Counts and Proportions

Thus far we have considered hypothesis testing and estimation with continuous variables. We now consider the case where there are consecutive experiments and where

1. There are only two possible (mutually exclusive) outcomes, often called a “success” and a “failure”.

2. Each experiment is identical to all the others, and the probability of a “success” is $p$. Thus, the probability of a “failure” is $1-p$. 
Each experiment is called a *Bernoulli* trial. Such experiments include throwing the die and observing whether or not it comes up six, tossing a coin, or investigating the survival of a cancer patient, etc.

For example, consider smoking status, and define $X=1$ for a smoker, and $X=0$ for a non-smoker. If “success” is the event that a randomly selected individual is a smoker and from previous research it is known that about 29% of the adults in the United States smoke, then

$$P(X=1)=p=0.29$$

and

$$P(X=0)=1−p=0.71$$

This is an example of a *Bernoulli* trial (we select just one individual at random, each selection is carried out *independently*, and, each time the probability of that individual to be a “success” is constant).
If we selected two adults, and looked at their smoking status, then the possible outcomes are:

- Neither is a smoker
- Only one is a smoker
- Both are smokers

If we define $X$ as the number of smokers between these two individuals, then

- $X=0$: Neither is a smoker
- $X=1$: Only one is a smoker
- $X=2$: Both are smokers
\[ P(X=0) = (1-p)^2 \]
\[ = (0.71)^2 \]
\[ = 0.5041 \]

\[ P(X=1) = P(\text{1st individual is a smoker OR 2nd individual is a smoker}) \]
\[ = p(1-p) + (1-p)p \]
\[ = 2p(1-p) \]
\[ = 0.4118 \]

\[ P(X=2) = p^2 \]
\[ = (0.29)^2 \]
\[ = 0.0841 \]

Notice that \( P(X=0) + P(X=1) + P(X=2) = 0.5041 + 0.4118 + 0.0841 = 1.000. \)
Smoking status of two US adults

Probability of being a smoker

Number of smokers

X=0

X=1

X=2
The binomial distribution

The bar chart above is a plot of the probability distribution of $X$, covering all of the possible numbers that $X$ can attain (in the previous example those were 0, 1, and $2=n$). The binomial distribution is a special distribution that closely models the behavior of variables that corresponds to a repeated Bernoulli experiments. When describing the binomial distribution, we need to specify two parameters:

- The probability of “success” $p$
- The number of Bernoulli experiments $n$

One way of looking at $p$ is as the proportion of time that an experiment is successful when repeated a large number of times.
Given the above parameters, the mean and standard deviation of a binomial distribution are:

| Mean: $np$ | Standard deviation: $\sqrt{np(1-p)}$ |

If $n$ is sufficiently large, then the statistic

$$Z = \frac{x - np}{\sqrt{np(1-p)}}$$

is approximately distributed as normal with mean 0 and standard deviation 1 (a std. normal distribution). A better approximation to the normal distribution is given by $Z = \frac{x - np + 0.5}{\sqrt{np(1-p)}}$ when $X < np$ and $Z = \frac{x - np - 0.5}{\sqrt{np(1-p)}}$ when $X > np$. This is called a continuity correction.
For example, suppose that we want to find the proportion of samples of size $n=30$ in which at most six individuals smoke. With $p=0.29$ and $n=30$, $X=6<np=8.7$. Thus, applying the continuity correction as shown above,

\[
P(X \leq 6) = P\left( Z \leq \frac{x - np + 0.5}{\sqrt{np(1-p)}} \right)
\]

\[
= P\left( Z \leq \frac{6 - (30)(0.29) + 0.5}{\sqrt{30(1-0.29)}} \right)
\]

\[
= P(Z \leq -0.89)
\]

\[
= 0.187
\]

The exact binomial probability is 0.190, which is very close to the approximate value given above.
Sampling distribution of a proportion

Our thinking in terms of estimation (including confidence intervals) and hypothesis testing does not change when dealing with proportions. Since the proportion of a success in the general population will not be known, we must estimate it. In general, such an estimate is derived by calculating the proportion of successes in a sample of \( n \) experiments (trials) as follows:

\[
\hat{p} = \frac{x}{n}
\]

where \( x \) is the number of successes in the sample of size \( n \).
The sampling distribution of a proportion $\hat{p}$ has mean and standard deviation

\begin{center}
\begin{tabular}{c|c}
Mean: $p$ & Standard deviation = $\sqrt{\frac{p(1-p)}{n}}$
\end{tabular}
\end{center}

and the statistic

\[ Z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \]

is distributed according to the standard normal distribution (this approximation is particularly good when $np>5$ and $n(1-p)>5$). Note here that if we multiply the numerator and denominator of $Z$ by $n$ we have,

\[ Z = \frac{n(\hat{p} - p)}{n\sqrt{p(1-p)/n}} = \frac{n(\hat{p} - p)}{\sqrt{np(1-p)}} = \frac{x-np}{\sqrt{np(1-p)}} \]

which is the familiar form encountered in our study of the binomial distribution.
**Example:** Consider the five-year survival among patients under 40 who have been diagnosed with lung cancer. The mean proportion of individuals surviving is $p = 0.10$ (implying that the standard deviation of the 5-year survival is $\sigma_p = \sqrt{p(1-p)} = \sqrt{(0.10)(0.90)} = 0.30$.

If we select repeated samples of size $n=50$ patients diagnosed with lung cancer, what fraction of the samples will have 20% or more survivors? That is, “what percent of the time 10 (=50(0.20)) or more patients will be alive after 5 years”? Since $np=(50)(0.1)=5$ and $n(1-p)=(50)(0.9)=45>5$ the normal approximation should be adequate. Then,

$$P(\hat{p} \geq 0.20) = P \left( Z \geq \frac{0.20 - p}{\sqrt{p(1-p)/n}} \right) = P \left( Z \geq \frac{0.20 - 0.10}{\sqrt{(0.10)(1-0.10)/50}} \right) = P(Z \geq 2.36) = 0.009$$

Only 0.9% of the time will the proportion of lung cancer patients surviving past five years be 20% or more.
Hypothesis testing involving proportions

In the previous example, we did not know the true proportion of 5-year survivors among individuals under 40 years of age that have been diagnosed with lung cancer.

If it is known from previous studies that the five-year survival rate of lung cancer patients that are older than 40 years old is 8.2%, we might want to test whether the five-year survival among the younger lung-cancer patients is the same as that of the older ones.
From a sample of $n=52$ patients under the age of 40 that have been diagnosed with lung cancer the proportion surviving after five years is $\hat{p} = 0.115$. Is this within sampling variability of the known 5-year survival of older patients?

The test of hypothesis is constructed as follows:

1. $H_0$: $p=0.082$
2. $H_A$: $p \neq 0.082$
3. The alpha level of this test is 0.01
4. The test statistic is $z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$
5. We will reject the null hypothesis if $P(|Z| \geq z) < \alpha$. 
Example (continued): Five-year survival of lung cancer patients

The test statistic is

\[
\hat{p} - p = \frac{0.115 - 0.082}{\sqrt{(0.082)(1 - 0.082)/52}} = 0.87
\]

Since \( P(|Z|\geq 0.87) = P(Z \geq 0.87) + P(Z \leq -0.87) = 0.384 \), we do not reject the null hypothesis.

That is, there is not sufficient evidence to indicate that the five-year survival of lung cancer patients who are younger than 40 years of age is different than that of the older patients.
Carrying out an exact binomial test in the problem above we have

<table>
<thead>
<tr>
<th>N</th>
<th>Observed k</th>
<th>Expected k</th>
<th>Assumed p</th>
<th>Observed p</th>
</tr>
</thead>
<tbody>
<tr>
<td>52</td>
<td>6</td>
<td>4.264</td>
<td>0.08200</td>
<td>0.11538</td>
</tr>
</tbody>
</table>

Pr(k >= 6) = 0.251946 (one-sided test)
Pr(k <= 6) = 0.868945 (one-sided test)
Pr(k <= 1 or k >= 6) = 0.317935 (two-sided test)

We see that the $p$ value associated with the two-sided test is 0.318, which is close to that calculated above.
Using the normal approximation we have:

| Variable | Mean   | Std. Err. | z   | P>|z| | [95% Conf. Interval] |
|----------|--------|-----------|-----|-----|----------------------|
| x        | 0.1153846 | 0.0443047 | 2.60434 | 0.0092 | 0.0285491 | 0.2022202 |

\[ \text{Ho: proportion}(x) = 0.082 \]

\[ \begin{align*}
\text{Ha: } x &< 0.082 \\
z & = 0.877 \\
P & < z = 0.8099
\end{align*} \]

\[ \begin{align*}
\text{Ha: } x &= 0.082 \\
\text{Ha: } x &> 0.082 \\
z & = 0.877 \\
P & > |z| = 0.3802 \\
P & > z = 0.1901
\end{align*} \]

Which closely matches our calculations. We do not reject the null hypothesis, since the \( p \) value associated with the two-sided test is \( 0.380 > \alpha \). Note that any differences with the hand calculations are due to round-off error.
Estimation

Similar to the testing of hypothesis involving proportions, we can construct confidence intervals for a population proportion.

Again these intervals will be based on the statistic

\[ Z = \frac{\hat{p} - p}{\sqrt{\hat{p}(1 - \hat{p})/n}} \]

where \( \hat{p} = \frac{x}{n} \) and \( \sqrt{\hat{p}(1 - \hat{p})/n} \) are the estimates of the proportion and its associated standard deviation respectively.
Two sided confidence intervals:

\[
\left[ \hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right]
\]

One-sided confidence intervals:

**Upper:** \[0, \hat{p} + z_{\alpha} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\]

**Lower:** \[\hat{p} - z_{\alpha} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, 1\]
In the previous example, if 6 out of 52 lung cancer patients under 40 years of age were alive after five years, and using the normal approximation (which is justified since \( np = 52(0.115) = 5.98 > n \), and \( 52(1-0.115) = 46.02 > n \)), an approximate 95% confidence interval for the true proportion \( p \) is given by

\[
\hat{p} - z_{a/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{a/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}
\]

\[
\downarrow
\]

\[
[0.115 - 1.96 \sqrt{\frac{0.115(1-0.115)}{52}}, 0.115 + 1.96 \sqrt{\frac{0.115(1-0.115)}{52}}]
\]

\[
\downarrow
\]

\[
[0.028, 0.202]
\]
In other words, we are 95% confident that the true five-year survival of lung-cancer patients < 40 years of age is between 2.8% and 20.2%. Note that this interval contains 8.2% (the five-year survival rate among lung cancer patients that are older than 40 years of age). Thus, it is equivalent to a hypothesis test that \textit{did not reject} the null hypothesis of equal five-year survival between lung cancer patients that are older than 40 years old versus younger subjects.
In the previous example, using exact binomial confidence intervals we have

```
<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Err.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>52</td>
<td>.1153846</td>
<td>.0443047</td>
<td>.0435439  .2344114</td>
</tr>
</tbody>
</table>
```

which is close to our calculations that used the normal approximation. Note that $0.31902 = \sqrt{0.115(1-0.115)}$, $0.31902 = \sqrt{0.115(1-0.115)}$
Comparison between two proportions

We proceed when comparing two proportions similar to a two-mean comparison. The sample proportion in the first and second groups are $\hat{p}_1 = \frac{x_1}{n_1}$ and $\hat{p}_2 = \frac{x_2}{n_2}$. Under assumptions of equality of the two population proportions, we may want to derive a pooled estimate of the sample proportion, using data from both groups, $\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$, that is divide the total number of successes in the two groups ($x_1 + x_2$), by the total sample size ($n_1 + n_2$). Using this pooled estimate, we can derive a pooled estimate of the standard deviation of the unknown proportion (assumed equal between the two groups as $s_p = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$.)
The hypothesis testing of comparisons between two proportions is based on the statistic 
\[ z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}(1 - \hat{p})(1/n_1 + 1/n_2)}} \]
and is carried out as follows:

1. \( H_0: p_1 = p_2 \) (or \( p_1 - p_2 = 0 \))

2. a. **Two sided tests:** \( H_A: p_1 \neq p_2 \) (or \( p_1 - p_2 \neq 0 \))
   b. **One sided tests:** \( H_A: p_1 > p_2 \) (or \( p_1 - p_2 > 0 \))
      \( H_A: p_1 < p_2 \) (or \( p_1 - p_2 < 0 \))

3. Determine the alpha level of the test

4. **Rejection rule:**
   a. **Two-sided tests:** Reject \( H_0 \) if \( P(|Z| \geq z) < \alpha \).
   b. **One-sided tests:**
      - Reject \( H_0 \) if \( P(Z \geq z) < \alpha \).
      - Reject \( H_0 \) if \( P(Z \leq -z) < \alpha \).
Comparison between two proportions

In a study investigating morbidity and mortality among pediatric victims of motor vehicles accidents, information regarding the effectiveness of seat belts was collected.

Two random samples were selected, one of size $n_1=123$ from a population of children that were wearing seat belts at the time of the accident, and another of size $n_2=290$ from a group of children that were not wearing seat belts at the time of the accident. In the first case, $x_1=3$ children died, while in the second $x_2=13$ died. Consequently, $\hat{p}_1=0.024$ and $\hat{p}_2=0.045$. We wish to determine if the death rate is different in the two groups.
Comparison between two proportions

Carrying out the test of hypothesis as proposed earlier,

1. \( H_0: p_1 = p_2 \) (or \( p_1 - p_2 = 0 \))
2. \( H_A: p_1 \neq p_2 \) (or \( p_1 - p_2 \neq 0 \))
3. The alpha level of the test is 5%.
4. \[ P(|Z| \geq z) = P\left(|Z| \geq \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n_2)}}\right) = P(Z < -0.98) + P(Z > 0.98) = 0.325 > \alpha. \]

Thus, there is not sufficient evidence to conclude that children not wearing seat belts are safer (die in different rates) than children wearing seat belts.
Confidence intervals of the difference between two proportions

Confidence intervals of the difference of two proportions are also based on the statistic

\[ z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}_1 (1-\hat{p}_1)/n_1 + \hat{p}_2 (1-\hat{p}_2)/n_2}}. \]

Note that since we no longer need to assume that the two proportions are equal, the estimate of the standard deviation in the denominator is not a pooled estimate, but rather simply the sum of the std. deviations in each group. That is, the standard deviation estimate is

\[ s_\delta = \sqrt{\hat{p}_1 (1-\hat{p}_1)/n_1 + \hat{p}_2 (1-\hat{p}_2)/n_2}. \]

This an important deviation from hypothesis testing and may lead to disagreements between decisions reached through usual hypothesis testing.
versus hypothesis testing performed using confidence intervals (but infrequently).
Two sided confidence intervals:

\[
\left( \hat{p}_1 - \hat{p}_2 \right) - z_{a/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}, \left( \hat{p}_1 - \hat{p}_2 \right) + z_{a/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}
\]

One-sided confidence intervals:

**Upper:**

\[
\left[ -1, \left( \hat{p}_1 - \hat{p}_2 \right) + z_\alpha \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \right]
\]

**Lower:**

\[
\left[ \left( \hat{p}_1 - \hat{p}_2 \right) - z_\alpha \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}, 1 \right]
\]
A two-sided 95% confidence interval for the true difference death rates among children wearing seat belts versus those that did not is given by

\[
\left( \hat{p}_1 - \hat{p}_2 \right) - z_{a/2} \sqrt{ \frac{ \hat{p}_1 (1 - \hat{p}_1) + \hat{p}_2 (1 - \hat{p}_2) }{ n_1 } }, \left( \hat{p}_1 - \hat{p}_2 \right) + z_{a/2} \sqrt{ \frac{ \hat{p}_1 (1 - \hat{p}_1) + \hat{p}_2 (1 - \hat{p}_2) }{ n_2 } },
\]

\[\downarrow\]

\[
(0.024 - 0.045) \pm 1.96 \sqrt{ \frac{0.024(1-0.024)}{123} + \frac{0.045(1-0.045)}{290} }.
\]

\[\downarrow\]

\[
[-0.057, 0.015]
\]

That is, the true difference between the two groups will be between 5.7% in favor of those children wearing seat belts, to 1.5% in favor of those children not wearing seat belts. In this regard, since the zero (hypothesized under the null hypothesis) difference is included in the confidence interval we fail to reject the null hypothesis. There is no evidence to suggest a benefit of seat belts.
STATA Output

Two-sample test of proportion

| Variable | Mean   | Std. Err. | z      | P>|z|  | [95% Conf. Interval] |
|----------|--------|-----------|--------|------|---------------------|
| x        | 0.0243902 | 0.0139089 | 1.75357 | 0.0795 | -0.0028707 to 0.0516512 |
| y        | 0.0448276 | 0.0121511 | 3.68919 | 0.0002 | 0.0210119 to 0.0686432 |
| diff     | -0.0204373 | 0.0184691 | -0.98421 | 0.3250 | -0.0566361 to 0.0157614 |

Ho: proportion(x) - proportion(y) = diff = 0

Ha: diff < 0         Ha: diff = 0         Ha: diff > 0
z = -0.984           z = -0.984           z = -0.984
P < z = 0.1625       P > |z| = 0.3250       P > z = 0.8375