§ 6.2.1 Means

First assume that the samples are randomly selected from two populations that are independent, i.e., no relation exists between individuals of one population and the other, relative to the random variable, or any lurking or confounding variables that might have an effect on this variable.

Model: Phase III Randomized Clinical Trial (RCT)

Measuring the effect of treatment (e.g., drug) versus control (e.g., placebo) on a response variable $X$, to determine if there is any significant difference between them.

- **Control Arm**
  - Assume $X_1 \sim N(\mu_1, \sigma_1)$
- **Treatment Arm**
  - Assume $X_2 \sim N(\mu_2, \sigma_2)$

Then… ↓ CLT ↓

- Sample, size $n_1$
  - $\overline{X}_1 \sim N(\mu_1, \frac{\sigma_1}{\sqrt{n_1}})$
- Sample, size $n_2$
  - $\overline{X}_2 \sim N(\mu_2, \frac{\sigma_2}{\sqrt{n_2}})$

So…

$$\overline{X}_1 - \overline{X}_2 \sim N(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}})$$

### Null Distribution

$H_0$: $\mu_1 - \mu_2 = 0$ (There is no difference in mean response between the two populations.)

$\overline{X}_1 - \overline{X}_2$

### Comments:

- Recall from 4.1: If $Y_1$ and $Y_2$ are independent, then $\text{Var}(Y_1 - Y_2) = \text{Var}(Y_1) + \text{Var}(Y_2)$.
- If $n_1 = n_2$, the samples are said to be (numerically) balanced.
- The null hypothesis $H_0$: $\mu_1 - \mu_2 = 0$ can be replaced by $H_0$: $\mu_1 - \mu_2 = \mu_0$ if necessary, in order to compare against a specific constant difference $\mu_0$ (e.g., 10 cholesterol points), with the corresponding modifications below.
- $\text{s.e.} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ can be replaced by $\text{s.e.} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$, provided $n_1 \geq 30, n_2 \geq 30$. 
Example: \( X = \text{“cholesterol level (mg/dL)”} \)

Test \( H_0: \mu_1 - \mu_2 = 0 \) vs. \( H_A: \mu_1 - \mu_2 \neq 0 \) for significance at the \( \alpha = .05 \) level.

<table>
<thead>
<tr>
<th>Placebo</th>
<th>Drug</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n_1 = 80 )</td>
<td>( n_2 = 60 )</td>
</tr>
<tr>
<td>( \bar{x}_1 = 240 )</td>
<td>( \bar{x}_2 = 229 )</td>
</tr>
<tr>
<td>( s_1^2 = 1200 )</td>
<td>( s_2^2 = 600 )</td>
</tr>
</tbody>
</table>

\[
\frac{s_1^2}{n_1} = \frac{1200}{80} = 15, \quad \frac{s_2^2}{n_2} = \frac{600}{60} = 10 \quad \Rightarrow \quad \text{s.e.} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{25} = 5
\]

\[
(1 - \alpha) \times 100\% \text{ Confidence Interval for } \mu_1 - \mu_2
\]

\[
\left( \bar{x}_1 - \bar{x}_2 \right) - z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \quad (\bar{x}_1 - \bar{x}_2) + z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}
\]

\[
\text{95\% Confidence Interval for } \mu_1 - \mu_2
\]

95\% limits = \( 11 \pm (1.96)(5) = 11 \pm 9.8 \) \( \leftarrow \text{margin of error} \)

\[
\therefore \text{95\% CI} = (1.2, 20.8), \text{ which does not contain 0} \quad \Rightarrow \quad \text{Reject } H_0. \text{ Drug works!}
\]

\[
(1 - \alpha) \times 100\% \text{ Acceptance Region for } H_0: \mu_1 - \mu_2 = \mu_0
\]

\[
\left( \mu_0 - z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \quad \mu_0 + z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right)
\]

\[
\text{95\% Acceptance Region for } H_0: \mu_1 - \mu_2 = 0
\]

95\% limits = \( 0 \pm (1.96)(5) = \pm 9.8 \) \( \leftarrow \text{margin of error} \)

\[
\therefore \text{95\% AR} = (-9.8, +9.8), \text{ which does not contain 11} \quad \Rightarrow \quad \text{Reject } H_0. \text{ Drug works!}
\]

\[
\text{Test Statistic}
\]

\[
Z = \frac{(\bar{X}_1 - \bar{X}_2) - \mu_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim N(0, 1)
\]

\[
\text{p-value} = 2P(\bar{X}_1 - \bar{X}_2 \geq 11)
\]

\[
= 2P(Z \geq \frac{11 - 0}{5})
\]

\[
= 2P(Z \geq 2.2)
\]

\[
= 2(.0139)
\]

\[
= .0278 < .05 = \alpha
\]

\[
\therefore \text{Reject } H_0. \text{ Drug works!}
\]
Null Distribution
\[ \bar{X}_1 - \bar{X}_2 \sim N(0, 5) \]

0 is not in the 95% Confidence Interval = (1.2, 20.8)

11 is not in the 95% Acceptance Region = (−9.8, 9.8)
Small samples: What if \( n_1 < 30 \) and/or \( n_2 < 30 \)? Then use the \( t \)-distribution, provided…

\[ H_0: \sigma_1^2 = \sigma_2^2 \]  
(equivariance, homoscedasticity)

Technically, this requires a formal test using the \( F \)-distribution; see next section (§ 6.2.2). However, an informal criterion is often used:

\[ \frac{1}{4} < F = \frac{s_1^2}{s_2^2} < 4. \]

If equivariance is accepted, then the common value of \( \sigma_1^2 \) and \( \sigma_2^2 \) can be estimated by the weighted mean of \( s_1^2 \) and \( s_2^2 \), the pooled sample variance:

\[ s_{pooled}^2 = \frac{d_1 s_1^2 + d_2 s_2^2}{d_1 + d_2}, \]  
where \( d_1 = n_1 - 1 \) and \( d_2 = n_2 - 1 \),

i.e.,

\[ s_{pooled}^2 = \frac{(n_1 - 1) s_1^2 + (n_2 - 1) s_2^2}{n_1 + n_2 - 2} = \frac{SS}{df}. \]

Therefore, in this case, we have \( \text{s.e.} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \) estimated by

\[ \hat{s.e.} = \sqrt{s_{pooled}^2/n_1 + s_{pooled}^2/n_2} \]

i.e.,

\[ \hat{s.e.} = \sqrt{s_{pooled}^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)} \]

\[ = s_{pooled} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}. \]

If equivariance (but not normality) is rejected, then an approximate \( t \)-test can be used, with the approximate degrees of freedom \( df \) given by

\[ \frac{\left( \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}. \]

This is known as the Smith-Satterwaite Test. (Also used is the Welch Test.)
Example:  \( X = \text{“cholesterol level (mg/dL)”} \)

Test \( H_0: \mu_1 - \mu_2 = 0 \) vs. \( H_A: \mu_1 - \mu_2 \neq 0 \) for significance at the \( \alpha = .05 \) level.

<table>
<thead>
<tr>
<th>Placebo</th>
<th>Drug</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n_1 = 8 )</td>
<td>( n_2 = 10 )</td>
</tr>
<tr>
<td>( \bar{x}_1 = 230 )</td>
<td>( \bar{x}_2 = 200 )</td>
</tr>
<tr>
<td>( s_1^2 = 775 )</td>
<td>( s_2^2 = 1175 )</td>
</tr>
</tbody>
</table>

\[ \bar{x}_1 - \bar{x}_2 = 30 \]

\[ F = \frac{s_1^2}{s_2^2} = 0.66, \]

which is between 0.25 and 4.

\[ \text{Equivariance accepted} \Rightarrow t\text{-test} \checkmark \]

**Pooled Variance**

\[ s_{pooled}^2 = \frac{(8 - 1)(775) + (10 - 1)(1175)}{8 + 10 - 2} = \frac{16000}{16} = 1000 \]

\[ \uparrow \]

\[ \text{df} \]

Note that \( s_{pooled}^2 = 1000 \) is indeed between the variances \( s_1^2 = 775 \) and \( s_2^2 = 1175 \).

**Standard Error**

\[ \text{s.e.} = \sqrt{1000 \left( \frac{1}{8} + \frac{1}{10} \right)} = 15 \]

\[ \text{Margin of Error} = (2.120)(15) = 31.8 \]

**Critical Value**

\[ t_{16,.025} = 2.120 \]
(1 - \alpha) \times 100\% Confidence Interval for \mu_1 - \mu_2

\left( \bar{x}_1 - \bar{x}_2 \right) - t_{df, \alpha/2} \sqrt{s_{pooled}^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}, \left( \bar{x}_1 - \bar{x}_2 \right) + t_{df, \alpha/2} \sqrt{s_{pooled}^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}

where df = n_1 + n_2 - 2

\begin{equation}
\begin{align*}
95\% \text{ Confidence Interval for } \mu_1 - \mu_2 \\
95\% \text{ limits} &= 30 \pm 31.8 \ \text{margin of error} \\
\therefore 95\% \text{ CI} &= (-1.8, 61.8), \text{ which contains 0 } \Rightarrow \text{ Accept } H_0.
\end{align*}
\end{equation}

(1 - \alpha) \times 100\% Acceptance Region for H_0: \mu_1 - \mu_2 = \mu_0

\left( \mu_0 - t_{df, \alpha/2} \sqrt{s_{pooled}^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}, \mu_0 + t_{df, \alpha/2} \sqrt{s_{pooled}^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)} \right)

where df = n_1 + n_2 - 2

\begin{equation}
\begin{align*}
95\% \text{ Acceptance Region for } H_0: \mu_1 - \mu_2 = 0 \\
95\% \text{ limits} &= 0 \pm 31.8 \ \text{margin of error} \\
\therefore 95\% \text{ AR} &= (-31.8, +31.8), \text{ which contains 30 } \Rightarrow \text{ Accept } H_0.
\end{align*}
\end{equation}

Test Statistic

\begin{equation}
T = \frac{\left( \bar{x}_1 - \bar{x}_2 \right) - \mu_0}{\sqrt{s_{pooled}^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim t_{df}
\end{equation}

where df = n_1 + n_2 - 2

\begin{equation}
p-value = 2 P(\bar{x}_1 - \bar{x}_2 \geq 30) = 2 P\left( T_{16} \geq \frac{30 - 0}{15} \right) = 2 P(T_{16} \geq 2.0) = 2(.0314) = .0628 > .05 = \alpha
\end{equation}

\Rightarrow \text{ Accept } H_0.

Once again, low sample size implies low power to reject the null hypothesis. The tests do not show significance, and we cannot conclude that the drug works, based on the data from these small samples. Perhaps a larger study is indicated…
Now consider the case where the two samples are dependent. That is, each observation in the first sample is paired, or matched, in a natural way on a corresponding observation in the second sample.

**Examples:**

- Individuals may be matched on characteristics such as age, sex, race, and/or other variables that might confound the intended response.
- Individuals may be matched on personal relations such as siblings (similar genetics, e.g., twin studies), spouses (similar environment), etc.
- Observations may be connected physically (e.g., left arm vs. right arm), or connected in time (e.g., before treatment vs. after treatment).

\[ H_0: \mu_1 - \mu_2 = 0 \]

\[ D = X - Y \sim N(\mu, \sigma) \]

where \( \mu_D = \mu_1 - \mu_2 \)

Calculate the difference \( d_i = x_i - y_i \) of each matched pair of observations, thereby forming a single collapsed sample \( \{d_1, d_2, d_3, \ldots, d_n\} \), and apply the appropriate one-sample Z- or t- test to the equivalent null hypothesis \( H_0: \mu_D = 0 \).
Checks for normality include normal scores plot (probability plot, Q-Q plot), etc., just as with one sample.

Remedies for non-normality include transformations (e.g., logarithmic or square root), or nonparametric tests.

- Independent Samples: Wilcoxon Rank Sum Test (= Mann-Whitney U Test)
- Dependent Samples: Sign Test, Wilcoxon Signed Rank Test (just as with one sample)
Step-by-Step Hypothesis Testing
Two Sample Means \( H_0: \mu_1 - \mu_2 \) vs. 0

Independent

Independent or Paired?

Yes

No, or don’t know

Are \( X_1 \) and \( X_2 \) approximately normally distributed (or mildly skewed)?

Yes

No

Are \( \sigma_1, \sigma_2 \) known?

Yes

No

Use Z-test (with \( \sigma_1, \sigma_2 \))

\[
z = \frac{\bar{X}_1 - \bar{X}_2 - \mu_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}
\]

Use t-test or Z-test (with \( \hat{\sigma}_1 = s_1, \hat{\sigma}_2 = s_2 \))

\[
Z = \frac{\bar{X}_1 - \bar{X}_2 - \mu_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}
\]

Yes

No

Are \( n_1 \geq 30 \) and \( n_2 \geq 30 \)?

Equivariance: \( \sigma_1^2 = \sigma_2^2 ? \)

Compute \( F = s_1^2 / s_2^2 \).

Is \( 1/4 < F < 4 ? \)

Yes

No

Use t-test (with \( \hat{\sigma}_1^2 = \hat{\sigma}_2^2 = s_\text{pooled}^2 \))

\[
T_{n_1+n_2-2} = \frac{\bar{X}_1 - \bar{X}_2 - \mu_0}{s_{\text{pooled}} \sqrt{1/n_1 + 1/n_2}}
\]

\[
s_{\text{pooled}}^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}
\]

Use an approximate t-test, e.g., Satterwaithe Test

Compute \( D = X_1 - X_2 \) for each \( i = 1, 2, \ldots, n \).
Then calculate…

- sample mean \( \bar{d} = \frac{1}{n} \sum d_i \)
- sample variance \( s_d^2 = \frac{1}{n-1} \sum (d_i - \bar{d})^2 \)

… and GO TO “One Sample Mean” testing of \( H_0: \mu_D = 0 \), section 6.1.1.
§ 6.2.2 Variances

Suppose \( X_1 \sim N(\mu_1, \sigma_1) \) and \( X_2 \sim N(\mu_2, \sigma_2) \).

Null Hypothesis \( H_0: \sigma_1^2 = \sigma_2^2 \)

versus

Alternative Hypothesis \( H_A: \sigma_1^2 \neq \sigma_2^2 \)

Test Statistic

\[
F = \frac{s_1^2}{s_2^2} \sim F_{\nu_1, \nu_2}
\]

where \( \nu_1 = n_1 - 1 \) and \( \nu_2 = n_2 - 1 \) are the corresponding numerator and denominator degrees of freedom, respectively.

Formal test: Reject \( H_0 \) if the \( F \)-statistic is significantly different from 1.

Informal criterion: Accept \( H_0 \) if the \( F \)-statistic is between 0.25 and 4.

Comment: Another test, more robust to departures from the normality assumption than the \( F \)-test, is Levene’s Test, a \( t \)-test of the absolute deviations of each sample. It can be generalized to more than two samples (see section 6.3.2).
§ 6.2.3 Proportions

POPULATION

<table>
<thead>
<tr>
<th>Binary random variable</th>
<th>Binary random variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>$I_1 = 1$ or $0$, with</td>
<td>$I_2 = 1$ or $0$, with</td>
</tr>
<tr>
<td>$P(I_1 = 1) = \pi_1$, $P(I_1 = 0) = 1 - \pi_1$</td>
<td>$P(I_2 = 1) = \pi_2$, $P(I_2 = 0) = 1 - \pi_2$</td>
</tr>
</tbody>
</table>

INDEPENDENT SAMPLES

Random Variable

$X_1 = \#(I_1 = 1) \sim \text{Bin}(n_1, \pi_1)$

Recall (assuming $n_1 \pi_1 \geq 15$, $n_1(1 - \pi_1) \geq 15$):

$\hat{\pi}_1 = \frac{X_1}{n_1} \sim N\left(\pi_1, \sqrt{\frac{\pi_1(1 - \pi_1)}{n_1}}\right)$, approx.

Random Variable

$X_2 = \#(I_2 = 1) \sim \text{Bin}(n_2, \pi_2)$

Recall (assuming $n_2 \pi_2 \geq 15$, $n_2(1 - \pi_2) \geq 15$):

$\hat{\pi}_2 = \frac{X_2}{n_2} \sim N\left(\pi_2, \sqrt{\frac{\pi_2(1 - \pi_2)}{n_2}}\right)$, approx.

Therefore, approximately...

$\hat{\pi}_1 - \hat{\pi}_2 \sim N\left(\pi_1 - \pi_2, \sqrt{\frac{\pi_1(1 - \pi_1)}{n_1} + \frac{\pi_2(1 - \pi_2)}{n_2}}\right)$.

↑

standard error s.e.

Confidence intervals are computed in the usual way, using the estimate

$s.e. = \sqrt{\frac{\hat{\pi}_1(1 - \hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1 - \hat{\pi}_2)}{n_2}}$,

as follows:
Unlike the one-sample case, the same estimate for the standard error can also be used in computing the acceptance region for the null hypothesis $H_0: \pi_1 - \pi_2 = \pi_0$, as well as the test statistic for the $p$-value, provided the null value $\pi_0 \neq 0$. HOWEVER, if testing for equality between two proportions via the null hypothesis $H_0: \pi_1 - \pi_2 = 0$, then their common value should be estimated by the more stable weighted mean of $\hat{\pi}_1$ and $\hat{\pi}_2$, the pooled sample proportion:

$$\hat{\pi}_{\text{pooled}} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{n_1 \hat{\pi}_1 + n_2 \hat{\pi}_2}{n_1 + n_2}.$$  

Substituting yields…

$$s.e.0 = \sqrt{\frac{\hat{\pi}_{\text{pooled}}(1 - \hat{\pi}_{\text{pooled}})}{n_1} + \frac{\hat{\pi}_{\text{pooled}}(1 - \hat{\pi}_{\text{pooled}})}{n_2}}.$$  

i.e.,

$$s.e.0 = \sqrt{\frac{\hat{\pi}_{\text{pooled}}(1 - \hat{\pi}_{\text{pooled}})}{n_1} + \frac{\hat{\pi}_{\text{pooled}}(1 - \hat{\pi}_{\text{pooled}})}{n_2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$  

Hence…

$$(1 - \alpha) \times 100\% \text{ Acceptance Region for } H_0: \pi_1 - \pi_2 = 0$$

$$\left( 0 - z_{\alpha/2} \sqrt{\frac{\hat{\pi}_{\text{pooled}}(1 - \hat{\pi}_{\text{pooled}})}{n_1} + \frac{1}{n_1} + \frac{1}{n_2}}, 0 + z_{\alpha/2} \sqrt{\frac{\hat{\pi}_{\text{pooled}}(1 - \hat{\pi}_{\text{pooled}})}{n_2} + \frac{1}{n_1} + \frac{1}{n_2}} \right)$$

Test Statistic for $H_0: \pi_1 - \pi_2 = 0$

$$Z = \frac{(\hat{\pi}_1 - \hat{\pi}_2) - 0}{\sqrt{\hat{\pi}_{\text{pooled}}(1 - \hat{\pi}_{\text{pooled}}) \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0, 1)$$
Example: Consider a group of 720 patients who undergo physical therapy for arthritis. A daily supplement of glucosamine and chondroitin is given to $n_1 = 400$ of them in addition to the physical therapy; after four weeks of treatment, $X_1 = 332$ show measurable signs of improvement (increased ROM, etc.). The remaining $n_2 = 320$ patients receive physical therapy only; after four weeks, $X_2 = 244$ show improvement. Does this difference represent a statistically significant treatment effect? Calculate the $p$-value, and form a conclusion at the $\alpha = .05$ significance level.

### $PT + Supplement$ vs. $PT only$

<table>
<thead>
<tr>
<th></th>
<th>$n_1 = 400$</th>
<th>$n_2 = 320$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
<td>332</td>
<td>244</td>
</tr>
</tbody>
</table>

$H_0$: $\pi_1 - \pi_2 = 0$ vs. $H_A$: $\pi_1 - \pi_2 \neq 0$ at $\alpha = .05$

$\hat{\pi}_1 = \frac{332}{400} = 0.83, \quad \hat{\pi}_2 = \frac{244}{320} = 0.7625 \quad \Rightarrow \quad \hat{\pi}_1 - \hat{\pi}_2 = 0.0675$

$\hat{\pi}_{pooled} = \frac{332 + 244}{400 + 320} = \frac{576}{720} = 0.8$

and thus $1 - \hat{\pi}_{pooled} = \frac{144}{720} = 0.2$

$s.e. = \sqrt{(0.8)(0.2)} \sqrt{\frac{1}{400} + \frac{1}{320}} = 0.03$

Therefore, $p$-value $= 2 P(Z \geq 2.25) = 2 \cdot 0.0122 = 0.0244$.

Conclusion: As this value is smaller than $\alpha = .05$, we can reject the null hypothesis that the two proportions are equal. There does indeed seem to be a moderately significant treatment difference between the two groups.
Exercise: Instead of $H_0: \pi_1 - \pi_2 = 0$ vs. $H_A: \pi_1 - \pi_2 \neq 0$, test the null hypothesis for a 5% difference, i.e., $H_0: \pi_1 - \pi_2 = .05$ vs. $H_A: \pi_1 - \pi_2 \neq .05$, at $\alpha = .05$. [Note that the pooled proportion $\hat{\pi}_{pooled}$ is no longer appropriate to use in the expression for the standard error under the null hypothesis, since $H_0$ is not claiming that the two proportions $\pi_1$ and $\pi_2$ are equal (to a common value); see notes above.] Conclusion?

Exercise: Instead of $H_0: \pi_1 - \pi_2 = 0$ vs. $H_A: \pi_1 - \pi_2 \neq 0$, test the one-sided null hypothesis $H_0: \pi_1 - \pi_2 \leq 0$ vs. $H_A: \pi_1 - \pi_2 > 0$ at $\alpha = .05$. Conclusion?

Exercise: Suppose that in a second experiment, $n_1 = 400$ patients receive a new drug that targets B-lymphocytes, while the remaining $n_2 = 320$ receive a placebo, both in addition to physical therapy. After four weeks, $X_1 = 376$ and $X_2 = 272$ show improvement, respectively. Formally test the null hypothesis of equal proportions at the $\alpha = .05$ level. Conclusion?

Exercise: Finally suppose that in a third experiment, $n_1 = 400$ patients receive “magnet therapy,” while the remaining $n_2 = 320$ do not, both in addition to physical therapy. After four weeks, $X_1 = 300$ and $X_2 = 240$ show improvement, respectively. Formally test the null hypothesis of equal proportions at the $\alpha = .05$ level. Conclusion?

See…

Appendix > Statistical Inference > General Parameters and FORMULA TABLES.
**Alternate Method:** Chi-Squared ($\chi^2$) Test

As before, let the *binary* variable $I = 1$ for improvement, $I = 0$ for no improvement, with probability $\pi$ and $1 - \pi$, respectively. Now define a second *binary* variable $J = 1$ for the “PT + Drug” group, and $J = 0$ for the “PT only” group. Thus, there are four possible disjoint events: “$I = 0$ and $J = 0$,” “$I = 0$ and $J = 1$,” “$I = 1$ and $J = 0$,” and “$I = 1$ and $J = 1$.” The number of times these events occur in the random sample can be arranged in a $2 \times 2$ contingency table that consists of four cells (NW, NE, SW, and SE) as demonstrated below, and compared with their corresponding expected values based on the null hypothesis.

### Observed Values

<table>
<thead>
<tr>
<th>Status ($)</th>
<th>Improvement</th>
<th>No Improvement</th>
<th>Total marginal totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group (J)</td>
<td>PT + Drug</td>
<td>PT only</td>
<td></td>
</tr>
<tr>
<td>Improvement</td>
<td>332</td>
<td>244</td>
<td>576</td>
</tr>
<tr>
<td>No Improvement</td>
<td>68</td>
<td>76</td>
<td>144</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Column marginal totals</th>
<th>Row marginal totals</th>
<th>Total Sample Size n</th>
</tr>
</thead>
<tbody>
<tr>
<td>400</td>
<td>320</td>
<td>720</td>
<td></td>
</tr>
</tbody>
</table>

| Informal reasoning: Consider the first cell, improvement in the 400 patients of the “PT + Drug” group. The null hypothesis conjectures that the probability of improvement is equal in both groups, and this common value is estimated by the pooled proportion $\frac{576}{720}$. Hence, the expected number (under $H_0$) of improved patients in the “PT + Drug” group is $400 \times \frac{576}{720}$, etc. |

### Expected Values

Under $H_0$: $\pi_1 = \pi_2$

\[ \hat{\pi}_{\text{pooled}} = \frac{576}{720} = 0.8 \]

<table>
<thead>
<tr>
<th></th>
<th>Column total $\times$ Row total</th>
<th>Total Sample Size n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Improvement</td>
<td>$\frac{400 \times 576}{720}$ = 320.0</td>
<td>$\frac{320 \times 576}{720}$ = 256.0</td>
</tr>
<tr>
<td>Status ($)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No Improvement</td>
<td>$\frac{400 \times 144}{720}$ = 80.0</td>
<td>$\frac{320 \times 144}{720}$ = 64.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Group (J)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PT + Drug</td>
<td>PT only</td>
</tr>
<tr>
<td>Improvement</td>
<td>$\frac{320}{400} = 256\div320 \checkmark$</td>
</tr>
<tr>
<td>No Improvement</td>
<td>$\frac{576}{720}$</td>
</tr>
</tbody>
</table>
Ideally, if all the observed values = all the expected values, then this statistic would = 0, and the corresponding \( p \)-value = 1. As it is,

\[
X^2 = \frac{(332 - 320)^2}{320} + \frac{(244 - 256)^2}{256} + \frac{(68 - 80)^2}{80} + \frac{(76 - 64)^2}{64} = 5.0625 \quad \text{on 1 df}
\]

Therefore, the \( p \)-value = \( P(\chi^2_1 \geq 5.0625) = .0244 \), as before. Reject \( H_0 \).

**Comments:**

- Chi-squared Test is valid, provided Expected Values \( \geq 5 \). (Otherwise, the score is inflated.) For small expected values in a \( 2 \times 2 \) table, defer to Fisher’s Exact Test.

- Chi-squared statistic with **Yates continuity correction** to reduce spurious significance:

\[
X^2 = \sum_{\text{all cells}} \frac{(\text{Obs} - \text{Exp}) - 0.5)^2}{\text{Exp}}
\]

- Chi-squared Test is strictly for the two-sided \( H_0: \pi_1 - \pi_2 = 0 \) vs. \( H_A: \pi_1 - \pi_2 \neq 0 \). It cannot be modified to a one-sided test, or to \( H_0: \pi_1 - \pi_2 = \pi_0 \) vs. \( H_A: \pi_1 - \pi_2 \neq \pi_0 \).
How could we solve this problem using R? The code (which can be shortened a bit):

```r
# Lines preceded by the pound sign are read as comments,
# and ignored by R.

# The following set of commands builds the 2-by-2 contingency table,
# column by column (with optional headings), and displays it as
# output (my boldface).

Tx.vs.Control = matrix(c(332, 68, 244, 76), ncol = 2, nrow = 2,
dimnames = list("Status" = c("Improvement", "No Improvement"),
"Group" = c("PT + Drug", "PT")))

Tx.vs.Control

<table>
<thead>
<tr>
<th>Group</th>
<th>PT + Drug</th>
<th>PT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Improvement</td>
<td>332</td>
<td>244</td>
</tr>
<tr>
<td>No Improvement</td>
<td>68</td>
<td>76</td>
</tr>
</tbody>
</table>

# A shorter alternative that outputs a simpler table:

Improvement = c(332, 244)
No_Improvement = c(68, 76)
Tx.vs.Control = rbind(Improvement, No_Improvement)

Tx.vs.Control

[,1] [,2]
| Improvement     | 332   | 244 |
| No_Improvement   | 68    | 76  |

# The actual Chi-squared Test itself. Since using a correction
# factor is the default, the F option specifies that no such
# factor is to be used in this example.

chisq.test(Tx.vs.Control, correct = F)

**Pearson's Chi-squared test**

data:  Tx.vs.Control
X-squared = 5.0625, df = 1, p-value = 0.02445

Note how the output includes the Chi-squared test statistic, degrees of freedom, and
p-value, all of which agree with our previous manual calculations.
**Application: Case-Control Study Design**

Determines if an association exists between disease $D$ and risk factor exposure $E$.

*Investigate:* Relation with $E^+$ and $E^-$

*Given:* Cases ($D^+$) and Controls ($D^-$)

---

### Chi-Squared Test

**$H_0$:** $\pi_{E^+ | D^+} = \pi_{E^+ | D^-}$

Randomly select a sample of cases and controls, and categorize each member according to whether or not he/she was exposed to the risk factor.

**SAMPLE**

<table>
<thead>
<tr>
<th></th>
<th>$n_1$ cases</th>
<th>$n_2$ controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D^+$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$E^+$</td>
<td>$a$</td>
<td>$b$</td>
</tr>
<tr>
<td>$E^-$</td>
<td>$c$</td>
<td>$d$</td>
</tr>
</tbody>
</table>

For each *case* ($D^+$), there are 2 disjoint possibilities for exposure: $E^+$ or $E^-$. For each *control* ($D^-$), there are 2 disjoint possibilities for exposure: $E^+$ or $E^-$. Calculate the $\chi^2$ statistic:

$$\chi^2 = \frac{(a + b + c + d)(ad - bc)^2}{(a + c)(b + d)(a + b)(c + d)}$$

---

### McNemar’s Test

**$H_0$:** $\pi_{E^+ | D^+} = \pi_{E^+ | D^-}$

Match each case with a corresponding control on age, sex, race, and any other confounding variables that may affect the outcome. Note that this requires a **balanced** sample: $n_1 = n_2$.

**SAMPLE**

<table>
<thead>
<tr>
<th></th>
<th>$n$ cases</th>
<th>$n$ controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D^+$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$E^+$</td>
<td>$a$</td>
<td>$b$</td>
</tr>
<tr>
<td>$E^-$</td>
<td>$c$</td>
<td>$d$</td>
</tr>
</tbody>
</table>

For each *matched case-control* ordered pair ($D^+, D^-$), there are 4 disjoint possibilities for exposure:

- **concordant pair**: $E^+$ and $E^+$ or $E^-$ and $E^-$
- **discordant pair**: $E^+$ and $E^-$ or $E^-$ and $E^+$

Calculate the $\chi^2$ statistic:

$$\chi^2 = \frac{(b - c)^2}{b + c}$$

See *Appendix > Statistical Inference > Means and Proportions, One and Two Samples.*
To quantify the *strength* of association between $D$ and $E$, we turn to the notion of…

**Odds Ratios – Revisited**

*Recall:*

**POPULATION**

**Case-Control Studies:**

$$OR = \frac{\text{odds(Exposure} \mid \text{Disease})}{\text{odds(Exposure} \mid \text{No Disease})} = \frac{P(E+ \mid D+) / P(E- \mid D+)}{P(E+ \mid D-) / P(E- \mid D-)}$$

**Cohort Studies:**

$$OR = \frac{\text{odds(Disease} \mid \text{Exposure})}{\text{odds(Disease} \mid \text{No Exposure})} = \frac{P(D+ \mid E+) / P(D- \mid E+)}{P(D+ \mid E-) / P(D- \mid E-)}$$

$H_0$: $OR = 1 \iff$ No association exists between $D, E$.

versus…

$H_A$: $OR \neq 1 \iff$ An association exists between $D, E$.

**SAMPLE, size $n$**

\[
\begin{array}{cc}
D+ & D- \\
E+ & a & b \\
E- & c & d \\
\end{array}
\]

$$\bar{OR} = \frac{ad}{bc}$$

Alas, the probability distribution of the odds ratio $OR$ is distinctly skewed to the right. However, its natural logarithm, $\ln(OR)$, is approximately normally distributed, which makes it more useful for conducting the **Test of Association** above. Namely…

$$\left(1 - \alpha\right) \times 100\% \text{ Confidence Limits for } \ln(OR)$$

$$\ln(\overline{OR}) \pm (z_{\alpha/2}) \text{ s.e.}, \quad \text{where } \hat{\text{s.e.}} = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

$$\left(1 - \alpha\right) \times 100\% \text{ Confidence Limits for } OR$$
Examples: Test $H_0: \ OR = 1$ versus $H_A: \ OR \neq 1$ at the $\alpha = .05$ significance level.

<table>
<thead>
<tr>
<th></th>
<th>$D+$</th>
<th>$D-$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E+$</td>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>$E-$</td>
<td>10</td>
<td>32</td>
</tr>
</tbody>
</table>

$\widehat{OR} = \frac{(8)(32)}{(10)(10)} = 2.56$

$\ln(2.56) = 0.94$

$\text{s.e.} = \sqrt{\frac{1}{8} + \frac{1}{10} + \frac{1}{10} + \frac{1}{32}} = 0.6 \Rightarrow 95\% \text{ Margin of Error} = (1.96)(0.6) = 1.176$

$95\% \text{ Confidence Interval for } \ln(OR) = (0.94 - 1.176, 0.94 + 1.176) = (-0.236, 2.116)$

and so...

$95\% \text{ Confidence Interval for } OR = (e^{-0.236}, e^{2.116}) = (0.79, 8.30)$

$\boxed{0.79 \ 1 \ 2.56 \ 8.30}$

Conclusion: As this interval does contain the null value $OR = 1$, we cannot reject the hypothesis of non-association at the 5% significance level.

<table>
<thead>
<tr>
<th></th>
<th>$D+$</th>
<th>$D-$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E+$</td>
<td>40</td>
<td>50</td>
</tr>
<tr>
<td>$E-$</td>
<td>50</td>
<td>160</td>
</tr>
</tbody>
</table>

$\widehat{OR} = \frac{(40)(160)}{(50)(50)} = 2.56$

$\ln(2.56) = 0.94$

$\text{s.e.} = \sqrt{\frac{1}{40} + \frac{1}{50} + \frac{1}{50} + \frac{1}{160}} = 0.267 \Rightarrow 95\% \text{ Margin of Error} = (1.96)(0.267) = 0.523$

$95\% \text{ Confidence Interval for } \ln(OR) = (0.94 - 0.523, 0.94 + 0.523) = (0.417, 1.463)$

and so...

$95\% \text{ Confidence Interval for } OR = (e^{0.417}, e^{1.463}) = (1.52, 4.32)

$\boxed{1 \ 1.52 \ 2.56 \ 4.32}$

Conclusion: As this interval does not contain the null value $OR = 1$, we can reject the hypothesis of non-association at the 5% level. With 95% confidence, the odds of disease are between 1.52 and 4.32 times higher among the exposed than the unexposed.

Comments:

- If any of $a, b, c, \text{ or } d = 0$, then use $\text{s.e.} = \sqrt{\frac{1}{a+0.5} + \frac{1}{b+0.5} + \frac{1}{c+0.5} + \frac{1}{d+0.5}}$.

- If $OR < 1$, this suggests that exposure might have a protective effect, e.g., daily calcium supplements (yes/no) and osteoporosis (yes/no).
Summary Odds Ratio

Combining $2 \times 2$ tables corresponding to distinct strata.

Examples:

<table>
<thead>
<tr>
<th></th>
<th>Males</th>
<th>Females</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$D^+$</td>
<td>$D^-$</td>
<td>$E^+$</td>
</tr>
<tr>
<td>$E^+$</td>
<td>10</td>
<td>50</td>
<td>80</td>
</tr>
<tr>
<td>$E^-$</td>
<td>10</td>
<td>150</td>
<td>20</td>
</tr>
</tbody>
</table>

$\hat{OR}_1 = 3$

$\hat{OR}_2 = 1$

$\hat{OR} = 3$

$\hat{OR}_2 = 2$

$\hat{OR} = 2$

$\hat{OR} = 5.0625$

$\hat{OR}_1 = 2$

$\hat{OR}_2 = 1$

$\hat{OR} = 1$

These examples illustrate the phenomenon known as Simpson’s Paradox.

Ignoring a confounding variable (e.g., gender) may obscure an association that exists within each stratum, but not observed in the pooled data, and thus must be adjusted for. When is it acceptable to combine data from two or more such strata? How is the summary odds ratio $OR_{\text{summary}}$ estimated? And how is it tested for association?
In general...

<table>
<thead>
<tr>
<th>Stratum 1</th>
<th>Stratum 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>D+</td>
<td>D+</td>
</tr>
<tr>
<td>E+</td>
<td>E+</td>
</tr>
<tr>
<td>( a_1 )</td>
<td>( a_2 )</td>
</tr>
<tr>
<td>( b_1 )</td>
<td>( b_2 )</td>
</tr>
<tr>
<td>( c_1 )</td>
<td>( c_2 )</td>
</tr>
<tr>
<td>( d_1 )</td>
<td>( d_2 )</td>
</tr>
</tbody>
</table>

\[ \hat{OR}_1 = \frac{a_1 d_1}{b_1 c_1} \]

\[ \hat{OR}_2 = \frac{a_2 d_2}{b_2 c_2} \]

I. Calculate the estimates of \( OR_1 \) and \( OR_2 \) for each stratum, as shown.

II. Can the strata be combined? Conduct a “Breslow-Day” (Chi-squared) Test of Homogeneity for

\[ H_0: OR_1 = OR_2 \]

III. If accepted, calculate the Mantel-Haenszel Estimate of \( OR_{\text{summary}} \):

\[ \hat{OR}_{\text{MH}} = \frac{a_1 d_1}{n_1} + \frac{a_2 d_2}{n_2} \cdot \frac{b_1 c_1}{n_1} + \frac{b_2 c_2}{n_2} \]

IV. Finally, conduct a Test of Association for the combined strata

\[ H_0: OR_{\text{summary}} = 1 \]

either via confidence interval, or special \( \chi^2 \)-test (shown below).

Example:

<table>
<thead>
<tr>
<th></th>
<th>Males</th>
<th>Females</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>D+</td>
<td>D−</td>
</tr>
<tr>
<td>E+</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>E−</td>
<td>30</td>
<td>90</td>
</tr>
</tbody>
</table>

\[ \hat{OR}_1 = 1.5 \]

\[ \hat{OR}_2 = 1.2 \]

Assuming that the Test of Homogeneity \( H_0: OR_1 = OR_2 \) is conducted and accepted,

\[ \hat{OR}_{\text{MH}} = \frac{(10)(90)}{150} + \frac{(40)(90)}{240} = \frac{6 + 15}{4 + 12.5} = \frac{21}{16.5} = 1.273 \]

Exercise: Show algebraically that \( \hat{OR}_{\text{MH}} \) is a weighted average of \( \hat{OR}_1 \) and \( \hat{OR}_2 \).
To conduct a formal Chi-squared Test of Association $H_0: OR_{\text{summary}} = 1$, we calculate, for the $2 \times 2$ contingency table in each stratum $i = 1, 2, \ldots, s$.

<table>
<thead>
<tr>
<th>Observed</th>
<th>Expected</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td># diseased</td>
<td>vs. # diseased</td>
<td></td>
</tr>
<tr>
<td>$D+$</td>
<td>$R_{1i}$</td>
<td>$E_{1i} = \frac{R_{1i}C_{1i}}{n_i}$</td>
</tr>
<tr>
<td>$D-$</td>
<td>$R_{2i}$</td>
<td>$E_{2i} = \frac{R_{2i}C_{1i}}{n_i}$</td>
</tr>
<tr>
<td>$E+$</td>
<td>$a_i$</td>
<td>$V_i = \frac{R_{1i}R_{2i}C_{1i}C_{2i}}{n_i^2(n_i - 1)}$</td>
</tr>
<tr>
<td>$E-$</td>
<td>$b_i$</td>
<td></td>
</tr>
<tr>
<td>$C_{1i}$</td>
<td>$C_{2i}$</td>
<td>$n_i$</td>
</tr>
</tbody>
</table>

Therefore, summing over all strata $i = 1, 2, \ldots, s$, we obtain the following:

<table>
<thead>
<tr>
<th>Observed total, Diseased</th>
<th>Expected total, Diseased</th>
<th>Total Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposed: $O_1 = \sum a_i$</td>
<td>Exposed: $E_1 = \sum E_{1i}$</td>
<td>$V = \sum V_i$</td>
</tr>
<tr>
<td>Not Exposed: $O_2 = \sum c_i$</td>
<td>Not Exposed: $E_2 = \sum E_{2i}$</td>
<td></td>
</tr>
</tbody>
</table>

and the formal test statistic for significance is given by

$$X^2 = \frac{(O_1 - E_1)^2}{V} \sim \chi^2_1.$$ 

This formulation will appear again in the context of the Log-Rank Test in the area of Survival Analysis (section 8.3).

Example (cont’d):

For stratum 1 (males), $E_{11} = \frac{(30)(40)}{150} = 8$ and $V_1 = \frac{(30)(120)(40)(110)}{150^2(149)} = 4.725$.

For stratum 2 (females), $E_{12} = \frac{(90)(100)}{240} = 37.5$ and $V_2 = \frac{(90)(150)(100)(140)}{240^2(239)} = 13.729$.

Therefore, $O_1 = 50$, $E_1 = 45.5$, and $V = 18.454$, so that $X^2 = \frac{(4.5)^2}{18.454} = 1.097$ on 1 degree of freedom, from which it follows that the null hypothesis $H_0: OR_{\text{summary}} = 1$ cannot be rejected at the $\alpha = .05$ significance level, i.e., there is not enough empirical evidence to conclude that an association exists between disease $D$ and exposure $E$.

Comment: This entire discussion on Odds Ratios $OR$ can be modified to Relative Risk $RR$ (defined only for a cohort study), with the following changes: $\text{s.e.} = \sqrt{\frac{1}{a - R_1} + \frac{1}{c - R_2}}$, as well as $b$ replaced with row marginal $R_1$, and $d$ replaced with row marginal $R_2$, in all other formulas. [Recall, for instance, that $\hat{OR} = \frac{ad}{bc}$, whereas $\hat{RR} = \frac{aR_2}{R_1c}$, etc.]