Homework Due Dec. 4
p. 313: 6.28, 6.36
p. 334: 6.83
p. 351: 7.5, 7.7, 7.8

Chapter 7: Inferences for the Population Variance σ^2

• As before, we are interested in making inferences about a population of individuals, units, objects, or subjects from a sample drawn from a population

• Previously, we discussed estimators and tests which involved the population means or, or more generally, their centers

• Now we turn our attention to measures of population spread, specifically, the population variance σ^2 and the population standard deviation $\sigma = \sqrt{\sigma^2}$

• The usual point estimator of σ^2 is the sample variance

$$s^2 = \frac{\sum_{i=1}^{n} (y_i - \overline{y})^2}{n-1},$$

and the usual estimator of the population standard deviation σ is $s = \sqrt{s^2}$.

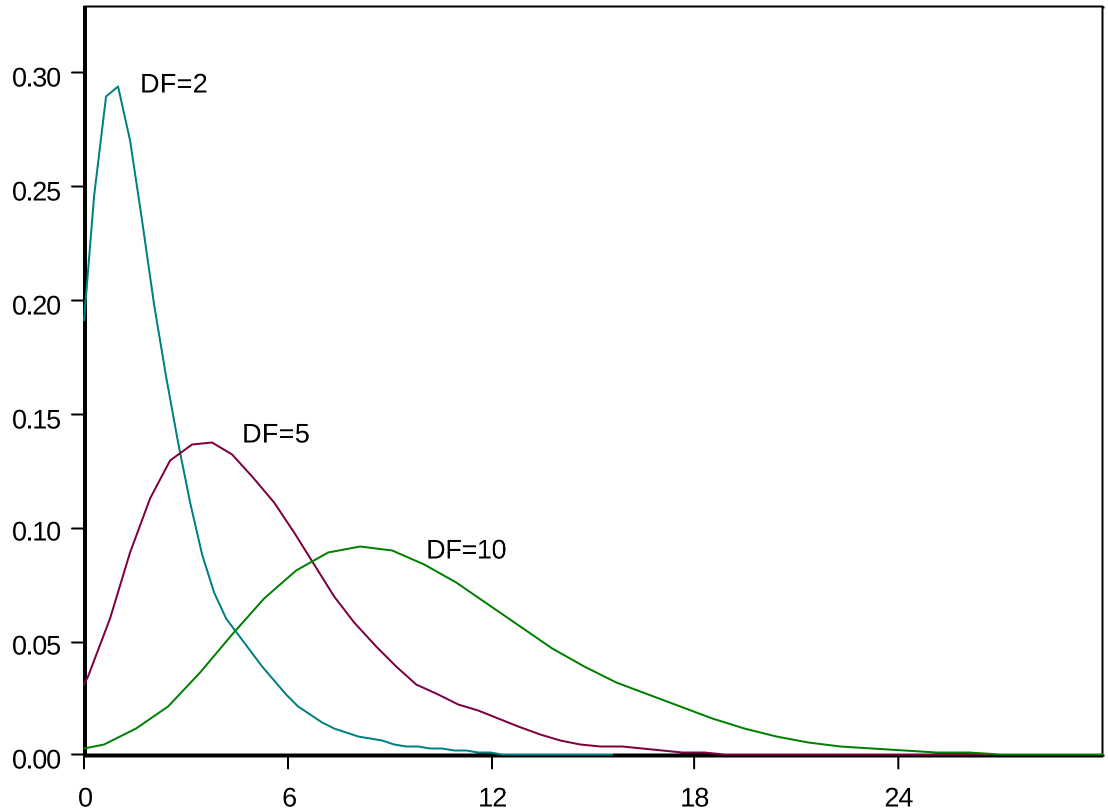
• A key distributional result is for this Chapter is: Given a random sample of size n from a normal population, then the sample variance, multiplied by a suitable constant, has a *chi-square* distribution.

• Specifically,

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

where χ^2_{n-1} is a chi-square random variable with $n-1$ degrees of freedom. The constant is $(n-1)/\sigma^2$.

Figure Estimated probability density functions (smoothed histograms) for three chi-square random variables. 1000 observations were used to generate of the curves



• The range of the chi-square random variable is $(0, \infty)$, and it is right-skewed unless n is largerthan 10

• The mean of the χ^2_{n-1} distribution is $n-1$ and the variance is $2(n-1)$. For the data used above, the sample statistics are

Degrees of Freedom	Sample Size	Mean	Sample Variance
2	1000	1.97	3.63
5	1000	5.08	10.3
10	1000	9.86	19.4

• It is sometimes insufficient to present only the point estimate s^2 of σ^2. Sometimes a confidence interval is needed.

• What is a confidence interval? How is it interpreted?

• A *confidence interval* is a set of values between two endpoints L (lower) and U (upper)

• The *procedure (or process)* of constructing a $100(1-\alpha)\%$ CI will contain the true unknown parameter (in this case, σ^2) $100(1-\alpha)\%$ of the time, if the process of collecting the data and computing the interval is repeated many times, the samples are truly independent, and the population is truly normal

• If the data are not a random sample from a normal distribution, then the procedure (or process) of constructing th $100(1-\alpha)\%$ CI will contain the true unknown parameter (in this case, σ^2) an unknown percentage of the time

• A $100(1-\alpha)\%$ confidence interval for σ^2 is derived from the probability statement

$$1-\alpha = P\Big(\chi^2_L < \frac{(n-1)s^2}{\sigma^2} < \chi^2_U\Big)$$

where χ^2_L and χ^2_U are the $100\alpha/2$th and $100(1-\alpha/2)$ percentiles of the χ^2_{n-1} distribution.

• We rearrange the probability statement as follows to get a CI for σ^2:

$$\begin{aligned} 1-\alpha &= P\Big(\chi^2_L < \frac{(n-1)s^2}{\sigma^2} < \chi^2_U\Big) \\ &= P\Big(\frac{\chi^2_L}{(n-1)s^2} < \frac{1}{\sigma^2} < \frac{\chi^2_U}{(n-1)s^2}\Big) \\ &= P\Big(\frac{(n-1)s^2}{\chi^2_U} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_L}\Big). \end{aligned}$$

• This calculation tells us what the lower (L) and upper (U) bounds are, namely,

$$L = \frac{(n-1)s^2}{\chi^2_U} \text{ and } U = \frac{(n-1)s^2}{\chi^2_L}$$

• Table 7, p. 1100, contains percentiles of some chi-square distributions

Example: Lamb tapeworms (see p. 272)

• 7 lambs receive a drug against tapeworms, 6 receive no drug. Responses were worm counts in slaughtered lambs after 6 months.

- The summary statistics were

$\bar{y}_1 = 9.00$, $s_1^2 = 38.67$,$n_1 = 7$,

$\bar{y}_2 = 40.17$, $s_2^2 = 258.2$, $n_2 = 6$.

• Recall that a separate variance t-test was used to test $H_0 : \mu_1 - \mu_2 = 0$ versus $H_2 : \mu_1 - \mu_2 > 0$. Was it necessary to use the separate variance t?

• To assess (informally) the necessity of the separate variance t, we can compare 95% confidence intervals for σ_1^2 and σ_2^2 and determine if they overlap.

• For $n_1 = 7$, we have df $= 6$. In addition, let $\alpha = 0.05$ so that $\alpha/2 = 0.025$. From Table 7, $\chi_L^2 = 1.237$ and $\chi_U^2 = 14.45$. Then,

$$L = \frac{(n-1)s_1^2}{\chi_U^2} = \frac{6 \times 38.67}{14.45} = 16.05$$

and

$$U = \frac{(n-1)s_1^2}{\chi_L^2} = \frac{6 \times 38.67}{1.237} = 187.5.$$

• Finally, the 95% confidence interval for σ_1^2 is 16.05 to 187.5

• I am 95% confident that the true value is in the interval

• I cannot say that there is a 0.95 probability that σ_1^2 is between 16.05 and 187.5. In truth, $0.95 \neq P(16.05 < \sigma_1^2 < 187.5)$ because there are no random variables involved in the statement. (σ_1^2 is a parameter)

• The confidence interval formed by L and U is

1) all those values of σ_1^2 that are consistent with the data. Any value outside the interval is not consistent with the data

2) constructed using a method that captures σ_1^2 95% of the time, assuming that population of interest is normal and sampled randomly.

• The same set of steps for the second data set yields a 95% confidence interval for σ_2^2 given by $(100.6, 1553)$.

• While it is true that the intervals overlap, the degree of overlap is small, indicating that there is some statistical evidence that $\sigma_1^2 \neq \sigma_2^2$

AStatistical Test for σ^2

• The test requires data that are a random sample from a normal distribution [to achieve the specified Type I error rate (α)]

• The hypotheses of interest are:

$H_0 : \sigma^2 = \sigma_0^2$ versus
1. $H_a : \sigma^2 > \sigma_0^2$, or
2. $H_a : \sigma^2 < \sigma_0^2$, or
3. $H_a : \sigma^2 \neq \sigma_0^2$

• The test statistic is the chi-square statistic

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}.$$

• We reject H_0 if χ^2 belongs to the rejection region R. R is determined by H_a, α, and the degrees of freedom df $= n - 1$

• For each of the three research hypotheses,
1. $H_a : \sigma^2 > \sigma_0^2$,

$$R = \left\{\chi^2 \mid \chi^2 > \chi_U^2\right\}$$

and χ_U^2 is the $1-\alpha$ percentile from the χ_{n-1}^2 distribution
2. $H_a : \sigma^2 < \sigma_0^2$,

$$R = \left\{\chi^2 \mid \chi^2 < \chi_L^2\right\}$$

and χ_L^2 is the α percentile from the χ_{n-1}^2 distribution
3. $H_a : \sigma^2 \neq \sigma_0^2$,

$$R = \left\{\chi^2 \mid \chi^2 > \chi_U^2 \text{ or } \chi^2 < \chi_L^2\right\}$$

and χ_L^2 and χ_U^2 are $\alpha/2$ and $1-\alpha/2$ percentiles from the χ_{n-1}^2 distribution, respectively

• A more general method of testing is to compute and interpret the p-value of associated with the test statistic χ^2. Specifically, for the three alternative hypotheses,
1. $H_a : \sigma^2 > \sigma_0^2$,

$$\text{p-value} = P\left(\chi_{n-1}^2 > \chi^2\right)$$

where χ_{n-1}^2 represents the chi-square random variable with $n-1$ degrees of freedom, and χ^2 is the observed value of the test statistic.
2. $H_a : \sigma^2 < \sigma_0^2$,

$$\text{p-value} = P\left(\chi_{n-1}^2 < \chi^2\right)$$

3. $H_a : \sigma^2 \neq \sigma_0^2$,

$$\text{p-value} = \begin{cases} 2P\left(\chi_{n-1}^2 < \chi^2\right) & \text{if } \chi^2 < n-1, \\ 2P\left(\chi_{n-1}^2 > \chi^2\right) & \text{if } \chi^2 \geq n-1 \end{cases}$$

Example test $H_0 : \sigma^2 = 20$ versus $H_a : \sigma^2 > 20$ using $s_1^2 = 38.67$ and $n_1 = 7$:

$$\chi^2 = \frac{(n-1)s_1^2}{\sigma_0^2} = \frac{6\times 38.67}{20} = 11.60$$

$$\Rightarrow \text{p-value} = P\left(\chi^2_{n-1} > 11.60\right) = 0.071.$$

• My interpretation is that there is *some* statistical evidence that $\sigma^2 = 20$ is false and $\sigma^2 > 20$ is true

• Unlike inferential procedures regarding the population mean(s), violations of the normality assumption are severe. This is because the Central Limit Theorem is about the sample mean $\overline{X}$, and not about s^2

7.4 Estimation and Tests for Comparing Two Population Variances $\boldsymbol{\sigma_1^2}$ and $\boldsymbol{\sigma_2^2}$

• The general objective is to compare two population variances σ_1^2 and σ_2^2. A central objective is to lay out a test of

$$H_0 : \sigma_1^2 = \sigma_2^2 \text{ versus } H_0 : \sigma_1^2 \neq \sigma_2^2$$

• We begin by assuming that the data are two independent samples of size n_1 and n_2, respectively, from normal populations with variances σ_1^2 and σ_2^2, respectively.

• The key distributional result is: if the data are two independent samples from normal populations with the same variance, i.e., $\sigma_1^2 = \sigma_2^2$, then

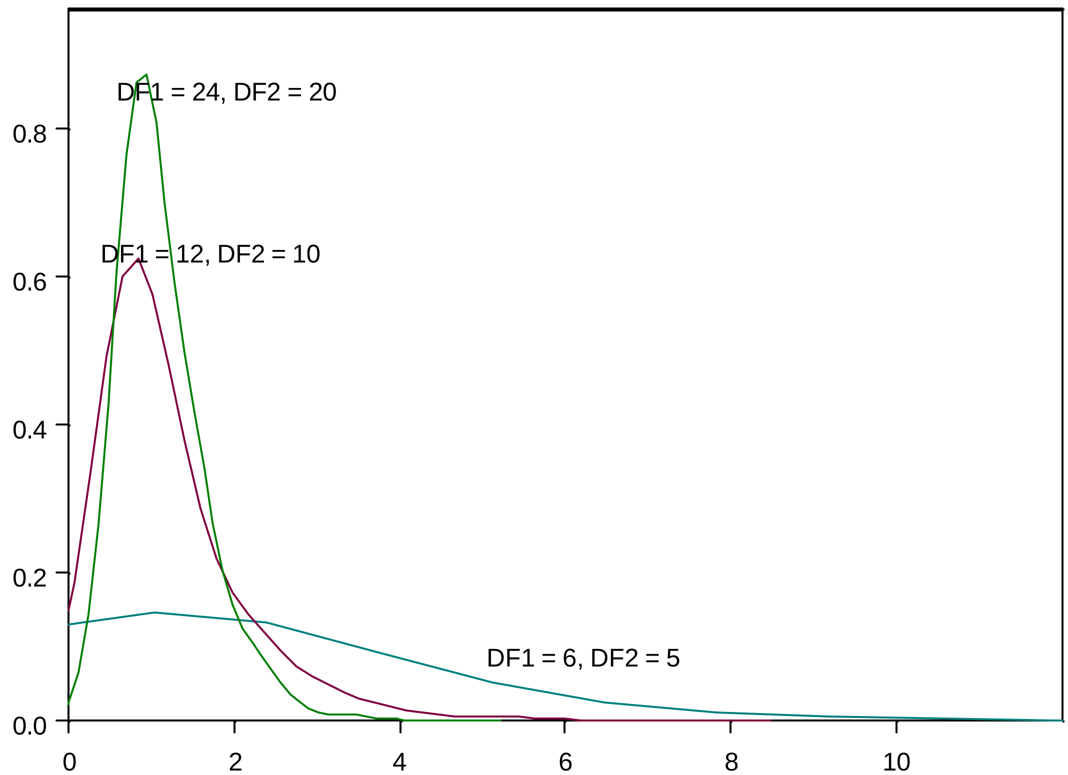
$$F = \frac{s_1^2}{s_2^2} \sim \mathcal{F}_{\text{df}_1,\text{df}_2}$$

where $\mathcal{F}_{\text{df}_1,\text{df}_2}$ represents the F-distribution with $\text{df}_1 = n_1 - 1$ numerator and $\text{df}_2 = n_2 - 1$ denominator degrees of freedom respectively

• However, if $\sigma_1^2 \neq \sigma_2^2$, then the distribution of s_1^2/s_2^2 is not $\mathcal{F}_{\text{df}_1,\text{df}_2}$, and the observed value of the ratio will inconsistent with the F-distribution.

• Therefore, the test is based on comparing s_1^2/s_2^2 to the $\mathcal{F}_{\text{df}_1,\text{df}_2}$ distribution

Figure Estimated probability density functions (smoothed histograms) for three F random variables. 1000 observations were used to generate of the curves



Notation: the distribution is $\mathcal{F}_{\text{df}_1,\text{df}_2}$ and the random variable is $F = s_1^2/s_2^2$

• $f_{\alpha,\text{df}_1,\text{df}_2}$ is the α-level upper tail value from the $\mathcal{F}_{\text{df}_1,\text{df}_2}$ distribution. That is, if $F \sim \mathcal{F}_{\text{df}_1,\text{df}_2}$, then

$$\alpha = P(F > f_{\alpha,\text{df}_1,\text{df}_2})$$

• There is an important relationship between $f_{\alpha,\mathrm{df}_1,\mathrm{df}_2}$ and $f_{1-\alpha,\mathrm{df}_2,\mathrm{df}_1}$, namely,

$$f_{\alpha,\mathrm{df}_1,\mathrm{df}_2} = \frac{1}{f_{1-\alpha,\mathrm{df}_2,\mathrm{df}_1}}$$

• Table 8, p. 1103-1113, is the F-table. It contains only left-tail areas and only for values of α less than or equal to 0.25. We will have to use the inverse relationship above

• For example, if $F \sim \mathcal{F}_{1,5}$, $\alpha = 0.05$, we find $F_{0.05,1,5} = 6.61$. This means that

$$0.05 = P(F > f_{0.05,1,5}) = P(F > 10.01)$$

• For example, if $\alpha = 0.95$ and we want to find $f_{\alpha,\mathrm{df}_1,\mathrm{df}_2} = f_{0.95,1,5}$, we look up

$$f_{1-\alpha,\mathrm{df}_2,\mathrm{df}_1} = f_{0.05,5,1} = 230.2,$$

and compute

$$f_{0.95,1,5} = \frac{1}{f_{005,5,1}} = \frac{1}{230.2} = 0.0043.$$

Then,

$$0.95 = P(F > f_{0.95,1,5}) = P(F > 0.0043)$$

A Statistical Test of $H_0 : \sigma_1^2 = \sigma_2^2$

• The test assumes that the data are independent random samples from two normal distributions

• We suppose that random samples of size n_1 and n_2 have been collected from populations with variances σ_1^2 and σ_2^2, respectively. Consider testing $H_0 : \sigma_1^2 = \sigma_2^2$ versus

1. $H_a : \sigma_1^2 > \sigma_2^2$
2. $H_a : \sigma_1^2 \neq \sigma_2^2$

• It is not necessary to lay out the third alternative hypothesis $H_a : \sigma_1^2 < \sigma_2^2$ because we can use the methods for 1 after relabeling σ_1^2 as σ_2^2, and σ_2^2 as σ_1^2

• For a test of $H_0 : \sigma_1^2 = \sigma_2^2$ versus $H_a : \sigma_1^2 > \sigma_2^2$, the test statistic is

$$F = \frac{s_1^2}{s_2^2}.$$

If H_0 is true, then $F \sim \mathcal{F}_{\text{df}_1,\text{df}_2}$

• To test $H_0 : \sigma_1^2 = \sigma_2^2$ versus $H_a : \sigma_1^2 \neq \sigma_2^2$, the F-statistic is computed according to

$$F = \begin{cases} \dfrac{s_1^2}{s_2^2}, & \text{if } s_1^2 \geq s_2^2; \\ \dfrac{s_2^2}{s_1^2}, & \text{if } s_1^2 < s_2^2. \end{cases}$$

• In other words, we divide the larger of s_1^2 and s_2^2 by the smaller of s_1^2 and s_2^2. Consequently, F will always be larger than 1.

• For a α-level test, and 1) $H_a : \sigma_1^2 > \sigma_2^2$, we reject H_0 in favor of H_a whenever $F > F_{\alpha,\text{df}_1,\text{df}_2}$

• For 2) $H_a : \sigma_1^2 \neq \sigma_2^2$, we reject H_0 in favor of H_a whenever $F > F_{\alpha/2,\text{df}_1,\text{df}_2}$

• It is sometime preferable to report a p-value rather than the accept/reject decision.

• If H_a is *one-tailed*, then p-value $= P(F > f)$ where f is the observed value of F. Use Table 8, p. 1103, df_1 and df_2, or the SPSS function CDF.F

• If H_a is *two-tailed*, then p-value $= 2P(F > f)$. Use Table 8, p. 1103, df_1 and df_2

Example: The tapeworm example (see p. 272)

• 7 lambs receive a drug against tapeworms, 6 receive no drug. Responses were worm counts in slaughtered lambs after 6 months.

$s_1^2 = 38.67, n_1 = 7,$

$s_2^2 = 258.2,\ n_2 = 6.$

• Test $H_0 : \sigma_1^2 = \sigma_2^2$ versus $H_a : \sigma_1^2 \neq \sigma_2^2$:

$$F = \frac{s_2^2}{s_1^2} = \frac{258.2}{38.67} = 6.68.$$

• Note that $\text{df}_1 = n_2 - 1 = 6 - 1 = 5$ is the numerator degrees of freedom, and $\text{df}_2 = n_1 - 1 = 7 - 1 = 6$

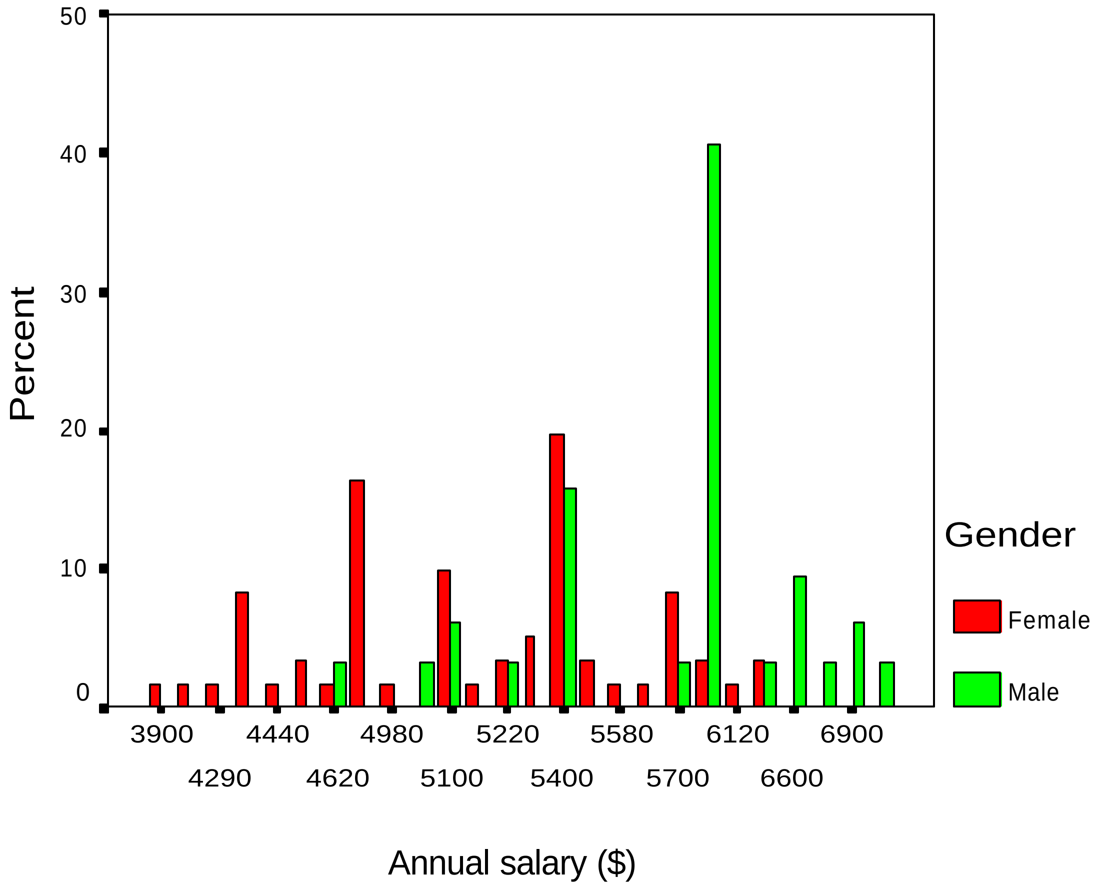
• Under $H_0,\ F \sim \mathcal{F}_{5,6} \Rightarrow$ p-value $= 2P(F > 6.68) = 2 \times 0.0194$ $= 0.039.$

• Based on this test, a separate-variance t-statistic is preferred to the pooled-two sample t-statistic for a test of equality of population means

Case Study From Roberts, H.V. 1979. "Harris Trust and Savings Bank: An Analysis of Employee Compensation," Report 7946, Center for Mathematical Studies in Business and Economics, University of Chicago Graduate School of Business. See Ramsey. F.L. and Schafer, D.W. 2001. *The Statistical Sleuth, 2nd ed*. Duxbury, for more details.

• A lawsuit was filed against the Harris Trust and Savings Bank claiming that as a group, new male employees received larger starting salaries than new women employees.

Figure. Bar chart showing male and female starting salaries of skilled, entry-level clerical workers hired by the bank from 1969 to 1977. (32 male starting salaries, and 61 female starting salaries)



- **Statistical Summary**

$\overline{X}_1 = \$5957,\ s_1 = \$690.7$ (Males)
$\overline{X}_2 = \$5138,\ s_2 = \$539.9$ (Females)

- Are these differences indicative of a difference among *all* (?) employees?

- Our primary interest is on the true population means, or more precisely, the expected starting salaries of male and female skilled, entry-level clerical workers hired by the Harris Trust and Savings

- A two-sample t-test (either the pooled or the separate variance t) can be used to assess the evidence supporting the plaintiffs' claim

- SPSS provides a test of equality of variances against the alternative that they are different, i.e.,

$$H_0 : \sigma_1^2 = \sigma_2^2 \text{ versus } H_a : \sigma_1^2 \neq \sigma_2^2.$$

The SPSS test is a *robust* test (Levene's), and generally preferable to the F-test above

- **Levene's test** is roughly equivalent to a two-sample t test where the observations of interest are the absolute deviations about the respective sample medians, i.e., $Z_{1i} = |X_{1i} - M_1|$, $i = 1, \ldots, n_1$, and $Z_{2j} = \left|X_{2j} - M_2\right|$, $j = 1, \ldots, n_2$
- Levene's test statistic is an F
- Levene's test is robust because the Central Limit Theorem insures that when n_1 and n_2 are large, the test statistic is approximately normal in distribution regardless of the distribution of Z_{1i}, $i = 1, \ldots, n_1$, and Z_{2j}, $j = 1, \ldots, n_2$
- The ordinary F-statistic above is sensitive to departures from normality
- SPSS yields $F = 0.344$ and p-value $= 0.56$. This result implies that the pooled variance t statistic is acceptable for a test of equality of means
- The test statistic for a test of

$$H_0 : \mu_1 = \mu_2 \text{ versus } H_a : \mu_1 \neq \mu_2$$

was $t = 6.29$ with 91 degrees of freedom. The p-value is less than 0.001, and we conclude that these data provide substantial statistical evidence of differences in the expected starting salary

A Confidence Interval for σ_1^2/σ_2^2

- See Ott and Longnecker, p. 359
- A $100(1-\alpha)\%$ confidence interval for σ_1^2/σ_2^2 is (L, U) where

$$L = \frac{s_1^2}{s_2^2} f_{1-\alpha/2,\mathrm{df}_2,\mathrm{df}_1} \text{ and } U = \frac{s_1^2}{s_2^2} f_{\alpha/2,\mathrm{df}_2,\mathrm{df}_1}$$

• Because Table 8 does not have values of $f_{1-\alpha/2,\text{df}_2,\text{df}_1}$, we look up $f_{\alpha/2,\text{df}_1,\text{df}_2}$ and compute

$$f_{1-\alpha/2,\text{df}_2,\text{df}_1} = \frac{1}{f_{\alpha/2,\text{df}_1,\text{df}_2}}$$

for determining L

Example: The tapeworm example

- $s_1^2 = 38.67, n_1 = 7$,
- $s_2^2 = 258.2,\ n_2 = 6$.
- To compute a 95% CI for σ_1^2/σ_2^2, note that $\text{df}_1 = 6$ and $\text{df}_2 = 5$. Then,

$$\begin{aligned} f_{1-\alpha/2,\text{df}_2,\text{df}_1} &= f_{0.975,5,6} \\ &= \frac{1}{f_{\alpha/2,\text{df}_1,\text{df}_2}} = \frac{1}{f_{0.025,6,5}} \\ &= \frac{1}{6.98} = 0.143. \end{aligned}$$

and

$$f_{\alpha/2,\text{df}_1,\text{df}_2} = f_{0.025,5,6} = 5.99.$$

Then, the confidence interval is

$$\left(\frac{38.67}{258.2} \times 0.143, \frac{38.67}{258.2} \times 5.99\right) = (0.021, 0.897).$$

- Note that $1 \notin (0.021, 0.897)$, so the CI provides evidence that $\sigma_1^2/\sigma_2^2 \neq 1$. Hence, the CI also provides evidence that $\sigma_1^2 \neq \sigma_2^2$
- A 95% CI for σ_1/σ_2 is $(\sqrt{0.021}, \sqrt{0.897})$.