Data Analysis
Spring Semester, 2023
February 14, 2023
Lecture 7

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

Chapter Six
Inferences Comparing Two Population Central Values

The random variables W_1 and W_2 are independent. Their expected values are $E(W_1) = \mu_1$ and $E(W_2) = \mu_2$. Their variances are $\text{var}(W_1) = \sigma_1^2 < \infty$, and $\text{var}(W_2) = \sigma_2^2 < \infty$.

First, $E(W_1 - W_2) = E(W_1) - E(W_2) = \mu_1 - \mu_2$.
Second, $\text{var}(W_1 - W_2) = \sigma_1^2 + \sigma_2^2$

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. In the case that the random variable *X* is not normal (or the random variable *B* is not normal), then the CLT implies that the distribution of *TS* is approximately normal. From the problem at the start of class, $E(TS) = E(\overline{X}_n - \overline{B}_m) = E(\overline{X}_n) - E(\overline{B}_m) = \mu_X - \mu_B$, 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 $H_0: \mu_X = \mu_B$,
$E_0(TS) = E_0(\bar{X}_n - \bar{B}_m) = \mu_X - \mu_B = 0$.

Under the null hypothesis, not only does $\mu_X = \mu_B$ under the null, but also $\sigma_X^2 = \sigma_B^2 = \sigma^2$. Using this assumption, $\mathrm{var}_0(TS) = \mathrm{var}_0(\bar{X}_n - \bar{B}_m) = \frac{\sigma^2}{n} + \frac{\sigma^2}{m} = \sigma^2(\frac{1}{n}+\frac{1}{m})$.

Test when variance known

As in Chapter 5, we test this null hypothesis by putting TS in standard score form:

$$Z = \frac{\bar{X}_n - \bar{B}_m - 0}{\sqrt{\sigma^2(\frac{1}{n}+\frac{1}{m})}}.$$

If $\alpha = 0.10$, we reject H_0 when $|Z| \geq 1.645$. If $\alpha = 0.05$, we reject H_0 when $|Z| \geq 1.960$. If $\alpha = 0.01$, we reject H_0 when $|Z| \geq 2.576$.

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, $S_X^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}{n-1}$ and $S_B^2 = \frac{\sum_{i=1}^{n}(B_i - \bar{B}_m)^2}{m-1}$ 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

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

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 $|T_{n+m-2}| \geq t_{2.576,n+m-2}$.

Test when variances unknown and unequal

This procedure is called the unequal variance t-test or unequal variance confidence interval. As before, 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 $N(\mu_B, \sigma_B^2)$. Here, $\sigma_X^2 \neq \sigma_B^2$. Then,

$$\operatorname{var}(\overline{X}_n - \overline{B}_m) = \frac{\sigma_X^2}{n} + \frac{\sigma_B^2}{m}.$$

The standard score form of the test statistic is then

$$Z = \frac{\overline{X}_n - \overline{B}_m - 0}{\sqrt{\dfrac{\sigma_X^2}{n} + \dfrac{\sigma_B^2}{m}}}.$$

The studentized standard score form of test statistic would be

$$T_? = \frac{\overline{X}_n - \overline{B}_m - 0}{\sqrt{\dfrac{S_X^2}{n} + \dfrac{S_B^2}{m}}}.$$

The problem is that the probability theory calculations that are the basis of the degrees of freedom in the pooled variance t-test are not valid. There is, however, a complicated formula (called Satterthwaite's formula) that produces an approximate degrees of freedom. I will not ask you to calculate this by hand in an examination. Fortunately, the statistical programs that you will use will calculate the unequal variance t-test and unequal variance confidence interval for you. Typically, the equal variance t-test and unequal variance t-tests have essentially equal p-values when the assumption of equal variance appears reasonable. When the assumption does not appear reasonable, the one should use the unequal variance calculations. My practice is to report the unequal variance results as calculated by a reputable statistics program.

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 $\alpha,\ \alpha \leq 0.5,$ 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, $\beta \le 0.5$, 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 $TS \ge 0 + |z_\alpha| \sqrt{\frac{\sigma_0^2}{J} + \frac{\sigma_0^2}{J}}$. 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 $\beta = \Pr_1\{\text{Accept } H_0\} \le 0.5$, 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.

Permutation Tests

Suppose that we are comparing two groups *A* and *B*, with a random sample of 2 observations from the *A* group and an independent random sample of 3 observations for the *B* group. The *A* group observations are {1,2}, and the *B* group observations are {3, 4, 5}, with an observed t-value of -3.0.
The concept is that the null hypothesis essentially is that the two groups being compared are identical so that the group labels are arbitrary. In that event, each permutation is equally likely. The two independent sample t-test is calculated for each value as in the table below.

Table 1
Permutations and t-test values

A group permutation	B group permutation	t-value
1, 2	3, 4, 5	-3.0
1, 3	2, 4, 5	-1.22
1, 4	2, 3, 5	-0.52
1, 5	2, 3, 4	0.00
2, 3	1, 4, 5	-0.52
2, 4	1, 3, 5	0.00
2, 5	1, 3, 4	0.52
3, 4	1, 2, 5	0.52
3, 5	1, 2, 4	1.22
4, 5	1, 2, 3	3.00

The permutation distribution of the t-test values for this example is the set: {-3.00, -1.22, -0.52, -0.52, 0.00, 0.00, 0.52, 0.52, 1.22, 3.00}. The two-sided permutation test p-value is equal to the number of permutations that have an absolute value of t greater than or equal to the absolute value of the t-test for the observed samples. Here, there are two permutations with absolute value of 3.0 or more. Then the two-sided permutation test p-value is 2/10=0.2, which is not a small value. This example is just to illustrate the concept. In practice, the number of permutations is very large. For example, if there are *n* observations in the *A* sample and *m* observations in the *B* sample, then there are $\binom{n+m}{n}$ permutations.

Researchers typically take a random sample of permutations to calculate an estimated permutation test p-value.

Pitman's Findings

E.J.G. Pitman studied the permutation distribution of the two independent sample t-test results. He proved that the first four moments of the permutation distribution are asymptotically the same as the first four moments of the Student's t distribution with $n+m-2$ degrees of freedom. Consequently, the p-value from Student's t-test is a good approximation to the permutation test p-value.

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, $\dfrac{Y - \mu_Y}{\sigma_Y} = Z$ 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 $\sum_{i=1}^{n} \left(\dfrac{Y_i - \mu_Y}{\sigma_Y}\right)^2$ is χ_n^2. After factoring out σ_Y^2, $\dfrac{\sum_{i=1}^{n}(Y_i - \mu_Y)^2}{\sigma_Y^2}$ 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 $\dfrac{\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2}{\sigma_Y^2}$ 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 $S^2 = \dfrac{\sum(Y_i - \bar{Y}_n)^2}{n-1}$. Since $(n-1)S^2 = (n-1)\dfrac{\sum(Y_i - \bar{Y}_n)^2}{n-1} = \sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2$, $\dfrac{(n-1)S^2}{\sigma_Y^2}$ 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.

没有用的信息

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.

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.

看表

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 $\text{var}(Y)$. 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 $s_X^2 = 917.6$, where $\bar{x}_{10}$ was the average of the ten observations sampled from X and s_X^2 was the unbiased estimate of $\mathrm{var}(X)$. Test the null hypothesis $H_0 : \mathrm{var}(X) = \mathrm{var}(Y)$ against the alternative $H_1 : \mathrm{var}(X) > \mathrm{var}(Y)$ 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 $\mathrm{var}(X)/\mathrm{var}(Y)$ 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.