Unit 7
Inference for Proportions

Distribution of a Sample Proportion

Recall from Unit 3, if $\hat{p}$ is the sample proportion of successes in an SRS drawn from a large population having population proportion p of successes, then the mean and standard deviation of $\hat{p}$ are

$$\mu_{\hat{p}} = p$$

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

Distribution of a Sample Proportion

When the sample size n is high,

$$\hat{p} \dot{\sim} N\left(p, \sqrt{\frac{p(1-p)}{n}}\right) \quad \Rightarrow \quad Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \dot{\sim} N(0,1)$$

We can safely use this approximation provided that

$$np \geq 10 \quad \text{and} \quad n(1-p) \geq 10$$

and that the population is very large compared to the sample.

Inference for a Population Proportion

We now examine inference methods for the case where the parameter of interest is some population proportion p.

The methods and interpretations are the same, but since we are now dealing with proportions rather than means, the formulas are different.

Inference for a Population Proportion

In order to use the normal distribution when doing probability calculations for $\hat{p}$, we required that both $np \geq 10$ and $n(1-p) \geq 10$, and that the population was large relative to the sample.

Although we won't formally verify this for each example, these assumptions will hold for all of them, and so the use of the normal distribution in our probability calculations is justified.

Confidence Interval for a Population Proportion

We take a simple random sample of n individuals and calculate the proportion $\hat{p}$ that possess some characteristic of interest. A $100(1 - \alpha)\%$ confidence interval for the population proportion p is

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Ideally, we would like to use the true standard deviation of $\hat{p}$ in the formula, but we don't know p (this is the reason for doing inference!), so we estimate it by $\hat{p}$, and we estimate the standard deviation of $\hat{p}$ by the **standard error** of $\hat{p}$.

Example

In a survey of 1,000 American voters, 687 said they support stronger gun control laws. The following is a 95% confidence interval for the true proportion of all American voters who support stronger gun laws:

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.687 \pm 1.96 \sqrt{\frac{0.687(0.313)}{1000}}$$

$$= 0.687 \pm 0.029 = (0.658, 0.716)$$

Example

We interpret the confidence interval as follows:

If we repeatedly selected random samples of 1,000 Americans and constructed the confidence interval in a similar manner, then 95% of such intervals would contain the true proportion of voters who support stronger gun control laws.

Practice Question

We would like to estimate the true proportion p of Winnipeggers who have been to a Blue Bombers game. In a random sample of 500 people in the city, 300 of them say they have been to a Bombers game. What is the margin of error for a 99% confidence interval for p?

(A) 0.0156
(B) 0.0245
(C) 0.0382
(D) 0.0473
(E) 0.0564

Practice Question

We would like to estimate the true proportion p of Canadian seniors who get the flu vaccine each year. In a random sample of 750 Canadian seniors, 480 of them say they get the flu vaccine each year. A 90% confidence interval for p is calculated to be (0.611, 0.669). What is the correct interpretation of this interval?

(A) 90% of samples of 750 seniors will give proportions between 0.611 and 0.669.

(B) 90% of similarly constructed intervals will contain the sample proportion of seniors who get the flu vaccine each year.

(C) 90% of similarly constructed intervals will contain the true proportion of seniors who get the flu vaccine each year.

(D) The probability that the true proportion is between 0.611 and 0.669 is 90%.

Sample Size Determination

Suppose we would like to select a sample of individuals large enough to estimate some population proportion p to within a specified margin of error m with a given level of confidence.

$$m = z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \quad \Rightarrow \quad n = \left(\frac{z^*}{m}\right)^2 \hat{p}(1-\hat{p})$$

Sample Size Determination

But we have a problem – we are at the stage where we have not yet selected the sample, and so we don't know the value of $\hat{p}$. We will estimate the population proportion p by some value p^*. We can either use an educated guess for p^*, or we can use a conservative estimate $p^* = 0.5$, which will result in a margin of error no greater than m, regardless of the sample proportion $\hat{p}$.

Sample Size Determination

We therefore use the following formula to determine the required sample size:

$$n = \left(\frac{z^*}{m}\right)^2 p^*(1 - p^*)$$

If we believe the value of p is relatively close to 0.5 (say, between 0.3 and 0.7), we should use $p^* = 0.5$.

Example

Suppose we would like to take a sample large enough to estimate the true proportion of all consumers who prefer Pepsi over Coke to within 3% with 95% confidence. We require a sample of size

$$n = \left(\frac{z^*}{m}\right)^2 p^*(1-p^*) = \left(\frac{1.96}{0.03}\right)^2 (0.5)(0.5) = 1067.1 \approx 1068$$

R Code

```
> nsize(0.03, p = 0.5, conf.level = 0.95, type = "pi")

The required sample size (n) to estimate the population
proportion of successes with a 0.95 confidence interval
so that the margin of error is no more than 0.03 is 1068 .
```

Example

Using the conservative estimate $p^* = 0.5$ in this case does not result in a much higher sample size than if we had used $p^* = 0.3$ or $p^* = 0.7$, for which the required sample size would be $n = 897$.

However, if we know the sample proportion will be quite far from 0.5, we may want to use an educated guess for p^*.

Example

Suppose we would like to estimate the true proportion of all Canadians who are military Veterans to within 0.005 with 90% confidence. If we use $p^* = 0.5$, we require a sample of size

$$n = \left(\frac{z^*}{m}\right)^2 p^*(1-p^*) = \left(\frac{1.645}{0.005}\right)^2 (0.5)(0.5) \approx 27,061$$

Example

But we know the proportion of people who are Veterans is much lower than 0.5. Suppose we believe the true proportion is somewhere close to 0.02. Using $p^* = 0.02$, we require a sample of size

$$n = \left(\frac{z^*}{m}\right)^2 p^*(1-p^*) = \left(\frac{1.645}{0.005}\right)^2 (0.02)(0.98) \approx 2,122$$

R Code

```
> nsize(0.005, p = 0.02, conf.level = 0.90, type = "pi")

The required sample size (n) to estimate the population
proportion of successes with a 0.9 confidence interval
so that the margin of error is no more than 0.005 is 2122 .
```

Example

We see that we would be taking much too large a sample if we used $p^* = 0.5$. This would give us a confidence interval with a margin of error much smaller than what we originally wanted.

A small margin of error is good, but we decided we were happy estimating the true proportion to within 0.5%, and we see that if we use a more reasonable value of p^* (i.e., 0.02), we need to sample more than ten times fewer individuals.

Practice Question

We would like to estimate the true proportion of people who text while driving to within 0.05 with 98% confidence. What sample size is required?

(A) 475 (B) 542 (C) 368 (D) 193 (E) 246

Practice Question

A researcher calculates that, in order to estimate the true proportion of Canadians who recycle regularly to within 0.03 with 90% confidence, she requires a sample of 360 Canadians.

What sample size would be required in order to estimate the true proportion of Canadians who recycle regularly to within 0.01 with 90% confidence?

(A) 40 (B) 120 (C) 360 (D) 1,080 (E) 3,240

Practice Question

A researcher calculates that, in order to estimate the true proportion of Canadians who recycle regularly to within 0.03 with 90% confidence, she requires a sample of 360 Canadians.

The population of the United States is ten times that of Canada. What sample size would be required in order to estimate the true proportion of Americans who recycle regularly to within 0.03 with 90% confidence?

(A) 36 (B) 360 (C) 1,139 (D) 3,600 (E) 36,000

Hypothesis Tests for a Population Proportion

We can also conduct hypothesis tests for a population proportion p.

A political party would like to know if their support has increased since the last election, when they received 18% of all votes. In a random sample of 750 voters, 165 say they support the party today. We will conduct a hypothesis test to determine if the party's popular support has increased since the last election.

Example

Step 1

Let $\alpha = 0.05$.

Step 2

We are testing the hypotheses

H_0: The proportion of voters who support the party is the same as last election.

H_a: The party's support has increased since the election.

Equivalently, H_0: $p = 0.18$ vs. H_a: $p > 0.18$.

Example

Step 3

Reject H_0 if the P-value $\leq \alpha = 0.05$.

Step 4

We calculate $\hat{p} = 165/750 = 0.22$. The test statistic is

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.22 - 0.18}{\sqrt{\frac{0.18(0.82)}{750}}} = 2.85$$

Example

Notice that we are assuming that $p = p_0$ (the value of p in the null hypothesis) in the calculation of the test statistic, just as we assumed $\mu = \mu_0$ in calculating the test statistic when conducting a test for a population mean.

We always calculate the test statistic assuming H_0 is true.

Example

Step 5

The P-value is $P(Z \geq 2.85) = 1 - P(Z < 2.85)$
$= 1 - 0.9978 = 0.0022.$

P-value Interpretation: If the party's support was the same as last election (i.e., if the proportion of voters who supported the party was 0.18), the probability of observing a sample proportion at least as high as 0.22 would be 0.0022.

Example

Step 6

Since the P-value = 0.0022 < α = 0.05, we reject the null hypothesis. At the 5% level of significance, we have sufficient evidence that the party's popular support has increased since the last election.

If we had conducted the test using the critical value method, the decision rule would be to reject H_0 if $z \geq z^* = 1.645$. The conclusion would be to reject H_0, since $z = 2.85 > z^* = 1.645$.

R Code

```
> prop.test(165, 750, 0.18, alternative = "greater",
            correct = FALSE)

data:  165 out of 750, null probability 0.18
X-squared = 8.1301, df = 1, p-value = 0.002177
alternative hypothesis: true p is greater than 0.18
95 percent confidence interval:
 0.1961505 1.0000000
sample estimates:
   p
0.22
```

Note that R gives the test statistic as X-squared. To find the value of our z test statistic, take the square root of X-squared.

Example

Shortly after the introduction of the Euro coin in Belgium, newspapers published articles claiming the coin was unfair (i.e., that when the coin was flipped, Heads and Tails were not equally likely). We investigate the claim by flipping the Belgian Euro coin 500 times, and we observe 265 Heads.

First, let us calculate a 95% confidence interval for the true proportion of all flips of the coin that land on Heads.

Example

We calculate $\hat{p} = 265/500 = 0.53$, so our 95% confidence interval is

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.53 \pm 1.96 \sqrt{\frac{0.53(0.47)}{500}}$$

$$= 0.53 \pm 0.044 = (0.486, 0.574)$$

We will now conduct a hypothesis test to determine whether there is evidence that the Belgian Euro coin really is unfair.

Example

Step 1

Let $\alpha = 0.05$.

Step 2

We are testing the hypotheses

H_0: The Belgian Euro coin is fair.

H_a: The Belgian Euro coin is unfair.

Equivalently, H_0: $p = 0.5$ vs. H_a: $p \neq 0.5$

Example

Step 3

Reject H_0 if the P-value $\leq \alpha = 0.05$.

Step 4

The test statistic is

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.53 - 0.50}{\sqrt{\frac{0.50(0.50)}{500}}} = 1.34$$

Example

Step 5

The P-value is $2P(Z \geq 1.34) = 2(1 - 0.9099)$
$= 2(0.0901) = 0.1802.$

Step 6

Since the P-value $= 0.1802 > \alpha = 0.05$, we fail to reject H_0. At the 5% level of significance, we have insufficient evidence to conclude that the Belgian Euro coin is unfair.

Example

P-value Interpretation: If the Belgian Euro coin was fair (i.e., if $p = 0.5$), the probability of observing a sample proportion of Heads at least as extreme as 0.53 would be 0.1802.

If we had conducted the test using the critical value method, the decision rule would be to reject H_0 if $|z| \geq 1.96$. The conclusion would be to fail to reject H_0, since $|z| = 1.34 < z^* = 1.96$.

R Code

```
> prop.test(265, 500, 0.50, alternative = "two.sided",
            correct = FALSE)

data:  265 out of 500, null probability 0.5
X-squared = 1.8, df = 1, p-value = 0.1797
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.4861906 0.5733519
sample estimates:
   p
0.53
```

Note that R gives the test statistic as X-squared. To find the value of our z test statistic, take the square root of X-squared.

Example

Note that we previously constructed a 95% confidence interval, and then we conducted a two-sided test with a 5% level of significance.

However, in hypothesis tests for p, the confidence interval **cannot** be used to conduct the test. This is because the formulas for the confidence interval and test statistic use different versions of the formula for the standard error/standard deviation.

Proportions

We have now seen four different formulas for proportions:

1) When we know the population proportion p and we want to find probabilities for the sample proportion $\hat{p}$:

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

2) Confidence interval for p:

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Proportions

3) Sample size calculation:

$$n = \left(\frac{z^*}{m}\right)^2 p^*(1-p^*)$$

4) Hypothesis test for p – test statistic:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Practice Question

A drug company manufactures antacid that is known to be successful in providing relief for 70% of people with heartburn. The company tests a new formula to determine whether it is better than the current one. A total of 150 people with heartburn try the new antacid and 120 report feeling some relief. Conduct a hypothesis test at the 5% level of significance to determine whether the new formula is more effective than the old formula.

Practice Question

In a random sample of 1,000 U of M students, it is found that 215 of them are international students.

a) Construct a 90% confidence interval for the true proportion of international students at the U of M.

b) Conduct a hypothesis test at the 10% level of significance to determine if the true proportion of international students at the University of Manitoba differs from 0.20.