STAT234: Lecture 8 - Analyzing Tables

Kushal K. Dey

Tables

Early in this course, we talked about contingency tables.

Tables

Early in this course, we talked about contingency tables.

An example - Yawning Contagious

Tables

Early in this course, we talked about contingency tables.

An example - Yawning Contagious

```
yawn.data <- matrix(c(10,4,24,12), nrow=2, byrow=TRUE,
                dimnames = list(c("Yawned", "No Yawn"),
                                c("Yawn Seed", "No Seed")))
thetable <-  addmargins(yawn.data)
thetable

        Yawn Seed No Seed Sum
Yawned         10       4  14
No Yawn        24      12  36
Sum            34      16  50
```

Is there any effect of planting yawn seed on yawning?

Tables

We already solved this problem before....

Tables

We already solved this problem before....

But that was using shuffles, similar to resampling or bootstrapping concept.

Tables

We already solved this problem before....

But that was using shuffles, similar to resampling or bootstrapping concept.

How about using a theoretical model?

Tables

We already solved this problem before....

But that was using shuffles, similar to resampling or bootstrapping concept.

How about using a theoretical model?

Our focus today - model based inference of tables.

Tables

For the 34 cases where yawn seed is planted, define random variables $X_1, X_2, \cdots, X_{34}$ such that

$$X_i = 1 \;\; \textit{if person yawns} \quad (1)$$
$$= 0 \;\; \textit{if person does not yawn} \quad (2)$$
$$(3)$$

Let p_1 be the probability

$$p_1 = Pr[\textit{a person yawns} | \textit{yawn seed is planted}]$$

then,

$$X_i \sim Bin(34, p_1)$$

The conditionality of yawn seed planted is there because all the X_i's are generated conditional on yawn seed being planted.

Tables

For the 16 cases where yawn seed is not planted, define random variables $Y_1, Y_2, \cdots, Y_{16}$ such that

$$
\begin{aligned}
Y_i &= 1 \;\; \textit{if person yawns} && (4)\\
&= 0 \;\; \textit{if person does not yawn} && (5)\\
& && (6)
\end{aligned}
$$

Let p_2 be the probability

$$p_2 = Pr[\textit{a person yawns}|\textit{yawn seed is not planted}]$$

then,

$$Y_i \sim Bin(16, p_2)$$

The conditionality of yawn seed planted is there because all the Y_i's are generated conditional on yawn seed being planted.

Tables

We reduce a contingency table to a 2-sample problem

$$X_1, X_2, \cdots, X_{34} \sim Bin(34, p_1)$$

$$Y_1, Y_2, \cdots, Y_{16} \sim Bin(16, p_2)$$

Tables

We reduce a contingency table to a 2-sample problem

$$X_1, X_2, \cdots, X_{34} \sim Bin(34, p_1)$$

$$Y_1, Y_2, \cdots, Y_{16} \sim Bin(16, p_2)$$

What is the hypothesis?

$$H_0 : p_1 = p_2 \quad H_1 : p_1 > p_2$$

Tables

We reduce a contingency table to a 2-sample problem

$$X_1, X_2, \cdots, X_{34} \sim Bin(34, p_1)$$

$$Y_1, Y_2, \cdots, Y_{16} \sim Bin(16, p_2)$$

What is the hypothesis?

$$H_0 : p_1 = p_2 \quad H_1 : p_1 > p_2$$

$$\hat{p}_1 = \frac{10}{34} \quad \hat{p}_2 = \frac{4}{16}$$

Tables

We reduce a contingency table to a 2-sample problem

$$X_1, X_2, \cdots, X_{34} \sim Bin(34, p_1)$$

$$Y_1, Y_2, \cdots, Y_{16} \sim Bin(16, p_2)$$

What is the hypothesis?

$$H_0 : p_1 = p_2 \quad H_1 : p_1 > p_2$$

$$\hat{p}_1 = \frac{10}{34} \quad \hat{p}_2 = \frac{4}{16}$$

$$\hat{p}_1 - \hat{p}_2 = \frac{10}{34} - \frac{4}{16} = 0.044$$

Tables

You know

$$E(\hat{p}_1 - \hat{p}_2) = E(\hat{p}_1) - E(\hat{p}_2) = p_1 - p_2$$

Tables

You know

$$E(\hat{p}_1 - \hat{p}_2) = E(\hat{p}_1) - E(\hat{p}_2) = p_1 - p_2$$

But to make inference on $p_1 - p_2$, we need to know the distribution of $\hat{p}_1 - \hat{p}_2$.

We know under **large sample assumption**

$$\hat{p}_1 \sim N(p_1, \frac{p_1(1-p_1)}{34})$$

$$\hat{p}_2 \sim N(p_2, \frac{p_2(1-p_2)}{16})$$

Note that X_i and Y_j are all independent, hence so are $\hat{p}_1$ and $\hat{p}_2$.

$$\hat{p}_1 - \hat{p}_2 \sim N\left(p_1 - p_2, \frac{p_1(1-p_1)}{34} + \frac{p_2(1-p_2)}{16}\right)$$

Tables

Under H_0, we have

$$p_1 = p_2 = p(say)$$

$$\hat{p}_1 - \hat{p}_2 \sim N\left(0, \frac{p(1-p)}{34} + \frac{p(1-p)}{16}\right)$$

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{p(1-p)}{34} + \frac{p(1-p)}{16}}} \sim N(0,1)$$

We replace p by $\hat{p}$,

$$\hat{p} = \frac{10+4}{36+14} = \frac{14}{50}$$

Tables

under large sample assumption

$$T = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{34} + \frac{1}{16})}} \sim N(0,1)$$

We find the realized value of this quantity from the sample after substituting the values of $\hat{p}_1 = \frac{10}{36}$, $\hat{p}_2 = \frac{4}{16}$ and $\hat{p} = \frac{14}{50}$,

Tables

under large sample assumption

$$T = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{34} + \frac{1}{16})}} \sim N(0,1)$$

We find the realized value of this quantity from the sample after substituting the values of $\hat{p}_1 = \frac{10}{36}$, $\hat{p}_2 = \frac{4}{16}$ and $\hat{p} = \frac{14}{50}$,

$$t = 0.275$$

Tables

under large sample assumption

$$T = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{34} + \frac{1}{16})}} \sim N(0,1)$$

We find the realized value of this quantity from the sample after substituting the values of $\hat{p}_1 = \frac{10}{36}$, $\hat{p}_2 = \frac{4}{16}$ and $\hat{p} = \frac{14}{50}$,

$$t = 0.275$$

The p-value of this quantity

$$Pr\left(Z > t\right) = 0.4 >> 0.05$$

Tables

under large sample assumption

$$T = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{34} + \frac{1}{16})}} \sim N(0,1)$$

We find the realized value of this quantity from the sample after substituting the values of $\hat{p}_1 = \frac{10}{36}$, $\hat{p}_2 = \frac{4}{16}$ and $\hat{p} = \frac{14}{50}$,

$$t = 0.275$$

The p-value of this quantity

$$Pr\,(Z > t) = 0.4 >> 0.05$$

We reject researchers' claim !!

Revisit Bootstrapping/ Resampling

We generate $N\hat{p} = 50 \times \frac{14}{50} = 14$ cards which are red (yawners) and $N \times (1 - \hat{p}) = 36$ cards which are black (non-yawners).

Revisit Bootstrapping/ Resampling

We generate $N\hat{p} = 50 \times \frac{14}{50} = 14$ cards which are red (yawners) and $N \times (1 - \hat{p}) = 36$ cards which are black (non-yawners).

Shuffle the pile of cards.

Revisit Bootstrapping/ Resampling

We generate $N\hat{p} = 50 \times \frac{14}{50} = 14$ cards which are red (yawners) and $N \times (1 - \hat{p}) = 36$ cards which are black (non-yawners).

Shuffle the pile of cards.

Take 34 cards from the pile and calculate the number of red cards (these are my X)

Revisit Bootstrapping/ Resampling

We generate $N\hat{p} = 50 \times \frac{14}{50} = 14$ cards which are red (yawners) and $N \times (1 - \hat{p}) = 36$ cards which are black (non-yawners).

Shuffle the pile of cards.

Take 34 cards from the pile and calculate the number of red cards (these are my X)

Take 16 cards from remaining pile and calculate the number of red cards (these are my Y).

Revisit Bootstrapping/ Resampling

We generate $N\hat{p} = 50 \times \frac{14}{50} = 14$ cards which are red (yawners) and $N \times (1 - \hat{p}) = 36$ cards which are black (non-yawners).

Shuffle the pile of cards.

Take 34 cards from the pile and calculate the number of red cards (these are my X)

Take 16 cards from remaining pile and calculate the number of red cards (these are my Y).

Revisit Bootstrapping/ Resampling

Calculate proportion of red cards in each of th two draws $\hat{p}_1$ and $\hat{p}_2$ anc calculate $\hat{p}_1 - \hat{p}_2$.

Revisit Bootstrapping/ Resampling

Calculate proportion of red cards in each of th two draws $\hat{p}_1$ and $\hat{p}_2$ anc calculate $\hat{p}_1 - \hat{p}_2$.

Repeat this 1000 times and record the number of times we get the value of $\hat{p}_1 - \hat{p}_2$ greater than 0.044. Thats my p-value

Revisit Bootstrapping/ Resampling

Calculate proportion of red cards in each of th two draws $\hat{p}_1$ and $\hat{p}_2$ anc calculate $\hat{p}_1 - \hat{p}_2$.

Repeat this 1000 times and record the number of times we get the value of $\hat{p}_1 - \hat{p}_2$ greater than 0.044. Thats my p-value

This is equivalent to construct the histogram of $\hat{p}_1 - \hat{p}_2$ values from the 10000 draws and counting the frequencies above 0.044

Summary

We saw how one can test for effectiveness of a hypothesis from contingency table.

Summary

We saw how one can test for effectiveness of a hypothesis from contingency table.

The Normal test is only applicable for large number of samples

Summary

We saw how one can test for effectiveness of a hypothesis from contingency table.

The Normal test is only applicable for large number of samples

The Bootstrap/Resampling test does not assume distribution but will give different results over different runs

Summary

We saw how one can test for effectiveness of a hypothesis from contingency table.

The Normal test is only applicable for large number of samples

The Bootstrap/Resampling test does not assume distribution but will give different results over different runs

But this was for a 2×2 table. Can we have a test for a table with 3 or more categories?

Summary

We saw how one can test for effectiveness of a hypothesis from contingency table.

The Normal test is only applicable for large number of samples

The Bootstrap/Resampling test does not assume distribution but will give different results over different runs

But this was for a 2×2 table. Can we have a test for a table with 3 or more categories?

Chi-square table

```
yawn.data <- matrix(c(10,4,24,12), nrow=2, byrow=TRUE,
                dimnames = list(c("Yawned", "No Yawn"),
                                c("Yawn Seed", "No Seed")))
O <-  addmargins(yawn.data)
O
```

```
        Yawn Seed No Seed Sum
Yawned         10       4  14
No Yawn        24      12  36
Sum            34      16  50
```

Chi-square table

```
yawn.data <- matrix(c(10,4,24,12), nrow=2, byrow=TRUE,
                  dimnames = list(c("Yawned", "No Yawn"),
                                  c("Yawn Seed", "No Seed")))
O <-  addmargins(yawn.data)
O

        Yawn Seed No Seed Sum
Yawned         10       4  14
No Yawn        24      12  36
Sum            34      16  50
```

What would be expected counts under the null, if we assume $p_1 = p_2 = p$, we should have

	Yawned	No Yawn	Total
Yawn Seed	$34 \times p$	$16 \times p$	$50 \times p$
No seed	$34 \times (1-p)$	$16 \times (1-p)$	$50 \times (1-p)$
Total	34	16	50

Chi-square table

Since we do not know p, we replace it by $\hat{p}$.

$$\hat{p} = 14/50$$

Chi-square table

Since we do not know p, we replace it by $\hat{p}$.

$$\hat{p} = 14/50$$

The table under the null expected then is

```
yawn.data <- matrix(c(9.52,4.48,24.48,11.52), nrow=2, byrow=TRUE
                dimnames = list(c("Yawned", "No Yawn"),
                                c("Yawn Seed", "No Seed")))
E <-  addmargins(yawn.data)
E
```

```
        Yawn Seed No Seed Sum
Yawned       9.52    4.48  14
No Yawn     24.48   11.52  36
Sum         34.00   16.00  50
```

Chi-square table

Take the observed table O and the expected table E. If the null hypothesis is not true, we will expect the squared differences

$$(O_{ij} - E_{ij})^2$$

to be high (i=1,2 , j=1,2).

Chi-square table

Take the observed table O and the expected table E. If the null hypothesis is not true, we will expect the squared differences

$$(O_{ij} - E_{ij})^2$$

to be high (i=1,2 , j=1,2).
Lets consider the expression

$$X = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

We need distribution for X. This distribution has a fancy name - **Chi-square distribution** with degrees of freedom $(r-1) \times (c-1)$ where r is the number of rows in table and c is the number of columns.

$$X \sim \chi^2_{(2-1)(2-1))} = \chi^2_1$$

Chi-Square Distribution

The chi-square distributions: χ_k^2

$$f(x) = \frac{x^{(k/2)-1} e^{-x/2}}{2^{k/2}\Gamma(k/2)} \qquad x > 0,\ k \geq 1$$

Chi-Square Distribution

The chi-square distributions: χ^2_k

$$f(x) = \frac{x^{(k/2)-1} e^{-x/2}}{2^{k/2}\Gamma(k/2)} \qquad x > 0,\ k \geq 1$$

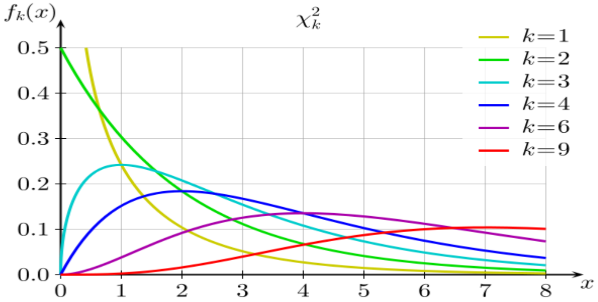
$$E(X) = k \qquad Var(X) = 2k$$

Chi-Square Distribution

The chi-square distributions: χ_k^2

$$f(x) = \frac{x^{(k/2)-1} e^{-x/2}}{2^{k/2}\Gamma(k/2)} \qquad x > 0,\ k \geq 1$$

$$E(X) = k \qquad Var(X) = 2k$$



Chi-Square Distribution

The chi-square distributions: χ_1^2

$$f(x) = \frac{x^{-1/2} e^{-x/2}}{\sqrt{2\pi}} \qquad x > 0$$

Chi-Square Distribution

The chi-square distributions: χ_1^2

$$f(x) = \frac{x^{-1/2} e^{-x/2}}{\sqrt{2\pi}} \qquad x > 0$$

Origin of the Chi-square distribution:
Let $Z \sim N(0,1)$ and define $Y = Z^2$. Then, $Y \sim \chi_1^2$.

Chi-Square Distribution

The chi-square distributions: χ_1^2

$$f(x) = \frac{x^{-1/2} e^{-x/2}}{\sqrt{2\pi}} \qquad x > 0$$

Origin of the Chi-square distribution:
Let $Z \sim N(0,1)$ and define $Y = Z^2$. Then, $Y \sim \chi_1^2$.

In general, χ_n^2 is a sum of n i.i.d. standard normal distributions.