

BIOS 600: Principles of Statistical Inference

Inference on Proportions

Fall 2012

Reading

- ▶ Pagano and Gauvreau, Chapter 14

Binomial Distribution

Recall the binomial distribution. If X is binomial with parameters n and π , then

$$Pr(X = x) = \binom{n}{x} \pi^x (1 - \pi)^{n-x}$$

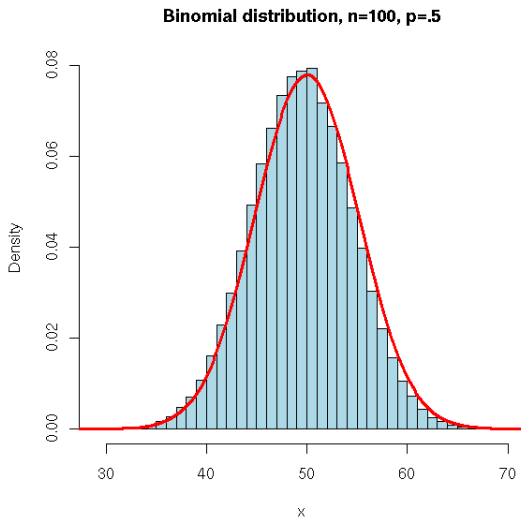
with mean $n\pi$ and standard deviation $\sqrt{n\pi(1 - \pi)}$.

Estimating a Proportion



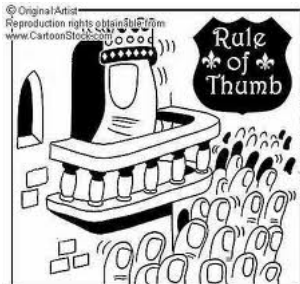
- ▶ Suppose we conduct a binomial experiment (coin toss) n times
- ▶ Each time let $h_i = 1$ if we get a head and $h_i = 0$ if we get a tail
- ▶ Then the number of successes $x = \sum_{i=1}^n h_i$
- ▶ $\hat{\pi} = \frac{x}{n} = \frac{1}{n} \sum_{i=1}^n h_i$ (note that this is just a sample mean)
- ▶ The next slide shows the values $\hat{\pi}$ we get from a series of $n = 100$ experiments

Binomial Distribution (Normal Superimposed)



Normal Approximation to the Binomial Distribution

For large n , the binomial distribution can be cumbersome (think about the combination term). Fortunately, for large n , the normal distribution provides a good approximation to the binomial distribution. In smaller samples, the approximation is better when π is close to 0.5. It is given by $Z = \frac{X - n\pi}{\sqrt{n\pi(1-\pi)}} \approx N(0, 1)$.



n is “large enough” for the approximation if both $n\pi$ and $n(1 - \pi)$ are greater than or equal to 5 (though some folks say 10)

Normal Approximation to the Binomial Distribution in Action

Normal Approximation Applet

Sampling Distribution of a Proportion

Suppose we take repeated samples of size n from the population, and obtain estimates of the population proportion $\hat{\pi}_1, \hat{\pi}_2$, etc. According to the Central Limit Theorem (P&G Chapter 8), the distribution of the sample proportions has the following properties

- ▶ Its mean is the population mean π
- ▶ Its standard deviation is given by $\frac{\sqrt{n\pi(1-\pi)}}{n} = \sqrt{\frac{\pi(1-\pi)}{n}}$
- ▶ Its shape is approximately normal for n “large enough”

Then we know

$$Z = \frac{\hat{\pi} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$$

is approximately $N(0, 1)$.

Using the Normal Approximation to Get CI's

We can use this result to get confidence intervals of the form estimate \pm margin of error. For example, a $100(1 - \alpha)\%$ CI would be given by

$$\hat{\pi} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\pi(1 - \pi)}{n}}.$$

Note that π is not known, so we usually estimate it with $\hat{\pi}$ and use

$$\hat{\pi} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{n}}$$

to obtain an approximate CI.

Example: Viagra Marketing

The drug Viagra became available in the U.S. in May, 1998, in the wake of an advertising campaign that was unprecedented in scope and intensity. A Gallup poll found that by the end of the first week in May, 643 out of a random sample of 1,005 adults were aware that Viagra was an impotency medication (based on “Viagra A Popular Hit,” a Gallup poll analysis by Lydia Saad, May 1998).

Let's estimate the proportion π of all adults in the U.S. who by the end of the first week of May 1998 were already aware of Viagra and its purpose by setting up a 95% confidence interval for π .

Example: Viagra Marketing

We first need to calculate the sample proportion $\hat{\pi}$. Out of 1,005 sampled adults, 643 knew what Viagra is used for, so

$$\hat{\pi} = \frac{643}{1005} = .64.$$

Is this good marketing?

- ▶ 59% of Americans know that humans and dinosaurs did not live at the same time

Example: Viagra Marketing

We first need to calculate the sample proportion $\hat{\pi}$. Out of 1,005 sampled adults, 643 knew what Viagra is used for, so

$$\hat{\pi} = \frac{643}{1005} = .64.$$

Is this good marketing?

- ▶ 59% of Americans know that humans and dinosaurs did not live at the same time
- ▶ 57% of Americans know that George Washington led the Continental Army

Example: Viagra Marketing

We first need to calculate the sample proportion $\hat{\pi}$. Out of 1,005 sampled adults, 643 knew what Viagra is used for, so

$$\hat{\pi} = \frac{643}{1005} = .64.$$

Is this good marketing?

- ▶ 59% of Americans know that humans and dinosaurs did not live at the same time
- ▶ 57% of Americans know that George Washington led the Continental Army
- ▶ 53% of Americans know how long it takes for Earth to revolve around the Sun

Example: Viagra Marketing

We first need to calculate the sample proportion $\hat{\pi}$. Out of 1,005 sampled adults, 643 knew what Viagra is used for, so

$$\hat{\pi} = \frac{643}{1005} = .64.$$

Is this good marketing?

- ▶ 59% of Americans know that humans and dinosaurs did not live at the same time
- ▶ 57% of Americans know that George Washington led the Continental Army
- ▶ 53% of Americans know how long it takes for Earth to revolve around the Sun
- ▶ 46% of Americans know that electrons are *smaller* than atoms

Example: Viagra Marketing

OK to use normal approximation? Sure, as
 $n\hat{\pi} = 1005(0.64) = 643.2$ and $n(1 - \hat{\pi}) = 361.8$, both $\gg 5$.

So our 95% CI is given by

$$\hat{\pi} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{n}}$$

$$0.64 \pm 1.96 \sqrt{\frac{0.64(1 - 0.64)}{1015}}$$

or (0.61, 0.67).

Hypothesis Testing

Suppose we have the null hypothesis $H_0 : \pi = \pi_0$ versus the alternative $H_A : \pi \neq \pi_0$. How do we test this hypothesis?

We draw a random sample from the underlying population of interest, estimate π using $\hat{\pi}$, and find the probability of getting a sample proportion as extreme, or more extreme than, $\hat{\pi}$ if the true population proportion is π . We do this by calculating the z-statistic

$$z = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

and comparing it to a $N(0, 1)$ distribution.

If the “rule of thumb” is not satisfied for the normal approximation, we can use the exact binomial distribution to conduct the test as we did before.

Hypothesis Testing

Why the different standard error estimate?

In the CI, we estimate the population proportion in the standard error by $\hat{\pi}$, but in the hypothesis test we assume the null is true and use π_0 . Because we use different standard error estimates for the CI and the test, the CI and hypothesis test results may not always agree as they did when we were estimating means.

You Try It: Legalizing Marijuana

Earlier in 2012, a poll of 1000 Americans found that 56% were in favor of legalizing marijuana use. Do the results provide evidence that the majority of Americans feel that marijuana use should be legalized?

Sample Size Estimation

Suppose we wish to test $H_0 : \pi = \pi_0$ versus the alternative $H_0 : \pi \neq \pi_0$. The required sample size depends on a number of factors:

- Desired type I error rate, α

$$n = \left[\frac{z_{\frac{\alpha}{2}} \sqrt{\pi_0(1 - \pi_0)} + z_{\beta} \sqrt{\pi_1(1 - \pi_1)}}{\pi_1 - \pi_0} \right]^2$$

Sample Size Estimation

Suppose we wish to test $H_0 : \pi = \pi_0$ versus the alternative $H_0 : \pi \neq \pi_0$. The required sample size depends on a number of factors:

- ▶ Desired type I error rate, α
- ▶ Desired power, $1 - \beta$

$$n = \left[\frac{z_{\frac{\alpha}{2}} \sqrt{\pi_0(1 - \pi_0)} + z_{\beta} \sqrt{\pi_1(1 - \pi_1)}}{\pi_1 - \pi_0} \right]^2$$

Sample Size Estimation

Suppose we wish to test $H_0 : \pi = \pi_0$ versus the alternative $H_0 : \pi \neq \pi_0$. The required sample size depends on a number of factors:

- ▶ Desired type I error rate, α
- ▶ Desired power, $1 - \beta$
- ▶ Smallest difference π_1 you wish to be able to detect given the specified type I and type II error rates

$$n = \left[\frac{z_{\frac{\alpha}{2}} \sqrt{\pi_0(1 - \pi_0)} + z_{\beta} \sqrt{\pi_1(1 - \pi_1)}}{\pi_1 - \pi_0} \right]^2$$

Sample Size Estimation

Suppose we wish to test $H_0 : \pi = \pi_0$ versus the alternative $H_0 : \pi \neq \pi_0$. The required sample size depends on a number of factors:

- ▶ Desired type I error rate, α
- ▶ Desired power, $1 - \beta$
- ▶ Smallest difference π_1 you wish to be able to detect given the specified type I and type II error rates
- ▶ **Stata sampsi**

$$n = \left[\frac{z_{\frac{\alpha}{2}} \sqrt{\pi_0(1 - \pi_0)} + z_{\beta} \sqrt{\pi_1(1 - \pi_1)}}{\pi_1 - \pi_0} \right]^2$$

Example: Power Calculation

Suppose we would like to estimate the proportion of vegetarians at UNC and that our null hypothesis is that this proportion is the same as that in the US (3.2%). We would like to have 80% power at a type I error rate of 5% to detect an alternative percentage double this one.

```
. sampsi 0.032 0.064, power(0.8) alpha(0.05) onesample
```

Estimated sample size for one-sample comparison of proportion
 to hypothesized value

Test Ho: $p = 0.0320$, where p is the proportion in the population

Assumptions:

```
alpha =    0.0500    (two-sided)
power =    0.8000
alternative p = 0.0640
```

Estimated required sample size:

```
n =      297
```

Example: Power Calculation

What if we just want to be able to detect an alternative percentage three times the size of our hypothesized value?

```
. sampsi 0.032 0.096, power(0.8) alpha(0.05) onsample
```

Estimated sample size for one-sample comparison of proportion
to hypothesized value

Test Ho: $p = 0.0320$, where p is the proportion in the population

Assumptions:

```
alpha = 0.0500 (two-sided)
power = 0.8000
alternative p = 0.0960
```

Estimated required sample size:

```
n = 86
```


Example: Power Calculation

What if we want to test whether the proportion of vegans here is the same as that in the US population (0.5%), using an alternative three times the size of that proportion?

```
. sampsi 0.005 0.015, power(0.8) alpha(0.05) onesample
```

Estimated sample size for one-sample comparison of proportion
to hypothesized value

Test Ho: $p = 0.0050$, where p is the proportion in the population

Assumptions:

```
alpha = 0.0500 (two-sided)
power = 0.8000
alternative p = 0.0150
```

Estimated required sample size:

```
n = 579
```

We generally need a larger sample size for these rare events.

Comparison of Two Proportions

As we did for means, we can generalize our testing procedure to accommodate comparison of two proportions. Suppose we wish to test the hypothesis that the proportions from two independent populations are identical:

$$H_0 : \pi_1 = \pi_2$$

against the alternative

$$H_A : \pi_1 \neq \pi_2.$$

Suppose we draw a sample of size n_1 from the first population and estimate $\hat{\pi}_1 = \frac{x_1}{n_1}$ and draw a sample of size n_2 from the second population and estimate $\hat{\pi}_2 = \frac{x_2}{n_2}$ where x is the number of “successes” in each sample.

Comparison of Two Proportions

As before, the p-value of this test will represent the probability of obtaining a discrepancy $\hat{\pi}_1 - \hat{\pi}_2$ as large as or larger than what we see, if the two population proportions are indeed identical.

If H_0 is indeed true, then we can estimate the overall proportion simply by combining the samples to get

$$\hat{\pi} = \frac{x_1 + x_2}{n_1 + n_2}.$$

In addition, if H_0 is true the standard error of $\hat{\pi}_1 - \hat{\pi}_2$ can be estimated by $\sqrt{\hat{\pi}(1 - \hat{\pi}) \left[\frac{1}{n_1} + \frac{1}{n_2} \right]}$.

Comparison of Two Proportions

Then our test statistic is given by

$$z = \frac{(\hat{\pi}_1 - \hat{\pi}_2) - (\pi_1 - \pi_2)}{\sqrt{\hat{\pi}(1 - \hat{\pi}) \left[\frac{1}{n_1} + \frac{1}{n_2} \right]}},$$

which is approximately $N(0, 1)$ when n_1 and n_2 are sufficiently large.



n is “large enough” for the approximation if $n\pi_1$, $n(1 - \pi_1)$, $n\pi_2$, and $n(1 - \pi_2)$ are all greater than or equal to 5

CI for Difference of Two Proportions

We generate confidence intervals as

$$\hat{\pi}_1 - \hat{\pi}_2 \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{\pi}_1(1 - \hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1 - \hat{\pi}_2)}{n_2}}$$

We do not use the standard error estimate $\sqrt{\hat{\pi}(1 - \hat{\pi}) \left[\frac{1}{n_1} + \frac{1}{n_2} \right]}$ here because that estimate assumes H_0 is true.

Example: Vegetarians

Suppose we wish to test the hypothesis that the proportions of male and female vegetarians are equal at UNC versus the alternative they are not. We have

$$H_0 : \pi_1 = \pi_2$$

and

$$H_A : \pi_1 \neq \pi_2.$$

Suppose we randomly sample 1000 UNC students, with 600 female and 400 male students. Suppose 7% of the female students are vegetarian while 3% of the males are.

Example: Vegetarians

```
. prtesti 600 .07 400 .03
```

Two-sample test of proportion

x: Number of obs = **600**
y: Number of obs = **400**

Variable	Mean	Std. Err.	z	P> z	[95% Conf. Interval]	
x	.07	.0104163			.0495844	.0904156
y	.03	.0085294			.0132828	.0467172
diff	.04	.0134629			.0136132	.0663868
	under Ho:	.0145894	2.74	0.006		

diff = prop(**x**) - prop(**y**) z = **2.7417**
Ho: diff = 0

Ha: diff < 0 Ha: diff != 0 Ha: diff > 0
Pr(Z < z) = **0.9969** Pr(|Z| < |z|) = **0.0061** Pr(Z > z) = **0.0031**

What do we conclude?