MATH 4720 / MSCS 5720

Instructor: Mehdi Maadooliat

Chapter 10(Part A)

Department of Mathematics, Statistics and Computer Science

CATEGORICAL DATA ANALYSIS (ANALYSIS FOR COUNT DATA)

MARQUETTE UNIVERSITY
Be The Difference.

- **Two Categorical Variables**
- **Example: (Evaluation of president's performance vs Gender)**

subject	President's Job Performance	Gender
1	Approve	F
2	Disapprove	M
.	No opinion	F
.	.	
100	Approve	M

- **Variables:**
  - **President's Job Performance:** **Approve, Disapprove, No Opinion**
  - **Gender:** **Male, Female**

- **First variable has three levels, and the second has two levels**

ANALYSIS OF COUNT DATA

MARQUETTE UNIVERSITY
Be The Difference.

- We can of course convert this data into count data

	President's Job Performance		
Gender	Approve	Disapprove	No Opinion
Male	20	25	5
Female	27	20	3

- One question may be to test whether the opinion on
  - President's Job Performance depends on Gender.
- How to formulate this problem in hypothesis testing?
- What is the probability distribution?
  - Of course, we cannot use normal distribution.

ANALYSIS OF COUNT DATA CONT'D

- **In general, for Categorical Data, what probability distribution should be considered?**

- **Categorical Variable is A with categories:** $A_1, A_2, \dots, A_k$

Subject	A_1	A_2	.	.	A_k
1	x				
2				x	
.		x			
.					x
n	x				

- **Switch to Count Data, and we get:**

A:	A_1	A_2	…	A_k
Count	y_1	y_2	…	y_k

- **where** $y_1 + y_2 + \cdots + y_k = n$

WHAT IS THE PROBABILITY DISTRIBUTION OF THIS COUNTS?

- **The probability distribution of $(Y_1, Y_2, \dots, Y_k)$ is Multinomial distribution**

$$P(y_1, y_2, \dots, y_k) = \frac{n!}{y_1!\, y_2! \dots y_k!} \pi_1^{y_1} \pi_2^{y_2} \dots \pi_k^{y_k}$$

- **Here**
  - $\pi_1 = P(A_1) =$ **Population proportion of category** A_1
  - $\pi_2 = P(A_2) =$ **Population proportion of category** A_2
  - $\vdots$
  - $\pi_k = P(A_k) =$ **Population proportion of category** A_k

- **We write** $(Y_1, Y_2, \dots, Y_k) \sim Multinomial(n; \pi_1, \pi_2, \dots, \pi_k)$

- **Note that this is a generalization of the binomial distribution. In the binomial distribution, you have two categories:** $A_1(success)\ \ and \ \ A_2(Failure)$

GOING BACK TO EXAMPLE

- **In the example of President's Job Performance, we have**
  - **President Job Performance:** **Approve** **Disapprove** **No Opinion**
  - **Count** Y_1 Y_2 Y_3
- $(Y_1, Y_2, Y_3) \sim \ Multinomial(n;\ \pi_1, \pi_2, \pi_3)$
- $\pi_1 = P(Approve),\ \pi_2 = P(Disapprove), \pi_3 = P(No\ Opinion)$
- **Any statistical inference, now, can be made in terms of**

$$(\pi_1, \pi_2, \pi_3)$$

- **So the statistical analysis for the categorical data is statistical analysis of multinomial distribution.**

SIMPLE EXAMPLE (BINOMIAL)

- **Example: Exit Poll**

- **Suppose, we collected data on 1,000 voters in election with only two candidates: R and D**

- **Data**

Voter	R	D
1	x	
2		x
⋮		
1,000	x	

- **Based on this data, we want to forecast who won the election.**

POLL EXAMPLE CONT'D

- Let $Y = \#$ of voters voted for R = 551
- $Y \sim Binimial(n = 1000,\ \pi)$
- $\pi =$ P(a voter voted for R)

  = proportion of all voters voted for R

- We want to predict that "R won the election"
- $H_0: \pi \leq \frac{1}{2}$
- $H_a: \pi > \frac{1}{2}$ (more than $\frac{1}{2}$ voted for R)
- **So, if we reject H_0 in favor of H_a at $\alpha = 0.05$, this would mean that our forecast that "R won" is with P(False Discovery)=0.05.**

HYPTHESIS TESTING FOR π

- $H_0: \pi = \pi_0$
  - $H_a: \pi > \pi_0$
  - $H_a: \pi < \pi_0$
  - $H_a: \pi \neq \pi_0$

- **T.S.** $z = \dfrac{\hat{\pi} - \pi_0}{\sqrt{\dfrac{\pi_0(1-\pi_0)}{n}}}$
  - **where** $\hat{\pi}$ $=$ **sample proportion** $= \dfrac{Y}{n}$

- **Assumption:**
  - $n\pi_0 \geq 5, \; n(1 - \pi_0) \geq 5$

- **Decision Rule: Reject** H_0 **in favor of** H_a **if**
  - $H_a: \pi > \pi_0$: **Reject** H_0 **in favor of** H_a **if** $z > z_\alpha$
  - $H_a: \pi < \pi_0$: **Reject** H_0 **in favor of** H_a **if** $z < -z_\alpha$
  - $H_a: \pi \neq \pi_0$: **Reject** H_0 **in favor of** H_a **if** $|z| > z_{\alpha/2}$

CONFIDENCE INTERVAL FOR π

- **Estimate π with a 100(1-α)% confidence interval**

$$\hat{\pi} \pm z_{\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$

- **Assumption:**

$$n\hat{\pi} \geq 5, \qquad n(1-\hat{\pi}) \geq 5$$

BACK TO EXAMPLE

MARQUETTE UNIVERSITY
Be The Difference.

- **In an exit poll of $1{,}000$ voters, 516 voted for R. Assume that there are only two candidates: R and D. Is there a sufficient evidence to conclude at $\alpha = 0.05$ that "R won" the election.**

- **If π is the proportion of all voters voted for R**

- $H_0 : \pi \leq \frac{1}{2}$
- $H_a : \pi > \frac{1}{2}$

- **Assumption:** $n\pi_0 = 500 \geq 5,\ n(1-\pi_0) \geq 5$ **(True)**

- **T.S.** $z = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$, **where** $\hat{\pi} = \frac{516}{1000} = 0.516,\ \pi_0 = \frac{1}{2}$

EXAMPLE CONT'D

- $$z = \frac{0.516 - 0.5}{\sqrt{\frac{0.5(1-0.5)}{1000}}} = 1.01$$

- **Decision Rule:**
  - **Reject H_0 in favor of H_a if $z > z_\alpha = 1.64$**

- **Conclusion: Is $z > 1.64$?**
  - **No. Fail to Reject H_0 in favor of H_a.**
  - **We do not have sufficient evidence to conclude that "R won."**

- **We can conclude the same based on p-value:**
- $$p - value = P(Z > 1.01) = 0.1562 > 0.05$$

EXAMPLE CONT'D

- **Estimate the proportion of all voters voted for R using 95% confidence interval**

$$\hat{\pi} \pm z_{\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$

  - **where** $\hat{\pi} = 0.516$

- **Assumption:** $n\hat{\pi} = 516 \geq 5 \quad n(1-\hat{\pi}) \geq 484 \geq 5$

- **95% CI of** π**:** $0.516 \pm 1.96\sqrt{\frac{0.516(1-0.516)}{1000}}$

$$0.516 \pm 0.031$$

$$0.485 < \pi < 0.547$$

SAMPLE SIZE DETERMINATION

MARQUETTE UNIVERSITY

Be The Difference.

- **Finding sample size so that π can be estimated with $100(1-\alpha)\%$ at a margin of error of E.**

$$\hat{\pi} \pm E$$

- **Formula:** $n = \dfrac{z_{\alpha/2}^2 \pi(1-\pi)}{E^2}$

- **Since π is unknown, a good guess can be used.**

- **Or, since $\max[\pi(1-\pi)] = \frac{1}{4}$, we can use**

- **Formula:** $n = \dfrac{z_{\alpha/2}^2}{4E^2}$

BACK TO EXIT POLL EXAMPLE:

- **We want to know how many voters to sample to estimate the proportion of voters voted for R with 95% confidence at 2% margin of error.**

- $n = \frac{z_{\alpha/2}^2 \pi(1-\pi)}{E^2}$
- $z_{\alpha/2} = 1.96,$
- $E = 0.02,$

- **Since π is unknown, use**
  - $\max[\pi(1-\pi)] = \frac{1}{4}$

- $n = \frac{z_{\alpha/2}^2}{4E^2} = \frac{1.96^2}{4*0.02^2} = 2401$

TWO POPULATION PROPORTION

- **Comparing Two Population Proportions**

	Group 1	**Group 2**
	n_1	n_2
# of success	Y_1	Y_2

- $Y_1 \sim Binomial(n_1, \pi_1)$ $\quad$ $Y_2 \sim Binomial(n_2, \pi_2)$

- π_1 - **Population proportion of success of Group 1**
- π_2 - **Population proportion of success of Group 2**

HYPTHESIS TESTING FOR π

- $H_0: \pi_1 = \pi_2$
  - $H_a: \pi_1 > \pi_2$
  - $H_a: \pi_1 < \pi_2$
  - $H_a: \pi_1 \neq \pi_2$

- **T.S.** $Z = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi})}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}}$

- **Assumption:**
  - $n_1\hat{\pi}_1 \geq 5, \;\; n_1(1-\hat{\pi}_1) \geq 5$
  - $n_2\hat{\pi}_2 \geq 5, \;\; n_2(1-\hat{\pi}_2) \geq 5$

- **Decision Rule: Reject H_0 in favor of H_a if**
  - $H_a: \pi_1 > \pi_2$: **Reject** H_0 **in favor of** H_a **if** $z > z_\alpha$
  - $H_a: \pi_1 < \pi_2$: **Reject** H_0 **in favor of** H_a **if** $z < -z_\alpha$
  - $H_a: \pi_1 \neq \pi_2$: **Reject** H_0 **in favor of** H_a **if** $|z| > z_{\alpha/2}$

CONFIDENCE INTERVAL FOR $\pi_1 - \pi_2$

- **Estimate $\pi_1 - \pi_2$ with a $100(1-\alpha)\%$ confidence interval**

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

- **Assumption:**
  - $n_1\hat{\pi}_1 \geq 5, \;\; n_1(1-\hat{\pi}_1) \geq 5$
  - $n_2\hat{\pi}_2 \geq 5, \;\; n_2(1-\hat{\pi}_2) \geq 5$

BOOK EXAMPLE 10.7

- A study was done on 300 students to **compare the effectiveness** of teaching English to non-English-speaking people by a **computer software program** and by a **traditional classroom system**.

- A randomly selected 125 students were assigned to computer program and the remaining 175 were assigned to traditional program.

- **Exam Results**

Exam Results	Computer	Traditional
Pass	94	113
Fail	31	62
Total	125	175

- Is there sufficient evidence to conclude that the computer program is more effective than the traditional at $\alpha = 0.05$?

EXAMPLE 10.7 CONT'D

- $H_0: \pi_1 = \pi_2$ **vs.** $H_a: \pi_1 > \pi_2$
- **Here**
  - $\pi_1 =$ **Pop. Prop. of students passing the exam under computer program**
  - $\pi_2 =$ **Pop. Prop. of students passing the exam under traditional program**

- $\hat{\pi}_1 = \frac{94}{125} = 0.752, \;\; \hat{\pi}_2 = \frac{113}{175} = 0.646$
- **Assumptions:**
  - $n_1\hat{\pi}_1 = 94 \geq 5, \;\; n_1(1 - \hat{\pi}_1) = 31 \geq 5$
  - $n_2\hat{\pi}_2 = 113 \geq 5, \;\; n_2(1 - \hat{\pi}_2) = 62 \geq 5$

- **T.S.** $$Z = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi})}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}} = 2.00$$

- **Decision Rule: Reject** H_0 **in favor of** H_a **if** $Z > z_\alpha = 1.64$
- **Conclusion: Is** $Z > 1.64$?
  - **Yes. Reject** H_0**. We have sufficient evidence to conclude that the computer program is more effective.**

EXAMPLE 10.7 CONT'D (CONFIDENCE INT.)

- **Now, suppose you want to know how much effective is the computer program?**
- **Estimate $\pi_1 - \pi_2$ using a 95% confidence interval.**

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

- $z_{\alpha/2} = 1.96.$
- $0.752 - 0.646 \pm 1.96\sqrt{\frac{0.752(1-0.752)}{125} + \frac{0.646(1-0.646)}{175}}$
- 0.106 ± 0.104
- **95% C.I.**

$$0.002 < \pi_1 - \pi_2 < 0.21$$

REMARK

- Assumption that
  - $$n_1\hat{\pi}_1 \geq 5, \;\; n_1(1-\hat{\pi}_1) \geq 5$$
  - $$n_2\hat{\pi}_2 \geq 5, \;\; n_2(1-\hat{\pi}_2) \geq 5$$

  is not satisfied for some experiment since $\hat{\pi}_1$ and $\hat{\pi}_2$ may be very smalls.

- Example: Certain car battery causes fire in engine.
-

	Test Battery	Good Battery
	$n_1 = 10$	$n_2 = 10$
# of cases fire occurred	$y_1 = 2$	$y_2 = 0$

- The above assumption is not satisfied. So, z-test **cannot** be used.

- In such cases, we use **Fisher's Exact test (See Book Example 10.8)**