Fundamentals of Hypothesis Testing

James H. Steiger

Department of Psychology and Human Development
Vanderbilt University

Fundamentals of Hypothesis Testing

Introduction

- In this module, we review the basics of hypothesis testing.

Introduction

- In this module, we review the basics of hypothesis testing.
- We shall develop the *binomial distribution* formulas, show how they lead to some important *sampling distributions*, and then investigate the key principles of hypothesis testing.

The Binomial Distribution

- A *binomial process* is characterized by the following:

The Binomial Distribution

- A *binomial process* is characterized by the following:
  1. There are n *independent* trials

The Binomial Distribution

- A *binomial process* is characterized by the following:
  1. There are n *independent* trials
  2. Only two things can happen on each trial. We might arbitrarily label them "success" and "failure."

The Binomial Distribution

- A *binomial process* is characterized by the following:
  1. There are *n independent* trials
  2. Only two things can happen on each trial. We might arbitrarily label them "success" and "failure."
  3. The probabilities of "success" and "failure" are π and $1 - \pi$.

The Binomial Distribution

- A *binomial process* is characterized by the following:
  1. There are n *independent* trials
  2. Only two things can happen on each trial. We might arbitrarily label them "success" and "failure."
  3. The probabilities of "success" and "failure" are π and $1 - \pi$.
- The binomial random variable Y is the number of successes in the n trials.

The Binomial Distribution

- A *binomial process* is characterized by the following:
  1. There are n *independent* trials
  2. Only two things can happen on each trial. We might arbitrarily label them "success" and "failure."
  3. The probabilities of "success" and "failure" are π and $1 - \pi$.
- The binomial random variable Y is the number of successes in the n trials.
- Of course, Y is a random variable, and the number of successes that actually occur in any sequence is uncertain unless $\pi = 0$ or $\pi = 1$.

The Binomial Distribution

- A *binomial process* is characterized by the following:
  1. There are n *independent* trials
  2. Only two things can happen on each trial. We might arbitrarily label them "success" and "failure."
  3. The probabilities of "success" and "failure" are π and $1 - \pi$.
- The binomial random variable Y is the number of successes in the n trials.
- Of course, Y is a random variable, and the number of successes that actually occur in any sequence is uncertain unless $\pi = 0$ or $\pi = 1$.
- The *binomial distribution* $p(y) = \Pr(Y = y)$ assigns probabilities to each (potential) number of successes.

The Binomial Distribution

Example (The Binomial Distribution)

A couple plans to have 4 children, and to allow the sex of the child to be determined randomly. Assume that the probability of any child being a boy is 0.51. What is the probability that of the 4 children, there are exactly 3 boys and 1 girl?

We'll load the code in *full.binomial.txt* and use the function to generate the entire probability distribution:

```
> full.binomial <- function(n, pi) {
+     a <- matrix(0:n, n + 1, 1)
+     b <- dbinom(a, n, pi)
+     c <- pbinom(a, n, pi)
+     result <- cbind(a, b, c, 1 - c)
+     rownames(result) <- rep("", n + 1)
+     colnames(result) <- c("y", "Pr(Y = y)", "Pr(Y <= y)", "Pr(Y > y)")
+     return(result)
+ }
```

The Binomial Distribution

Example (The Binomial Distribution)

As you can see, the probability of having exactly 3 boys is just a smidgen below 0.26. The probability of having more girls than boys is $\Pr(Y \leq 1)$, or roughly 0.298.

```
> full.binomial(4, 0.51)

 y Pr(Y = y) Pr(Y <= y) Pr(Y > y)
 0   0.05765    0.05765   0.94235
 1   0.24000    0.29765   0.70235
 2   0.37470    0.67235   0.32765
 3   0.26000    0.93235   0.06765
 4   0.06765    1.00000   0.00000
```

Derivation of the Binomial Distribution Formula

- We shall develop the binomial distribution formula in terms of the preceding example.

Derivation of the Binomial Distribution Formula

- We shall develop the binomial distribution formula in terms of the preceding example.
- In this example, there are 4 "trials", and the probability of "success" is 0.51.

Derivation of the Binomial Distribution Formula

- We shall develop the binomial distribution formula in terms of the preceding example.
- In this example, there are 4 "trials", and the probability of "success" is 0.51.
- We wish to know $\Pr(Y = 3)$.

Derivation of the Binomial Distribution Formula

- We shall develop the binomial distribution formula in terms of the preceding example.
- In this example, there are 4 "trials", and the probability of "success" is 0.51.
- We wish to know $\Pr(Y = 3)$.
- To begin, we recognize that there are several ways the event $Y = 3$ might occur. For example, the first child might be a girl, and the next 3 boys, i.e., the sequence GBBB. What is the probability of this *particular* sequence?

Derivation of the Binomial Distribution Formula

- Since the trials are independent, we can say $\Pr(GBBB) = \Pr(G)\Pr(B)\Pr(B)\Pr(B)$. This probability is $0.49 \times 0.51 \times 0.51 \times 0.51 = .51^3 \times .49^1 = 0.06499899$

Derivation of the Binomial Distribution Formula

- Since the trials are independent, we can say $\Pr(GBBB) = \Pr(G)\Pr(B)\Pr(B)\Pr(B)$. This probability is $0.49 \times 0.51 \times 0.51 \times 0.51 = .51^3 \times .49^1 = 0.06499899$
- This is just a smidgen below .065.

Derivation of the Binomial Distribution Formula

- Since the trials are independent, we can say $\Pr(GBBB) = \Pr(G)\Pr(B)\Pr(B)\Pr(B)$. This probability is $0.49 \times 0.51 \times 0.51 \times 0.51 = .51^3 \times .49^1 = 0.06499899$
- This is just a smidgen below .065.
- Since the order of multiplication doesn't matter, we quickly realize that any other sequence involving 3 boys and 1 girl will have this same probability.

Derivation of the Binomial Distribution Formula

- Since the trials are independent, we can say $\Pr(GBBB) = \Pr(G)\Pr(B)\Pr(B)\Pr(B)$. This probability is $0.49 \times 0.51 \times 0.51 \times 0.51 = .51^3 \times .49^1 = 0.06499899$
- This is just a smidgen below .065.
- Since the order of multiplication doesn't matter, we quickly realize that any other sequence involving 3 boys and 1 girl will have this same probability.
- Suppose there are k such sequences. Then the total probability of having exactly 3 boys is $k \times .51^3 \times .49^1$. More generally, we can say that the probability of any *particular* sequence involving y successes is $\pi^y \times (1-\pi)^{n-y}$, and so

$$\Pr(Y = y) = k \times \pi^y \times (1-\pi)^{n-y}$$

Derivation of the Binomial Distribution Formula

- Since the trials are independent, we can say $\Pr(GBBB) = \Pr(G)\Pr(B)\Pr(B)\Pr(B)$. This probability is $0.49 \times 0.51 \times 0.51 \times 0.51 = .51^3 \times .49^1 = 0.06499899$
- This is just a smidgen below .065.
- Since the order of multiplication doesn't matter, we quickly realize that any other sequence involving 3 boys and 1 girl will have this same probability.
- Suppose there are k such sequences. Then the total probability of having exactly 3 boys is $k \times .51^3 \times .49^1$. More generally, we can say that the probability of any *particular* sequence involving y successes is $\pi^y \times (1-\pi)^{n-y}$, and so

$$\Pr(Y = y) = k \times \pi^y \times (1-\pi)^{n-y}$$

- But what is k?

Derivation of the Binomial Distribution Formula

Combinations

- In Psychology 310, we learned the basic combinatorial formulas. A key formula is *the number of ways y objects can be selected from n distinctly different objects without respect to order*.

Derivation of the Binomial Distribution Formula

Combinations

- In Psychology 310, we learned the basic combinatorial formulas. A key formula is *the number of ways y objects can be selected from n distinctly different objects without respect to order*.
- For example, you have the 4 letters A,B,C,D. How many different sets of size 2 may be selected from these 4 letters?

Derivation of the Binomial Distribution Formula

Combinations

- In Psychology 310, we learned the basic combinatorial formulas. A key formula is *the number of ways y objects can be selected from n distinctly different objects without respect to order*.
- For example, you have the 4 letters A,B,C,D. How many different sets of size 2 may be selected from these 4 letters?
- This is called "the number of combinations of 4 objects taken 2 at a time," or "4 choose 2."

Derivation of the Binomial Distribution Formula

Combinations

- In general, we ask, what is n choose y.

Derivation of the Binomial Distribution Formula

Combinations

- In general, we ask, what is n choose y.
- This quantity is often symbolized with the notations $\binom{n}{y}$ or (less frequently) ${}_nC_y$.

Derivation of the Binomial Distribution Formula

Combinations

- In general, we ask, what is n choose y.
- This quantity is often symbolized with the notations $\binom{n}{y}$ or (less frequently) ${}_nC_y$.
- This can be computed as the following ratio of two products.

$$\binom{n}{y} = \frac{\text{The product of the } y \text{ integers counting down from } n}{\text{The product of the } y \text{ integers counting up from } 1}$$

Derivation of the Binomial Distribution Formula

Combinations

- In general, we ask, what is n choose y.
- This quantity is often symbolized with the notations $\binom{n}{y}$ or (less frequently) ${}_nC_y$.
- This can be computed as the following ratio of two products.

$$\binom{n}{y} = \frac{\text{The product of the } y \text{ integers counting down from } n}{\text{The product of the } y \text{ integers counting up from } 1}$$

- In the preceding example, this is

$$\frac{4 \times 3 \times 2}{3 \times 2 \times 1} = \frac{24}{6} = 4$$

Derivation of the Binomial Distribution Formula

Combinations

- There are several relationships involving combinations.

Derivation of the Binomial Distribution Formula

Combinations

- There are several relationships involving combinations.
- The most important one is that

$$\binom{n}{y} = \binom{n}{n-y}$$

because, for every selection of y objects, there is a corresponding (de-)selection of $n - y$ objects.

Derivation of the Binomial Distribution Formula

Combinations

- There are several relationships involving combinations.
- The most important one is that

$$\binom{n}{y} = \binom{n}{n-y}$$

because, for every selection of y objects, there is a corresponding (de-)selection of $n - y$ objects.

- So, when solving for $\binom{n}{y}$, choose $w = \min(y, n-y)$ and compute $\binom{n}{w}$.

Derivation of the Binomial Distribution Formula

Combinations

- There are several relationships involving combinations.
- The most important one is that

$$\binom{n}{y} = \binom{n}{n-y}$$

  because, for every selection of y objects, there is a corresponding (de-)selection of $n - y$ objects.

- So, when solving for $\binom{n}{y}$, choose $w = \min(y, n-y)$ and compute $\binom{n}{w}$.
- Although the preceding formula is computationally much more efficient, many textbooks prefer to present

$$\binom{n}{y} = \frac{n!}{y!(n-y)!} \tag{1}$$

  where $y!$ is the product of the integers from y to 1.

Derivation of the Binomial Distribution Formula

Combinations

- The combinations formula relates to the binomial distribution.
- Recall that we were interested in computing k, the number of different sequences of n trials that produce exactly y successes.
- This can be computed as follows. Suppose we code each sequence by listing the trials on which the "successes" occur.
- For example, the sequence BBGB can be coded as 1,2,4.
- It then becomes clear that the number of different 4-trial sequences yielding exactly 3 successes is equal to the number of ways we can select 3 trial numbers out of 4. This is, of course $\binom{4}{3}$, or, more generally, $\binom{n}{y}$. So the final binomial distribution formula is

$$p(y|n,\pi) = \Pr(Y = y|n,\pi) = \binom{n}{y}\pi^y(1-\pi)^{n-y} \tag{2}$$

- Fortunately, this is computed for us with the R function `dbinom`.

The Binomial Distribution as a Sampling Distribution

- The binomial distribution gives probabilities for the number of successes in n binomial trials.

The Binomial Distribution as a Sampling Distribution

- The binomial distribution gives probabilities for the number of successes in n binomial trials.
- However, since each number of successes y_i corresponds to exactly one *sample proportion* of successes y_i/n,we see that we also have derived, in effect, the distribution of the sample proportion p.

The Binomial Distribution as a Sampling Distribution

- The binomial distribution gives probabilities for the number of successes in n binomial trials.
- However, since each number of successes y_i corresponds to exactly one *sample proportion* of successes y_i/n,we see that we also have derived, in effect, the distribution of the sample proportion p.
- For example, we previously determined that the probability of exactly 3 boys out of 4 is roughly 0.26, and this implies that the probability of a proportion of $3/4 = .75$ is also 0.26.

Hypothesis Testing

Parameters, Statistics, Estimators, and Spaces

- A *parameter*, loosely speaking, as a numerical characteristic of a statistical population.

Hypothesis Testing

Parameters, Statistics, Estimators, and Spaces

- A *parameter*, loosely speaking, as a numerical characteristic of a statistical population.
- A *statistic* is any function of the sample.

Hypothesis Testing

Parameters, Statistics, Estimators, and Spaces

- A *parameter*, loosely speaking, as a numerical characteristic of a statistical population.
- A *statistic* is any function of the sample.
- An *estimator* of a parameter is a statistic that is used to approximate the parameter from sample data. The observed value of an estimator is an *estimate* of the parameter.

Hypothesis Testing

Parameters, Statistics, Estimators, and Spaces

- A *parameter*, loosely speaking, as a numerical characteristic of a statistical population.
- A *statistic* is any function of the sample.
- An *estimator* of a parameter is a statistic that is used to approximate the parameter from sample data. The observed value of an estimator is an *estimate* of the parameter.
- The *parameter space* is the set of all possible values of the parameter.

Hypothesis Testing

Parameters, Statistics, Estimators, and Spaces

- A *parameter*, loosely speaking, as a numerical characteristic of a statistical population.
- A *statistic* is any function of the sample.
- An *estimator* of a parameter is a statistic that is used to approximate the parameter from sample data. The observed value of an estimator is an *estimate* of the parameter.
- The *parameter space* is the set of all possible values of the parameter.
- The *sample space* is the set of all possible values of the statistic employed as an estimator of the parameter.

Hypothesis Testing

Null and Alternative Hypotheses

- A *statistical hypothesis* is a statement that specifies a region of the parameter space.

Hypothesis Testing

Null and Alternative Hypotheses

- A *statistical hypothesis* is a statement that specifies a region of the parameter space.
- A hypothesis test is a procedure that defines rules for deciding, on the basis of an estimate, between two or more mutually exclusive statistical hypotheses.

Hypothesis Testing

Null and Alternative Hypotheses

- A *statistical hypothesis* is a statement that specifies a region of the parameter space.
- A hypothesis test is a procedure that defines rules for deciding, on the basis of an estimate, between two or more mutually exclusive statistical hypotheses.
- Often, but not always, the hypothesis involves two mutually exclusive and exhaustive hypotheses.

Hypothesis Testing

Null and Alternative Hypotheses

- A *statistical hypothesis* is a statement that specifies a region of the parameter space.
- A hypothesis test is a procedure that defines rules for deciding, on the basis of an estimate, between two or more mutually exclusive statistical hypotheses.
- Often, but not always, the hypothesis involves two mutually exclusive and exhaustive hypotheses.
- In the classic *Reject-Support* hypothesis-testing framework, one of the hypotheses, H_1, represents the experimenter's belief (or what the experimenter is trying to demonstrate. This hypothesis is called the *alternative hypothesis*.

Hypothesis Testing

Null and Alternative Hypotheses

- A *statistical hypothesis* is a statement that specifies a region of the parameter space.
- A hypothesis test is a procedure that defines rules for deciding, on the basis of an estimate, between two or more mutually exclusive statistical hypotheses.
- Often, but not always, the hypothesis involves two mutually exclusive and exhaustive hypotheses.
- In the classic *Reject-Support* hypothesis-testing framework, one of the hypotheses, H_1, represents the experimenter's belief (or what the experimenter is trying to demonstrate. This hypothesis is called the *alternative hypothesis*.
- The statistical *null hypothesis*, H_0, is actually the opposite of what the experimenter believes, and so rejecting this hypothesis supports the experimenter's belief.

Hypothesis Testing

An Example

Example (A Hypothesis Test)

In section 4.1 RDASA3 presents an introductory example involving guessing in an ESP experiment. A subject, Rachel, attempts to guess which of 4 cards has been selected, and performs the guessing task for a sequence of 20 trials. The experimenter chooses one of the 4 cards *randomly* on each trial, and so, in the example, MWL state the null and alternative hypotheses are

$$H_0 : \pi = 0.25, \text{ and } \; H_1 : \pi > 0.25$$

How would you describe these hypotheses *substantively*? (C.P.)

Hypothesis Testing

An Example

Example (A Hypothesis Test (ctd))

One might ponder this choice of hypotheses. Clearly, if no information is being transmitted to Rachel, and the cards are truly selected independently and at random by the experimenter, then her long run probability of success, no matter what strategy she employs, is $\pi = 0.25$. However, it is possible that information is transmitted to her, but, because she has "negative ESP," she achieves a success rate lower than 0.25.

With this in mind, I prefer a pair of mutually exclusive and exhaustive hypotheses, such as

$$H_0 : \pi = 0.25, \text{ and } \; H_1 : \pi \neq 0.25$$

or

$$H_0 : \pi \leq 0.25, \text{ and } \; H_1 : \pi > 0.25$$

How would you describe these hypotheses *substantively*? (C.P.)

Hypothesis Testing

The Critical Region Approach

- MWL discuss (boxes 4.1–4.2, pages 75–76) two approaches to hypothesis testing.

Box 4.1 Steps for Testing Hypotheses Using the *p*-Value Approach

1. State the null and alternative hypotheses, H_0 and H_1.

2. Decide on the *test statistic* that will be used to assess the evidence against H_0.

3. Decide, making reasonable assumptions, what *sampling distribution* the test statistic should have if H_0 is true.

4. Decide on the *significance level* that will be used as the criterion for deciding whether or not to reject the null hypothesis. We will reject H_0 only if our result is very unlikely under the assumption that H_0 is true. The significance level (denoted by α, the Greek letter alpha) specifies exactly how unlikely the result must be.

5. Use the sampling distribution that assumes H_0 is true to find the probability of getting a value for the statistic that is at least as "extreme" as what was actually obtained in our sample of data—call this probability the *p-value*. In finding the p-value, use only the part or parts of the sampling distribution that are consistent with H_1.

6. Reject H_0 in favor of H_1 if $p \leq \alpha$. If we reject H_0, we say that our result is "statistically significant at level α." If $p > \alpha$, we say that we have failed to reject H_0 or that we have insufficient evidence to reject H_0.

Hypothesis Testing

The Critical Region Approach

- MWL discuss (boxes 4.1–4.2, pages 75–76) two approaches to hypothesis testing.
- One approach is the *p-value* approach, described in Box 4.1.

Box 4.1 Steps for Testing Hypotheses Using the *p*-Value Approach

1. State the null and alternative hypotheses, H_0 and H_1.
2. Decide on the *test statistic* that will be used to assess the evidence against H_0.
3. Decide, making reasonable assumptions, what *sampling distribution* the test statistic should have if H_0 is true.
4. Decide on the *significance level* that will be used as the criterion for deciding whether or not to reject the null hypothesis. We will reject H_0 only if our result is very unlikely under the assumption that H_0 is true. The significance level (denoted by α, the Greek letter alpha) specifies exactly how unlikely the result must be.
5. Use the sampling distribution that assumes H_0 is true to find the probability of getting a value for the statistic that is at least as "extreme" as what was actually obtained in our sample of data—call this probability the *p-value*. In finding the *p*-value, use only the part or parts of the sampling distribution that are consistent with H_1.
6. Reject H_0 in favor of H_1 if $p \leq \alpha$. If we reject H_0, we say that our result is "statistically significant at level α." If $p > \alpha$, we say that we have failed to reject H_0 or that we have insufficient evidence to reject H_0.

Hypothesis Testing

The *p*-Value Approach

- Let's work the problem in terms of the null and alternative hypotheses stated by MWL, namely

$$H_0 : \pi = 0.25, \text{ and } \; H_1 : \pi > 0.25$$

Hypothesis Testing

The p-Value Approach

- Let's work the problem in terms of the null and alternative hypotheses stated by MWL, namely

$$H_0 : \pi = 0.25, \text{ and } \ H_1 : \pi > 0.25$$

- Let's assume our test statistic is Y, the number of correct responses.

Hypothesis Testing

The *p*-Value Approach

- Let's work the problem in terms of the null and alternative hypotheses stated by MWL, namely

$$H_0 : \pi = 0.25, \text{ and } \; H_1 : \pi > 0.25$$

- Let's assume our test statistic is Y, the number of correct responses.
- Furthermore, assume that the significance level is $\alpha = .05$.

Hypothesis Testing

The *p*-Value Approach

- Let's work the problem in terms of the null and alternative hypotheses stated by MWL, namely

$$H_0 : \pi = 0.25, \text{ and } \quad H_1 : \pi > 0.25$$

- Let's assume our test statistic is Y, the number of correct responses.
- Furthermore, assume that the significance level is $\alpha = .05$.
- We've already decided that, under H_0, a reasonable assumption is that trials are independent and random, and that $\pi = .25$, and so it is implied that Y has a distribution that is $B(20, 0.25)$, i.e, binomial with parameters $n = 20$ and $\pi = 0.25$.

Hypothesis Testing

The *p*-Value Approach

- Let's work the problem in terms of the null and alternative hypotheses stated by MWL, namely

$$H_0 : \pi = 0.25, \text{ and } \; H_1 : \pi > 0.25$$

- Let's assume our test statistic is Y, the number of correct responses.
- Furthermore, assume that the significance level is $\alpha = .05$.
- We've already decided that, under H_0, a reasonable assumption is that trials are independent and random, and that $\pi = .25$, and so it is implied that Y has a distribution that is $B(20, 0.25)$, i.e, binomial with parameters $n = 20$ and $\pi = 0.25$.
- The *p*-value of the observed result y is the probability of obtaining a result as extreme as y *and be consistent with* H_1. To be consistent with H_1, y needs to be large.

Hypothesis Testing

The *p*-Value Approach

- Let's work the problem in terms of the null and alternative hypotheses stated by MWL, namely

$$H_0 : \pi = 0.25, \text{ and } \; H_1 : \pi > 0.25$$

- Let's assume our test statistic is Y, the number of correct responses.
- Furthermore, assume that the significance level is $\alpha = .05$.
- We've already decided that, under H_0, a reasonable assumption is that trials are independent and random, and that $\pi = .25$, and so it is implied that Y has a distribution that is $B(20, 0.25)$, i.e, binomial with parameters $n = 20$ and $\pi = 0.25$.
- The *p*-value of the observed result y is the probability of obtaining a result as extreme as y *and be consistent with* H_1. To be consistent with H_1, y needs to be large.
- Therefore, we use the binomial distribution calculator to compute the probability of obtaining $Y \geq y$ if the distribution is $B(20, 0.25)$.

Hypothesis Testing

The *p*-Value Approach

- Let's work the problem in terms of the null and alternative hypotheses stated by MWL, namely

$$H_0 : \pi = 0.25, \text{ and } \quad H_1 : \pi > 0.25$$

- Let's assume our test statistic is Y, the number of correct responses.
- Furthermore, assume that the significance level is $\alpha = .05$.
- We've already decided that, under H_0, a reasonable assumption is that trials are independent and random, and that $\pi = .25$, and so it is implied that Y has a distribution that is $B(20, 0.25)$, i.e, binomial with parameters $n = 20$ and $\pi = 0.25$.
- The *p*-value of the observed result y is the probability of obtaining a result as extreme as y *and be consistent with* H_1. To be consistent with H_1, y needs to be large.
- Therefore, we use the binomial distribution calculator to compute the probability of obtaining $Y \geq y$ if the distribution is $B(20, 0.25)$.
- If this $p - value$ is less than or equal α, then we say that our result is "significant at the α level."

Hypothesis Testing

The p-Value Approach

- Let's see how that works. We need to compute the total probability of obtaining a result as extreme or more than the obtained value.
- That's *really* easy to do in R, because its probability functions are vectorized, and will operate simultaneously on a range of values.
- Suppose Rachel answers 9 out of 20 correct. We compute

```
> options(scipen = 9, digits = 4)
> sum(dbinom(9:20, 20, 0.25))

[1] 0.04093
```

- Since the p-value of 0.0409 is less than 0.05, we reject the null hypothesis "at the .05 significance level."
- Note — some people would say the result is "significant beyond the .05 level."
- Note also that, because the binomial distribution is discrete, only $n + 1$ p-values are possible.

Hypothesis Testing

The Critical (Rejection) Region Approach

- With the Critical Region approach, we specify, in advance, which values of the test statistic will cause us to reject the statistical null hypothesis.
- To have a "significance level" (α) of 0.05, we must control the probability of incorrectly rejecting a true H_0 *at or below .05*.
- When the test statistic distribution is discrete, it is usually impossible to control the probability of an incorrect rejection at exactly 0.05.

Hypothesis Testing

The Critical (Rejection) Region Approach

- So, in practice, what we do in the discrete case
  1. Start at the most extreme possible value ($y = n$ in this case) in the direction of H_1.
  2. Start adding up the $p(y)$ values, moving in from the end.
  3. Stop as soon as the current sum of the $p(y)$ values exceeds α. This means that the preceding y value demarcates the critical region. Values of the statistic at or above that value are in the rejection region.
  4. An easy way to do this is to use the `full.binomial` function, and look in the column labeled `Pr(Y > y)`. Find the largest value in that column that is still below .05. Then, choose the value of y immediately above that to demarcate the rejection region.
- To see if you are catching on, answer the following. What would be the critical value of y if a significance level of 0.01 is desired? If that value of y is used, what is the true probability of incorrectly rejecting a true H_0?

Hypothesis Testing

Null and Alternative Hypotheses

- In Psychology 310, we discussed in detail the 2×2 table representing the standard decision possibilities, and their probabilities that hold when the null and alternative hypotheses and the decision regions partition the sample space into mutually exclusive and exhaustive regions.

	State of the World	
Decision	H_0 True	H_0 False
Accept H_0	Correct Acceptance $(1-\alpha)$	Type II Error (β)
Reject H_0	Type I Error (α)	Correct Rejection $(1-\beta)$

One-Tailed vs. Two-Tailed Tests

- The significance test we discussed in the preceding section was designed in a situation where only one rejection region was required. Such a test is referred to as *one-tailed* or *one-sided*.

One-Tailed vs. Two-Tailed Tests

- The significance test we discussed in the preceding section was designed in a situation where only one rejection region was required. Such a test is referred to as *one-tailed* or *one-sided*.
- However, many traditional significance tests in the social sciences and education involve two rejection regions, and are therefore referred to as *two-tailed* or *two-sided* tests.

One-Tailed vs. Two-Tailed Tests

- The significance test we discussed in the preceding section was designed in a situation where only one rejection region was required. Such a test is referred to as *one-tailed* or *one-sided*.
- However, many traditional significance tests in the social sciences and education involve two rejection regions, and are therefore referred to as *two-tailed* or *two-sided* tests.
- As an example, suppose you flip a fair coin 20 times to see if it is not "fair." In this case, we operationalize the notion of fairness in the null hypothesis as

$$H_0 : \pi = 0.50$$

One-Tailed vs. Two-Tailed Tests

- The significance test we discussed in the preceding section was designed in a situation where only one rejection region was required. Such a test is referred to as *one-tailed* or *one-sided*.
- However, many traditional significance tests in the social sciences and education involve two rejection regions, and are therefore referred to as *two-tailed* or *two-sided* tests.
- As an example, suppose you flip a fair coin 20 times to see if it is not "fair." In this case, we operationalize the notion of fairness in the null hypothesis as

$$H_0 : \pi = 0.50$$

- Note that the coin is unfair if π is any value other than 0.50, so we state the alternative hypothesis as

$$H_1 : \pi \neq 0.50$$

One-Tailed vs. Two-Tailed Tests

- In this situation, values of y either much lower than 10 (out of 20) or much higher than 10 can be cause to reject H_0. So how do we handle this situation to produce a significance level (α) of 0.05?

One-Tailed vs. Two-Tailed Tests

- In this situation, values of y either much lower than 10 (out of 20) or much higher than 10 can be cause to reject H_0. So how do we handle this situation to produce a significance level (α) of 0.05?
- In this case, we start counting in from both sides (up from 0, down from 20)

One-Tailed vs. Two-Tailed Tests

- In this situation, values of y either much lower than 10 (out of 20) or much higher than 10 can be cause to reject H_0. So how do we handle this situation to produce a significance level (α) of 0.05?
- In this case, we start counting in from both sides (up from 0, down from 20)
  1. The total probability of rejecting a true H_0 is as close to 0.05 as possible without exceeding 0.05.

One-Tailed vs. Two-Tailed Tests

- In this situation, values of y either much lower than 10 (out of 20) or much higher than 10 can be cause to reject H_0. So how do we handle this situation to produce a significance level (α) of 0.05?
- In this case, we start counting in from both sides (up from 0, down from 20)
  1. The total probability of rejecting a true H_0 is as close to 0.05 as possible without exceeding 0.05.
  2. The probabilities in the two rejection regions are as close to each other as possible. (Note that in this case, the binomial distribution is perfectly symmetric and this is relatively easy to do.)

One-Tailed vs. Two-Tailed Tests

- We generate the B(20,0.50) distribution.

```
> full.binomial(20, 0.5)

  y    Pr(Y = y)   Pr(Y <= y)    Pr(Y > y)
  0 0.0000009537 0.0000009537 0.9999990463
  1 0.0000190735 0.0000200272 0.9999799728
  2 0.0001811981 0.0002012253 0.9997987747
  3 0.0010871887 0.0012884140 0.9987115860
  4 0.0046205521 0.0059089661 0.9940910339
  5 0.0147857666 0.0206947327 0.9793052673
  6 0.0369644165 0.0576591492 0.9423408508
  7 0.0739288330 0.1315879822 0.8684120178
  8 0.1201343536 0.2517223358 0.7482776642
  9 0.1601791382 0.4119014740 0.5880985260
 10 0.1761970520 0.5880985260 0.4119014740
 11 0.1601791382 0.7482776642 0.2517223358
 12 0.1201343536 0.8684120178 0.1315879822
 13 0.0739288330 0.9423408508 0.0576591492
 14 0.0369644165 0.9793052673 0.0206947327
 15 0.0147857666 0.9940910339 0.0059089661
 16 0.0046205521 0.9987115860 0.0012884140
 17 0.0010871887 0.9997987747 0.0002012253
 18 0.0001811981 0.9999799728 0.0000200272
 19 0.0000190735 0.9999990463 0.0000009537
 20 0.0000009537 1.0000000000 0.0000000000
```

One-Tailed vs. Two-Tailed Tests

- We start working up from the bottom, looking for a cumulative probability that is close to $\alpha/2 = 0.025$ without exceeding it. We see that a lower rejection region of $y \leq 5$ has a total probability of 0.0207.
- Careful examination of the upper end of the distribution shows that an upper rejection region of $y \geq 15$ will also have a total probability of 0.0207.
- So with these two rejection regions, the total probability is 0.0414.
- But — what about the p-value approach?
- The tradition there is to compute the p-value of an observation as if the test were one-sided (using whichever rejection region is closer to the observed value of y, and then double it.
- So, if a value of 7 is observed, you compute the p-value as

```
> 2 * sum(dbinom(0:7, 20, 0.5))

[1] 0.2632
```

- Since this value is higher than 0.05, H_0 cannot be rejected at the 0.05 level.

The Power of a Statistical Test

- The *power* of a statistical test for a state of the world in which H_0 is false is defined as the probability of rejecting H_0 under that state of the world.

Box 4.3 Steps in Computing the Power of a Test

1. Determine the theoretical sampling distribution of Y assuming H_0 to be true.
2. Determine the rejection region.
3. Assume that the null hypothesis is incorrect and that some specific alternative hypothesis, H_A, is correct.
4. Compute the probability of a result in the rejection region using the sampling distribution specified by the alternative hypothesis. The resulting value is the conditional probability of observing an outcome in the rejection region given that H_A is true. This is the *power* of the test.

The Power of a Statistical Test

- The *power* of a statistical test for a state of the world in which H_0 is false is defined as the probability of rejecting H_0 under that state of the world.
- MWL summarize the general approach to power computation in Box 4.3 of RDASA3.

Box 4.3 Steps in Computing the Power of a Test

1. Determine the theoretical sampling distribution of Y assuming H_0 to be true.
2. Determine the rejection region.
3. Assume that the null hypothesis is incorrect and that some specific alternative hypothesis, H_A, is correct.
4. Compute the probability of a result in the rejection region using the sampling distribution specified by the alternative hypothesis. The resulting value is the conditional probability of observing an outcome in the rejection region given that H_A is true. This is the *power* of the test.

Power Calculation

An Example

Example (Power Calculation)

Suppose we are testing $H_0 : \pi = 0.50$ with $n = 20$ and $\alpha = 0.05$, with resulting dual rejection regions of $0 \leq Y \leq 5$ and $15 \leq Y \leq 20$.

What is the statistical power if the true state of the world is that $\pi = .80$?

Solution. We use R to compute the probability of a rejection

```
> sum(dbinom(0:5, 20, 0.8)) + sum(dbinom(15:20, 20, 0.8))
[1] 0.8042
```

In this case, power is 0.8042. The fact that the null hypothesis is false by a large amount is enough to offset the very small sample size of $n = 20$.

Factors Affecting Power

A General Perspective

- All other things being equal, there are several factors that affect statistical power:

Factors Affecting Power

A General Perspective

- All other things being equal, there are several factors that affect statistical power:
    1. *The amount by which the null hypothesis is false.* In Reject-Support testing, this is often referred to as the "effect size." The larger the effect size, the larger the power.

Factors Affecting Power

A General Perspective

- All other things being equal, there are several factors that affect statistical power:

  1. *The amount by which the null hypothesis is false.* In Reject-Support testing, this is often referred to as the "effect size." The larger the effect size, the larger the power.
  2. *Sample size*. The larger the sample size, the larger the power.

Factors Affecting Power

A General Perspective

- All other things being equal, there are several factors that affect statistical power:

  1. *The amount by which the null hypothesis is false.* In Reject-Support testing, this is often referred to as the "effect size." The larger the effect size, the larger the power.
  2. *Sample size.* The larger the sample size, the larger the power.
  3. *Significance level.* The larger ("more liberal") the α, the larger the power.

Factors Affecting Power

A General Perspective

- All other things being equal, there are several factors that affect statistical power:

  1. *The amount by which the null hypothesis is false.* In Reject-Support testing, this is often referred to as the "effect size." The larger the effect size, the larger the power.
  2. *Sample size.* The larger the sample size, the larger the power.
  3. *Significance level.* The larger ("more liberal") the α, the larger the power.
  4. *Number of tails.* A one-tailed hypothesis, provided the directionality is correct, puts a larger rejection region on the side of the true state of the world (for a given α), thereby increasing power.

Factors Affecting Power

A General Perspective

- All other things being equal, there are several factors that affect statistical power:

  1. *The amount by which the null hypothesis is false.* In Reject-Support testing, this is often referred to as the "effect size." The larger the effect size, the larger the power.
  2. *Sample size.* The larger the sample size, the larger the power.
  3. *Significance level.* The larger ("more liberal") the α, the larger the power.
  4. *Number of tails.* A one-tailed hypothesis, provided the directionality is correct, puts a larger rejection region on the side of the true state of the world (for a given α), thereby increasing power.
  5. *Reducing error variance.* Error is like noise in an experimental design, and the experimental effect is like a signal. With careful, efficient experimental design, aspects of a study that might be lumped in with "error" get partialled out as a planned source of variation. This reduction of noise makes it easier to "receive the signal," and results in higher statistical power for the test of interest.