

## Hypothesis Testing and Power with the Binomial Distribution

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

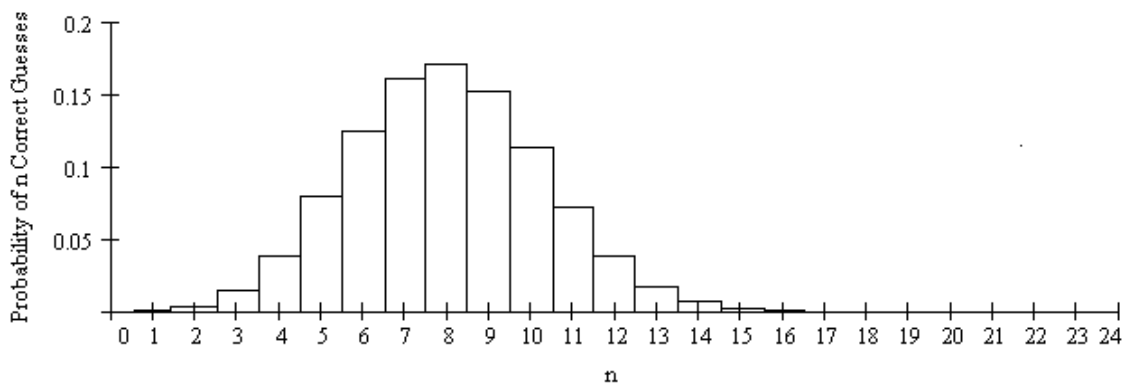
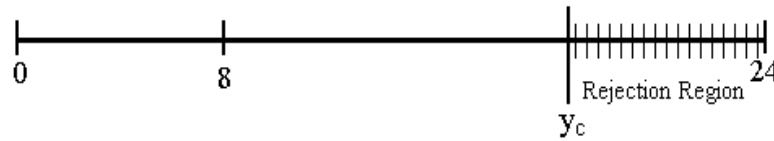


Figure 6: Distribution of Number of Correct Guesses with  $p = \frac{1}{3}$

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  $\alpha$ , 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  $y_c$ :

$$\begin{aligned}
 y_c = 12, & P(\text{type I error}) = 0.0677 > 0.05 \\
 y_c = 13, & P(\text{type I error}) = 0.0284 < 0.05
 \end{aligned}$$

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.

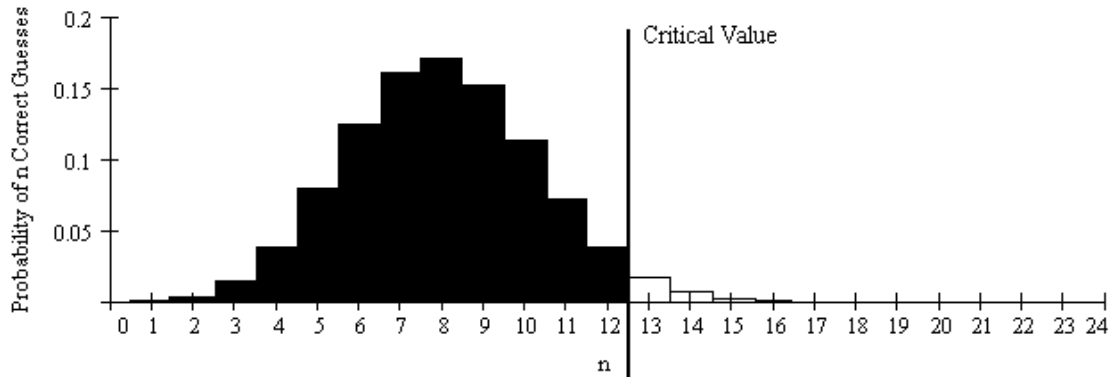


Figure 7: Distribution of Number of Correct Guesses with critical value at 13

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  $\beta$ , is a function of  $p$ ,  $n$ , and  $\alpha$ . In this example,  $n$  is fixed at 24 and  $\alpha = 0.05$ .  $\alpha$  is, in fact, 0.0284.

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

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

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

For example:

If the true value of  $p$  is 0.5,

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

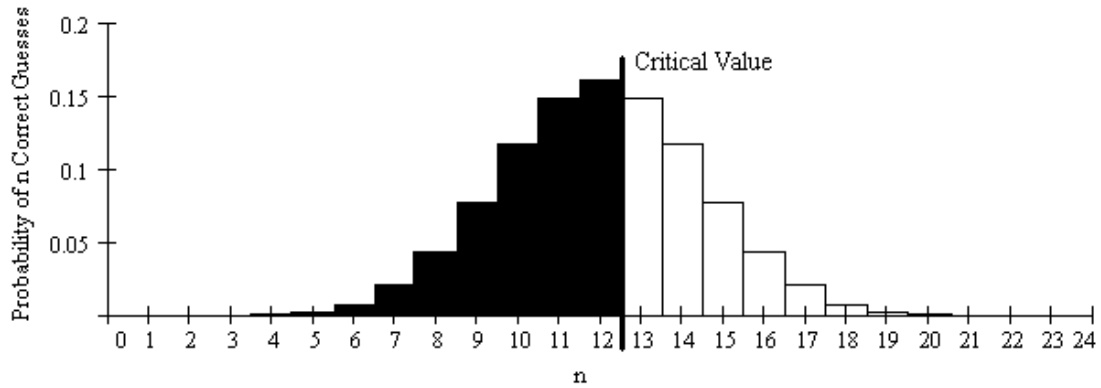


Figure 8: Distribution of Number of Correct Guesses with  $p = \frac{1}{2}$

If the true value of  $p$  is 0.7,

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

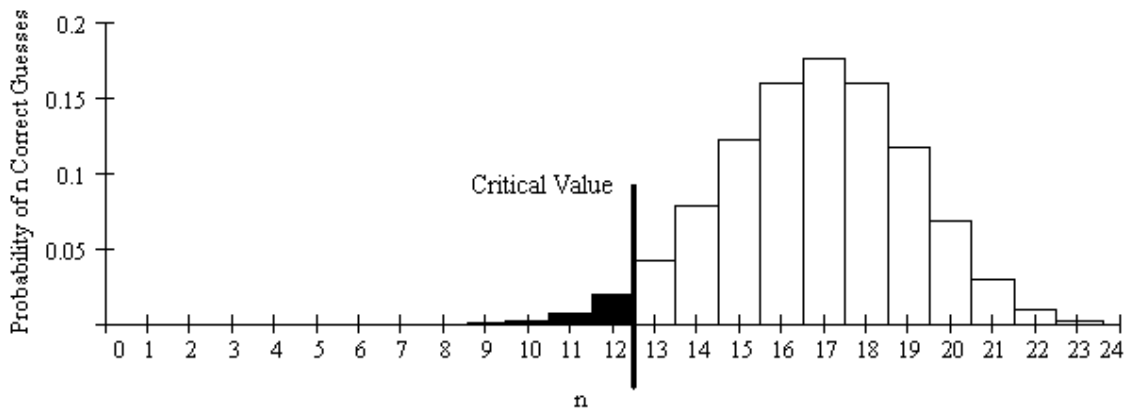


Figure 9: Distribution of Number of Correct Guesses with  $p = 0.7$

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

We usually calculate the *power* of a statistical test against a specific alternative by subtraction, thus power is 1 – 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  $\beta$  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

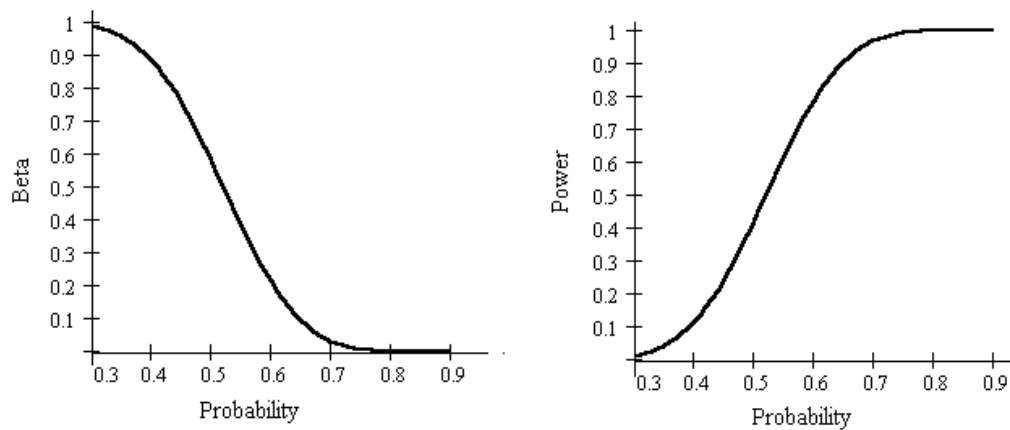


Figure 10: Beta and Power Curves