11: Comparing Group Variances

Review of Variance

Parametric measures of variability are often based on **sum of squares (*SS*)** around the mean:

$$SS = \sum (x_i - \bar{x})^2 \qquad \textbf{(1)}$$

For the **data set {3, 4, 5, 8},** $\bar{x}$ = 5 and $SS = (3-5)^2 + (4-5)^2 + (5-5)^2 + (8-5)^2 = 4 + 1 + 0 + 9 = 14$.

The **variance** is the mean sum or squares ("mean square"). The symbol σ^2 denotes the population variance (parameter) and s^2 denotes the sample variance. The population variance is seldom known, so we calculate the sample variance:

$$s^2 = \frac{SS}{(n-1)} \qquad \textbf{(2)}$$

For **{3, 4, 5, 8}**, $s^2 = 14 / 3 = 4.667$. This is an unbiased estimate of σ^2.

The **standard deviation** is the square root of the variance ("root mean square"):

$$s = \sqrt{s^2} = \sqrt{\frac{SS}{n-1}} \qquad \textbf{(3)}$$

For **{3, 4, 5, 8}**, $s = \sqrt{(4.667)} = 2.16$. This is an unbiased estimate of σ.

Interpretation of the standard deviation is tricky. One thing to keep in mind is that big standard deviations indicate big "spreads" and small standard deviations indicate small spreads. For example, if one group has a standard deviation of 15 and an other has a standard deviation of 2, there is much greater variability in the first group.

There are also rules for interpreting standard deviations. For Normally distributed data, 68% of values lie in $\mu \pm \sigma$, 95% lie in $\mu \pm 2\sigma$, and nearly all values lie in $\mu \pm 3\sigma$. Most data are not Normally distributed, in which case we can use **Chebychev's rule** which states that *at least* 75% of the values lie within $\mu \pm 2\sigma$.

Illustrative example (data set = agebycen.sav for center 1). Ages (years) of patients at a given center are:

60, 66, 65, 55, 62, 70, 51, 72, 58, 61, 71, 41, 70, 57, 55, 63, 64, 76, 74, 54, 58, 73

The mean $\bar{x}$ = 62.545
The sum of squares $SS = (60-62.545)^2 + (66-62.545)^2 + \cdots + (73-62.545)^2 = 1579.45$.
The sample variance $s^2 = 1579.455/ (22-1) = 75.212$
The standard deviation $s = \sqrt{75.212} \doteq 8.7$.
From Chebychev's rule we predict that *at least* 75% of the ages lies in the interval $62.5 \pm (2)(8.7) = 62.5 \pm 17.4$, or between 45.1 and 79.9.

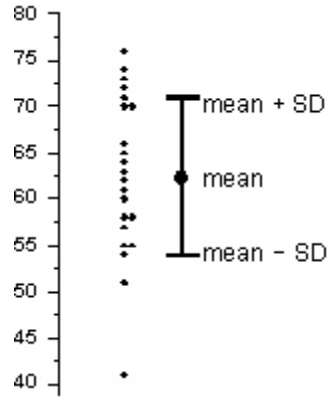
Graphical Representations of Spread

The data in the prior illustrative example (n = 22) are displayed as a stemplot:

```
4|1
4|
5|14
5|55788
6|01234
6|56
7|001234
7|6
×10 (years)
```

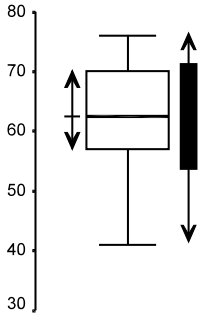
We can see from this plot that data spread from 41 to 76 around a center of about 60-something. There is a low outlier, and the distribution have a negative skew.

You can also display the data with a **dot plot** or **mean ± standard deviation plot** as follows:



Be aware that error bars in statistical charts may represent standard deviations, standard errors, or margins of error, which are quite different things.

A **boxplot** is also nice:The plot below shows the hinge-spread (smaller vertical arrow) and a "whiskers-spread" (larger vertical arrow) of the data. Chapter 3 provides instruction on construction and interpretation of boxplots.



Confidence Limits for a Variance

95% confidence interval for the population variance. The sample variance s^2 is an unbiased estimator of σ^2. The 95% confidence intervals for ^{2} (for Normal populations) is given by:

$$\left(\frac{SS}{\chi^2_{n-1,.975}}, \frac{SS}{\chi^2_{n-1,.025}} \right) \quad \textbf{(4)}$$

where *SS* represents the sum of square (formula 1), $\chi^2_{n\text{-}1,.975}$ is the 97.5*th* percentile on a chi-square distribution with $n-1$ degrees of freedom, and $\chi^2_{n\text{-}1,.025}$ is the 2.5*th* percentile on a chi-square distribution with $n-1$ degrees of freedom.

Going from the variance to the sum of squares. If the sums of squares is not available, but the variance (or standard deviation) and sample size is, the sum of squares is

$$SS = (n-1)s^2 \quad \textbf{(5)}$$

For the illustrative data, $s^2 = 4.667$ and $n = 4$. Therefore, $SS = (4 \quad 1)(4.667) = 14.00$.

Illustrative example. Consider the small data set on page 1 of this chapter in which $SS = 14$ and $n = 4$. For 95% confidence use $\chi^2_{3,.975} = 9.3484$, $\chi^2_{3,.025} = 0.2158$. Thus, a 95% confidence interval for ^{2} is $\left(\frac{14}{9.35}, \frac{14}{0.22} \right)$ = (1.50, 64.87).

95% confidence interval for the population standard deviation. The 95% confidence limits for population standard deviation is derived by taking the square root of the confidence limits for the variance. For the illustrative example, the 95% confidence interval for $= (\sqrt{1.50}, \sqrt{64.87}) = (1.22, 8.05)$.

Testing Variances for a Significant Difference

When we have two independent samples, we might ask if the variances of the two populations differ. Consider the following **fictitious data**:

Sample 1: {3, 4, 5, 8} Sample 2: {4, 6, 7, 9}

The first sample has s^2_1= 4.667 and the second has s^2_2 = 4.333. Is it possible the observed difference reflects random variation and the variances in the population are the same? We test null hypothesis H_0: $\sigma^2_1 = \sigma^2_2$.

We ask "What is the probability of taking samples from two populations with identical variances while observing sample variances as different as s^2_1 and s^2_2? If this probability is low (say, less than .06*), we will reject H_0 and conclude the two samples came from populations with unequal variances. If this probability is *not* too low, we will say there is insufficient evidence to reject the null hypothesis.

The most common procedure for the test is the ***F* ratio test**, which has this test statistic:

$$F_{stat} = \frac{s_1^2}{s_2^2} \text{ or } \frac{s_2^2}{s_1^2}, \text{ whichever is larger.} \qquad \textbf{(6)}$$

Notice that the larger variance is placed in the numerator of the statistic, and the smaller variance is placed in the denominator. The statistic is associated with numerator and denominator degrees of freedom. The **numerator degrees of freedom** is $df_1 = n_1 - 1$, where group 1 is the group with the larger variance. The **denominator degrees of freedom** is $df_2 = n_2 - 1$. It is important to keep these degrees of freedom in the correct numerator-denominator order.

Illustrative example (fictitious data). Group 1 is {3, 4, 5, 8} and group 2 is {4, 6, 7, 9}; $s^2_1 = 4.667$ and $s^2_2 = 4.333$. We calculate $F_{stat} = s^2_1 / s^2_2 = 4.667 / 4.333 = 1.08$ with $df_1 = 4 \quad 1 = 3$ and $df_2 = 4 \quad 1 = 3$.

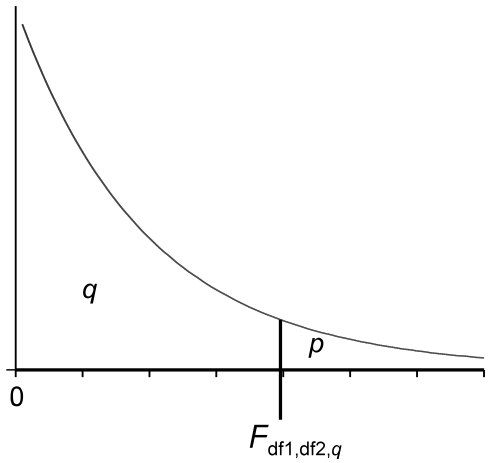
We ask whether the F statistic sufficiently far from 1.0 to reject H_0? To answer this question, we convert the F_{stat} to a p value through use of the appropriate F distribution. The test is one-tailed focusing on the upper extent of the $F_{df1, df2}$ distribution.

* Not a typo. Surely god loves $\alpha = 0.06$ as much as $\alpha = 0.05$.

The *F* Distribution

The ***F* distribution** is a family of distributions initially described by Fisher and Snedecor. Each member of the *F* distribution is identified by two parameters: df_1 (numerator degrees of freedom) and df_2 (denominator degrees of freedom). *F* distributions are positively skewed, with the extent of skewness determined by the distributions degrees of freedoms.

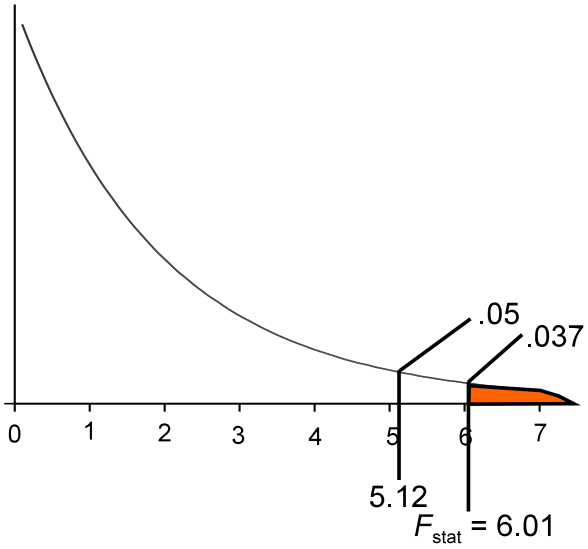
Notation: Let $F_{df1,df2,q}$ denote the q^{th} percentile an *F* distribution with df_1 and df_2 degrees of freedom. (The q^{th} percentile is greater than or equal to q% of the distribution.) As always, the area under the curve represents probability, and the total area under the curve sums to 1. The area under the curve to the right of the point $F_{df1,df2,q}$ is $p = 1-q$:



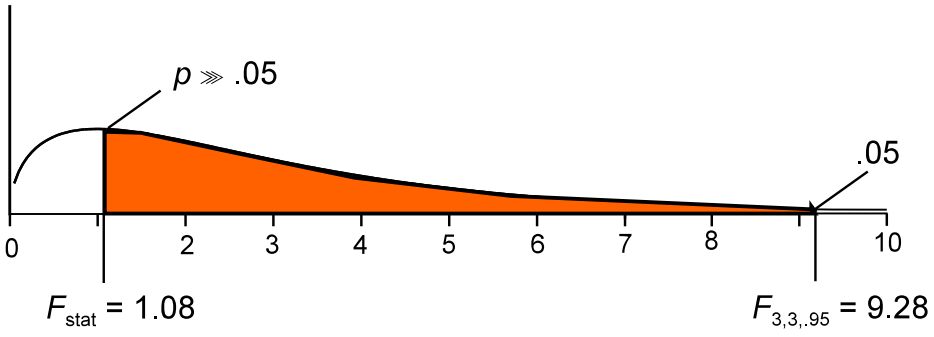
Our ***F* table** in is organized with numerator degrees of freedom (df_1) shown in the first row of the table and denominator degrees of freedom (df_2) shown in the first column. Only 95th percentile points are provided. Here's the first ten lines in the table:

		Numerator Degrees of Freedom									
	*	1	2	3	4	5	6	7	8	9	10	*
	1	161	199	216	225	230	234	237	239	241	242	1
	2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	2
D	3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	3
e	4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	4
n	5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	5
o												
m	6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	6
i	7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	7
n	8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	8
a	9	*5.12*	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	9
t	10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	10

Example: The 95th percentile on $F_{1,9,}$ is 5.12 (*italicize*). Suppose we calculate F_{stat} = 6.01 with df_1 = 1 and df_2 = 9. Since this F_{stat} is situated to the right of the 95th percentile point, we know the *P*-value is less than 0.05. More precise *P*-values can be derived with the proper statistical utility (e.g., *StaTable, WinPepi*). For example, we can use StaTable to determine that an F_{stat} of 6.01 with df_1 = 1 and df_2 = 9 is equivalent to *P* = 0.037.



Illustrative example. Recall that with our illustrative data calculated F_{stat} = 1.08 with df_1 = 4 1 = 3 and df_2 = 4 1 = 3. Using the *F* table we find that the 95th percentile on $F_{3,3}$ is 9.28. Therefore, *P* > 0.05. Using a software utility determine *P* = 0.48. Thus, H_0 is retained. [Keep in mind that a non-significant test does not provide evidence of equal variances, especially when samples are small. It merely says the evidence is insufficient to reject H_0: equal variances]



Pooling Variances

It is common in statistical practice to pool (average) group variances to come up with a more reliable estimate of variability. This should be done only if data suggest $^2_1 \approx \ ^2_2$ (e.g., with a negative F ratio test).

The **pooled estimate of variance (s^2_p)** is:

$$s_p^2 = \frac{(df_1)(s_1^2) + (df_2)(s_2^2)}{df} \quad \textbf{(7)}$$

where **$df = df_1 + df_2$.**

Illustrative example. Group 1 is {3, 4, 5, 8} and group 2 is {4, 6, 7, 9}. We've already calculated $s^2_1 = 4.667$ and $s^2_2 = 4.333$. Note $df_1 = 4 - 1 = 3$, $df_2 = 4 - 1 = 3$, and $df = 3 + 3 = 6$. Thus: $s_p^2 = \frac{(3)(4.667) + (3)(4.333)}{6} = 4.50$.

The pooled estimate of the variance can be used to derive the pooled **standard error of the mean difference**:

$$se_{\bar{x}_1 - \bar{x}_2} = \sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \quad \textbf{(8)}$$

Use this standard error to test H_0: $\mu_1 = \mu_2$ when and calculate a confidence interval for $\mu_1 - \mu_2$ when population variances are equal. The **equal variance test statistic** is:

$$t_{stat} = \frac{(\bar{x}_1 - \bar{x}_2)}{se_{\bar{x}_1 - \bar{x}_2}} \quad \textbf{(9)}$$

The **equal variance 95% confidence interval for $\mu_1 \quad \mu_2$** is:

$$(\bar{x}_1 - \bar{x}_2) \pm (t_{df,.975})(se_{\bar{x}_1 - \bar{x}_2}) \quad \textbf{(10)}$$

Illustrative example. For the illustrative data $se_{\bar{x}_1 - \bar{x}_2} = \sqrt{4.5\left(\frac{1}{4} + \frac{1}{4}\right)} = 1.5$, $\bar{x}_1 = 5.0, \bar{x}_2 = 6.5$, and $t_{stat} = \frac{5.0 - 6.5}{1.5} \quad -1.00$ with $df = 4 + 4 \quad 2 = 6$. A t_{stat} with 6 degrees of freedom ($p = .36$). The 95% confidence $\mu_1 - \mu_2$ = (5.0 − 6.5) ± (2.48)(1.5) = −1.5 ± 3.7 = (−5.2 to 2.2).

When Variances Should Not be Pooled

If EDA and *F* testing suggest variances differ, sample variance should *not* be pooled and the equal variance *t* test should be avoided. Inference can still be pursed while assuming $\sigma^2_1 \neq \sigma^2_2$ using the **Behrens-Fisher procedures** (also called the **unequal variance *t* procedure**).

The t statistic and confidence interval formulas for this procedure are the same as that for the equal variance procedures. What differs, however, is the standard error, which is:

$$se_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \qquad \textbf{(11)}$$

There are two ways to calculate the **df** for this statistic. SPSS uses this formula:

$$df' = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{(n_1 - 1)} + \frac{(s_2^2/n_2)^2}{(n_2 - 1)}}$$

When working by hand, use the smaller of df_1 or df_2.

A **95% confidence interval** for $\mu_1 - \mu_2$ without assuming equal variance is $(\bar{x}_1 - \bar{x}_2) \pm (t''_{df'',.975})(se_{\bar{x}_1 - \bar{x}_2})$.

Illustrative example (Familial blood glucose levels – hypothetical data). Blood glucose levels are determined for twenty-five ($n_1 = 25$) 5-year olds whose fathers have type II diabetes ("cases"). That cases have mean fasting blood glucose levels ($\bar{x}_1$) of 107.3 mg/dl with a standard deviation (s_1) of 9.6 mg/dl. A group of control group of 20 children ($n_2 = 20$) from the same census track whose fathers have no history of type II diabetes demonstrate $\bar{x}_2$ = 99.7 mg/dl and $s_2 = 5.2$ mg/dl. We want to test the means for inequality but suspect population variances differ significantly. In testing, H_0: $\sigma^2_1 = \sigma^2_2$, we calculate $F_{stat} = 9.6^2 / 5.2^2 = 3.41$ with $df_1 = 24$ and $df_2 = 19$ ($p = .0084$), concluding variances differ significantly. An unequal t procedure will be pursued.

In applying an unequal variance *t* test, $se_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{9.6^2}{25} + \frac{5.2^2}{20}} = 2.245$, $t_{stat} = \frac{107.3 - 99.7}{2.245} = 3.39$. The df" is a tedious to calculate. Without software, we use the smaller of $n_1 - 1$ or $n_2 - 1$, which in this case is 19. (The actual df" = 38). This will provide conservative inference for the two populations. Using a *t* distribution with df = 19, we derive $.002 < p < .01$. The 95% confidence interval for the data= (107.3 − 99.7) ± (2.09)(2.245) = 7.6 ± 4.7 = (2.9, 12.3).