Statistics: Used and Abused Tools of the Trade

Scott Staggenborg
Kansas State University
Statistics

• Tools for making decisions
  – “Is the untreated different than the treated?”
  – “What is the optimum fertilizer rate or seeding rate?”
  – “How much does delayed planting affect corn yields?”

We want to do this with some confidence about our final decision.

We also want to make informed decisions
“If you cannot measure it, you cannot manage it”
Data and Distributions

The basis for statistical decisions

- **Yield data**
  - **A population of** 11,196 points
  - Mean = 123.8 bu/a
  - Min = 2.1 bu/a
  - Max = 276.7 bu/a
  - Standard Deviation = 35.1
Data and Distributions
The basis for statistical decisions

- Yield data
  - 11,196 points
  - Mean = 123.8 bu/a
  - Min = 2.1 bu/a
  - Max = 276.7 bu/a
  - Standard Deviation = 35.1
Confidence or Risk in Statistics
The basis for statistical decisions

Mean is often best forecast

<table>
<thead>
<tr>
<th></th>
<th>farm A</th>
<th>farm B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year 1</td>
<td>$80</td>
<td>$5</td>
</tr>
<tr>
<td>Year 2</td>
<td>$200</td>
<td>$30</td>
</tr>
<tr>
<td>Year 3</td>
<td>$50</td>
<td>$20</td>
</tr>
<tr>
<td>Year 4</td>
<td>$270</td>
<td>$25</td>
</tr>
<tr>
<td>Year 5</td>
<td>$300</td>
<td>$30</td>
</tr>
<tr>
<td>Average</td>
<td>$20</td>
<td>$20</td>
</tr>
</tbody>
</table>

Five year average return per acre

How much confidence do you have in the $20 estimate?
When analyzing data

• The mean is a powerful measure/concept
• However, the mean does not convey all important and relevant information.
• We often also want to consider the variability in the data.
Measures of variability

• **Range** -- the difference between the largest reading and the smallest reading.

• **Standard deviation** -- a measurement of the total variability of the data. It is an average of deviations from the mean.

\[
s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}}
\]

• **Coefficient of variation (CV)** -- normalized measure of variability equal to standard deviation divided by the mean.
Standard deviation

• Same unit of measure as the original data

• Affected by extreme values, which tend to enlarge the standard deviation.

• Larger values for standard deviation indicate the data are more widely dispersed around the mean (i.e., more variable).

• Generally has the most meaning when data are normally distributed.
Probabilities and Statistics

- Definitely more than others
- Definitely less than others
- Probably less than others
- Same as others
- Probably more than others

SD = Standard Deviation
Statistical Tests Often Used

• Mean Comparison
  – We have two or more products and we want to know if the response of one is different than the other

• Trend Analysis
  – We have a range of treatments of the same product and we want to know the optimum

• Spatial Analysis
  – Not going to talk about today
## Mean Comparisons

### Analysis of Variance

<table>
<thead>
<tr>
<th>BRAND</th>
<th>NAME</th>
<th>YIELD (bu/a)</th>
<th>PAVG (%)</th>
<th>TW (lb/bu)</th>
<th>MOIST (%)</th>
<th>DAYS (silk)</th>
<th>LDG (%)</th>
<th>100pp</th>
</tr>
</thead>
<tbody>
<tr>
<td>RENZE</td>
<td>1357YGPL/RR</td>
<td>188</td>
<td>99</td>
<td>57</td>
<td>17</td>
<td>64</td>
<td>0</td>
<td>24.9</td>
</tr>
<tr>
<td>RENZE</td>
<td>5347HX1/LL</td>
<td>192</td>
<td>101</td>
<td>56</td>
<td>18</td>
<td>66</td>
<td>2</td>
<td>25.7</td>
</tr>
<tr>
<td>RENZE</td>
<td>8386YGCB</td>
<td>186</td>
<td>98</td>
<td>55</td>
<td>19</td>
<td>67</td>
<td>3</td>
<td>24.4</td>
</tr>
<tr>
<td>RENZE</td>
<td>8428YGCB</td>
<td>177</td>
<td>93</td>
<td>56</td>
<td>18</td>
<td>65</td>
<td>1</td>
<td>25.7</td>
</tr>
<tr>
<td>RENZE</td>
<td>9328YGCB/RR</td>
<td><strong>195</strong></td>
<td>102</td>
<td>57</td>
<td>19</td>
<td>65</td>
<td>2</td>
<td>24.4</td>
</tr>
<tr>
<td>RENZE</td>
<td>9386YGCB/RR</td>
<td>178</td>
<td>93</td>
<td>57</td>
<td>18</td>
<td>65</td>
<td>0</td>
<td>25.7</td>
</tr>
<tr>
<td>TAYLOR</td>
<td>77640 RR</td>
<td>185</td>
<td>97</td>
<td>58</td>
<td>16</td>
<td>64</td>
<td>1</td>
<td>24.4</td>
</tr>
<tr>
<td>TAYLOR</td>
<td>930 RR/Bt</td>
<td>181</td>
<td>95</td>
<td>56</td>
<td>16</td>
<td>63</td>
<td>2</td>
<td>24.4</td>
</tr>
<tr>
<td>TRIUMPH</td>
<td>1608VT3</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>TRIUMPH</td>
<td>1866Bt</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>TRIUMPH</td>
<td>1977CbRR</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td></td>
<td>AVERAGE</td>
<td><strong>190</strong></td>
<td><strong>190</strong></td>
<td>56</td>
<td>18</td>
<td><strong>65</strong></td>
<td>2</td>
<td>25.7</td>
</tr>
<tr>
<td></td>
<td>CV (%)</td>
<td>9</td>
<td>9</td>
<td>2</td>
<td>9</td>
<td>2</td>
<td>--</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>LSD (.05)</td>
<td><strong>24</strong></td>
<td>14</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>2.2</td>
</tr>
</tbody>
</table>

* Seed treatments and hybrid traits located in Table 16.
** Yields in bold are in the top LSD group.
*** Unless two hybrids differ by more than the LSD, little confidence can be placed in one being superior.
Mean Comparisons
Analysis of Variance

<table>
<thead>
<tr>
<th>Replication</th>
<th>Hyb A</th>
<th>Hyb B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>159</td>
<td>158</td>
</tr>
<tr>
<td>2</td>
<td>161</td>
<td>155</td>
</tr>
<tr>
<td>3</td>
<td>159</td>
<td>156</td>
</tr>
<tr>
<td>4</td>
<td>165</td>
<td>153</td>
</tr>
<tr>
<td>Average</td>
<td>161</td>
<td>156</td>
</tr>
<tr>
<td>St. Dev</td>
<td>2.8</td>
<td>2.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>Df</th>
<th>SS</th>
<th>F</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rep</td>
<td>3</td>
<td>2.37</td>
<td>2.37/34.38</td>
<td>0.97</td>
</tr>
<tr>
<td>Hybrid</td>
<td>1</td>
<td>136.12</td>
<td>136.12/34.38</td>
<td>0.04</td>
</tr>
<tr>
<td>Error</td>
<td>3</td>
<td>34.38</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

“Probability of F”
- Where the calculated F-Statistic falls in the distribution. F is a normal distribution and can be used to estimate probabilities.
- 1-(Pr>F) could be thought of as the “Confidence”
LSDs and Alphas

• We see LSD\(_{(0.05)}\) and “significant at the alpha = 0.10” or “not statistically significant” in research reports.

• What does it mean?

• “Significant “ in statistics means “at the level of risk we are willing to accept that the evidence (sampled data) is sufficient to accept or discredit our H\(_0\)”
Analysis of Variance

Between Group Variation

Within Group Variation

Total Variation
Analysis of Variance

Assertion of null hypothesis

H₀: All samples drawn from the same population
(μ₁ = μ₂ = μ₃)

Assertion of alternate hypothesis

Hₐ: At least one sample drawn from a different population
(μ₁ ≠ μ₂ ≠ μ₃)

Case 1:
Small apparent difference between sample means
Likely decision: do not reject H₀

Case 2:
Large apparent difference between sample means
Likely decision: reject H₀

FIGURE 10.1
Null and Alternate Hypotheses in Analysis of Variance (ANOVA)

## Mean Comparisons

### Analysis of Variance

<table>
<thead>
<tr>
<th>Replication</th>
<th>Hyb A</th>
<th>Hyb B</th>
<th>Hyb A</th>
<th>Hyb B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>159</td>
<td>158</td>
<td>155</td>
<td>145</td>
</tr>
<tr>
<td>2</td>
<td>161</td>
<td>155</td>
<td>160</td>
<td>150</td>
</tr>
<tr>
<td>3</td>
<td>159</td>
<td>156</td>
<td>162</td>
<td>160</td>
</tr>
<tr>
<td>4</td>
<td>165</td>
<td>153</td>
<td>168</td>
<td>169</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td>161</td>
<td>156</td>
<td>161</td>
<td>156</td>
</tr>
<tr>
<td><strong>St. Dev</strong></td>
<td>2.8</td>
<td>2.1</td>
<td>5.4</td>
<td>10.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>Prob &gt; F</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rep</td>
<td>0.97</td>
<td>0.11</td>
</tr>
<tr>
<td>Hybrid</td>
<td>0.04</td>
<td>0.21</td>
</tr>
</tbody>
</table>
Analysis of Variance

“But it is 6 bu/a different”
Risk and Research

• Academics and researchers have been conditioned to be low risk takers because our results turn into recommendations in real life.

• We want to be VERY certain that narrow rows are better than wide rows before a farmer spends big $$$ to switch.

• However, in real life, 95% confidence (5% Pr > F) is not always likely necessary.
Probabilities and Decisions?

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Cost ($/acre)</th>
<th>Benefit</th>
<th>Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soybean Inoculant</td>
<td>$ 1.00</td>
<td>2 bushel</td>
<td>51%</td>
</tr>
</tbody>
</table>

Low Cost and high return…do not need much confidence.

Did not find a big advantage to BT in two years tested, but infestation the next year could pay big.

New unproven product that costs $10.00/acre has a lot of risk.

Little risk here because you are going to plant something.
# Analysis of Variance Examples

## Plant Population (32,000 and 40,000 plants/acre)

<table>
<thead>
<tr>
<th></th>
<th>LY1</th>
<th>LY2</th>
<th>LY3</th>
<th>LY4</th>
<th>LY5</th>
<th>LY6</th>
<th>LY7</th>
<th>LY8</th>
<th>LY9</th>
<th>LY10</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>245</td>
<td>187</td>
<td>223</td>
<td></td>
<td>207</td>
<td>166</td>
<td>132</td>
<td>199</td>
<td>109</td>
<td>210</td>
<td>186</td>
</tr>
<tr>
<td>High</td>
<td>237</td>
<td>172</td>
<td>226</td>
<td></td>
<td>196</td>
<td>175</td>
<td>127</td>
<td>191</td>
<td>95</td>
<td>209</td>
<td>181</td>
</tr>
</tbody>
</table>

Bold numbers indicate years when means were different at the 10% level

## Fertilizer

<table>
<thead>
<tr>
<th></th>
<th>LY1</th>
<th>LY2</th>
<th>LY3</th>
<th>LY4</th>
<th>LY5</th>
<th>LY6</th>
<th>LY7</th>
<th>LY8</th>
<th>LY9</th>
<th>LY10</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>230</td>
<td>176</td>
<td>221</td>
<td>230</td>
<td>195</td>
<td>165</td>
<td>122</td>
<td>195</td>
<td>98</td>
<td>211</td>
<td>184</td>
</tr>
<tr>
<td>High</td>
<td>252</td>
<td>183</td>
<td>228</td>
<td>248</td>
<td>208</td>
<td>175</td>
<td>137</td>
<td>196</td>
<td>105</td>
<td>208</td>
<td>194</td>
</tr>
</tbody>
</table>

Bold numbers indicate years when means were different at the 10% level

---

E. Nafziger: “Managing continuous corn for high yields” white paper
Comparing Treatments

• Often we get data from a large number of different studies and want to use them to give us a better picture of the situation.
• How do we compare all of these results?
• Simple comparisons are often very useful.
Soybean Row Spacing Example

• When narrow row soybeans were being studied, a lot of results were being generated by universities.
• It seemed that some environments worked well, others did not?
• A difference plot can often be useful in determining environmental impacts.
Soybean Row Spacing Example

- Accumulate data
- Calculate the differences between the treatments
- Plot the data
  - Difference as the Y
  - Location average as the X.

Excel Table:

<table>
<thead>
<tr>
<th>Location</th>
<th>Year</th>
<th>7.5 in</th>
<th>30 in</th>
<th>Difference</th>
<th>Loc Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manhattan</td>
<td>1997</td>
<td>10.6</td>
<td>13.2</td>
<td>-2.6</td>
<td>11.9</td>
</tr>
<tr>
<td>Ottawa</td>
<td>1998</td>
<td>19.4</td>
<td>19.7</td>
<td>0.3</td>
<td>19.6</td>
</tr>
<tr>
<td>Lincoln</td>
<td>1998</td>
<td>25.5</td>
<td>24.3</td>
<td>1.2</td>
<td>24.9</td>
</tr>
<tr>
<td>SW IA</td>
<td>1999</td>
<td>31.8</td>
<td>29.8</td>
<td>2.0</td>
<td>30.8</td>
</tr>
<tr>
<td>Lincoln</td>
<td>1999</td>
<td>32.4</td>
<td>29.8</td>
<td>2.6</td>
<td>31.1</td>
</tr>
<tr>
<td>Mead</td>
<td>2000</td>
<td>34.0</td>
<td>31.5</td>
<td>2.5</td>
<td>32.8</td>
</tr>
<tr>
<td>Manhattan</td>
<td>2000</td>
<td>34.4</td>
<td>36.3</td>
<td>-1.9</td>
<td>35.4</td>
</tr>
<tr>
<td>Columbia</td>
<td>2000</td>
<td>38.1</td>
<td>37.2</td>
<td>0.9</td>
<td>37.7</td>
</tr>
<tr>
<td>MO</td>
<td>2000</td>
<td>40.9</td>
<td>37.8</td>
<td>3.1</td>
<td>39.4</td>
</tr>
<tr>
<td>VA</td>
<td>2001</td>
<td>45.0</td>
<td>40.8</td>
<td>4.2</td>
<td>42.9</td>
</tr>
<tr>
<td>Ark</td>
<td>2001</td>
<td>44.6</td>
<td>43.3</td>
<td>1.3</td>
<td>44.0</td>
</tr>
<tr>
<td>Manhattan</td>
<td>2001</td>
<td>48.0</td>
<td>44.8</td>
<td>3.2</td>
<td>46.4</td>
</tr>
<tr>
<td>Ottawa</td>
<td>2001</td>
<td>51.0</td>
<td>45.8</td>
<td>5.2</td>
<td>48.4</td>
</tr>
</tbody>
</table>
Soybean Row Spacing Example

Average Site Yield (bu/acre)

Row Spacing Yield Difference (bu/acre)

7.5 - 30 in
Overall yield response = 3.3 bu/acre
Positive yield response = 77%
Economically beneficial = 27%

- Fungicide X hybrid (replicated)
- Fungicide (replicated)
- Fungicide (side-by-side)
Foliar Fungicides – Corn

Northeast Rotated Corn

Corn Yield (bu/a)

- Manhattan
- Rossville
- Scandia
- Meadowlark

UTC
HL6
HL9
HL6-NP
HL4.5-C
Corn Fungicides in Kansas

Fungicide costs = $25/acre

- $3/bu
- $4/bu
- $5/bu

Treatment – Untreated Control (bu/a)

Untreated Control (bu/a)

y = 0.161x - 32.336
R² = 0.324
Regression Analysis

• “Line fitting” Analysis
• Is the most appropriate analysis method for rate related data such as planting date, planting rate, fertilizer rates.
• Easy to conduct
• Requires less knowledge of “statistics” as we are often looking for optimums.
# OverWorked and Mis-Used LSDs

<table>
<thead>
<tr>
<th>N Rate (lbs/acre)</th>
<th>Corn Yield (bu/acre)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>99 c</td>
</tr>
<tr>
<td>40</td>
<td>151 b</td>
</tr>
<tr>
<td>80</td>
<td>192 a</td>
</tr>
<tr>
<td>120</td>
<td>204 a</td>
</tr>
<tr>
<td>160</td>
<td>207 a</td>
</tr>
</tbody>
</table>

Based on these data, most people would assume that 80 lbs/acre is the optimum amount.
Corn Yield and N

\[ y = 0.656x + 115.68 \]

\[ R^2 = 0.6472 \]
$y = -0.0059x^2 + 1.6236x + 94.291$

$R^2 = 0.7664$

Calculated optimum N rate = 138 lbs N/acre
Corn Yield and N

Calculated optimum N rate = 100 lbs N/acre
$y = -0.0733x^2 + 0.3329x + 0.601$

$R^2 = 0.7987$

Calculated optimum 2.25 in
Calculated Plateau = 2.00 in
Wheat - Date of Planting

Grain Yield (% of Maximum)

Days Relative to Optimum Planting Date

\[ y = 100 + 0.5x \]

\[ y = 100 - 2.9x \]

- 1995-96
- 1997-98
- Trendline
Data Sources

• Universities
  – Our job is to collect data, report it, AND give our opinion on what it means. Usually pretty conservative.

• Seed and Chemical Companies
  – Pioneer, Monsanto, and Syngenta (to name a few) have or are adding Crop Management personnel to collect production data. Ask for it.

• Collect your own…it is not difficult
Variable Rate Technology
<table>
<thead>
<tr>
<th>Rep 1</th>
<th>Rep 2</th>
<th>Rep 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hybrid 2</td>
<td>Hybrid 3</td>
<td>Hybrid 1</td>
</tr>
<tr>
<td>Hybrid 3</td>
<td>Hybrid 1</td>
<td>Hybrid 2</td>
</tr>
<tr>
<td>Hybrid 1</td>
<td>Hybrid 2</td>
<td>Hybrid 3</td>
</tr>
</tbody>
</table>

Guidance Bars = Replication
Important Points

• Replication
  – Necessary for analysis and required to have confidence in results.
  – Often does not require a great deal of extra time if planned correctly.

• Yield Monitors and Yields
  – Use yield monitor to replace weigh wagon (measure grain mass).
  – Measure plot width and length manually.
  – Calculate yields and adjust for moisture as you normally would.
  – Caution with unfiltered yield monitor calculated yields!! Errors could exist in plot length and individual yield point estimates.
Summary

• Statistics are a tool that help you make informed decisions.
  – You must decide on the “risk” you are willing to take

• The key is to make sure that you are using real data to make decisions. “Plant health” does not increase price or decrease costs.
  – If you cannot measure it, you cannot manage it”

• Analysis of Variance or Mean Separations work for treatments that have yes/no decisions
  – treated vs untreated; Hyb A vs Hyb B
Summary

• Regression or trend analysis is what you want to evaluate rate or response data
  – yield response to fertilizer or to plant population.

• Get as much data on a subject as you can prior to making a decision.
  – Informed vs uninformed decisions

• Do not be afraid to use statistics and if needed, ask for help. There are a lot of people who can help you