Stat 13 Lecture 22 comparing proportions

- Estimation of population proportion
- Confidence interval; hypothesis testing
- Two independent samples
- One sample, competitive categories (negative covariance)
- One sample, non-competitive categories (usually, positive covariance)
An Example

Assume the sample is simple random.

<table>
<thead>
<tr>
<th>1996 US</th>
<th>Pre-election</th>
<th>polls</th>
<th>Election result</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>Clinton</td>
<td>Dole</td>
</tr>
<tr>
<td>New Jersey</td>
<td>1000</td>
<td>51%</td>
<td>33%</td>
</tr>
<tr>
<td>New York</td>
<td>1000</td>
<td>59%</td>
<td>25%</td>
</tr>
<tr>
<td>Connecticut</td>
<td>1000</td>
<td>51%</td>
<td>29%</td>
</tr>
</tbody>
</table>

Does the poll result significantly show the majority favor Clinton in New Jersey? For Dole, is there is a significant difference between NY and Conn? Find a 95% confidence interval for the difference of support between Clinton and Dole in New Jersey?
Do you play

- Tennis? Yes, No
- Golf? Yes, No
- Basketball? Yes, No

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>30%</td>
<td>70%</td>
</tr>
<tr>
<td>G</td>
<td>25%</td>
<td>75%</td>
</tr>
<tr>
<td>B</td>
<td>40%</td>
<td>60%</td>
</tr>
</tbody>
</table>

n=100 persons are involved in the survey
Gene Ontology
200 genes randomly selected

<table>
<thead>
<tr>
<th>cytoplasm</th>
<th>nucleus</th>
<th>others</th>
<th>unknown</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td>35</td>
<td>20</td>
<td>100</td>
</tr>
</tbody>
</table>
Central limit theorem implies that binomial is approximately normal when $n$ is large

- Sample proportion is approximately normal
- The variance of sample proportion is equal to $p(1-p)/n$
- If two random variables, $X$, $Y$ are independent, then variance of $(X-Y) = \text{var}(X) + \text{var}(Y)$
- If two random variables, $X,Y$ are dependent, then variance of $(X-Y)=\text{var}(X) + \text{var}(Y)-2\text{cov}(X,Y)$
- May apply the z-score formula to obtain confidence interval as done before.
One sample, Competitive categories

- $X=$ votes for Clinton, $Y=$ votes for Dole
- Suppose sample size is $n=1$, then only three possibilities $P(X=1, Y=0)=p_1; P(X=0, Y=1)=p_2; P(X=0, Y=0)=1-p_1-p_2$
- $E(X)=p_1; E(Y)=p_2$
- $\text{Cov}(X,Y) = E(X-p_1)(Y-p_2) = (1-p_1)(0-p_2)p_1 + (0-p_1)(1-p_2)p_2 + (0-p_1)(0-p_2)(1-p_1-p_2)$
- $= -p_1p_2$, which is negative
- In general, $\text{cov}(X,Y)= -n p_1 p_2$; therefore
- $\text{Var} \ (X/n - Y/n)=n^{-2} (\text{Var} \ X + \text{Var} \ Y + 2np_1p_2)$
- $=n^{-2}(np_1(1-p_1) + np_2(1-p_2) + 2np_1p_2)=
- (p_1 + p_2- p_1^2- p_2^2 + 2p_1p_2)/n = (p_1+p_2- (p_1-p_2)^2)/n$
Formula for confidence interval

• Let \( \hat{p}_1 = \frac{X}{n} \), \( \hat{p}_2 = \frac{Y}{n} \)

• Then the interval runs from

• \( \hat{p}_1 - \hat{p}_2 - z \text{ sd}(\hat{p}_1 - \hat{p}_2) \), to

• \( \hat{p}_1 - \hat{p}_2 + z \text{ sd}(\hat{p}_1 - \hat{p}_2) \)

• Where \( \text{sd} \) is the square root of variance, plug in the variance formula