# STAT 333
# Chapter 4
# Contingency Tables

• Contingency tables are summaries (in matrix form) of categorical data, where the entries in the table are counts of how many observations fell into specific categories (and combinations of categories).

• A one-way contingency table summarizes data on a single categorical variable and has only one row.

• A two-way contingency table summarizes data on two categorical variables and may have several rows and several columns.

• Data on several categorical variables can be summarized by multi-way contingency tables.

# Section 4.1: Tests for 2 × 2 Tables

• Consider the simplest form of two-way table:

2x2 table (2 rows , 2 columns)

• Such a table could summarize data arising from

- Having a single sample in which two binary variables are measured on each individual
- Having two samples in which the same binary variable is measured on each individual in each sample.

## Comparing Two Probabilities, Independent Samples

• Suppose we have two independent samples, with respective sizes $n_1$ and $n_2$. We classify each individual in each sample into class 1 or class 2.

• Our data could be arranged in a 2 × 2 table as follows:

| | Class 1 | Class 2 | Total |
|---|---|---|---|
| Sample from population 1 | $O_{11}$ | $O_{12}$ | $n_1$ |
| Sample from population 2 | $O_{21}$ | $O_{22}$ | $n_2$ |
| Total | $C_1$ | $C_2$ | N |

• The total number of observations is $N = n_1 + n_2$.

• Our goal is to compare the probability of "success" (Class 1) across the two populations:

$P_1$=Proportion an observation from population 1 will be in class 1.
$P_2$=Proportion an observation from population 2 will be in class 1.

**Hypotheses:**

| Two -tailed | Lower -tailed | Upper-tailed |
|---|---|---|
| $H_0$: $p_1 = p_2$<br>$H_1$: $p_1 \neq p_2$ | $H_0$: $p_1 \geq p_2$<br>$H_1$: $p_1 < p_2$ | $H_0$: $p_1 \leq p_2$<br>$H_1$: $p_1 > p_2$ |

## Development of the Test Statistic

As estimators of $p_1$ and $p_2$, we have:

$$\hat{p}_1 = \frac{O_{11}}{n_1} \quad , \quad \hat{p}_2 = \frac{O_{21}}{n_2} \quad (\, n_1 = O_{11} + O_{12} \;\; , n_2 = O_{21} + O_{22})$$

$$\hat{p}_1 - \hat{p}_2 = \frac{O_{11}}{n_1} - \frac{O_{21}}{n_2} = \frac{O_{11}n_2 - O_{21}n_1}{n_1 n_2} = \frac{O_{11}(O_{21}+O_{22}) - O_{21}(O_{11}+O_{12})}{n_1 n_2}$$

$$= \frac{O_{11}O_{21} + O_{11}O_{22} - O_{21}O_{11} - O_{21}O_{12}}{n_1 n_2} = \frac{O_{11}O_{22} - O_{21}O_{12}}{n_1 n_2}$$

• This estimates how far apart $p_1$ and $p_2$ are.

• Scaling this by dividing by the estimated standard error (see Eq. 5, p. 187), we get the test statistic

$$T_1 = \frac{\sqrt{N}(O_{11}O_{22} - O_{21}O_{12})}{\sqrt{n_1 n_2 c_1 c_2}}$$

which has a standard normal distribution for large samples when $H_0$ is true.

• If $T_1$ is far from zero, this indicates that $p_1 \neq p_2$

• If $T_1$ is far below zero, this indicates that $p_1 < p_2$

• If $T_1$ is far above zero, this indicates that $p_1 > p_2$

Decision Rules

| Alternative hypothesis | **$H_1$:** $p_1 \neq p_2$ | **$H_1$:** $p_1 < p_2$ | **$H_1$:** $p_1 > p_2$ |
|---|---|---|---|
| Reject $H_0$ if | $\lvert T_1 \rvert > Z_{1-\frac{\alpha}{2}}$ | $T_1 < -Z_{1-\alpha}$ | $T_1 > Z_{1-\alpha}$ |
| P-value | $2[min\{P(Z \leq T_1^{obs}), P(Z \geq T_1^{obs})\}]$ | $P(Z \leq T_1^{obs})$ | $P(Z \geq T_1^{obs})$ |

Note : • In all cases, reject $H_0$ if the p-value $\leq \alpha$

• Note: The normal approximation for $T_1$ is valid for large samples, say, if

Each of $O_{11}$ , $O_{12}$ , $O_{21}$, $O_{22}$ are at least 5

**Example 1:**

A survey was conducted of 160 rural households and 261 urban households with Christmas trees. Of interest was whether the tree was natural or artificial. Is the probability of natural trees different for rural and urban households? Use α = 0.05.

**Data:**

Tree

| Population | Natural | Artificial | Total |
|---|---|---|---|
| Rural | 64 | 96 | 160 |
| Urban | 89 | 172 | 261 |
| Total | 153 | 268 | 421 |

**Hypothesis:**

**H₀:** $p_1 = p_2$ **H₁:** $p_1 \neq p_2$

**Test statistic:**

$$T_1 = \frac{\sqrt{N}(O_{11}O_{22} - O_{21}O_{12})}{\sqrt{n_1 n_2 c_1 c_2}}$$

| | Natural | Artificial | Total |
|---|---|---|---|
| Rural | 64=$O_{11}$ | 96=$O_{12}$ | 160=$n_1$ |
| Urban | 89=$O_{21}$ | 172=$O_{22}$ | 261=$n_2$ |
| Total | 153=$c_1$ | 268=$c_2$ | 421=N |

$$T_1 = \frac{\sqrt{421}\,(64x172 - 89x96)}{\sqrt{160x261x153x268}} = 1.22$$

**Reject H₀ if** $|T_1| > Z_{1-\frac{\alpha}{2}}$

**Since** $|1.22| > 1.96$

$1.22 \not> 1.96$ (condition not satisfies)

where α=0.05
1-α/2=0.975
$Z_{1-\frac{\alpha}{2}} = Z_{0.975}$
$= 1.96$

**Decision** : Fail to reject $H_0$

**Conclusion:** Can't conclude that, the probability of natural tree differs for urban and rural households.

**P-value** $=2[min\{P(Z < T_1^{obs}), P(Z > T_1^{obs})\}]$

$= 2[\min\{P(Z < 1.22), P(Z > 1.22)\}]$

$= 2[\min\{0.88877, 0.11123\}] = 2x0.11123 = 0.22246 > \alpha$ (Accept $H_0$)

**Example 2:**

Page 184 gives data from a study to determine whether a new lighting system worsened midshipmen's vision.

**Data:**

Vision

| Lighting | Good | Poor | Total |
|---|---|---|---|
| Old | 714 | 111 | 825 |
| New | 662 | 154 | 816 |
| Total | 1376 | 265 | 1641 |

**Hypothesis:**

**$H_0$:** $p_1 \leq p_2$ **$H_1$:** $p_1 > p_2$

**Test statistic:**

$$T_1 = \frac{\sqrt{N}(O_{11}O_{22} - O_{21}O_{12})}{\sqrt{n_1 n_2 c_1 c_2}}$$

| | Good | Poor | Total |
|---|---|---|---|
| Old | 714=$O_{11}$ | 111=$O_{12}$ | 825=$n_1$ |
| New | 662=$O_{21}$ | 154=$O_{22}$ | 816=$n_2$ |
| Total | 1376=$c_1$ | 265=$c_2$ | 1641=N |

$$T_1 = \frac{\sqrt{1641}\,(714x154 - 662x111)}{\sqrt{825x816x1376x265}} = 2.982$$

**Reject $H_0$ if** $T_1 > Z_{1-\alpha}$

**Since** $2.982 > 1.645$

(condition satisfies)

where $\alpha=0.05$

$1-\alpha=0.95$

$Z_{1-\alpha} = Z_{0.95}$

$= 1.645$

**Decision** : Reject $H_0$

**Conclusion:** Conclude that, the old lighting produced a better of good vision that new lighting.

**P-value=**

$P(Z > T_1^{obs}) = P(Z > 2.89) = 0.00193$

## Fisher's Exact Test

• In the previous examples, the row totals were the sizes of the two samples, which are fixed before the data are examined (i.e., they are not random).

• When we have a single sample in which two binary variables are measured on each individual, the resulting 2 × 2 table has random row totals and random column totals.
• We will cover that scenario in Section 4.2.

• In other situations, both the row totals and the column totals may be fixed prior to the data being examined.

• In this case of "fixed margins", Fisher's Exact Test is ideal.

Data setup:

| | Column 1 | Column2 | Total |
|---|---|---|---|
| Row 1 | x | r-x | r |
| Row2 | c-x | N-r-c+x | N-r |
| Total | c | N-c | N |

• We again wish to compare:

$P_1$=Probability of an observation in row1 being classified into column 1.
$P_2$=Probability of an observation in row 2 being classified into column 1.

## Test statistic

$T_2$ = x = number of observations in (1,1) cell

## Null Distribution

• Let $p$ = probability an observation is in Column 1.
• Under $H_0$, this probability is the same whether the observation is in Row 1 or Row 2. Then:

P(table results | row totals) = $\binom{r}{x}\binom{N-r}{c-x}p^c(1-p)^{N-c}$

P(column totals) = $\binom{N}{c}p^c(1-p)^{N-c}$

→ P(table results | row totals & column totals) =

$$\frac{\binom{r}{x}\binom{N-r}{c-x}p^c(1-p)^{N-c}}{\binom{N}{c}p^c(1-p)^{N-c}} = \frac{\binom{r}{x}\binom{N-r}{c-x}}{\binom{N}{c}}$$

• The decision is based on the P-value, which is found differently depending on the alternative hypothesis:

| Alternative hypothesis | **$H_1$:** $p_1 \neq p_2$ | **$H_1$:** $p_1 < p_2$ | **$H_1$:** $p_1 > p_2$ |
|---|---|---|---|
| Reject $H_0$ if | $\lvert T_1 \rvert > Z_{1-\frac{\alpha}{2}}$ | $T_1 < -Z_{1-\alpha}$ | $T_1 > Z_{1-\alpha}$ |
| P-value | $2[min\{P(T_2 \leq T_2^{obs}), P(T_2 \geq T_2^{obs})\}]$ | $P(T_2 \leq T_2^{obs})$ | $P(T_2 \geq T_2^{obs})$ |

• In all cases, reject $H_0$ if the p-value $\leq \alpha$

**Example 3:**
Fourteen new hires (10 male and 4 female) are being assigned to bank positions (there are 4 account representative positions open and 10 (less desirable) teller positions open. The data on page 190 summarize the assignments. If all new employees are equally qualified, is there evidence that female hires were more likely to get the account representative jobs?

**Data:**

| | Account rep. | Teller | Total |
|---|---|---|---|
| Male | 1 | 9 | 10 |
| Female | 3 | 1 | 4 |
| Total | 4 | 10 | 14 |

**Hypothesis:**
**$H_0$:** $p_1 \geq p_2$ **$H_1$:** $p_1 < p_2$

**Test statistic:** $T_2^{obs} = x = 1$

| | Account rep. | Teller | Total |
|---|---|---|---|
| Male | x | r-x | r |
| Female | c-x | N-r-c+x | N-r |
| Total | c | N-c | N |

**P-value:**

$$P(T_2 \leq T_2^{obs}) = P(T_2 \leq 1) = P(T_2 = 0) + P(T_2 = 1)$$

$$= \frac{\binom{10}{0}\binom{4}{4-0}}{\binom{14}{4}} + \frac{\binom{10}{1}\binom{4}{4-1}}{\binom{14}{4}} = \frac{\binom{10}{0}\binom{4}{4}}{\binom{14}{4}} + \frac{\binom{10}{1}\binom{4}{3}}{\binom{14}{4}}$$

$$= \frac{1x1}{1001} + \frac{10x4}{1001} = \frac{41}{1001} = 0.041$$

**Decision :** P-value < α (Reject $H_0$)

**Conclusion :** Female hires more likely to get account representative positions

• See fisher.test function in R to perform this test.

• Fisher's Exact Test may be used if the row totals and/or column totals are random, but in this case it is _more conservative than the z-test.

• Fisher's Exact Test can also be viewed as an alternative to the z-test when the large-sample rule is not met, but the Exact Test lacks power when the sample size is very small.

• Suppose we have several related (but not identical) conditions in which sub-experiments are conducted, each of which produces a $2 \times 2$ table.

• It is of interest to see whether rows and columns are independent in each table.

## Mantel-Haenszel Test

• We assume we have $k \geq 2$ such $2 \times 2$ tables, each with fixed row and column totals (although the test can be done even with random totals).

Let $p_{1i}$ = Probability of an observation in row 1 being classified into column 1,in the i-th table.

And $p_{2i}$ = Probability of an observation in row 2 being classified into column 1,in the i-th table.

**Hypotheses:**

| | | |
|---|---|---|
| **$H_0$:** $p_{1i} = p_{2i}$ for all i=1,2…k | **$H_0$:** $p_{1i} \geq p_{2i}$ for all i=1,2..k | **$H_0$:** $p_{1i} \leq p_{2i}$ for all i=1,2…k |
| **$H_1$:** Either $p_{1i} > p_{2i}$ for some i Or either $p_{1i} < p_{2i}$ for some i , but not both | **$H_1$:** $p_{1i} \leq p_{2i}$ for all i, and $p_{1i} < p_{2i}$ for some i | **$H_1$:** $p_{1i} \geq p_{2i}$ for all i , and $p_{1i} > p_{2i}$ for some i |

**Test statistic:**

$$T_4 = \frac{\sum_{i=1}^{k} x_i - \sum_{i=1}^{k} \frac{r_i c_i}{N_i}}{\sqrt{\sum_{i=1}^{k} \frac{r_i c_i (N_i - r_i)(N_i - c_i)}{{N_i}^2 (N_i - 1)}}}$$

• The null distribution is approximately standard normal, tabulated in Table A1.
Decision Rules and P-value:

| | Two- tailed | Lower-tailed | Upper-tailed |
|---|---|---|---|
| Reject $H_0$ if | $T_4 > Z_{1-\frac{\propto}{2}}$ $or$ $T_4 < -Z_{1-\frac{\propto}{2}}$ | $T_4 < -Z_{1-\propto}$ | $T_4 > Z_{1-\propto}$ |
| P-value | $2[min\{P(Z \leq T_4^{obs}), P(Z \geq T_4^{obs})\}]$ | $P(Z \leq T_4^{obs})$ | $P(Z \geq T_4^{obs})$ |

**Example 4:**

Three groups of cancer patients were given either a drug treatment or a control, and for each patient, whether the outcome was successful was recorded. Is there evidence that in at least one group, the treatment produces a better chance of success than the control? (Use $\alpha = 0.05$.)

**Data:**

| | Group 1 | | Group 2 | | Group 3 | |
|---|---|---|---|---|---|---|
| | Success | Failure | Success | Failure | Success | Failure |
| Treatment | 10 = $x_1$ | 1 | 9 =$x_2$ | 0 | 8= $x_3$ | 0 |
| Control | 12 | 1 | 11 | 1 | 7 | 3 |
| Total | 24=$N_1$ | | 21=$N_2$ | | 18=$N_3$ | |

**Hypothesis:**

$H_0$: $p_{1i} \leq p_{2i}$ for all i=1,2…k $H_1$: $p_{1i} > p_{2i}$ and $p_{1i} \geq p_{2i}$ for all i=1,2…k

**Test statistic:** ( $r_i$ =sum of row i , $c_i$ =sum of column i)

$$T_4 = \frac{\sum_{i=1}^{k} x_i - \sum_{i=1}^{k} \frac{r_i c_i}{N_i}}{\sqrt{\sum_{i=1}^{k} \frac{r_i c_i (N_i - r_i)(N_i - c_i)}{{N_i}^2 (N_i - 1)}}}$$

$$= \frac{(10+9+8) - \left(\frac{11x22}{24} + \frac{9x20}{21} + \frac{8x15}{18}\right)}{\sqrt{\frac{11x22x(24-11)x(24-22)}{24^2x(24-1)} + \frac{9x20x(21-9)x(21-20)}{21^2x(21-1)} + \frac{8x15x(18-8)x(18-15)}{18^2x(18-1)}}}$$

$$= \frac{27 - 25.3214}{\sqrt{\frac{11132}{24^2x23} + \frac{2160}{21^2x20} + \frac{3600}{18^2x17}}} = \frac{1.6786}{1.31862} = 1.272997$$

$T_4$= 1.0057 (R reports $T_4^2$)

**P-value: =**

$$P\left(Z \geq T_4^{obs}\right) = P(Z \geq 1.01) = 1 - P(Z < 1.01) = 1 - 0.84375 = 0.15625$$

P-value > α (accept $H_0$)

**Conclusion:**

There is no evidence that the success probability is better for treatment than for control. (in any group)

• See mantelhaen.test function in R to perform this test.