TESTS FOR CATEGORICAL DATA

ONE-SAMPLE TEST FOR A BINOMIAL PROPORTION

$H_0: p = p_0$ vs. $H_0: p \neq p_0$

Bernoulli trials: 0, 1, 0, 0, 1, ... - independent trials $Pr\{x=1\}=p$
Number of successes in a series of $n$ trials - Binomial distribution
mean = $np$, variance = $np(1-p)$

Proportion is the mean number of successes
Sample mean is normally distributed => z-test

We can use the normal approximation if $np_0q_0 > 5$.  

$q_0 = 1 - p$
Under the null hypothesis

\[ \hat{p} = \frac{1}{n} \sum_{i} x_i = \frac{\# \text{ successes}}{\# \text{ trials}} \]

\[ x_i = \begin{cases} 1 & \text{success} \\ 0 & \text{failure} \end{cases} \]

\[ \text{E}(\hat{p}) = \frac{n \cdot p_0}{n} = p_0 \quad \text{Under } H_0 \]

\[ \text{Var}(\hat{p}) = \frac{1}{n^2} \cdot \text{Var}(\# \text{ successes}) = \frac{n \cdot p_0 (1 - p_0)}{n^2} = \frac{p_0 (1 - p_0)}{n} \]

\[ \hat{p} \mid_{H_0} \sim \text{N}(p_0, \frac{p_0 (1 - p_0)}{n}) \]
Why z-test rather than t-test?

- Binomial distribution is approximated by normal distribution.
- Binomial distribution has only one parameter, Pr{success}.
- Variance is pre-determined by the mean.
- T-test is associated with estimation of variance - not needed here.
Cardiovascular Disease Example

Suppose the incidence rate of MI per year was 5 per 1000 among 45-54-year-old males in 1970.

5000 45-54-year-old men were followed for 1-year starting 1980. Fifteen new cases of MI were found.

Did the incidence rates of MI change from 1970 to 1980?

\[ H_0: p = \frac{5}{1000} = 0.005 \]
\[ H_1: p \neq 0.005 \]

\[ np_0q_0 = 5000 \times 0.005 \times 0.995 = 24.88 > 5 \]

\[ \hat{p} = \frac{15}{5000} = 0.003 \]

=> Use the normal approximation method
Conclusion: The incidence rate of MI among 45-54-year-old males in 1980 appears to be significantly lower than that in 1970.

Test group allocation probability for the tennis elbow example vs. 0.5

```
> prop.test(rbind(.Table), alternative='two.sided', p=.5, conf.level=.95, correct=FALSE)

1-sample proportions test without continuity correction

data:  rbind(.Table), null probability 0.5
X-squared = 0, df = 1, p-value = 1
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval: 0.3977417 0.6022583
sample estimates:
p 0.5
```

\[ \begin{align*}
  Z &= -2.01 \\
  p &= 0.045 < 0.05 \\
  \text{Significant}
\end{align*} \]
Two-Sample Test for Binomial Proportions

Data Type: Two independent samples with binary outcome

Hypothesis: \( H_0: \ p_1 = p_2 = p \) vs \( H_1: \ p_1 \neq p_2 \)

Normal Theory Method

\( x_1 = \) the number of successes among \( n_1 \) trials in Group 1
\( x_2 = \) the number of successes among \( n_2 \) trials in Group 2

The probability of success for the two groups can be estimated by \( \hat{p}_1, \hat{p}_2 \).
Under $H_0: p_1 = p_2 = p$,

$$\mathbb{E} \left\{ \hat{p}_1 - \hat{p}_2 \right\} = 0$$

$$\text{Var} \left\{ \hat{p}_1 - \hat{p}_2 \right\} = \text{Var}(\hat{p}_1) + \text{Var}(\hat{p}_2) = \frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2} = p(1-p) \left( \frac{1}{n_1} + \frac{1}{n_2} \right) = \frac{1}{n_1 + n_2} \hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

$$\text{STD}^2(\hat{p}_1 - \hat{p}_2)$$

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\text{STD}(\hat{p}_1 - \hat{p}_2)} \sim \mathcal{N}(0,1)$$
Gynecology example

Are the proportion of women who used IUD in their lifetime different between infertile women and control women? Is IUD use associated with infertility?

<table>
<thead>
<tr>
<th></th>
<th>IUD Use</th>
<th>Non IUD Use</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Infertile</td>
<td>89</td>
<td>194</td>
<td>283</td>
</tr>
<tr>
<td>Control</td>
<td>640</td>
<td>3193</td>
<td>3833</td>
</tr>
</tbody>
</table>

\[ z = 6.29 \]

\[ p_{value} < 0.0001 \]

\[ \hat{p}_1 = \frac{89}{283} = 0.3145 \text{ Rate of IUD use in Infertile} \]

\[ \hat{p}_2 = \frac{640}{3833} = 0.1670 \text{ Rate of IUD use in Controls} \]

\[ \hat{\phi} = \frac{(89 + 640)}{(283 + 3833)} = 0.1771 \]
## Contingency Table Method

**Setup:** 2 x 2 Contingency Table

<table>
<thead>
<tr>
<th></th>
<th>Exposed</th>
<th>Unexposed</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disease</td>
<td>O&lt;sub&gt;11&lt;/sub&gt;</td>
<td>O&lt;sub&gt;12&lt;/sub&gt;</td>
<td>n&lt;sub&gt;1&lt;/sub&gt;</td>
</tr>
<tr>
<td>No Disease</td>
<td>O&lt;sub&gt;21&lt;/sub&gt;</td>
<td>O&lt;sub&gt;22&lt;/sub&gt;</td>
<td>n&lt;sub&gt;2&lt;/sub&gt;</td>
</tr>
<tr>
<td>Total</td>
<td>m&lt;sub&gt;1&lt;/sub&gt;</td>
<td>m&lt;sub&gt;2&lt;/sub&gt;</td>
<td>n</td>
</tr>
</tbody>
</table>

O<sub>ij</sub> = observed cell counts in the (i,j) cell  

n<sub>i</sub> = row total (row margin)  
m<sub>j</sub> = column total (column margin)  
n = grand total  

E<sub>ij</sub> = expected cell counts

Exp. under H<sub>0</sub> of no assoc.
Under $H_0$:

\[ E_{ij} = \frac{n_i \cdot m_j}{n} = \frac{n}{n} \cdot \frac{n_i}{n} \cdot \frac{m_j}{n} \]

\[ P(A \text{ and } B) = P(A) \cdot P(B) - \text{Independence} \]

\[ T = \sum \left( (O_{ij} - E_{ij}) - 0.5 \right)^2 \]

\[ \chi^2 \quad n \quad df = 1 \]
HYPOTHESIS OF HOMOGENEITY

Is the chance of Disease different for Exposed as compared to the Unexposed?

Is the proportion of Exposed different in those who get the Disease as compared to those who do not?

Test if some two proportions are equal:
Chi-square test of homogeneity
HYPOTHESIS OF INDEPENDENCE

Two binary random variables

$X \rightarrow \{\text{Exposed, Unexposed}\} = \{0, 1\}$

$Y \rightarrow \{\text{Disease, No Disease}\} = \{0, 1\}$

Population/Sample $(X_i, Y_i) \Rightarrow$ Summarize in a two-by-two table

Test if $X$ is independent of $Y$

Chi-square test of independence
$z \sim N(0,1) \Rightarrow z^2 \sim \chi^2_1$

$p-value = P_{\chi^2} \left\{ z^2 > \frac{y}{n} \right\} = 2P_{\chi^2} \left\{ z > \sqrt{\frac{y}{n}} \right\}$

Use only if $E_{ij} > 5$ for all $i, j$

Fisher's exact

$T = 38.34 \Rightarrow p-value < 0.0001$

If any $E_{ij} \leq 5 \Rightarrow$ Use Fisher's exact test
Two-Sample Test for Binomial Proportions for Matched-Pair Data
(McNemar's Test)

Data type: Paired data with binary outcome

Example: Sexually Transmitted Disease
1 Comparison of two different antibiotics A, B for the treatment of gonorrhea

2 Each person receiving antibiotic A is matched with an equivalent person (i.e., same age, sex and clinical condition) to whom antibiotic B is given

Information is in the difference
1 = success; 0 = failure
Difference (A,B) => (0,1) (1,0)
No Difference => (0,0) (1,1)
Difference: \(-1\)  \(0\)  \(+1\)  
Pair \((0,1)\)  \((1,1);(0,0)\)  \((1,0)\)

H0: ED=0 => equal numbers of discordant pairs  
H1: ED ≠ 0 => unequal numbers of discordant pairs

Concordant pairs bring no information on the effect (nuisance)  
Set up a table in a way that discordant/concordant pairs are visible

<table>
<thead>
<tr>
<th></th>
<th>Antibiotic B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Success</td>
<td></td>
</tr>
<tr>
<td>Failure</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Antibiotic A</th>
<th>Success</th>
<th>Failure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Success</td>
<td>40</td>
<td>20</td>
</tr>
<tr>
<td>Failure</td>
<td>16</td>
<td>3</td>
</tr>
</tbody>
</table>
Concordant pairs: Pair in which the outcome is the same for each member of the pair (43 pairs)

Discordant pairs: Pair in which the outcomes are different for the members of the pair (36 pairs)

Type A discordant pair = A: success B: failure (20 pairs)
Type B discordant pair = A: failure B: success (16 pairs)

Let $p = \Pr (\text{a discordant pair is Type A})$. Then testing whether the proportion of success is same for two groups is equivalent to testing:

\[ H_0: p = .5 \text{ vs } H_1: p \neq .5 \]
Under \( H_0 \):

\[ n_A \sim \text{Binomial}(0.5, n_D) \]

\[ n_D = n_A + n_B \]

\[ T = \frac{(n_A - 0.5 n_D) - 0.5}{\sqrt{\frac{n_D}{4}}} \sim N(0,1) \]

\[ \frac{1}{4} = 0.5 \cdot (1 - 0.5) \]

Back to example \( T = 1.96 \) \( \bar{p} = 0.61 \)

\( n_D = 36 \), \( n_A = 20 \)
EXACT TEST FOR PROPORTIONS

A test is set up to distinguish a normal person from a schizophrenic based on handwriting. A graphologist is given a set of 10 folders, each containing handwriting samples of two persons, one “normal” and the other schizophrenic. Her task is to identify which of the writings is the work of the schizophrenics. When this experiment was actually performed (Journal of Personality, 16 (1947), 192-197) the graphologist made 6 correct identifications.

1. Set up an appropriate H0 and H1 for this situation.
2. Find the decision rule at 0.05 significance level.
3. Did the graphologist at that time demonstrate a statistically significant ability to distinguish the writing of a schizophrenic from the writing of a normal person?
4. What is the p-value? Use binomial distribution.
5. What is the power of the test. (Assuming that 6/10 is the true probability of correct identification).
H₀: She is just guessing,\nPr\{Correct classification\}=p=0.5

H₁: Her judgment is informative\nPr\{Correct classification\}=p>0.5

Binomial random variable\nν = # Correct classifications in n = 10 trials

Under the null hypothesis

\[ \alpha = \Pr \{v \geq C \mid H₀: p=0.5\} \]

↑ Critical value

0.05 significance level target
Computing binomial coefficients: Pascal Triangle

<table>
<thead>
<tr>
<th>n</th>
<th>1</th>
<th></th>
<th></th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td></td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>1</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>1</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>1</td>
<td>6</td>
<td>15</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>1</td>
<td>7</td>
<td>21</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>1</td>
<td>8</td>
<td>28</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>1</td>
<td>9</td>
<td>36</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>1</td>
<td>10</td>
<td>45</td>
</tr>
<tr>
<td>K</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

\[ P \left( v = k \mid p = 0.5 \right) = \binom{n}{k} \cdot \left(1-p\right)^{n-k} \binom{n}{k} \]

\[ \text{Binomial coefficient} \]

\[ (a + b)^n = \sum_{k=0}^{n} \binom{n}{k} \cdot a^{n-k} \cdot b^k \]
| C   | Pr{v≥C|p=0.5} |
|-----|---------------|
| 6   | 0.38          |
| 7   | 0.17          |
| 8   | 0.055         |
| 9   | 0.01<0.05     | => C=9

**Decision rule**

Reject $H_0$ if $v \geq 9$

$v = 6 < 9$ => No evidence of predictive potential of handwriting

*Critical value* and type I error $\leq 0.05$
Power

What is power to detect a difference in $p=0.5$ vs. 0.6?

$\text{Power} = 1\times 0.6^1\times 0.4^0 + 10\times 0.6^9\times 0.4^1 = 0.046$

$p$-Value

$p$-Value $= \Pr\{v \geq 6 \mid p=0.5\} = 0.38 \Rightarrow H_0 \text{ cannot be rejected}$

As extreme as observed (6) or more

$\chi^2$-distribution

$\chi^2$-distribution

$H_0$

$v = 6$

$c = 9$

$v \geq 9$
A review of a clinic's records of diabetic patients revealed approximately 3,000 cases for which data on maximum weight were available. It was found that approximately two-thirds of the patients had been, at some time, 11 percent or more overweight. This provides evidence of an association between obesity and diabetes.
Data provide evidence on the chance to be overweight among diabetics. Association can only be concluded if a similar chance is estimated in non-diabetic controls and significant differences between proportions in diabetics vs. controls are shown.
Data on 123 consecutive unselected female patients with hyperparathyroidism at a university hospital revealed that 36 were under 45 years old and 87 were 45 years old and older. This led the author to conclude that, in women, hyperparathyroidism was more common in the menopausal and postmenopausal age groups.
Age may be unrelated to the disease, and the observed age distribution does not in itself say anything about the association and may be just a usual composition of hospital admissions. In order to establish the association disease risk among pre vs. post menopausal women could be evaluated for significant differences. Alternatively age distribution in hyperparathyroidism patients could be compared with age distribution of the population served by the hospital.
Although the layman generally associates heart attacks with overexertion, they are far more likely to occur during periods of rest. More than one-half the victims of heart attacks are stricken while sleeping or resting and less than 2 percent are afflicted while engaged in strenuous activity such as sports, running, lifting, or moving a heavy load.
Confounding: The chance of anything happening depends on the length of observation period. We generally spend more time sleeping than moving heavy stuff. Need to divide the number of attacks by person years rather than by people and treat it as count data or standardize observation period.
Differential admission rates


An investigation of a possible association between tuberculosis and cancer began with identification of 816 instances of cancer among the necropsy protocols at a large metropolitan teaching hospital during a particular time period.

For controls, 816 necropsy protocols were obtained during the same time period for patients dying from a wide variety of causes except any cancer. For both cancer patients and controls, the records were reviewed to determine whether or not the patients had had tuberculosis.
The results are summarized in the following 2x2 table:

<table>
<thead>
<tr>
<th>TB</th>
<th>Cancer</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Present</td>
<td>Absent</td>
</tr>
<tr>
<td>Present</td>
<td>54</td>
<td>133</td>
</tr>
<tr>
<td>Absent</td>
<td>762</td>
<td>683</td>
</tr>
<tr>
<td>Total</td>
<td>816</td>
<td>816</td>
</tr>
<tr>
<td>Percent with TB</td>
<td>6.6%</td>
<td>16.3%</td>
</tr>
</tbody>
</table>

Chi-square test, Fisher's exact test => $p < 0.0000001$

Association may be spurious if admission rate depends on the response

Given data at hand it is not possible to test whether this is happening
SPSS: 2x2 Tables

Data in case form: Weight, Cancer(Yes or No), TB(Yes or No)

Data/Weight cases

Analyze/Descriptive stats/cross tabs
Check Statistics/ Chi-square

TB * Cancer Crosstabulation

Count

<table>
<thead>
<tr>
<th>TB</th>
<th>Cancer</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>683</td>
<td>1445</td>
</tr>
<tr>
<td>Yes</td>
<td>133</td>
<td>187</td>
</tr>
<tr>
<td>Total</td>
<td>816</td>
<td>1632</td>
</tr>
</tbody>
</table>

Chi-Square Tests

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
<th>df</th>
<th>Asymp. Sig. (2-sided)</th>
<th>Exact Sig. (2-sided)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson Chi-Square</td>
<td>37.693(b)</td>
<td>1</td>
<td>.000</td>
<td></td>
</tr>
<tr>
<td>Continuity Correction(a)</td>
<td>36.745</td>
<td>1</td>
<td>.000</td>
<td></td>
</tr>
<tr>
<td>Likelihood Ratio</td>
<td>38.767</td>
<td>1</td>
<td>.000</td>
<td></td>
</tr>
<tr>
<td>Fisher's Exact Test</td>
<td></td>
<td></td>
<td></td>
<td>.000</td>
</tr>
<tr>
<td>Linear-by-Linear Association</td>
<td>37.670</td>
<td>1</td>
<td>.000</td>
<td></td>
</tr>
<tr>
<td>N of Valid Cases</td>
<td>1632</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a Computed only for a 2x2 table
b 0 cells (.0%) have expected count less than 5. The minimum expected count is 93.50.
### Column Percentages

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.62</td>
<td>16.3</td>
<td>100.0</td>
</tr>
<tr>
<td>2</td>
<td>93.38</td>
<td>83.7</td>
<td>100.0</td>
</tr>
</tbody>
</table>

**Count**

- 816.00
- 816.00

**X-squared** = 37.6934, df = 1, p-value = 8.279e-10

**Expected Counts**

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>93.5</td>
<td>93.5</td>
</tr>
<tr>
<td>2</td>
<td>722.5</td>
<td>722.5</td>
</tr>
</tbody>
</table>

**Fisher's Exact Test for Count Data**

p-value = 8.102e-10

alternative hypothesis: true odds ratio is not equal to 1

95 percent confidence interval:

- 0.2559018
- 0.5123850

sample estimates:
### Worst case scenario

Population data: No association between TB and Cancer

<table>
<thead>
<tr>
<th></th>
<th>Present</th>
<th>Absent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Present</td>
<td>79</td>
<td>409</td>
</tr>
<tr>
<td>Absent</td>
<td>1,318</td>
<td>6,830</td>
</tr>
<tr>
<td>Total</td>
<td>1,397</td>
<td>7,239</td>
</tr>
<tr>
<td>Percent with TB</td>
<td>5.6%</td>
<td>5.6%</td>
</tr>
</tbody>
</table>

Hypothetical admission probabilities:

- Cancer: 0.578
- Other diseases (excl TB and Cancer): 0.100
- TB: 0.250
# MEASURE OF EFFECT

<table>
<thead>
<tr>
<th></th>
<th>Exposed</th>
<th>Unexposed</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disease</td>
<td>$O_{11}$</td>
<td>$O_{12}$</td>
<td>$n_1$</td>
</tr>
<tr>
<td>No Disease</td>
<td>$O_{21}$</td>
<td>$O_{22}$</td>
<td>$n_2$</td>
</tr>
<tr>
<td>Total</td>
<td>$m_1$</td>
<td>$m_2$</td>
<td>$n$</td>
</tr>
</tbody>
</table>

$O_{ij} = \text{observed cell counts in the (i,j) cell}$

$n_i = \text{row total (row margin)}$

$m_j = \text{column total (column margin)}$

$n = \text{grand total}$

## Odds

The odds of an event, $Pr\{\text{event}\}=p$:

$$\text{Odds} = \frac{1-p}{p} = \text{Chance of not happening / Chance of happening}$$
Odds ( Disease | Exposed ) = ( O_{21} / m_1) / ( O_{11} / m_1)

Odds ( Disease | Non-Exposed ) = ( O_{22} / m_2) / ( O_{12} / m_2)

Odds Ratio
OR ( Disease | Exposed vs. Non-Exposed ) = ( O_{21} O_{12} ) / ( O_{11} O_{22} )

Test for association in a 2x2 table or
Test for equality of proportions is a test for OR=1
Back to differential admission, assuming independence:

<table>
<thead>
<tr>
<th>Reason for admission/Condition</th>
<th>Chance of admission</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>Not C</td>
</tr>
<tr>
<td>Cancer and TB</td>
<td>TB</td>
</tr>
<tr>
<td>C</td>
<td>Not C</td>
</tr>
<tr>
<td>Other disease and TB</td>
<td>TB</td>
</tr>
<tr>
<td>C</td>
<td>Not C</td>
</tr>
<tr>
<td>Cancer, No TB, no Other</td>
<td>.578</td>
</tr>
<tr>
<td>O</td>
<td>Not O</td>
</tr>
<tr>
<td>Other, No TB, no Cancer</td>
<td>.100</td>
</tr>
</tbody>
</table>

Population Admission Study group
79 409 x .684 .325 = 54 133
1,318 6,830 x .578 .100 = 762 683

OR Population x OR Admission = OR Study group
1.000 x 2.75 = 2.75

OR Admission = 1 if Pr{Admission due to C} = Pr{Admission due to other}

\[ \frac{2.75}{1} \]

Cases have a higher chance of admission than any other diseases.