MULTIPLE COMPARISONS IN QOE ANALYSIS

Kjell Brunnström
STATISTICAL ANALYSIS OF EXPERIMENTAL DATA

• Data is collected, then what?
• Mean and standard deviation is usually easily computed
• Common question – are two means the same or different?
EXAMPLE

- Example: One video with an error. A number of people in France and an equal number in Sweden rates the quality on a five graded scale
STATISTICAL TEST OR HYPOTHESIS TEST

• Null hypothesis (H₀): One possible arrangement
  • H₀: μ₁ = μ₂ (in this case)
• Alternative hypothesis (H₁): all other arrangement
  • H₁: μ₁ ≠ μ₂ (in this case)
• Student T-test
  • If $t_{obs} \geq t_{critical}$ reject the null hypothesis
  • For example (unequal sample sizes, unequal variance)

$$t_{obs} = \frac{\mu_1 - \mu_2}{s_{diff}}; \\ s_{diff} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}; \\ df = (n_1 - 1) + (n_2 - 1)$$

• Our example:
  • $\mu_1 = 1.77; \mu_2 = 2.00; s_1 = 0.869; s_2 = 0.667; n_1 = 22; n_2 = 19;$
  • $t = -0.927; \text{df} = 39 + \text{Table} \Rightarrow t(40) = 2.02$ at 0.05 significance level
  • $\text{Abs}(t) < t(40)$
  • Statistics packages gives $p = 0.359$
MORE THAN TWO MEAN?

• What if there are more than two means?
• Are the mean the same?
• If we test all the pairs with a t-test, then we know they are the same?
• Unfortunately, there is an increased risk of type-I errors
  • Reject the null hypothesis, although it is true
  • At each pairwise test there is small risk
  • Eg significance level $\alpha = 0.05$ gives 5% risk one comparison and about $n*\alpha$ for $n$ comparisons
• Thus we need to handle this
BONFERRONI CORRECTION METHOD

• the significance level ($\alpha$) should be divided by the number of comparisons $N$

• $\alpha_{\text{comp}} = \frac{\alpha_{\text{total}}}{N}$

• 10 different mean => 10*9/2 comparison 45

• $\alpha_{\text{comp}} = \frac{0.05}{45} = 0.0011$
EXAMPLE

• Compare exp 1 and exp 2
  • Same PVSs used in both exp
  • N_PVS = 100
  • Interesting comparison
    • PVS_exp1 (i) with PVS_exp2 (i)
  • If pre-planned gives 100 comparisons
  • If post-hoc gives 100*99/2 = 4950
RECENT RESULT
INFLUENCE OF THE NUMBER OF SUBJECTS

The graph illustrates the probability of significance with varying numbers of subjects. The x-axis represents the number of subjects, ranging from 1 to 1000, while the y-axis shows the probability of significance. Different lines indicate various conditions, such as different standard deviations (std=1) and differences (diff=0.5, diff=1), as well as different alpha levels (alpha = 0.05, alpha = 0.0005, alpha = 0.000001). The graph also includes a line for 24 subjects.

The probability of significance decreases as the number of subjects increases, indicating a higher likelihood of obtaining statistically significant results with more subjects.
INFLUENCE OF DIFFERENCE
CORRELATION SIGNIFICANCE TESTING

• Fisher z-statistics
• Compare with two-tailed Student statistics
• If $Z_N \geq t_{\text{critical}}$ reject the null hypothesis

\[
Z_N = \frac{z_1 - z_2 - \mu_{(z_1-z_2)}}{\sigma_{(z_1-z_2)}}
\]

\[
\mu_{(z_1-z_2)} = 0
\]

\[
\sigma_{(z_1-z_2)} = \sqrt{\sigma_{z_1}^2 + \sigma_{z_2}^2}
\]

\[
\sigma_z = \sqrt{\frac{1}{N-3}}
\]
INFLUENCE ON CORRELATION

Probability of significance vs. Correlation for different conditions:
- Diff 0.5 N=100
- Diff = 0.5 N = 10
- Diff 0.5 N=1000
- 10 comp
- 45 comp
- 190 comp
INFLUENCE ON CORRELATION

![Graph showing the influence of correlation on probability of significance]

- **Correlation**
- **Probability of significance**

Legend:
- **diff 1,0 N=100**
- **Diff = 1.0 N = 10**
- **Diff 0.5 N=1000**
## INFLUENCE OF RECENT TEST

### TABLE II. TEST OF SIGNIFICANT DIFFERENCE FOR SROCC.

<table>
<thead>
<tr>
<th></th>
<th>PSNR</th>
<th>SSIM</th>
<th>MS-SSIM</th>
<th>MARZ</th>
<th>VAR</th>
<th>VQM</th>
<th>VQM-VFD</th>
<th>PEVQ</th>
<th>V-BLIINDS</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSNR</td>
<td>-</td>
<td>+</td>
<td></td>
<td>+</td>
<td>-</td>
<td>+</td>
<td></td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>SSIM</td>
<td></td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td>+</td>
<td></td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>MS-SSIM</td>
<td>+</td>
<td></td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>MARZ</td>
<td></td>
<td></td>
<td></td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>VAR</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VQM</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+</td>
<td></td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>VQM-VFD</td>
<td>+</td>
<td></td>
<td></td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>PEVQ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V-BLIINDS</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>
INFLUENCE OF RECENT TEST

TABLE II. TEST OF SIGNIFICANT DIFFERENCE FOR SROCC.

<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>
</tr>
</thead>
<tbody>
<tr>
<td>PSNR</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSIM</td>
<td></td>
<td></td>
<td></td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MS-SSIM</td>
<td>+</td>
<td></td>
<td></td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VQM</td>
<td></td>
<td></td>
<td></td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VQM-VFD</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>PEVQ</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>V-BLIINDS</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
DISCUSSION

• Not correcting for Type-I error may see effect that are not there.
• Correcting will lower efficiency
• What can be done:
  • More test subjects may be needed
  • Reduce variance
  • Plan comparisons ahead