USC Department Of Translational Genomics
Keck Medicine of USC

TRGN 599: Applied Data Science and Bioinformatics

UNIT V. Unsupervised Analysis, linear regression, enrichment analysis

Week 10 - Lecture 1

Enrique I. Velazquez Villarreal, M.D., Ph.D., M.P.H., M.S. | **Assistant Professor**

Dept. of Translational Genomics

USC | Keck School of Medicine | Norris Comprehensive Cancer Center

Leader of the USC Bioinformatics Core – *USC CaRE2 Health Equity Center*

David W. Craig, Ph.D. | **Professor and Vice Chair**

Dept. of Translational Genomics

USC | Keck School of Medicine | Norris Comprehensive Cancer Center

Co-Director, Institute of Translational Genomics

Topics

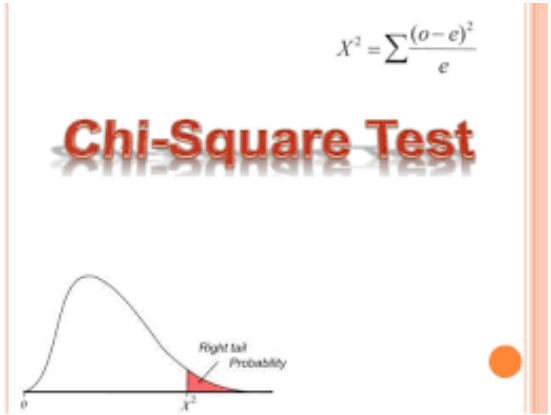
- Analysis of Categorical Variables, Pearson's X^2 Test for One Categorical Variable, Binary Variables, Categorical Variables with Multiple Categories, Pearson's X^2 Test of Independence, Contingency Tables, Fisher's Exact Test.

Introduction

- In previous classes we discussed about about hypothesis testing regarding population proportions.

- It was used the central limit theorem (for large enough sample sizes) to obtain an approximate normal distribution of the sample proportion, which we used as the test statistic.

- It was followed a similar approach in order to test hypotheses regarding the relationship between two binary random variables.



Introduction

- This class we will discuss **Pearson's $\chi 2$ (chi-squared) test** for testing hypotheses regarding the distribution of a categorical variable or the relationship between two categorical variables.

- Pearson's test evaluates whether the probabilities for different categories are equal to the values specified by the null hypothesis.

- Although it is not necessary, we can think of the probability of each category as its population proportion

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

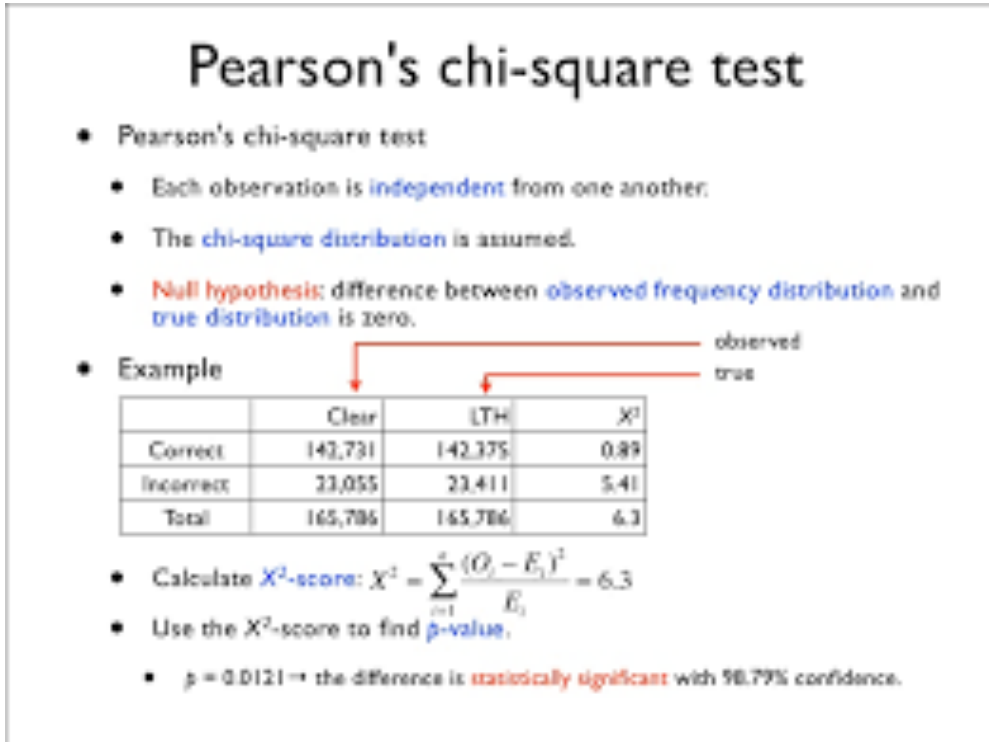
χ^2 = the test statistic $\sum$ = the sum of

O = Observed frequencies E = Expected frequencies

Introduction

- For example, when we will talk about the probability of heart attack survival being 0.7, we can interpret this as 70% of heart attack patients (i.e., 70% of the entire population of people suffered from heart attack) survive.

- As we calculated before, we use the sample proportion of each category as a point estimate for its probability (i.e., its population proportion).



Introduction

- Pearson's $\chi 2$ test uses a test statistic, which we denote as Q, to measure the discrepancy between the observed data and what we expect to observe under the null hypothesis (i.e., assuming the null hypothesis is true).

- Higher levels of discrepancy between data and H0 results in higher values of Q.

- We use q to denote the observed value of Q based on a specific sample of observed data.

- As usual, we need to find the null distribution of Q (i.e., its sampling distribution assuming that H0 is true) and measure the observed significance level pobs by calculating the probability of values as or more extreme than the observed value q.

Pearson's χ2 Test for One Categorical Variable - Binary Variables

- Let's start reviewing Pearson's method by focusing on binary random variables first and then extend this approach where categorical variables have more than two possible values.

- Let us denote the binary variable of interest as X, based on which we can divide the population into two groups depending on whether X = 1 or X = 0.

- Further, suppose that the null hypothesis H0 states that the probability of group 1 (i.e., the probability that an individual belongs to the group 1) is μ01 and the probability of group 2 is μ02.

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

Pearson's χ2 Test for One Categorical Variable - Binary Variables

- The sum of probabilities adds up to one, $\mu 02 = 1 - \mu 01$.
- As a running example, we use the heart attack survival rate (i.e., the probability of survival after heart attack) within one year after hospitalization.
- Suppose that H0 specifies that the probability of surviving is $\mu 01 = 0.70$ and the probability of not surviving is $\mu 02 = 0.30$.
- If we take a random sample of size $n = 40$ from the population (people who suffer from heart attack), we expect that 70% of them survive and 30% of them die within one year from the time of hospitalization if in fact the null hypothesis is true.

Pearson's $\chi 2$ Test for One Categorical Variable - Binary Variables

- That is, we expect that $0.70 \times 40 = 28$ of subjects belong to the first group (survived) and $0.30 \times 40 = 12$ of subjects belong to the second group (non-survived).

- If the null hypothesis is true, we expect that, out of n randomly selected individuals, $E1 = n\mu 01$ belong to the first group, and $E2 = n(1 - \mu 0)$ belong to the second group. We refer to E1 and E2 as the **expected frequencies** under the null.

- In our example, $E1 = 28$ and $E2 = 12$.

Pearson's $\chi 2$ Test for One Categorical Variable - Binary Variables

- Now suppose that we randomly select 40 people who have suffered from heart attack.

- After one year from the time of hospitalization, we find that 24 of them have survived and 16 of them did not survive.

- We refer to the observed number of people in each group as the **observed frequencies** and denote them O1 and O2 for group 1 and group 2, respectively.

- In our example, O1 = 24 and O2 = 16.

Pearson's χ2 Test for One Categorical Variable - Binary Variables

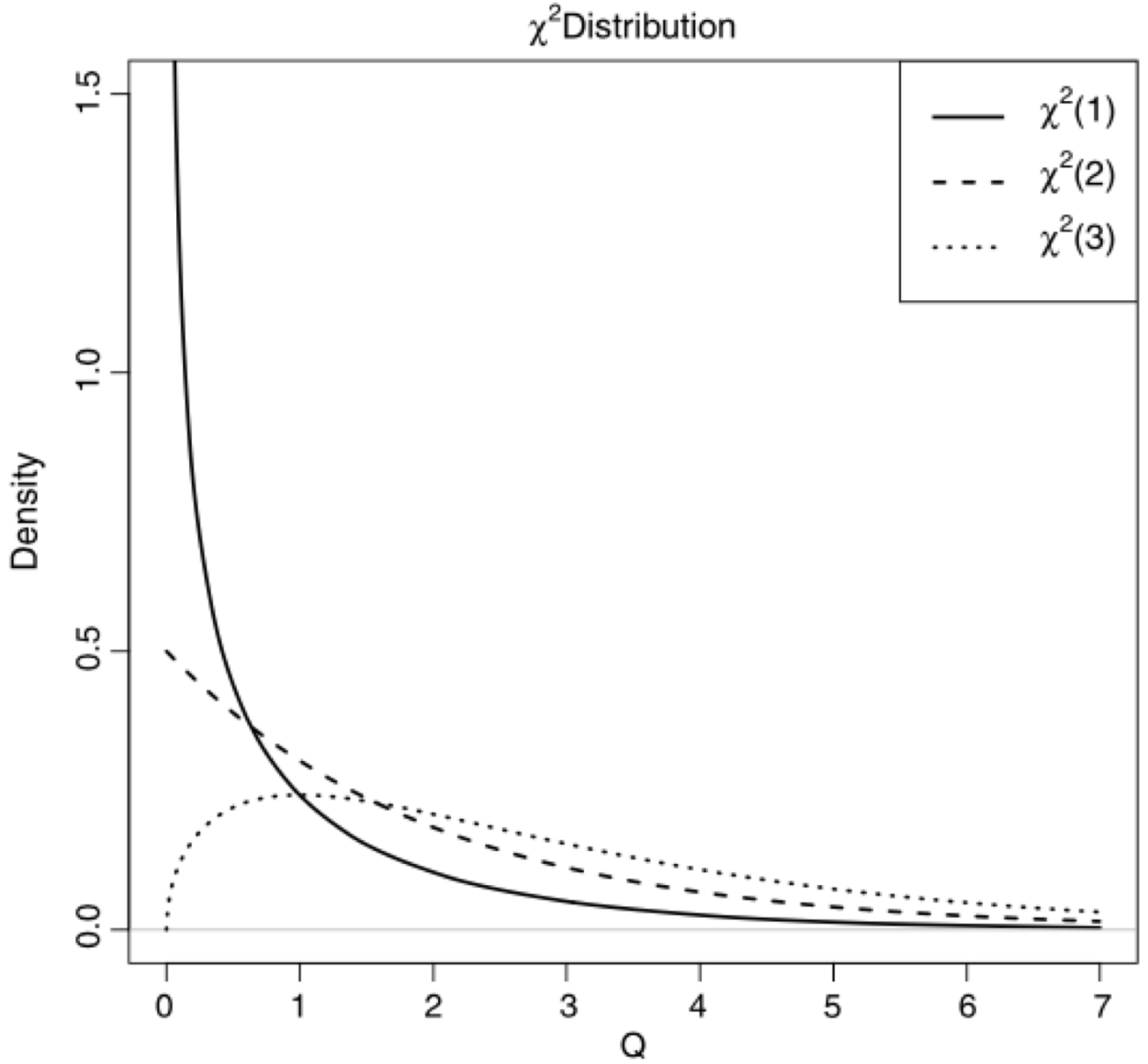
- Pearson's χ2 test measures the discrepancy between the observed data and the null hypothesis based on the difference between the observed and expected frequencies as follows:

$$Q = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2}.$$

- We use q to denote the observed value of the test statistic Q.

Pearson's χ2 Test for One Categorical Variable - Binary Variables

- The plot of the pdf for a χ2 distribution with various degrees of freedom



Pearson's χ2 Test for One Categorical Variable - Binary Variables

- For the heart attack survival example, the observed value of the test statistic is:

$$q = \frac{(24-28)^2}{28} + \frac{(16-12)^2}{12} = 1.90.$$

- The value of Q will be zero only when the observed data matches our expectation under the null exactly.
- When there is some discrepancy between the data and the null hypothesis, Q becomes greater than zero.
- The higher discrepancy between our data and what is expected under H0, the larger Q and therefore the stronger the evidence against H0.

Pearson's $\chi 2$ Test for One Categorical Variable - Binary Variables

- To evaluate the null hypothesis, we need to find the p-value, which is, as usual, the probability of observing as or more extreme values compared to the observed value of the test statistic.

- For this, we first need to find the sampling distribution of the test statistic Q under the null and calculate the probability of observing as large or larger values than q.

- If the null hypothesis is true, then the approximate distribution of Q is $\chi 2$. Like the t -distribution, the χ 2 -distribution is commonly used for hypothesis testing.

- Also, similar to the t distribution, the $\chi 2$ distribution is defined by its degrees of freedom df (which is a positive number) and is denoted $\chi 2(df)$.

Pearson's $\chi 2$ Test for One Categorical Variable - Binary Variables

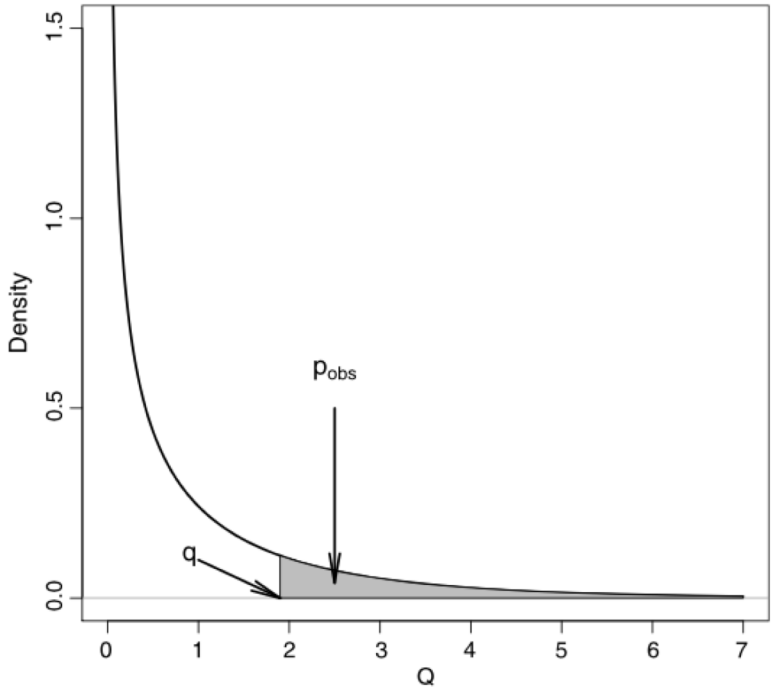
- For binary random variables (i.e., when there are two possible groups), the approximate distribution of Q is $\chi 2(1)$ distribution.

- To evaluate the null hypothesis regarding the probabilities of two groups, we determine the observed significance level pobs by calculating the probability of Q values as or more extreme than the observed value q using the $\chi 2$ distribution with 1 degree of freedom.

- This corresponds to the upper tail probability of q from the $\chi 2(1)$ distribution.

Pearson's χ2 Test for One Categorical Variable - Binary Variables

- For the heart attack survival example, where q = 1.90, this probability is shown as the shaded area:



- The sampling distribution for Q under the null hypothesis: Q ~ χ2(1).
- The p-value is the upper tail probability of observing values as extreme or more extreme than q = 1.90.

Pearson's χ2 Test for One Categorical Variable - Binary Variables

- The probability of observing values as extreme or more extreme than 1.90 based on a χ 2 distribution with 1 degree of freedom is **Pobs = 0.17.**

- Therefore, the results are not statistically significant, and we cannot reject the null hypothesis at commonly used significance levels (e.g., 0.01, 0.05, and 0.1).

- In this case, we believe that the difference between observed and expected frequencies could be due to chance alone.

Categorical Variables with Multiple Categories

- Pearson's $\chi 2$ test can be generalized to situations where the categorical random variable can take more than two values.

- Let us reconsider the heart attack example.

- This time, suppose that we monitor heart attack patients for one year and divide them into three groups:
  - 1. patients who did not have another heart attack and survived.
  - 2. patients who had another heart attack and survived.
  - 3. patients who did not survive.

Categorical Variables with Multiple Categories

- Now suppose that the probabilities of these three groups according to the null is μ01 = 0.5, μ02 = 0.2, and μ03 = 0.3.

- That is, among 70% of patients who survive, 20% of them have another heart attack within a year from their first hospitalization.

- As before, we can find the expected frequencies of each category for a sample of n = 40 patients assuming that the null hypothesis is true:

$$E_1 = 0.5 \times 40 = 20, \qquad E_2 = 0.2 \times 40 = 8, \qquad E_3 = 0.3 \times 40 = 12.$$

Categorical Variables with Multiple Categories

- This time, suppose that the actual observed frequencies based on a sample of size n = 40 for the three groups are:

$$O_1 = 13, \qquad O_2 = 11, \qquad O_3 = 16.$$

- Again, we measure the amount of discrepancy between the observed data and the null hypothesis based on the difference between the observed and expected frequencies:

$$Q = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \frac{(O_3 - E_3)^2}{E_3}.$$

Categorical Variables with Multiple Categories

- For the heart attack survival example, the observed value of this test statistic is:

$$q = \frac{(13-20)^2}{20} + \frac{(11-8)^2}{8} + \frac{(16-12)^2}{12} = 4.91.$$

- For the above example, the p-value is the upper tail probability of 4.91 for a $\chi 2(2)$distribution.

- Calculating the p-value: pobs = $P(Q \geq 8.67) = 0.086$ using the $\chi 2$ distribution with 2 degrees of freedom. Therefore, we can reject the null hypothesis at 0.1 level but not at 0.05 level.

- At the 0.1 significance level, we can conclude that the difference between observed and expected frequencies is statistically significant, and it is probably not due to chance alone.

Categorical Variables with Multiple Categories

- In general, for a categorical random variable with I possible categories, we calculate the test statistic Q as:

$$Q = \sum_{i=1}^{I} \frac{(O_i - E_i)^2}{E_i}.$$

- The approximate distribution of Q is $\chi 2$ with the degrees of freedom equal to the number of categories minus 1: df = I − 1.
- Therefore, to find pobs, we calculate the upper tail probability of q (the observed value of Q) from the $\chi 2(I - 1)$ distribution.

Pearson's $\chi 2$ Test of Independence

- We now discuss the application of Pearson's $\chi 2$ test for evaluating a hypothesis regarding possible relationship between two categorical variables.
- As before, we measure the discrepancy between the observed data and the null hypothesis.
- More specifically, we measure the difference between the observed frequencies and expected frequencies under the null.

Pearson's $\chi 2$ Test of Independence

- The null hypothesis in this case states that the two categorical random variables are independent.
- Recall that two random variables are independent if they do not affect each other's probabilities.
- For two independent random variables, the joint probability is equal to the product of their individual probabilities.
- In what follows, we use this rule to find the expected frequencies.

Pearson's χ2 Test of Independence

- In the following example we will investigate the relationship between smoking and low birth-weight.
- When investigating the relationship between two categorical variables, we typically start by creating a contingency table to summarize the data .
  - In R birthwt data set (variables low and smoke are converted to factors (categorical) variables).
- The contingency table shows the observed frequency of each cell (i.e., each combination of mother's smoking status and baby's birthweight status).
- We denote the observed frequency in row i and column j as Oij .

Contingency table of `low` by `smoke`

Observed frequency		`low`	
		0	1
`smoke`	0	86	29
	1	44	30

Pearson's χ2 Test of Independence

- The following table shows the proportion of observations, out of the total number of observations, that fall within each cell (i.e., each possible combination of smoking status and birth-weight status).

Sample proportions for each combination of `low` and `smoke`

Proportion		low		Total
		0	1	
smoke	0	0.455	0.153	0.608
	1	0.233	0.159	0.392
	Total	0.688	0.312	1

Pearson's χ2 Test of Independence

- Recalling the calculations of expected frequencies using two independent variables, the probability of the intersection of events is the product of their individual probabilities.

- Therefore, for example, the probability that the mother is smoker (i.e., smoke=1) and the baby has low birthweight (i.e., low=1) is the product of smoker and low-birthweight probabilities.

- We use sample proportions to estimate these probabilities.

Pearson's $\chi 2$ Test of Independence

- The proportion of observations with smoke=1 according to last table is 0.392, and the proportion of observations with low=1 is 0.312.

- Therefore, our estimate of the joint probability (under the null) is $0.392 \times 0.312 = 0.122$.

- Consequently, out of 189 babies, we expect $0.122 \times 189 = 23.1$ babies to have smoker mother and have low birthweight if the null hypothesis is true and the two variables are in fact independent.

Pearson's χ2 Test of Independence

- The following table shows the expected frequency of each cell if the null hypothesis was true and the two random variables were independent.
- We denote the expected frequency in row i and column j as Eij .

Expected frequencies for different combinations of low and smoke assuming the null hypothesis is true

Expected frequency		low	
		0	1
smoke	0	79.1	35.9
	1	50.9	23.1

Pearson's χ2 Test of Independence

- If the null hypothesis is true, the observed frequencies would be close to the expected frequencies under the null.
- We therefore use the difference between the observed and expected frequencies as a measure of disagreement between the observed data and what we expected under the null.
- This would be interpreted as evidence against the null hypothesis.
- For this, we use the following general form of Pearson's χ2 test, which summarizes the differences between the expected frequencies (under the null hypothesis) and the observed frequencies over all cells of the contingency table:

$$Q = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

Pearson's χ2 Test of Independence

- Where Oij and Eij are the observed and expected values in the ith row and jth column of the contingency table.
- The double sum simply means that we add the individual measures of discrepancies for cells by going through all cells in the contingency table.
- As before, higher values of Q provide stronger evidence against H0.
- For I × J contingency tables (i.e., I rows and J columns), the Q statistic has approximately the χ 2 distribution with (I − 1) × (J − 1) degrees of freedom under the null.

Pearson's χ2 Test of Independence

- Therefore, we can calculate the observed significance level by finding the upper tail probability of the observed value for Q, which we denote as q, based on the χ2 distribution with (I − 1) × (J − 1) degrees of freedom.
- For the baby weight example, we can summarize the observed and expected frequencies in the contingency tables.

Comparing the observed and expected (under the null hypothesis) frequencies for different combinations of birthweight status and smoking status

	Observed			Expected	
	Normal	Low		Normal	Low
Nonsmoking	86	29	Nonsmoking	79.1	35.9
Smoking	44	30	Smoking	50.9	23.1

Pearson's χ2 Test of Independence

- Then Pearson's test statistic is:

$$Q = \frac{(O_{11} - E_{11})^2}{E_{11}} + \frac{(O_{12} - E_{12})^2}{E_{12}} + \frac{(O_{21} - E_{21})^2}{E_{21}} + \frac{(E_{22} - E_{22})^2}{E_{22}},$$

$$q = \frac{(86 - 79.1)^2}{79.1} + \frac{(29 - 35.9)^2}{35.9} + \frac{(44 - 50.9)^2}{50.9} + \frac{(30 - 23.1)^2}{23.1} = 4.9.$$

Pearson's χ2 Test of Independence

- Because the table has I = 2 rows and J = 2 columns, the approximate null distribution of Q is χ 2 with (2 − 1) × (2 − 1) = 1 degrees of freedom.
- Consequently, the observed p-value is the upper tail probability of 4.9 using the χ2(1) distribution.
- We find pobs = P (Q ≥ 4.9) = 0.026.

```
> .Test

        Pearson's Chi-squared test

data:  .Table
X-squared = 4.9237, df = 1, p-value = 0.02649
```

- Therefore, at the 0.05 significance level (but not at 0.01 level), we can reject the null hypothesis that the mother's smoking status and the baby's birthweight status are independent.

Pearson's χ2 Test of Independence

- In summary we have reviewed how to test a hypothesis regarding the relationship between two categorical variables by summarizing the observed and expected frequencies in contingency tables, calculating the value of Pearson's test statistic, and then finding pobs from χ2 distribution.

Pearson's χ2 Test Using R

- To test a hypothesis regarding the probabilities (population proportions) of different categories for a categorical variable, we can use the chisq.test() function to perform Pearson's χ2 test in R:

````
13 ```{R}
14 chisq.test(x = c(24, 16), p = c(0.7, 0.3))
15 ```
````

```
        Chi-squared test for given probabilities

data:  c(24, 16)
X-squared = 1.9048, df = 1, p-value = 0.1675
```

Pearson's χ2 Test Using R

- The first argument to the chisq.test() function provides the observed frequencies for each possible category.
- Here, there are two categories:
  - The second argument, p, specifies the corresponding probabilities under the null hypothesis.
  - In the output, X-squared provides the observed value of the test statistics (which we denoted Q).

````
13 ```{R}
14 chisq.test(x = c(24, 16), p = c(0.7, 0.3))
15 ```
````

```
        Chi-squared test for given probabilities

data:  c(24, 16)
X-squared = 1.9048, df = 1, p-value = 0.1675
```

Pearson's χ2 Test Using R

- We can also use the chisq.test() function for categorical variables with multiple categories.
- For the heart attack example discussed previously, the null hypothesis was H0 : μ01 = 0.5, μ02 = 0.2, μ03 = 0.3.
- The observed frequencies were O1 = 13, O2 = 11, and O3 = 16.
- Therefore, we can perform Pearson's χ2 test as follows:

````
18 ```{R}
19 chisq.test(x = c(13, 11, 16), p = c(0.5, 0.2, 0.3))
20 ```
````

```
        Chi-squared test for given probabilities

data:  c(13, 11, 16)
X-squared = 4.9083, df = 2, p-value = 0.08593
```

Pearson's χ2 Test Using R

- As before, x provides the number of observations in each group, and p provides the corresponding probability of each group under the null hypothesis.

- To test the relationship between two binary random variables, we use the χ2 test to compare the observed frequencies to the expected frequencies based on the null hypothesis.

- To use chisq.test() for this purpose, we first create the contingency table using the table function and then pass the resulting contingency table to chisq.test().

Pearson's χ2 Test Using R

- For example, the following code creates the contingency table for smoke by low from the birthwt data set and then performs the χ 2 test to examine their relationship:

````
```{R}
# install.packages("MASS")
library(MASS)
birthwt.tab <- table(birthwt$smoke, birthwt$low)
birthwt.tab
```


# Running chisq.test
```{R}
chisq.test(birthwt.tab, correct = FALSE)
```
````

```
        Pearson's Chi-squared test

data:  birthwt.tab
X-squared = 4.9237, df = 1, p-value = 0.02649
```

Pearson's χ2 Test Using R

- Note that we have set the option correct to FALSE to obtain the same results as they were obtained before.

- If we only have the summary of the data in the form of a contingency table as oppose to individual observations, we can enter the contingency table in R and perform the χ2 test as before.

- For example, consider the study investigating the relationship between aspirin intake and the risk of a heart attack.

Pearson's χ2 Test Using R

- We can enter the given contingency table directly in R.

```
# Entering a Contingency table directly in R
```{R}
contTable <- matrix(c(189, 10845, 104, 10933), nrow=2,ncol=2,byrow=TRUE)
rownames(contTable) <- c("Placebo", "Aspirin")
colnames(contTable) <- c("No heart attack", "Heart attack")
contTable
```
```

```
         No heart attack Heart attack
Placebo              189        10845
Aspirin              104        10933
```

Pearson's χ2 Test Using R

- From the last slide's code, the first parameter to the matrix() function is a vector of values.
- We also specify the number of rows (nrow) and the number of columns (ncol).
- The byrow option tells R to fill the matrix by rows.
- We then use the rownames() and colnames() to add names to the rows and columns, respectively.

Pearson's χ2 Test Using R

- To examine the relationship between heart attack and aspirin intake, we use the chisq.test() function as before:

```
# Using chisq.test() function to examine the relationship between two variables
```{R}
output <- chisq.test(contTable, correct = FALSE)
output
```
```

```
        Pearson's Chi-squared test

data:  contTable
X-squared = 25.014, df = 1, p-value = 5.692e-07
```

Pearson's χ2 Test Using R

- The argument to the chisq.test() function is the contingency table of observed values.
- We have assigned the output of the function to a new object called output.
- From this object, we can obtain the observed and expected frequencies with the $ operator:

````
51
52 # Obtaining the observed frequencies with the $ operator
53 ```{R}
54 output$observed
55 ```

         No heart attack Heart attack
Placebo              189        10845
Aspirin              104        10933

56
57 # Obtaining the expected frequencies with the $ operator
58 ```{R}
59 output$expected
60 ```

         No heart attack Heart attack
Placebo         146.4801     10887.52
Aspirin         146.5199     10890.48
````

Exercises

- Import the birthwt dataset from the MASS package, convert ht and low to factors, and use Statistics Contingency tables to examine the relationship between these two categorical variables.

````
72
73 # Examine the relationship between these two categorical variables (ht and low)
74 ```{R}
75 chisq.test(birthHT.tab, correct = FALSE)
76 ```
````

```
Chi-squared approximation may be incorrect
        Pearson's Chi-squared test

data:  birthHT.tab
X-squared = 4.388, df = 1, p-value = 0.03619
```

- The results show that the relationship is statistically significant (*p*-value = 0.036) at 0.05 level so we can reject the null hypothesis.

Exercises

- Create a 4×2 table, enter the frequencies as shown below and run Pearson's χ2 test.

Snoring Severity	Heart Disease	Total
Never	24	1379
Occasionally	35	638
Nearly every night	21	213
Every night	30	254

Exercises

```
# Creating a 4x2 tab.
```{R}
contTable <- matrix(c(24, 1379, 35, 638, 21, 213, 30, 254), nrow=4,ncol=2, byrow=TRUE)
rownames(contTable) <- c("Never", "Occasionally", "Nearly every night", "Every night")
colnames(contTable) <- c("Heart Disease", "Total")
contTable
```
```

```
                   Heart Disease Total
Never                         24  1379
Occasionally                  35   638
Nearly every night            21   213
Every night                   30   254
```

```
# Running Pearson's χ2 test in the 4x2 tab.
```{R}
chisq.test(contTable)
```
```

```
        Pearson's Chi-squared test

data:  contTable
X-squared = 64.515, df = 3, p-value = 6.37e-14
```

- The results of Pearson's χ2 test of independence show that the relationship between snoring and heart disease is statistically significant (*p*-value = 6.37×10^{-14}).

Fisher's Exact Test

- We will discuss Fisher's exact test for analyzing contingency tables from small data sets.

- For Person's χ 2 test to be valid, the expected frequencies (Eij) under the null should be at least 5, so we can assume that the distribution of Q is approximately $\chi 2$ under the null.

- Occasionally, this requirement is violated (especially when the sample size is small, or the number of categories is large, or some of the categories are rare), and some of the expected frequencies become small (less than 5).

Fisher's Exact Test

- If you have a 4 × 2 contingency table similar to the one below (Note that there are only two underweight women in this sample. (This seems to be a rare event in the population) the expected frequencies E1,1 = 1.32 and E12 = 0.68 (The remaining expected frequencies are above 5.)
- If use R for this data, R will give a warning message indicating that "2 expected frequencies are less than 5".
- In this case, instead of using Pearson's χ2 test (which assumes that Q statistic has an approximate χ 2 distribution), we should use Fisher's exact test (which is based on the exact distribution of a test statistic that captures the deviation from the null).

Contingency table of diabetes status by weight status

Frequency	type		Total
	No	Yes	
weight.status			
Underweight	2	0	2
Normal	21	2	23
Overweight	35	8	43
Obese	74	58	132
Total	132	68	200

Fisher's Exact Test

- If we use R, creating a contingency table for weight.status and type, but this time, select Fisher's exact test, instead of the default option Chi-square test of independence, under the Hypothesis tests. The resulting p-value would be 0.0002, which is slightly lower than the p-value of 0.0004 based on $\chi 2$ test.
- At 0.01 level, we can reject the null hypothesis, which indicates that the disease status is independent from the weight status, and conclude that the relationship between the two variables is statistically significant.

Contingency table of diabetes status by weight status

Frequency	type		Total
	No	Yes	
weight.status			
Underweight	2	0	2
Normal	21	2	23
Overweight	35	8	43
Obese	74	58	132
Total	132	68	200

Fisher's Exact Test

- When examining the relationship between weight.status and type, R give the warning message: "Chi-squared approximation may be incorrect"

```
92
93  # Using Fisher exact test
94  # weight.status and type
95
96  ```{R}
97  contTable <- matrix(c(2, 0, 21, 2, 35, 8, 74, 58), nrow=4,ncol=2, byrow=TRUE)
98  rownames(contTable) <- c("Underweight", "Normal", "Overweight", "Obese")
99  colnames(contTable) <- c("No", "Yes")
100 contTable
101 ```
```

```
            No Yes
Underweight  2   0
Normal      21   2
Overweight  35   8
Obese       74  58
```

```
102
103 # Running Pearson's χ2 test.
104 ```{R}
105 chisq.test(contTable)
106 ```
```

```
Chi-squared approximation may be incorrect
        Pearson's Chi-squared test

data:  contTable
X-squared = 17.946, df = 3, p-value = 0.0004512
```

Fisher's Exact Test

- Therefore, Fischer's exact test is more appropriate for analyzing the contingency table.

```
107
108 # Running Fisher's exact test.
109 ```{R}
110 fisher.test(contTable)
111 ```
```

```
        Fisher's Exact Test for Count Data

data:  contTable
p-value = 0.0001842
alternative hypothesis: two.sided
```

- The resulting p-value is 0.0001, which is slightly lower than the p-value of 0.0004 based on $\chi 2$ test. At 0.01 level, we can reject the null hypothesis, which indicates that the disease status is independent from the weight status, and conclude that the relationship between the two variables is statistically significant.