6.3 Several Samples

§ 6.3.1 Proportions

*General formulation:* Suppose that \( I \) and \( J \) are two general *categorical* variables, having \( r \) and \( c \) categories, respectively. Then there is a total of \( r \times c \) possible disjoint events: “\( I = i \) and \( J = j \)” for \( i = 1, 2, \ldots, r \) and \( j = 1, 2, \ldots, c \). Let \( \pi_{i|j} \) be the conditional probability \( P(I = i \mid J = j) = P(\{I = i\} \cap \{J = j\}) / P(J = j) \), so that \( \pi_{1|j} + \pi_{2|j} + \ldots + \pi_{r|j} = 1 \), for each \( j = 1, 2, \ldots, c \). We wish to test the null hypothesis that, for each category \( i \) of \( I \), the probabilities \( \pi_{i|j} \) are equal, over all the categories \( j \) of \( J \). That is,

\[
H_0: \quad \pi_{1|1} = \pi_{1|2} = \pi_{1|3} = \ldots = \pi_{1|c} \quad \text{and} \quad \pi_{2|1} = \pi_{2|2} = \pi_{2|3} = \ldots = \pi_{2|c} \quad \text{and} \quad \ldots \\
\ldots \\
\ldots \\
\ldots \\
\pi_{r|1} = \pi_{r|2} = \pi_{r|3} = \ldots = \pi_{r|c}
\]

versus...

\[
H_A: \quad \text{At least one of these equalities is false, i.e., } \pi_{i|j} \neq \pi_{i|k} \text{ for some } i.
\]

*Important Note:* We can formulate a similar but distinct null hypothesis, by considering probabilities \( \pi_{j|i} \) conditioned on the categories of \( I \), rather than \( J \) as above, but with a basically equivalent interpretation of non-association between them. Examples below...

Much as before, we can construct an \( r \times c \) *contingency table* of \( n \) observed values, where \( r = \# \text{ rows} \), and \( c = \# \text{ columns} \).
For $i = 1, 2, \ldots, r$ and $j = 1, 2, \ldots, c$, the following are obtained:

**Observed Values**  
$O_{ij} = \#(I = i, J = j)$  
*whole numbers $\geq 0$*

**Expected Values**  
$E_{ij} = \frac{R_i C_j}{n}$,  
*real numbers (i.e., with decimals) $\geq 0$*

where the **row marginals** $R_i = O_{i1} + O_{i2} + O_{i3} + \ldots + O_{ic}$,  
and the **column marginals** $C_j = O_{1j} + O_{2j} + O_{3j} + \ldots + O_{rj}$

**Test Statistic**  
$X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{\nu}$

where $\nu = df = (r - 1)(c - 1)$

**Comments:**

- Chi-squared Test is valid, provided 80% or more of $E_{ij} \geq 5$. For small expected values, **lumping** categories together increases the numbers in the corresponding cells. **Example:** The five age categories “18-24,” “25-39,” “40-49,” “50-64,” and “65+” in a might be lumped into three categories “18-39,” “40-64,” and “65+” if appropriate. **Caution:** Categories should be deemed contextually meaningful *before* using $\chi^2$.

- Remarkably, the same Chi-squared statistic can be applied in different scenarios, including tests of different null hypotheses $H_0$ on the same contingency table, as shown in the following examples.

- If $Z_1, Z_2, \ldots, Z_d$ are $N(0, 1)$ random variables, then $Z_1^2 + Z_2^2 + \ldots + Z_d^2 \sim \chi^2_d$. 
Example: Suppose that a study, similar to the previous one, compares \( r = 4 \) improvement responses of \( c = 3 \) groups of \( n = 600 \) patients: one group of 250 receives physical therapy alone, a second group of 200 receives an over-the-counter supplement in addition to physical therapy, and a third group of 150 receives a prescription medication in addition to physical therapy. The following \( 4 \times 3 \) contingency table of observed values is generated.

\[
\begin{array}{|c|c|c|c|}
\hline
\text{Improvement Status (I)} & \text{PT + Rx} & \text{PT + OTC} & \text{PT only} \\
\hline
\text{None} & 6 & 14 & 40 & 60 \\
\text{Minor} & 9 & 30 & 81 & 120 \\
\text{Moderate} & 15 & 60 & 105 & 180 \\
\text{Major} & 120 & 96 & 24 & 240 \\
\hline
\end{array}
\]

\text{fixed column marginal totals}

Upon inspection, it seems obvious that there are clear differences, but determining whether or not these differences are statistically significant requires a formal test. For instance, consider the null hypothesis that “there is no significant difference in each improvement response rate, across the treatment populations” – i.e., for each improvement category \( I \), the conditional probabilities \( P(I \mid J) \) over all treatment categories \( J \), are equal.

\[ H_0: \text{“Treatment populations are homogeneous with respect to each response.”} \]

\[
\begin{align*}
\pi_{\text{None} \mid \text{PT + Rx}} &= \pi_{\text{None} \mid \text{PT + OTC}} &= \pi_{\text{None} \mid \text{PT only}}, \\
\pi_{\text{Minor} \mid \text{PT + Rx}} &= \pi_{\text{Minor} \mid \text{PT + OTC}} &= \pi_{\text{Minor} \mid \text{PT only}}, \\
\pi_{\text{Moderate} \mid \text{PT + Rx}} &= \pi_{\text{Moderate} \mid \text{PT + OTC}} &= \pi_{\text{Moderate} \mid \text{PT only}}, \\
\pi_{\text{Major} \mid \text{PT + Rx}} &= \pi_{\text{Major} \mid \text{PT + OTC}} &= \pi_{\text{Major} \mid \text{PT only}}.
\end{align*}
\]

If the null hypothesis is true, then the expected table would consist of the values below,

\[
\begin{array}{|c|c|c|c|}
\hline
\text{Improvement Status (I)} & \text{PT + Rx} & \text{PT + OTC} & \text{PT only} \\
\hline
\text{None} & 15 & 20 & 25 & 60 \\
\text{Minor} & 30 & 40 & 50 & 120 \\
\text{Moderate} & 45 & 60 & 75 & 180 \\
\text{Major} & 60 & 80 & 100 & 240 \\
\hline
\end{array}
\]

\text{fixed column marginal totals}
because then...

\[
\begin{array}{ccc}
15 & 20 & 25 \\
150 & 200 & 250 \\
30 & 40 & 50 \\
45 & 60 & 75 \\
60 & 80 & 100 \\
150 & 200 & 250 \\
\end{array}
\]

(= pooled proportion \( \hat{\pi}_{\text{None}} = \frac{60}{600} \), true since all = 0.1 ✓

(= pooled proportion \( \hat{\pi}_{\text{Minor}} = \frac{120}{600} \), true since all = 0.2 ✓

(= pooled proportion \( \hat{\pi}_{\text{Mod}} = \frac{180}{600} \), true since all = 0.3 ✓

(= pooled proportion \( \hat{\pi}_{\text{Major}} = \frac{240}{600} \), true since all = 0.4. ✓

If the null hypothesis is rejected based on the data, then the alternative is that at least one of its four statements is false. For that corresponding improvement category, one of the three treatment populations is significantly different from the others. This is referred to as a Chi-squared Test of Homogeneity.

However, we can also consider another null hypothesis: “Within each treatment group, there is no significant difference in the improvement response rates.” That is, for each treatment category \( J \), the conditional probabilities \( P(J \mid I) \) down all improvement categories \( I \) are equal, and in particular, equal to the corresponding fixed column marginal probability \( P(J) \), that is, \( P(J \mid I) = P(J) \). But recall that, by definition, any two events \( A \) and \( B \) are statistically independent if \( P(A \mid B) = P(A) \). Therefore...

\( H_0: \) “Treatment and response are statistically independent.”

\[
\begin{array}{ccc}
\pi_{\text{PT + Rx | None}} & \text{and} & \pi_{\text{PT + OTC | None}} & \text{and} & \pi_{\text{PT only | None}} \\
\pi_{\text{PT + Rx | Minor}} & \text{and} & \pi_{\text{PT + OTC | Minor}} & \text{and} & \pi_{\text{PT only | Minor}} \\
\pi_{\text{PT + Rx | Moderate}} & \text{and} & \pi_{\text{PT + OTC | Moderate}} & \text{and} & \pi_{\text{PT only | Moderate}} \\
\pi_{\text{PT + Rx | Major}} & \text{and} & \pi_{\text{PT + OTC | Major}} & \text{and} & \pi_{\text{PT only | Major}} \\
\end{array}
\]

[Also, recall that an equivalent definition of independence is \( P(A \cap B) = P(A) \times P(B) \), so that, for example, \( P(\text{“None”} \cap \text{“PT + Rx”}) = P(\text{None}) \times P(\text{PT + Rx}) \), and likewise for all the other cells: \( P(\text{“} I = i \text{”} \cap \text{“} J = j \text{”}) = P( I = i ) \times P( J = j ) \). Rewriting this yields “expected value \( E_{ij} / n = (\text{“Row marginal } R_i \text{”} / n) \times (\text{“Column marginal } C_j \text{”} / n) \), i.e., the familiar formula \( E_{ij} = R_i C_j / n \).]
If the null hypothesis is true, then the same expected table would yield the following.

\[
\begin{array}{c|c|c|c}
\hat{\pi}_{PT+Rx} & \hat{\pi}_{PT+OTC} & \hat{\pi}_{PT\text{only}} \\
15/60 & 20/60 & 25/60 \\
= 30/120 & = 40/120 & = 50/120 \\
= 45/180 & = 60/180 & = 75/180 \\
= 60/240 & = 80/240 & = 100/240 \\
\end{array}
\]

\( (\hat{\pi}_{PT+Rx} = \frac{150}{600}) \), \( (\hat{\pi}_{PT+OTC} = \frac{200}{600}) \), \( (\hat{\pi}_{PT\text{only}} = \frac{250}{600}) \),

true, since all = 1/4.  true, since all = 1/3.  true, since all = 5/12.

This is referred to as a Chi-squared Test of Independence.

The same Chi-squared statistic

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

on df = \((r - 1)(c - 1)\) is used for both null hypothesis tests! The exact interpretation depends on the design of the experiment, i.e., whether two or more populations are being compared for homogeneity with respect to a set of responses, or whether two categorical variables are independent of one another. **MORAL:** In general, if the null hypothesis is rejected in either scenario, then there is an association between the two categorical variables I and J.

**Exercise:** Conduct (both versions of) the Chi-squared Test for this 4 × 3 table.
As a final application, consider one of the treatment categories alone, say “PT + Rx,” written below as a row, for convenience.

<table>
<thead>
<tr>
<th>PT + Rx</th>
<th>None</th>
<th>Minor</th>
<th>Moderate</th>
<th>Major</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed Values</td>
<td>6</td>
<td>9</td>
<td>15</td>
<td>120</td>
</tr>
</tbody>
</table>

$n = 150$

Suppose we wish to test the null hypothesis that there is “no significant difference in improvement responses,” i.e., the probabilities of all the improvement categories are equal. That is, $H_0: \pi_{\text{None}} = \pi_{\text{Minor}} = \pi_{\text{Moderate}} = \pi_{\text{Major}}$ (thus, $= 0.25$ each). Therefore, under this null hypothesis (and changing notation slightly), these $n = 150$ patients should be equally divided into the $k = 4$ response categories, i.e., $H_0$: “For this treatment category, the responses follow a uniform distribution ($= n/k$)” as illustrated.

<table>
<thead>
<tr>
<th>PT + Rx</th>
<th>None</th>
<th>Minor</th>
<th>Moderate</th>
<th>Major</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expected Values</td>
<td>37.5</td>
<td>37.5</td>
<td>37.5</td>
<td>37.5</td>
</tr>
</tbody>
</table>

$n = 150$

Of course, even a cursory comparison of these two distributions strongly suggests that there is indeed a significant difference. Remarkably, the same basic test statistic can be used in this Chi-squared Goodness-of-Fit Test. The “degrees of freedom” is equal to one less than $k$, the number of response categories being compared; in this case, df = 3.

In general, this test can be applied to determine if data follow other probability distributions as well. For example, suppose it is more realistic to believe that the null distribution is not uniform, but skewed, i.e., $H_0: \pi_{\text{None}} = .10, \pi_{\text{Minor}} = .20, \pi_{\text{Moderate}} = .30, \pi_{\text{Major}} = .40$. Then the observed values above would instead be compared with...

<table>
<thead>
<tr>
<th>PT + Rx</th>
<th>None</th>
<th>Minor</th>
<th>Moderate</th>
<th>Major</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expected Values</td>
<td>15</td>
<td>30</td>
<td>45</td>
<td>60</td>
</tr>
</tbody>
</table>

$n = 150$

**In general,**

**Goodness-of-Fit Test**

$H_0$: $\pi_1 = \pi_1^0, \pi_2 = \pi_2^0, \pi_3 = \pi_3^0, \ldots, \pi_k = \pi_k^0$

For $i = 1, 2, 3, \ldots, k =$ # groups, $n =$ sample size:

**Observed Values $O_i$** Expected Values $E_i = n \pi_i^0$

**Test Statistic**

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{df}$$

where $\nu = df = k - 1$

**Exercise:** Conduct this test for the “PT + Rx” data given, under both null hypotheses.
Inherited biological traits among humans (e.g., right- or left-handedness) and other organisms are transmitted from parents to offspring via “unit factors” called genes, discrete regions of DNA that are located on chromosomes, which are tightly coiled within the nucleus of a cell. Most human cells normally contain 46 chromosomes, arranged in 23 pairs (“diploid”); hence, two copies of each gene. Each copy can be either dominant (say, $A = \text{right-handedness}$) or recessive ($a = \text{left-handedness}$) for a given trait. The trait that is physically expressed in the organism—i.e., its phenotype—is determined by which of the three possible combinations of pairs $AA$, $Aa$, $aa$ of these two “alleles” $A$ and $a$ occurs in its genes—i.e., its genotype—and its interactions with environmental factors: $AA$ is “homozygous dominant” for right-handedness, $Aa$ is “heterozygous dominant” (or “hybrid”) for right-handedness, and $aa$ is “homozygous recessive” for left-handedness. However, reproductive cells (“gametes”: egg and sperm cells) only have 23 chromosomes, thus a single copy of each gene (“haploid”). When male and female parents reproduce, the “zygote” receives one gene copy—either $A$ or $a$—from each parental gamete, restoring diploidy in the offspring. With two traits, say handedness and eye color ($B = \text{brown}$, $b = \text{blue}$), there are nine possible genotypes: $AABB$, $AABb$, $AAbb$, $AaBB$, $Aabb$, $aaBB$, $aaBb$, $aabb$, resulting in four possible phenotypes. ($AaBb$ is known as a “dihybrid.”)

According to Mendel’s Law of Independent Assortment, segregation of the alleles of one allelic pair during gamete formation is independent of the segregation of the alleles of another allelic pair. Therefore, a homozygous dominant parent $AABB$ has gametes $AB$, and a homozygous recessive parent $aabb$ has gametes $ab$; crossing them consequently results in all dihybrid $AaBb$ offspring in the so-called $F_1$ (or “first filial”) generation, having gametes $AB$, $Ab$, $aB$, and $ab$, as shown below.
It follows that further crossing two such $\text{AaBb}$ genotypes results in expected genotype frequencies in the $F_2$ (“second filial”) generation that follow a 9:3:3:1 ratio, shown in the $4 \times 4$ Punnet square below.

<table>
<thead>
<tr>
<th>Male Gametes</th>
<th>Female Gametes</th>
</tr>
</thead>
<tbody>
<tr>
<td>AB</td>
<td>AB</td>
</tr>
<tr>
<td>AABB$^1$</td>
<td>AABB$^1$</td>
</tr>
<tr>
<td>AAbb$^2$</td>
<td>AaBb$^1$</td>
</tr>
<tr>
<td>AaBB$^1$</td>
<td>AaBb$^1$</td>
</tr>
<tr>
<td>aB</td>
<td>AaBb$^1$</td>
</tr>
<tr>
<td>ab</td>
<td>AaBb$^1$</td>
</tr>
</tbody>
</table>

Phenotypes | Expected Frequencies |
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1 = Right-handed, Brown-eyed</td>
<td>$9/16 = 0.5625$</td>
</tr>
<tr>
<td>2 = Right-handed, Blue-eyed</td>
<td>$3/16 = 0.1875$</td>
</tr>
<tr>
<td>3 = Left-handed, Brown-eyed</td>
<td>$3/16 = 0.1875$</td>
</tr>
<tr>
<td>4 = Left-handed, Blue-eyed</td>
<td>$1/16 = 0.0625$</td>
</tr>
</tbody>
</table>

For example, in a random sample of $n = 400$ such individuals, the expected phenotypic values under the null hypothesis $H_0$: $\pi_1 = 0.5625$, $\pi_2 = 0.1875$, $\pi_3 = 0.1875$, $\pi_4 = 0.0625$ are as follows.

<table>
<thead>
<tr>
<th>Expected Values</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>225</td>
<td>75</td>
<td>75</td>
<td>25</td>
</tr>
</tbody>
</table>

These would be compared with the observed values, say

<table>
<thead>
<tr>
<th>Observed Values</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>234</td>
<td>67</td>
<td>81</td>
<td>18</td>
</tr>
</tbody>
</table>

via the Chi-squared Goodness of Fit Test: $X^2 = \frac{(234-225)^2}{225} + \frac{(67-75)^2}{75} + \frac{(81-75)^2}{75} + \frac{(18-25)^2}{25} = 3.653$ on df = 3.

Because this is less than the .05 Chi-squared score of 7.815, the $p$-value is greater than .05 (its exact value = 0.301), and hence the data provide evidence in support of the 9:3:3:1 ratio in the null hypothesis, at the $\alpha = .05$ significance level. If this model had been rejected however, then this would suggest a possible violation of the original assumption of independent assortment of allelic pairs. This is indeed the case in genetic linkage, where the two genes are located in close proximity to one other on the same chromosome.

If two alleles $\text{A}$ and $\text{a}$ occur with respective frequencies $p$ and $q (= 1 – p)$ in a population, then observed genotype frequencies can be compared with those expected from The Hardy-Weinberg Law (namely $p^2$ for $\text{AA}$, $2pq$ for $\text{Aa}$, and $q^2$ for $\text{aa}$) via a similar Chi-squared Test.
§ 6.3.2 Variances

Consider $k$ independent groups $X_1 \sim N(\mu_1, \sigma_1)$, $X_2 \sim N(\mu_2, \sigma_2)$, ..., $X_k \sim N(\mu_k, \sigma_k)$. We wish to conduct a formal test for equivariance, or homogeneity of variances.

**Null Hypothesis** $H_0$: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = ... = \sigma_k^2$

**versus**

**Alternative Hypothesis** $H_A$: At least one of the $\sigma_i^2$ is different from the others.

**F-distribution**

The **Test Statistic** is defined as:

$$ F = \frac{s_{\text{max}}^2}{s_{\text{min}}^2} \sim F_{\nu_1, \nu_2} $$

where $\nu_1$ and $\nu_2$ are the corresponding numerator and denominator degrees of freedom, respectively.

**Formal test:** Reject $H_0$ if the $F$-statistic is significantly $> 1$.

**Comments:**

- Other tests: Levene (see § 6.2.2), Hartley, Cochran, Bartlett, and Scheffé.

- For what follows (ANOVA), moderate heterogeneity of variances is permissible, especially with large, approximately equal sample sizes $n_1, n_2, \ldots, n_k$. Hence this test is often not even performed in practice, unless the sample variances $s_1^2, s_2^2, \ldots, s_k^2$ appear to be greatly unequal.
§ 6.3.3  Means

Assume \( k \) independent, normally-distributed groups \( X_1 \sim N(\mu_1, \sigma_1), X_2 \sim N(\mu_2, \sigma_2), \ldots, X_k \sim N(\mu_k, \sigma_k) \), e.g., corresponding to different treatments. We wish to compare the treatment means with each other in order to determine if there is a significant difference among any of the groups. Hence...

\[
\begin{align*}
&\Leftrightarrow \\
&\quad \quad H_0: \quad \mu_1 = \mu_2 = \ldots = \mu_k
\end{align*}
\]

\( \Leftrightarrow \) \( \quad \quad H_0: \quad \text{"There is no difference in treatment means, i.e., no treatment effect."} \)

vs.

\( \quad \quad H_A: \quad \text{"There is at least one treatment mean } \mu_k \text{ that is different from the others."} \)

Key Strategy

The “total variation” in this system can be decomposed into two disjoint sources:

- variation between the groups (via a “treatment” \( s^2 \) measure)
- variation within the groups (as measured by \( s_{\text{pooled}}^2 \)).

If the former is significantly larger than the latter (i.e., if the ratio is significantly \( > 1 \)), then there must be a genuine treatment effect, and the null hypothesis can be rejected.

Recall (from the comment at the end of 2.3) that sample variance has the general form

\[
s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1} = \frac{\text{Sum of Squares}}{\text{degrees of freedom}} = \frac{\text{SS}}{\text{df}}.
\]

That is, \( \text{SS} = (n - 1) s^2 \). Using this fact, the powerful technique of Analysis of Variance (ANOVA) separates the total variation of the system into its two disjoint sources (known as “partitioning sums of squares”), so that a formal test statistic can then be formulated, and a decision regarding the null hypothesis ultimately reached. However, in order to apply this, it is necessary to make the additional assumption of equivariance, i.e, \( \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \ldots = \sigma_k^2 \), testable using the methods of the preceding section.
Example: For simplicity, take $k = 2$ balanced samples, say of size $n_1 = 3$ and $n_2 = 3$, from two independent, normally distributed populations:

\[ X_1: \{50, 53, 71\} \quad X_2: \{1, 4, 25\} \]

The null hypothesis $H_0: \mu_1 = \mu_2$ is to be tested against the alternative $H_A: \mu_1 \neq \mu_2$ at the $\alpha = .05$ level of significance, as usual. In this case, the difference in magnitudes between the two samples appears to be sufficiently substantial, that significance seems evident, despite the small sample sizes.

The following summary statistics are an elementary exercise:

\[ \bar{x}_1 = 58 \quad \bar{x}_2 = 10 \]
\[ s_1^2 = 129 \quad s_2^2 = 171 \]

Therefore,

\[ s_{pooled}^2 = \frac{(3 - 1)(129) + (3 - 1)(171)}{3 + 3} = \frac{600}{4} = 150. \]

We are now in a position to carry out formal testing of the null hypothesis.

**Method 1.** (Old way: two-sample $t$-test) In order to use the $t$-test, we must first verify equivariance $\sigma_1^2 = \sigma_2^2$. The computed sample variances of 129 and 171 are certainly sufficiently close that this condition is reasonably satisfied. (Or, check that the ratio 129/171 is between 0.25 and 4.) Now, recall from the formula for standard error, that:

\[ \text{s.e.} = \sqrt{150} \sqrt{1/3 + 1/3} = 10. \]

Hence,

\[ p\text{-value} = 2 \cdot P(\bar{X}_1 - \bar{X}_2 \geq 48) = 2 \cdot P\left(T_4 \geq \frac{48 - 0}{10}\right) = 2 \cdot P(T_4 \geq 4.8) = 2(0.0043) = 0.0086 < .05 \]

so the null hypothesis is (strongly) rejected; a significant difference exists at this level.
Method 2. (New way: ANOVA $F$-test) We first calculate three “Sums of Squares (SS)” that measure the variation of the system and its two component sources, along with their associated degrees of freedom (df).

1. **Total Sum of Squares** = sum of the squared deviations of each observation $x_{ij}$ from the grand mean $\bar{x}$.

\[
SS_{\text{Total}} = (50 - 34)^2 + (53 - 34)^2 + (71 - 34)^2 + (1 - 34)^2 + (4 - 34)^2 + (25 - 34)^2 = 4056
\]

\[
df_{\text{Total}} = (3 + 3) - 1 = 5
\]

2. **Treatment Sum of Squares** = sum of the squared deviations of each group mean $\bar{x}_i$ from the grand mean $\bar{x}$.

**Motivation**: In order to measure pure treatment effect, imagine two ideal groups with no “within group” variation, i.e., replace each sample value by its sample mean $\bar{x}_i$:

- $X_1'$: \{58, 58, 58\}
- $X_2'$: \{10, 10, 10\}

\[
SS_{\text{Trt}} = 3(58 - 34)^2 + 3(10 - 34)^2 = 3456
\]

\[
df_{\text{Trt}} = 1 \quad \text{Reason: As with any deviations, these must satisfy a single constraint:} \quad \text{namely, their sum} = 3(58 - 34) + 3(10 - 34) = 0. \quad \text{Hence their degrees of freedom} = \text{one less} \quad \text{than the number of treatment groups} (k = 2).
\]

3. **Error Sum of Squares** = sum of the squared deviations of each observation $x_{ij}$ from its group mean $\bar{x}_i$.

\[
SS_{\text{Error}} = (50 - 58)^2 + (53 - 58)^2 + (71 - 58)^2 + (1 - 10)^2 + (4 - 10)^2 + (25 - 10)^2 = 600
\]

\[
df_{\text{Error}} = (3 - 1) + (3 - 1) = 4
\]

\[SS_{\text{Total}} = SS_{\text{Trt}} + SS_{\text{Error}}
\]

\[df_{\text{Total}} = df_{\text{Trt}} + df_{\text{Error}}\]
### ANOVA Table

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS = SS / df</th>
<th>F = MS_{Trt} / MS_{Err}</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>1</td>
<td>3456</td>
<td>3456 ( = s_{between}^2)</td>
<td>23.04</td>
<td>.0086</td>
</tr>
<tr>
<td>Error</td>
<td>4</td>
<td>600</td>
<td>150 ( = s_{within}^2)</td>
<td>23.04</td>
<td>.0086</td>
</tr>
<tr>
<td>Total</td>
<td>5</td>
<td>4056</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The $F_{1, 4}$-score of **23.04** is certainly much greater than 1 (the expected value under the null hypothesis of no treatment difference), and is in fact greater than 7.71, the $F_{1, 4}$ critical value for $\alpha = .05$. Hence the small $p$-value, and significance is established. In fact, the ratio of $\frac{SS_{Trt}}{SS_{Total}} = \frac{3456}{4056} = 0.852$ indicates that **85.2% of the total variation in response is due to the treatment effect!**

**Comment:** Note that **23.04 = (± 4.8)^2**, i.e., $F_{1, 4} = t_4^2$. In general, $F_{1, df} = t_{df}^2$ for any df. Hence the two tests are mathematically equivalent to each other. Compare Figs 1 and 2.
General ANOVA formulation

Consider now the general case of \( k \) independent, normally-distributed, equivariant groups.

<table>
<thead>
<tr>
<th>Treatment Groups</th>
<th>( X_1 \sim N(\mu_1, \sigma_1) )</th>
<th>( X_2 \sim N(\mu_2, \sigma_2) )</th>
<th>( \ldots )</th>
<th>( X_k \sim N(\mu_k, \sigma_k) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Sizes</td>
<td>( n_1 )</td>
<td>( n_2 )</td>
<td>( \ldots )</td>
<td>( n_k )</td>
</tr>
<tr>
<td>Group Means</td>
<td>( \bar{x}_1 )</td>
<td>( \bar{x}_2 )</td>
<td>( \ldots )</td>
<td>( \bar{x}_k )</td>
</tr>
<tr>
<td>Group Variances</td>
<td>( s_1^2 )</td>
<td>( s_2^2 )</td>
<td>( \ldots )</td>
<td>( s_k^2 )</td>
</tr>
</tbody>
</table>

Grand Mean
\[
\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2 + \ldots + n_k \bar{x}_k}{n}
\]

Pooled Variance
\[
s_{\text{within}}^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \ldots + (n_k - 1)s_k^2}{n - k}
\]

Null Hypothesis \( H_0: \mu_1 = \mu_2 = \ldots = \mu_k \) \( \iff \) “No treatment difference exists.”

Alternative Hypothesis \( H_A: \mu_i \neq \mu_j \) for some \( i \neq j \) \( \iff \) “A treatment difference exists.”

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>( F )-statistic</th>
<th>( p )-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>( k - 1 )</td>
<td>( \sum_{i=1}^{k} n_i (\bar{x}_i - \bar{x})^2 )</td>
<td>( s_{\text{between}}^2 )</td>
<td>( F_{k-1, n-k} )</td>
<td>( 0 \leq p \leq 1 )</td>
</tr>
<tr>
<td>Error</td>
<td>( n - k )</td>
<td>( \sum_{i=1}^{k} (n_i - 1)s_i^2 )</td>
<td>( s_{\text{within}}^2 )</td>
<td>( F_{k-1, n-k} )</td>
<td>( 0 \leq p \leq 1 )</td>
</tr>
<tr>
<td>Total</td>
<td>( n - 1 )</td>
<td>( \sum_{all , i, j} (x_{ij} - \bar{x})^2 )</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Comments:

- This is referred to as the overall \( F \)-test of significance. If the null hypothesis is rejected, then (the mean value of at least) one of the treatment groups is different from the others. But which one(s)?
- Nonparametric form of ANOVA: Kruskal-Wallis Test
- Appendix > Geometric Viewpoint > ANOVA
**Multiple Comparison Procedures**

Comparisons between two means of individual groups can be $t$-tested, e.g., $H_0: \mu_{\text{placebo}} = \mu_{\text{drug dose 1}}$, or $H_0: \mu_{\text{placebo}} = \mu_{\text{drug dose 2}}$, or between certain “linear combinations” of means, e.g., $H_0: \frac{1}{2}(\mu_{\text{drug dose 1}} + \mu_{\text{drug dose 2}})$.

**A Priori Analysis (Planned Comparisons)**

- Investigator wishes to perform pairwise $t$-test comparisons on a fixed number $m$ of specific groups of interest, which are decided *before* any data are collected.

  **Example**: Suppose Groups 2, …, $k$ are to be compared with Group 1, which acts as a control. Then there will be $m = k - 1$ pairwise comparisons, with corresponding null hypotheses $H_0: \mu_1 = \mu_2$, $H_0: \mu_1 = \mu_3$, $H_0: \mu_1 = \mu_4$, …, $H_0: \mu_1 = \mu_k$.

  **Example**: Suppose it is decided to compare all possible pairs among Groups 1, …, $k$, i.e., $H_0: \mu_i = \mu_j$ for all $i \neq j$. Then there will be $m = \binom{k}{2} = \frac{k(k-1)}{2}$ such $t$-tests.

**CAUTION**: With a large number $m$ of such comparisons, there is an increased probability of finding a *spurious* significance between two groups, just by chance.

**Exercise**: Show that this probability $= 1 - (1 - \alpha)^m$, which goes to 1 as $m$ gets large.

**Bonferroni correction** reduces this risk by lowering the significance level of each $t$-test to $\alpha^* = \frac{\alpha}{m}$, thereby making each one more conservative, while at the same time preserving the overall “experiment-wise” Type I error rate $\alpha$. Thus, if comparing all $\binom{5}{2} = 10$ pairs of $k = 5$ groups with an overall significance level of $\alpha = .05$, then set $\alpha^* = \frac{.05}{10} = .005$ for each $t$-test. (Note: Use the ANOVA MS_{Error} term for $s_{\text{pooled}}^2$.) Note however, that Bonferroni correction can be overly conservative, and fail to reject differences known to be statistically significant; see **Holm’s method** for a modification.

**A Posteriori (or Post Hoc) Analysis (Unplanned Comparisons)**

- “Data-mining,” “data-dredging,” “fishing expedition,” done *after* sample data are collected. Unlike above, can only be used if the ANOVA overall $F$-test is significant.

- Fisher’s Least Significant Difference (LSD) Test, Tukey’s Honest Significant Difference (HSD) Test, Scheffé’s Test, Neumann-Keuls Test, Dunnett’s Test (only for control group versus treatment groups),…