11: Variances and Means

Review of variance and standard deviation

Variability measures are often based on **sum of squares**:

$$SS = (x_i - \bar{x})^2$$

The variance is the **mean sum of squares**. We rarely know population variance σ^2, so we estimate it with the **sample variance:**

$$s^2 = \frac{SS}{df}$$

where df is the degrees of freedom. For a single sample, $df = n - 1$. We lose 1 degree of freedom in estimating μ with $\bar{x}$; every time you use an estimate for a parameter to estimate something else, you lose one degree of freedom.

The **standard deviation** is the square root of the variance, or **root mean square**. The direct formula is:

$$s = \sqrt{\frac{SS}{df}}$$

I'm going to use a very small data set to demonstrate the sum of squares. Here it is: **{3, 4, 5, 8}.** This data set has $\bar{x}$ = 5, $SS = (3-5)^2 + (4-5)^2 + (5-5)^2 + (8-5)^2 = 4 + 1 + 0 + 9 =$ 14, and $df = 4 - 1$. Therefore, $s = \sqrt{\frac{SS}{df}} = \sqrt{\frac{14}{3}}$ = 2.16. The variance is just the square of the standard deviation, which in this case is $s^2 = 2.16^2 = 4.67$.

We usually report the standard deviation or variance (not both), as these are redundant.

Interpretation: (a) The variance and standard deviation are measures of spread. The bigger they are, the more variability there is in the data. (b) For Normal populations, you can use the 68-95-99.7 rule to predict ranges of values. (c) Most data are not Normal. (Sampling distribution of means tend to be Normal, but samples and populations do not.) You can use **Chebychev's rule** for non-Normal data; we can safely says that *at least* 75% of a population will lie within $\mu \pm 2\sigma$.

Illustrative data: Age of participants, center 1. Ages in years of participants at a certain center are as follows: {60, 66, 65, 55, 62, 70, 51, 72, 58, 61, 71, 41, 70, 57, 55, 63, 64, 76, 74, 54, 58, 73}. The data can be displayed graphically with the stemplot:

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

This data set has 22 observations. Data spread from 41 to 76 with a median of 62.5 (underlined). There is a low outlier, and the distribution has a mild negative skew. The sample mean is 62.545, $SS = (60-62.545)^2 + (66-62.545)^2 + \cdots + (73-62.545)^2 = 1579.45$, and $s^2 = 1579.455/(22-1) = 75.212$. The standard deviation = $\sqrt{75.212} = 8.7$. Exploratory graphs can be used to visualize the variance of the data. Figure 1 is a **dot plot** and **mean ± standard deviation plot**. **Boxplot** are nice, too (Figure 2).

{Figure 1}
{Figure 2}

Testing two variances for inequality

When we have two independent samples, we can ask if the variances of the two underlying populations differ. Consider the following **very small samples**:

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 is random and the variances in the populations are the same? We can test null hypothesis H_0: $\sigma^2_1 = \sigma^2_2$ and 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 different variances. If this probability is *not* too low, we will say there is insufficient evidence to reject the null hypothesis. We can test H_0: $\sigma^2_1 = \sigma^2_2$ with the ***F*** statistic:

$$F_{stat} = \frac{s_1^2}{s_2^2} \text{ or } \frac{s_1^2}{s_2^2}\text{, whichever is larger}$$

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

For the very small data sets above (i.e., {3, 4, 5, 8} vs. {4, 6, 7, 9}), $s^2_1 = 4.667$ and $s^2_2 = 4.333$. The test statistic $F_{stat} = s^2_1 / s^2_2 = 4.667 / 4.333 = 1.08$ with $df_1 = 4 - 1 = 3$ and df_2

= 4 − 1 = 3. Now we ask whether the observed F statistic is sufficiently far from 1.0 to reject H_0. To answer this question, we convert the F_{stat} to a P value with Fisher's F distribution.

Fisher's F Distributions

***F* distributions** are a family of distributions with each member identified by numerator (df_1) and denominator (df_2) degrees of freedom. They are positively skewed, with the extent of skewness determined by the *df*s.

Let $F_{df1,df2,q}$ denote the q^{th} percentile of an F distribution with df_1 and df_2 degrees of freedom. As usual, the area under the curve represents probability, and the total area under the curve sums to 1. Figure 3 shows an F value with cumulative probability q and right-tail region p.

{Figure 3}

Our ***F* table** (available online) lists critical value for tail regions of 0.10, 0.05, 0.025, 0.01, and 0.0001 for various combinations of df_1 and df_2. Therefore, for most problems, you will need to wedge observed F statistics between landmarks in this table to find the P value. As an example, an F_{stat} of 6.01 with $df_1 = 1$ and $df_2 = 9$ falls between *P*s of 0.05 and 0.025. A more precise P values can be derived with *StaTable* or *WinPepi*, which in this case derives $P = 0.037$.

Testing two means for inequality without assuming $\sigma^2_1 = \sigma^2_2$

Recall that in Unit 8 we tested two independent means for inequality with **Student's *t* test**. This test required us to pool variances from the two samples to come up with a pooled estimate of variance (s^2_{pooled}). We then used this variance to calculate a pooled standard error. This approach is fine for groups with (nearly) equal variances, but can be unreliable when group variances differ widely. In such instances, it is best to use a test of means that does not assume equal variances. The problem of testing group means from populations with different variances is known as the **Fisher-Behrens problem**.

There are several different procedures that can be used to test means in the face of unequal population variances. You are probably familiar with the SPSS output that says "variances not assumed to be equal." This determines inferential statistics with the **Welch procedure**, which uses this standard error of the mean difference:

$$SE_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

The degrees of freedom associated with this estimate is

$$df_{\textbf{Welch}} = \frac{\left(SE_{\bar{x}_1}^2 + SE_{\bar{x}_2}^2\right)^2}{\dfrac{SE_{\bar{x}_1}^4}{n_1 - 1} + \dfrac{SE_{\bar{x}_2}^4}{n_2 - 1}}$$

where $SE_{\bar{x}_1} = \dfrac{s_1}{\sqrt{n_1}}$ and $SE_{\bar{x}_2} = \dfrac{s_2}{\sqrt{n_2}}$.

Because calculation of df_{Welch} is tedious, we may when working by hand use the smaller of $df_1 = n_1 - 1$ or $df_2 = n_2 - 1$ as a conservative approximation for the degrees of freedom:

$$df_{\text{conserv}} = \text{the smaller of } df_1 \text{ and } df_2$$

The degrees of freedom are never less than the smaller of df_1 and df_2 (Welch, 1938, p. 356). Using df_{conserv} creates a $(1 - \alpha)100\%$ confidence interval will capture the parameter more than $(1-\alpha)100\%$ of the time.

A 95% confidence interval for $\mu_1 - \mu_2$ is calculated with the usual formula:

$$(\text{point estimate}) \pm t \cdot SE$$

where point estimate = $\bar{x}_1 - \bar{x}_2$, t = the t percentile corresponding to the desired confidence level (using df_{Welch} or df_{conserv}, whichever is handy), and SE is the standard error of the difference in means shown just above.

The t test can be performed with the test statistic

$$t_{\text{stat}} = \frac{\text{point estimate}}{SE}$$

Illustrative example (Familial blood glucose). Blood glucose is determined in twenty-five 5-year-olds whose fathers have type II diabetes ("cases") and twenty comparable controls whose fathers have no history of type II diabetes. Cases have mean fasting blood glucose of 107.3 mg/dl (standard deviation = 9.6 mg/dl). The controls have a mean of 99.7 mg/dl (standard deviation = 5.2 mg/dl).

Note that the cases have about twice the standard deviation of control. Also note that an F test of, H_0: $\sigma^2_1 = \sigma^2_2$ determines $F_{\text{stat}} = 9.6^2 / 5.2^2 = 3.41$ with $df_1 = 24$ and $df_2 = 19$ (P = 0.0084). Because variances seem to differ significantly, we apply unequal variance t procedures. Intermediate calculations for the problem are:

- Point estimate of mean difference = $\bar{x}_1 - \bar{x}_2$ = 107.3 – 99.7 = 7.6

- $SE_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} = \sqrt{\frac{9.6^2}{25}+\frac{5.2^2}{20}} = 2.245.$
- Since df_{Welch} is a tedious to calculate, we use the smaller of $n_1 - 1$ or $n_2 - 1$ as the degrees of freedom, which in this case is 19.
- For at least 95% confidence, use $t_{19,\ 0.975} = 2.09$ (from the *t* Table)

95% confidence interval for $\mu_1 - \mu_2$ = (point estimate) ± $t \cdot SE$ = 7.6 ± (2.09)(2.245) = (2.9, 12.3) mg / dl. This indicates that there is a small (but detectable) difference in the two populations.

In testing $\mu_1 - \mu_2 = 0$, $t_{\text{stat}} = \frac{\text{point estimate}}{SE} = \frac{7.6}{2.245} = 3.39$ with dfconserv = 19, P = 0.0031, confirming that the observed difference is highly unlikely to be a chance observations (so-called statistical significance).