STAT115: Introduction to Biostatistics

University of Otago
Ōtākou Whakaihu Waka

Lecture 28: Inference for Proportions

Outline

- A closer look at hypothesis tests for p
- Compare probabilities between two (independent) groups
- Difference in proportions: $p_1 - p_2$
  - Confidence interval
  - Hypothesis test

Hypothesis Testing for a Proportion

- We may wish to test the hypotheses:
  - $\mathrm{H}_0 : p = p_0$
  - $\mathrm{H}_A : p \neq p_0$
- A test statistic can be found using:

$$z = \frac{\text{estimate} - \text{null}}{\text{standard error}} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

- Two things to note:
  - Find standard error assuming null hypothesis is true: $\sigma_{\hat{p}} = \sqrt{\frac{p_0(1-p_0)}{n}}$
  - Find p-value from a (standard) normal distribution
    - That's why the test statistic is z, not t

Looking again at hypothesis test for myopia

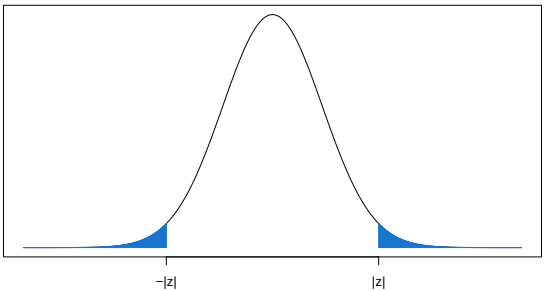
- To test if probability of myopia in randomly chosen Australian (aged 18–22) is different from 0.23: set $p_0 = 0.23$

```
# estimate of p
x = 342; n=1344
phat = x/n
p0 = 0.23
# Find standard error under H0
se = sqrt(p0*(1-p0)/n)
# Find test statistic
z = (phat - p0)/se
# Find pvalue
pval = 2*pnorm(-abs(z))
pval

## [1] 0.03307
```

```
z

## [1] 2.131
```



Hypothesis testing in R

- `prop.test` conducts the hypothesis test in a slightly different way
- By default it uses a continuity correction, etc.
- Details outside the scope of the course
- Differences in methodology have a relatively small impact
  - P-values 0.033 versus 0.036 in myopia example

A Theoretical Aside

Normal approximation

- Binomial: The sampling distribution for $\hat{p}$ was approximated by a normal
  - Provided n is large, and p is not too close to 0 or 1
- This formed the basis for finding confidence intervals and doing hypothesis tests
- Why does this work, and does it generalise?
  - I.e. will this also work for other 'non-normal' distributions?

A Theoretical Aside

Central Limit Theorem

- If we collect a large sample of independent observations from a population with mean μ and standard deviation σ, the sampling distribution of $\bar{Y}$ will be approximately normal
  - Mean μ
  - Standard error $\frac{\sigma}{\sqrt{n}}$
- This is known as the Central Limit Theorem

A Theoretical Aside

What has the Central Limit Theorem got to do with proportions?

- A proportion can be viewed as a mean
- Consider $n = 5$ binary observations: 0, 0, 1, 1, 0.
- The sample mean is $\bar{y} = \frac{1}{5}(0 + 0 + 1 + 1 + 0) = 2/5 = 0.4$
- In other words, $\bar{y} = \hat{p}$ (sample proportion)
- So the Central Limit Theorem justifies use of normal distribution when working with proportions

A Theoretical Aside

Wider implications

- Implications are far wider.
- The Central Limit Theorem justifies methodology for confidence intervals and tests for all of the following, even if the data themselves are not notmal:
  - Population mean μ with one sample: use `t.test`
  - Difference in two means $\mu_1 - \mu_2$: use `t.test`
  - ANOVA: use `aov`
  - Linear regression: use `lm`
- Just need sufficiently large sample size; $n > 30$ a very rough rule of thumb.

Data: Smallpox in Boston

- Data are 6224 observations from individuals in Boston in 1721 who were exposed to smallpox[1]
  - Inoculated: yes or no
  - Result: lived or died
- We are interested in comparing the probability of death for those who were inoculated to those who were not

		inoculated		
		yes	no	Total
result	lived	238	5136	5374
	died	6	844	850
	Total	244	5980	6224

[1] This is the same data that we saw earlier.

Models for binomial data

- We don't yet have the tools to answer the question
  - We only know how to estimate p, not compare p across two groups
- We can look at model extensions for binomial data that parallel those we explored for normal models, e.g.
  - Comparing two or more independent groups
  - Regression-type models: probability of success depends on predictor variables
    - Called logistic regression
  - Defer many of these extensions to later courses (i.e. STAT 210)
- For smallpox data: two independent binomials
  - Inoculated: modelled as binomial with probability of death p_1
    - $x_1 = 6$, $n_1 = 244$
  - Not inoculated: modelled as binomial with probability of death p_2
    - $x_2 = 844$, $n_2 = 5980$

Big picture

- We want to compare the survival between inoculated and uninoculated
- There are multiple ways we could do this, e.g.
  - Difference in probabilities: $p_1 - p_2$
  - Ratio of probabilities (also called relative risk): p_1/p_2
- For now we will focus on $p_1 - p_2$
- It is straightfoward to estimate this difference
  - $\hat{p}_1 - \hat{p}_2$
- We also know those estimates are uncertain
  - Found from data (a sample from the population)
  - Find a confidence interval

Confidence interval for $p_1 - p_2$

- Find a confidence interval using

$$\text{estimate} \pm \text{multiplier} \times \text{standard error}$$

- Estimate: $\hat{p}_1 - \hat{p}_2$
- Multiplier: we again approximate the sampling distribution with normal
  - Multiplier is $z_{1-\alpha/2}$
- Standard error: $\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}$
  - Estimate this with: $s_{\hat{p}_1 - \hat{p}_2} = \sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

Wald confidence interval for $p_1 - p_2$

- Putting this together we have the $100(1-\alpha)\%$ Wald confidence interval:

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

- This is the interval returned by `prop.test` when we have two groups
- As with the Wald interval for p
  - The interval is not that reliable if either n_1 or n_2 is small and either p_1 or p_2 is close to 0 or 1
  - Improved confidence intervals do exist
    - Such intervals can be found in other R packages
- We will use the Wald interval in `prop.test`

In R

```
x = c(6, 844); n = c(244, 5980) # smallpox data
prop.test(x, n)

##
##  2-sample test for equality of proportions with continuity correction
##
## data:  x out of n
## X-squared = 26, df = 1, p-value = 3e-07
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.14002 -0.09307
## sample estimates:
##  prop 1  prop 2
## 0.02459 0.14114
```

- We are 95% confident that the probability of death was between 0.0931 and 0.14 lower for those who were inoculated compared to those who were not

Hypothesis test

- Both p_1 and p_2 are conditional probabilities
  - p_1 is the survival probability given inoculated
  - p_2 is the survival probability given not inoculated
- If $p_1 = p_2$ then survival does not depend on inoculation
  - Survival and inoculation are independent
- We can test the hypotheses:
  - $\mathrm{H}_0 : p_1 - p_2 = 0$ (this is equivalent to $p_1 = p_2$)
  - $\mathrm{H}_A : p_1 - p_2 \neq 0$ (this is equivalent to $p_1 \neq p_2$)

Hypothesis test

- A test statistic can be found using:

$$z = \frac{\text{estimate} - \text{null}}{\text{standard error}}$$

- Estimate is $\hat{p}_1 - \hat{p}_2$
- Null value is 0
- We need the standard error assuming null hypothesis is true
  - The two groups have the same probability: $p_1 = p_2$
  - The null hypothesis doesn't specify what this value is
    - Let's call it p^*

Hypothesis test

- The standard error is: $\sigma_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{p^*(1-p^*)}{n_1} + \frac{p^*(1-p^*)}{n_2}}$
  - This is the standard error above evaluated at $p_1 = p_2 = p^*$
- We don't know p^*
  - Estimate it: $\hat{p}^* = \frac{\text{total success}}{\text{total trials}} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$
  - $\hat{p}^*$ is sometimes call the pooled proportion
- Use this to estimate the standard error: $s_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{\hat{p}^*(1-\hat{p}^*)}{n_1} + \frac{\hat{p}^*(1-\hat{p}^*)}{n_2}}$
- This hypothesis test is found using `prop.test`. As with the test for p:
  - The implementation in R contains minor refinements (e.g. continuity correction)

Hypothesis test: in R

- Using `prop.test` to find the p-value

```
prop.test(x,n)

##
##  2-sample test for equality of proportions with continuity correction
##
## data:  x out of n
## X-squared = 26, df = 1, p-value = 3e-07
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.14002 -0.09307
## sample estimates:
##  prop 1  prop 2
## 0.02459 0.14114
```

Interpretation

- The p-value quantifies the incompatibility between the null hypothesis and the data
  - The p-value $< \alpha = 0.05$, which suggests the data are unusual if the two groups (inoculated and uninoculated) truly had the same probability of survival

Alternate Estimands I: Relative Risk

- There are other ways we could have compared between groups
  - Consider a broad overview of two alternatives, widely used in medicine.
- We could use the relative risk: $RR = \dfrac{p_1}{p_2}$
  - Ratio of the probabilities
  - Think of p_i as the risk in group i
- For smallpox example: $RR = \dfrac{p_1}{p_2} = \dfrac{\text{risk of death for those inoculated}}{\text{risk of death for those uninoculated}}$
- The main advantage of the RR is interpretability
  - A $RR = 1.5$ means the risk is 50% higher in group 1 than group 2
- It is possible to find estimates, confidence intervals, etc.

Alternative Estimands II: Odds Ratio

- What are odds?
  - In popular usage: "What are the odds of that?!?"
  - Even so, odds are not well understood
- If probability of event A is p, then the odds of event A are $p/(1-p)$
- If the probability of rain tomorrow is 0.5, the odds are 1 (to 1)
- If the probability of rain tomorrow is 0.8, the odds are 4 (to 1)

Alternative Estimands II: Odds Ratio

continued

- We can compare two groups with an odds ratio: $OR = \dfrac{p_1/(1-p_1)}{p_2/(1-p_2)}$
  - Harder to interpret, but more reliable in certain situations
  - Useful when comparing to regression-like models for binomial data
- When p_1 and p_2 small (e.g. rare disease), $OR = \dfrac{p_1/(1-p_1)}{p_2/(1-p_2)} \approx \dfrac{p_1}{p_2} = RR$
- Will see relative risk and odds ratio 'in action' in later lectures.

Relative Risk and Odds Ratio for Smallpox Data (in R)

```
p1hat = 6/244 # Estimated prob death for innoculated
p2hat = 844/5980 # Estimated prob death for not innoculated
odds1 = p1hat/(1-p1hat); odds2 = p2hat/(1-p2hat)
c(odds1,odds2)

## [1] 0.02521 0.16433

RR = p1hat/p2hat
RR

## [1] 0.1742

OR = odds1/odds2
OR

## [1] 0.1534
```

Summary

- Look at inference for a proportion p
- Methods theoretically supported by Central Limit Theorem
- Explored comparison between two groups: $p_1 - p_2$
  - Confidence intervals
  - Hypothesis test
- Compared relative risk and odds ratio