6.2 Two Samples → Independent / Dependent (Paired, Matched)

§ 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 | ***Treatment Arm***

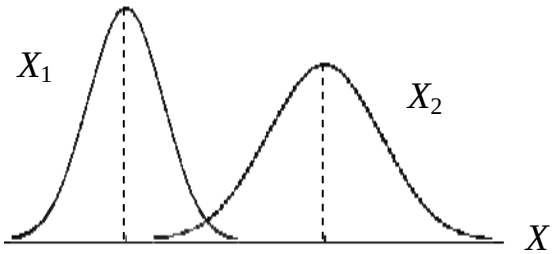
Assume	Assume
$X_1 \sim N(\mu_1, \sigma_1)$	$X_2 \sim N(\mu_2, \sigma_2)$

Then… ↓ **CLT** ↓

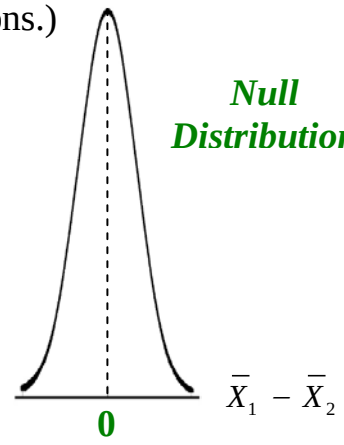
Sample, size n_1 | **Sample, size n_2**

$$\bar{X}_1 \sim N\left(\mu_1, \frac{\sigma_1}{\sqrt{n_1}}\right) \qquad \bar{X}_2 \sim N\left(\mu_2, \frac{\sigma_2}{\sqrt{n_2}}\right)$$

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



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



Comments:

- Recall from 4.1: If Y_1 and Y_2 are independent, then $\mathrm{Var}(Y_1 - Y_2) = \mathrm{Var}(Y_1) + \mathrm{Var}(Y_2)$.
- 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 μ_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 $\widehat{\text{s.e.}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$, provided $n_1 \geq 30$, $n_2 \geq 30$.

Example: X = "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.

Placebo	Drug
$n_1 = 80$	$n_2 = 60$
$\bar{x}_1 = 240$	$\bar{x}_2 = 229$
$s_1^2 = 1200$	$s_2^2 = 600$

$\rightarrow \quad \bar{x}_1 - \bar{x}_2 = 11$

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

$(1 - \alpha) \times 100\%$ Confidence Interval for $\mu_1 - \mu_2$

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

95% Confidence Interval for $\mu_1 - \mu_2$

95% limits = $11 \pm (1.96)(5) = 11 \pm 9.8$ ← **margin of error**

$\therefore$ 95% CI = **(1.2, 20.8)**, which does not contain **0** $\Rightarrow$ Reject H_0. Drug works!

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

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

95% Acceptance Region for $H_0: \mu_1 - \mu_2 = 0$

95% limits = $0 \pm (1.96)(5) = \pm 9.8$ ← **margin of error**

$\therefore$ 95% AR = **(–9.8, +9.8)**, which does not contain **11** $\Rightarrow$ Reject H_0. Drug works!

Test Statistic

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

***p*-value** $= 2\, P(\bar{X}_1 - \bar{X}_2 \geq 11)$

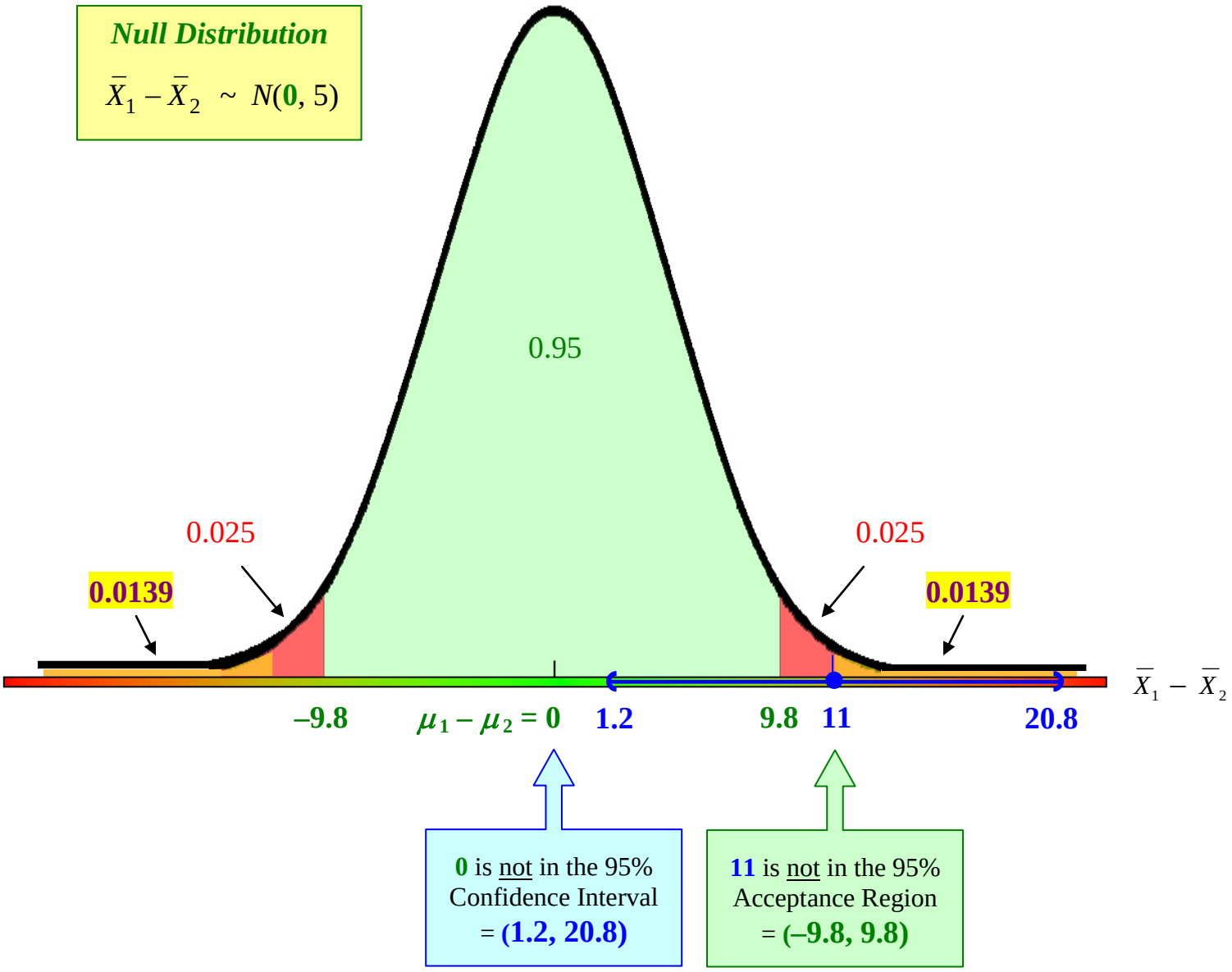
$= 2\, P\left(Z \geq \frac{11 - 0}{5}\right)$

$= 2\, P(Z \geq 2.2)$

$= 2(.0139)$

$= .0278 < .05 = \alpha$

$\Rightarrow$ Reject H_0. Drug works!


Null Distribution
$\bar{X}_1 - \bar{X}_2 \sim N(0, 5)$
0.95
0.025
0.025
0.0139
0.0139
$\bar{X}_1 - \bar{X}_2$
–9.8
$\mu_1 - \mu_2 = 0$
1.2
9.8
11
20.8
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*…

$$H_0: \ \sigma_1^2 = \sigma_2^2 \qquad \textbf{(equivariance, homoscedasticity)}$$

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 σ_1^2 and σ_2^2 can be estimated by the *weighted* mean of s_1^2 and s_2^2, the **pooled sample variance:**

$$s_{\text{pooled}}^2 = \frac{\text{df}_1 \, s_1^2 + \text{df}_2 \, s_2^2}{\text{df}_1 + \text{df}_2}, \ \text{where} \ \ \text{df}_1 = n_1 - 1 \ \ \text{and} \ \ \text{df}_2 = n_2 - 1,$$

i.e.,

$$s_{\text{pooled}}^2 = \frac{(n_1 - 1) \, s_1^2 + (n_2 - 1) \, s_2^2}{n_1 + n_2 - 2} = \frac{\textbf{SS}}{\textbf{df}} \, .$$

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

$$\widehat{\text{s.e.}} = \sqrt{\frac{s_{\text{pooled}}^2}{n_1} + \frac{s_{\text{pooled}}^2}{n_2}}$$

i.e.,

$$\widehat{\text{s.e.}} = \sqrt{s_{\text{pooled}}^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

$$= s_{\text{pooled}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \, .$$

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

$$\frac{\left(\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2} \right)^2}{\dfrac{(s_1^2/n_1)^2}{n_1 - 1} + \dfrac{(s_2^2/n_2)^2}{n_2 - 1}} \, .$$

This is known as the **Smith-Satterwaithe Test**. (Also used is the **Welch Test**.)

Example: X = "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.

Placebo	***Drug***
$n_1 = 8$	$n_2 = 10$
$\bar{x}_1 = 230$	$\bar{x}_2 = 200$
$s_1^2 = 775$	$s_2^2 = 1175$

$\rightarrow$ $\bar{x}_1 - \bar{x}_2 = \mathbf{30}$

$\rightarrow$ $F = s_1^2 / s_2^2 = 0.66$,
which is between 0.25 and 4.
Equivariance accepted $\Rightarrow$ ***t*-test** ✓

Pooled Variance

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

$\uparrow$
df

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

Standard Error

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

Critical Value

$t_{16,\,.025} = \mathbf{2.120}$

Margin of Error $= (2.120)(15) = \mathbf{31.8}$

$(1 - \alpha) \times 100\%$ Confidence Interval for $\mu_1 - \mu_2$

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

where $\text{df} = n_1 + n_2 - 2$

95% Confidence Interval for $\mu_1 - \mu_2$

95% limits = 30 ± 31.8 ← **margin of error**
$\therefore$ 95% CI = **(–1.8, 61.8)**, which contains **0** $\Rightarrow$ Accept H_0.

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

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

where $\text{df} = n_1 + n_2 - 2$

95% Acceptance Region for H_0: $\mu_1 - \mu_2 = 0$

95% limits = 0 ± 31.8 ← **margin of error**
$\therefore$ 95% AR = **(–31.8, +31.8)**, which contains **30** $\Rightarrow$ Accept H_0.

Test Statistic

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - \mu_0}{\sqrt{s_{\text{pooled}}^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim t_{\text{df}}$$

where $\text{df} = n_1 + n_2 - 2$

***p*-value** $= 2\,P(\bar{X}_1 - \bar{X}_2 \geq 30)$
$= 2\,P\left(T_{16} \geq \frac{30 - 0}{15}\right)$
$= 2\,P(T_{16} \geq 2.0)$
$= 2(.0314)$
$= .0628 > .05 = \alpha$

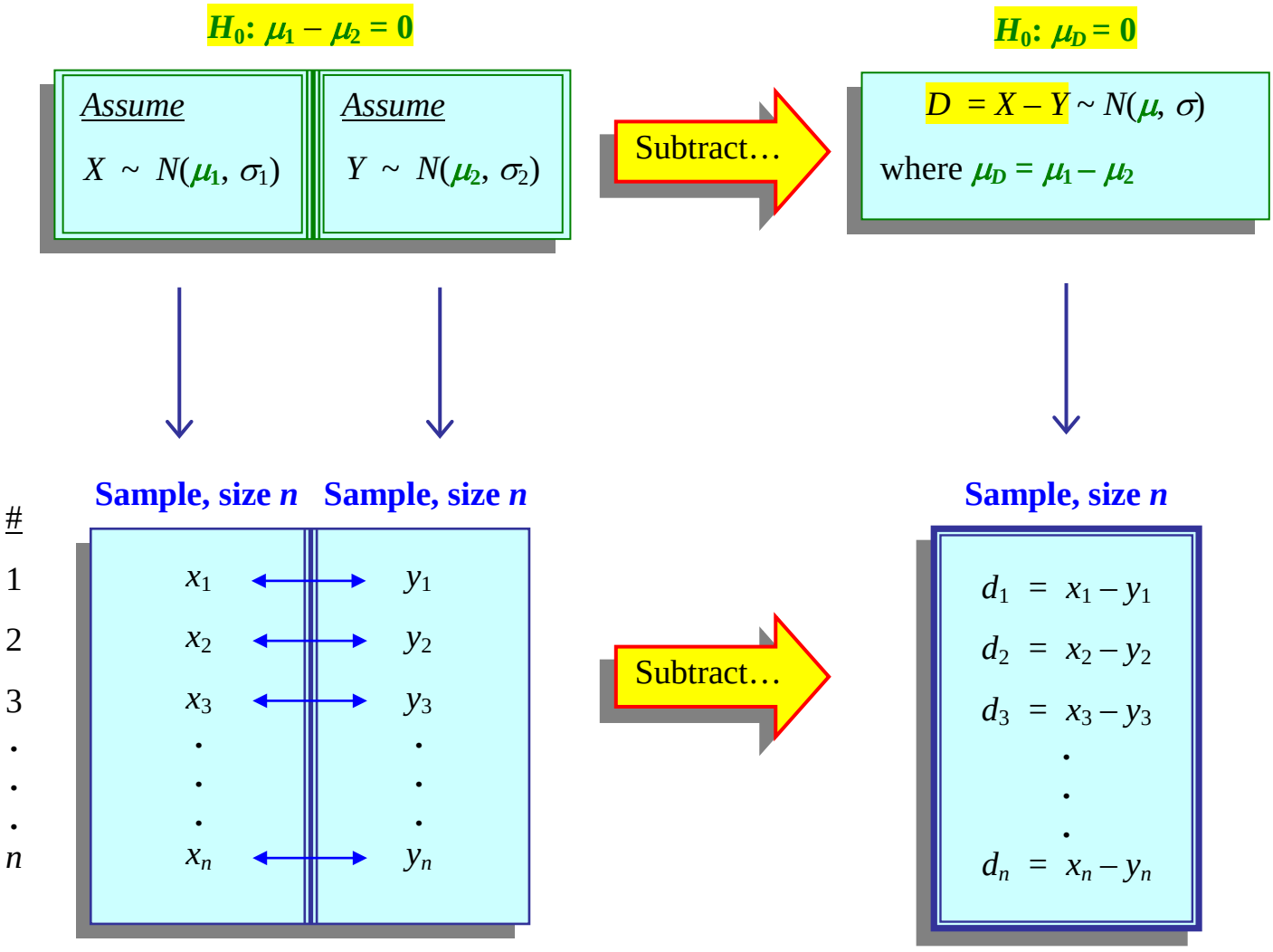
$\Rightarrow$ Accept H_0.

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).



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: $\mu_D = 0$.

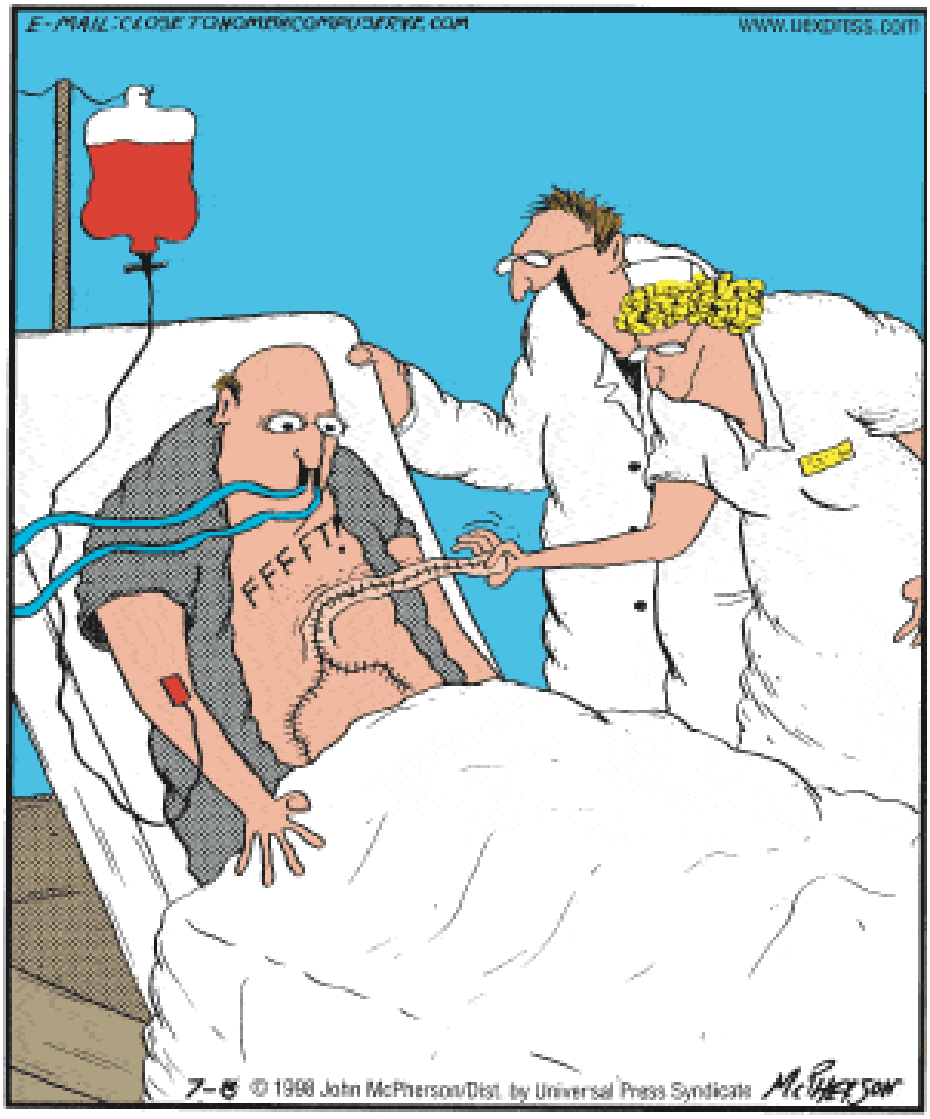
Checks for normality

include **normal scores plot** (**probability plot**, **Q-Q plot**), etc., just as with one sample.

Remedies for non-normality

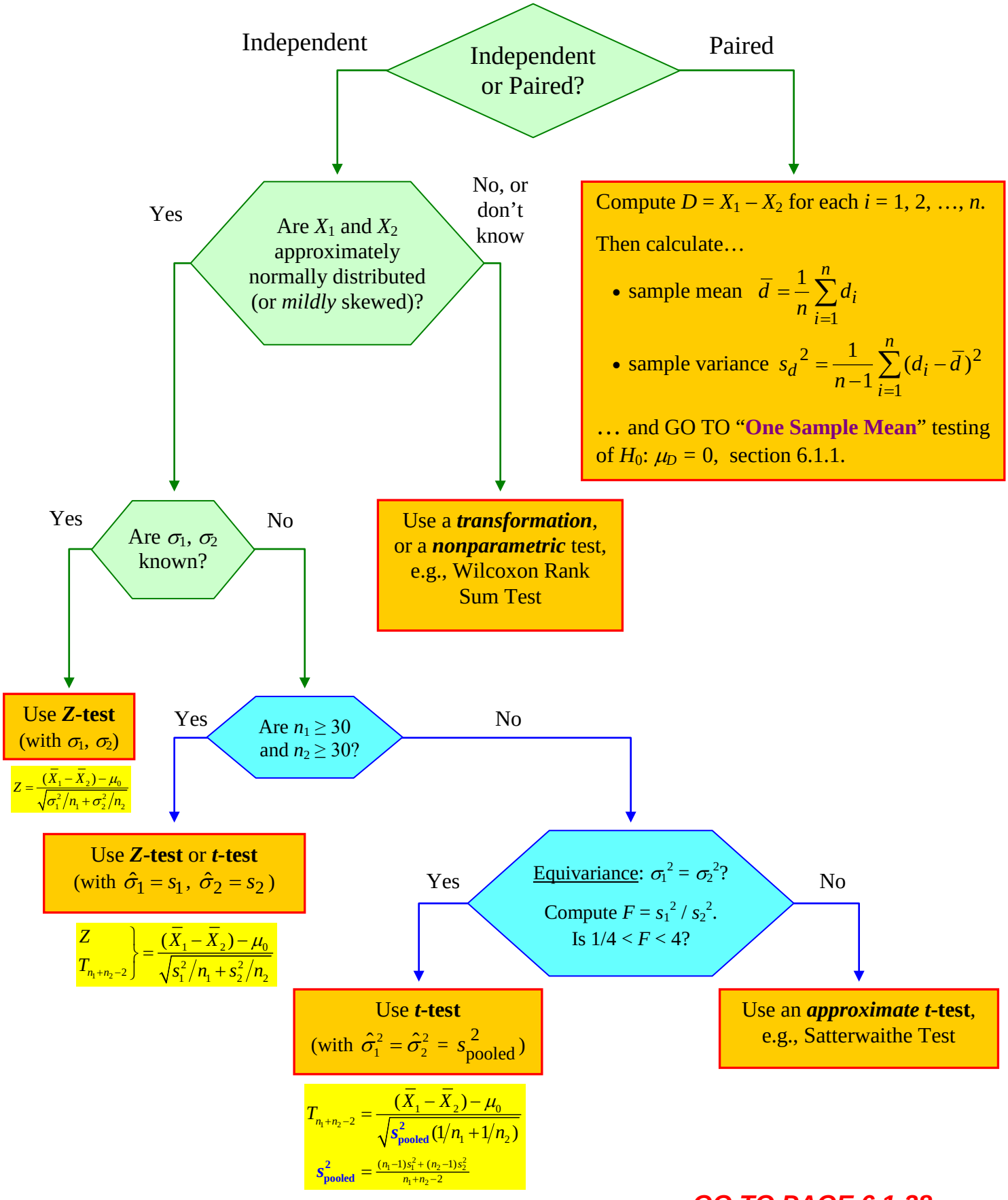
include **transformations** (e.g., logarithmic or square root), or **nonparametric tests.**

- **Independent Samples:** **Wilcoxon Rank Sum Test** (= **Mann-Whitney U Test**)
- **Dependent Samples:** **Sign Test, Wilcoxon Signed Rank Test**
  (just as with one sample)



"Now that you're fully recovered, Mr. Dawkins, we can tell you the truth. The12-hour operation, the intravenous meals, the three weeks of bed rest ... all were part of an elaborate placebo effect."

Step-by-Step Hypothesis Testing
Two Sample Means H_0: $\mu_1 - \mu_2$ vs. 0



…GO TO PAGE 6.1-28

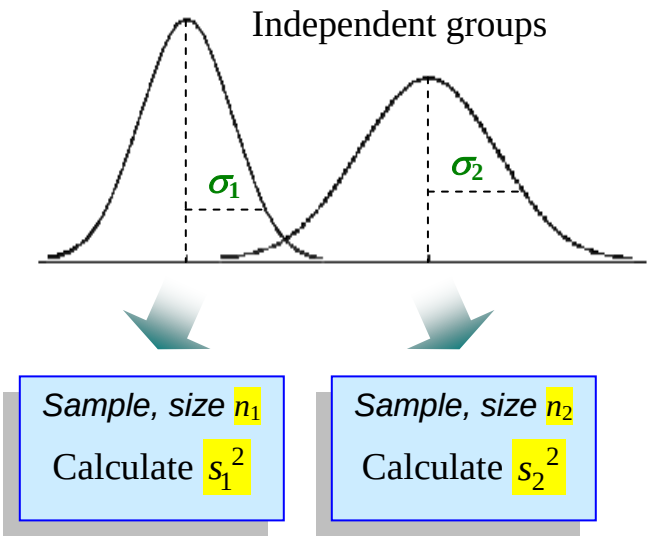
§ 6.2.2 Variances

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

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

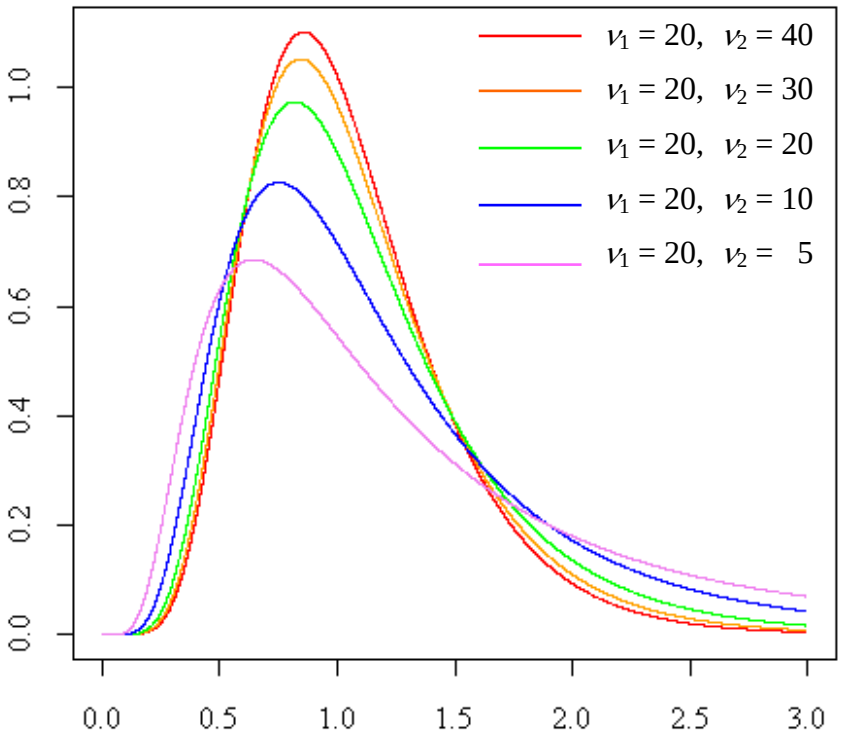
where $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$ are the corresponding numerator and denominator **degrees of freedom**, respectively.

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.

F-distribution

$$f(x) = \frac{1}{\mathrm{B}(\nu_1/2,\ \nu_2/2)} \left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2} x^{\nu_1/2 - 1} \left(1 + \frac{\nu_1}{\nu_2} x\right)^{-\nu_1/2 - \nu_2/2}$$



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

***Binary* random variable** I_1 = 1 or 0, with	***Binary* random variable** I_2 = 1 or 0, with
$P(I_1 = 1) = \pi_1$, $P(I_1 = 0) = \mathbf{1 - \pi_1}$	$P(I_2 = 1) = \pi_2$, $P(I_2 = 0) = \mathbf{1 - \pi_2}$

↓

INDEPENDENT SAMPLES

$n_1 \geq 30$	$n_2 \geq 30$
Random Variable $X_1 = \#(I_1 = 1) \sim \text{Bin}(n_1, \pi_1)$	**Random Variable** $X_2 = \#(I_2 = 1) \sim \text{Bin}(n_2, \pi_2)$
Recall (assuming $n_1\pi_1 \geq 15$, $n_1(1 - \pi_1) \geq 15$):	*Recall* (assuming $n_2\pi_2 \geq 15$, $n_2(1 - \pi_2) \geq 15$):
$\hat{\pi}_1 = \frac{X_1}{n_1} \sim N\left(\pi_1, \sqrt{\frac{\pi_1 (1 - \pi_1)}{n_1}}\right)$, approx.	$\hat{\pi}_2 = \frac{X_2}{n_2} \sim N\left(\pi_2, \sqrt{\frac{\pi_2 (1 - \pi_2)}{n_2}}\right)$, 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

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

as follows:

$(1 - \alpha) \times 100\%$ Confidence Interval for $\pi_1 - \pi_2$

$$\left((\hat{\pi}_1 - \hat{\pi}_2) - z_{\alpha/2} \sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}} \;,\; (\hat{\pi}_1 - \hat{\pi}_2) + z_{\alpha/2} \sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}} \right)$$

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…

$$\text{s.e.}_0 = \sqrt{\frac{\hat{\pi}_{\text{pooled}}\,(1 - \hat{\pi}_{\text{pooled}})}{n_1} + \frac{\hat{\pi}_{\text{pooled}}\,(1 - \hat{\pi}_{\text{pooled}})}{n_2}}$$

i.e.,

$$\text{s.e.}_0 = \sqrt{\hat{\pi}_{\text{pooled}}\,(1 - \hat{\pi}_{\text{pooled}})} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

Hence…

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

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

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	***PT only***
$n_1 = 400$	$n_2 = 320$
$X_1 = 332$	$X_2 = 244$

$$H_0: \pi_1 - \pi_2 = 0 \quad \textit{vs.} \quad H_A: \pi_1 - \pi_2 \neq 0 \quad \text{at } \alpha = .05$$

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

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

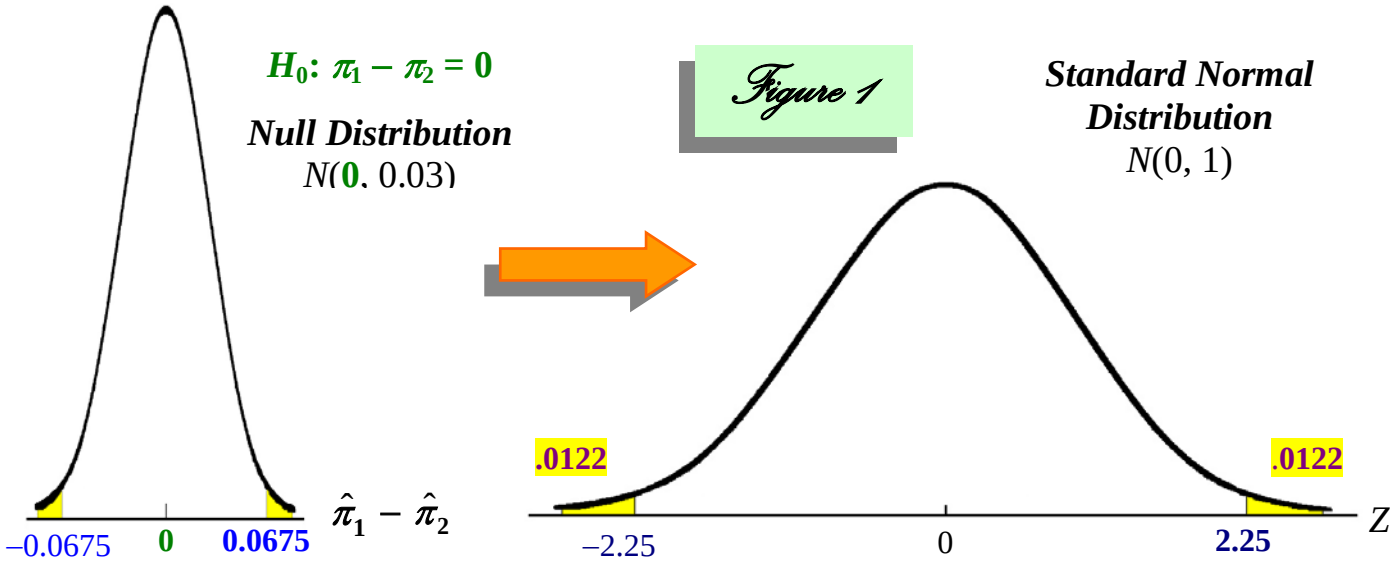
and thus $\mathbf{1 - \hat{\pi}_{\text{pooled}}} = \dfrac{\mathbf{144}}{\mathbf{720}} = \mathbf{0.2}$

$$\text{s.e.}_0 = \sqrt{(0.8)(0.2)}\sqrt{\frac{1}{400} + \frac{1}{320}} = \mathbf{0.03}$$

Therefore, p-value $=$

$$2\,P(\hat{\pi}_1 - \hat{\pi}_2 \geq 0.0675) = 2\,P\left(Z \geq \frac{0.0675 - 0}{0.03}\right) = 2\,P(Z \geq \mathbf{2.25}) = 2(.0122) = \mathbf{.0244}.$$

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 $\hat{\pi}_{\text{pooled}}$ 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 π_1 and π_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 (χ^2) Test**

Note: "Chi" is pronounced "kye"

As before, let the ***binary*** variable $I = 1$ for improvement, $I = 0$ for no improvement, with probability π 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 × 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

Group (J)

Status (I)	**PT + Drug**	**PT only**	
Improvement	332	244	576
No Improvement	68	76	144
	400	320	720

Row marginal totals

Column marginal totals

versus…

Expected Values = $\dfrac{\textit{Column total} \times \textit{Row total}}{\text{Total Sample Size } n}$

under H_0: $\underbrace{\pi_1 = \pi_2}$

$\hat{\pi}_{\text{pooled}} = 576/720 = 0.8$

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 **576/720**. Hence, the **expected** number (under H_0) of improved patients in the "PT + Drug" group is 400 × **576/720**, etc.

Group (J)

Status (I)	**PT + Drug**	**PT only**	
Improvement	$\frac{400 \times 576}{720} = 320.0$	$\frac{320 \times 576}{720} = 256.0$	576
No Improvement	$\frac{400 \times 144}{720} = 80.0$	$\frac{320 \times 144}{720} = 64.0$	144
	400.0	320.0	720

Note that, by construction,

H_0: $\frac{320}{400} = \frac{256}{320}$ ✓

$= \frac{576}{720}$, the pooled proportion.

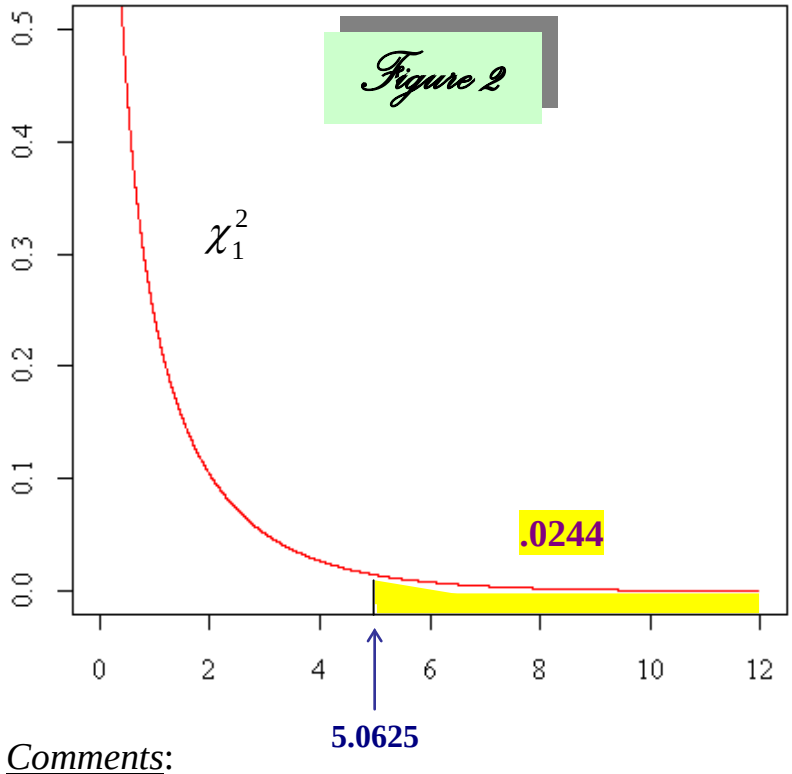
Test Statistic for H_0: $\pi_1 - \pi_2 = 0$

$$X^2 = \sum_{\text{all cells}} \frac{(\text{Obs} - \text{Exp})^2}{\text{Exp}} \sim \chi^2_1$$

Ideally, if all the observed values = all the expected values, then this statistic would = 0, and the corresponding p-value = 1. As it is,

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

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



Note that

$$\mathbf{5.0625} = (\pm\, \mathbf{2.25})^2,$$

i.e.,

$$\chi^2_1 = Z^2.$$

The two test statistics are mathematically equivalent! (Compare Figures 1 and 2.)

Comments:

- Chi-squared Test is valid, provided Expected Values ≥ 5. (Otherwise, the score is *inflated*.) For small expected values in a 2 × 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: $\pi_1 - \pi_2 = \mathbf{0}$ *vs.* H_A: $\pi_1 - \pi_2 \neq \mathbf{0}$. It cannot be modified to a one-sided test, or to H_0: $\pi_1 - \pi_2 = \boldsymbol{\pi_0}$ *vs.* H_A: $\pi_1 - \pi_2 \neq \boldsymbol{\pi_0}$.

How could we solve this problem using R? The code (which can be shortened a bit):

```
# 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
```

```
               Group
Status           PT + Drug  PT
  Improvement          332 244
  No Improvement        68  76
```

```
# 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

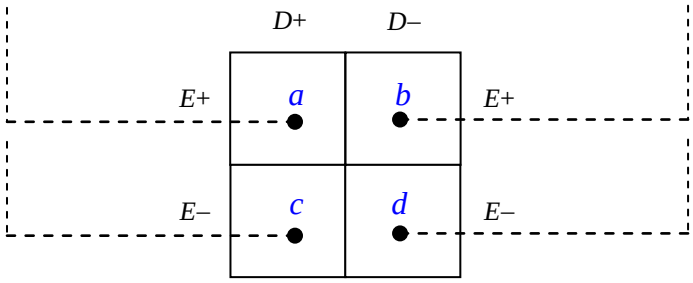
SAMPLE

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.

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 χ^2_1 statistic:

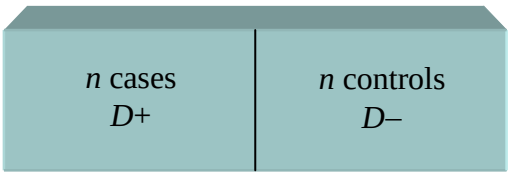
$$\frac{(a+b+c+d)\,(ad-bc)^2}{(a+c)\,(b+d)\,(a+b)\,(c+d)}$$

McNemar's Test

SAMPLE

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

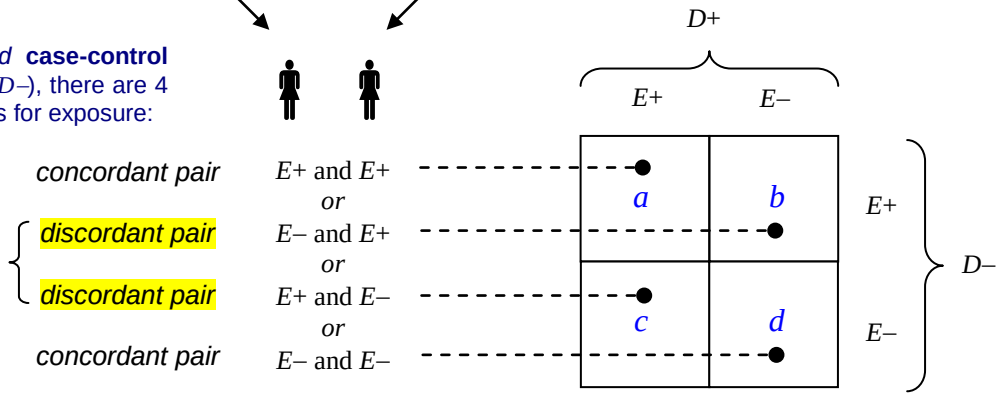
Match each case with a corresponding control on age, sex, race, and any other **confounding variables** that may affect the outcome. Note that this requires a **balanced** sample: $n_1 = n_2$.



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

Calculate the χ^2_1 statistic:

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



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:

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

Cohort Studies:

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

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

versus…

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

SAMPLE, size n

	$D+$	$D-$
$E+$	a	b
$E-$	c	d

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

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…

$(1 - \alpha) \times 100\%$ Confidence Limits for $\ln(OR)$

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

$(1 - \alpha) \times 100\%$ Confidence Limits for OR

Examples: Test $\boldsymbol{H_0}$**: *OR* = 1** versus $\boldsymbol{H_A}$**: *OR* ≠ 1** at the $\alpha = .05$ significance level.

	D+	*D*–
E+	8	10
E–	10	32

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

$\ln(2.56) = \mathbf{0.94}$

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

95% Confidence Interval for ln(*OR*) = **(0.94 – 1.176, 0.94 + 1.176)** = **(–0.236, 2.116)**

and so… **95% Confidence Interval for *OR*** = $(e^{-0.236}, e^{2.116})$ = **(0.79, 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.

	D+	*D*–
E+	40	50
E–	50	160

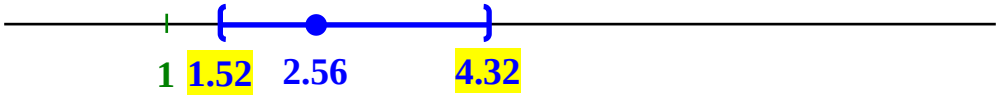
$$\widehat{OR} = \frac{(40)(160)}{(50)(50)} = \mathbf{2.56}$$

$\ln(2.56) = \mathbf{0.94}$

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

95% Confidence Interval for ln(*OR*) = **(0.94 – 0.523, 0.94 + 0.523)** = **(0.417, 1.463)**

and so… **95% Confidence Interval for *OR*** = $(e^{0.417}, e^{1.463})$ = **(1.52, 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*, or *d* = 0, then use $\widehat{\text{s.e.}} = \sqrt{\frac{1}{a+0.5}+\frac{1}{b+0.5}+\frac{1}{c+0.5}+\frac{1}{d+0.5}}$.
- 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 2×2 tables corresponding to distinct **strata**.

Examples:

Males

	$D+$	$D-$
$E+$	10	50
$E-$	10	150

$\widehat{OR}_1 = 3$

Females

	$D+$	$D-$
$E+$	10	10
$E-$	60	60

$\widehat{OR}_2 = 1$

$\rightarrow$

All

	$D+$	$D-$
$E+$	20	60
$E-$	70	210

$\widehat{OR} = 1$

???

Males

	$D+$	$D-$
$E+$	80	20
$E-$	20	10

$\widehat{OR}_1 = 2$

Females

	$D+$	$D-$
$E+$	10	20
$E-$	20	80

$\widehat{OR}_2 = 2$

$\rightarrow$

All

	$D+$	$D-$
$E+$	90	40
$E-$	40	90

$\widehat{OR} = 5.0625$

???

Males

	$D+$	$D-$
$E+$	60	100
$E-$	10	50

$\widehat{OR}_1 = 3$

Females

	$D+$	$D-$
$E+$	50	10
$E-$	100	60

$\widehat{OR}_2 = 3$

$\rightarrow$

All

	$D+$	$D-$
$E+$	110	110
$E-$	110	110

$\widehat{OR} = 1$

???

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_{summary} estimated? And how is it tested for association?

In general…

Stratum 1

	D+	D–
E+	a_1	b_1
E–	c_1	d_1

$$\widehat{OR}_1 = \frac{a_1\, d_1}{b_1\, c_1}$$

Stratum 2

	D+	D–
E+	a_2	b_2
E–	c_2	d_2

$$\widehat{OR}_2 = \frac{a_2\, d_2}{b_2\, c_2}$$

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

$$H_0\text{: } OR_1 = OR_2 \text{ .}$$

III. If accepted, calculate the **Mantel-Haenszel Estimate** of OR_{summary}:

$$\widehat{OR}_{\text{MH}} = \frac{\dfrac{a_1\, d_1}{n_1} + \dfrac{a_2\, d_2}{n_2}}{\dfrac{b_1\, c_1}{n_1} + \dfrac{b_2\, c_2}{n_2}} \text{ .}$$

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

$$H_0\text{: } OR_{\text{summary}} = 1$$

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

Example:

Males

	D+	D–
E+	10	20
E–	30	90

$$\widehat{OR}_1 = 1.5$$

Females

	D+	D–
E+	40	50
E–	60	90

$$\widehat{OR}_2 = 1.2$$

Assuming that the **Test of Homogeneity** H_0: $OR_1 = OR_2$ is conducted and accepted,

$$\widehat{OR}_{\text{MH}} = \frac{\dfrac{(10)(90)}{150} + \dfrac{(40)(90)}{240}}{\dfrac{(20)(30)}{150} + \dfrac{(50)(60)}{240}} = \frac{6 + 15}{4 + 12.5} = \frac{21}{16.5} = \mathbf{1.273} \text{ .}$$

Exercise: Show algebraically that $\widehat{OR}_{\text{MH}}$ is a *weighted average* of $\widehat{OR}_1$ and $\widehat{OR}_2$.

To conduct a formal Chi-squared **Test of Association** H_0: $OR_{summary} = 1$, we calculate, for the 2×2 contingency table in each stratum $i = 1, 2, \ldots, s$.

	Observed # diseased		*vs.*	**Expected # diseased**		**Variance**
	$D+$	$D-$				
$E+$	a_i	b_i	R_{1i}	$\rightarrow$	$E_{1i} = \frac{R_{1i}\, C_{1i}}{n_i}$	$V_i = \frac{R_{1i}\, R_{2i}\, C_{1i}\, C_{2i}}{n_i^2\,(n_i - 1)}$
$E-$	c_i	d_i	R_{2i}	$\rightarrow$	$E_{2i} = \frac{R_{2i}\, C_{1i}}{n_i}$	
	C_{1i}	C_{2i}	n_i			

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

Observed total, Diseased	**Expected total, Diseased**	**Total Variance**
Exposed: $O_1 = \sum a_i$	**Exposed:** $E_1 = \sum E_{1i}$	$V = \sum V_i$
Not Exposed: $O_2 = \sum c_i$	**Not Exposed:** $E_2 = \sum E_{2i}$	

and the formal test statistic for significance is given by

$$X^2 = \frac{(O_1 - E_1)^2}{V} \sim \chi_1^2.$$

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: $OR_{summary} = 1$ 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: $\widehat{\text{s.e.}} = \sqrt{\frac{1}{a} - \frac{1}{R_1} + \frac{1}{c} - \frac{1}{R_2}}$, 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 $\widehat{OR} = ad/bc$, whereas $\widehat{RR} = aR_2/R_1c$, etc.]