Introduction to Statistical Inference (QTM 100 Lab)

Lecture 6: Inference for a Single Proportion

Justin Eloriaga — Emory University

Fall 2024

Gameplan

Data Preliminaries

One Sample z-test

Data Preliminaries

Inference for a Single Proportion

For today...

- Inference on a dichotomous categorical variable (i.e. binary variable) using a one-sample proportion
- We will apply the *z-score* and the *binomial distribution*

Gardasil

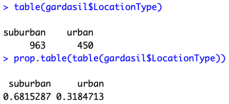
1 x 0.5 mL Single-Dose Syringe
Sterile
DIN 02437058
GARDASIL® 9
[Human Papillomavirus 9-valent
Vaccine, Recombinant]
Suspension for injection
Each 0.5 mL dose contains 30/40/60/40/20/20/20/20/20 mcg
of HPV Types 6/11/16/18/31/33/45/52/58 L1 proteins.
MERCK
Seringue à dose unique
GARDASIL® 9
[Human Papillomavirus 9-valent
Vaccine, Recombinant / vaccin
recombinant nonavalent contre
LOT
L006406

Gardasil, developed by Merck Laboratories, was licensed by the U.S. Food and Drug Administration in 2006 to vaccinate against HPV.

- The “typical” Gardasil regimen consists of a sequence of three shots, which should be completed within 12 months.
- The dataset tries to characterize young female patients who complete the anti-HPV Gardasil vaccination sequence.

One Sample z-test

Zeroing in on `LocationType`



```
> table(gardasil$LocationType)

suburban    urban 
     963      450 
> prop.table(table(gardasil$LocationType))

 suburban     urban 
0.6815287 0.3184713 
```

- First, let's explore the location where the clinic that patients visit would be in.

```
table(gardasil$LocationType)
prop.table(table(gardasil$LocationType))
```

- 963 out of 1413 participants live in an urban environment (68.2%)
- **QUESTION**: Do 70% of Gardasil patients visit suburban clinics?

Testing the Hypothesis

We can test the hypothesis below with a **one-sample z test**.

$$H_0 : p = 0.70$$

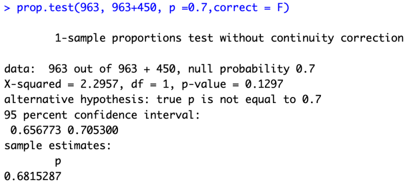
$$H_a : p \neq 0.70$$

- 2 ways to do this in `R`: (1) you can manually input the counts or use the (2) variables in the data set

Manual Input of Counts

- To manually input the counts, we include the number of subjects who experienced the event of interest, the total number of subjects studied, and the value tested in the null hypothesis, $p = 0.70$

```
prop.test(963,963+450, p=0.7,correct=F)
```



```
> prop.test(963, 963+450, p =0.7,correct = F)

	1-sample proportions test without continuity correction

data:  963 out of 963 + 450, null probability 0.7
X-squared = 2.2957, df = 1, p-value = 0.1297
alternative hypothesis: true p is not equal to 0.7
95 percent confidence interval:
 0.656773 0.705300
sample estimates:
        p 
0.6815287 
```

- 68.2% of participants use a suburban clinic (95% CI: 65.7% - 70.5%). At the $\alpha = 0.05$ level of significance, the proportion of this population that visit a suburban clinic is **not** significantly different than 0.7 ($p > 0.05$)

Manual Input of Counts

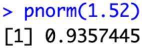
- The argument `correct=F` indicates not to use a 'continuity correction' (these results should exactly match the results you would have from hand calculation.
- The one-sample proportion test provides a χ^2 test stat with one degree of freedom. A z stat squared is the equivalent to a χ^2 test stat with one degree of freedom.

$$z = \sqrt{\chi^2} = \sqrt{2.3} = 1.52$$

Use `sqrt(2.3)` to get this

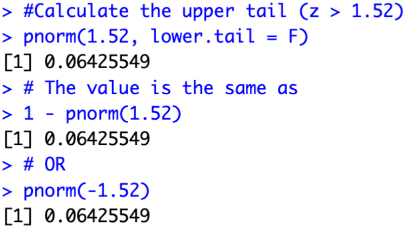
Using `pnorm()` again

- Recall that we can use a *z* test stat with the standard normal to determine *probability*. Hence, using `pnorm()` we can use the *z* score we just calculated (1.52) in the `pnorm()` function



```
> pnorm(1.52)
[1] 0.9357445
```

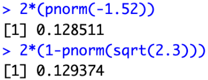
- Recall, the `pnorm()` calculates the **lower tail** by default. When estimating a test stat, we are looking at the probability of a sample proportion falling outside of that value.



```
> #Calculate the upper tail (z > 1.52)
> pnorm(1.52, lower.tail = F)
[1] 0.06425549
> # The value is the same as
> 1 - pnorm(1.52)
[1] 0.06425549
> # OR
> pnorm(-1.52)
[1] 0.06425549
```

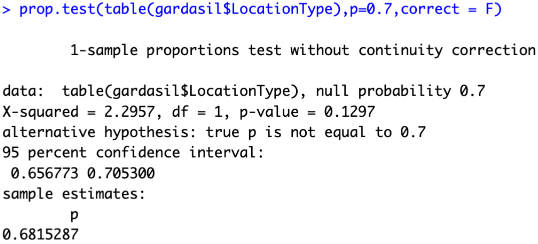
BUT WAIT, we need a two-tailed test!

- In our case, our one-sample proportion test is a **two-tailed** test. We need to double the probability of the upper tail. This will yield the correct proportion. (we will have a slight difference due to rounding errors)



```
> 2*(pnorm(-1.52))
[1] 0.128511
> 2*(1-pnorm(sqrt(2.3)))
[1] 0.129374
```

- You can also conduct the test by inputting a table of the variable of interest. However, in this case, you need to be careful with ordering factor-level variables



```
> prop.test(table(gardasil$LocationType),p=0.7,correct = F)

        1-sample proportions test without continuity correction

data:  table(gardasil$LocationType), null probability 0.7
X-squared = 2.2957, df = 1, p-value = 0.1297
alternative hypothesis: true p is not equal to 0.7
95 percent confidence interval:
 0.656773 0.705300
sample estimates:
        p
0.6815287
```

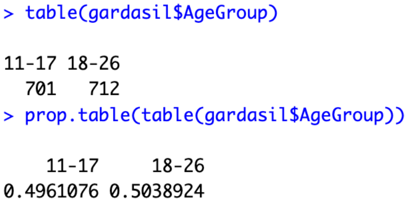
Let's Look at `AgeGroup`

Question: Do 53% of women who receive the Gardasil vaccine in the population belong to the sub-18 age group?

$$H_0 : p = 0.53$$

$$H_a : p \neq 0.53$$

Let's try to look at the data first using the `table()` and `prop.table()` commands.



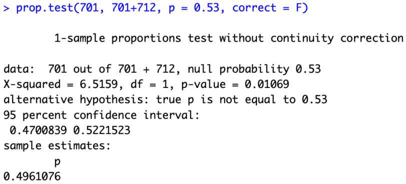
```
> table(gardasil$AgeGroup)

11-17 18-26
  701   712
> prop.table(table(gardasil$AgeGroup))

    11-17     18-26
0.4961076 0.5038924
```

Running a one-sample z test

- We see that 701 out of 1413 patients are below 18 years old or 49.6%.
- Running a one-sample z test

```
prop.test(701,701+712, p=0.53,correct=F)
```

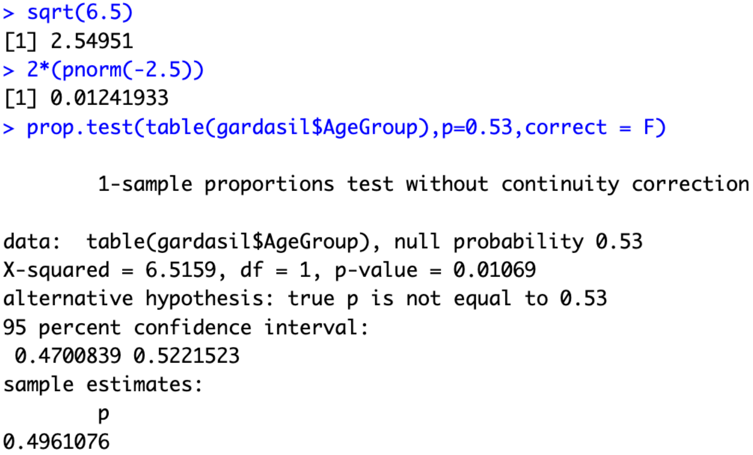
- We should get the following results



```
> prop.test(701, 701+712, p = 0.53, correct = F)

        1-sample proportions test without continuity correction

data:  701 out of 701 + 712, null probability 0.53
X-squared = 6.5159, df = 1, p-value = 0.01069
alternative hypothesis: true p is not equal to 0.53
95 percent confidence interval:
 0.4700839 0.5221523
sample estimates:
        p
0.4961076
```

- 49.2% of participants are below 18 (95% CI: 47.0% - 52.2%). At the $\alpha = 0.05$ level of significance, the proportion of this population that are below 17 is significantly different from zero ($p \approx> 0.01$)

Obtaining the z score and double checking the p-value



```
> sqrt(6.5)
[1] 2.54951
> 2*(pnorm(-2.5))
[1] 0.01241933
> prop.test(table(gardasil$AgeGroup),p=0.53,correct = F)

	1-sample proportions test without continuity correction

data:  table(gardasil$AgeGroup), null probability 0.53
X-squared = 6.5159, df = 1, p-value = 0.01069
alternative hypothesis: true p is not equal to 0.53
95 percent confidence interval:
 0.4700839 0.5221523
sample estimates:
        p 
0.4961076 
```