Comparing Systems Using Sample Data

CS 700

Comparing alternatives

- Today’s lecture: comparing two alternatives
  - use confidence intervals
- Comparing more than two alternatives
  - ANOVA
    - Analysis of Variance
  - Will discuss later this semester
Comparing Two Alternatives

- Suppose you want to compare two cache replacement policies under similar workloads.
- Metric of interest: cache hit ratio.
- Types of comparisons:
  - Paired observations
  - Unpaired observations.

Paired Observations

![Diagram showing paired observations between System A and System B](image-url)
Example of Paired Observations

- Six similar workloads were used to compare the cache hit ratio obtained under object replacement policies A and B on a Web server. Is A better than B?

<table>
<thead>
<tr>
<th>Workload</th>
<th>Cache Hit Ratio</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Policy A</td>
<td>Policy B</td>
</tr>
<tr>
<td>1</td>
<td>0.35</td>
<td>0.28</td>
</tr>
<tr>
<td>2</td>
<td>0.46</td>
<td>0.37</td>
</tr>
<tr>
<td>3</td>
<td>0.29</td>
<td>0.34</td>
</tr>
<tr>
<td>4</td>
<td>0.54</td>
<td>0.60</td>
</tr>
<tr>
<td>5</td>
<td>0.32</td>
<td>0.22</td>
</tr>
<tr>
<td>6</td>
<td>0.15</td>
<td>0.18</td>
</tr>
</tbody>
</table>

Sample mean: 0.02000
Sample variance: 0.00552
Sample standard dev.: 0.07430

In Excel:
\[ TINV(1-0.9, 5) \]
0.02
0.0743
2.015

0.95 quantile of t-variable with 5 degrees of freedom
90% confidence interval
lower bound: -0.0411
upper bound: 0.0811
Example of Paired Observations

In Excel:

\[
TINV(1-0.9,5)
\]

The interval includes zero, so we cannot say that policy A is better than policy B.

Unpaired Observations

input values for A

System A

unpaired output values

input values for B

System B
Inferences concerning two means

For large samples, we can statistically test the equality of the means of two samples by using the statistic

\[
Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}
\]

- Z is a random variable having the standard normal distribution.
- We need to check if the confidence interval of Z at a given level includes zero.
- We can approximate the population variances above with sample variances when \(n_1\) and \(n_2\) are greater than 30.

Inferences concerning two means (cont’d)

For small samples, if the population variances are unknown, we can test for equality of the two means using the t-statistic below, provided we can assume that both populations are normal with equal variances

\[
t = \frac{\bar{X}_1 - \bar{X}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}
\]

- t is a random variable having the t-distribution with \(n_1 + n_2 - 2\) degrees of freedom and \(S_p\) is the square root of the pooled estimate of the variance of the two samples.

\[
S_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}
\]
Inferences concerning two means (cont’d)

- The pooled-variance t test can be used if we assume that the two population variances are equal
  - In practice, we can use it if one sample variance is less than 4 times the variance of the other sample
- If this is not true, we need another test
  - Smith-Satterthwaite test described in Jain (with some errors)

Unpaired Observations (t-test)

1. Size of samples for A and B: \(n_A\) and \(n_B\)
2. Compute sample means:

\[
\bar{x}_A = \frac{1}{n_A} \sum_{i=1}^{n_A} x_{iA}
\]

\[
\bar{x}_B = \frac{1}{n_B} \sum_{i=1}^{n_B} x_{iB}
\]
Unpaired Observations (t-test)

1. Compute the sample standard deviations:

\[ s_A = \sqrt{\frac{\sum_{i=1}^{n_A} x_{iA}^2 - n_A \left( \bar{x}_A \right)^2}{n_A - 1}} \]

\[ s_B = \sqrt{\frac{\sum_{i=1}^{n_B} x_{iB}^2 - n_B \left( \bar{x}_B \right)^2}{n_B - 1}} \]

Unpaired Observations (t-test)

1. Compute the mean difference: \( \bar{x}_a - \bar{x}_b \)
2. Compute the standard deviation of the mean difference:
\[ s = \sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}} \]
4. Compute the effective number of degrees of freedom.

\[ \nu = \frac{\left( \frac{s_a^2}{n_a} + \frac{s_b^2}{n_b} \right)^2}{\frac{1}{n_a - 1} \left( \frac{s_a^2}{n_a} \right)^2 + \frac{1}{n_b - 1} \left( \frac{s_b^2}{n_b} \right)^2} \]
Unpaired Observations (t-test)

1. Compute the confidence interval for the mean difference:

\[(\bar{x}_a - \bar{x}_b) \pm t_{[1-\alpha/2,\nu]} \times s\]

4. If the confidence interval includes zero, the difference is not significant at 100(1-\(\alpha\))% confidence level.

Example of Unpaired Observations

Two cache replacement policies A and B are compared under similar workloads. Is A better than B?

<table>
<thead>
<tr>
<th>Workload</th>
<th>Cache Hit Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Policy A</td>
</tr>
<tr>
<td>1</td>
<td>0.35</td>
</tr>
<tr>
<td>2</td>
<td>0.23</td>
</tr>
<tr>
<td>3</td>
<td>0.29</td>
</tr>
<tr>
<td>4</td>
<td>0.21</td>
</tr>
<tr>
<td>5</td>
<td>0.21</td>
</tr>
<tr>
<td>6</td>
<td>0.15</td>
</tr>
<tr>
<td>7</td>
<td>0.42</td>
</tr>
<tr>
<td>8</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>0.2657</td>
</tr>
<tr>
<td>St. Dev</td>
<td>0.0934</td>
</tr>
</tbody>
</table>
Example of Unpaired Observations

| na     | 7 |
| nb     | 9 |
| mean diff | -0.135 |
| st.dev diff. | 0.059776 |
| Eff. Deg. Freed. | 13 |
| alpha  | 0.1 |
| 1-alpha/2 | 0.95 |
| \( t[1-alpha/2,v] \) | 1.782287 |

90% Confidence Interval

| lower bound | -0.24193 |
| upper bound  | -0.02886 |

At a 90% confidence level the two policies are not identical since zero is not in the interval. With 90% confidence, the cache hit ratio for policy A is smaller than that for policy B. So, policy B is better at that confidence level.

Approximate Visual Test

CIs do not overlap: A is higher than B

CIs overlap and mean of A is in B’s CI: A and B are similar

CIs overlap and mean of A is not in B’s CI: need to do t-test
Example of Visual Test

<table>
<thead>
<tr>
<th>Workload</th>
<th>Cache Hit Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Policy A</td>
</tr>
<tr>
<td>1</td>
<td>0.35</td>
</tr>
<tr>
<td>2</td>
<td>0.23</td>
</tr>
<tr>
<td>3</td>
<td>0.29</td>
</tr>
<tr>
<td>4</td>
<td>0.21</td>
</tr>
<tr>
<td>5</td>
<td>0.21</td>
</tr>
<tr>
<td>6</td>
<td>0.15</td>
</tr>
<tr>
<td>7</td>
<td>0.42</td>
</tr>
<tr>
<td>8</td>
<td>0.35</td>
</tr>
<tr>
<td>9</td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>0.2657</td>
</tr>
<tr>
<td>St. Dev</td>
<td>0.0934</td>
</tr>
</tbody>
</table>

Non-parametric tests

- The unpaired t-tests can be used if we assume that the data in the two samples being compared are taken from normally distributed populations.

- What if we cannot make this assumption?
  - We can make some normalizing transformations on the two samples and then apply the t-test.
  - Some non-parametric procedure such as the Wilcoxon rank sum test that does not depend upon the assumption of normality of the two populations can be used.

CIs overlap but mean of A is not in CI of B and vice-versa. Need to do a t-test.
Rank-sum (Wilcoxon test)

- Non-parametric test, i.e., does not depend upon distribution of population, for comparing two samples
- Example:
  - Suppose the time between two successive crashes are recorded for two competing computer systems as follows (time in weeks):
    - System I: 0.63 0.17 0.35 0.49 0.18 0.43 0.12 0.20 0.47 1.36 0.51 0.45 0.84 0.32 0.40
    - System II: 1.13 0.54 0.96 0.26 0.39 0.88 0.92 0.53 1.01 0.48 0.89 1.07 1.11 0.58
  - The problem is to determine if the two populations are the same or if one is likely to produce larger observations than the other

Rank-sum test (cont’d)

- U-test is a non-parametric alternative to the paired and unpaired t-tests
- First step in the U-test is to rank the data jointly, in increasing order of magnitude
  - 0.12 0.17 0.18 0.20 0.26 0.32 0.35 0.39 0.40 0.43
  - I I I I II I I II I I
  - 0.45 0.47 0.48 0.49 0.51 0.53 0.54 0.58 0.63 0.84
  - I I II I I II II I I
  - 0.88 0.89 0.92 0.96 1.01 1.07 1.11 1.13 1.36
  - II II II II II II II I
- Assign each data item a rank in this order
  - If there are ties among values, the rank assigned to each observation is the mean of the ranks which they jointly occupy
Rank-sum test (cont’d)

- The values in the first sample occupy ranks 1, 2, 3, 4, 6, 7, 9, 10, 11, 12, 14, 15, 19, 20 and 29
- The sum of the ranks for the two samples, \( W_1 = 162 \) and \( W_2 = 273 \)
- The U-test is based on the statistics
  \[ U_1 = W_1 - \frac{n_1(n_1 + 1)}{2} \]
  
or
  \[ U_2 = W_2 - \frac{n_2(n_2 + 1)}{2} \]
  
or on the statistic \( U \) which is the smaller of the two

Rank-sum test (cont’d)

- Under the null hypothesis that the two samples come from identical populations, it can be shown that the mean and variance of the sampling distribution of \( U_1 \) are
  \[ \mu_{U_1} = \frac{n_1 n_2}{2} \]
  and
  \[ \sigma^2_{U_1} = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12} \]
- Numerical studies have shown that the sampling distribution of \( U_1 \) can be approximated closely by the normal distribution when \( n_1 \) and \( n_2 \) are both greater than 8
Rank-sum test (cont’d)

- Thus, the test of the null hypothesis that both samples come from identical populations can be based on
  \[ Z = \frac{U_i - \mu_{U_i}}{\sigma_{U_i}} \]
  which is a random variable having approximately the standard normal distribution
- The alternative hypothesis is either:
  - Population 2 is stochastically larger than Population 1
    - We reject the null hypothesis if \( Z < -z_\alpha \)
  - Or, Population 1 is stochastically larger than Population 2
    - We reject the null hypothesis if \( Z > -z_\alpha \)

Example cont’d

- At the 0.01 level of significance, test the null hypothesis that the two samples in our example come from the same population
  - Alternative hypothesis, populations are not identical
  - For \( \alpha = 0.01 \), we can reject the null hypothesis if \( Z < -2.575 \) or \( Z > 2.575 \)
    - Calculations: \( n1 = 15, n2 = 14, W1 = 162 \)
      \( U1 = 162 - 15 \times 16/2 = 42 \)
      \( Z = (42 - 15 \times 14/2)/\sqrt((15 \times 14 \times 30)/12) = -2.75 \)
    - Since \( Z \) is less than -2.575, we reject the null hypothesis; we conclude there is a difference between the two systems