

Chapter 1

Making Connections: The two-sample t-test, Regression and ANOVA

As a final project in an introductory statistics class, several students decided to conduct a study to test the strength of paper towels. Several television advertisements had claimed that a certain brand of paper towel was the strongest and these students wanted to determine if there really was a difference. The students sampled 26 towels from two brands of paper towels, Comfort and Decorator.

Before any data were collected, these students determined that the following should be held as constant as possible throughout the study:

- 15 drops of water were applied to the center of each towel,
- paper towels were selected that had the same size
- the towels were held at all four corners by two people, and
- weights (10, 25, 50 or 100 grams) were slowly added to the center of each towel by a third person until it broke.

The file, `PaperTowelData1`, shows the breaking strength of each towel. Breaking strength is defined as the total weight which each towel successfully held. The next additional weight caused the towel to break.

[On Your Own:]

1. For this paper towel study, identify the explanatory variable, the observational units (experimental unit or subject), and the response variable. Write out in words and symbols an appropriate null and alternative hypothesis.

Since it is not possible to test the entire population of Comfort (Called Brand C) and Decorator (called Brand D) paper towels, a representative sample was collected for each brand, and summary statistics calculated.

The **population mean** for Brand C, μ_1 , is estimated by the **sample mean**, which is the average of a **random sample** of n_1 responses $(y_{11}, y_{12}, \dots, y_{1n_1})$. The formula for the sample mean of Brand C is $\bar{y}_1 = \frac{1}{n_1} \sum_{j=1}^{n_1} y_{1j}$.

Notice the double set of subscripts on the observed responses. y_{11} is not “y eleven”, but “y one one”. The first subscript tells us that the observation was from the first group (Brand C). The second subscript tells the observation number. y_{1j} is the j^{th} observation from Brand C.

The **sample standard deviation**, $s_1 = \sqrt{\frac{1}{n_1-1} \sum_{j=1}^{n_1} (y_{1j} - \bar{y}_1)^2}$, is also calculated to estimate the **population standard deviation** of Brand C (σ_1). Similar calculations can be done to estimate the mean, μ_2 , and standard deviation, σ_2 , of the breaking strength of Brand D paper towels.

[KEY CONCEPT] **Sample statistics (\bar{x} and s) are used to estimate population parameters (μ and σ).** A statistic is any mathematical function of the sample data. Parameters are actual population values, that cannot be known unless the entire population is sampled.

[On Your Own:]

2. Calculate \bar{y}_1 , \bar{y}_2 , s_1 , and s_2 from the paper towel data: [In Minitab, open the file named PaperTowelData1. Click Stat → Basic Statistics → Display Descriptive Statistics. Double click weight and by brand . Click the Statistics button and verify that mean and standard deviation are selected. Click OK. Click OK].

Clearly the sample means (\bar{y}_1 and \bar{y}_2) are not equal, but how certain can we be that the true population means (μ_1 and μ_2) are different?

One student in the class was an economics major. He had seen hypothesis tests used in regression to analyze whether a certain variable belonged in a regression model. Could a hypothesis test in regression be used to test if a variable like brand was important in modeling paper towel strength?

Another student, a biology major, had previously completed an experiment in ecology where her professor told her to use an F test and **AN**alysis **O**f **VA**riance (ANOVA) to determine if a certain variable was significant. A third student had seen the t-test explained in her textbook and believed she should do whatever the textbook example did.

Each of these approaches seem reasonable, so how can you decide which technique to use? Instead of simply stating which student is right, this lab will discuss the two-sample t-test, linear regression and ANOVA while emphasizing the statistical modeling process.

Once a statistician is able to develop an appropriate statistical model, determining which test to use becomes much more evident. As studies get more complex, modeling variability becomes much more important to ensure that the appropriate conclusions are drawn from a

sample set of data.

Generally, **statistical models** have the following form:

$$\text{DATA} = \text{DETERMINISTIC MODEL} + \text{RANDOM ERROR},$$

while a **mathematical model** has the form

$$\text{DATA} = \text{DETERMINISTIC MODEL}.$$

Statistical modeling consists of using sample data or prior knowledge to develop deterministic models and then using appropriate statistical techniques to determine how confident we are that the models represent the population.

The statistical model includes a **random error** component that includes phenomena such as measurement error, variation caused by unobserved (or unobservable) components of the deterministic model, and natural variation. While we cannot predict each individual error term, in the long run there is a regular pattern.

This model shows two types of variability: variability we want (from the deterministic model) and variability we can live with (the random error). In the next sections we will show that the deterministic model represents changes in brand. If the two brands truly do have different breaking strengths, the deterministic model will show different expected means for each brand. The next section will also show that statisticians can model the chance-like behavior of the random error term.

There is one more type of variability that is not included in this model: **systematic errors** that were not included in the design of the study. This is the type of variability that can bias the results. For example, if the students had used different techniques to add weights (or water) to the Brand C towels than the Brand D towels, there would be no way to identify if the observed difference was due to brand or due to the different techniques.

[Key Concept] There are three types of variability that need to be considered in every statistical model. The deterministic model should represent the planned variability that is of interest in our study. The random error should follow an overall pattern and will be modeled with a distribution, like the normal distribution. Any other source of variability that may potentially bias the results must be held constant, incorporated into the model (becoming part of the deterministic model), or randomized (randomization will convert the bias terms into a chance-like variation that can be modeled with a distribution).

[Key Concept] In most studies the systematic errors must be accounted for before the data is collected. In our example, the paper towel study would have been meaningless if they had not controlled their techniques for adding water and weights while collecting the data.

1.1 The Two-Sample t-Test to Compare Population Means

1.1.1 The Statistical Model

The key question in this paper towels study is whether or not the two brands of paper towels have different mean breaking strengths. We will model every observation as being composed of a group mean, which is the deterministic component, and a random error term.

$$\begin{array}{ccccc}
 & & \text{deterministic} & & \\
 & & \text{model for} & & \\
 \text{observed} & & \text{the response} & & \text{error} \\
 \text{response} & & \text{(not random)} & & \text{(random)} \\
 \text{(random)} & & & & \\
 \downarrow & & \downarrow & & \downarrow \\
 y_{1j} & = & \mu_1 & + & \epsilon_{1j} \quad \text{for } j = 1, 2, \dots, n_1
 \end{array}$$

This model states that every observation sampled from the Brand C population is expected to be centered at the constant value μ_1 . In addition to being centered at μ_1 , the j^{th} observation is expected to have some variability that is modeled by ϵ_{1j} . The ϵ_{1j} are modeled as independent and identically distributed (*i.i.d.*) random variables. Independent means that each sample is independent of the other samples and identically distributed states that each sample comes from a population with the same mean and the same variance. Most often this is a normally distributed population with mean zero and a fixed variance σ_1^2 .

Similarly, any observation from the Brand D population can be modeled as a deterministic component, μ_2 , plus a random error term, ϵ_{2j} :

$$y_{2j} = \mu_2 + \epsilon_{2j} \quad \text{for } j = 1, 2, \dots, n_2$$

where the ϵ_{2i} are *i.i.d* random variables with mean zero and variance σ_2^2 . Often, this statistical model is more succinctly written as

$$y_{ij} = \mu_i + \epsilon_{ij} \quad \text{for } i = 1, 2 \quad \text{and } j = 1, 2, \dots, n_j$$

1.1.2 Model Assumptions

While this model is fairly straight forward, there are several conditions that need to be satisfied before a hypothesis test can be developed:

- the error terms are independent and identically distributed
- the population variances are equal ($\sigma_1^2 = \sigma_2^2$) and
- the error terms follow a normal probability distribution: denoted as $\epsilon_{ij} \sim N(0, \sigma^2)$. Note that $\sigma^2 = \sigma_1^2 = \sigma_2^2$.

The “independent” assumption states that there is no relationship between one observation and the next. For example, knowing that the 8th sample towel was stronger than average does not provide any information about whether the 7th or 9th samples will be above or below the average.

The “identically distributed” assumption also states that each observation sampled within each brand is from a “population” with the same mean and variance. If any sample observation comes from a different population, the two-sample t-test will not be valid. For example, if the students accidentally spilled too much water on a paper towel, we would expect that the mean breaking strength of that paper towel may be different than the others. It would be inappropriate to include this towel in the study.

To check the assumption that the two populations have the same variance an informal test can be used. If the ratio of the sample standard deviations is less than 2, we can proceed with the analysis.

$$\text{If } \frac{\max(s_1, s_2)}{\min(s_1, s_2)} < 2, \quad \text{then we do not have enough evidence to reject } \sigma_1^2 = \sigma_2^2$$

The only model assumption left to check is that the random errors follow a normal distribution. If sample sizes are similar and large ($n_1 \geq 20$ and $n_2 \geq 20$) and without extreme skewness or outliers, then the assumption of normality is not critical to the validity of the t-test because of the Central Limit Theorem.

[KEY CONCEPT] Every statistical hypothesis test has basic underlying conditions that need to be checked before any conclusions can be drawn. Even the most complex statistical formulas may be useless if the data was not collected properly.

[On Your Own:]

3. Determine if it is appropriate to assume equal variances for the paper towel data.
4. Create one or two graphs to determine if extreme skewness or outliers are present in these data. Does it seem appropriate to assume that the error terms are *i.i.d.* and $N(0, \sigma^2)$?

1.1.3 The Two-Sample t-Test

The two-sample t-test can be used to test whether the two population means are equal. The null hypothesis about the populations ($H_0 : \mu_1 - \mu_2 = 0$) is rejected if the sample means, \bar{y}_1 and \bar{y}_2 , are far enough apart compared to the spread in the data.

The test statistic for the two-sample t-test is:

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \text{where } s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}. \quad (1.1.1)$$

We calculate $\bar{y}_1, \bar{y}_2, s_1^2$, and s_2^2 from the samples, while $(\mu_1 - \mu_2)$ comes from the statement of the null hypothesis: $\mu_1 - \mu_2 = 0$ (or, equivalently, $\mu_1 = \mu_2$). Thus the test statistic is simply a ratio of the distance between the two sample means over a measure of variation.

The **pooled standard deviation**, denoted s_p , is used to estimate the population standard deviation. Since we assume that $\sigma_1 = \sigma_2$, s_p uses a weighted average of the two sample variances in order to estimate the common variance σ^2 , and then we take the square root to get an estimate for the standard deviation.

Probability theory can be used to prove that if the model assumptions are true, the t-statistic in equation (1.1.1) follows a t-distribution with $(n_1 + n_2 - 2)$ degrees of freedom.

We will reject the null hypothesis that the two means are equal ($H_0 : \mu_1 = \mu_2$) if the value of this t-statistic corresponds to a small enough p-value, which is found with software or in a t-table.

[On Your Own:]

5. Calculate the test statistic (t) shown in equation (1.1.1) and find the p-value corresponding to this statistic. Use statistical software to validate your work. [In Minitab select **Stat** → **Basic Statistics** → **2-sample t**. Select **Samples in one column** and double click **Weight** and **Brand**. Select **Assume equal Variances**. Select **Options** and set **Test difference** to 0 and **Alternative** to **not equal**. Click **OK**. Click **OK**.] If $H_0 : \mu_1 = \mu_2$ is true, the **p-value** states how likely that just random sampling variability would create a difference between two sample means $(\bar{y}_1 - \bar{y}_2)$ at least as large as we observed. Based on the p-value, what can you conclude about these two brands of paper towels?

6. **OPTIONAL** Look up the test-statistic in a standard normal Z-table (or use Minitab) and find the p-value. Is this the p-value similar to that found using the t-table? Is it larger? Smaller? Explain why you would expect this to occur.

1.2 The Regression Model to Compare Population Means

1.2.1 The Linear Regression Model

The simple linear regression model discussed in introductory statistics courses typically has the following form:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad \text{for } i = 1, 2, \dots, n$$

For this linear regression model, the deterministic component $(\beta_0 + \beta_1 x_i)$ is a linear function of two parameter(s), β_0 and β_1 and an explanatory variable x . The random error terms, ϵ_i , are assumed to be independent and to follow a $N(0, \sigma^2)$ distribution.

Later chapters will describe how least squares techniques can be used to calculate the following statistics to estimate β_0 and β_1 :

$$b_1 = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2} \quad b_0 = \frac{\sum_{i=1}^n y_i - b_1 \sum_{i=1}^n x_i}{n} . \quad (1.2.1)$$

In most introductory statistics texts, the above equations for the slope and intercept are simplified to:

$$b_1 = r \frac{s_y}{s_x} \quad \text{and} \quad b_0 = \bar{y} - b_1 \bar{x}, \quad \text{where } r = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{y_i - \bar{y}}{s_y} \right) \left(\frac{x_i - \bar{x}}{s_x} \right)$$

is the sample correlation coefficient.

For the paper towel data, $n = 52 = n_1 + n_2$. Instead of writing two distinct models, as in the two-sample t-test, there is one regression model where the explanatory variable, x , will be used to indicate brand.

A common procedure used to incorporate categorical explanatory variables, such as brand, into a regression model is to define **dummy variables**, or **indicator variables** that will take on the role of the x variable in the model. Creating dummy variables is a process of mapping the one column of categorical data into several columns of 0 and 1 data. For example, the two possible levels, Brand C and Brand D, can be recoded to create two distinct variables: one for each brand. For example, the dummy variable for Brand C will have the value 1 for every Brand C paper towel and the value 0 for each Brand D paper towel. Most statistical software packages have a command for creating the dummy variables automatically.

[KEY CONCEPT] Dummy variables can be created to incorporate categorical explanatory variables into a regression model.

[On Your Own:]

7. Create dummy variables for Brand C and Brand D [In Minitab, Calc → Make Indicator Variables → Indicator Variable for Brand → Store Results into c5-c6 (select any 2 unused columns here) → OK. Now label the columns, in order brandC and brandD.] Look at the new data columns to understand how the dummy variables are defined.

These dummy variables can now be incorporated into a regression model. We assign x to be the dummy variable that indicates Brand C, the “slope” coefficient (β_1) for this dummy variable is an estimate of the average amount by which the response variable will change when $x = 1$ (Brand C) compared when $x = 0$ (Brand D).

Notice that there is complete redundancy between these dummy variables; knowing the dummy value of the **brandC** variable automatically tells us the value of **brandD** variable for the same observation (paper towel), so we actually only need one of the dummy variables in this model.

While the notation has changed, the regression model and the model used in the two-sample t-test are mathematically equivalent. When an observation is from Brand C, the deterministic model is μ_1 in the t-test and the deterministic model sets $x = 1$ in regression.

$$\mu_1 = \beta_0 + \beta_1(1) = \beta_0 + \beta_1 \quad (1.2.2)$$

When an observation is from Brand D, the deterministic model is μ_2 in the t-test and the deterministic model sets $x = 0$ in regression.

$$\mu_2 = \beta_0 + \beta_1(0) = \beta_0 \quad (1.2.3)$$

Equations (1.2.2) and (1.2.3) can be combined to show the relationship between the hypothesis being tested.

$$\mu_1 - \mu_2 = (\beta_0 + \beta_1) - \beta_0 = \beta_1 .$$

Thus, testing the null hypothesis $H_0 : \beta_1 = 0$ is equivalent to testing the two-sample t-test hypothesis, $H_0 : \mu_1 - \mu_2 = 0$.

1.2.2 Checking the Validity of the Regression Model Assumptions

As in the two-sample t-procedure it is necessary to check the following model assumptions:

- the error terms are independent and identically distributed

- the population variances are equal. In other words the errors in the regression model (also called residuals) are assumed to come from a single population with variance σ^2 , and
- the error terms follow a normal probability distribution: denoted as $\epsilon_i \sim N(0, \sigma^2)$.

In regression, these assumptions are generally checked looking at the **residuals** from the data: $y_i - \hat{y}_i$. Here, y_i are the observed responses and $\hat{y}_i = b_0 + b_1 x_i$ are the estimated responses given the data and the model. So, the “errors” made by using simply the deterministic portion of the model to estimate the responses are the residuals, $e_i = y_i - \hat{y}_i$.

[On Your Own:]

8. Create a plot of the residuals versus x and comment on the appropriateness of the equal variances assumption. [In Minitab, **Stat** → **Regression** → **Regression Response is weight Predictors is BrandC**, Click **Graphs** and in **Residual versus the variables:** select **BrandC**]. Explain how you can visually compare variances of the brands using a residual plot.
9. Create a normal probability plot or a histogram of the residuals and comment on the normality assumption for the random error term. [In Minitab, **Stat** → **Regression** → **Regression Response is Weight Predictors is BrandC**, Click **Graphs** and in **Residual Plots** select **Normal plot of residuals**].
10. Create a normal probability plot or a histogram of the observations (y_i $i = 1, 2, \dots, n$). [In Minitab, **Graph** → **Probability Plot** → **Single** in **Graph variables:** **Weight**]. Explain why residuals should be used to test the normality assumption instead of the observations.

1.2.3 A t-Test for a Difference in Means Using Linear Regression

Using the regression model specifications in Section 1.2.1, it can be shown that the t-statistic for the slope coefficient is

$$t = \frac{b_1 - \beta_1}{\hat{\sigma} \sqrt{\frac{1}{n \sum_{i=1}^n (x_i - \bar{x})^2}}} \quad \text{where} \quad \hat{\sigma} = \frac{\sum_{i=1}^n (y_i - (b_0 + b_1 x_i))^2}{n - 2} \quad (1.2.4)$$

Probability theory can be used to prove that if the regression model assumptions are true, the t-statistic in equation (1.2.4) follows a t-distribution with $(n - 2) = (n_1 + n_2 - 2)$ degrees of freedom.

[On Your Own:]

11. Fit the data to a regression model using `weight` as the response and the dummy variable `brandC` as the explanatory variable. Calculate the least squares regression line, the t-statistic and p-value for $H_0 : \beta_1 = 0$. Based on these statistics, can you conclude that the `brandA` coefficient, β_1 , is significantly different from 0? [In Minitab, click `Stat` → `Regression` → `regression`].

Do the p-value and test statistic for the `brandA` coefficient look familiar? Explain.

1.3 ANOVA to Compare Population Means

The paper towel breaking strength study uses data from one quantitative response variable and one categorical explanatory variable (**factor**) with two conditions (**levels**). These levels were previously described as the two populations of interest. ANOVA models will describe variables in terms of factors and levels. The factor of interest in this study is the brand of paper towel. The levels of the factor are Brand C and Brand D.

1.3.1 The ANOVA Model

The ANOVA model used to develop an appropriate hypothesis test is:

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij} \quad \text{where } \epsilon_{ij} \stackrel{\text{iid}}{\sim} N(0, \sigma^2) \quad \text{for } i = 1, 2 \quad \text{and } j = 1, 2, \dots, n_i. \quad (1.3.1)$$

In the paper towel example the deterministic model is $\mu + \alpha_1$ for Brand C and $\mu + \alpha_2$ for Brand D. μ is the population mean of the breaking strength all paper towels involved in our study. α_1 is the **effect** of using only Brand C instead of the entire population. The **(main) effect of Brand C**, α_1 describes how much you would expect to adjust the overall mean to get to the Brand C mean.

Again, the notation varies, but the deterministic models for the t-test and ANOVA are equivalent.

$$\mu_1 = \mu + \alpha_1 \quad \text{level mean for Brand C paper towels and}$$

$$\mu_2 = \mu + \alpha_2 \quad \text{level mean for Brand D paper towels}$$

To summarize, the symbols in the model represent

y_{ij}	observed breaking strength for towel j from brand i
μ	overall population mean (the benchmark value)
α_i	brand effect ($i = 1, 2$) tells which brand was used
ϵ_{ij}	error for towel j ($j = 1, 2, \dots, 26$) from brand i

[On Your Own:]

12. Explain why the ANOVA hypothesis, $H_0 : \alpha_1 = \alpha_2 = 0$ is equivalent to the two-sample t-test hypothesis $H_0 : \mu_1 = \mu_2$.
13. Write the proper ANOVA model for the observation representing the third paper towel selected from Brand C. Also give the notation for the observation representing the twentieth paper towel selected from Brand D. In other words, use the appropriate numbers for the subscripts defined in equation (1.3.1).
14. Why doesn't μ have any subscript in the ANOVA model?

After the data are collected, the averages for all three meaningful groupings of the data can be calculated. The following mathematical notation is often used to represent the calculated sample averages:

- $\bar{y}_{..}$ the **grand mean** (the overall average of the combined results),
- $\bar{y}_{1.}$ the level average for the Brand C sample results, and
- $\bar{y}_{2.}$ the level average for the Brand D sample results.

The effect of Brand C, α_1 , can be estimated by $\bar{y}_{1.} - \bar{y}_{..}$. Similarly, $\bar{y}_{2.} - \bar{y}_{..}$ estimates the Brand D effect, α_2 .

[On Your Own:]

15. Calculate $\bar{y}_{..}$, $\bar{y}_{1.}$, and $\bar{y}_{2.}$ for the paper towel data. [In Minitab, Click **Stat** → **Basic Statistics** → **Display Descriptive Statistics**. Double click **weight** and by **brand**. Repeat the process without selecting by **brand**].
16. Use the paper towel sample data to estimate the effect sizes for Brand C and Brand D.
17. The main effects are often visualized with a **main effects plot**. [In Minitab, Click **Stat** → **ANOVA** → **Main Effects Plot**. Select **weight** as the Responses and **brand** as the Factors]. This graph simply plots the averages for each factor level and in this example shows that Brand C paper towels exhibited a higher average breaking strength than the Brand D towels. However, are the two sample averages far enough apart to conclude the population means are different? ANOVA will calculate a p-value to determine if the difference in level averages for Brand C are large enough to be confident that they did not occur by chance.

As in regression, the **residuals** ϵ_{ij} are the difference between the actual observation and the estimated value found from the deterministic model. Calculate the ij^{th} residual value by

subtracting all influencing factors from each observation.

$$\begin{aligned}\epsilon_{ij} &= \text{observed} - \text{effect of Brand}_i - \text{grand mean} \\ &= y_{ij} - (\bar{y}_i - \bar{y}_{..}) - \bar{y}_{..} \\ &= y_{ij} - \bar{y}_i.\end{aligned}$$

The residual values sum to zero within each level combination. This is not surprising since residuals measure the deviations from the observed values to the deterministic model value.

1.3.2 The Model Assumptions

The model assumptions are essentially the same as the two previous tests:

- the error terms are independent and identically distributed,
- the population variances within each factor level are equal (i.e. the sample variances can be pooled), and
- the error terms follow a normal probability distribution: denoted as $\epsilon_{ij} \sim N(0, \sigma^2)$.

1.3.3 An F test in ANOVA to Compare Population Means

Several calculations will be made to test the hypothesis $H_0 : \alpha_1 = \alpha_2 = 0$ in ANOVA: **Sum of Squares** (SS) is a measure of spread for

1. the factor of interest,
2. the residuals, and
3. the total sample.

SS can be calculated by squaring each effect and summing over every cell.

$$\begin{aligned}
 SS_{Brand} &= \sum_{i=1}^{(\text{number of Brand levels})} (\text{sample size}_i) \times (\text{effect}_i)^2 \\
 &= \sum_{i=1}^1 n_i \times (\text{effect}_i)^2 \\
 &= \sum_{i=1}^2 26 \times (\bar{y}_{i.} - \bar{y}_{..})^2 \\
 &= 26 \times (\bar{y}_{1.} - \bar{y}_{..}) + 26 \times (\bar{y}_{2.} - \bar{y}_{..}) \\
 SS_{Error} &= \sum (\text{each residual effect})^2 \\
 &= \sum_{i=1}^2 \left[\sum_{j=1}^{26} (y_{ij} - \bar{y}_{i.})^2 \right] \\
 &= \sum_{i=1}^2 [(n_i - 1) \times s_i^2] = 25 \times s_1^2 + 25 \times s_2^2 \\
 SS_{Total} &= (\text{total number of samples} - 1) \times (\text{overall sample variance}) \\
 &= (n_1 + n_2 - 1) \times s^2 \\
 &= \sum_{i=1}^2 \left[\sum_{j=1}^{26} (y_{ij} - \bar{y}_{..})^2 \right]
 \end{aligned}$$

It can be shown that $SS_{Total} = SS_{Brand} + SS_{Error}$. While the specific formula for SS_{Error} is provided above, it is most easily calculated by subtracting SS_{Brand} from SS_{Total} .

Degrees of Freedom (df) In this example there are two levels of the brand factor. The brand effects sum to 0. Thus, knowing the effect of Brand C automatically forces a known effect for Brand D. Thus SS_{Brand} only has 1 df. SS_{Total} , like the usual, one-sample variance, has $(n - 1)$ df. It can also be shown that $df_{Total} = df_{Brand} + df_{Error}$, thus $df_{Error} = df_{Total} - df_{Brand} = (n_1 + n_2 - 1) - 1 = n_1 + n_2 - 2$. In this example, this is $26 + 26 - 2$.

Mean Squares Brand (MS_{Brand}) $= SS_{Brand}/df_{Brand}$. MS_{Brand} is a measure of variability between the levels of each factor and is often called **between-level variability**.

Mean Square Error (MSE) $= SS_{Error}/df_{Error}$. MSE is a pooled measure of the variability within each level, or the **within-level variability**. Remember that the variance is assumed to be the same for the responses within each level of the factor: $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

F-statistic $= MS_{Brand}/MSE$. The F-statistic is a ratio of the between-level variability over the within-level variability. If indeed $\mu_1 = \mu_2$, (i.e., $H_0 : \alpha_1 = \alpha_2 = 0$ is true), then we would expect the variation between the level means in our sample data to be about the same as the typical variation within levels and the F-statistic would be close to 1.

If MS_{Brand} is much larger than MSE, it is reasonable to conclude that there truly is a difference between level means and the difference we observed in our sample runs was not simply due to chance (i.e., sample-to-sample variability).

These calculations are often summarized in an **ANOVA table** as shown in Table 1.1.

Source	df	SS	MS	F-Statistic
Brand	1	$\sum_{i=1}^2 n_i (\bar{y}_{i.} - \bar{y}_{..})^2$	$\frac{SS_{Brand}}{df_{Brand}}$	$\frac{MS_{Brand}}{MSE}$
Error	$n_1 + n_2 - 2$	$\sum_{i=1}^2 (n_i - 1) s_i^2$	$\frac{SS_{Error}}{df_{Error}}$	
Total	$n_1 + n_2 - 1$	$\sum_{i=1}^2 \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2$		

Table 1.1: One-Factor ANOVA Table (One Factor With 2 Levels).

Mathematical theory can be used to prove that if the model assumptions are true, the F-statistic follows an F distribution with df_{Brand} and df_{Error} degrees of freedom. The **p-value** gives the likelihood of observing an F-statistic at least this extreme (at least this large), assuming that the population means of Brand C and Brand D are equal.. Thus, when the p-value is small (e.g., less than 0.05 or 0.1), the effect size of the brand factor is statistically significant.

Figure 1.1 shows dotplots for two fictitious data sets, Results and Results(2). These plots illustrate the comparison of the between-level variation and the within-level variation.

Both sets of data have the same means, thus the between-level variation (MS_{Brand}) is the same in both Results and Results(2). However the within-level variation (MSE) is much larger for Results than for Results(2). Thus, Results(2) provides much stronger evidence that the factor effects are not simply due to chance. The Results(2) data set will correspond to a larger F-statistic, smaller p-value, and we are more likely to reject the null hypothesis and conclude that the factor effects are significant.

[On Your Own:]

- OPTIONAL Use \bar{y}_1 , \bar{y}_2 , s_1 , and s_2 calculated in Question 2 to calculate mean square error (MSE) and (MS_{Brand}). Calculate the overall variance of the entire data set and use it to find the total sum of squares (SST). Calculate the F-statistic and find the p-value. Use Minitab to check you calculations. [In Minitab, click **Stat** → **ANOVA** → **1-way ANOVA**]. Use the p-values, SS and F statistics to summarize the results.
- Check the model assumptions using Minitab. Are the variances equal? Are the residuals *i.i.d.* and follow a normal distribution?

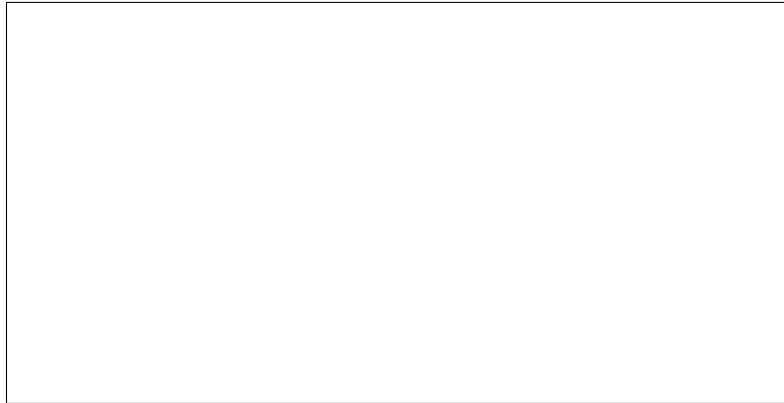


Figure 1.1: (a) Data demonstrating higher variability of the response within factor levels than between level means. (b) Data demonstrating lower variability of the response within factor levels than between level means.

20. Describe how MSE is related to the pooled variance in the two-sample t-test given in equation (1.1.1).
21. How does the p-value in ANOVA compare to the p-value you found for the two-sample t-test and the regression model?
22. Take the square root of the F-statistic in the ANOVA table. Does this value look familiar? Explain.
23. Compare the three statistical models in this lab. Describe the connections between the t-test, ANOVA and regression. Why are the p-values the same for all three models.