Lesson 2.1

Sampling Distributions Based on the Normal Distribution

Introduction

Recall that a random sample of observations is also referred to as an "iid" (*independent and identically distributed*) sample of observations $Y_1, Y_2, \cdots, Y_n$. That is, these observations are independent and come from the same probability distribution.

Definition

A **statistic**, say T, is a function of the random variables $Y_1, Y_2, \cdots, Y_n$. A statistic can depend on known constants, but it cannot depend on unknown parameters.

To denote the dependence of T on $Y_1, Y_2, \cdots, Y_n$, we may write

$$T = T(Y_1, Y_2, \cdots, Y_n)$$

In addition, while it often be the case that $Y_1, Y_2, \cdots, Y_n$ constitute a random sample, the above definition of T holds in more general setting. In practice, it is common to view $Y_1, Y_2, \cdots, Y_n$ as **data** from an experiment or observational study and T as some summary measure (such as sample mean, sample variance, etc.).

Example 2.1.1

Suppose that $Y_1, Y_2, \cdots, Y_n$ is an iid sample from $f_Y(y)$. The following are statistics:

- $T = T(Y_1, Y_2, \cdots, Y_n) = \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
- $T = T(Y_1, Y_2, \cdots, Y_n) = \frac{1}{2}[Y_{(n/2)} + Y_{(n/2+1)}]$
- $T = T(Y_1, Y_2, \cdots, Y_n) = Y_{(1)}$

- $T = T(Y_1, Y_2, \cdots, Y_n) = Y_{(n)} - Y_{(1)}$
- $T = T(Y_1, Y_2, \cdots, Y_n) = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$

It is very important to note that since $Y_1, Y_2, \cdots, Y_n$ are random variables, any statistic $T = T(Y_1, Y_2, \cdots, Y_n)$, being a function of random variables, is also a random variable. Thus, T has its own distribution.

Definition

The probability distribution of a statistic T is called its **sampling distribution**. The sampling distribution of T describes mathematically how the values of T vary in repeated sampling from the population distribution $f_Y(y)$. Sampling distributions play a crucial role in statistics

Example 2.1.2

Suppose $Y_1, Y_2, \cdots, Y_n$ is an iid sample from $N(\mu, \sigma^2)$ and consider the statistic

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

the sample mean. It can be shown (via MGF technique) that

$$\overline{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Furthermore, the quantity

$$Z = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

Example 2.1.3

In the interest of pollution control, an experimenter records Y, the amount of bacteria per unit volume of water (measured in mg/cm^3). The population distribution for Y is assumed to be normal with mean $\mu = 48$ and variance $\sigma^2 = 100$, that is, $Y \sim N(48, 100)$.

a. What is the probability that the amount of bacteria in single water sample exceeds $50\, mg/cm^3$?

b. Suppose the experimenter takes a random sample of $n = 100$ water samples and denote the observations by $Y_1, Y_2, \cdots, Y_{100}$. What is the probability that the sample mean $\overline{Y}$ will exceed $50\, mg/cm^3$?

c. How large should the sample size n be so that $P(\overline{Y} > 50) < 0.01$?

SOLUTION: [Left as a classroom exercise!]

The Chi-square distribution

Recall that a chi-square distribution with 1 degree of freedom is a special type of Gamma distribution with $\alpha = 1/2$ and $\beta = 2$. We next show that we can also generate a random variable with a chi-square distribution from a normal distribution.

Example 2.1.3

Suppose that $Y_1, Y_2, \cdots, Y_n$ are independent observations from $N(\mu_i, \sigma_i^2)$. Find the distribution of

$$U = \sum_{i=1}^{n} \left(\frac{Y_i - \mu_i}{\sigma_i} \right)^2$$

SOLUTION

Define for each $i = 1, 2, \cdots, n$,

$$Z_i = \frac{Y_i - \mu_i}{\sigma_i}$$

Note of the following facts:

1. $Z_1, Z_2, \cdots, Z_n$ are independent $N(0, 1)$ random variables
2. $Z_1^2, Z_2^2, \cdots, Z_n^2$ are independent random variable each with $\chi^2(1)$ [from *Example 1.2.3*]

Therefore, $U = \sum_{i=1}^{n} \left(\frac{Y_i - \mu_i}{\sigma_i}\right)^2 = \sum_{i=1}^{n} Z_i^2$ has a $\chi^2(n)$ distribution.

REMARK

The case where $Y_1, Y_2, \cdots, Y_n$ are iid from $N(\mu, \sigma^2)$ directly follows from the above result. That is,

$$\sum_{i=1}^{n} \left(\frac{Y_i - \mu}{\sigma}\right)^2 \sim \chi^2(n)$$

Example 2.1.4

Suppose that $Y_1, Y_2, \cdots, Y_n$ are iid observations from $N(\mu, \sigma^2)$. Prove that

$$\frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} \left(\frac{Y_i - \overline{Y}}{\sigma}\right)^2 \sim \chi^2(n-1)$$

PROOF

First we write

$$\begin{aligned}
\underbrace{\sum_{i=1}^{n} \left(\frac{Y_i - \mu}{\sigma}\right)^2}_{W_1} &= \sum_{i=1}^{n} \left(\frac{Y_i - \overline{Y} + \overline{Y} - \mu}{\sigma}\right)^2 \\
&= \underbrace{\sum_{i=1}^{n} \left(\frac{Y_i - \overline{Y}}{\sigma}\right)^2}_{W_2} + \underbrace{\sum_{i=1}^{n} \left(\frac{\overline{Y} - \mu}{\sigma}\right)^2}_{W_3}
\end{aligned}$$

Now, we know that $W_1 \sim \chi^2(n)$, and we can also rewrite W_3 as follows:

$$\begin{aligned}
W_3 = \sum_{i=1}^{n} \left(\frac{\overline{Y} - \mu}{\sigma}\right)^2 &= n \left(\frac{\overline{Y} - \mu}{\sigma}\right)^2 \\
&= \left(\frac{\overline{Y} - \mu}{\sigma/\sqrt{n}}\right)^2 \sim \chi^2(1)
\end{aligned}$$

So, now we have

$$\begin{aligned}
W_1 &= W_2 + W_3 \\
&= \frac{(n-1)S^2}{\sigma^2} + W_3
\end{aligned}$$

Note that $\overline{Y}$ and S^2 are independent [*proof deferred to advance courses in statistics*] and since W_3 and W_2 are functions of $\overline{Y}$ and S^2, respectively, then W_3 and W_2 are independent.

Since $W_1 \sim \chi^2(n)$, thus, $m_{W_1}(t) = (1-2t)^{-n/2}$. Similarly, since $W_3 \sim \chi^2(1)$, thus, $m_{W_3}(t) = (1-2t)^{-1/2}$.

Now,

$$\begin{aligned} m_{W_1}(t) = E[e^{tW_1}] &= E[e^{t(W_2+W_3)}] \\ &= E[e^{tW_2+tW_3}] \\ &= E[e^{tW_2}] \times E[e^{tW_3}] \\ &= m_{W_2}(t) \times m_{W_3}(t) \end{aligned}$$

This means that

$$\begin{aligned} m_{W_2}(t) &= \frac{m_{W_1}(t)}{m_{W_3}(t)} \\ &= \frac{(1-2t)^{-n/2}}{(1-2t)^{-1/2}} \\ &= (1-2t)^{-(n-1)/2} \end{aligned}$$

Therefore, $W_2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$. QED

Example 2.1.5

In an ecological study examining the effects of a typhoon, researchers choose 9 plots and for each plot record the amount of dead weight material (Y, in grams). Denote the 9 dead weights as $Y_1, Y_2, \cdots, Y_9$. Assume that these observations are a random sample from $N(100, 32)$.

a. What is the probability that the sample variance S^2 of the 9 observations is less than 20?

b. How large should the sample size n be so that $P(S^2 < 20) < 0.01$?

SOLUTION

a. Recall that
$$\frac{(n-1)S^2}{\sigma^2} = \frac{8S^2}{32} \sim \chi^2(8).$$

Hence,

$$
\begin{aligned}
P(S^2 < 20) &= P\left[\frac{8S^2}{32} < \frac{8(20)}{32}\right] \\
&= P\left[\chi^2(8) < 5\right] \\
&\approx 0.24
\end{aligned}
$$

This probability can be obtained using R or MS Excel.

b. [Left as a classroom exercise!]

The t distribution

Suppose that $Z \sim N(0,1)$ and that $W \sim \chi^2(\nu)$. If Z and W are independent, then the random variable

$$
T = \frac{Z}{\sqrt{W/\nu}}
$$

has a **t** distribution with ν degrees of freedom. This is denoted as $T \sim t(\nu)$.

The PDF of $T \sim t(\nu)$ is given by

$$
f_T(t) = \begin{cases} \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi\nu}\Gamma(\nu/2)}(1+t^2/\nu)^{-(\nu+1)/2}, & -\infty < t < \infty \\ 0, & \text{elsewhere} \end{cases}
$$

<u>Derivation</u>:

Let $Z \sim N(0,1)$ and $W \sim \chi^2(\nu)$ be independent random variables. The joint PDF of Z and W is

$$
f_{Z,W}(z,w) = \underbrace{\frac{1}{\sqrt{2\pi}}e^{-z^2/2}}_{N(0,1)} \times \underbrace{\frac{1}{\Gamma(\nu/2)2^{\nu/2}}w^{(\nu/2)-1}e^{-w/2}}_{\chi^2(\nu)}
$$

for $-\infty < z < \infty$ and $w > 0$.

Consider the bivariate transformation

$$
\begin{aligned}
T &= g_1(Z,W) = \frac{Z}{\sqrt{W/\nu}} \\
U &= g_2(Z,W) = W
\end{aligned}
$$

The support of (Z, W) is $R_{Z,W} = \{(z, w) : -\infty < z < \infty, w > 0\}$, while the support of (T, U) is $R_{T,U} = \{(t, u) : -\infty < t < \infty, u > 0\}$. Obviously, the vector-valued function g is one-to-one, so the inverse transformations exists and is given by

$$\begin{aligned} z &= g_1^{-1}(t, u) = t\sqrt{u/\nu} \\ w &= g_2^{-1} = u \end{aligned}$$

The Jacobian of the transformation is

$$\begin{aligned} J &= det \begin{bmatrix} \frac{\partial g_1^{-1}(t,u)}{\partial t} & \frac{\partial g_1^{-1}(t,u)}{\partial u} \\ \frac{\partial g_2^{-1}(t,u)}{\partial t} & \frac{\partial g_2^{-1}(t,u)}{\partial u} \end{bmatrix} \\ &= det \begin{bmatrix} \sqrt{u/\nu} & t/2\sqrt{u\nu} \\ 0 & 1 \end{bmatrix} \\ &= \sqrt{u/\nu} \end{aligned}$$

Hence, the joint PDF of (T, U) is,

$$\begin{aligned} f_{T,U}(t, u) &= f_{Z,W}[g_1^{-1}(t, u), g_2^{-1}(t, u)]|J| \\ &= \frac{1}{\sqrt{2\pi}} e^{-\frac{(t\sqrt{u/\nu})^2}{2}} \times \frac{1}{\Gamma(\nu/2)2^{\nu/2}} u^{(\nu/2)-1} e^{-u/2} \times \left|\sqrt{u/\nu}\right| \\ &= \frac{1}{\sqrt{2\pi\nu}\Gamma(\nu/2)2^{\nu/2}} u^{[(\nu+1)/2]-1} e^{-\frac{u}{2}\left(1+\frac{t^2}{\nu}\right)} \end{aligned}$$

To get the PDF of T, we integrate the above joint PDF:

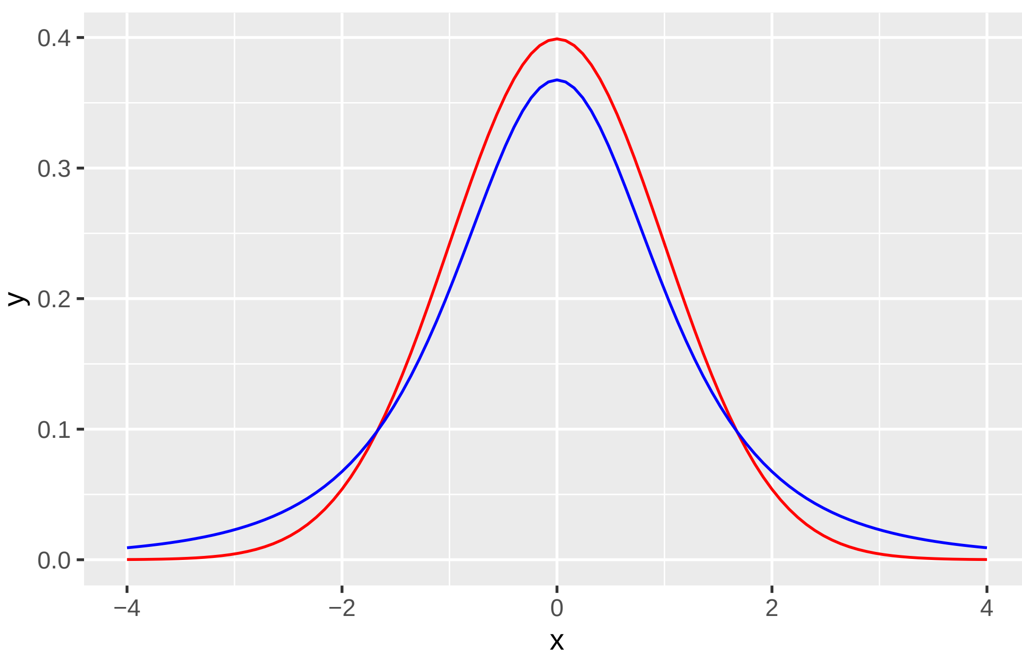
$$\begin{aligned} f_T(t) &= \int_0^\infty f_{T,U}(t, u)\, du \\ &= \int_0^\infty \frac{1}{\sqrt{2\pi\nu}\Gamma(\nu/2)2^{\nu/2}} u^{[(\nu+1)/2]-1} e^{-\frac{u}{2}\left(1+\frac{t^2}{\nu}\right)}\, du \\ &= \frac{1}{\sqrt{2\pi\nu}\Gamma(\nu/2)2^{\nu/2}} \int_0^\infty \underbrace{u^{[(\nu+1)/2]-1} e^{-\frac{u}{2}\left(1+\frac{t^2}{\nu}\right)}}_{\text{Gamma(a,b) kernel}}\, du \end{aligned}$$

where $a = (\nu + 1)/2$ and $b = 2\left(1 + \frac{t^2}{\nu}\right)^{-1}$. Thus,

$$
\begin{aligned}
f_T(t) &= \frac{1}{\sqrt{2\pi\nu}\Gamma(\nu/2)2^{\nu/2}} \int_0^\infty u^{[(\nu+1)/2]-1} e^{-\frac{u}{2}\left(1+\frac{t^2}{\nu}\right)}\, du \\
&= \frac{\Gamma[(\nu+1)/2]\left[2\left(1+\frac{t^2}{\nu}\right)^{-1}\right]^{(\nu+1)/2}}{\sqrt{2\pi\nu}\Gamma(\nu/2)2^{\nu/2}} \int_0^\infty \frac{1}{\Gamma[(\nu+1)/2]\left[2\left(1+\frac{t^2}{\nu}\right)^{-1}\right]^{(\nu+1)/2}} u^{[(\nu+1)/2]-1} e^{-\frac{u}{2}\left(1+\frac{t^2}{\nu}\right)}\, du \\
&= \frac{\Gamma[(\nu+1)/2]\left[2\left(1+\frac{t^2}{\nu}\right)^{-1}\right]^{(\nu+1)/2}}{\sqrt{2\pi\nu}\Gamma(\nu/2)2^{\nu/2}} \\
&= \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi\nu}} \left(1+t^2/\nu\right)^{-(\nu+1)/2} \quad QED
\end{aligned}
$$

FACTS ABOUT THE t DISTRIBUTION:

- continuous and symmetric about 0
- indexed by a parameter called the **degrees of freedom**, denoted by ν (an integer which is related to sample size)
- as $\nu \to \infty$, $t(\nu) \to N(0,1)$; in general the t distribution is less peaked and has more mass in the tails than the standard normal distribution
- $E(T) = 0$ and $V(T) = \frac{\nu}{\nu-2}, \nu > 2$



Suppose $Y_1, Y_2, \cdots, Y_n$ is an iid sample from $N(\mu, \sigma^2)$. We know that

$$\frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

and

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

Since $\overline{Y}$ and S^2 are independent so are the above quantities. Thus,

$$t = \frac{\frac{\overline{Y}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}} = \frac{\overline{Y} - \mu}{S/\sqrt{n}}$$

has a $t(n-1)$ distribution.

The F distribution

Suppose that $W_1 \sim \chi^2(\nu_1)$ and that $W_2 \sim \chi^2(\nu_2)$. If W_1 and W_2 are independent, then the quantity

$$F = \frac{W_1/\nu_1}{W_2/\nu_2}$$

has an F distribution with ν_1 and ν_2 degrees of freedom. We call ν_1 and ν_2 as the numerator and denominator degrees of freedom, respectively.

Definition

If $W \sim F(\nu_1, \nu_2)$, then the PDF of W, is given by

$$f_W(w) = \begin{cases} \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)\left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2} w^{(\nu_1-2)/2}}{\Gamma(\frac{\nu_1}{2})\Gamma(\frac{\nu_2}{2})\left(1+\frac{\nu_1 w}{\nu_2}\right)^{(\nu_1+\nu_2)/2}}, \; w > 0 \\ 0, \text{otherwise} \end{cases}$$

PROOF: [Left as a challenge!].

FACTS ABOUT THE F DISTRIBUTION:

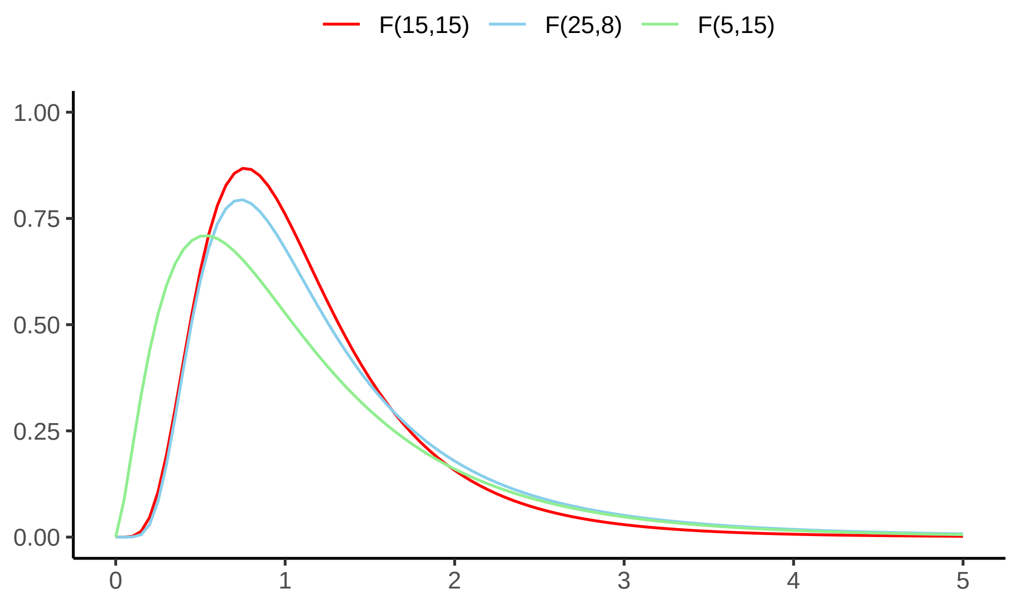
- It is continuous and skewed to the right.

- It is indexed by two degrees of freedom, ν_1 and ν_2, which are both integers and related to the sample sizes.
- If $W \sim F(\nu_1, \nu_2)$, then

$$E(W) = \frac{\nu_2}{\nu_2 - 2},\ \nu_2 > 2$$

and

$$V(W) = \frac{2\nu_2^2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 2)^2(\nu_2 - 4)},\ v_2 > 4$$



FUNCTIONS of t and F

The following results are useful. Each of the following facts can be proven using the method of transformations.

1. If $W \sim F(\nu_1, \nu_2)$, then $1/W \sim F(\nu_2, \nu_1)$.
2. If $T \sim t(\nu)$, then $T^2 \sim F(1, \nu)$.
3. If $W \sim F(\nu_1, \nu_2)$, then $\frac{(\nu_1/\nu_2)W}{1+(\nu_1/\nu_2)W} \sim Beta(\nu_1/2, \nu_2/2)$.

Example 2.1.6

Suppose $Y_1, Y_2, \cdots, Y_n$ is a random sample from a $N(\mu, \sigma^2)$ distribution. Recall that

$$Z = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

and

$$T = \frac{\overline{Y} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

Now,

$$\begin{aligned}
T^2 &= \left(\frac{\overline{Y} - \mu}{S/\sqrt{n}}\right)^2 \\
&= \left(\frac{\overline{Y} - \mu}{\sigma/\sqrt{n}}\right)^2 \frac{\sigma^2}{S^2} \\
&= \frac{\left(\frac{\overline{Y} - \mu}{\sigma/\sqrt{n}}\right)^2 /1}{\frac{(n-1)S^2}{\sigma^2}/(n-1)} \\
&= \frac{"\chi 2(1)"/1}{"\chi^2(n-1)"/(n-1)} \\
&\sim F(1, n-1)
\end{aligned}$$

Example 2.1.7

Consider two independent random samples of sizes n_1 and n_2

$$Y_{11}, Y_{12}, \cdots, Y_{1n} \sim N(\mu_1, \sigma_1^2) Y_{21}, Y_{22}, \cdots, Y_{2n} \sim N(\mu_2, \sigma_2^2)$$

Define the statistics

$$
\begin{aligned}
\overline{Y}_{1.} &= \frac{1}{n_1}\sum_{j=1}^{n_1} Y_{1j} \equiv \text{mean of sample 1} \\
\overline{Y}_{2.} &= \frac{1}{n_2}\sum_{j=1}^{n_2} Y_{2j} \equiv \text{mean of sample 2} \\
S_1^2 &= \frac{1}{n_1 - 1}\sum_{j=1}^{n_1} (Y_{1j} - \overline{Y}_{1.})^2 \equiv \text{variance of sample 1} \\
S_2^2 &= \frac{1}{n_2 - 1}\sum_{j=1}^{n_2} (Y_{2j} - \overline{Y}_{2.})^2 \equiv \text{variance of sample 2}
\end{aligned}
$$

We know that

$$\frac{(n_1 - 1)S_1^2}{\sigma_1^2} \sim \chi^2(n_1 - 1)$$

and

$$\frac{(n_2 - 1)S_2^2}{\sigma_2^2} \sim \chi^2(n_2 - 1)$$

Since the samples are independent, then

$$
\begin{aligned}
F &= \frac{\frac{(n_1-1)S_1^2}{\sigma_1^2}/(n_1 - 1)}{\frac{(n_2-1)S_2^2}{\sigma_2^2}/(n_2 - 1)} \\
&= \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \\
&\sim F(n_1 - 1, n_2 - 1)
\end{aligned}
$$

Now, if the two population variances are equal. that is, $\sigma_1^2 = \sigma_2^2$, then

$$F = \frac{S_1^2}{S_2^2} \sim F(n_1 - 1, n_2 - 1)$$

The above result is the basis of the F test for two population variances.