Statistical Computing with R: Masters in Data Sciences 503 (S20) Fourth Batch, SMS, TU, 2025

Shital Bhandary
Associate Professor
Statistics/Bio-statistics, Demography and Public Health Informatics
Patan Academy of Health Sciences, Lalitpur, Nepal
Faculty, Masters in Medical Research, NHRC/Kathmandu University
Faculty, FAIMER Fellowship in Health Professions Education, India/USA

Review Preview

- One sample proportion test
- Two samples proportions test
- One-way ANOVA
  - Classical 1-way ANOVA
  - Welch 1-way ANOVA
- Multiple proportion test
- Post-hoc tests!

One-sample proportion test

- It is used to test a claim/hunch that a categorical variable has certain categories in terms of proportion

- For instance, we can claim that there are equal proportion of smokers in a sample of people responding in a random survey

- In other words,

- H0: $P = 0.5$ (Assumed/literature)
- H1: $P \neq 0.5$

- We can test this in R using built-in “prop.test” function

- It needs: x (freq) and n (total)

Let's do it in R: Out of 32 randomly selected people, 19 are smokers! (**Check claim of P=0.5**)

- prop.test(x=19, n=32, p=0.5)
- 1-sample proportions test **with continuity correction**
- data: 19 out of 32, null probability 0.5
- X-squared = 0.78125, df = 1, **p-value = 0.3768**
- alternative hypothesis: true p is not equal to 0.5
- 95 percent confidence interval:
- 0.4078543 0.7578086
- **sample estimates:**
- **p**
- **0.59375 (i.e. 19/32)**

The claim that 50% of the population are smokers i.e. P=0.5 is true in this case!

Questions?

1. Why continuity correction? Not need when:

- n*p = 32*0.5975 = 19.12 > 10
- n*q = 32*(1-0.5975) = 12.88 > 10
- **Lesson learned?**

2. Why chi-squared (X-squared) test is used here (instead of z-test)?

Answer: Normal approximation of binomial distribution? **What is it?**

- **How was the 95% CI was computed for this proportion?**

In theory, proportion test is done with z-test but as there are some inherence problem with it, R uses Chi-square test as both give the same results i.e. p-value.

Let’s do it in R: Without continuity correction now! (Do not use continuity correction without testing!)

- prop.test(x=19, n=32, p=0.5, correct=F)
- 1-sample proportions test without continuity correction
- data: 19 out of 32, null probability 0.5
- X-squared = 1.125, df = 1, **p-value = 0.2888**
- alternative hypothesis: true p is not equal to 0.5
- 95 percent confidence interval:
- 0.4226002 0.7448037
- sample estimates:
- p
- 0.59375

- Interpretation:
- Decision: Since p-value is greater than 0.05, we fail to reject (accept) null hypothesis
- Conclusion: This means that there are equal proportion of smokers in the random sample of 32 people
- However, total sample size is less than 50. This is a major violation of chi-square test as the key assumptions for using this test is that the total sample size must be 50 and above.
- **What to do now?**
  - We need to use exact “binomial” test!

Let’s do it in R: **Exact ‘Binomial’ Proportion Test!**

https://statstutorial.com/one-proportion-z-test-in-r-with-examples/

- If we want the “exact” solution based on **binomial distribution (0 and 1)** then we must use:
- binom.test(x=19, n=32, p=0.5, alternative="two.sided")
- number of successes = 19, number of trials = 32, **p-value = 0.3771**
- **2*sum(dbinom(19:32,32,0.5))?**

- Interpretation:
- Decision: Since p-value is greater than 0.05, we fail to reject (accept) null hypothesis
- Conclusion: This means that there are equal proportion of smokers in the randomly selected sample of people!

No need to test whether n*p and n*q > 10 or not to use chi-square test based on normal approximation!

Two sample proportion test:
H0: P1=P2 vs H1: P1 ≠ P2

- Test the claim that proportion of automatic and manual transmission vehicles are equal in the mtcars data

- prop.test(x=c(19,13), n=c(32,32), alternative="two.sided", correct=F)

- **Why correct=F used here?**

What happens if we do as follows:

- **df <- cbind(x=c(19,13), y=c(13,19))**
- **chisq.test(df, correct=F)**

- 2-sample test for equality of proportions without continuity correction

- data: c(19, 13) out of c(32, 32)
- X-squared = 2.25, df = 1, **p-value = 0.1336**
- alternative hypothesis: two.sided
- 95 percent confidence interval **(zero?)**:

  -0.05315041 0.42815041
- sample estimates:

prop 1 **(P1)**	prop 2 **(P2)**
0.59375	0.40625

95%CI of P1? **95% CI of P2?**

Comparing means of an outcome variable across another variable with more than two categories:

- **One-way ANOVA**

- H_0: $\mu_1 = \mu_2 = \mu_3$
- H_1: **At least one pair of means are not equal**

- **If H_1 is accepted, pairwise comparison (post-hoc) test must be done to find the significant pairs!**

- Compare mpg (miles per gallon) by cars with different gear (numbers of gears) using "mtcars" data

- Dependent variable = mpg
- Independent variable = gear (must be a factor variable)

Assumptions of 1-way ANOVA:

- Same as two-samples t-test:

- Dependent variable must be "normally distributed" for each categories

- Variance across categories must be same

- Normally distributed:
  - Test of normality by each category

- Homogenous variance:
  - **var.test is not useful (>2 groups)**
  - Levene's Variance test is preferred
  - It is available in the "car" package
  - library(car)
  - leveneTest(y~x, data=data)
  - **x must be categorical i.e. factor!**

1-way ANOVA assumptions checks:

Normality by categories:

- with(mtcars, shapiro.test(mpg[gear == 3]))

W = 0.95833, p-value = 0.6634

- with(mtcars, shapiro.test(mpg[gear == 4]))

W = 0.90908, p-value = 0.2076

- with(mtcars, shapiro.test(mpg[gear == 5]))

W = 0.90897, p-value = 0.4614

Equal variance among categories:

library(car)

leveneTest(mpg ~ gear, data=mtcars)

Result:

Levene's Test for Homogeneity of Variance (center = median)

	Df	F value	Pr(>F)
group	2	1.4886	**0.242429**

Levene's Test is a GOF test, so group variances are equal as p-value>0.05.

So, Classical 1-way ANOVA can be used now!

- summary(aov(mpg ~ gear, data = mtcars))
- Since F-test p-value <0.05, we accept H1. At least one of the mean pairs are not equal!
- This means, post-hoc test or pairwise comparison is required!
- **Fisher's LSD uses pairwise t-tests (not good)!**
- For classical 1-way ANOVA, Tukey HSD is the best post-hoc test!
- TukeyHSD (aov(mpg ~ gear, data = mtcars))

	Df	SumSq	MeanSq	Fvalue	Pr(>F)
gear	2	483.2	241.62	10.9	0.000295
Residuals	29	642.8	22.17		

Tukey multiple comparisons of means

95% family-wise confidence level

Fit: aov(formula = mpg ~ gear, data = mtcars)

$gear

	diff	lwr	upr	p adj
4-3	**8.426667**	3.9234704	12.929863	0.0002088
5-3	5.273333	-0.7309284	11.277595	0.0937176
5-4	-3.153333	-9.3423846	3.035718	0.4295874

Check this result with the simple linear model (regression):

- summary(lm(mpg ~ gear, data = mtcars))

- P-value are reported without correcting them i.e. simple t-test were used, which can be checked with this command in R/R Studio:

- **pairwise.t.test(mtcars$mpg, mtcars$gear, p.adj = "none")**
- 3 4
- 4 7.3e-05 (3 vs 4) --
- 5 0.038 (3 vs 5) 0.218 (4 vs 5)

- What is the interpretation?
- **Why gear = 3 category is omitted in the result?**

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	16.107	1.216	13.250	7.87e-14 ***
gear[T.4]	**8.427**	1.823	4.621	7.26e-05 ***
gear[T.5]	**5.273**	2.431	2.169	0.0384 * (why?)

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

- R automatically creates 3 dummy variables for 3 categories of gear variable i.e. 3, 4 and 5 and uses only last two of them in the model and takes the first one as reference!
- gear[T.3] = 1 if gear = 3, else 0
- gear[T.4] = 1 if gear = 4, else 0
- gear[T.5] = 1 if gear = 5, else 0

Multiple proportion test

- H0: P1 = P2 = P3 …. = Pn

- H1: At least one of the proportion pairs are not equal

- Lets do it for gear variable of mtcars data
- table(mtcars$gear)

- prop.test(x=c(15,12,5), n=c(32,32,32)) **#Correct=F?**

- pairwise.prop.test(x=c(15,12,5), n=c(32,32,32), correct=F)

Suggested Book: Elementary Statistics: Step-by-step guide, 9th Edition by Alan Bluman

Question/queries so far?

Thank you!

@shitalbhandary