MATH 4720 / MSSC 5720

Instructor: Mehdi Maadooliat

Chapter 10 (Part B)

Department of Mathematical and Statistical Sciences

INFERENCE FOR MORE THAN TWO PROPORTIONS

MARQUETTE UNIVERSITY
Be The Difference.

- **Chi-Square Goodness of Fit Test**

Category	Observed Count	Expected Count
A_1	O_1	$E_1 = n\pi_1^0$
A_2	O_2	$E_2 = n\pi_2^0$
$\vdots$	$\vdots$	$\vdots$
A_k	O_k	$E_k = n\pi_k^0$

- $H_0: \pi_1 = \pi_1^0,\ \pi_2 = \pi_2^0, \ldots, \pi_k = \pi_k^0$
- $H_a: \pi_i \neq \pi_i^0$ **for some** i

- **Assumption:** $E_i \geq 5$, **for** $i = 1, \cdots, k$
  - **No** E_i **is less than** 1, **and nor more than** 20% **of** E_i**s are less than** 5 **(Cochran Suggestion)**

- **T.S.** $\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$

- **Decision Rule: Reject** H_0, **if** $\chi^2 > \chi_\alpha^2\ \ (df = k - 1)$

BOOK EXAMPLE 10.10:

MARQUETTE UNIVERSITY
Be The Difference.

- **A test drug is investigated to see its effectiveness in reducing blood pressure among hypertensive patients. Suppose the improvement in patients condition is characterized as**

Category	Standard Treatment	Observe count
Marked decrease in BP	50%	120
Moderate decrease in BP	25%	60
Slight decrease in BP	10%	10
No decrease in BP	15%	10

- **Does this data provide sufficient evidence that the new treatment is different than the standard at $\alpha = 0.05$?**

EXAMPLE 10.10 CONT'D

- $H_0: \pi_1 = 0.5, \pi_2 = 0.25, \pi_3 = 0.10, \pi_4 = 0.15$
- H_a: **one of the above is not true**

- **Assumption** $E_i \geq 5$

Category	Observe count	Expected Count
Marked decrease in BP	120	200*0.5=100
Moderate decrease in BP	60	200*0.25=50
Slight decrease in BP	10	200*0.10=20
No decrease in BP	10	200*0.15=30
Total	200	

- **T.S.** $\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$
- $\chi^2 = \frac{(120-100)^2}{100} + \frac{(60-50)^2}{50} + \frac{(10-20)^2}{20} + \frac{(10-30)^2}{30} = 24.33$

EXAMPLE 10.10 CONT'D

MARQUETTE UNIVERSITY
Be The Difference.

- **Decision Rule:**
  - **Reject H_0 if $\chi^2 > \chi^2_\alpha\ (df = 4 - 1 = 3) = 7.815$**

- **Conclusion: Is $\chi^2 > 7.815$.**
  - **Yes, since $\chi^2 = 24.33$.**
  - **Thus, we can conclude that the new treatment is different than the standard.**

- **Note that, from the expected and observe counts, observed counts are higher than the expected counts for Marked decrease and the Moderate decrease in BP. So, it is reasonable to conclude that the new treatment is better than the standard.**

- **Chi-Squared Calculator**

GENERALIZE
CHI-SQUARE GOODNESS OF FIT TEST

MARQUETTE UNIVERSITY
Be The Difference.

- The Chi-Square test of goodness of fit can be used for many different types of inference. Here we give an example to test if a random number generator of Minitab truly generate random numbers.

- Example: Let us look at the uniform random number in [0, 1] generated by Minitab. Note that any random number can be described in a decimal form, and it is sufficient to test the randomness of numbers at each decimal places.

- Let us look at the tenth decimal place. The random numbers are

- {0, 1, 2, 3, …, 9}

- Q. Are these numbers generated by Minitab truly random?

EXAMPLE CONT'D

MARQUETTE UNIVERSITY
Be The Difference.

Digit:	0	1	2	3	4	5	6	7	8	9
Count:	8	15	6	9	12	11	9	10	7	13

- $H_0: \pi_i = \frac{1}{10}$ **for all** $i = 1, 2, \ldots, 10$

  $H_a: \pi_i \neq \frac{1}{10}$ **for some** i

- **Assumption: Expected count** $E_i = n\pi_i = 10 \geq 5$ **for all** i

- **T.S.** $\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = \frac{(8-10)^2}{10} + \frac{(15-10)^2}{10} + \cdots + \frac{(13-10)^2}{10}$

  $= 7.00$

- **Decision Rule: Reject** H_0 **if** $\chi^2 > \chi^2_\alpha(df = 10 - 1) = 16.92$

- **Conclusion: Is** $\chi^2 > 16.92$**. No, since** $\chi^2 = 7.00$**. Thus, we fail to reject** H_0**. We cannot conclude that number generated by Minitab are not random.**

EXAMPLE CONT'D AND P-VALUES

MARQUETTE UNIVERSITY
Be The Difference.

- $p - value = P(\chi^2 > 7.00) = 0.64$
- **Based on p-value, we reach the same conclusion since**
  - $p - value > 0.05.$
- **We can use this p-value to determine the goodness of fit.**
- **Q. To what degree can we say that the numbers are truly random?**
- **Some Guideline:**
  - $p - value \geq 0.25 \quad \rightarrow$ **Excellent Fit**
  - $0.15 \leq p - value < 0.25 \quad \rightarrow$ **Good Fit**
  - $0.05 \leq p - value < 0.15 \quad \rightarrow$ **Moderately good fit**
  - $p - value < 0.05 \quad \rightarrow$ **Reject the fit.**
- **In our case, we have excellent fit.**

BOOK EXAMPLE 10.11:

- **In one investigation, a lake sample was analyzed under a microscope to determine the number of clumps of cells per microscope field. These data are summarized here for 150 fields examined under a microscope. Here y_i denotes the number of cell clumps per field and n_i denotes the number of fields with y_i cell clumps.**

y_i	0	1	2	3	4	5	6	≥7
n_i	6	23	29	31	27	13	8	13

- **Use $\alpha = 0.05$ to test the null hypothesis that the sample data were drawn from a Poisson probability distribution.**

- **Solution:**
  - **Let Y follows $Poisson(\mu)$**
  - **μ is the average number of cell clumps per field**

EXAMPLE 10.11 CONT'D

- **Observed Counts:**

y_i	0	1	2	3	4	5	6	≥7
n_i	6	23	29	31	27	13	8	13

- **Sample mean, $\bar{y} = \frac{\sum_i n_i y_i}{\sum_i n_i} = 3.3$ is a good estimate for μ**
  - $\hat{\mu} = 3.3$

- **Note that the sample mean was computed to be 3.3 by using all the sample data before the 13 largest values were collapsed into the final cell.**

- **Given: $P(Y = y) = \frac{\hat{\mu}^y e^{-\hat{\mu}}}{y!}, \quad y = 0,1,2, \ldots$**

y_i	0	1	2	3	4	5	6	≥ 7
$P(y_i)$ for $\mu = 3.3$	.0369	.1217	.2008	.2209	.1823	.1203	.0662	.0509

EXAMPLE 10.11 CONT'D

MARQUETTE UNIVERSITY
Be The Difference.

- **Given:**

y_i	0	1	2	3	4	5	6	≥ 7
$P(y_i)$ for $\mu = 3.3$	.0369	.1217	.2008	.2209	.1823	.1203	.0662	.0509

- **Note that expected cell count $E_i = nP(y_i)$,**
  - **where $n = \sum_i n_i = 150$**

y_i	0	1	2	3	4	5	6	≥ 7
E_i	5.54	18.26	30.12	33.14	27.35	18.05	9.93	7.63

- **Observed Counts:**

y_i	0	1	2	3	4	5	6	≥ 7
n_i	6	23	29	31	27	13	8	13

- **Assumption: Expected count $E_i \geq 5$ for all i**
- $\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = \frac{(6-5.54)^2}{5.54} + \frac{(23-18.26)^2}{18.26} + \cdots + \frac{(13-7.63)^2}{7.63}$

$= 7.02$

EXAMPLE 10.11 CONT'D

MARQUETTE UNIVERSITY
Be The Difference.

- H_0: Data are drawn from Poisson distribution
- H_a: Data are **NOT** drawn from Poisson distribution

- **T.S.** $\chi^2 = 7.02$

- **Decision Rule:** Reject H_0 if $\chi^2 > \chi^2_\alpha(df = 8 - 2) = 12.59$

- For the null hypothesis with μ unspecified, it is necessary to reduce the degrees of freedom from $k - 1$ to $k - 2$ because we must first estimate the Poisson parameter μ prior to obtaining the cell probabilities.

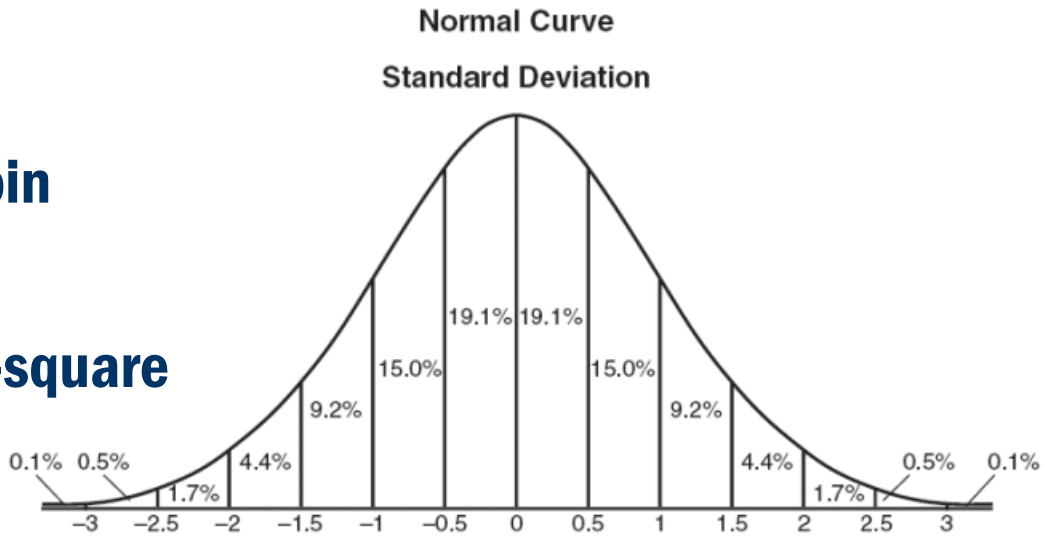
- **Conclusion:** Is $\chi^2 > 12.59$. No, since $\chi^2 = 7.02$. Thus, we fail to reject H_0. We cannot conclude that data are NOT from Poisson distribution.

MORE USE OF CHI-SQUARE GOODNESS OF FIT TEST

MARQUETTE UNIVERSITY
Be The Difference.

- **Suppose you want to test if the data is generated from a normal distribution.**

- **Convert the data into z-scores**
  - $z_i = \frac{y_i - \bar{y}}{s_y}$
- **From the data, we get the observed counts in each bin**



- **From this, we can use chi-square goodness of fit to test**

- $H_0: \pi_1 = 0.001, \pi_2 = 0.005, \pi_3 = 0.017, \ldots, \pi_{14} = 0.001$
- H_a: **one of the above is not true**

CHI-SQUARE TEST OF INDEPENDENCE

MARQUETTE UNIVERSITY
Be The Difference.

- **Suppose, we have two categorical Variables**
  - **Australian Institute of Sports: Gender and Sport**

- **Categories**

A: $A_1, A_2, \ldots A_c$

B: $B_1, B_2, \ldots, B_r$

- **Count Data**
  - **Observed Counts**

	A_1	A_2	$\cdots$	A_c	**Total**
B_1	O_{11}	O_{12}	$\cdots$	O_{1c}	$O_{1.}$
B_2	O_{21}	O_{22}	$\cdots$	O_{2c}	$O_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
B_r	O_{r1}	O_{r2}	$\cdots$	O_{rc}	$O_{r.}$
Total	$O_{.1}$	$O_{.2}$	$\cdots$	$O_{.c}$	$O_{..}$

CHI-SQUARE TEST OF INDEPENDENCE CONT'D

MARQUETTE UNIVERSITY
Be The Difference.

- H_0: A and B are independent
- H_a: A and B are dependent

- Note that we can write this hypothesis in terms of
  - multinomial proportions.

- Let π_{ij} - population proportion for (i, j)- cell, then

- $H_0: \pi_{ij} = \pi_{i.}\pi_{.j}$ vs. $H_a: \pi_{ij} \neq \pi_{i.}\pi_{.j}$,
  - where $\pi_{i.}$ and $\pi_{.j}$ are the marginal probabilities.

- Under $H_0: \pi_{ij} = \frac{O_{i.}}{O_{..}} \times \frac{O_{.j}}{O_{..}}$
- Moreover $\mathrm{E}_{ij} = \pi_{ij} \times O_{..}$
- Therefore $\mathrm{E}_{ij} = \frac{O_{i.} \times O_{.j}}{O_{..}}$

	A_1	A_2	$\cdots$	A_c	Total
B_1	O_{11}	O_{12}	$\cdots$	O_{1c}	$O_{1.}$
B_2	O_{21}	O_{22}	$\cdots$	O_{2c}	$O_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
B_r	O_{r1}	O_{r2}	$\cdots$	O_{rc}	$O_{r.}$
Total	$O_{.1}$	$O_{.2}$	$\cdots$	$O_{.c}$	$O_{..}$

CHI-SQUARE TEST OF INDEPENDENCE CONT'D

MARQUETTE UNIVERSITY
Be The Difference.

- **Observed Counts**

	A_1	A_2	$\cdots$	A_c	**Total**
B_1	O_{11}	O_{12}	$\cdots$	O_{1c}	$O_{1.}$
B_2	O_{21}	O_{22}	$\cdots$	O_{2c}	$O_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
B_r	O_{r1}	O_{r2}	$\cdots$	O_{rc}	$O_{r.}$
Total	$O_{.1}$	$O_{.2}$	$\cdots$	$O_{.c}$	$O_{..}$

- **Expected Counts**

	A_1	A_2	$\cdots$	A_c	**Total**
B_1	$E_{11} = \frac{O_{1.} \times O_{.1}}{O_{..}}$	$E_{12} = \frac{O_{1.} \times O_{.2}}{O_{..}}$	$\cdots$	$E_{1c} = \frac{O_{1.} \times O_{.c}}{O_{..}}$	$O_{1.}$
B_2	$E_{21} = \frac{O_{2.} \times O_{.1}}{O_{..}}$	$E_{22} = \frac{O_{2.} \times O_{.2}}{O_{..}}$	$\cdots$	$E_{2c} = \frac{O_{2.} \times O_{.c}}{O_{..}}$	$O_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
B_r	$E_{r1} = \frac{O_{r.} \times O_{.1}}{O_{..}}$	$E_{r2} = \frac{O_{r.} \times O_{.2}}{O_{..}}$	$\cdots$	$E_{rc} = \frac{O_{r.} \times O_{.c}}{O_{..}}$	$O_{r.}$
Total	$O_{.1}$	$O_{.2}$	$\cdots$	$O_{.c}$	$O_{..}$

CHI-SQUARE TEST OF INDEPENDENCE CONT'D

MARQUETTE UNIVERSITY
Be The Difference.

- **Since, this does not show a practical meaning, we write the hypothesis in the form stated earlier.**

- H_0: **A and B are independent**
- H_a: **A and B are dependent**

- **T.S.** $\chi^2 = \sum\sum \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$

- **Decision Rule: Reject** H_0 **in favor of** H_a **if**
  - $\chi^2 > \chi^2_\alpha \; (df = (r-1)(c-1))$

- **Here**
  - r **= number of rows**
  - c **= number of columns**

EXAMPLE

MARQUETTE UNIVERSITY
Be The Difference.

- **Does the "Opinion on President's Job Performance" depends on "Gender"?**

- **Observed Counts:**

	President's Job Performance			
Gender	Approve	Disapprove	No Opinion	Total
Male	20	25	5	50
Female	27	20	3	50
Total	47	45	8	100

- H_0:**Opinions does not depend on Gender**
- H_a:**Opinion depend on Gender**
- **T.S.** $\chi^2 = \sum\sum \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$

EXAMPLE CONT'D

MARQUETTE UNIVERSITY
Be The Difference.

- **Observed Counts:**

	President's Job Performance			
Gender	Approve	Disapprove	No Opinion	Total
Male	20	25	5	50
Female	27	20	3	50
Total	47	45	8	100

- **Expected Counts**

	President's Job Performance		
Gender	Approve	Disapprove	No Opinion
Male	$\frac{(47)(50)}{100} = 23.5$	$\frac{(45)(50)}{100} = 22.5$	$\frac{(8)(50)}{100} = 4$
Female	$\frac{(47)(50)}{100} = 23.5$	$\frac{(45)(50)}{100} = 22.5$	$\frac{(8)(50)}{100} = 4$

- **T.S.** $\chi^2 = \sum\sum \frac{(O_{ij}-E_{ij})^2}{E_{ij}} = \frac{(20-23.5)^2}{23.5} + \frac{(25-22.5)^2}{22.5} + \frac{(5-4)^2}{4}$

$$+ \frac{(27-23.5)^2}{23.5} + \frac{(20-22.5)^2}{22.5} + \frac{(3-4)^2}{4} = 2.098$$

EXAMPLE CONT'D

MARQUETTE UNIVERSITY
Be The Difference.

- **Decision Rule: Reject H_0 if**
  - $\chi^2 > \chi^2_\alpha\big(df = (2-1)(3-1)\big) = 5.991$

- **Conclusion: Is $\chi^2 > 5.991$?**
  - **No, since $\chi^2 = 2.098$.**
  - **Fail to reject H_0. Thus, we cannot conclude the "Opinion" on President's Job Performance depends on "Gender".**

- **You can also answer this by saying this:**
  - **You cannot conclude that the ways the Male population and the Female population respond are different.**

- **In other words, the problem can be stated in terms of the Homogeneity of two populations.**

CHI-SQUARE TEST WITH CONFOUNDING FACTOR

MARQUETTE UNIVERSITY
Be The Difference.

- Some time, Chi-Square Test of Independence may be misleading if there is a confounding factor. For example, when testing the President's "Job Performance" and "Gender", we did not take the Age into account. It could be that younger population and older population respond differently. In that case Age would be a confounding factor. Suppose we divide the count data into "Young Adult (Age≤ 50)" and "Senior (Age> 50)". In this case we will have two count tables. The Chi-Square test statistics in this case is called Cochran-Mantel-Haenszel statistics.

- For details, see Example 10.17.