Statistics & Experimental Design with R

Barbara Kitchenham

Keele University

Comparing two or more groups

Part 5

Aim

- To cover standard approaches for independent and dependent groups
  - For two groups
    - Student's "t" test (parametric)
    - Mann-Whitney Wicoxon (non-parametric)
  - For multiple groups
    - ANOVA
    - Kruskal-Wallis
- To introduce more modern approaches for 2 and more groups
  - Non-parametric
  - Robust

Student's "t"

- Standard classical method
- Two independent groups
  - Size n_1 and n_2
  - Some measure of interest x_{ij}
  - i=1 or 2 specifying group
    - $j=1,\dots n_1$ if $i=1$
    - $j=1,\dots n_2$ if $i=2$
- Assumptions
  - x_{ij} are iid
  - $x_{ij} \sim N(\mu_i,\sigma^2)$
- H0: $\mu_1 = \mu_2$, H1: $\mu_1 \neq \mu_2 \mid \mu_1 < \mu_2 \mid \mu_1 > \mu_2$

Justification

- Normal distribution means: $\bar{x}_i \sim N(\mu_i, \sigma^2/n_i)$
- Since individual x_{ij} independent μ_i in each group are independent
- Variance of $\bar{x}_1 - \bar{x}_2$ is $\sigma^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)$
- Estimate of σ^2 is

$$s^2 = \frac{(n_1-1)\sum_{j=1}^{n_1}(x_{1,j}-\bar{x}_1) + (n_2-1)\sum_{j=1}^{n_2}(x_{2,j}-\bar{x}_1)}{(n_1+n_2-2)}$$

- Under null hypothesis $t = \frac{(\bar{x}_1-\bar{x}_2)}{s\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} = 0$
  - With n_1+n_2-2 degrees of freedom

Variation1

- Paired values
  - $n_1=n_2=n$
  - Paired values are not independent so

$$var(\overline{x}_1 - \overline{x}_2) = \frac{2(s^2 - cov(x_1, x_2))}{n}$$

  - Difference
    - $d_j=x_{1j}-x_{2j}$
  - Paired values reduces variance
    - More likely to find a significant difference
    - Reason why repeat measure experiments are considered useful
  - Degrees of freedom=n-1

Variation 2

- Variance of groups differ
  - Welch's test (default in R)

$$var(\bar{x}_1 - \bar{x}_2) = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)}$$

  - Changes degrees of freedom (ν)

$$\nu = \frac{(q_1 + q_2)^2}{\left(\frac{s_1^2}{(n_1 - 1)} + \frac{s_2^2}{(n_2 - 1)}\right)}$$

  - where $q_i = \frac{s_i^2}{n_1}$

Problems with t-test

- Mean is not robust
  - Single large value can inflate mean
- Estimate of variance may be very poor
  - If there are outlier values that inflate mean they will also inflate variance
  - Estimate of variance is not robust
- If outliers in the data real effects may not be found
  - i.e. power of t-test is low if there are outliers
- In the presence of outliers, the outliers may not be easily detected (i.e. masked)

Mann-Whitney-Wilcoxon test

- Non-parametric test
  - Used very frequently in SE studies because datasets are often not Normal
- Usually estimated via ranks
  - Values measured on items in two groups
    - Rank values across all values
  - Mann-Whitney $U = \sum_{i=1}^{m} \sum_{j=1}^{n} \phi(x_i, y_j)$
  - where

$$\phi(x_i, y_i) = \begin{cases} 1 \; if \; x_i < y_i \\ otherwise \end{cases}$$

  - Wilcoxon, W=Sum of ranks from G2
    - W=U+n (n+1)/2

Testing process

- Large sample approximation
- Converts into standard normal deviate
  - $E_0(W)=n(m+n+1)/2$
    - Sum of all ranks $=(n+m)\times(n+m+1)/2$
    - Under H_0 Proportion of ranks in Group 2= $n/(n+m)$
  - $Var_0(W)=mn(n+1)/12$
  - Standardized $(W)=[W-E_0(W)]/[Var_0]^{0.5}$
  - For U
    - $E_0(U)=mn/2$
    - $Var_0(U)=mn(m+n+1)/12$
- R function: wilcox.test reports U (but says W)

Problems with Mann-Whitney

- Has poor power if:
  - Ties among data
  - When distribution of two groups differs, uses the wrong standard error
- Alternative methods available
  - Mann-Whitney test is related to probability (p) than random observation from group 1 <random observation from group 2
    - H0: p=0.5 $$\hat{p} = \frac{U}{n_1 n_2}$$
  - Other methods based on this viewpoint

Alternative "New" Nonparametric Methods

- Cliff's method (1996)
  - $p_1=P(X_{l1}>X_{i2})$, $p_2=P(X_{l1}=X_{i2})$, $p_3=P(X_{l1}<X_{i2})$
  - $P=p_3+0.5p_2$
  - $\delta=p_3-p_{1,}$ H0: $\delta=0$ giving $\delta=1-2P$
- Brunner-Munzel (2000)
  - When tied values average rank of tied values

$$\overline{R}_j = \frac{1}{n_j}\sum_{1}^{n_j} R_{ij} \qquad \hat{p} = \frac{1}{n}\left(\overline{R}_1 - \overline{R}_2\right) + 0.5$$

- R functions in WRS package
  - Load library WRS

Advantages of New methods

- $\hat{P}$ provides a sensible non-parametric effect size
- Have well-defined process for handling tied data
- Version of both Cliff & Brunner-Munzel available for three or more groups
  - Although tests suggest Cliff is slightly better at achieving specified alpha level

Permutation test

- Useful when data sets are small
- Calculate test statistic based on actual data T_0
    - Could be "t" value, the Mann-Whitney statistics or another test statistic e.g. sum of ranks of smallest group
- Resample data **without replacement**
    - Calculate and record new sum (T_1)
- Repeat for every possible way of arrangement of data
- Arrange T_i in ascending order
- If T_0 fall outside the middle 95% of values, reject hypothesis
- If too many permutations, take sample

R Permutation Test facility

- Load packages
  - coin & lmPerm
- library(coin)
  - For t-test
    - oneway_test(y~A)
  - For Wilcoxon test
    - wilcox_test(y~A)
  - A must be defined as a factor with two levels

Other robust approaches

Keele University

- Use differences between medians and standard error of medians, then

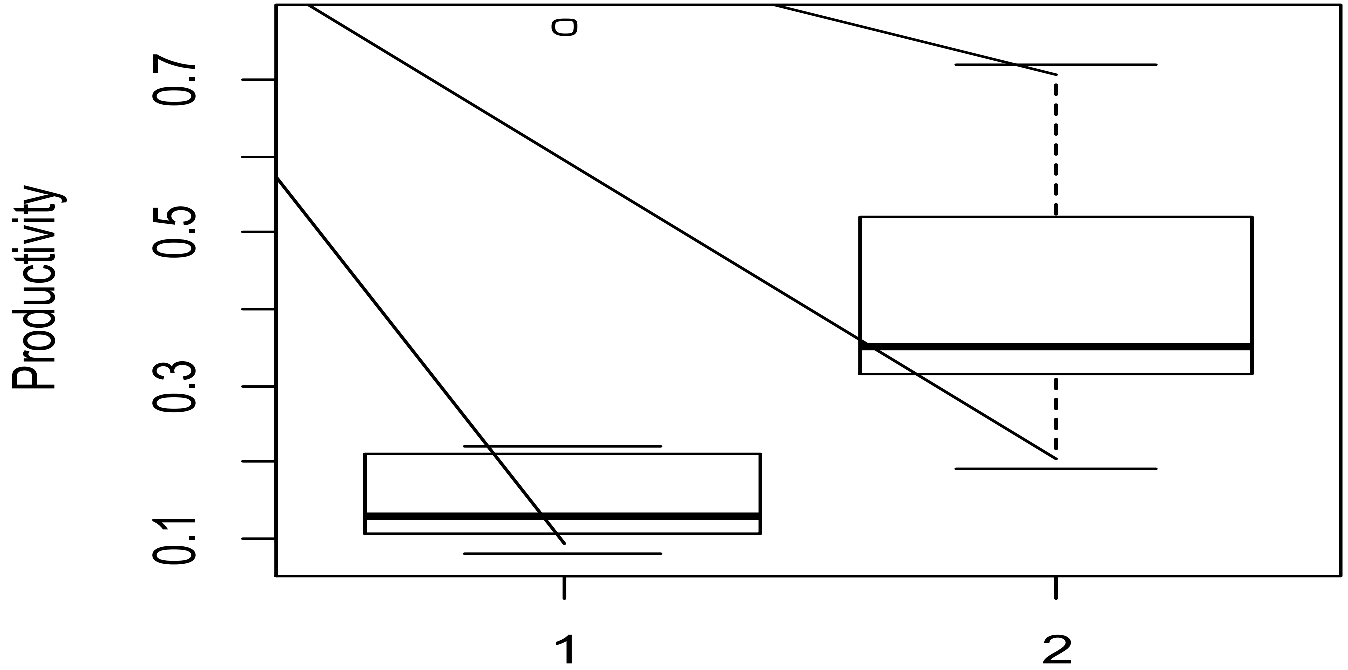
$$M1 - M2 \mp c\sqrt{S_1^2 + S_2^2}$$

  - where c=(1-α/2) quantile of unit normal distribution
  - But which estimate of SE of median?
- Version of t-test based on 20% trimmed means
  - Allowing for unstable variances
  - Yuen-Welch method available in R package WRS
    - Library(WRS)
    - yuen(y,x,tr=0.2,alpha=0.05)

Comparing Two Groups

- From COCOMO dataset
- Productivity (KLoc/MM) of organic projects that used different amounts of tool support
- GR1 (Low): {0.09, 0.13, 0.77,0.08, 0.20, 0.22, 0.12}
- GR2 (Average): {0.19,0.48,0.72,0.31,0.34,0.34,0.45,0.64, 0.35,0.56 }

Box plot



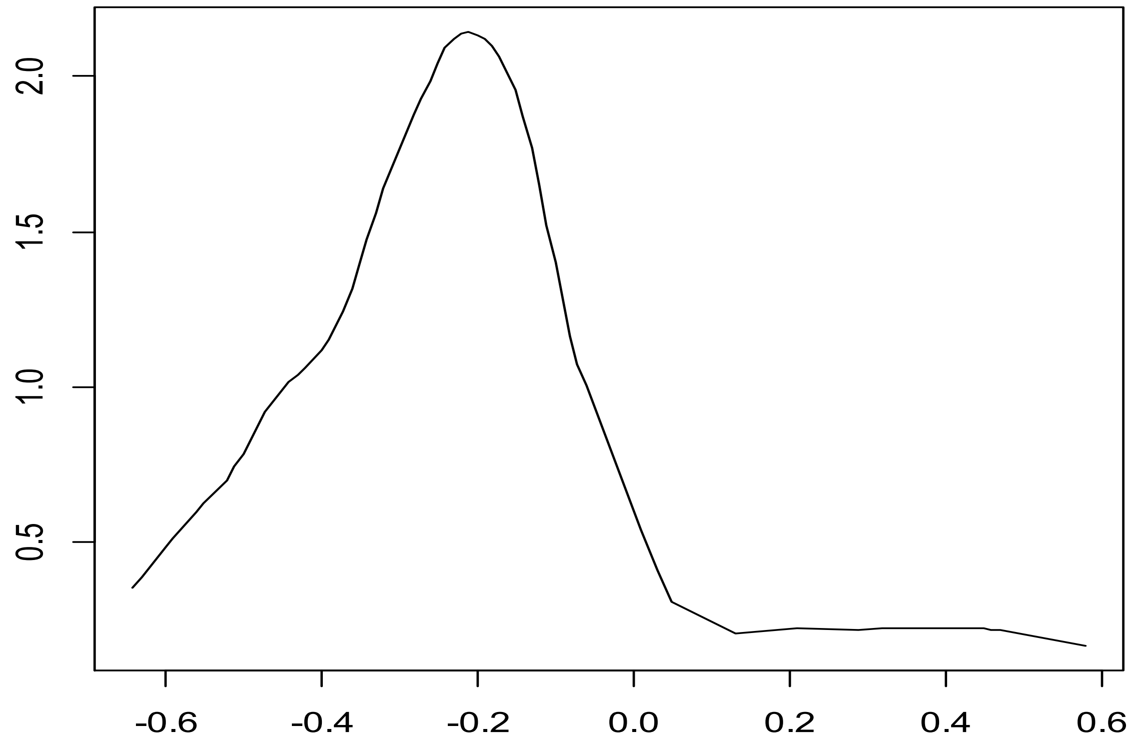
Are groups different?

- Basic statistics
  - Mean G1=0.23 (n_1=7)
  - Mean G2=0.4236 (n_2=11)
  - StDev1=0.2439
  - StDev2=0.1622
  - Median G1=0.13
  - Median G2=0.35

Difference Test Results

- t-test, t=2.0348, df=16, p=0.05879
- Welch test, t=1.8558, df=9.406, p= 0.09503,
- Wilcoxon rank test p=0.0204
- Yuen-Welch test for trimmed means
  - 20% Trimmed means G1=0.152, G2=0.4014
  - p=0.0029, df=9.3
- Cliff, $\hat{p}$ =0.8312, CI (0.46131, 0.9659), p=0.081
- Brunner-Munzel, $\hat{P}$ =0.8312, CI (0.4894, 1.1729), p=0.056, df=6.42
- Permutation t-test, z=1.8694, p=0.062
- Permutation Wilcoxon test ,z=2.3095, p=0.019

Robust methods plot difference



Reasons for Disagreement

- Outlier in Group 1
  - Group 1 Mean and Variance appear inflated
- Box plots suggest groups do not have the same variance
  - Variance inflation has masked difference
  - Ordinary t-test close to significant because degree of freedom greater than for Welch test
- Trimmed means remove outlier, reduce group1 variance and find significant difference
- Standard robust measures fairly resilient to outlier
- New methods do not find a significant effect
- Permutation methods mimic their base test

Issues with Robust methods

- The main problem with using more appropriate methods
- Major reduction with degrees of freedom
- One approach is to use bootstrap to calculate
  - Standard error
  - Confidence limits

Example

- Yuen-Welch (catering for heteroscedasity)
  - No trimming
    - Without bootstrap CI (-0.4281, 0.04085)
    - Bootstrap CI (-0.4820, 0.09478)
  - 20% Trimming
    - Without bootstrap CI(-0.3901, -0.1088)
    - With bootstrap CI (-0.3807, -0.1187)
- No major difference but
  - Bootstrap values probably more reliable

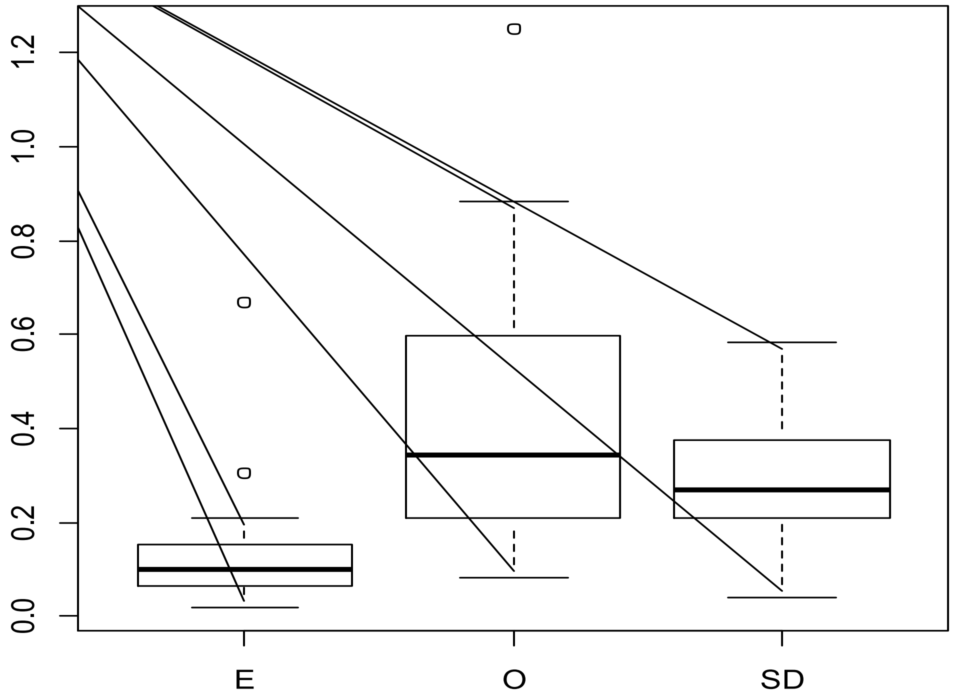
Conclusions

- Always inspect your data
- Different results from different methods need to be investigated
- Permutation method mimics the standard test statistic it uses
  - Still may be useful if no standard statistic exists!
- We need to be able to identify outliers
  - Also need to know what we do about them

Multiple Group Methods

Non-parametric and Robust

COCOMO Productivity for each Mode



Summary Statistics

Mode	Projects	Mean Productivity	St Dev Productivity	20%Trimmed mean
Embedded (E)	28	0.1296	0.1232	0.1052
Semi-Detached (SD)	12	0.2910	0.1670	0.2850
Organic (O)	23	0.4368	0.2998	0.3900

Keele University

Robust methods

- Yuen-Welch method for trimmed means
  - Allowing for heteroscedasticty
  - Has been adapted for three or more groups
- Also possible to estimate linear combinations of means
  - E.g can check whether effect of three treatments is linear
    - If effect of T1>T2>T3, linear increase can be tested with linear combination
      - Mean(T3)-Mean(T2)=Mean(T2)-Mean(T1)
      - Mean(T3)-2Mean(T2)+Mean(T1)=0

Keele University

Yeun-Welch Results

- Use R Function lincon(w,con=0, tr=0.2, alpha=0.05)
  - con describes the linear combination
    - If 0 all pair-wise contrasts performed

Group 1	Group 2	Test statistic	Critical value	se	df
E	SD	4.8904	2.8967	0.03677	8.8933
E	O	4.1887	2.6690	0.0680	14.8940
SD	O	1.3961	2.5945	0.7523	19.6753

Linear Combinations

- COCOMO cost drivers are supposed to have an increasing impact on effort/productivity
  - TOOLcat recoded to
    - low=very low or low (20 projects)
    - Normal (28 projects)
    - High= High, Very High, Extra High (14 projects)
  - Linear Contrast: low-2 × normal-high=0
  - Using lincon(x,con=vec,tr=.20) where vec=c(1,-2,1)
    - x is list variable containing Productivity values for each TOOLcat group
    - Lc=0.0352 with s.e.=0.1295
    - Test value=0.2523, with df=19.93, p=0.803
  - Results consistent with linear relationship between levels

Standard Non-Parametric Method

- Kruskall-Wallis
  - Standard Analysis of Variance
  - Using Ranks not raw data
  - kruskal.test(Productivity~Modecat,cocomo)
- Finds significant difference between productivity for different Modes
  - Test statistic=24.1368
  - p-value=5.738e-06

Robust Non-Parametric Methods

- Brunner, Dette & Munk (BDM) method
  - Based on ranks
  - Allows tied values
  - R Function bdm(w)
  - Finds significant difference between productivity for different modes p=.000295
  - **Relative** effect sizes reported when more than two groups
    - Mode E RES=0.3033
    - Mode SD RES=0.5860
    - Mode O RES=0.6946

Relative Effect Size

- BDM method reports relative effect size if more than two groups
- The relative effect size is

$$\mathrm{RES} = \hat{p}_i = \frac{\bar{R}_{i\cdot} - 0.5}{N}$$

- Where $\bar{R}_{i\cdot}$ is mean rank of group i
- N is total number of observations
- If H0 true all groups have a similar RES

Robust Non-Parametric Methods - Continued

- Cliff method with Hochberg's method for controlling multiple tests
- R function cidmulv2(w)

Group 1	Group 2	phat Prob(G1<G2)	p-value	Critical value
E	SD	0.8036	0.017	0.025
E	O	0.8804	0.001	0.0167
SD	O	0.6341	0.200	0.05

Recommendation

- With obviously non-Normal data
  - Cliff's test is an appropriate choice
    - Provides a robust, non-parametric effect size
    - Test that is reliable when there are tied values
- If both data sets are symmetric
  - But heavy tails (i.e. many outliers)
  - Interested in whether central location is different
    - Consider trimmed means
      - Yuen-Welch method