



# Live Classroom



Module 6 –Tests for Proportions and Logistic Regression

# Topics

---

- ▶ One Sample Tests for Proportions
- ▶ Two Sample Tests for Proportions (based on risk difference)
- ▶ Calculation of Effect Measures/Interpretation
- ▶ Logistic Regression



# One-Sample Tests for Proportions

---

- ▶ Interested in the proportion of the population that has a particular outcome (dichotomous outcome).
- ▶ Seek to make conclusions about the population parameter ( $p$  in this case) by using information from a sample.
- ▶ The sample proportion ( $\hat{p}$ ) is an estimate of the population parameter,  $p$ .
- ▶ What can we say about the sampling distribution of the sample proportion in order to see how well it estimates the population proportion?



# One-Sample Tests for Proportions

---

1. As the sample size increases, the sampling distribution of  $\hat{p}$  becomes approximately normal.
2. The mean of the sampling distribution is  $p$ .
3. The standard deviation of the sampling distribution decreases as the sample size increases. In fact, the variability (as measured by the standard deviation) of  $\hat{p}$  is given by

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



# One-Sample Tests for Proportions

---

- ▶ Tests of hypotheses about a population proportion are based on the sampling distribution of the sample proportion.
- ▶ The value of the sample proportion is compared to the value of the population proportion in the null hypothesis.
- ▶ If the sample proportion is far from the expected value of the population proportion under the null hypothesis, then we have evidence against the null hypothesis.

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

**This test is valid if the sample size is sufficiently large. Rule of thumb is  $np_0$  and  $n(1 - p_0)$  are both  $> 10$ .**



# One-Sample Tests for Proportions

---

- ▶ The test statistic measures how far  $\hat{p}$  is from the value of  $p$  (under the null hypothesis) in standard deviation units.
- ▶ Given the properties discussed regarding the sample distribution, the z-statistic is approximately normally distributed with a mean of 0 and a standard deviation of 1.
- ▶ This result allows us to quantify how far the point estimate is from the expected value of the population parameter under the null hypothesis and to make inference about the population proportion.

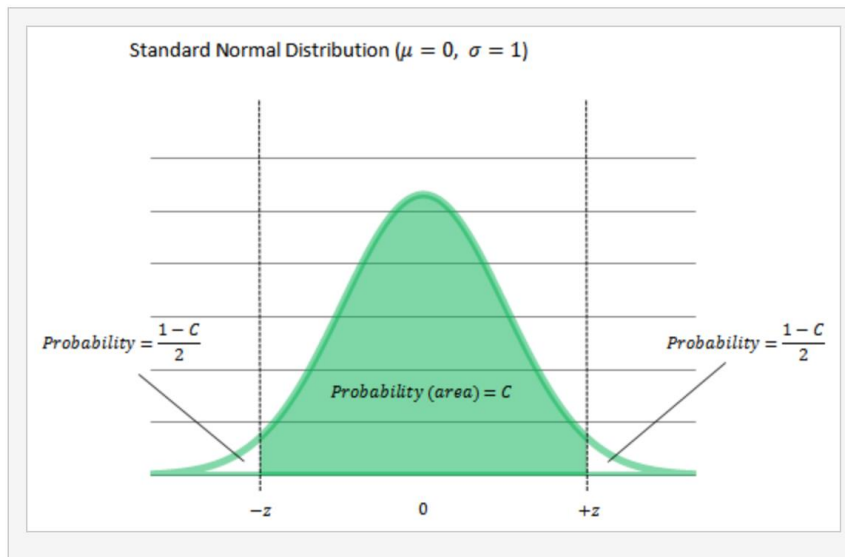


# Confidence Interval for a Proportion

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

Values of  $z$  can be determined using the standard normal distribution as depicted in the figure below. Values of the critical value,  $z$ , correspond with the points ( $-z$  and  $+z$ ) on the standard normal curve that give a central area of size  $C$  under the normal curve. Typical values of  $C$  with their associated values of  $z$  are shown in the table below.

Confidence Level, $C$	90%	95%	99%
Critical Value, $z$	1.645	1.960	2.576



# One Sample Tests for Proportions

---

## ► Use the `prop.test()` function

- `prop.test([s], [n], p = [p0],  
            alternative = [alternative],  
            conf.level = [confidence level],  
            correct = FALSE)`
- `[s]` = number of successes
- `[n]` = sample size
- `[p0]` = proportion under the null hypothesis
- `[alternative]` = "two.sided" (default), "greater" or "less"
- `[confidence level]` = 0.95, etc
- `correct = FALSE` specifies not to use a continuity correction
  - Continuity correction is a very slight alteration of the formula for the test statistic and for the confidence interval that makes the test a little more conservative (and thus the confidence interval a little wider). May be used anytime, but is especially helpful when the sample size is small





## Example – Vaccinations

---

We are interested in estimating the proportion of children in the county that are vaccinated for measles. We suspect that it may be as low as 80%. A random sample is taken of 100 children from the county. Of those sampled, only 70 were vaccinated. Formally test if the proportion of vaccinated children is different than 80%. Calculate the 95% confidence interval for the proportion of children vaccinated in the county.

► `prop.test(70, 100, p = 0.80,  
          alternative = "two.sided",  
          conf.level = 0.95, correct = FALSE)`

---



# Two-Sample Tests for Proportions

---

- Based on the null hypothesis that  $p_1 = p_2$  ( $p_1 - p_2 = 0$ )

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

Check out CI formula  
in the notes!

Inference is valid when the counts of  
successes and failures are  $\geq 5$  in both  
samples.

**Pooled sample proportion**

This quantity represents how far the difference in sample proportions is from 0 in standard deviation units under the null hypothesis where the underlying population proportion is the same in the two groups.



# Two Sample Tests for Proportions

---

## ► Use the `prop.test()` function

- `prop.test([s], [n],  
              alternative = [alternative],  
              conf.level = [confidence level],  
              correct = FALSE)`
- `[s]` = number of successes in each group:  $c(s_1, s_2)$
- `[n]` = sample size in each group:  $c(n_1, n_2)$
- `[alternative]` = "two.sided" (default), "greater" or "less"
- `[confidence level]` = 0.95, etc
- `correct = FALSE` specifies not to use a continuity correction
  - Continuity correction is a very slight alteration of the formula for the test statistic and for the confidence interval that makes the test a little more conservative (and thus the confidence interval a little wider). May be used anytime, but is especially helpful when the sample size is small

## ► Calculate Risk Difference by hand

---



# Example – Need for Social Services

---

An investigator is interested in the long term effects of preschool programs on low income children. A study was conducted where by two groups of children were followed overtime. The first group of 61 children did not attend preschool. The second group of 62 children (from similar areas and with similar backgrounds of those in the first sample) attended preschool as 3- and 4- year olds. The need for social programs as adults was the outcome of interest. Of the group who did not attend preschool, 49 of them needed social services (mainly welfare) between the ages of 18 and 30. In the preschool group, 38 required social services in the same age range.

Formally test whether or not preschool attendance reduces the need for social services in adult years at the  $\alpha = 0.05$  level of significance. Calculate the 95% confidence interval for the difference in proportions of adults requiring social services between those who did not attend preschool versus those who did attend preschool.

---



# Example – Need for Social Services

---

Population	Population description	Sample size	Count of Successes	Sample proportion
1	No Preschool	$n_1 = 61$	49	$\hat{p}_1 = \frac{49}{61} = 0.803$
2	Preschool	$n_2 = 62$	38	$\hat{p}_2 = \frac{38}{62} = 0.613$

- ▶ `prop.test(c(49,38), c(61,62), alternative = "two.sided",  
conf.level = 0.95, correct = FALSE)`
- ▶ Risk difference = 0.19 (indicating that the need for services was higher in the group that didn't attend preschool)



# Example – Need for Social Services

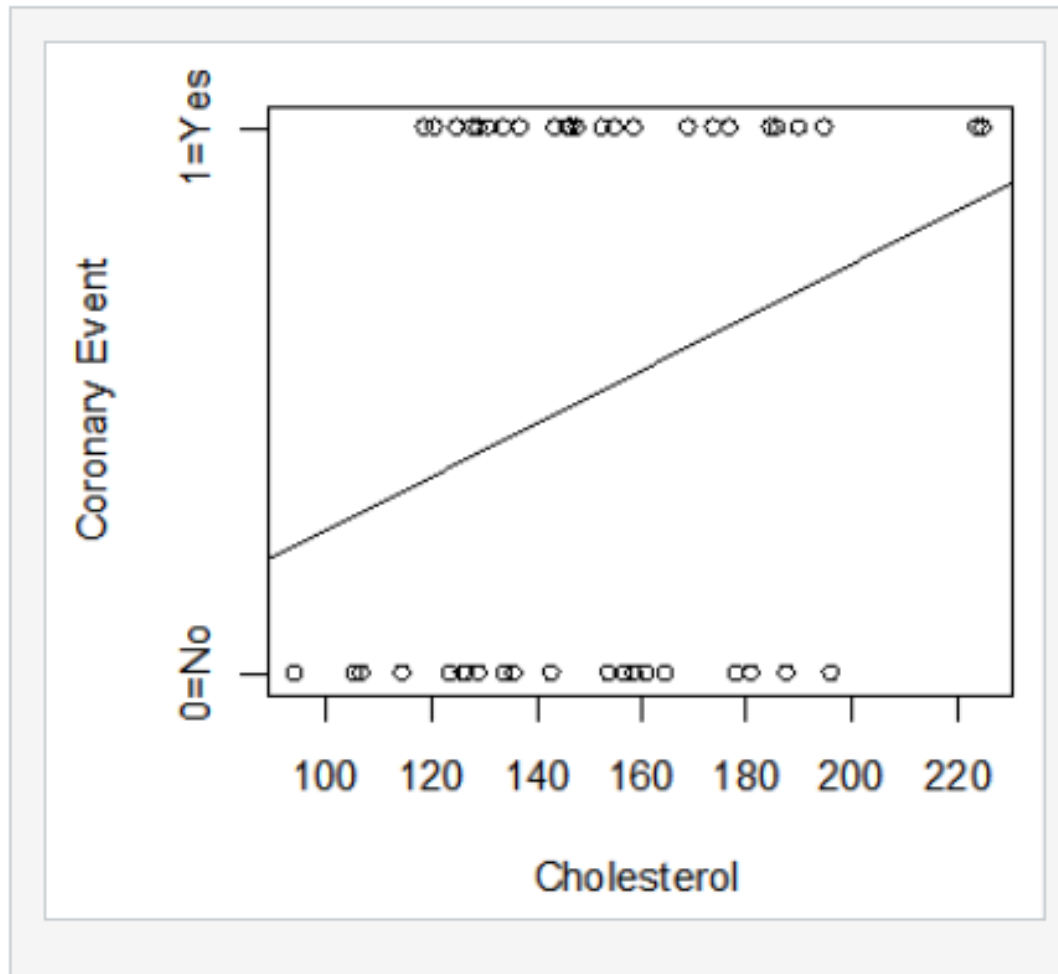
---

- ▶ Risk difference =  $\hat{p}_1 - \hat{p}_2 = 0.803 - 0.613 = 0.19$ 
  - ▶ The need for services was nearly 20% **higher** among those who did not attend preschool as compared with those who attended preschool
- ▶ Risk Ratio =  $\hat{p}_1 / \hat{p}_2 = 0.803 / 0.613 = 1.31$ 
  - ▶ The need for services is 1.3 **times higher** among those who did not attend preschool as compared with those who attended preschool.
- ▶ Odds Ratio =  $\frac{\hat{p}_1 / 1 - \hat{p}_1}{\hat{p}_2 / 1 - \hat{p}_2} = \frac{0.803 / 1 - 0.803}{0.613 / 1 - 0.613} = 2.6$ 
  - ▶ The **odds** of needing services is 2.6 **times higher** among those who did not attend preschool as compared with those who attended preschool.



# Logistic Regression

---



# Logistic Regression

---

$$L = \ln \frac{p}{1-p} = \beta_0 + \beta_1 x + \epsilon$$

where

- $L$  is the log odds of the event
- $p$  is the probability of a success
- $\frac{p}{1-p}$  are the odds of the event
- $x$  is the explanatory variable
- $\beta_0$  is the intercept
- $\beta_1$  is the regression coefficient
- $\epsilon$  is the random error





# Logistic Regression

---

- ▶ We can predict the risk of the outcome of interest, using the estimates of the slope and intercept.

$$\hat{p} = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x}}$$



# Logistic Regression

---

- The interpretation of the regression coefficients are generally based on odds ratios. Consider the odds ratio for an event given  $x=A$  vs.  $x=B$ .

$$\widehat{\text{odds}}_a = e^{\hat{\beta}_0 + \hat{\beta}_1 x_a} \quad \widehat{\text{odds}}_b = e^{\hat{\beta}_0 + \hat{\beta}_1 x_b}$$

$$\widehat{OR}_{x_a \text{ versus } x_b} = \frac{\widehat{\text{odds}}_a}{\widehat{\text{odds}}_b} = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_a}}{e^{\hat{\beta}_0 + \hat{\beta}_1 x_b}}$$




# Logistic Regression

---

- ▶ The interpretation of the regression coefficients are generally based on odds ratios. Consider the odds ratio for an event given  $x=A$  vs.  $x=B$ .

$$\begin{aligned}\widehat{OR}_{x_a \text{ versus } x_b} &= \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_a}}{e^{\hat{\beta}_0 + \hat{\beta}_1 x_b}} \\ &= e^{(\hat{\beta}_0 + \hat{\beta}_1 x_a) - (\hat{\beta}_0 + \hat{\beta}_1 x_b)} \\ &= e^{\hat{\beta}_0 + \hat{\beta}_1 x_a - \hat{\beta}_0 - \hat{\beta}_1 x_b} \\ &= e^{\hat{\beta}_1 x_a - \hat{\beta}_1 x_b} \\ &= e^{\hat{\beta}_1 (x_a - x_b)}\end{aligned}$$

The odds of the event are this much higher for every  $x_a - x_b$  unit increase in  $x$ .



Check out section in notes on confidence intervals for the OR!



# Logistic Regression

---

- ▶ The null hypothesis is generally of the form  $\beta_1 = 0$  ( $H_0$ : there is no association between  $x$  and the odds of the outcome) versus  $\beta_1 \neq 0$  ( $H_1$ : there is an association between  $x$  and the odds of the outcome) .
- ▶ Note that  $\beta_1 = 0$  means  $OR = 1$ .
- ▶ The general principle is that the null hypothesis is rejected if  $\widehat{\beta}_1$  is sufficiently far from 0.



# Logistic Regression

---

- ▶ Use the `glm()` function with binomial option
  - ▶ `glm(data$event ~ data$explanatory1 + data$explanatory2 + ... , family = binomial)`
  - ▶ Ensure that your event is coded 1 = Event and 0 = Non Event (numeric, as opposed to a factor variable)
  - ▶ If one of the variables in the model is a factor variable, it is best to create dummy variables (1/0) so that you know exactly what the reference group is
- ▶ Use the `summary()` function on saved regression result to get regression equation and associated tests for each regression coefficient
- ▶ Use the `exp()` function on the resulting coefficients to obtain odds ratios for each regression coefficient
- ▶ Use the `predict()` function on the saved regression result to get the predicted risks for each observation (or calculate manually)
- ▶ In multiple logistic regression, use the `wald.test()` function (from “aod” package) to get p-value for the global test (of all beta coefficients = 0)
- ▶ Use the `roc()` function (from the “pROC” package) to get c-statistic (area under the curve) and to generate the ROC curve



## Example – Risk of a Coronary Event

---

Suppose we are interested in the association between cholesterol levels, age and gender on having a coronary event in a high risk patient population (who have had an event in the past). We collect information on these parameters for 50 subjects and then follow each for a year to see if they have another coronary event.



# Example – Risk of a Coronary Event

---

- ▶ Simple logistic Regression (predicting risk from cholesterol only)
  - ▶ Odds Ratios and confidence intervals (1 and 10 unit increase)
  - ▶ Risk prediction
- ▶ Multiple logistic Regression (predicting risk from age, gender and cholesterol)
  - ▶ Global test and individual tests



# Area under the ROC Curve

---

- ▶ In linear regression, the R squared value was helpful in quantifying the proportion of variability in the outcome explained by the regression model.
- ▶ We need to use the predicted values in the logistic regression setting.
- ▶ Sensitivity: Proportion of true events classified correctly.
- ▶ Specificity: Proportion of true non-events classified correctly.
- ▶ Area under the ROC (receiver operating characteristic) curve (also known as the c-statistic) is a measure of the sensitivity and specificity across a range of possible cutoffs.
- ▶ Often measures GOF for logistic model.

