Data Analysis
Spring Semester, 2023
February 16, 2023
Lecture 8

Feb 23: Examination One: Chapters 3, 4, 5, 6, and 7

Chapter Six
Inferences Comparing Two Population Central Values

Two Independent Sample Test

Let $X_1, X_2, \ldots, X_n$ be a random sample of size *n* from the random variable *X*, which is $N(\mu_X, \sigma_X^2)$. Let $B_1, B_2, \ldots, B_m$ be a random sample of size *m* from the random variable *B*, which is $N(\mu_B, \sigma_B^2)$. The two samples are independent of each other. The random variable $\overline{X}_n$ is $N(\mu_X, \frac{\sigma_X^2}{n})$, and $\overline{B}_m$ is $N(\mu_B, \frac{\sigma_B^2}{m})$ The two sample averages are independent. We seek to use this data to test $H_0: E(X-B) = 0$ against the alternative hypothesis $H_1: E(X-B) \neq 0$ at the α level of significance. Our test statistic is $TS = \overline{X}_n - \overline{B}_m$.

Distribution of the Test Statistic

The distribution of $TS = \overline{X}_n - \overline{B}_m$ is normal; and
$E(TS) = E(\overline{X}_n - \overline{B}_m) = E(\overline{X}_n) - E(\overline{B}_m) = \mu_X - \mu_B$, and
$\text{var}(TS) = \text{var}(\overline{X}_n - \overline{B}_m) = \text{var}(\overline{X}_n) + \text{var}(\overline{B}_m) = \frac{\sigma_X^2}{n} + \frac{\sigma_B^2}{m}$.

Null Distribution of the Test Statistic

Since $H_0: E(X-B) = 0$ which is equivalent to $H_0: \mu_X = \mu_B$,
$E_0(TS) = E_0(\overline{X}_n - \overline{B}_m) = \mu_X - \mu_B = 0$.

Under the null hypothesis that the distribution of *X* and *B* are identical, not only does $\mu_X = \mu_B$ under the null, but also $\sigma_X^2 = \sigma_B^2 = \sigma^2$. Using this assumption,
$\text{var}_0(TS) = \text{var}_0(\overline{X}_n - \overline{B}_m) = \frac{\sigma^2}{n} + \frac{\sigma^2}{m} = \sigma^2(\frac{1}{n} + \frac{1}{m})$.

Test when variances unknown but equal

Just as in Chapter 5, we use an estimate of σ^2 and stretch the critical values an amount determined by the degrees of freedom of our estimate. There are many estimates of σ^2. For example, $S_X^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}{n-1}$ and $S_B^2 = \frac{\sum_{i=1}^{n}(B_i - \overline{B}_m)^2}{m-1}$ are unbiased estimates of σ^2, with $n-1$ and $m-1$ degrees of freedom respectively. That is, $E(S_X^2) = E(S_B^2) = \sigma^2$. We use both estimates. Let $S_P^2 = \frac{(n-1)S_X^2 + (m-1)S_B^2}{n+m-2}$.

This estimator has $n+m-2$ degrees of freedom. Then our studentized statistic is

$$T_{n+m-2} = \frac{\overline{X}_n - \overline{B}_m - 0}{\sqrt{S_P^2(\frac{1}{n} + \frac{1}{m})}}.$$

If $\alpha = 0.10$, we reject H_0 when $|T_{n+m-2}| \geq t_{1.645,n+m-2}$. Similarly, if $\alpha = 0.05$, we reject H_0 when $|T_{n+m-2}| \geq t_{1.960,n+m-2}$. If $\alpha = 0.01$, we reject H_0 when $|T_{n+m-2}| \geq t_{2.576,n+m-2}$.

Type II Error Rate and Power Calculations

The definition of the Type II error rate is $\beta = \Pr_1\{\text{Accept } H_0\}$. The power of a statistical test is defined to be $\text{Power} = 1 - \beta$. This is the probability that the null hypothesis is correctly rejected. A large Type II error rate indicates a study that is "underpowered." The calculation of β is just a normal probability calculation. The specification of the normal distribution is based on the alternative specified in the problem. Typically, the values of the variances are assumed known.

Example Problem: from Chapter 6 Study Guide, Problem 3

In a clinical trial, 50 patients suffering from an illness will be randomly assigned to one of two groups so that 25 receive an experimental treatment and 25 receive the best available treatment. The random variable *X* is the response of a patient to the experimental medicine, and the random variable *B* is the response of a patient to the best currently available treatment. The random variables *X* and *B* are normally distributed with $\sigma_X = \sigma_B = 500$ under both the null and alternative distributions. The null hypothesis to be tested is that $E(X) - E(B) = 0$ against the alternative that $E(X) - E(B) > 0$ at the 0.01 level of significance. What is the probability of a Type II error for the test of the null hypothesis when $E(X) - E(B) = 500$?

Solution: The standard score form of the TS is $Z = \dfrac{\overline{X}_n - \overline{B}_m - 0}{\sqrt{\sigma^2(\frac{1}{n} + \frac{1}{m})}}$, and

$H_0 : E(X - B) = 0$ is rejected when $Z \geq 2.326$, remembering that the problem asks for a one-sided test at level of significance 0.01. Using the TS directly,
$H_0 : E(X - B) = 0$ is rejected when
$\overline{X}_{25} - \overline{B}_{25} \geq 0 + 2.326\sqrt{\dfrac{500^2}{25} + \dfrac{500^2}{25}} = 0 + 2.326 \bullet 141.42 = 328.95$. For the alternative specified in the problem $\overline{X}_{25} - \overline{B}_{25}$ is $N(500, 141.42^2)$. Then

$$\beta = \Pr_1\{\text{Accept } H_0\} = \Pr_1\{\overline{X}_{25} - \overline{B}_{25} < 328.95\} = \Pr\{\frac{\overline{X}_{25} - \overline{B}_{25} - E_1(\overline{X}_{25} - \overline{B}_{25})}{\sigma_1(\overline{X}_{25} - \overline{B}_{25})} < \frac{328.95 - 500}{141.42}\}.$$

That is, $\beta = \Pr\{Z < \dfrac{328.95 - 500}{141.42} = -1.210\} = \Phi(-1.210) = 0.113$.

Sample size for two sample test:

The bad news is that the mathematics of sample size calculations is relatively complex. The good news is that one can solve a wide range of sample size problems once one knows how to solve this one. The argument is essentially the same.

Problem 6 in Chapter 6 Study Guide

> In a clinical trial, *2J* patients suffering from an illness will be randomly assigned to one of two groups so that *J* will receive an experimental treatment and *J* will receive the best available treatment. The random variable *X* is the response of a patient to the experimental medicine, and the random variable *B* is the response of a patient to the best currently available treatment. The random variables *X* and *B* are normally distributed. The null hypothesis to be tested is that $E(X) - E(B) = 0$ against the alternative that $E(X) - E(B) > 0$ at the α, $\alpha \leq 0.5$, level of significance. When the null hypothesis is true, $\text{var}(X) = \text{var}(B) = \sigma_0^2$. When the alternative hypothesis is true, $\text{var}(B) = \sigma_0^2$, but $\text{var}(X) = \sigma_1^2 > \sigma_0^2$. What is the number *J* in each group that would have to be taken so that the probability of a Type II error for the test of the null hypothesis specified in the common section is β, $\beta \leq 0.5$, when $E(X) - E(B) = \Delta > 0$?

Solution: The test statistic is $TS = \overline{X}_J - \overline{B}_J$, and *TS* is

$N(E(X)-E(B), \frac{\text{var}(X)}{J}+\frac{\text{var}(B)}{J})$. The null distribution of *TS* is then $N(0, \frac{\sigma_0^2}{J}+\frac{\sigma_0^2}{J})$. Hence, we reject $H_0: E(X)-E(B)=0$ against $H_1: E(X)-E(B)>0$ at the α level of significance when $TS \geq 0+|z_\alpha|\sqrt{\frac{\sigma_0^2}{J}+\frac{\sigma_0^2}{J}}$. When $E(X)-E(B)=\Delta>0$ and $\text{var}(X)=\sigma_1^2>\sigma_0^2$. $\text{var}(B)=\sigma_0^2$, the (alternative) distribution of *TS* is then $N(\Delta, \frac{\sigma_0^2}{J}+\frac{\sigma_1^2}{J})$. Then, the probability of a Type II error is $\beta = \Pr_1\{\text{Accept } H_0\} = \Pr_1\{TS < 0+|z_\alpha|\sqrt{\frac{2\sigma_0^2}{J}}\}$. That is,

$\beta = \Pr_1\{TS < 0+|z_\alpha|\sqrt{\frac{2\sigma_0^2}{J}}\} = \Pr\{Z = \frac{TS-\Delta}{\sigma_1(\overline{X}_J-\overline{B}_J)} < \frac{0+|z_\alpha|\sqrt{\frac{2\sigma_0^2}{J}}-\Delta}{\sqrt{\frac{\sigma_0^2}{J}+\frac{\sigma_1^2}{J}}}\}$. Since $\beta = \Pr_1\{\text{Accept } H_0\} \leq 0.5$, it is true that $\beta = \Pr\{Z < -|z_\beta|\}$. We now have two equations:

$$\beta = \Pr\{Z < \frac{0+|z_\alpha|\sqrt{\frac{2\sigma_0^2}{J}}-\Delta}{\sqrt{\frac{\sigma_0^2}{J}+\frac{\sigma_1^2}{J}}}\}, \text{ and}$$

$$\beta = \Pr\{Z < -|z_\beta|\}.$$

The problem is to choose *J* so that the probability of a Type II error is a specified value. That is, we should choose *J* so that the right-hand sides of the two equations are equal: $\frac{0+|z_\alpha|\sqrt{\frac{2\sigma_0^2}{J}}-\Delta}{\sqrt{\frac{\sigma_0^2}{J}+\frac{\sigma_1^2}{J}}} = -|z_\beta|$.

We have to solve for *J* in the equation above. This reduces to: $0+|z_\alpha|\sqrt{\frac{2\sigma_0^2}{J}}-\Delta = -|z_\beta|\sqrt{\frac{\sigma_0^2}{J}+\frac{\sigma_1^2}{J}}$. That is, $|z_\alpha|\sqrt{\frac{2\sigma_0^2}{J}}+|z_\beta|\sqrt{\frac{\sigma_0^2}{J}+\frac{\sigma_1^2}{J}} = \Delta$. Next, solve for *J* to get $\sqrt{J} = \frac{|z_\alpha|\sqrt{2\sigma_0^2}+|z_\beta|\sqrt{\sigma_0^2+\sigma_1^2}}{\Delta}$. Since *J* has to be an integer, we increase *J* to the next integer value.

This calculation assumes that there is no attrition of subjects. Typically, study attrition is large. An attrition rate less than 15% at a follow-up three or more years later is a very low attrition rate. Accounting for attrition is its own modeling effort, which can be very difficult to do well.

Problem 4 in Chapter 6 Study Guide

In a clinical trial, *2J* patients suffering from an illness will be randomly assigned to one of two groups so that *J* will receive an experimental treatment and *J* will receive the best available treatment. The random variable *X* is the response of a patient to the experimental medicine, and the random variable *B* is the response of a patient to the best currently available treatment. The random variables *X* and *B* are normally distributed and have $\sigma_X = \sigma_B = 500$ under both the null and alternative distributions. The null hypothesis to be tested is that $E(X) - E(B) = 0$ against the alternative that $E(X) - E(B) > 0$ at the 0.005 level of significance. What is the number *J* in each group that would have to be taken so that the probability of a Type II error for the test of the null hypothesis specified in the common section is 0.01 when $E(X) - E(B) = 250$?

Solution: For this specification, $\alpha = .005$, so that $|z_\alpha| = 2.576$. Also, $\beta = .01$, so that $|z_\beta| = 2.326$. With regard to variances, $\sigma_0^2 = \sigma_1^2 = 500^2$. Finally, $\Delta = 250$. Then, the design equation is

$$\sqrt{J} = \frac{2.576\sqrt{2 \bullet 500^2} + 2.326\sqrt{2 \bullet 500^2}}{250} = \frac{3466.24}{250} = 13.865 = \sqrt{192.24}\ .$$

That is, there should be at least 193 in each group.

The magnitude of the difference in the expected values is 250, which is half of the assumed standard deviation of each group. This is called a one-half standard deviation effect. To detect a difference equal to 1/2 of the standard deviation, researchers need about 200 in each group. This is 400 total observations. One needs about 50 observations per group to detect a one standard deviation effect. That is, fewer observations are needed to detect a larger effect size.

Chapter 7
Inferences about Population Variances

Probability Theory Facts

Let Z be $N(0,1)$. Then Z^2 has the (central) chi-squared distribution with 1 degree of freedom. This is denoted χ_1^2.

Let $Z_1, Z_2, \ldots, Z_n$ be $NID(0,1)$. Then $S_n = Z_1^2 + Z_2^2 + \cdots + Z_n^2 = \sum_{i=1}^{n} Z_i^2$ follows the (central) chi-square distribution with n degrees of freedom, denoted χ_n^2. The expected value of a χ_n^2 is n: $E(S_n) = E(Z_1^2) + E(Z_2^2) + \cdots + E(Z_n^2)$. Since $\mathrm{var}(Z) = 1 = E(Z^2) - [E(Z)]^2 = 1$, then $E(Z^2) - [0]^2 = 1$. Using this in $E(S_n) = E(Z_1^2) + E(Z_2^2) + \cdots + E(Z_n^2) = n$. Further, the variance of a chi-square distribution with n degrees of freedom is $2n$: $\mathrm{var}(S_n) = 2n$

Let Y be $N(\mu_Y, \sigma_Y^2)$. Then, $\frac{Y - \mu_Y}{\sigma_Y} = Z$ is $N(0,1)$. Let $Y_1, Y_2, \ldots, Y_n$ be a random sample from Y, which is $N(\mu_Y, \sigma_Y^2)$. Then $\sum_{i=1}^{n} (\frac{Y_i - \mu_Y}{\sigma_Y})^2$ is χ_n^2. After factoring out σ_Y^2,

$\frac{\sum_{i=1}^{n} (Y_i - \mu_Y)^2}{\sigma_Y^2}$ is also χ_n^2.

Since μ_Y is not known in applications, it must be estimated. An important property of a sample from a normal distribution is that $\frac{\sum_{i=1}^{n} (Y_i - \bar{Y}_n)^2}{\sigma_Y^2}$ is distributed as χ_{n-1}^2. That is, using the sample mean has reduced the degrees of freedom by one. From AMS 310, the unbiased estimator of the sample variance is $S^2 = \frac{\sum (Y_i - \bar{Y}_n)^2}{n-1}$. Since $(n-1)S^2 = (n-1)\frac{\sum (Y_i - \bar{Y}_n)^2}{n-1} = \sum_{i=1}^{n} (Y_i - \bar{Y}_n)^2$, $\frac{(n-1)S^2}{\sigma_Y^2}$ has a central chi-squared distribution with $n-1$ degrees of freedom when $Y_1, Y_2, \ldots, Y_n$ is a sample of size n from a $N(\mu_Y, \sigma_Y^2)$ distribution. This is our first important use of the chi-squared distribution. The tests in this chapter are usually one-sided.

Problem 1 from Chapter 7 Study Guide

A research team took a sample of 8 observations from the random variable *Y*, which had a normal distribution $N(\mu,\sigma^2)$. They observed $\bar{y}_8 = 43.2$, where $\bar{y}_8$ is the average of the eight sampled observations and $s^2 = 517.5$ is the observed value of the unbiased estimate of σ^2, based on the sample values. Test the null hypothesis that $H_0 : \sigma^2 = 400$ against the alternative $H_1 : \sigma^2 > 400$ at the 0.10, 0.05, and 0.01 levels of significance.

这是单边检测

Solution: The test statistic is $TS = \dfrac{(n-1)S^2}{\sigma_Y^2}$, which has a central χ^2_{n-1}, where $n-1=8-1=7$. Since $H_0 : \sigma^2 = 400$, the null distribution of $TS = \dfrac{(n-1)S^2}{\sigma_Y^2} = \dfrac{(n-1)S^2}{400}$ is χ^2_7. The problem specifies a right-sided test with levels of significance 0.10, 0.05, and 0.01. From Table 7, the right-sided critical values are 12.02 (for the 0.10 level), 14.07 (for 0.05), and 18.48 (for 0.01). The value of the chi-squared test statistic is $ts == \dfrac{(8-1)\bullet 517.5}{400} = 9.056$. Since this is less than each of the critical values, we accept the null hypothesis at the 0.10, 0.05, and 0.01 levels. The value of the sample mean is not relevant. A common mistake is for a student to take the sample mean as a cue and answer with a one-sample t-test. This is not correct.

As usual, a confidence interval may be more informative than a statistical test. The next problem is an example of finding a confidence interval for a variance.

Problem 2 from Chapter 7 Study Guide

> A research team took a sample of 7 observations from the random variable *Y*, which had a normal distribution $N(\mu,\sigma^2)$. They observed $\bar{y}_7 = 93.4$, where $\bar{y}_7$ was the average of the sampled observations, and $s^2 = 47.5$ was the observed value of the unbiased estimate of σ^2, based on the sample values. Find the 99% confidence interval for σ^2.

Solution: Since the estimated variance has 6 degrees of freedom,

$\Pr\{0.6757 < \frac{\sum_{i=1}^{7}(Y_i - \overline{Y}_7)^2}{\sigma^2} < 18.55\} = \Pr\{0.6757 < \frac{6S^2}{\sigma^2} < 18.55\} = 0.99$. Then

$\Pr\{0.6757 < \frac{6S^2}{\sigma^2} < 18.55\} = \Pr\{\frac{1}{18.55} < \frac{\sigma^2}{6S^2} < \frac{1}{0.6757}\} = \Pr\{\frac{6S^2}{18.55} < \sigma^2 < \frac{6S^2}{0.6757}\}$.

The interval from $\frac{6S^2}{18.55}$ to $\frac{6S^2}{0.6757}$ is then the basis of a confidence interval for σ^2. In this problem, the left end of the confidence interval is

$\frac{6s^2}{18.55} = 0.3235s^2 = 0.3235 \bullet 47.5 = 15.36$, and the right end is

$\frac{6s^2}{0.6757} = 8.880s^2 = 8.880 \bullet 47.5 = 421.785$. The confidence interval extends from a factor of about 3 less than $s^2 = 47.5$ to a factor of about 9 greater than $s^2 = 47.5$. Again, the sample mean is not needed to answer the question.

When you work these problems, examine your answer and notice that the confidence interval for σ^2 is very wide. Specifically examine the ratio of the upper limit to the lower limit, here almost 28. It is remarkable that the t distribution stretches are as small as they are. One gets percentiles of the chi-squared distribution from Table 7. The Excel spreadsheet and all statistical packages have the percentiles available as well.

The F-distribution

Let X be $N(\mu_X, \sigma_X^2)$. Using the standard score transformation, $\frac{X - \mu_X}{\sigma_X} = Z$ is $N(0,1)$.

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from X, which is $N(\mu_X, \sigma_X^2)$. Then,

$\frac{\sum_{i=1}^{n}(X_i - \mu_X)^2}{\sigma_X^2}$ is χ_n^2.

Let Y be $N(\mu_Y, \sigma_Y^2)$. Then, $\frac{Y - \mu_Y}{\sigma_Y} = Z$ is also $N(0,1)$. Let $Y_1, Y_2, \ldots, Y_m$ be a random sample of size m from Y, which is $N(\mu_Y, \sigma_Y^2)$. Then, $\frac{\sum_{i=1}^{m}(Y_i - \mu_Y)^2}{\sigma_Y^2}$ is χ_m^2. The definition of the central F distribution is that the random variable

$$F_{n,m} = \frac{\{[\sum_{i=1}^{n}(X_i - \mu_X)^2]/[n\sigma_X^2]\}}{\{[\sum_{i=1}^{m}(Y_i - \mu_Y)^2]/[m\sigma_Y^2]\}}$$ has a (central) F distribution with *n* numerator and *m* denominator degrees of freedom.

Application of the F distribution

The problem with this random variable is that the expected values are not known. As before, we use the sample averages as estimates of the expected values. The penalty for using sample data rather than expected values is a one degree reduction in both the numerator and denominator degrees of freedom. That is,

$$\frac{\{[\sum_{i=1}^{n}(X_i - \bar{X}_n)^2]/[(n-1)\sigma_X^2]\}}{\{[\sum_{i=1}^{m}(Y_i - \bar{Y}_m)^2]/[(m-1)\sigma_Y^2]\}} = \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} = F_{n-1,m-1}$$ has a central F distribution with $n-1$ numerator and $m-1$ denominator degrees of freedom. Of course, there is still the issue of the unknown variances of *X* and *Y* that has to be dealt with.

One use of this random variable is to test the null hypothesis $H_0: \sigma_X^2 = \sigma_Y^2$. The most common alternative hypothesis is $H_1: \sigma_X^2 > \sigma_Y^2$. The test statistic for this hypothesis is $TS = \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2}$. Under the null hypothesis $H_0: \sigma_X^2 = \sigma_Y^2$, $TS = \frac{S_X^2}{S_Y^2}$, and its null distribution is central $F_{n-1,m-1}$. Under the null hypothesis $E(S_X^2) = \sigma_X^2$ and $E(S_Y^2) = \sigma_Y^2$, so that $E_0(TS) \cong \frac{E(S_X^2)}{E(S_Y^2)} = 1$. Under the alternative hypothesis, $E_1(TS) \cong \frac{E(S_X^2)}{E(S_Y^2)} > 1$. That is, the test of $H_0: \sigma_X^2 = \sigma_Y^2$ against the alternative $H_1: \sigma_X^2 > \sigma_Y^2$ is a right-sided test. A value of *TS* near 1 (modulo statistical variation) supports the null hypothesis, and a value of *TS* much greater than 1 supports the alternative. The next problem illustrates the test.

Problem 3 from Chapter 7 Study Guide

A research team took a random sample of 9 observations from a normally distributed random variable *Y* and observed that $\bar{y}_9 = 91.2$ and $s_Y^2 = 229.6$, where $\bar{y}_9$ was the average of the nine observations sampled from *Y* and s_Y^2 was the unbiased estimate of $\operatorname{var}(Y)$. A second research team took a random sample of 10 observations from a normally distributed random variable *X* and observed that

$\bar{x}_{10} = 103.5$ and $s_X^2 = 917.6$, where $\bar{x}_{10}$ was the average of the ten observations sampled from X and s_X^2 was the unbiased estimate of $\operatorname{var}(X)$. Test the null hypothesis $H_0 : \operatorname{var}(X) = \operatorname{var}(Y)$ against the alternative $H_1 : \operatorname{var}(X) > \operatorname{var}(Y)$ at the 0.10, 0.05, and 0.01 levels of significance.

Solution: One has a choice of which sample variance to put in the numerator. When one puts the variance *hypothesized* to be larger in the numerator, then the test is right-sided. Here $ts = \frac{s_X^2}{s_Y^2} = \frac{917.6}{229.6} = 3.9965$, with 9 numerator and 8 denominator degrees of freedom. The critical value for the 0.10 level is 2.56; for the 0.05 level, 3.39; and for the 0.01 level, 5.91. The correct decision is to reject the null hypothesis at the 0.10 and 0.05 levels and accept it at the 0.01 level. As before, the sample means are not needed for the problem. Students who use the information given as the cue to their choice of statistical tests sometimes respond to a question like this with a two-sample t-test. This is incorrect.

Confidence interval for the ratio of variances $\frac{\sigma_X^2}{\sigma_Y^2}$

For this task, we use $TS = \frac{S_Y^2 / \sigma_Y^2}{S_X^2 / \sigma_X^2}$, which has an F-distribution with $m-1$ numerator and $n-1$denominator degrees of freedom. This choice may be counter-intuitive, but is necessary. The percentage points in Table 8 are based on right sided tail areas, so that $\Pr\{F_{1-\alpha/2,m-1,n-1} < \frac{S_Y^2 / \sigma_Y^2}{S_X^2 / \sigma_X^2} < F_{\alpha/2,m-1.n-1}\} = 1-\alpha$, and

$\Pr\{F_{1-\alpha/2,m-1,n-1} < \frac{\sigma_X^2}{\sigma_Y^2} \bullet \frac{S_Y^2}{S_X^2} < F_{\alpha/2,m-1.n-1}\} = 1-\alpha$.

Then, $\Pr\{(F_{1-\alpha/2,m-1,n-1})\frac{S_X^2}{S_Y^2} < \frac{\sigma_X^2}{\sigma_Y^2} < (F_{\alpha/2,m-1.n-1})\frac{S_X^2}{S_Y^2}\} = 1-\alpha$.

The values of $F_{\alpha/2,m-1,n-1}$ are given in Table 8. These tables do not explicitly give $F_{1-\alpha/2,m-1,n-1}$. One needs to use a property of the F distribution to get this value. Since

$\Pr\{F_{1-\alpha/2,m-1,n-1} < \frac{S_Y^2 / \sigma_Y^2}{S_X^2 / \sigma_X^2}\} = 1-\frac{\alpha}{2}$, $\Pr\{\frac{S_Y^2 / \sigma_Y^2}{S_X^2 / \sigma_X^2} < F_{1-\alpha/2,m-1,n-1}\} = \frac{\alpha}{2}$

$\Pr\{[1/[\frac{S_Y^2 / \sigma_Y^2}{S_X^2 / \sigma_X^2}] > [1/F_{1-\alpha/2,m-1,n-1}]\} = \frac{\alpha}{2}$. That is

$\Pr\{\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} > \frac{1}{F_{1-\alpha/2,m-1,n-1}}\} = \frac{\alpha}{2}$. The distribution of $\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2}$ is a central F with $n-1$ numerator degrees of freedom and $m-1$ denominator degrees of freedom. From the definition of the F percentage points in Table 8, $\Pr\{\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} > F_{\alpha/2,n-1,m-1}\} = \frac{\alpha}{2}$.

Equating the two right hand sides of these inequalties shows that $F_{\alpha/2,n-1,m-1} = \frac{1}{F_{1-\alpha/2,m-1,n-1}}$, or equivalently, $F_{1-\alpha/2,m-1,n-1} = \frac{1}{F_{\alpha/2,n-1,m-1}}$. Then $\Pr\{(F_{1-\alpha/2,m-1,n-1})\frac{S_X^2}{S_Y^2} < \frac{\sigma_X^2}{\sigma_Y^2} < (F_{\alpha/2,m-1.n-1})\frac{S_X^2}{S_Y^2}\} = 1-\alpha$ reduces to $\Pr\{(\frac{1}{F_{\alpha/2,n-1,m-1}})\frac{S_X^2}{S_Y^2} < \frac{\sigma_X^2}{\sigma_Y^2} < (F_{\alpha/2,m-1.n-1})\frac{S_X^2}{S_Y^2}\} = 1-\alpha$. The interval that contains $\frac{\sigma_X^2}{\sigma_Y^2}$ with probability $1-\alpha$ is $(\frac{1}{F_{\alpha/2,n-1,m-1}})\frac{S_X^2}{S_Y^2}$ to $(F_{\alpha/2,m-1.n-1})\frac{S_X^2}{S_Y^2}$. We use the observed sample variances in the $1-\alpha$% confidence interval for $\frac{\sigma_X^2}{\sigma_Y^2}$: $(\frac{1}{F_{\alpha/2,n-1,m-1}})\frac{s_X^2}{s_Y^2}$ to $(F_{\alpha/2,m-1.n-1})\frac{s_X^2}{s_Y^2}$.

Problem 4 from Chapter 7 Study Guide

A research team took a random sample of 9 observations from a normally distributed random variable Y and observed that $\bar{y}_9 = 91.2$ and $s_Y^2 = 529.6$, where $\bar{y}_9$ was the average of the nine observations sampled from Y and s_Y^2 was the unbiased estimate of $\text{var}(Y)$. A second research team took a random sample of 10 observations from a normally distributed random variable X and observed that $\bar{x}_{10} = 103.5$ and $s_X^2 = 894.3$, where $\bar{x}_{10}$ was the average of the ten observations sampled from X and s_X^2 was the unbiased estimate of $\text{var}(X)$. Find the 95% confidence interval for $\text{var}(X)/\text{var}(Y)$.

Solution: The sample variance $s_X^2 = 894.3$ is based on 9 degrees of freedom, and the sample variance $s_Y^2 = 529.6$ is based on 8 degrees of freedom. From Table 8, $F_{\alpha/2,m-1.n-1} = F_{0.025,8,9} = 4.10$, and $F_{\alpha/2,n-1.m-1} = F_{0.025,9,8} = 4.36$. The ratio $\frac{s_X^2}{s_Y^2} = \frac{894.3}{529.6} = 1.689$. The left endpoint is given by $\frac{1}{4.36}\frac{s_X^2}{s_Y^2} = 0.229 \bullet 1.689 = 0.387$

The right endpoint is given by $4.10\frac{s_X^2}{s_Y^2} = 4.10 \bullet 1.689 = 6.92$. The 95% confidence interval for $\text{var}(X)/\text{var}(Y)$ is from 0.387 to 6.92. Since the confidence interval for $\text{var}(X)/\text{var}(Y)$ includes 1, we would accept the null hypothesis that the ratio of the variances was 1 at the two-sided 0.05 level of significance. The confidence interval for the ratio of the variances has a right endpoint that is a factor of roughly 18 times the left endpoint. The sample averages do not enter into the solution of this problem. Some students respond incorrectly with a 95% confidence interval for the difference in means. This is not correct.

Chapter 11
Linear Regression and Correlation

The research context is that two variables have been observed for each of *n* participants. The research team then has a spreadsheet with *n* pairs of observations $(x_i, y_i),\ i = 1,\ldots,n$. One of the variables (here *y*) is the outcome variable or dependent variable. This is the variable hypothesized to be affected by the other variable in scientific research. The other variable (here *x*) is the independent variable. It may be hypothesized to predict the outcome variable or to cause a change in the outcome variable. The research task is to document the association between independent and dependent variables. An example of a research project seeking to document a causal association would be a clinical trial in which x_i was the dosage of a medicine randomly assigned to a participant (say simvastatin) and y_i was the participant's response after a specified period taking the medicine (say cholesterol reduction after 3 months). An example of a study seeking to document the value of a predictive association would be an observational study in which x_i was the score of a statistics student on the first examination in a course and y_i was the student's score on the final examination in the course.

A recommended first step is to create the scatterplot of observations, with the vertical axis representing the dependent variable and the horizontal axis representing the independent variable. The "pencil test" is to hold up a pencil to the scatterplot and examine whether that describes the data well. If so, then it is reasonable to assume that a ***linear model*** (such as $\beta_0 + \beta_1 x$) describes the data. The linear model is reasonable for many data sets in observational studies. A more object procedure is to use a "nonlinear smoother" such as LOWESS to estimate the association. If the LOWESS curve is not well approximated by a line, then the assumption of linearity is not reasonable.

Estimating the Linear Model Parameters (section 11.2)

OLS (ordinary least squares) is the most used method to estimate the parameters of the linear model. An arbitrary linear model $b_0 + b_1 x$ is used as a *fit* for the dependent variable values. The method uses the *residual* $y_i - b_0 - b_1 x_i$. The fitting model is judged by how small the set of residuals is. OLS uses each residual and focuses on the magnitude of the residuals by examining the sum of squares function $SS(b_0, b_1) = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$. The OLS method is to find the arguments $(\hat{\beta}_0, \hat{\beta}_1)$ that make $SS(b_0, b_1)$ as small as possible. This minimization is a standard calculus problem. Step 1 is to calculate the partial derivatives of $SS(b_0, b_1)$ with respect to each argument. First, the partial with respect to b_0:

$$\frac{\partial SS(b_0,b_1)}{\partial b_0} = \frac{\partial}{\partial b_0}\sum_{i=1}^{n}(y_i - b_0 - b_1x_i)^2 = \sum_{i=1}^{n}\frac{\partial}{\partial b_0}(y_i - b_0 - b_1x_i)^2 = \sum_{i=1}^{n}2(y_i - b_0 - b_1x_i)\frac{\partial(y_i - b_0 - b_1x_i)}{\partial b_0}$$

$$\frac{\partial SS(b_0,b_1)}{\partial b_0} = \sum_{i=1}^{n}(-2)(y_i - b_0 - b_1x_i).$$

Second, the partial with respect to b_1:

$$\frac{\partial SS(b_0,b_1)}{\partial b_1} = \frac{\partial}{\partial b_1}\sum_{i=1}^{n}(y_i - b_0 - b_1x_i)^2 = \sum_{i=1}^{n}\frac{\partial}{\partial b_1}(y_i - b_0 - b_1x_i)^2 = \sum_{i=1}^{n}2(y_i - b_0 - b_1x_i)\frac{\partial(y_i - b_0 - b_1x_i)}{\partial b_1}$$

$$\frac{\partial SS(b_0,b_1)}{\partial b_1} = \sum_{i=1}^{n}(-2x_i)(y_i - b_0 - b_1x_i).$$

Step 2 is to find the arguments $(\hat{\beta}_0, \hat{\beta}_1)$ that make the two partial derivatives zero. The resulting equations are called the *normal equations*:

$\sum_{i=1}^{n}(-2)(y_i - \hat{\beta}_0 - \hat{\beta}_1x_i) = 0$ and

$\sum_{i=1}^{n}(-2)(y_i - \hat{\beta}_0 - \hat{\beta}_1x_i)x_i = 0$.

These equations have a very important interpretation. Let $r_i = y_i - \hat{\beta}_0 - \hat{\beta}_1x_i,\ i = 1,\ldots,n$. The first normal equation is equivalent to $\sum_{i=1}^{n} r_i = 0$, and the second is $\sum_{i=1}^{n} r_ix_i = 0$. That is, there are two constraints on the n residuals. The OLS residuals must sum to zero, and the OLS residuals are orthogonal to the independent variable values. The n residuals then have $n-2$ degrees of freedom.

Step 3 is to solve this two linear equation system in two unknowns. Start by using the first normal equation to solve for $\hat{\beta}_0$:

$\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1x_i) = \sum_{i=1}^{n}y_i - \sum_{i=1}^{n}\hat{\beta}_0 - \sum_{i=1}^{n}\hat{\beta}_1x_i = n\bar{y}_n - n\hat{\beta}_0 - \hat{\beta}_1(n\bar{x}_n) = 0$. Solving for $\hat{\beta}_0$ yields $\hat{\beta}_0 = \bar{y}_n - \hat{\beta}_1\bar{x}_n$. Next, insert the solution for $\hat{\beta}_0$ in the second normal equation and solve for $\hat{\beta}_1$:

$$0 = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1x_i)x_i = \sum_{i=1}^{n}\{[y_i - (\bar{y}_n - \hat{\beta}_1\bar{x}_n) - \hat{\beta}_1x_i]x_i\} = \sum_{i=1}^{n}[(y_i - \bar{y}_n)x_i] - \sum_{i=1}^{n}[\hat{\beta}_1(x_i - \bar{x}_n)x_i],$$

The solution is $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y}_n)x_i}{\sum_{i=1}^{n}(x_i - \bar{x}_n)x_i}$. There are several modifications of this formula that are helpful. The first results from noting that $\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 = \sum_{i=1}^{n}(x_i - \bar{x}_n)x_i - \sum_{i=1}^{n}(x_i - \bar{x}_n)\bar{x}_n = \sum_{i=1}^{n}(x_i - \bar{x}_n)x_i$ and $\sum_{i=1}^{n}(y_i - \bar{y}_n)(x_i - \bar{x}_n) = \sum_{i=1}^{n}(y_i - \bar{y}_n)x_i - \sum_{i=1}^{n}(y_i - \bar{y}_n)\bar{x}_n = \sum_{i=1}^{n}(y_i - \bar{y}_n)x_i$. The OLS solution is then $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y}_n)(x_i - \bar{x}_n)}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}$. This is a very commonly quoted formula.

The second shows the relation of $\hat{\beta}_1$ and the Pearson product moment correlation. The Pearson product moment correlation is a dimensionless measure of association. The formula is $r(x,y) = \frac{\sum_{i=1}^{n}(y_i - \bar{y}_n)(x_i - \bar{x}_n)}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y}_n)^2}}$. The Cauchy-Schwartz inequality shows that $|r(x,y)| \le 1$. A correlation of +1 or -1 shows a perfect linear association. A correlation of 0 means no linear association. The numerator of $\hat{\beta}_1$ and $r(x,y)$ are the same. Starting with $\hat{\beta}_1$,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y}_n)(x_i - \bar{x}_n)}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2} = \frac{\sum_{i=1}^{n}(y_i - \bar{y}_n)(x_i - \bar{x}_n)}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2} \bullet \sqrt{\sum_{i=1}^{n}(y_i - \bar{y}_n)^2}} \bullet \frac{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y}_n)^2}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}}.$$

That is,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y}_n)(x_i - \bar{x}_n)}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2} \bullet \sqrt{\sum_{i=1}^{n}(y_i - \bar{y}_n)^2}} \bullet \frac{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y}_n)^2}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}} = r(x,y) \bullet \frac{\sqrt{(n-1)s_Y^2}}{\sqrt{(n-1)s_X^2}} = \frac{s_Y}{s_X} \bullet r(x,y).$$

The second formula is then $\hat{\beta}_1 = \frac{s_Y}{s_X} \bullet r(x,y)$.

The next variation will be used in calculating the distributional properties of $\hat{\beta}_1$ and uses the identity that

$\sum_{i=1}^{n}(y_i - \bar{y}_n)(x_i - \bar{x}_n) = \sum_{i=1}^{n}[(x_i - \bar{x}_n)y_i] - \sum_{i=1}^{n}[(x_i - \bar{x}_n)\bar{y}_n] = \sum_{i=1}^{n}(x_i - \bar{x}_n)y_i$. Then $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x}_n)y_i}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}$

Fisher's Decomposition of the Total Sum of Squares

The total sum of squares of the dependent variable is defined to be $TSS = \sum_{i=1}^{n}(y_i - \bar{y}_n)^2$ with $n-1$ degrees of freedom. The *i*th residual was defined above to be $r_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i, i = 1,\ldots,n$. After substituting for $\hat{\beta}_0$, $r_i = y_i - \bar{y}_n - \hat{\beta}_1(x_i - \bar{x}_n), i = 1,\ldots,n$.

Fisher's decomposition is a fundamental tool for the analysis of the linear model. It starts with $TSS = \sum_{i=1}^{n}(y_i - \bar{y}_n)^2 = \sum_{i=1}^{n}[y_i - \bar{y}_n - \hat{\beta}_1(x_i - \bar{x}_n) + \hat{\beta}_1(x_i - \bar{x}_n)]^2 = \sum_{i=1}^{n}[r_i + \hat{\beta}_1(x_i - \bar{x}_n)]^2$.

Next $TSS = \sum_{i=1}^{n}[r_i + \hat{\beta}_1(x_i - \bar{x}_n)]^2 = \sum_{i=1}^{n}[r_i^2 + \hat{\beta}_1^2(x_i - \bar{x}_n)^2 + 2\hat{\beta}_1 r_i(x_i - \bar{x}_n)]$, and

$TSS = \sum_{i=1}^{n} r_i^2 + \sum_{i=1}^{n}\hat{\beta}_1^2(x_i - \bar{x}_n)^2 + 2\hat{\beta}_1\sum_{i=1}^{n} r_i(x_i - \bar{x}_n)$. The first sum $\sum_{i=1}^{n} r_i^2 = SSE$, the sum of squared errors and has $n-2$ degrees of freedom. The second sum $\sum_{i=1}^{n}\hat{\beta}_1^2(x_i - \bar{x}_n)^2$ is called the regression sum of squares and has 1 degree of freedom. It can be simplified:

$$\text{RegSS} = \sum_{i=1}^{n}\hat{\beta}_1^2(x_i - \bar{x}_n)^2 = \hat{\beta}_1^2\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 = [r(x,y)]^2\left[\frac{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y}_n)^2}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}}\right]^2\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 \text{ and}$$

$$\text{RegSS} = [r(x,y)]^2\left[\frac{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y}_n)^2}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}}\right]^2\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 = [r(x,y)]^2\sum_{i=1}^{n}(y_i - \bar{y}_n)^2 = [r(x,y)]^2 TSS.$$

Finally, the third sum $2\hat{\beta}_1\sum_{i=1}^{n} r_i(x_i - \bar{x}_n) = 2\hat{\beta}_1(\sum_{i=1}^{n} r_i x_i - \sum_{i=1}^{n} r_i\bar{x}_n) = 2\hat{\beta}_1(0-0) = 0$.

This is conventionally displayed in an Analysis of Variance Table as below:

Analysis of Variance Table
One Predictor Linear Regression

Source	DF	Sum of Squares	Mean Square	F
Regression	1	$\sum_{i=1}^{n} \hat{\beta}_1^2 (x_i - \bar{x}_n)^2 = [r(x,y)]^2 TSS$	$[r(x,y)]^2 TSS$	$\frac{(n-2)[r(x,y)]^2}{1-[r(x,y)]^2}$
Error	$n-2$	$\sum_{i=1}^{n} r_i^2 = \{1-[r(x,y)]^2\} TSS$	$\frac{\{1-[r(x,y)]^2\} TSS}{(n-2)}$	
Total	$n-1$	$TSS = (n-1) s_{DV}^2$		