Unit 11
Comparing Variances & Comparing Means

Including a Review of EDA
Review of Variance

(p. 11.1)

- Data set: \{3, 4, 5, 8\}
- Mean
  - \( n = 4 \)
  - \( \sum x = 3 + 4 + 5 + 8 = 20 \)
  - \( \bar{x} = 20 / 4 = 5 \)
- Variance and standard deviation
  - \( SS = (x_i - \bar{x})^2 = (3-5)^2 + (4-5)^2 + (5-5)^2 + (8-5)^2 = 14 \)
  - Variance \( (s^2) = SS / (n-1) = 14 / (4-1) = 4.667 \)
  - Standard deviation \( (s) = \sqrt{s^2} = \sqrt{4.667} \approx 2.2 \)
- Interpretation of standard deviation
  - Best when used to compare 2 groups
  - For normal distributions: 95\% of data within \( \mu \pm 2\sigma \)
  - For all distributions: \( \geq 75\% \) within \( \mu \pm 2\sigma \) (Chebychev)
Graphs
p. 11.2

- **Illustrative data:** `agebycen.sav`, group 1 (age)
  - $n = 22$
  - $\bar{x} = 62.5$
  - $s = 8.7$
- **Stem-and-leaf (“stemplot”)**
  - Draw stem first
  - Use between 4 and 12 bins

```
| 4 | 1  
| 4 |    
| 5 | 14 
| 5 |  5788
| 6 | 01234
| 6 | 56  
| 7 | 001234
| 7 | 6   
(x10)
```

*This stemplot uses split (double) stem values to draw out the shape of the distribution.*
Dot Plot and Mean±SD Plot
(p. 11.2)

Dot Plot

Mean ± Standard Deviation Bar

Age (years)

xbar + s = 62.5 + 8.7 = 71.2

xbar = 62.5

xbar - s = 62.5 - 8.7 = 53.8
5-Point Summary and Boxplot
p. 11.2 (cont.)

- Data (agebycen.sav, group 1)

<table>
<thead>
<tr>
<th>41</th>
<th>51</th>
<th>54</th>
<th>55</th>
<th>55</th>
<th>57</th>
<th>58</th>
<th>58</th>
<th>60</th>
<th>61</th>
<th>62</th>
<th>63</th>
<th>64</th>
<th>65</th>
<th>66</th>
<th>70</th>
<th>70</th>
<th>71</th>
<th>72</th>
<th>73</th>
<th>74</th>
<th>76</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q0</td>
<td>Q1</td>
<td>Q2</td>
<td>Q3</td>
<td>Q4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- 5-point summary
  - Q0 = 41
  - Q1 = 57
  - Q2 = 62.5
  - Q3 = 70
  - Q4 = 76

- IQR = Q3 − Q1 = 70 − 57 = 13

- Check for upper outside values
  - Fence_{upper} = Q3 + (1.5)(IQR) = 70 + (1.5)(13) = 89.5
  - no upper outside values
  - upper inside value = 76

- Check for lower outside values
  - Fence_{lower} = Q1 − (1.5)(IQR) = 57 − (1.5)(13) = 37.5
  - no lower outside values
  - lower inside value = 41
Boxplot

p. 11.2 (cont.)

Upper inside value = 76

Upper hinge = 70

Median = 62.5

Lower hinge = 57

Lower inside value = 41
Confidence Interval for $\sigma^2$

p. 11.3

- We won’t cover this technique but note:
  - $\sigma^2 = \text{population variance}$
    - parameter
    - unknown
    - what we want to know
  - $s^2 = \text{sample variance}$
    - estimate
    - calculated
    - What we have
  - $\sigma^2$ and $s^2$ are related but are not the same!
Illustrative data: independent samples
- $s^2_1 = 4.667, n_1 = 4$
- $s^2_2 = 4.333, n_2 = 4$
- Do samples come from populations with equal variance?

Test procedure (F ratio test)
- $H_0: \sigma^2_1 = \sigma^2_2$
- Flexible significance testing requires no preset $\alpha$
- $F_{\text{stat}} = s^2_1 / s^2_2$ or $s^2_2 / s^2_1$, whichever larger
- Convert $F_{\text{stat}}$ to $p$ value (next page)
  - Reject $H_0$ when $p$ is “small”
  - $p < .1 \rightarrow$ some evidence against $H_0$
  - $p < .05 \rightarrow$ stronger evidence against $H_0$
  - $p < .01 \rightarrow$ still stronger evidence against $H_0$
$F$ distributions

p. 11.5

- Used to determine probability of observing $F_{stat}$ or greater
- $df_1 =$ numerator degrees of freedom
- $df_2 =$ denominator degrees of freedom
- Notation $F_{df_1, df_2, q} = F$ percentile with $df_1$, $df_2$, and cumulative probability of $q$
Table provides 95th percentile landmarks (p. 11.5)

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

\[ F_{1,9,.95} = 5.12 \]
Suppose $F_{\text{stat}} = 6.01$ with $df_1 = 1$ and $df_2 = 9$

(p. 11.5)

- $p$ value = shaded area in right tail
  - Method 1: use landmark from F table: $p < .05$
  - Method 2: use computer (e.g., StaTable): $p = .037$
Suppose $F_{\text{stat}} = 1.08$ with $df_1 = 3$ and $df_2 = 3$

(p. 11.6)
Pooling Variances

(p. 11.7)

- If $\sigma^2_1 \approx \sigma^2_2$ (equal variance, homoscedasticity) → combine (“pool”) sample variances to estimate $\sigma^2$

- If $\sigma^2_1 \neq \sigma^2_2$ (unequal variance, heteroscedasticity) → do not pool sample variances

- How do you determine if $\sigma^2_1 \approx \sigma^2_2$?
  - Compare $s_1$ to $s_2$
  - EDA (e.g., side-by-side boxplots)
  - $F$ ratio test
Example of Homoscedasticity

(not in Reader)

<table>
<thead>
<tr>
<th></th>
<th>Group 1</th>
<th>Group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

\( s_1 = 15.8 \)

\( s_2 = 15.8 \)

(x10)
Example of Heteroscedasticity

(Not in Reader)

Group 1          Group 2
0 | 9 | 0
0 | 8 | 0
0 | 7 | 0
0 | 6 | 0
0 | 5 | 0
4 |
3 |
2 |
1 |
\(x10\)

\(s_1 = 15.8\)

\(s_2 = 31.6\)
Compare Means
(equal variance assumed)

(p. 11.7)

- $s^2_p = \text{formula 7}$
  - But what is it?
  - ANS: weighted average of variances from the two samples

- $se_{\text{mean dif}} = \text{formula 8}$
  - But what is it?
  - ANS: measure of precision of estimated mean dif

- $t_{\text{stat}} = \text{formula 9}$
  - But what question does it answer?
  - ANS: Can we say with confidence that populations means differ?
Comparing Means
(equal variance assumed)

p. 11.7

- Simple illustrative example
  - xbar\(_1\) = 5, s\(_1\)^2 = 4.667, n\(_1\) = 4
  - xbar\(_2\) = 6.5, s\(_2\)^2 = 4.333, n\(_2\) = 4

- s\(_p\)^2 = [(3)(4.667) + (3)(4.333)] / 6 = 4.5

- se_{\text{mean dif}} = \sqrt{4.5 \times (1/4 + 1/4)} = 1.5

- H\(_0\): \mu_1 = \mu_2
  - t_{\text{stat}} = (5 - 6.5) / 1.5 = -1.00
  - df = (4 - 1) + (4 - 1) = 6
  - p value = .356 (see next slide)
  - Conclude: Difference is not significant.
Converting $t_{\text{stat}}$ to $p$ value

- $t_{\text{stat}} = -1.00$ with 6 df
- By hand ($t$ table)
  - landmarks are $t_{6,.90} = 1.44$ and $t_{6,.80} = 0.91$
  - thus, $0.10 < \frac{1}{2}p < 0.20$ and $0.20 < p < 0.40$
- By computer: $p = 0.356$
Comparing Means
(equal variance *not* assumed)

Do not pool variances
 Behrens-Fisher unequal variance t test statistic (formula 11)
 Standard Error (formula 12)
 if you do not have a computer that does the test for you use lesser of df\(_1\) or df\(_2\)
 Illustrative example done on board