MATH 7343 Applied Statistics

Prof. (Aidong) Adam Ding

Northeastern University

Review

- Last time, we already covered the inference methods for one populations proportion. The standard methods based on normal approximation are straightforward.
- Improved methods: Wilson interval (inverse the hypothesis test) and Exact test (Binomial test).

- Today we go over more aspects of proportions inference: sample size calculation, two population proportions comparison.

Chapter 14 Inferences on Proportions

Sample size calculation. Same as for the inference of population mean, we can do this for either confidence interval or power of hypothesis testing.

- **(1) Confidence interval sample size calculation**:
- Example: we want to estimate the approval rating of President Trump within a margin of error of 5% at 90% confidence. The margin of error refers to the half length of the confidence interval : $\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

Sample size calculation (margin of error)

- Approval Rating Example: we want a margin of error of 5% at 90% confidence.

90% confidence $\Rightarrow \alpha = 1-0.9 = 0.1 \Rightarrow z_{\alpha/2} = z_{0.05} = 1.645$

Want $0.05 = 1.645\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \Rightarrow \sqrt{n} = \frac{1.645}{0.05}\sqrt{\hat{p}(1-\hat{p})}$

$\Rightarrow \quad n = \left(\frac{1.645}{0.05}\right)^2 \hat{p}(1-\hat{p})$

- There is an issue: we do not know the $\hat{p}$ value since we do not have any data yet. How to deal with this?

Sample size calculation (margin of error)

- **Approval Rating Example**: $n=\left(\frac{1.645}{0.05}\right)^2 \hat{p}(1-\hat{p})$

We do not know the $\hat{p}$ value. Two ways to deal with it:

- (1) Do a preliminary study, plug-in an estimate value.
- (2) Using the mathematical bound:

$$p(1-p) \le \frac{1}{2}\left(1-\frac{1}{2}\right) = \frac{1}{4} \text{ for all } 0 \le p \le 1.$$

Hence we can take $n=\left(\frac{1.645}{0.05}\right)^2 \frac{1}{4} = 271$

We need to sample at least 271 persons.

Sample size calculation (margin of error)

- Generally, to estimate within a margin of error at 1-α confidence, we solve the equation

$$margin = z_{\alpha/2}\sqrt{\frac{1}{4}\frac{1}{n}}.$$

Hence $$n = \left(\frac{z_{\alpha/2}}{margin}\right)^2 \frac{1}{4}.$$

As usual, we should round up the answer to the next integer.

Sample size calculation

- **(2) Power approach**:
- Test $\mathbf{H_0}$: $p \le p_0$ versus $\mathbf{H_A}$: $p > p_0$ at α level.
- Also, we want power of at least $1-\beta$ when $p = p_1$.

Rejection rule: Reject $\mathbf{H_0}$ when $\dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}} > z_\alpha$

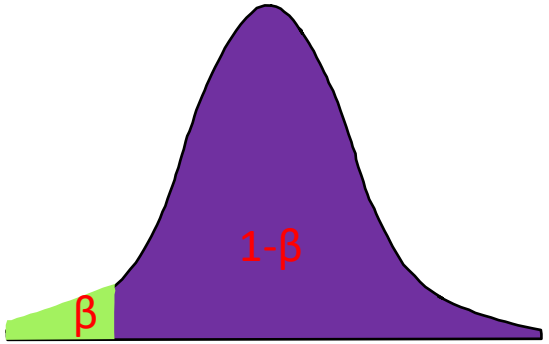
$$\Leftrightarrow \hat{p} > p_0 + z_\alpha \sqrt{\frac{p_0(1-p_0)}{n}}$$

Thus we solve $1-\beta = P\left(\hat{p} > p_0 + z_\alpha \sqrt{\dfrac{p_0(1-p_0)}{n}} \mid p = p_1\right)$

Sample size calculation

(2) Power approach: Test $\mathbf{H_0}$: $p \le p_0$ versus $\mathbf{H_A}$: $p > p_0$ at α level. Also, we want power of at least $1-\beta$ when $p=p_1$.

Solve $1-\beta = P(\hat{p} > p_0 + z_\alpha \sqrt{\frac{p_0(1-p_0)}{n}} \mid p=p_1)$



$$= P\left(\frac{\hat{p}-p_1}{\sqrt{\frac{p_1(1-p_1)}{n}}} > \frac{p_0 + z_\alpha \sqrt{\frac{p_0(1-p_0)}{n}} - p_1}{\sqrt{\frac{p_1(1-p_1)}{n}}} \mid p=p_1 \right)$$

$$= P\left(Z > z_\alpha \sqrt{\frac{p_0(1-p_0)}{p_1(1-p_1)}} + \frac{p_0-p_1}{\sqrt{\frac{p_1(1-p_1)}{n}}}\right)$$

Sample size calculation

(2) Power approach: Test $\mathbf{H_0}$: $p \le p_0$ versus $\mathbf{H_A}$: $p > p_0$ at α level. Also, we want power of $1-\beta$ at least when $p = p_1$.

Solve $$-z_\beta = z_\alpha \sqrt{\frac{p_0(1-p_0)}{p_1(1-p_1)}} + \frac{p_0 - p_1}{\sqrt{\frac{p_1(1-p_1)}{n}}}$$

$$\Leftrightarrow -z_\beta \sqrt{p_1(1-p_1)} - z_\alpha \sqrt{p_0(1-p_0)} = (p_0 - p_1)\sqrt{n}$$

$$\Leftrightarrow n = \left(\frac{z_\beta \sqrt{p_1(1-p_1)} + z_\alpha \sqrt{p_0(1-p_0)}}{p_1 - p_0} \right)^2$$

This is the formula for the sample size of an α level test to achieve 1-sided power of $1-\beta$.

Sample size calculation

Approval Rating Example : Test $\mathbf{H_0}$: $p \le 0.5$ versus $\mathbf{H_A}$: $p > 0.5$ at $\alpha = 0.05$ level. Also, we want at least 80% power if the approve rate is indeed 51%. How many voters do we need to sample?

Solution: $\alpha = 0.05$, $z_{0.05} = 1.645$, $1-\beta = 0.8$, $z_{0.2} = 0.84$

$$\Leftrightarrow n = \left(\frac{z_\beta \sqrt{p_1(1-p_1)} + z_\alpha \sqrt{p_0(1-p_0)}}{p_1 - p_0} \right)^2 = \left(\frac{0.84\sqrt{0.51\,(0.49)} + 1.645\sqrt{0.5(0.5)}}{0.01} \right)^2$$

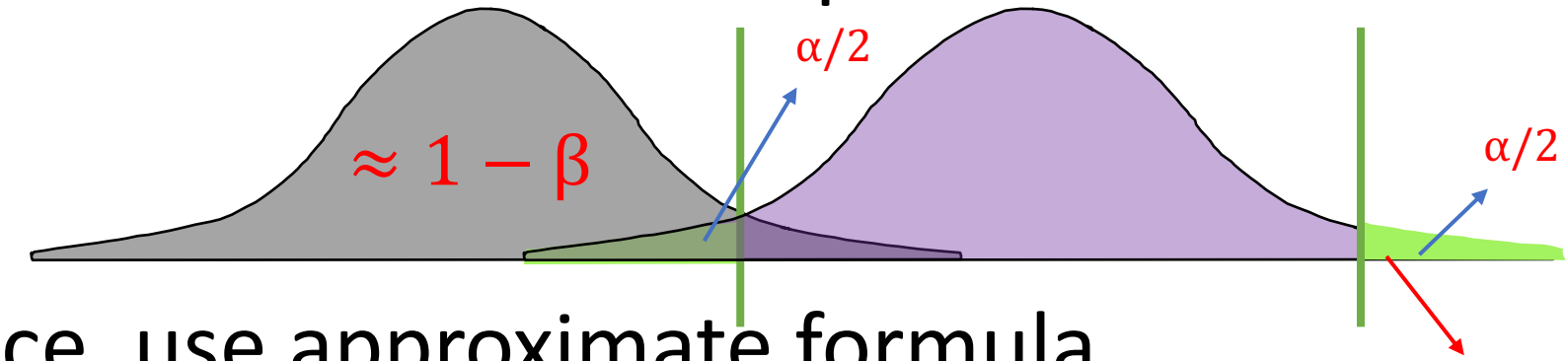
$$= 15436$$

We need to sample 15436 voters.

Sample size calculation

(2) Power approach:

What is the formula for the sample size of an α level test to achieve *2-sided power* of $1-\beta$? The exact mathematical solution is complicated.



In practice, use approximate formula

$$n = \left(\frac{z_{\beta} \sqrt{p_1(1-p_1)} + z_{\alpha/2} \sqrt{p_0(1-p_0)}}{p_1 - p_0} \right)^2$$

Sample size calculation

Approval Rating Example : Test H_0: p=0.5 versus H_A: p≠0.5 at α=0.05 level. Also, we want at least 80% power to detect the approve rate difference of 1% from 50%. How many votes do we need to sample?

Solution: α=0.05, $z_{0.025}$=1.96, 1-β=0.8, $z_{0.2}$=0.84

$$\Leftrightarrow n = \left(\frac{z_\beta \sqrt{p_1(1-p_1)} + z_{\alpha/2}\sqrt{p_0(1-p_0)}}{p_1 - p_0}\right)^2 = \left(\frac{0.84\sqrt{0.51\,(0.49)} + 1.96\sqrt{0.5(0.5)}}{0.01}\right)^2$$

$$= 19598$$

We need to sample 19598 voters.

Two Population Proportions Comparison

- **(A) Data coming from two independent populations**

Thus $X_1 \sim Bin(n_1, p_1)$ independent of $X_2 \sim Bin(n_2, p_2)$

- **(1) Point estimator for p_1-p_2 :** $\hat{p}_1 - \hat{p}_2 = \frac{X_1}{n_1} - \frac{X_2}{n_2}$

- **(2) Confidence interval for p_1-p_2 :**

Since $Var(\hat{p}_1 - \hat{p}_2) = Var(\hat{p}_1) + Var(\hat{p}_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$

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

Two Population Proportions Comparison

- **(A) Two independent populations**
- Example: Insecticide's impact on starling mortality.

  Group 1: diet treated with insecticide. n_1=80, X_1=11

  Group 2: normal diet. n_2=80, X_2=3
- **(1) Point estimator for** p_1-p_2 : $\hat{p}_1$-$\hat{p}_2$= $\frac{11}{80}$ - $\frac{3}{80}$= 0.10
- **(2) A 99% C.I. for** p_1-p_2 : $\hat{p}_1$=$\frac{11}{80}$ =0.14, $\hat{p}_2$= $\frac{3}{80}$ = 0.04

$$0.1 \pm 2.576\sqrt{\frac{0.14(1-0.14)}{80} + \frac{0.04(1-0.04)}{80}} = (-0.013, 0.213)$$

Two Population Proportions Comparison

- **(A) Data coming from two independent populations**
- **(3) Hypothesis test:** Test $\mathbf{H_0}$: $p_1 = p_2 = p$ versus $\mathbf{H_A}$: $p_1 \neq p_2$.
  Under $\mathbf{H_0}$, $X_1 \sim \text{Bin}(n_1, p)$ independent of $X_2 \sim \text{Bin}(n_2, p)$
  $\Rightarrow X_1 + X_2 \sim \text{Bin}(n_1+n_2, p)$

Hence $\text{Var}(\hat{p}_1 - \hat{p}_2) = \frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2} = p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$.

p is estimated by pooling $\hat{p} = \frac{X_1 + X_2}{n_1+n_2}$.

Reject $\mathbf{H_0}$ when $Z_{obs} = \left| \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \right| > z_{\alpha/2}$

Two Population Proportions Comparison

- **(A) Data coming from two independent populations**
- **(3) Hypothesis test:**

Starling example. Test $\mathbf{H_0}$:p_1=p_2=p versus $\mathbf{H_A}$:$p_1 \neq p_2$.

$$\hat{p} = \frac{14}{160} = 0.0875.$$

$$Z_{obs} = \left| \frac{0.1}{\sqrt{0.0875(1-0.0875)\left(\frac{1}{80}+\frac{1}{80}\right)}} \right| = 2.238$$

From Table A.3, p-value = 2(0.013)=0.026

Reject $\mathbf{H_0}$ at α=0.05 level. Fail to reject $\mathbf{H_0}$ at α=0.01 level. (Recall 0 is in the 99% C.I.)

Two Population Proportions Comparison

- **(B) Paired proportions comparison test** (This test is not in textbook, but the same as McNemar test in section 15.2)
- **Example:** The association between Myocardial Infarction (MI) and diabetes. Data on page 350 of the textbook.

  Diabetes among 144 MI victims is 46,

  Diabetes among 144 matched persons without MI is 25.

So $\hat{p}_1 = \frac{46}{144} = 0.3194$, $\hat{p}_2 = \frac{25}{144} = 0.1736$.

Var($\hat{p}_1$-$\hat{p}_2$) = ? (we do not know). Notice it $\neq \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$ since the data is NOT independent (matched pairs).

Two Population Proportions Comparison

- **(B) Paired proportions comparison**
- **MI/Diabetes Example:** Diabetes among 144 MI victims is 46; Diabetes among 144 matched persons without MI is 25.

We do not know Var($\hat{p}_1$-$\hat{p}_2$), can not do C.I. and test.

We need more detailed data!

MI	No MI		Total
	Diabetes	No Diabetes	
Diabetes	9	37	46
No Diabetes	16	82	98
Total	25	119	144

From p350 of textbook,

the second table on that page.

Of the 46 Navajos who had experienced acute myocardial infarction and who were diabetic, 9 were matched with individuals who had diabetes and 37 with individuals who did not. Of the 98 infarction victims who did not suffer from diabetes, 16 were paired with diabetics and 82 were not. Each entry in the table corresponds to the combination of responses for a matched pair rather than an individual person.

Paired proportions comparison

Generally, the paired data is

Observed

Second sample \ First sample	True	False	
True	n_{TT}	n_{FT}	X_2
False	n_{TF}	n_{FF}	$n-X_2$
	X_1	$n-X_1$	n

Expected

Second sample \ First sample	True	False	
True	$n*p_{TT}$	$n*p_{FT}$	$n*p_2$
False	$n*p_{TF}$	$n*p_{FF}$	$n(1-p_2)$
	$n*p_1$	$n(1-p_1)$	n

$\hat{p}_1=\frac{X_1}{n}$, $\hat{p}_2=\frac{X_2}{n}$. Let $\hat{p}_{TT}=\frac{X_{TT}}{n}$, $\hat{p}_{FT}=\frac{X_{FT}}{n}$, $\hat{p}_{TF}=\frac{X_{TF}}{n}$ and $\hat{p}_{FF}=\frac{X_{FF}}{n}$.

Then $\hat{p}_1-\hat{p}_2=\hat{p}_{TF}-\hat{p}_{FT}$; $\mathrm{Var}(\hat{p}_1-\hat{p}_2)=\frac{p_{TF}+p_{FT}}{n}-\frac{(p_{TF}-p_{FT})^2}{n}$

The (1-α) C.I. is $\hat{p}_1-\hat{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_{TF}+\hat{p}_{FT}}{n}-\frac{(\hat{p}_{TF}-\hat{p}_{FT})^2}{n}}$

Two Population Proportions Comparison

- **(B) Paired proportions comparison**

MI/Diabetes Example: $\hat{p}_1$ =0.3194,$\hat{p}_2$=0.1736. $\hat{p}_1$-$\hat{p}_2$=0.146

We also need $\hat{p}_{TF}=\frac{16}{144}$ =0.1111, $\hat{p}_{FT}=\frac{37}{144}$ = 0.2569.

A 95% C.I. for p_1-p_2 :

$$0.146 \pm 1.96\sqrt{\frac{0.111+0.257}{144} - \frac{(0.111-0.257)^2}{144}}$$

= (0.069, 0.223)

MI	No MI		Total
	Diabetes	No Diabetes	
Diabetes	9	37	46
No Diabetes	16	82	98
Total	25	119	144

Hence the proportions indeed differ. The diabetes and heart attacks are associated. (More likely MI for persons with diabetes).

Two Population Proportions Comparison

- **(B) Paired proportions comparison**
- **Hypothesis test:** Test $\mathbf{H_0}$: $p_1 = p_2 = p$ versus $\mathbf{H_A}$: $p_1 \neq p_2$.

Under $\mathbf{H_0}$, $p_{TF} = p_{FT}$; $\text{Var}(\hat{p}_1 - \hat{p}_2) = \frac{2p_{TF}}{n}$ $\left(= \frac{p_{TF}+p_{FT}}{n} - \frac{(p_{TF}-p_{FT})^2}{n}\right)$

p_{TF} is estimated by pooling $\tilde{p}_{TF} = \frac{\hat{p}_{TF} + \hat{p}_{FT}}{2}$.

Reject $\mathbf{H_0}$ when $Z_{obs} = \left| \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_{TF} + \hat{p}_{FT}}{n}}} \right| > z_{\alpha/2}$

$$\Leftrightarrow Z_{obs}^2 > (Z_{\alpha/2})^2 = \chi^2_{\alpha, df=1}$$

We will come back to this test in Chapter15.

Module 10 Contingency Tables (Chapter 15)

- The tests in this chapter are all versions of the χ^2-test.
- The χ^2-test is a general goodness-of-fit test, not restricted to usage on contingency tables. We first look at the setup of general χ^2-test (this part is not covered in textbook).

χ^2-test as a general goodness-of-fit test

- The χ^2-test for data in finite many K categories (cells):

(1) Under null hypothesis $\mathbf{H_0}$: find the best estimated frequencies E_i ;

(2) Compare with the observed frequencies O_i to get

$$\chi^2_{Obs} = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i}$$

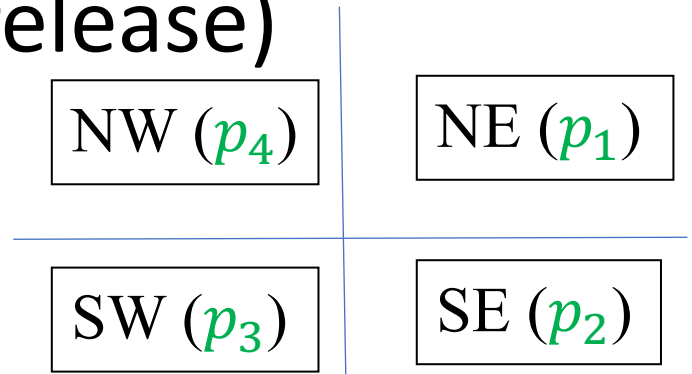
(3) Degree of freedom is given by

df= number of cells - number of estimated parameters -1

Reject $\mathbf{H_0}$ if $\chi^2_{Obs} > \chi^2_{\alpha,df}$

χ^2-test as a general goodness-of-fit test

- Example: Do pigeons know their way to home when released? Record the flight angles of homing pigeons (at the time of release)

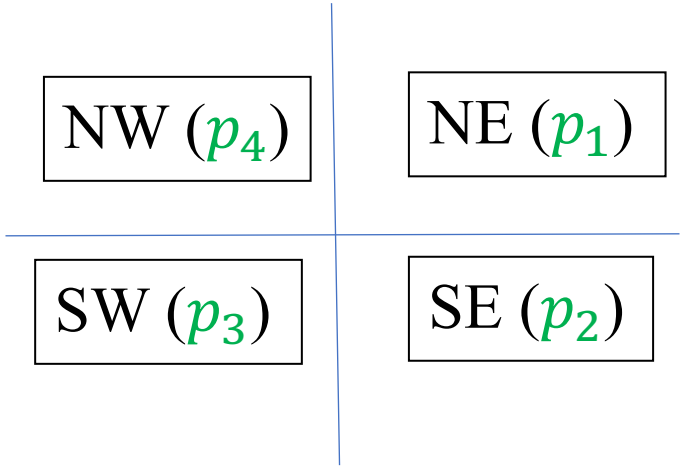


- **H_0**: Pigeons have no idea which to go at the time of release. $p_1 = p_2 = p_3 = p_4 = 1/4$.

  versus **H_A**: Not all p_i equal.

χ^2-test as a general goodness-of-fit test

- Pigeons Example: $\mathbf{H_0}$: p_1=p_2= p_3=p_4 =1/4.



	X_1	X_2	X_3	X_4	Total
Observed	18	24	36	22	100
Expected under $\mathbf{H_0}$	25	25	25	25	

$$\chi^2_{Obs} = \sum_{i=1}^{K} \frac{(X_i - n/4)^2}{n/4}$$

$$= \frac{(18-25)^2}{25} + \frac{(24-25)^2}{25} + \frac{(36-25)^2}{25} + \frac{(22-25)^2}{25} = 7.20$$

d.f. = # of cells - # of est. para -1 = 4-0-1 =3

χ^2-test as a general goodness-of-fit test

- Pigeons Example: $\mathbf{H_0}$: p_1=p_2= p_3=p_4 =1/4.

χ^2_{Obs} =7.20, d.f. =3

From Table A.8, 0.05<p-value<0.10.

Or use R: 1-pchisq(7.2, df=3) to get p-value=0.06578905

Fail to reject $\mathbf{H_0}$ at α=0.05 level.

Conclusion: Pigeons do not know their direction when released

- The χ^2-test is an approximate test. As a rule of thumb, we may use it when each cell has ≥5 observations.

Summary

Today, we finished Module 9 Inferences on proportions

- Confidence intervals and hypothesis test for one population proportion through normal approximation.
- Wilson interval and exact test.
- Two proportions comparison: paired and unpaired.

- We started the general χ^2-test. Next lecture we will use it on the contingency tables (Module 10).