Robust Estimation Of Treatment Effect

Wim Krijnen
Lector Analyse Technieken voor Praktijk Onderzoek
Lectoraat Transparante Zorgverlening
Academie voor Gezondheisstudies
Eyssoniusplein 18
9714 CE Groningen

January 21, 2015
Outline

1. Robust Comparison of Two Independent Groups
   - Energy Expenditure Lean and Obese Women
   - Normality Tests
   - Descriptive Statistics
   - Two-sample t-test Without Assuming Equal Variance
   - Exact Wilcoxon Mann-Whitney Rank Sum Test
   - Medians Test
   - Robust Estimation of Effect by Linear Model
   - Overview of Results on two groups

2. Robust Comparison of More Than Two Independent Groups
   - Age of Walking of Children
   - Normality Tests
   - Descriptive Statistics
   - Kruskal-Wallis Rank Sum Test
   - Robust Estimation of Linear model
Energy Expenditure Lean and Obese Women

Experiment:

- Energy expenditure (MJ) measured during 24 hour
- 13 lean and 9 obese woman
- See box-and-whiskers-plot
P-values from various Normality tests

Initial conclusion from boxplot:
1. Lean has outliers, Obese not
2. Lean seems symmetrically distributed, Obese not

What do normality tests say?

<table>
<thead>
<tr>
<th>Test</th>
<th>Lean</th>
<th>Obese</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shapiro-Wilk</td>
<td>0.0482</td>
<td>0.1426</td>
</tr>
<tr>
<td>Shapiro-Francia</td>
<td>0.0292</td>
<td>0.1549</td>
</tr>
<tr>
<td>Anderson-Darling</td>
<td>0.0186</td>
<td>0.1089</td>
</tr>
<tr>
<td>Cramer-von Mises</td>
<td>0.0139</td>
<td>0.0820</td>
</tr>
<tr>
<td>Lilliefors</td>
<td>0.0205</td>
<td>0.0820</td>
</tr>
</tbody>
</table>

Remark: Lilliefors is based on Kolmogorov-Smirnov

Conclusion: Normality violated by lean, not by Obese

What to do?
Let's look at descriptive statistics first.
P-values from various Normality tests

Initial conclusion from boxplot:
1. Lean has outliers, Obese not
2. Lean seems symmetrically distributed, Obese not

What do normality tests say?

<table>
<thead>
<tr>
<th>Test</th>
<th>Lean</th>
<th>Obese</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shapiro-Wilk</td>
<td>0.0482</td>
<td>0.1426</td>
</tr>
<tr>
<td>Shapiro-Francia</td>
<td>0.0292</td>
<td>0.1549</td>
</tr>
<tr>
<td>Anderson-Darling</td>
<td>0.0186</td>
<td>0.1089</td>
</tr>
<tr>
<td>Cramer-von Mises</td>
<td>0.0139</td>
<td>0.0820</td>
</tr>
<tr>
<td>Lilliefors</td>
<td>0.0205</td>
<td>0.0820</td>
</tr>
</tbody>
</table>

Remark: Lilliefors is based on Kolmogorov-Smirnov

Conclusion: Normality violated by lean, not by Obese

What to do?
Let's look at descriptive statistics first.
P-values from various Normality tests

Initial conclusion from boxplot:
1. Lean has outliers, Obese not
2. Lean seems symmetrically distributed, Obese not

What do normality tests say?

<table>
<thead>
<tr>
<th>Test</th>
<th>Lean</th>
<th>Obese</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shapiro-Wilk</td>
<td>0.0482</td>
<td>0.1426</td>
</tr>
<tr>
<td>Shapiro-Francia</td>
<td>0.0292</td>
<td>0.1549</td>
</tr>
<tr>
<td>Anderson-Darling</td>
<td>0.0186</td>
<td>0.1089</td>
</tr>
<tr>
<td>Cramer-von Mises</td>
<td>0.0139</td>
<td>0.0820</td>
</tr>
<tr>
<td>Lilliefors</td>
<td>0.0205</td>
<td>0.0820</td>
</tr>
</tbody>
</table>

Remark: Lilliefors is based on Kolmogorov-Smirnov

1. Conclusion: Normality violated by lean, not by Obese
2. What to do?
Initial conclusion from boxplot:

1. Lean has outliers, Obese not
2. Lean seems symmetrically distributed, Obese not

What do normality tests say?

<table>
<thead>
<tr>
<th>Test</th>
<th>Lean</th>
<th>Obese</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shapiro-Wilk</td>
<td>0.0482</td>
<td>0.1426</td>
</tr>
<tr>
<td>Shapiro-Francia</td>
<td>0.0292</td>
<td>0.1549</td>
</tr>
<tr>
<td>Anderson-Darling</td>
<td>0.0186</td>
<td>0.1089</td>
</tr>
<tr>
<td>Cramer-von Mises</td>
<td>0.0139</td>
<td>0.0820</td>
</tr>
<tr>
<td>Lilliefors</td>
<td>0.0205</td>
<td>0.0820</td>
</tr>
</tbody>
</table>

Remark: Lilliefors is based on Kolmogorov-Smirnov

1. Conclusion: Normality violated by lean, not by Obese
Descriptive Statistics Energy Expenditure Lean and Obese Women

<table>
<thead>
<tr>
<th></th>
<th>lean</th>
<th>obese</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>8.07</td>
<td>10.30</td>
</tr>
<tr>
<td>median</td>
<td>7.90</td>
<td>9.69</td>
</tr>
<tr>
<td>Huber</td>
<td>7.86</td>
<td>9.99</td>
</tr>
<tr>
<td>SD</td>
<td>1.24</td>
<td>1.40</td>
</tr>
<tr>
<td>MAD</td>
<td>0.62</td>
<td>0.74</td>
</tr>
<tr>
<td>IQR</td>
<td>0.47</td>
<td>1.70</td>
</tr>
</tbody>
</table>

Mean Absolute Deviation (MAD); Inter Quartile Range (IQR) both adapted for normal distribution

- small within group differences in mean, median, Huber
- within groups difference on SD, MAD, IQR considerable
For completeness: Definition of Huber mean

\[
\min \rho(X_1, \ldots, X_k, \mu) ; \rho(x) = \begin{cases} 
  x^2 & \text{if } |x| \leq k \\
  2k|x| - k^2 & \text{if } |x| > k 
\end{cases}
\]

where \( k = 1.345 \)

Two-sample t-test Without Assuming Equal Variance

```r
> t.test(expend ~ stature, var.equal = FALSE, data=energy)

Welch Two Sample t-test
data:  expend by stature
t = -3.8555, df = 15.919, p-value = 0.001411
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:  
  -3.459167  -1.004081
sample estimates:  
  mean in group lean mean in group obese
```

- t-test robust against mild violation from normality
- t-test optimal under assumption of normality
- null hypothesis of no effect rejected
- effect size -2.23 with 95% CI (-3.45; -1.00)
- Effect size confirmed by tests robust against violations of normality?
Exact Wilcoxon Mann-Whitney Rank Sum Test

```r
> wilcox_test(expend ~ stature, data=energy, 
+       distribution = "exact", conf.int = TRUE)

  Exact Wilcoxon Mann-Whitney Rank Sum Test

data:  expend by stature (lean, obese)
Z = -3.1061, p-value = 0.001039
alternative hypothesis: true mu is not equal to 0
95 percent confidence interval:
  -3.56  -1.26
sample estimates: difference in location
  -1.91
```

- Assumption of independent continuous measurements!
- SPPS does not give effect size or CI
- null hypothesis of no effect rejected
- effect size -1.91 with 95% CI (-3.56; -1.26)
> median_test(expend ~ stature, data=energy,
+    distribution = "exact", conf.int = TRUE)

Exact Median Test
data:  expend by stature (lean, obese)
Z = 3.8129, p-value = 0.0002211
alternative hypothesis: true mu is not equal to 0
95 percent confidence interval:
  -5.31  -1.08
sample estimates: difference in location
  -1.825

- Assumption independent continuous measurements!
- SPPS does not give effect size or CI
- null hypothesis of no effect rejected
- effect size -1.825 with 95% CI (-5.31; -1.08)
Robust estimation of linear model

```r
> library(robustbase)
> mod1 <- lmrob(expend ~ stature, data=energy) # MM method
> summary(mod1)

Coefficients:

|                      | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------------|----------|------------|---------|----------|
| (Intercept)           | 7.7746   | 0.2481     | 31.334  | < 2e-16  *** |
| statureobese          | 2.1087   | 0.6600     | 3.195   | 0.00455 **  |
```

- Assumption independent continuous measurements
- Extreme outliers are down weighted for maximum likelihood type of estimation (widely accepted)
- null hypothesis of no effect rejected
- effect size -2.10 with 95% CI (-3.4023; -0.8151)
### Overview of Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Effect</th>
<th>CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Independent t-test</td>
<td>-2.23</td>
<td>(-3.45; -1.00)</td>
</tr>
<tr>
<td>Median</td>
<td>-1.83</td>
<td>(-5.31; -1.08)</td>
</tr>
<tr>
<td>Wilcoxon</td>
<td>-1.91</td>
<td>(-3.56; -1.26)</td>
</tr>
<tr>
<td>Robust Regression</td>
<td>-2.10</td>
<td>(-3.40; -0.82)</td>
</tr>
</tbody>
</table>

- Identical conclusions existence of effect
- Effect estimates relative close
- CI differ moderately
Outline

1 Robust Comparison of Two Independent Groups
   - Energy Expenditure Lean and Obese Women
   - Normality Tests
   - Descriptive Statistics
   - Two-sample t-test Without Assuming Equal Variance
   - Exact Wilcoxon Mann-Whitney Rank Sum Test
   - Medians Test
   - Robust Estimation of Effect by Linear Model
   - Overview of Results on two groups

2 Robust Comparison of More Than Two Independent Groups
   - Age of Walking of Children
   - Normality Tests
   - Descriptive Statistics
   - Kruskal-Wallis Rank Sum Test
   - Robust Estimation of Linear model
Age of Walking of children

- Age of walking in months
- Conditions: control (5), none (6), passive (6), zactive (6)
- Observe from box-and-wiskers-plot below that there are drastic outliers (deviation from normality)

P-values from various Normality tests

<table>
<thead>
<tr>
<th></th>
<th>control</th>
<th>none</th>
<th>passive</th>
<th>zactive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shapiro-Wilk</td>
<td>0.1374</td>
<td>0.2939</td>
<td>0.0408</td>
<td>0.0087</td>
</tr>
<tr>
<td>Shapiro-Francia</td>
<td>0.1909</td>
<td>0.2116</td>
<td>0.0330</td>
<td>0.0084</td>
</tr>
<tr>
<td>Anderson-Darling</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cramer-von Mises</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lilliefors</td>
<td>0.4353</td>
<td>0.1520</td>
<td>0.2560</td>
<td>0.0110</td>
</tr>
</tbody>
</table>

- Some normality tests require at least 7 data points
- Partial non-normality for passive
- Clear non-normality for active
## Descriptive Statistics Age of Walking of children

<table>
<thead>
<tr>
<th></th>
<th>control</th>
<th>none</th>
<th>passive</th>
<th>zactive</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>mean</strong></td>
<td>12.35</td>
<td>11.71</td>
<td>11.38</td>
<td>10.12</td>
</tr>
<tr>
<td><strong>median</strong></td>
<td>12.00</td>
<td>11.75</td>
<td>10.75</td>
<td>9.62</td>
</tr>
<tr>
<td><strong>Huber</strong></td>
<td>12.34</td>
<td>11.92</td>
<td>10.98</td>
<td>9.69</td>
</tr>
<tr>
<td><strong>SD</strong></td>
<td>0.96</td>
<td>1.52</td>
<td>1.90</td>
<td>1.45</td>
</tr>
<tr>
<td><strong>MAD</strong></td>
<td>0.74</td>
<td>1.11</td>
<td>1.11</td>
<td>0.37</td>
</tr>
<tr>
<td><strong>IQR</strong></td>
<td>1.30</td>
<td>0.93</td>
<td>1.07</td>
<td>0.32</td>
</tr>
</tbody>
</table>

Remark: MAD and IQR adapted for normal distribution

- small within group differences in mean, median, Huber mean
- considerable within group differences in SD, MAD, IQR
> anova(mod1)
Analysis of Variance Table

Response: x

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>fac</td>
<td>3</td>
<td>14.778</td>
<td>4.9259</td>
<td>2.1422</td>
<td>0.1285</td>
</tr>
<tr>
<td>Residuals</td>
<td>19</td>
<td>43.690</td>
<td>2.2995</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- No significant difference in Mean between groups
Kruskal-Wallis Rank Sum Test

```r
> kruskal.test(x ~ fac, data=dfa)

Kruskal-Wallis rank sum test

data:  x by fac
Kruskal-Wallis chi-squared = 6.8805, df = 3,
  p-value = 0.0758

- Null hypothesis of no effect not rejected!
- Note small sample size
- Neither CI nor effect size!
```
> mod1 <- lmrob(x ~ fac, data=dfa)
> summary(mod1)

Coefﬁcients:

| Coefﬁcient   | Estimate | Std. Error | t value | Pr(>|t|) |
|--------------|----------|------------|---------|----------|
| (Intercept)  | 12.3317  | 0.4403     | 28.010  | < 2e-16  *** |
| facnone      | -0.3814  | 0.7744     | -0.492  | 0.628047 |
| facpassive   | -1.6311  | 0.5994     | -2.721  | 0.013550 * |
| faczactive   | -2.5537  | 0.5347     | -4.776  | 0.000131 *** |

- Unequal standard errors more realistic
- Passive group and active group have effect wrt control
- Confirms our intuition
- Better testing due to down weighting of outlying data points
### Confidence Intervals of Effects Compared with Control

<table>
<thead>
<tr>
<th>Group Effect</th>
<th>2.5 %</th>
<th>97.5 %</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>12.33</td>
<td>11.47</td>
</tr>
<tr>
<td>facnone</td>
<td>-0.38</td>
<td>-1.90</td>
</tr>
<tr>
<td>facpassive</td>
<td>-1.63</td>
<td>-2.81</td>
</tr>
<tr>
<td>faczactive</td>
<td>-2.55</td>
<td>-3.60</td>
</tr>
</tbody>
</table>

- CI of passive/active group effect wrt control does not contain zero