3. SAS procedures for t-tests and ANOVA

Karl B Christensen

http://192.38.117.59/~kach/SAS
Comparing two samples

- Two groups: data $x_{11}, \ldots, x_{1n_1}$ and $x_{21}, \ldots, x_{2n_2}$
- Theoretical distributions $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$
- Empirical mean and variance $(\bar{x}_1, s_1^2)$ and $(\bar{x}_2, s_2^2)$
- Significant difference between $\bar{x}_1$ and $\bar{x}_2$?
- Are $\mu_1$ and $\mu_2$ different?
- Null hypothesis $H_0 : \mu_1 = \mu_2$
Comparing two samples

Theoretical distributions $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$
Comparing two samples

Empirical mean and variance \((\bar{x}_1, s_1^2)\) and \((\bar{x}_2, s_2^2)\)
Two-sample $t$-test

- Standard error of mean $SEM = \frac{s}{\sqrt{n}}$.
- Standard error of difference of means
  \[ SEDM = \sqrt{SEM^2_1 + SEM^2_2}. \]
- $T$-test statistic
  \[ t = \frac{\bar{x}_2 - \bar{x}_1}{SEDM} \]
  measures disagreement between data and $H_0$

- If $H_0$ is true, then the distribution of $t$ is symmetric around 0

 reject if prob. of observing a more extreme value $p < 5\%$. 
Equal variances?

Assume $\sigma_1^2 = \sigma_2^2$ before testing $\mu_1 = \mu_2$?

$\sigma_1^2 = \sigma_2^2$ Natural assumption under the $H_0$ (distributions are equal). Nice theory.

$\sigma_1^2 \neq \sigma_2^2$ Looks specifically for difference in means. Approximative theory.

Test for equal variances: Compute test statistic (Note: 2-sided test)

$$F = \frac{s_1^2}{s_2^2}$$

F-distribution with $(f_1, f_2)$ degrees of freedom, where $f_1 = n_1 - 1$ and $f_2 = n_2 - 1$
We want to compare men and women with respect to log SIGF1. We only do the analysis for those between 20 and 30 years of age.

- Open dataset: SET statement
- Compute lsigf1=LOG(sigf1)
- Start PROC TTEST
- Use WHERE to select subgroup
- Specify dependent variable and classification
SAS code for the t-test

LIBNAME RH 'C:\Dropbox\sascourse';
DATA WORK.juul;
   SET RH.juu12;
   lsigf1=LOG(sigf1);
RUN;

PROC TTEST DATA=WORK.juul;
   WHERE age > 20 and age < 30;
   VAR lsigf1;
   CLASS sexnr;
RUN;

The output has three parts:

(i) Statistics showing the mean, the standard error, and the standard error of the mean. Confidence limits are also included.

(ii) T-tests showing two t-tests (one that assumes equal variances and one that doesn’t).

(iii) Equality of Variances showing a test of equal variances. newer version of SAS also have graphical output as default
The TTEST Procedure

Variable: 1sigf1

<table>
<thead>
<tr>
<th>sexnr</th>
<th>N</th>
<th>Mean</th>
<th>Std Dev</th>
<th>Std Err</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>23</td>
<td>5.5992</td>
<td>0.1981</td>
<td>0.0413</td>
<td>5.2470</td>
<td>6.0355</td>
</tr>
<tr>
<td>2</td>
<td>18</td>
<td>5.6306</td>
<td>0.2551</td>
<td>0.0601</td>
<td>5.1059</td>
<td>6.0064</td>
</tr>
<tr>
<td>Diff (1-2)</td>
<td>-0.0314</td>
<td>0.2247</td>
<td>0.0707</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>sexnr</th>
<th>Method</th>
<th>Mean</th>
<th>95% CL Mean</th>
<th>Std Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>5.5992</td>
<td>5.5136</td>
<td>5.6849</td>
</tr>
<tr>
<td>2</td>
<td>Pooled</td>
<td>5.6306</td>
<td>5.5038</td>
<td>5.7575</td>
</tr>
<tr>
<td>Diff (1-2)</td>
<td></td>
<td>-0.0314</td>
<td>-0.1745</td>
<td>0.1116</td>
</tr>
<tr>
<td>Diff (1-2)</td>
<td>Satterthwaite</td>
<td>-0.0314</td>
<td>-0.1801</td>
<td>0.1173</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>sexnr</th>
<th>Method</th>
<th>95% CL Std Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>0.1532 0.2804</td>
</tr>
<tr>
<td>2</td>
<td>Pooled</td>
<td>0.1914 0.3825</td>
</tr>
<tr>
<td>Diff (1-2)</td>
<td>Pooled</td>
<td>0.1841 0.2886</td>
</tr>
<tr>
<td>Diff (1-2)</td>
<td>Satterthwaite</td>
<td></td>
</tr>
</tbody>
</table>

Method           | Variances | DF  | t Value | Pr > |t|
-----------------|-----------|-----|---------|-------|
Pooled           | Equal     | 39  | -0.44   | 0.6594|
Satterthwaite    | Unequal   | 31.422 | -0.43   | 0.6697|
### Equality of Variances

<table>
<thead>
<tr>
<th>Method</th>
<th>Num DF</th>
<th>Den DF</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Folded F</td>
<td>17</td>
<td>22</td>
<td>1.66</td>
<td>0.2629</td>
</tr>
</tbody>
</table>

#### Distribution of Isigf1

![Distribution of Isigf1](image)

Karl B Christensen [http://192.38.117.59/~kach/SAS](http://192.38.117.59/~kach/SAS)
Exercise: t-test

Compare the SIGF1-level in boys and girls *above the age of 5 years* in Juul data set.

1. Use graphical methods and descriptive statistics to compare boys and girls for each Tanner stage.

2. The distribution of SIGF1 is skewed, but for the purpose of this exercise log(SIGF1) can be assumed to follow a normal distribution. Test if the SIGF1-level is the same in boys and girls using *t*-tests in each Tanner group.

3. Quantify the difference. Remember confidence intervals.

4. Can you interpret these differences on the original scale?
Interpretation of difference on original scale

Absolute difference in $\log(SIGF1)$:

-0.0314 (95% CI -0.1801 to 0.1173)

- $t$-test on $x = SIGF1$

$$\bar{x}_B - \bar{x}_G = \mu_B - \mu_G$$

- $t$-test on $y = \log(SIGF1)$

$$\bar{y}_B - \bar{y}_G = \log(\mu_B) - \log(\mu_G) = \log(\mu_B/\mu_G)$$

so $\exp(\bar{y}_B - \bar{y}_G)$ is an estimate of the ratio $\mu_B/\mu_G$.

- Compute

$$\exp(-0.0314) \approx 0.97, \ \exp(-0.1801) \approx 0.84 \ \text{and} \ \exp(0.1173) \approx 1.12$$

and interpret this as a relative difference in $SIGF1$ of

-3% (95% CI -16% to +12%).
Beyond the t-test

- The t-test compares two groups based on an assumption of normality, but what if data are not normally distributed or if we want to compare three or more groups?
- The t-test is robust - because means tend to be normally distributed, sometimes transformation ($x \mapsto \sqrt{x}$ or $\log(x)$) can help. Otherwise nonparametric methods.
- Compare more than three groups using analysis of variance (ANOVA).
Comparing more than two groups

\[ \bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k \quad s_1, s_2, \ldots, s_k \]

Joint test for any differences between the groups.

Why not just pairwise t-tests?

- Mass significance (type I error)
- Loss of overview

The fewer tests, the better.
$x_{ij}$ observation no. $j$ in group no. $i$, e.g. $x_{35}$ the 5th observation in group 3. Model

$$X_{ij} = \mu_i + \epsilon_{ij} \quad \epsilon_{ij} \sim N(0, \sigma^2)$$

The null hypothesis (no differences between groups)

$$\mu_1 = \mu_2 = \ldots = \mu_k$$
Main idea behind analysis of variance (ANOVA): If the variation between group means is large compared to the variation within groups, it is a sign that the null hypothesis is wrong.

The model (grouping) *explains* part of the variation.

\[ \text{Variation} = \text{between gr.} + \text{within gr.} \]
Sums of squares

Let $\bar{x}_i$ denote the mean for group $i$ and let $\bar{x}$ denote the total (grand) mean

**Variation Within groups:**

$$SSD_W = \sum_i \sum_j (x_{ij} - \bar{x}_i)^2$$

**Variation Between groups:**

$$SSD_B = \sum_i \sum_j (\bar{x}_i - \bar{x})^2$$

Can be mathematically proven that

$$SSD_B + SSD_W = SSD_{total} = \sum_i \sum_j (x_{ij} - \bar{x})^2$$
Var. between groups large compared to var. within groups

Small variation within groups

Large variation within groups

A  B  C

high 'between' variation
small 'within' variation

F is large

H₀ is rejected

A  B  C

high 'within' variation
small 'between' variation

F is small

H₀ is not rejected

Karl B Christensen http://192.38.117.59/~kach/SAS

3. SAS procedures for t-tests and ANOVA
F-test for identical group means

Reject the hypothesis if the variation between groups is large compared to the variation within groups.

\[ F = \frac{SSD_B/(k - 1)}{SSD_W/(N - k)} \]

If null hypothesis is true we know distribution of \( F \)

Reject hypothesis that group means are identical if \( F \) too large.
One-way ANOVA in SAS

Compare boys in different Tanner stage with respect to their log SIGF1

1. Generate a new data set
2. Select (sexnr=1, age<20)
3. Use PROC ANOVA or PROC GLM
4. MODEL statement: What is described by what?
5. Tell SAS that tanner is a grouping (CLASS statement)
SAS code

DATA WORK.juulboys;
    SET RH.juul2;
    lsigf1 = LOG(sigf1);
    if sexnr = 1 and 0 < age < 20;
RUN;
PROC ANOVA DATA=WORK.juulboys;
    CLASS tanner;
    MODEL lsigf1 = tanner;
RUN; QUIT;

output

The ANOVA Procedure

    Class Level Information

Class     Levels  Values
    tanner     5    1 2 3 4 5
The ANOVA Procedure

Dependent Variable: lsigf1

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>4</td>
<td>91.3807356</td>
<td>22.8451839</td>
<td>117.57</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>395</td>
<td>76.7537608</td>
<td>0.1943133</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>399</td>
<td>168.1344964</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

R-Square  Coeff Var  Root MSE  lsigf1 Mean
0.543498  7.836512   0.440810   5.625077

Source                      | DF | Anova SS       | Mean Square | F Value | Pr > F |
|-----------------------------|----|----------------|-------------|---------|--------|
tanner                       | 4  | 91.38073563    | 22.84518391 | 117.57  | <.0001 |
Graphical output

![Boxplot of isigf1 distribution across tanner stages]

- **F**: 117.57
- **Prob > F**: <.0001

---

3. SAS procedures for t-tests and ANOVA
PROC GLM DATA=WORK.juulboys;
   CLASS tanner;
   MODEL lsigf1 = tanner / SOLUTION;
RUN; QUIT;

also provides graphical output
Mann-Whitney test, Wilcoxon test, Kruskal-Wallis test

Nonparametric statistics: ‘t-test’ (Mann-Whitney test, Wilcoxon test) or ‘ANOVA’ (Kruskal-Wallis test) on ranks

\[
\begin{array}{c|c}
\text{Data} & \text{Rank} \\
3, 9, 17, 50 & 1, 5, 6, 7 \\
4, 7, 8, 7, 200 & 2, 3, 4, 8 \\
\end{array}
\]

Distribution is (in principle) known under null hypothesis. Does not depend on data following a normal distribution. Other “scores” than ranks can also be used

PROC NPAR1WAY - also provides graphical output
Non parametric tests in PROC NPAR1WAY

The WILCOXON option selects rank scores = Kruskal-Wallis\(^1\)

```
PROC NPAR1WAY DATA=WORK.juul WILCOXON;
   WHERE sexnr = 1 and 0 < age < 20;
   VAR sigf1;
   CLASS tanner;
RUN;
```

Output

The NPAR1WAY Procedure

<table>
<thead>
<tr>
<th>tanner</th>
<th>N</th>
<th>Sum of Scores</th>
<th>Expected Under H0</th>
<th>Std Dev Under H0</th>
<th>Mean Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>192</td>
<td>20758.00</td>
<td>38496.00</td>
<td>1155.20917</td>
<td>108.114583</td>
</tr>
<tr>
<td>2</td>
<td>38</td>
<td>8222.00</td>
<td>7619.00</td>
<td>677.99178</td>
<td>216.368421</td>
</tr>
<tr>
<td>3</td>
<td>23</td>
<td>6569.50</td>
<td>4611.50</td>
<td>538.28582</td>
<td>285.630435</td>
</tr>
<tr>
<td>4</td>
<td>32</td>
<td>10387.00</td>
<td>6416.00</td>
<td>627.30283</td>
<td>324.593750</td>
</tr>
<tr>
<td>5</td>
<td>115</td>
<td>34263.50</td>
<td>23057.50</td>
<td>1046.52522</td>
<td>297.943478</td>
</tr>
</tbody>
</table>

Kruskal-Wallis Test

<table>
<thead>
<tr>
<th>Chi-Square</th>
<th>Pr &gt; Chi-Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>254.3465</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

\(^1\)Mann-Whitney obtained if there are only two groups to compare
Omitting WILCOXON yields many different tests and many plots.
More non parametric tests. Sign test

Paired data where patients rate two drugs $A$ and $B$

$$M_i = \begin{cases} 1, & \text{if } A_i > B_i; \\ 0, & \text{if } B_i < A_i \end{cases}$$

for $i = 1, \ldots, 20$. Under the null hypothesis

$$H_0 : P(A_i > B_i) = \frac{1}{2}$$

the test statistic $M = \sum_{i=1}^{20} M_i$ is binomially distributed

```
PROC UNIVARIATE DATA=WORK.drugs;
   VAR D;
RUN;
```

(look for

Tests for Location: Mu0=0

reports z-score rather than $M$)
Nonparametric tests in SAS III

- Signed rank test (Tests for Location: $\mu_0=0$).
- Friedmans test.
- Jonckheere-Terpstra test

http://192.38.117.59/~kach/SAS/nonpar
The data set

http://192.38.117.59/~kach/SAS/RCT.sas7bdat
http://192.38.117.59/~kach/SAS/RCT.txt
http://192.38.117.59/~kach/SAS/RCT.xlsx

contains data from an RCT where a physical exercise intervention in cancer patients was evaluated. Consider the variables

**ID**  id number

**_VO2**  Aerobic capacity (VO$_2$max)

**group**  Intervention/control group assignment

**time**  Time (1: baseline data 3: after intervention)
Exercise: RCT data

1. Compare baseline aerobic capacity across the two groups using histograms, boxplots and descriptive statistics.

2. Compare aerobic capacity after the intervention across the two groups using histograms, boxplots and descriptive statistics.

3. Suggest how change scores
   \[(\text{VO}_2\text{max after intervention}) - (\text{baseline VO}_2\text{max})\]
   could be evaluated.