability that we would get a result as extreme or more extreme than we did, if the null hypothesis is true. For the present study, this would be calculated:

$$p\text{-value} = \sum_{y=11}^{24} \binom{24}{y} \left(\frac{1}{3}\right)^y \left(\frac{2}{3}\right)^{24-y} = 0.14$$

Now we want to consider the possibility that we made a type II error in deciding not to reject H_0 . The probability of a type II error, commonly denoted by β , is a function of p , n , and α . In this example, n is fixed at 24 and $\alpha = 0.05$. β is, in fact, 0.0284.

$$\begin{aligned} \beta &= P(\text{type II error}) \\ &= P(\text{Fail to reject } H_0 \mid p) \\ &= P[y \leq (y_c - 1) \mid p]. \end{aligned}$$

In the last probability statement, $y_c - 1$ is used because the distribution is discrete. For example, $\{y < 13\} = \{y \leq 12\}$. For a continuous distribution, we would have

$$P[y \leq y_c \mid p].$$

For example:

If the true value of p is 0.5,

$$\begin{aligned}
 \mathbf{b} &= P[Y \leq (y_c - 1) | p = 0.5] \\
 &= P[Y \leq 12 | p = 0.5] \\
 &= 0.581
 \end{aligned}$$

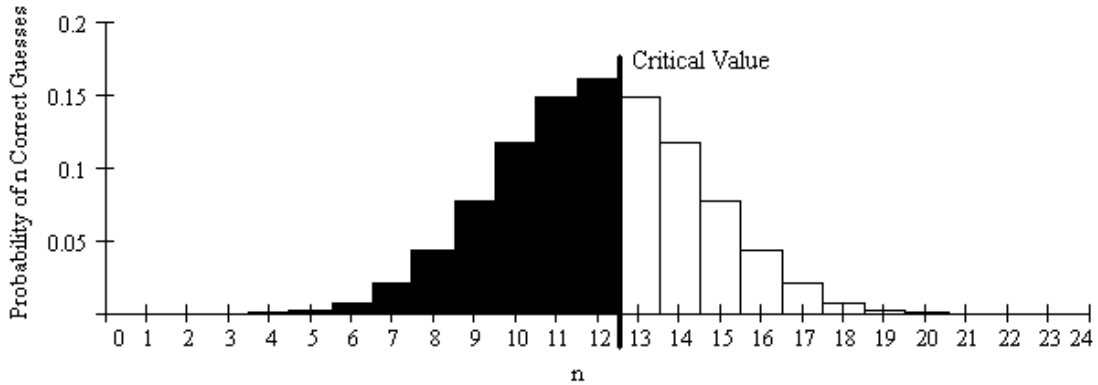


Figure 8: Distribution of Number of Correct Guesses with $p = \frac{1}{2}$

If the true value of p is 0.7,

$$\begin{aligned}
 \mathbf{b} &= P[Y \leq (y_c - 1) | p = 0.7] \\
 &= P[Y \leq 12 | p = 0.7] \\
 &= 0.031
 \end{aligned}$$

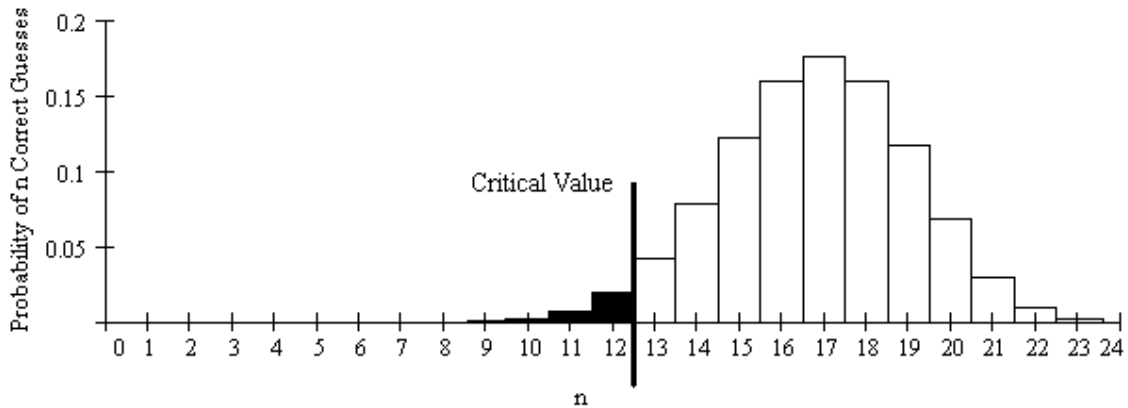


Figure 9: Distribution of Number of Correct Guesses with $p = 0.7$

These examples show that the probability of type II error is affected by the true value of the parameter. Other factors which affect the type II error are the level of the test, \mathbf{a} , and the sample size, n .

Power

Suppose that W is the test statistic and RR is the rejection region for a test of a hypothesis involving the value of a parameter \mathbf{q} . Then the *power* of the test is the probability that the test will lead to rejection of H_0 when the actual

parameter value is q . That is, $power(q) = P(W \text{ in RR when the parameter value is } q)$.

We usually calculate the *power* of a statistical test against a specific alternative by subtraction, thus power is $1 - \text{the probability of a type II error}$, or $1 - \beta$. Therefore, the power of the test against the alternative $p = 0.5$ is 0.419; the power of the test against the alternative $p = 0.7$ is 0.969. We can think of the power of a test as measuring the ability of the test to detect that the null hypothesis is false.

By repeating the calculations above for different assumed true values of p , we can create a table of values for β and power, and construct a graph of the power function for $n = 24, \alpha = 0.05$.

Probability	Beta	Power
0.40	0.886	0.114
0.45	0.758	0.242
0.50	0.581	0.419
0.55	0.385	0.615
0.60	0.213	0.787
0.65	0.094	0.906
0.70	0.031	0.969
0.75	0.007	0.993
0.80	0.001	0.999

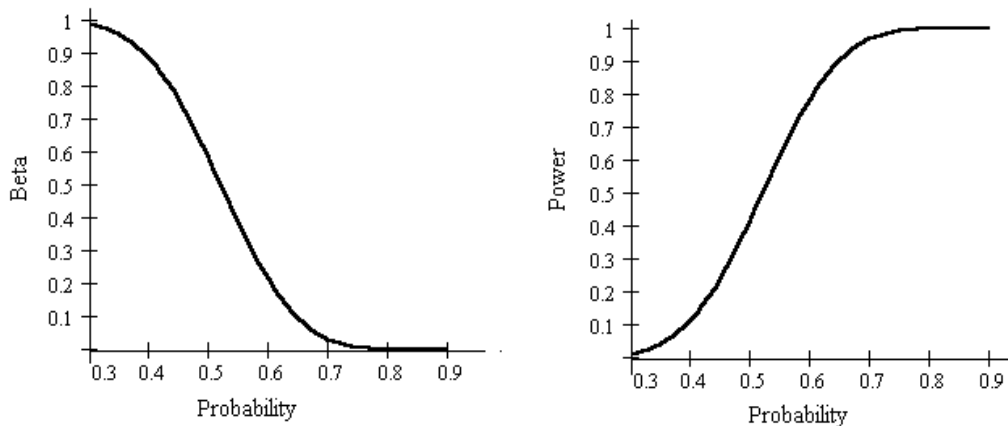


Figure 10: Beta and Power Curves

Sign Test

The Sign Test is a nonparametric test that deals with continuous random variables by converting them to binary alternatives. We take our data values and classify each as greater or less than the hypothesized median, i.e. we look at the sign of the value minus the median (stated in the null hypothesis), which we denote as m_0 . The structure of the hypothesis is:

H_0 : population median = m_0

H_a : population median $\neq m_0$

We define a random variable, $Y_i = \begin{cases} 1, & \text{if } y_i \geq m_0 \\ 0, & \text{if } y_i < m_0 \end{cases}$, where Y_i counts the number of

observations which are greater than or equal to the median. Then, $Y_i \sim$ binomial with probability of "success" equal to one half under the null hypothesis. The disadvantage to using the sign test instead of a t -test, for example, is a loss of power in the test.

Most Powerful Tests: Neyman-Pearson Lemma

In hypothesis testing situations, there are often several statistical tests from which to choose. Ideally we prefer tests with small probabilities of type I and type II error and high power. That is, we prefer to use the test with maximum power, often referred to as the *most powerful* test. When we are testing a *simple* null hypothesis $H_0: \mathbf{q} = \mathbf{q}_0$ versus a simple alternative hypothesis $H_a: \mathbf{q} = \mathbf{q}_a$, the following theorem provides the method for deriving the most powerful test. In the previous example, we considered $H_0: \mathbf{q}_0 = \frac{1}{3}$ and

$H_a: \mathbf{q}_a = \frac{7}{10}$. Are the data more likely with \mathbf{q}_0 or \mathbf{q}_a ?

Theorem: The Neyman-Pearson Lemma:

Suppose we wish to test the simple null hypothesis $H_0: \mathbf{q} = \mathbf{q}_0$ versus the simple alternative hypothesis $H_a: \mathbf{q} = \mathbf{q}_a$, based on a random sample Y_1, Y_2, \dots, Y_n from a distribution with parameter \mathbf{q} . Let $L(\mathbf{q})$ denote the likelihood of the sample when the value of the parameter is \mathbf{q} . Then for a given \mathbf{a} , the test that maximizes the power at \mathbf{q}_a has a rejection region determined

by $\frac{L(\mathbf{q}_0)}{L(\mathbf{q}_a)} < k$.

If the data are unlikely at \mathbf{q}_0 relative to \mathbf{q}_a , reject H_0 . We will not prove the Neyman-Pearson Lemma, but we will provide several examples to illustrate application of this theorem.

Essentially we will need to examine the ratio of likelihoods $\frac{L(\mathbf{q}_0)}{L(\mathbf{q}_a)}$ and choose the rejection

region that makes this ratio "small."

Example 1

Suppose that Y represents an observation from the probability density function given by

$$f(y|\mathbf{q}) = \begin{cases} \mathbf{q} y^{\mathbf{q}-1}, & 0 < y < 1 \\ 0, & \text{elsewhere.} \end{cases}$$

We want to test $H_0 : \mathbf{q} = 1$ versus $H_a : \mathbf{q} = 2$. Find the most powerful test with significance level $\alpha = .10$.

Solution:

Since we have only one observation, the likelihood function is simply the probability density function. So $\frac{L(\mathbf{q}_0)}{L(\mathbf{q}_a)} = \frac{f(y|\mathbf{q}_0)}{f(y|\mathbf{q}_a)} = \frac{f(y|\mathbf{q}=1)}{f(y|\mathbf{q}=2)} = \frac{1}{2y}$ for $0 < y < 1$. According to the Neyman-

Pearson Lemma, the form of the rejection region for the most powerful test is $\frac{1}{2y} < k$, or

$2y > \frac{1}{k}$. This inequality can be rewritten as $y > c$ where $c = \frac{1}{2k}$ and can be interpreted as “reject H_0 if and only if y is too large.” The exact value of c depends on the level of the test.

A graph showing the two density functions $f(y|\mathbf{q}=1) = 1$ and $f(y|\mathbf{q}=2) = 2y$, is shown below.

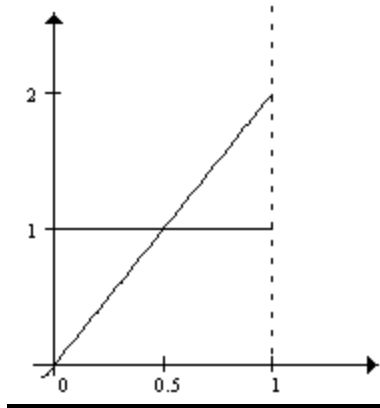


Figure 11: Density functions $f(y|\mathbf{q}=1) = 1$ and $f(y|\mathbf{q}=2) = 2y$

Looking at the graph we see that small values of y are more consistent with the uniform density function and that large values of y are more consistent with the $f(y) = 2y$ density curve.

To determine the value of c for which $y > c$ leads to rejection if $H_0 : \mathbf{q} = 1$ is true, we set $\alpha = .10$ equal to the probability of rejecting the null hypothesis when it is true. That is,

$$.10 = P(y > c | \mathbf{q} = 1) = \int_c^1 1 dy = y \Big|_c^1 = 1 - c \text{ and solve for } c = 0.9 .$$

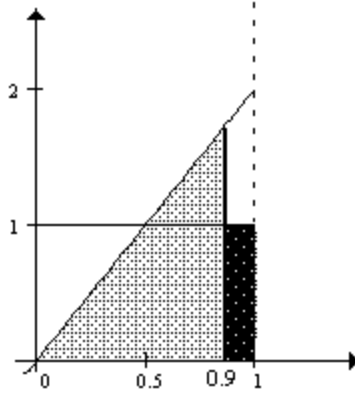


Figure 12: Rejection region

Notice that the probability \mathbf{b} of type II error for this test is

$$P(\text{fail to reject } H_0 : \mathbf{q} = 1 \text{ when } \mathbf{q} = 2) = P(y \leq 0.9 | \mathbf{q} = 2) = \int_0^{0.9} 2y dy = y^2 \Big|_0^{0.9} = 0.81.$$

The large value of \mathbf{b} for this test results from the small sample size used ($n = 1$). Though \mathbf{b} is large, this test has the smallest possible \mathbf{b} among all tests for $H_0 : \mathbf{q} = 1$ versus $H_a : \mathbf{q} = 2$ based on sample size one and $\mathbf{a} = 0.10$.

Example 2

Let Y_1, Y_2, \dots, Y_n denote a random sample from a Bernoulli-distributed population with parameter \mathbf{q} . Let $U = \sum_{i=1}^n Y_i$. Derive the most powerful test for testing $H_0 : \mathbf{q} = 1/2$ versus $H_a : \mathbf{q} = 1/3$.

Solution

$$L(\mathbf{q}_0) = L(1/2) = \left(\frac{1}{2}\right)^u \left(\frac{1}{2}\right)^{n-u} \quad L(\mathbf{q}_a) = L(1/3) = \left(\frac{1}{3}\right)^u \left(\frac{2}{3}\right)^{n-u}$$

$$\frac{L(\mathbf{q}_0)}{L(\mathbf{q}_a)} = \frac{(1/2)^u (1/2)^n (1/2)^{-u}}{(1/3)^u (2/3)^n (2/3)^{-u}} = \left(\frac{3}{4}\right)^n \frac{1}{2^{-u}} = \left(\frac{3}{4}\right)^n 2^u$$

$$\text{Reject } H_0 \text{ if and only if } \left(\frac{3}{4}\right)^n 2^u < k \text{ or } 2^u < \left(\frac{4}{3}\right)^n k = k_1$$

$$\text{Taking logarithms, we reject if and only if } u \ln 2 < \ln k_1 = k_2 \text{ or } u < \frac{k_2}{\ln 2} = k_3.$$

$$\therefore \text{Reject } H_0 \text{ if and only if } U = \sum_{i=1}^n Y_i \text{ is too small.}$$

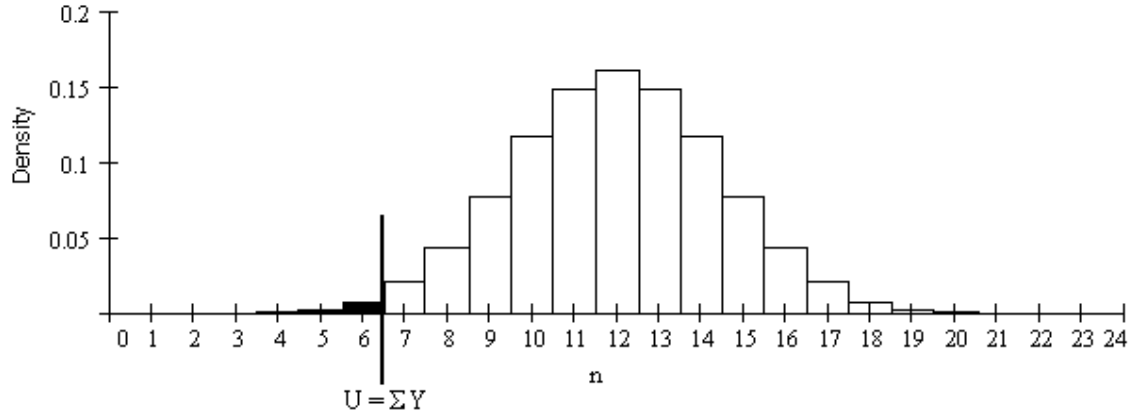


Figure 13: Reject H_0 if and only if $U = \sum_{i=1}^n Y_i$ is too small.

Notice that we have used the Neyman-Pearson Lemma to derive the *form* of the rejection region. The actual rejection region will depend upon the specified value for α . With large samples, instead of examining $U = \sum_{i=1}^n Y_i$ to decide whether or not to reject H_0 when working

with Bernoulli trials, we typically consider $Z = \frac{\hat{q} - q_0}{\sqrt{\frac{q_0(1-q_0)}{n}}}$ and reject H_0 if Z is too small.

It can be shown that the most powerful test for $H_0 : q = 1/2$ versus $H_a : q = 1/4$ leads to the same conclusion to reject H_0 if $U = \sum_{i=1}^n Y_i$ is too small. Similarly, whenever testing $H_0 : q = 1/2$ versus $H_a : q = q_a$ with $q_a < 1/2$, the same rejection region decision applies. That is, the form of the rejection region depends only on the fact that $q_a < q_0$, not on the particular choice of q_a . Thus the test result found above (to reject H_0 if U or Z is sufficiently small) is referred to as *uniformly most powerful* since it maximizes the power for every value of q less than q_0 .

Example 3

Let Y_1, Y_2, \dots, Y_n denote a random sample from a population having a Poisson distribution with mean λ . Find the form of the rejection region for a most powerful test of $H_0 : \lambda = \lambda_0$ versus $H_a : \lambda = \lambda_a$ when $\lambda_a > \lambda_0$.

Solution

$$f(y|\lambda) = \frac{\lambda^y e^{-\lambda}}{y!}$$

$$L(I_0) = f(y_1|I_0)f(y_2|I_0)\dots f(y_n|I_0) = \frac{I_0^{y_1} e^{-I_0}}{y_1!} \cdot \frac{I_0^{y_2} e^{-I_0}}{y_2!} \dots \frac{I_0^{y_n} e^{-I_0}}{y_n!}$$

$$= \frac{I_0^{\sum y_i} e^{-nI_0}}{\prod y_i!}$$

Similarly, $L(I_a) = \frac{I_a^{\sum y_i} e^{-nI_a}}{\prod y_i!}$

So $\frac{L(I_0)}{L(I_a)} = \frac{I_0^{\sum y_i} e^{-nI_0}}{I_a^{\sum y_i} e^{-nI_a}} = \left(\frac{I_0}{I_a}\right)^{\sum y_i} \frac{e^{nI_a}}{e^{nI_0}} = \left(\frac{I_0}{I_a}\right)^{\sum y_i} e^{n(I_a - I_0)}$

Reject H_0 if and only if $\left(\frac{I_0}{I_a}\right)^{\sum y_i} e^{n(I_a - I_0)} < k$ or

$$\left(\frac{I_0}{I_a}\right)^{\sum y_i} < \frac{k}{e^{n(I_a - I_0)}} = k_1$$

Taking logarithms, we reject H_0 if and only if $\sum_{i=1}^n y_i \cdot \ln\left(\frac{I_0}{I_a}\right) < \ln(k_1) = k_2$

Recall that $I_a > I_0$, so $\frac{I_0}{I_a} < 1$ and $\ln\left(\frac{I_0}{I_a}\right) < 0$.

\therefore Reject H_0 if and only if $\sum_{i=1}^n y_i > \frac{k_2}{\ln(I_0/I_a)}$.

That is, reject H_0 if and only if $\sum_{i=1}^n y_i$ is too large. Thus, if we want to test $I = 2$ messages per hour on e-mail against the alternative $I = 4$ messages per hour, we would reject the null hypothesis if we have too many messages. If we had tested, instead, $I_a < I_0$, for example, $I = 2$ messages per hour on e-mail against the alternative $I = 0.5$ messages per hour, all of the work above would be the same except for reversing the inequality in the last line. The result would be to reject the null hypothesis if the number of messages is too small. However, there is no uniformly most powerful test for $H_a : I \neq I_0$.

In each situation, the results of the Neyman-Pearson Lemma match our intuition. It doesn't give us any counter-intuitive or surprising result, but adds support to the methods that are taught in the first year statistics course. As we shall see in the next section, the standard t -test is an example of a generalized likelihood ratio test.

Likelihood Ratio Tests

The Neyman-Pearson Lemma provides a method of constructing most powerful tests in very limited situations. This method can be used only when the distribution of the observations is known except for a single unknown parameter and when there is a single alternative value. In

many situations the distribution of the observations has more than one unknown parameter and/or the alternative hypothesis is more general than a statement of a single alternative value. In these situations the rejection region can be determined by a likelihood ratio test.

Suppose we want to test $H_0 : \mathbf{q} \in \Omega_0$ versus $H_a : \mathbf{q} \in \Omega_a$ where \mathbf{q} may involve several parameters. We will let $\Omega = \Omega_0 \cup \Omega_a$. We want to maximize the likelihood subject to H_0 being true.

$$\text{Let } L(\hat{\Omega}_0) = \max_{\mathbf{q} \in \Omega_0} L(\mathbf{q}).$$

We also want to maximize the likelihood without the constraint that H_0 is true. That is, we want to maximize the likelihood subject to H_0 being true or false.

$$\text{Let } L(\hat{\Omega}) = \max_{\mathbf{q} \in \Omega} L(\mathbf{q})$$

If H_0 is true, the ratio of these likelihoods should be close to 1. If H_0 is "very false", the ratio of these likelihoods should be close to 0. So we will look at $I = \frac{L(\hat{\Omega}_0)}{L(\hat{\Omega})}$ (note $0 \leq I \leq 1$) and, according to the likelihood ratio test, reject H_0 if and only if $I \leq k$. (That is, if and only if I is sufficiently small.)

Suppose that Y_1, Y_2, \dots, Y_n constitute a random sample from a normal population with unknown mean m and unknown variance s^2 . We want to test $H_0 : m = m_0$ versus $H_a : m > m_0$. It can be shown that solving $I = \frac{L(\hat{\Omega}_0)}{L(\hat{\Omega})} \leq k$ leads to $\frac{\bar{y} - m_0}{s/\sqrt{n}} > \text{constant}$.

Note: A uniformly most powerful test does not exist for this situation.

(Reference: *Mathematical Statistics with Applications*, Wackerly, Mendenhall, and Scheaffer, Duxbury Press, 1996, pages 464-466.) Thus the one-sample t -test is the generalized likelihood ratio test.

Equivalence of Chi-Square Tests and Z-Tests for Proportions

In this section we will show that Z -tests for proportions are algebraically equivalent to chi-square tests. We will begin with the test for a single proportion where the hypotheses are

$$H_0 : p = p_0 \text{ and } H_a : p \neq p_0.$$

To perform a chi-square test, we would have observed and expected values as shown in the table below:

	Yes	No	Total
Observed	y	$n - y$	n
Expected	$n p_0$	$n(1 - p_0)$	n

Since $\hat{p} = y/n$, $y = n\hat{p}$, this table is equivalent to

	Yes	No	Total
Observed	$n \hat{p}$	$n - n \hat{p}$	n
Expected	$n p_0$	$n(1 - p_0)$	n

The chi-square statistic with one degree of freedom is

$$\begin{aligned}
 c^2 &= \sum_{i=1}^2 \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \\
 &= \frac{(y - np_0)^2}{np_0} + \frac{[(n - y) - n(1 - p_0)]^2}{n(1 - p_0)} \\
 &= \frac{(n\hat{p} - np_0)^2}{np_0} + \frac{[(n - n\hat{p}) - n(1 - p_0)]^2}{n(1 - p_0)} \\
 &= \frac{n^2(\hat{p} - p_0)^2}{np_0} + \frac{n^2[(1 - \hat{p}) - (1 - p_0)]^2}{n(1 - p_0)} \\
 &= \frac{n(\hat{p} - p_0)^2}{p_0} + \frac{n(-\hat{p} + p_0)^2}{(1 - p_0)} \\
 &= n(\hat{p} - p_0)^2 \left\{ \frac{1}{p_0} + \frac{1}{1 - p_0} \right\} \\
 &= n(\hat{p} - p_0)^2 \left\{ \frac{1}{p_0(1 - p_0)} \right\} \\
 &= \frac{(\hat{p} - p_0)^2}{p_0(1 - p_0)/n}
 \end{aligned}$$

This last expression we recognize as Z^2 with $Z^2 = \left[\frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}} \right]^2$

Note that large values of n are required for $\sum_{i=1}^2 \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$ to be approximately

χ^2 distributed and for $\frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$ to be approximately Z . We already know that Z^2 is

χ^2 with 1 degree of freedom. It should be no surprise that the critical value for χ^2 with 1 degree of freedom with $\alpha = 0.05$ is $1.96^2 = 3.84$.

Two Proportions Tests

Now consider the test for two proportions where the hypotheses are

$$H_0 : p_1 = p_2 \text{ and } H_a : p_1 \neq p_2$$

For this test, $\hat{p}_1 = \frac{y_1}{n_1}$, $\hat{p}_2 = \frac{y_2}{n_2}$, and pooled $\hat{p} = \frac{y_1 + y_2}{n_1 + n_2} = \frac{y}{n}$.

To perform a chi-square test, we have values as shown in the tables below:

	Group 1	Group 2	Total
Success	y_1	y_2	$y_1 + y_2 = y$
Failure	$n_1 - y_1$	$n_2 - y_2$	$n_1 + n_2 - y_1 - y_2 = n - y$
Total	n_1	n_2	$n_1 + n_2 = n$

The expected number of successes is $\frac{n_1 y}{n}$ for Group 1 and $\frac{n_2 y}{n}$ for Group 2. The expected number of failures is $\frac{n_1(n-y)}{n}$ for Group 1 and $\frac{n_2(n-y)}{n}$ for Group 2. These values are summarized in the table below.

	Group 1	Group 2	Total
Success	$\frac{n_1 y}{n} = n_1 \hat{p}$	$\frac{n_2 y}{n} = n_2 \hat{p}$	y
Failure	$\frac{n_1(n-y)}{n} = n_1(1-\hat{p})$	$\frac{n_2(n-y)}{n} = n_2(1-\hat{p})$	$n - y$
Total	n_1	n_2	n

The chi-square statistic with one degree of freedom is

$$\begin{aligned}
\mathbf{c}^2 &= \sum_{i=1}^4 \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \\
&= \frac{(y_1 - n_1 \hat{p})^2}{n_1 \hat{p}} + \frac{(y_2 - n_2 \hat{p})^2}{n_2 \hat{p}} + \frac{[(n_1 - y_1) - n_1(1 - \hat{p})]^2}{n_1(1 - \hat{p})} + \frac{[(n_2 - y_2) - n_2(1 - \hat{p})]^2}{n_2(1 - \hat{p})} \\
&= \frac{(n_1 \hat{p}_1 - n_1 \hat{p})^2}{n_1 \hat{p}} + \frac{(n_2 \hat{p}_2 - n_2 \hat{p})^2}{n_2 \hat{p}} + \frac{[n_1(1 - \hat{p}_1) - n_1(1 - \hat{p})]^2}{n_1(1 - \hat{p})} + \frac{[n_2(1 - \hat{p}_2) - n_2(1 - \hat{p})]^2}{n_2(1 - \hat{p})} \\
&= \frac{n_1(\hat{p}_1 - \hat{p})^2}{\hat{p}} + \frac{n_2(\hat{p}_2 - \hat{p})^2}{\hat{p}} + \frac{n_1(\hat{p} - \hat{p}_1)^2}{(1 - \hat{p})} + \frac{n_2(\hat{p} - \hat{p}_2)^2}{(1 - \hat{p})} \\
&= n_1(\hat{p}_1 - \hat{p})^2 \left\{ \frac{1}{\hat{p}} + \frac{1}{1 - \hat{p}} \right\} + n_2(\hat{p}_2 - \hat{p})^2 \left\{ \frac{1}{\hat{p}} + \frac{1}{1 - \hat{p}} \right\} \\
&= \left\{ \frac{1}{\hat{p}} + \frac{1}{1 - \hat{p}} \right\} \left[n_1(\hat{p}_1 - \hat{p})^2 + n_2(\hat{p}_2 - \hat{p})^2 \right] \\
&\quad \text{note that } \hat{p} = \frac{y_1 + y_2}{n_1 + n_2} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2} \\
&= \left\{ \frac{1}{\hat{p}} + \frac{1}{1 - \hat{p}} \right\} \left[n_1 \left(\hat{p}_1 - \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2} \right)^2 + n_2 \left(\hat{p}_2 - \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2} \right)^2 \right] \\
&= \frac{1}{\hat{p}(1 - \hat{p})} \left[n_1 \left(\frac{n_2 \hat{p}_1 - n_2 \hat{p}_2}{n_1 + n_2} \right)^2 + n_2 \left(\frac{n_1 \hat{p}_2 - n_1 \hat{p}_1}{n_1 + n_2} \right)^2 \right] \\
&= \frac{1}{\hat{p}(1 - \hat{p})} \left[\frac{n_1 n_2^2 (\hat{p}_1 - \hat{p}_2)^2}{(n_1 + n_2)^2} + \frac{n_2 n_1^2 (\hat{p}_1 - \hat{p}_2)^2}{(n_1 + n_2)^2} \right] \\
&= \frac{1}{\hat{p}(1 - \hat{p})} \frac{(\hat{p}_1 - \hat{p}_2)^2}{(n_1 + n_2)^2} [n_1 n_2 (n_2 + n_1)] \\
&= \frac{1}{\hat{p}(1 - \hat{p})} \frac{(\hat{p}_1 - \hat{p}_2)^2}{(n_1 + n_2)} (n_1 n_2) \\
&= \frac{(\hat{p}_1 - \hat{p}_2)^2}{\hat{p}(1 - \hat{p}) \left(\frac{n_1 + n_2}{n_1 n_2} \right)} = \frac{(\hat{p}_1 - \hat{p}_2)^2}{\hat{p}(1 - \hat{p}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} \\
&= \left[\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})} \sqrt{1/n_1 + 1/n_2}} \right]^2 = Z^2
\end{aligned}$$

Note that this Z is approximate since the variance is estimated from the data. As before, we note that \mathbf{c}^2 on a 2 x 2 table has 1 degree of freedom, and Z^2 is a \mathbf{c}^2 with 1 degree of freedom.

In our earlier work with moment generating functions we showed that the probability distribution of the square of a $N(0,1)$ random variable, that is Z^2 , is χ^2 with one degree of freedom. In this section we have demonstrated the equivalence (that we would expect) when we use a chi-square test and a Z -test for a single proportion and for equality of two proportions.