Inference for Proportions

So, now we switch gears, and look at qualitative variables. We "make" them quantitative by measuring the proportion of the sample that has the quality that we're investigating—the parameter is p (or π), the proportion of the population; and the statistic is $\hat{p}$, the proportion of the sample.

Do we need a new distribution?

In order to conduct inference, we need to know something about the distribution of $\hat{p}$. We did that earlier, but back then, we had a value for p. Will the loss of p cause us to switch to a new distribution?

What caused us to switch from z to t?

The loss of another parameter, σ_X, which prevented us from knowing $\sigma_{\bar{x}}$.

Does the loss of p cause a similar problem?

Yes. $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$, so without p we can't find $\sigma_{\hat{p}}$.

When we switched from z to t, what did we do to replace σ_X?

We swapped it with s_x. In general, s_x is larger than σ_X, so that forced us to switch distributions.

So, let's exchange $\hat{p}$ for p. Is $\hat{p}$ generally larger (or smaller) than p?

No! So there is no need to change the distribution. Whenever we don't know p, we'll use $\hat{p}$, and everything should work out fine.

So, in summary: the sample statistic $\hat{p}$ will have an approximately normal distribution with mean $\mu_{\hat{p}} = p$ and $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$, provided the sample is random, and that np and $n(1-p)$ are each at least 10 (remember to use $\hat{p}$ if you don't know p).

One Sample Procedures

Confidence Interval

A level C confidence interval for p is $\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, where z^* is the upper $\frac{1-C}{2}$ critical value from N(0,1). We definitely won't know the value of p, so we'll use $\hat{p}$.

Test of Significance

The test statistic is $z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$ (note the use of p in the denominator—we know the value of p from our null hypothesis!). The p-value is calculated in the usual way.

Examples

[1.] A Harris Poll from June 2000 reported that 79% of U.S. citizens (based on a random sample of 2000 people) thought that elected officials should be subjected to random drug tests. Let's construct a 90% confidence interval for the true population proportion that agree with this idea.

First, the requirements: the sample must be a random sample from the population—this is given.

Second, the distribution of $\hat{p}$ must be approximately normal—$n\hat{p}$ is 1580 and $n(1 - \hat{p})$ is 420; each of these is at least 10, so we may proceed.

90% confidence gives $z^* = 1.645$. The interval is $0.79 \pm 1.645\sqrt{\frac{0.79(0.21)}{2000}} = 0.79 \pm 0.0149 =$ (0.7750, 0.8049).

I am 90% confident that the true proportion of U.S. citizens that agree with this statement is between 77.5% and 80.5%.

[2.] A casino operator is checking to see if the Roulette wheel is working correctly. A sample of 200 spins reveals that 93 of them resulted in red. Theoretically, the probability of a spin resulting in red is $\frac{18}{38} \approx 0.474$. Is there evidence at the 10% level that this wheel is not working properly?

The variables are:

p = population proportion of spins resulting in red

$\hat{p}$ = sample proportion of spins resulting in red

H_0: $p = 0.474$ (the wheel is working properly)

H_a: $p \neq 0.474$ (the wheel is not working properly)

This calls for a one-sample z test for proportions. There are some requirements to conduct this test.

(a) The sample must be a random sample from the population. This was not stated; it must be assumed.

(b) np and $n(1 - p)$ must each be at least 10. $np = 94.74$ and $n(1 - p) = 105.26$.

I've been told to use $\alpha = 0.10$.

$\hat{p} = \frac{93}{200} = 0.465$.

$z = \frac{93/200 - 18/38}{\sqrt{\frac{18/38\left(20/38\right)}{200}}} = -0.2460$. $2P(\hat{p} < \frac{93}{200}) = 2P(Z < -0.2460) = 0.8057$.

If $\frac{18}{38}$ spins result in red, then I can expect to find fewer than 0.465, or more than 0.482, in about 80.57% of samples. This happens often enough to attribute to chance at the 10% level; it is not significant, and I fail to reject the null hypothesis.

It appears that the wheel is working correctly.

Two Sample Procedures

We can combine independent samples just like we did when we learned the *t* procedures.

The parameter is $p_1 - p_2$, and the statistic is $\hat{p}_1 - \hat{p}_2$. The distribution of $\hat{p}_1 - \hat{p}_2$ is approximately normal with mean $\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2$ and standard deviation $\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$, provided the samples are random, and that each of $n_1 p_1$, $n_1(1 - p_1)$, $n_2 p_2$, and $n_2(1 - p_2)$ are at least 5.

Confidence Interval

A level *C* confidence interval for $p_1 - p_2$ is $(\hat{p}_1 - \hat{p}_2) \pm z^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$, where z^* is the upper $\frac{1-C}{2}$ critical value from N(0,1).

Test of Significance

We begin with the null hypothesis $p_1 = p_2$ (or $p_1 - p_2 = 0$). Notice, though, that this hypothesis does not tell us anything about the common value of p_1 and p_2. So, we cannot use them in the calculation of $\sigma_{\hat{p}_1 - \hat{p}_2}$. However, we do have $\hat{p}_1$ and $\hat{p}_2$, which we could use in their stead. That would give us $z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$, which is the **unpooled two sample z statistic**…which is fine; but by our null hypothesis, $p_1 = p_2$. Shouldn't we be using the same value for p_1 and p_2?

We have $\hat{p}_1$ and $\hat{p}_2$…can we use them to estimate the common value of p_1 and p_2?

Sure! Just average them (correctly).

Set $\hat{p} = \frac{\text{total number of successes}}{\text{total sample size}}$ (the pooled sample proportion), which gives us

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}}, \text{ or } z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{(\hat{p})(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

This is the **pooled two sample *z* statistic** for proportions. This is the test that the TI calculators will perform (without asking your opinion!). Note, however, that this is only appropriate if the sample sizes are approximately equal—otherwise, use the unpooled version.

Be careful to use the pooled sample proportion when checking requirements. Other than that, the remainder of the test is the same as always.

Examples

[3.] An experiment sought to determine if folic acid, taken during pregnancy, can reduce the occurrence of major birth defects. 2701 women took folic acid, and 35 of their children had major birth defects. Another 2052 women did not receive folic acid, and 47 of their children had major birth defects. Is there evidence that folic acid reduces the occurrence of major birth defects?

The variables are p_f = proportion mothers taking folic acid whose children have birth defects
p_c = proportion mothers NOT taking folic acid whose children have birth defects
H_0: $p_f = p_c$ (folic acid has no effect on birth defects)
H_a: $p_f < p_c$ (folic acid reduces birth defects)
This calls for a two-sample z test (pooled) for proportions. There are some requirements for this test:
(a) Each sample must be a random sample from the population. As long as the others were randomly assigned to one of the treatments (folic acid, or not), this will suffice.
(b) $n_f\hat{p}$, $n_f(1 - \hat{p})$, $n_c\hat{p}$, $n_c(1 - \hat{p})$ must each be at least 5. $\hat{p} = \dfrac{35+47}{2701+2052} \approx 0.0173$. $n_f\hat{p} = 46.6$, $n_f(1 - \hat{p}) = 2654.4$, $n_c\hat{p} = 35.4$, $n_c(1 - \hat{p}) = 2016.6$.
I'll choose a 1% level of significance (I want to see a lot of evidence before I start recommending women take folic acid).

$$z = \frac{0.013 - 0.0229}{\sqrt{0.0173(0.9827)\left(\frac{1}{2701}+\frac{1}{2052}\right)}} \approx -2.6085.$$

$P(\hat{p}_f - \hat{p}_c < -0.0099) = P(Z < -2.6085) = 0.0045$.
If there is no difference in the proportions of births with major defects, then I can expect to find the proportion of birth defects to mothers who took folic acid to be at least 0.99% lower than the proportion among mothers who did not take folic acid in about 0.45% of samples. This occurs too rarely to attribute to chance at the 1% level of significance; it is significant, and I reject the null hypothesis.
It appears that taking folic acid does reduce the occurrence of serious birth defects.

[4.] A poll in 1999 asked U.S. adults whether or not they would order a vegetarian meal when at a restaurant. Of 747 men, 276 said that they would order vegetarian; of 434 women, 195 said that they would. Estimate the difference in the proportions of men and women who will order vegetarian with 90% confidence.

This calls for a two sample z interval for proportions. There are some requirements that must be met before the interval can be constructed.
(a) The samples must be random. This is not given; it will be assumed that the samples of men and women are random samples of all U.S. men and women.
(b) $n_m\hat{p}_m$, $n_m(1 - \hat{p}_m)$, $n_f\hat{p}_f$, $n_f(1 - \hat{p}_f)$ must all be at least 5. $n_m\hat{p}_m = 276$, $n_m(1 - \hat{p}_m) = 471$, $n_f\hat{p}_f = 195$, $n_f(1 - \hat{p}_f) = 239$; this requirement is met.

90% confidence gives $z^* = 1.645$. The interval is $\left(\hat{p}_m - \hat{p}_f\right) \pm z^* \sqrt{\frac{\hat{p}_m\left(1-\hat{p}_m\right)}{n_m} + \frac{\hat{p}_f\left(1-\hat{p}_f\right)}{n_f}}$ = $\left(0.3695 - 0.4493\right) \pm 1.645\sqrt{\frac{0.3695(0.6305)}{747} + \frac{0.4493(0.5507)}{434}}$ = -0.0798 ± 1.645·0.0297 = (-0.1287, -0.0310).

I am 90% confident that the true difference in proportions (male – female) is between -12.87% and -3.1%. In other words, I am 90% confident that the proportion of women who will order vegetarian is between 3.1% and 12.87% higher than the proportion of men.