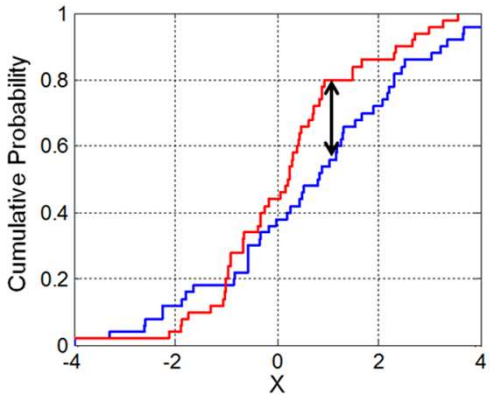


Estatística não-paramétrica

- Teste de correlação de Spearman
- IC95 da mediana populacional
- Testes para uma mediana populacional: teste do sinal e de Wilcoxon
- Teste U de Mann-Whitney
- Teste W de Wilcoxon
- Teste H de Kruskal-Wallis
- ANOVA de Friedman

Panorama

- Há pesquisas nas quais os testes paramétricos não são adequados, pois os dados não satisfazem às condições e suposições para o seu uso
- Lembrar que as suposições em geral são condições suficientes para os testes estatísticos

3

Panorama

- Correlação de Pearson
  - Correlação de Spearman
- t para uma condição
  - Teste do sinal para mediana populacional
  - Teste de Wilcoxon para mediana populacional
- t para duas condições
  - U de Mann-Whitney
  - W de Wilcoxon
- ANOVA unifatorial
  - H de Kruskal-Wallis
  - ANOVA de Friedman

4

Estatística não-paramétrica

- Os testes não-paramétricos permitem que a VI e a covariável sejam ordinais e/ou que a VD tenha qualquer distribuição
- Os testes robustos geralmente prescindem da suposição de homocedasticidade da VD (e.g., teste t de Welch, ANOVA de Welch etc.)
- Os testes não-paramétricos usam os postos (*ranks*) das VDs e das covariáveis como um artificio para comparar distribuições de probabilidade da VD nas condições e assim testar a hipótese nula
- Os testes não-paramétricos com VDs quantitativas, em geral, necessitam das suposições de igualdade dos tipos das distribuições e simetria das distribuições das VDs nas condições do estudo

6

Estatística robusta vs. não-paramétrica

- Estatísticas robustas e não-paramétricas existem para lidar com distribuição desconhecida, heterocedasticidade, não-linearidade, outlier, amostra pequena e desbalanceada

- Estatística robusta
  - Cálculos complexos com dados brutos ou transformações não-lineares
    - Transformações potência de Tukey e de Box-Cox (*transformation*)
    - Aparamento (*trimming*)
    - Ponderação (*weighting*)
    - Reamostragem (*bootstrapping*)

- Estatística não-paramétrica
  - Cálculos simples com postos (*ranking*)
  - Distribuição desconhecida (*distribution free*)

WONNACOTT, T & WONNACOTT, R (1990) *Introductory statistics for business and economics*, 4th ed, p. 536.

7

16/07/10 CLINICAL

Sample size estimation and statistical power analyses

Bhavna Prajapati, Mark Dunne & Richard Armstrong

correspondente

Parametric test	~~Equivalent~~ non-parametric test	ARE
One sample t test	Wilcoxon One sample test	0.955
Paired t test	Wilcoxon Signed-ranks test	0.955
Unpaired t test	Mann Whitney U test	0.955
Pearson's correlation coefficient	Spearman and Kendal's correlation coefficient	0.910
One way ANOVA	Kruskal-Wallis test	0.955
Repeated measures ANOVA	Friedman test	$\frac{0.955k}{k+1}$

Table 3

Asymptotic relative efficiency (ARE) of some common non-parametric tests. "k" is the number of groups

8

Teste estatístico paramétrico ou não-paramétrico?

- Para um dado número de unidades experimentais no estudo, N, testes paramétricos são mais poderosos do que os não-paramétricos correspondentes, desde que todas as suposições dos testes paramétricos e dos não-paramétricos sejam satisfeitas (e.g.: o teste t de Student tem as suposições de normalidade e homocedasticidade a mais que que o correspondente teste não-paramétrico U de Mann-Whitney)
- Toda a informação concernente às magnitudes das observações quantitativas é perdida ao convertê-las em postos (*ranks*) **[sic] (os postos são usados como artifício para comparar duas distribuições quaisquer)**
- RUNYON, R. & HABER, A. (1973) *Fundamentals of behavioral statistics*. USA: Addison-Wesley, p. 235-236)

9

Testes paramétricos vs. não-paramétricos

- Se os testes não-paramétricos têm menos suposições sobre os dados, por que não usar apenas eles?
- R.: Os testes paramétricos, tais como t, ANOVA e ANCOVA, são naturalmente robustos para normalidade, desde que a distribuição dos dados seja simétrica e tenha poucos outliers; além disso, se a amostra é grande, o TCL funciona. Os testes não-paramétricos ignoram a informação de distribuição exata dos dados gerando, e.g., IC95% mais largos, i.e., com menos poder, que os paramétricos.
- NORUSIS, M (1998) *SPSS 8 Guide to data analysis*. NJ: Prentice-Hall, p. 332

10

Testes paramétricos vs. não-paramétricos

- O que eu deveria fazer se não estou certo se eu tenho que usar um teste paramétrico ou não-paramétrico?
- R.: Na dúvida, use ambos! Se conseguir a mesma decisão sobre a hipótese nula nos testes paramétrico e não-paramétrico, não há nada com o que se preocupar. Se o teste não-paramétrico é estatisticamente não-significante e o paramétrico é significante, tente descobrir o motivo. Há outliers? Valores influentes? A distribuição da VD nos grupos é simétrica? Normal? Há desbalanceamento? Há heterocedasticidade? Se a VD é intervalar e a amostra é grande, tente transformação potência de Tukey para simetrizar as distribuições da VD nas condições, homegeneizar as variâncias das condições e linearizar as relações entre as variáveis.
- NORUSIS M. (1998) *SPSS 8 Guide to data analysis*. NJ: Prentice-Hall, p. 335

11

Estatística não-paramétrica

- “Os testes não paramétricos **não exigem [sic]** condições dos dados e você pode usar os descritos neste capítulo com segurança para analisar dados quando achar que não conseguirá satisfazer as condições dos testes paramétricos.”
- (Dancey & Ready, 2013, p. 518, 5e, Português)

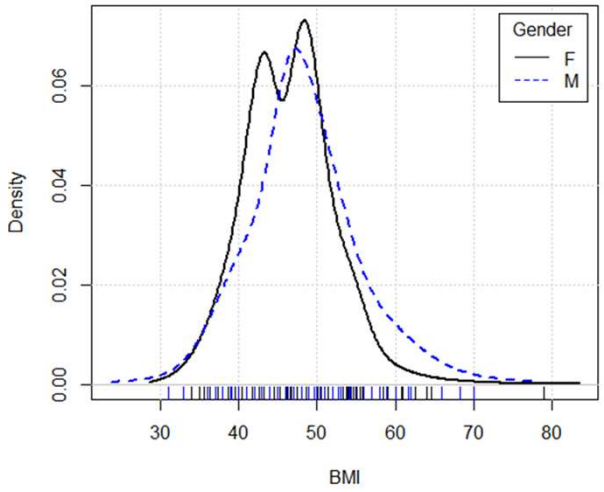
12

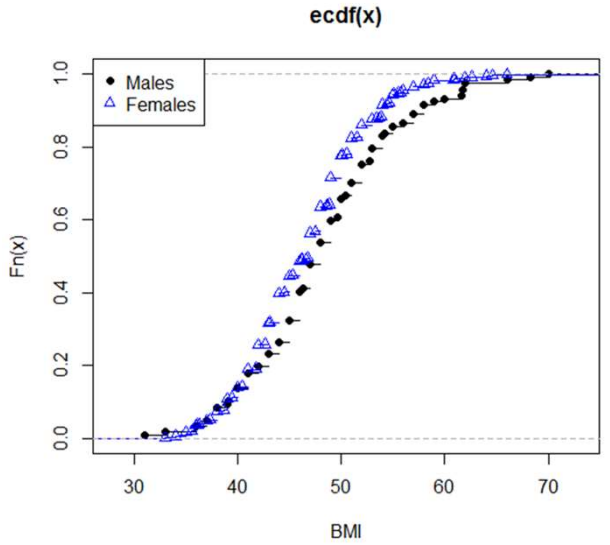
Estatística não-paramétrica

- “Esses testes (Mann-Whitney e Wilcoxon) são muito mais simples do que os testes t, pois **não envolvem [*sic*]** cálculos de médias, desvios-padrão e erros-padrão.” (Dancey & Ready, 2013, p. 524, 5e, Português)
- “Os testes Mann-Whitney e Wilcoxon avaliam se existe diferença estatística significativa entre as **médias dos postos [*sic*]** de duas condições.” (Dancey & Ready, 2013, p. 524, 5e, Português)
- “Estamos somente interessados em U, embora a conversão para um valor-z seja útil, pois **o valor-z dá uma medida do tamanho do efeito [*sic*]** (veja a Seção 4.2).” (Dancey & Ready, 2013, p. 528, 5e, Português)
- “**Histogramas [*sic*]** para as duas condições foram inspecionados separadamente. Como os dados eram assimétricos e o número de participantes pequeno, o teste estatístico mais apropriado foi o de Mann-Whitney.” (Dancey & Ready, 2013, p. 528, 5e, Português)

O que é o teste de hipótese nula?

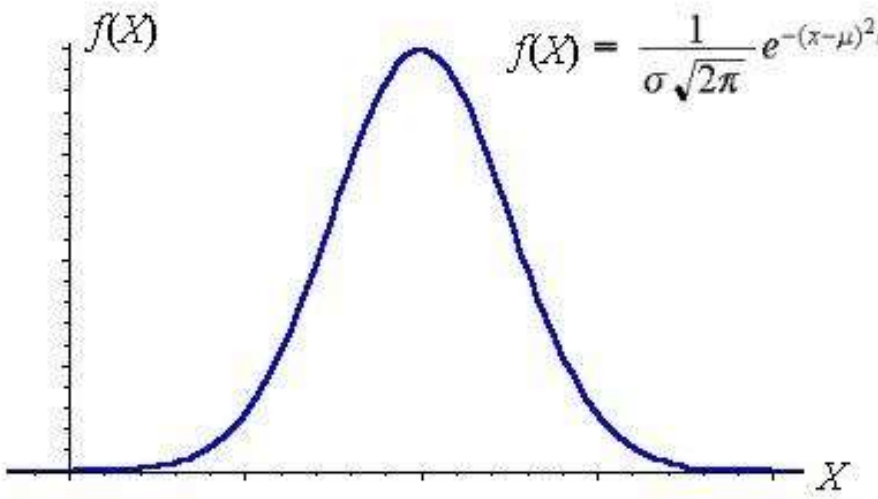
- Comparação de distribuições de probabilidade populacionais das condições do estudo

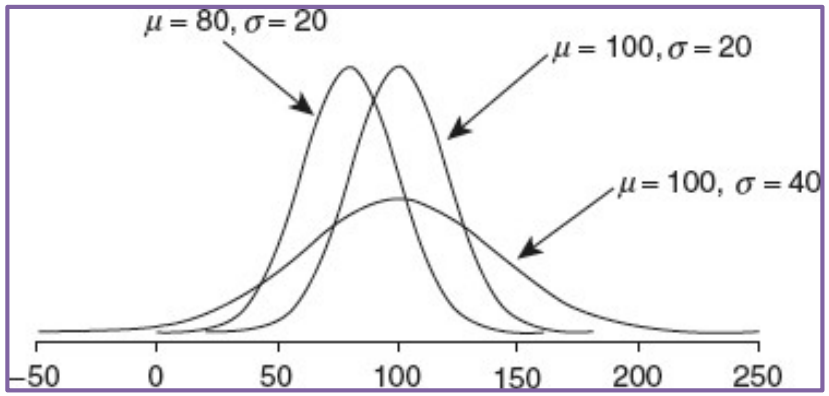




14

Teste paramétrico

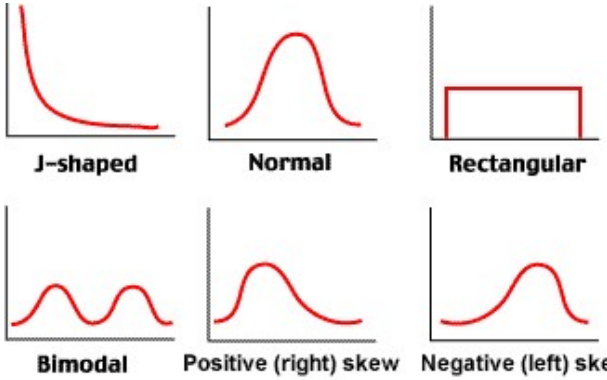


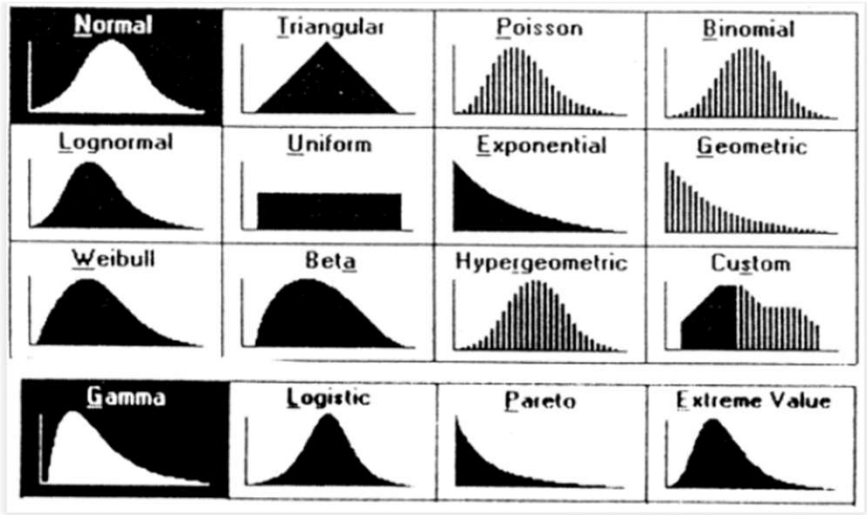


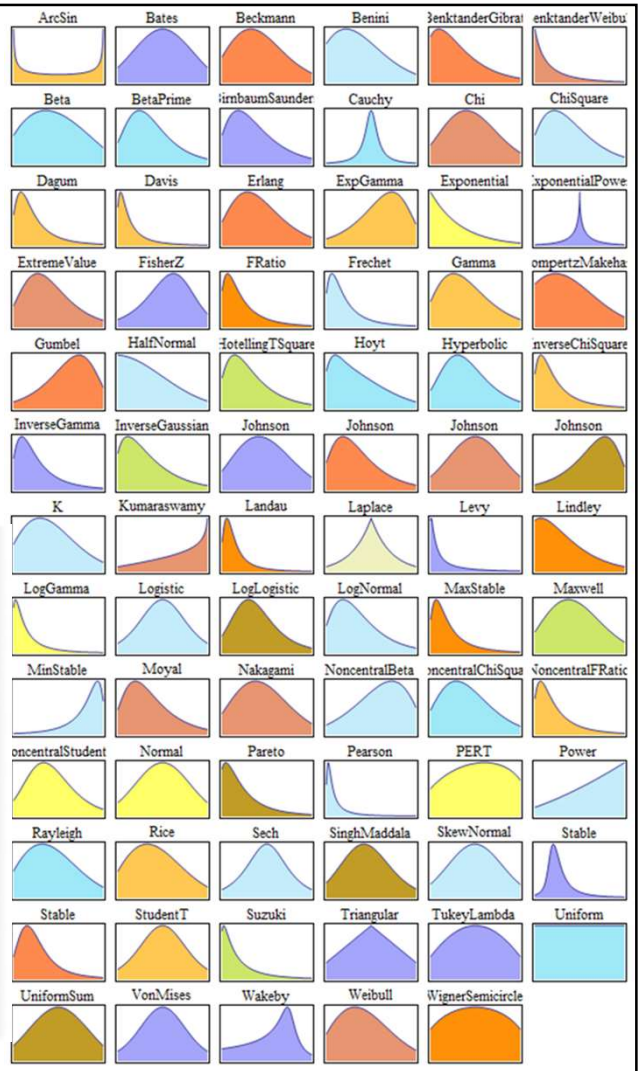
- Média μ e desvio-padrão σ são os parâmetros da normal
- Outras medidas NÃO são parâmetros da normal: mediana, percentil, moda, intervalo interquartílico etc.

15

Distribuições da VD



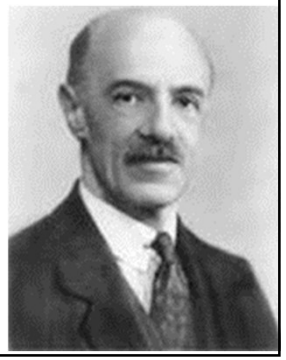




16

Teste de correlação de Spearman

- Avalia a relação de (de)crescimento não necessariamente linear (relação de monotonicidade) entre duas variáveis pelo menos ordinais
- Cálculo da correlação de Spearman: correlação de Pearson entre os postos de duas variáveis pelo menos ordinais



18

6. CORRELATION

One of the earliest applications of a rank transformation involves computing Pearson's product moment correlation coefficient

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{[\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2]^{1/2}} \qquad (6.1)$$

on ranks to obtain Spearman's rho

$$\rho = \frac{\sum \left(R(X_i) - \frac{n+1}{2}\right)\left(R(Y_i) - \frac{n+1}{2}\right)}{\left[\sum \left(R(X_i) - \frac{n+1}{2}\right)^2 \sum \left(R(Y_i) - \frac{n+1}{2}\right)^2\right]^{1/2}} \qquad (6.2)$$

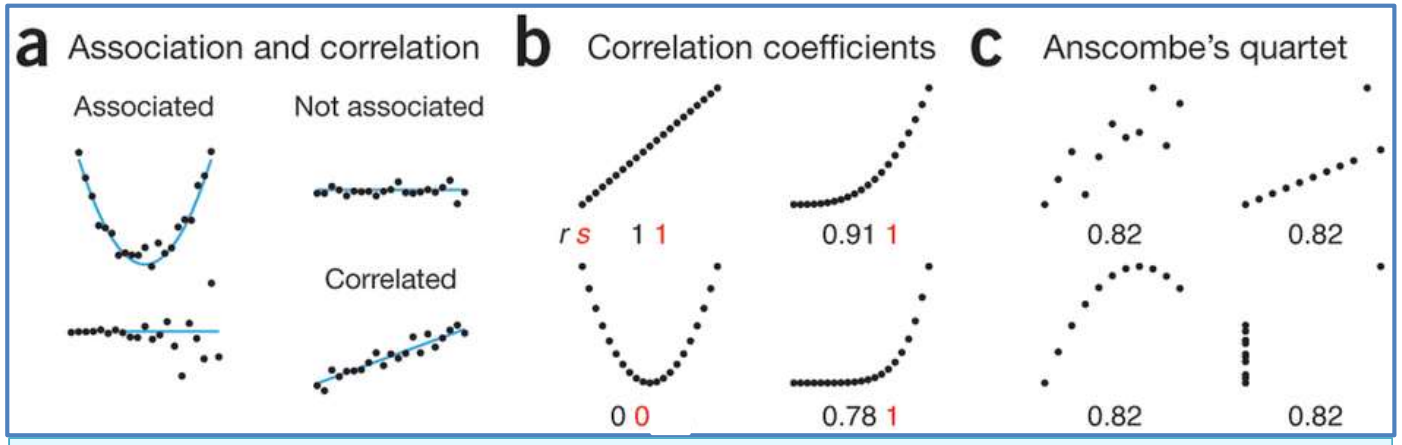
for paired data $(X_1, Y_1), \ldots, (X_n, Y_n)$. Since the observations within the subset $\{X_i\}_{i=1}^n$ are ranked within themselves, and the same is true for the subset $\{Y_i\}_{i=1}^n$, this is an example of an RT-2 type procedure. Just as r is a measure of linearity of the relationship between X and Y, so is ρ a measure of the linearity between the ranks of X and the ranks of Y, which translates as a measure of monotonicity in the relationship between X and Y.

Rank Transformations as a Bridge Between Parametric and Nonparametric Statistics - Conover & Iman - 1981

19

ALTMAN, N. & KRZYWINSKI, M. (2015) Association, correlation and causation. *Nature Methods* (12): 899-900
http://www.nature.com/nmeth/journal/v12/n10/full/nmeth.3587.html

Correlation implies association, but not causation.
Causation implies association, but not correlation.



(a) Scatter plots of associated (but not correlated), non-associated and correlated variables. In the lower association example, variance in y is increasing with x.
(b) The Pearson correlation coefficient (r, black) measures linear trends, and the Spearman correlation coefficient (s, red) measures increasing or decreasing trends.
(c) Very different data sets may have similar r values. Descriptors such as curvature or the presence of outliers can be more specific.

20

Avaliação de atratividade

- A maneira como avaliamos a atratividade em outros se relaciona a quão atraentes acreditamos sermos?
- Pesquisadores pediram a 9 pessoas que avaliassem a sua atratividade e a de outra pessoa usando itens de diferencial semântico de 10 pontos: 1 equivale a horroroso e 10, a maravilhoso.
- Os dois itens são considerados ordinais.
- O pequeno número de participantes, a natureza dos dados e o fato de muitos participantes terem avaliado a si próprios com valores próximos ao maravilhoso (assimetria) devem fazer você suspeitar que os dados não satisfaçam as condições para o cálculo da correlação de Pearson ou a realização de um teste paramétrico de correlação.

21

Teste de correlação de Spearman

- Se as duas variáveis são intervalares, o cálculo da correlação de Pearson é válido.
- Se as duas variáveis intervalares têm distribuição binormal ou o tamanho da amostra é maior que 30, o teste de hipótese da correlação de Pearson clássico é válido.

Avaliação de atratividade

Dados | SpearmanCorr_AvaliacaoAtratividade.R

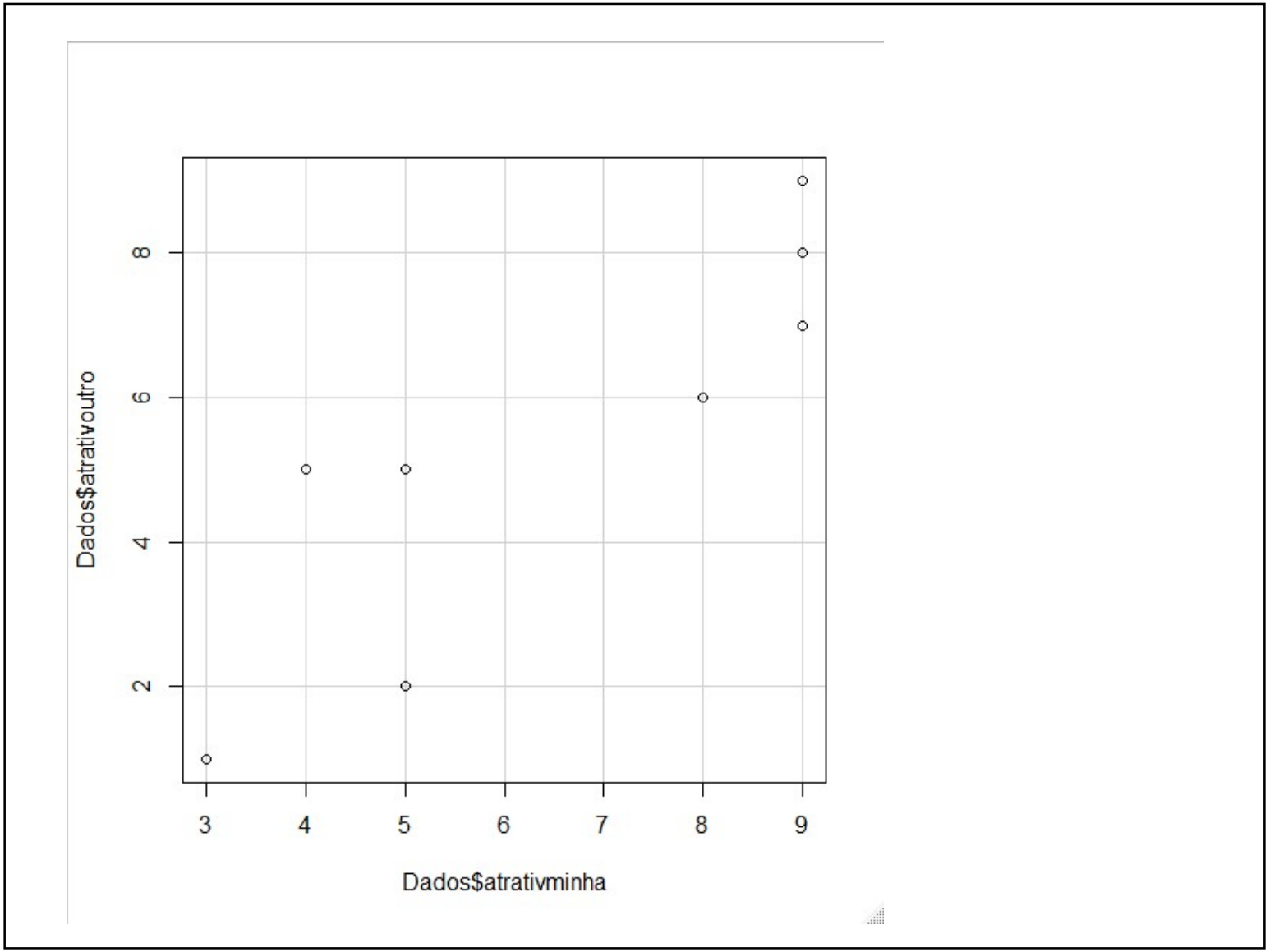
Filter

	Caso	atrativoutro 1=horroroso 10=maravilhoso	atrativminha	atrativoutro.posto	atrativminha.posto
1	1	7	9	6.0	7.5
2	2	5	4	3.5	2.0
3	3	5	5	3.5	3.5
4	4	8	9	7.5	7.5
5	5	9	9	9.0	7.5
6	6	6	8	5.0	5.0
7	7	1	3	1.0	1.0
8	8	2	5	2.0	3.5
9	9	8	9	7.5	7.5

SpearmanCorr_AvaliacaoAtratividade.R

```
library(haven)
library(car)
library(RVAideMemoire)
Dados <- haven::read_sav("Atratividade do outro e de si mesmo.sav")
plot(car::scatterplot(Dados$atrativoutro ~ Dados$atrativminha,
                  regLine=FALSE, smooth=FALSE, boxplots=TRUE,
                  jitter=list(x=0, y=0), col="black", data=Dados))
cor.test(Dados$atrativoutro, Dados$atrativminha,
         method = "spearm", exact=TRUE, na.rm=TRUE)
RVAideMemoire::spearman.ci(Dados$atrativoutro, Dados$atrativminha,
                           nrep = 1e6)
```

25



26

Correlação de Spearman

```
        Spearman's rank correlation rho

data:  Dados$atrativoutro and Dados$ativminha
S = 9.4285, p-value = 0.0004143
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.9214293
```

N = 9

A relação entre as atratividades própria e do outro é crescente.

```
        Spearman's rank correlation

data:  Dados$atrativoutro and Dados$ativminha
10000 replicates

95 percent confidence interval:
 0.6697248 0.9954023
sample estimates:
      rho
0.9214293
```

SpearmanCorr_AvaliacaoAtratividade.R
Comparison of several Spearman's rank correlation coefficients

```
set.seed(1510)
var1 <- c(1:15+rnorm(15,0,2),1:15+rnorm(15,0,2),1:15+rnorm(15,0,2))
var2 <- c(-1:-15+rnorm(15,0,2),1:15+rnorm(15,0,2),1:15+rnorm(15,0,2))
fact <- gl(3,15,labels=LETTERS[1:3])
Dados <- data.frame(var1, var2, fact)
RVAideMemoire::spearman.cor.multcomp(var1,var2,fact,nrep = 1e4)
# B and C similar but different from A

        Comparison of 3 Spearman's correlation coefficients

data:  var1 and var2 by fact
Bonferroni-adjusted  95 % confidence intervals
10000  replicates

      inf       r     sup
A -0.9847 -0.8464 -0.4128
B  0.2536  0.7893  0.9638
C  0.6410  0.9036  0.9856
```

29

SpearmanCorr_AvaliacaoAtratividade.R
Tests for (semi-)partial association/correlation between paired samples

```
set.seed(1444)
x <- 1:30
y <- 1:30+rnorm(30,0,2)
z1 <- runif(30,0,4)
z2 <- 30:1+rnorm(30,0,3)
RVAideMemoire::pcor.test(x,y,z1,method = "spearman",nrep = 1e4)
RVAideMemoire::pcor.test(x,y,list(z1,z2),method = "spearman",nrep = 1e4)

> RVAideMemoire::pcor.test(x,y,z1,method = "spearman",nrep = 1e4)

           Spearman's rank partial correlation

data:  x and y, controlling for z1
S = 134, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
95 percent confidence interval:
 0.8999330 0.9843505
sample estimates:
      rho
0.9701891

> RVAideMemoire::pcor.test(x,y,list(z1,z2),method = "spearman",nrep = 1e4)

           Spearman's rank partial correlation

data:  x and y, controlling for z1, z2
S = 1150, p-value = 5.729e-05
alternative hypothesis: true rho is not equal to 0
95 percent confidence interval:
 0.4965440 0.8732709
sample estimates:
      rho
0.7441602
```

30

Reamostragem (*bootstrapping*) da média amostral em R
Bootstrap_Media.R

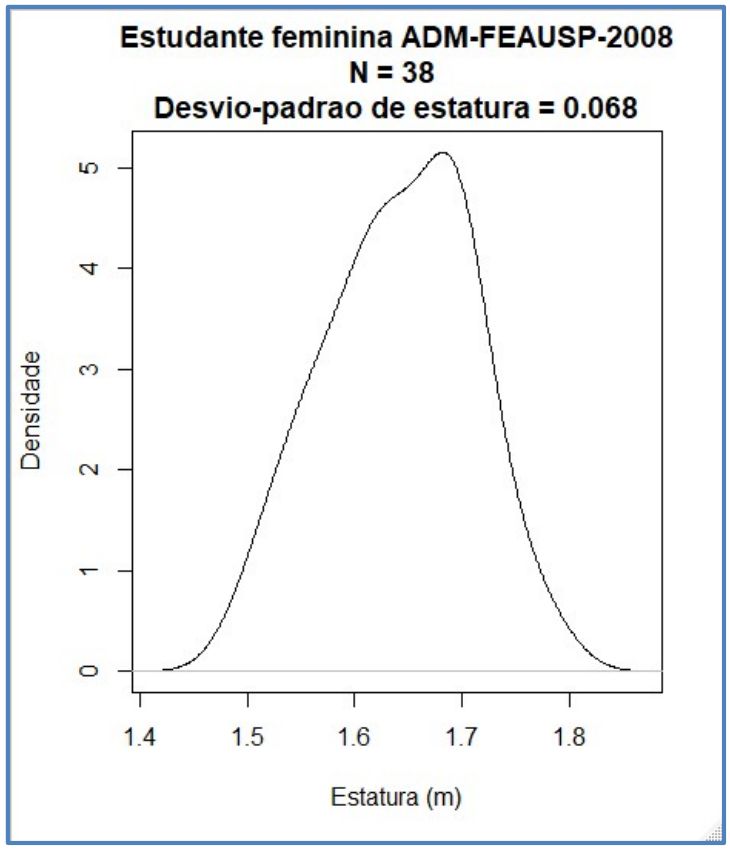
```
library(readxl)
B <- 1e6;  alfa <- 0.05; set.seed(123)
Dados <- readxl::read_excel("Adm2008.xlsx")
Dados <- Dados[, 1:4]
Matriz.Fem <- as.matrix(Dados[Dados$Genero=="Feminino", 3:4])
N.Fem <- nrow(Matriz.Fem)
plot(density(Matriz.Fem[,1], na.rm=TRUE),
     main=paste("Estudante feminina ADM-FEAUSP-2008\nN =", N.Fem,
                "\nDesvio-padrao de estatura =",
                round(sd(Matriz.Fem[,1], na.rm=TRUE), 4)),
     xlab="Estatura (m)", ylab="Densidade")
estat.media.boot.Fem <- replicate(B, mean(sample(Matriz.Fem[,1], replace=TRUE)))
print(mean(Matriz.Fem[,1], na.rm=TRUE))
print(mean(estat.media.boot.Fem, na.rm=TRUE))
quantile(estat.media.boot.Fem, probs=c(alfa/2, 1 - alfa/2))
t.test(Matriz.Fem[,1])$conf.int
plot(density(estat.media.boot.Fem, na.rm=TRUE),
     main=paste("Estudante feminina ADM-FEAUSP-2008\nN =", N.Fem,
                "\nErro-padrao da media amostral =",
                round(sd(estat.media.boot.Fem, na.rm=TRUE), 4)),
     xlab="Media amostral da estatura (m)", ylab="Densidade")
mi <- mean(estat.media.boot.Fem, na.rm=TRUE)
EP <- sd(estat.media.boot.Fem, na.rm=TRUE)
x <- seq(from=mi-5*EP, to=mi+5*EP, by=1e-5)
y <- dnorm(x, mean=mi, sd=EP)
lines(x,y,lwd=2,lty=2)
legend("topright", c("Media amostral","Normal"), lty=1:2, cex=.5)
Matriz.Masc <- as.matrix(Dados[Dados$Genero=="Masculino", 3:4])
N.Masc <- nrow(Matriz.Masc)
plot(density(Matriz.Masc[,1], na.rm=TRUE),
     main=paste("Estudante masculino ADM-FEAUSP-2008\nN =", N.Masc,
                "\nDesvio-padrao de estatura =",
                round(sd(Matriz.Masc[,1], na.rm=TRUE), 4)),
     xlab="Estatura (m)", ylab="Densidade")
estat.media.boot.Masc <- replicate(B, mean(sample(Matriz.Masc[,1], replace=TRUE)))
print(mean(Matriz.Masc[,1], na.rm=TRUE))
print(mean(estat.media.boot.Masc, na.rm=TRUE))
quantile(estat.media.boot.Masc, probs=c(alfa/2, 1 - alfa/2))
t.test(Matriz.Masc[,1])$conf.int
plot(density(estat.media.boot.Masc, na.rm=TRUE),
     main=paste("Estudante masculino ADM-FEAUSP-2008\nN =", N.Masc,
                "\nErro-padrao da media amostral =",
                round(sd(estat.media.boot.Masc, na.rm=TRUE), 4)),
     xlab="Media amostral da estatura (m)", ylab="Densidade")
mi <- mean(estat.media.boot.Masc, na.rm=TRUE)
EP <- sd(estat.media.boot.Masc, na.rm=TRUE)
x <- seq(from=mi-5*EP, to=mi+5*EP, by=1e-5)
y <- dnorm(x, mean=mi, sd=EP)
lines(x,y,lwd=2,lty=2)
legend("topright", c("Media amostral","Normal"), lty=1:2, cex=.5)
```

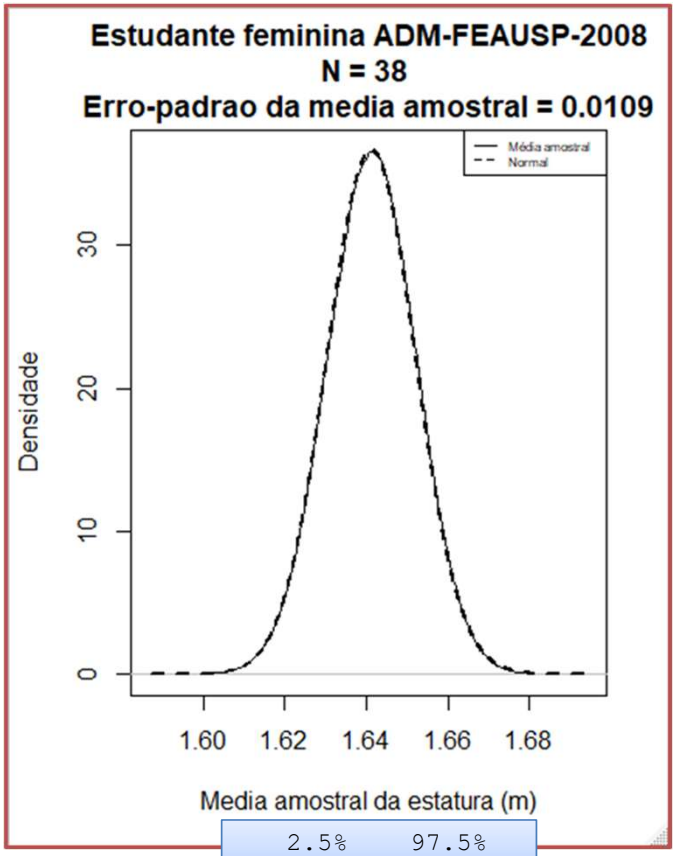
```
[1] 1.641316
> print(mean(Matriz.Fem[,1], na.rm=TRUE))
[1] 1.641316
> print(mean(estat.media.boot.Fem, na.rm=TRUE))
[1] 1.641323
> quantile(estat.media.boot.Fem, probs=c(alfa/2, 1 - alfa/2))
    2.5%    97.5%
1.620000 1.662368
> t.test(Matriz.Fem[,1])$conf.int
[1] 1.618968 1.663663
attr(,"conf.level")
[1] 0.95
```

```
> print(mean(Matriz.Masc[,1], na.rm=TRUE))
[1] 1.763529
> print(mean(estat.media.boot.Masc, na.rm=TRUE))
[1] 1.763543
> quantile(estat.media.boot.Masc, probs=c(alfa/2, 1 - alfa/2))
    2.5%    97.5%
1.741373 1.785294
> t.test(Matriz.Masc[,1])$conf.int
[1] 1.740813 1.786245
attr(,"conf.level")
[1] 0.95
```

31

Reamostragem (*bootstrapping*) da média amostral em R
Bootstrap_Media.R

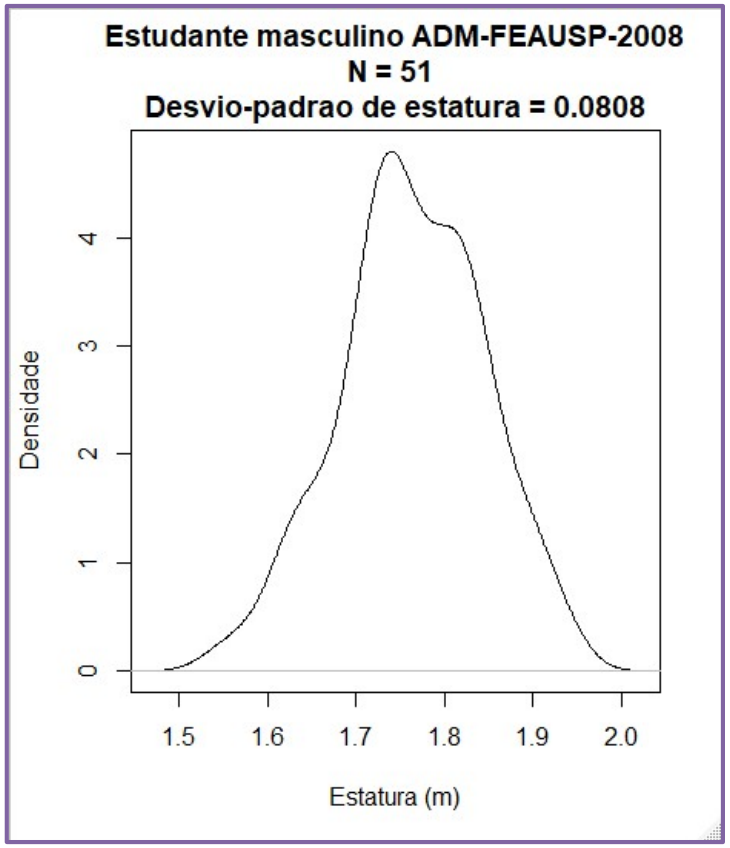


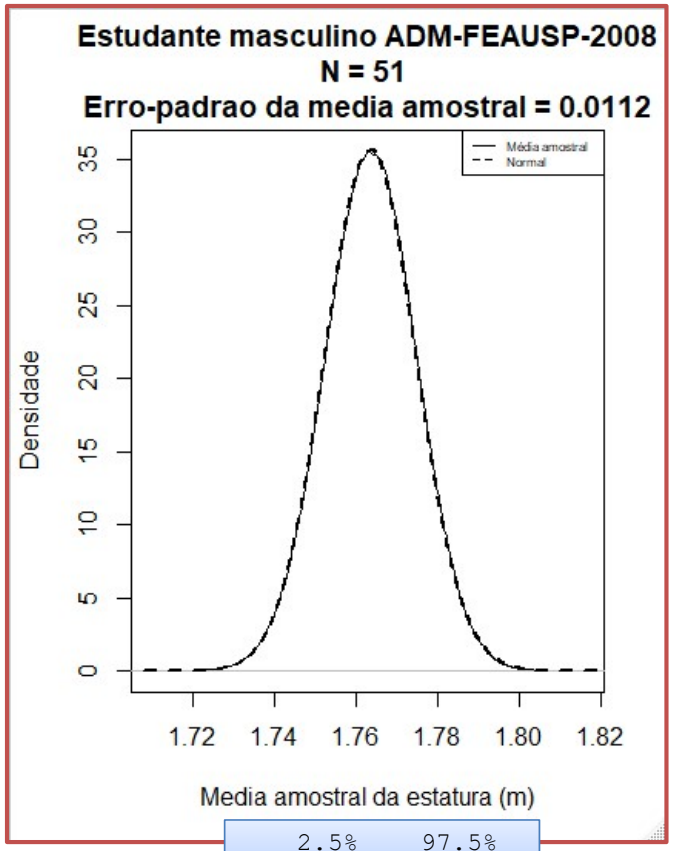



```
    2.5%    97.5%
1.620000 1.662368
```

32

Reamostragem (*bootstrapping*) da média amostral em R
Bootstrap_Media.R






```
    2.5%    97.5%
1.741373 1.785294
```

33

Reamostragem (*bootstrapping*) da mediana amostral em R
Bootstrap_Mediana.R

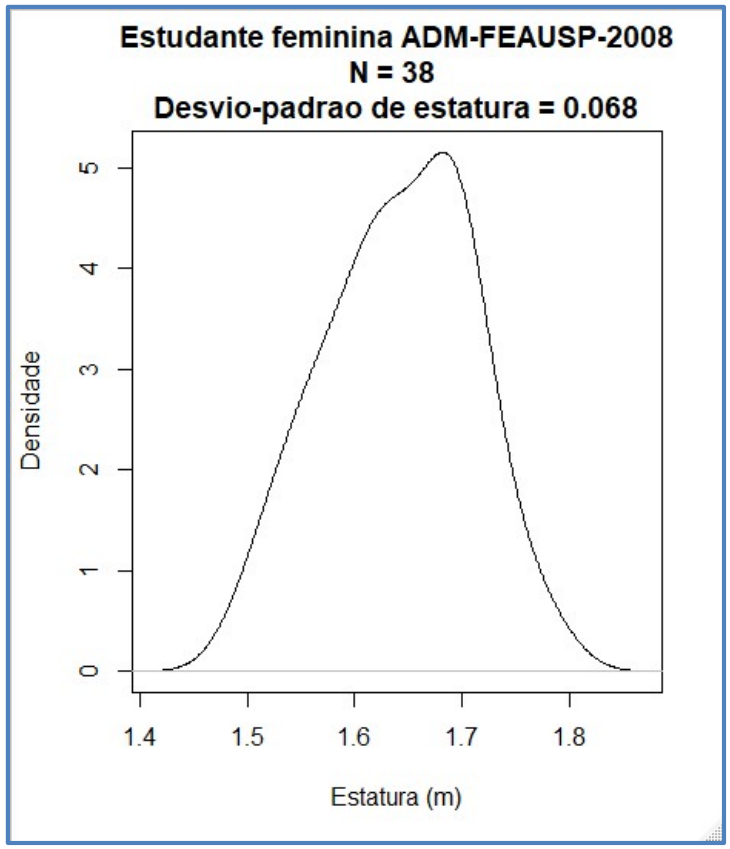
```
library(readxl)
B <- 1e6;  alfa <- 0.05; precisao <- 0.01/2; set.seed(123)
Dados <- readxl::read_excel("Adm2008.xlsx")
Dados <- Dados[, 1:4]
Matriz.Fem <- as.matrix(Dados[Dados$Genero=="Feminino", 3:4])
N.Fem <- nrow(Matriz.Fem)
plot(density(Matriz.Fem[,1], na.rm=TRUE),
     main=paste("Estudante feminina ADM-FEAUSP-2008\nN =", N.Fem,
                "\nDesvio-padrao de estatura =",
                round(sd(Matriz.Fem[,1], na.rm=TRUE), 4)),
     xlab="Estatura (m)", ylab="Densidade")
estat.media.boot.Fem <- replicate(B, median(sample(Matriz.Fem[,1], replace=TRUE)))
estat.media.boot.Fem <- estat.media.boot.Fem+runif(B,-precisao,precisao)
print(median(Matriz.Fem[,1], na.rm=TRUE))
print(mean(estat.media.boot.Fem, na.rm=TRUE))
quantile(estat.media.boot.Fem, probs=c(alfa/2, 1 - alfa/2))
wilcox.test(Matriz.Fem[,1], exact=FALSE, conf.int=TRUE)$conf.int
plot(density(estat.media.boot.Fem, na.rm=TRUE),
     main=paste("Estudante feminina ADM-FEAUSP-2008\nN =", N.Fem,
                "\nErro-padrao da mediana amostral =",
                round(sd(estat.media.boot.Fem, na.rm=TRUE), 4)),
     xlab="Mediana amostral da estatura (m)", ylab="Densidade")
mi <- mean(estat.media.boot.Fem, na.rm=TRUE)
EP <- sd(estat.media.boot.Fem, na.rm=TRUE)
x <- seq(from=mi-5*EP, to=mi+5*EP, by=1e-5)
y <- dnorm(x, mean=mi, sd=EP)
lines(x,y,lwd=2,lty=2)
legend("topright", c("Mediana amostral","Normal"), lty=1:2, cex=.5)
Matriz.Masc <- as.matrix(Dados[Dados$Genero=="Masculino", 3:4])
N.Masc <- nrow(Matriz.Masc)
plot(density(Matriz.Masc[,1], na.rm=TRUE),
     main=paste("Estudante masculino ADM-FEAUSP-2008\nN =", N.Masc,
                "\nDesvio-padrao de estatura =",
                round(sd(Matriz.Masc[,1], na.rm=TRUE), 4)),
     xlab="Estatura (m)", ylab="Densidade")
estat.media.boot.Masc <- replicate(B, median(sample(Matriz.Masc[,1], replace=TRUE)))
estat.media.boot.Masc <- estat.media.boot.Masc+runif(B,-precisao,precisao)
print(median(Matriz.Masc[,1], na.rm=TRUE))
print(mean(estat.media.boot.Masc, na.rm=TRUE))
quantile(estat.media.boot.Masc, probs=c(alfa/2, 1 - alfa/2))
wilcox.test(Matriz.Masc[,1], exact=FALSE, conf.int=TRUE)$conf.int
plot(density(estat.media.boot.Masc, na.rm=TRUE),
     main=paste("Estudante masculino ADM-FEAUSP-2008\nN =", N.Masc,
                "\nErro-padrao da mediana amostral =",
                round(sd(estat.media.boot.Masc, na.rm=TRUE), 4)),
     xlab="Mediana amostral da estatura (m)", ylab="Densidade")
mi <- mean(estat.media.boot.Masc, na.rm=TRUE)
EP <- sd(estat.media.boot.Masc, na.rm=TRUE)
x <- seq(from=mi-5*EP, to=mi+5*EP, by=1e-5)
y <- dnorm(x, mean=mi, sd=EP)
lines(x,y,lwd=2,lty=2)
legend("topright", c("Mediana amostral","Normal"), lty=1:2, cex=.5)
```

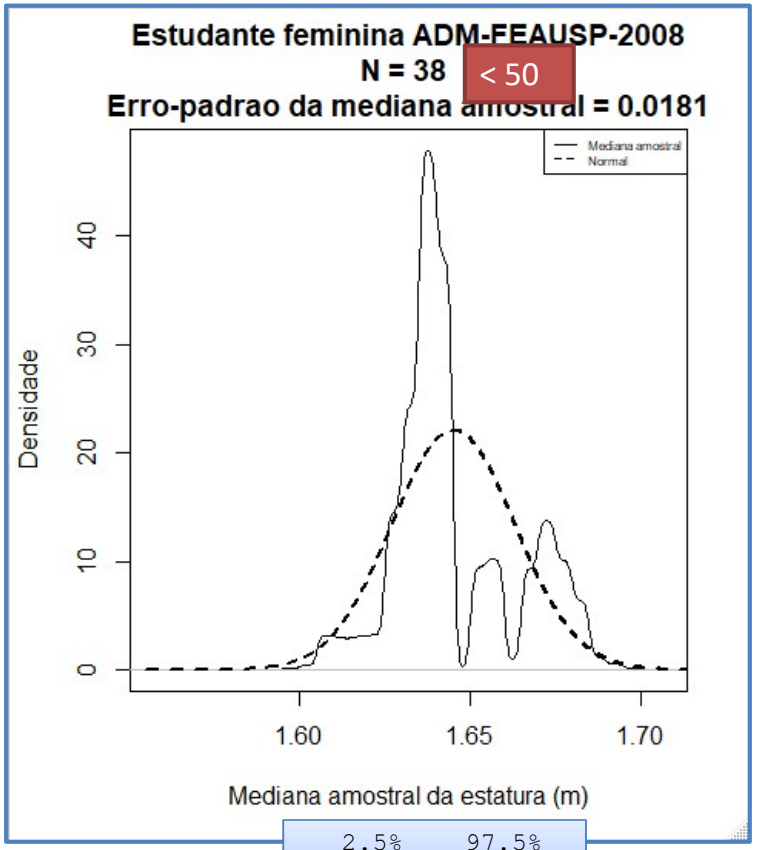
```
> print(median(Matriz.Fem[,1]
[1] 1.64
> print(mean(estat.media.boot
[1] 1.645207
> quantile(estat.media.boot.
    2.5%    97.5%
1.611974 1.682547
> wilcox.test(Matriz.Fem[,1]
[1] 1.619984 1.665002
attr(,"conf.level")
[1] 0.95
```

```
> print(median(Matriz.Masc[,1]
[1] 1.76
> print(mean(estat.media.boot.
[1] 1.760086
> quantile(estat.media.boot.Ma
    2.5%    97.5%
1.732912 1.797664
> wilcox.test(Matriz.Masc[,1],
[1] 1.739951 1.785059
attr(,"conf.level")
[1] 0.95
```

34

Reamostragem (*bootstrapping*) da mediana amostral em R
Bootstrap_Mediana.R

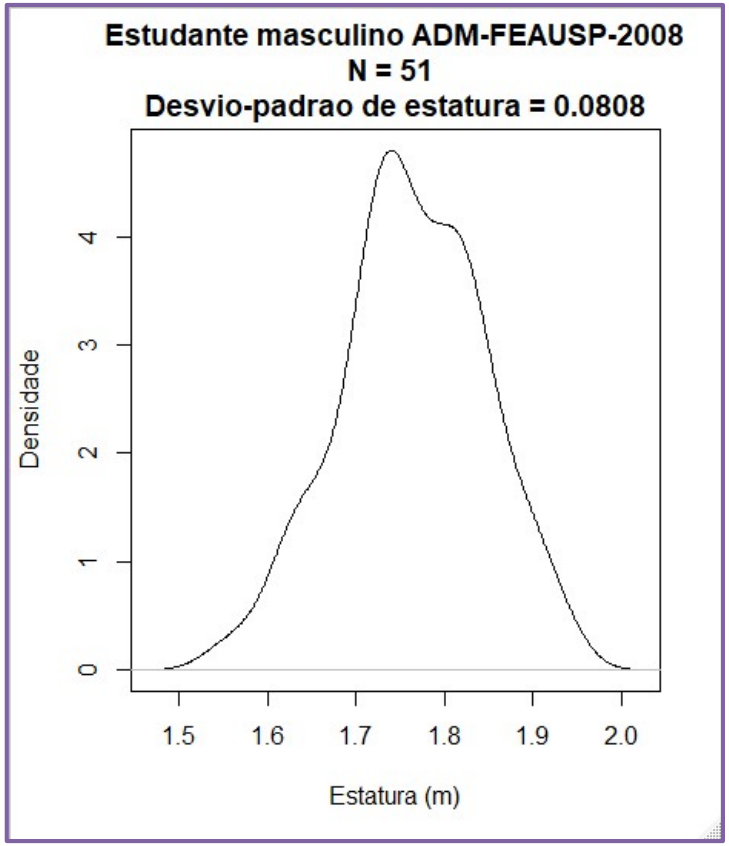


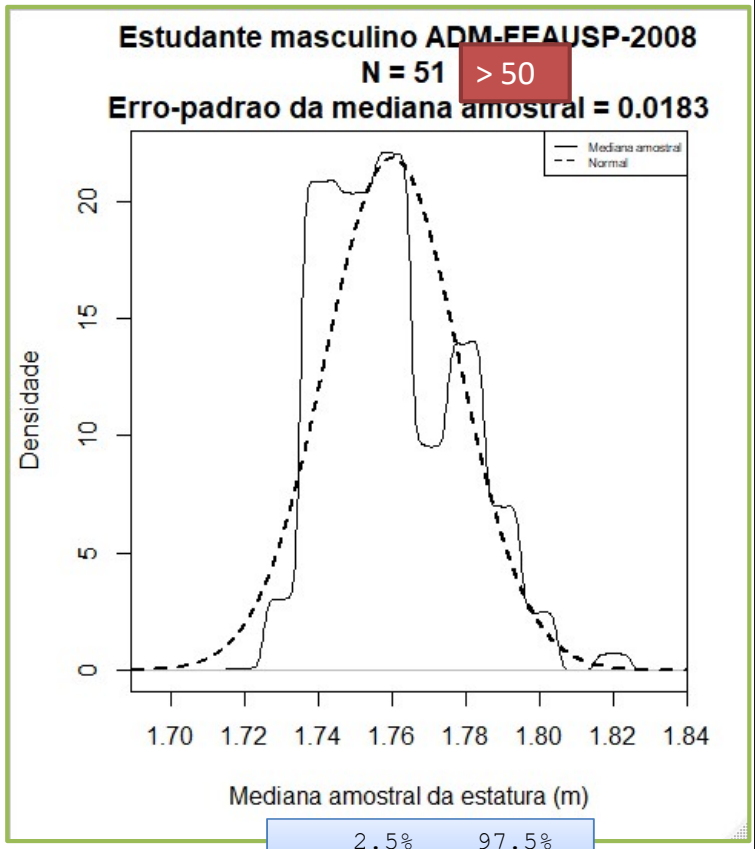



```
   2.5%    97.5%
1.611974 1.682547
```

35

Reamostragem (*bootstrapping*) da mediana amostral em R
Bootstrap_Mediana.R






```
     2.5%    97.5%
1.732912 1.797664
```

Teste do sinal para mediana populacional

- O teste usa o sinal da diferença não nula – ignorando sua magnitude – entre o valor observado da variável ordinal ou intervalar X e a mediana populacional hipotetizada ν_0
- Suposição: a variável ordinal ou intervalar X tem uma distribuição desconhecida
- Teste bilateral

$H_0: P(X > \nu_0) = 0{,}5$
$H_1: P(X > \nu_0) \neq 0{,}5$

37

Duração da consulta

- Estudos recentes de consultórios particulares de médicos que não atendem pacientes do SUS sugeriram que a duração *mediana* de cada consulta do paciente era de 22 minutos.
- Acredita-se que o tempo *mediano* de consultas em práticas com uma grande carga de SUS seja menor que 22 minutos.
- Uma amostra aleatória de 20 consultas em consultórios com uma grande carga do SUS produziu, em ordem, os seguintes tempos de consulta
  - 9.4 13.4 15.6 16.2 16.4 16.8 18.1 18.7 18.9 19.1
  - 19.3 20.1 20.4 21.6 21.9 23.4 23.5 24.8 24.9 26.8
- Com base nesses dados, existem evidências suficientes para concluir que a duração *mediana* da consulta em consultórios com uma grande carga de SUS é menor que 22 minutos?

Resposta

- Teste do sinal: Para alfa = 5%, sim. Para alfa = 1%, não.
- Teste de Wilcoxon: Para alfa = 5%, sim. Para alfa = 1%, sim.

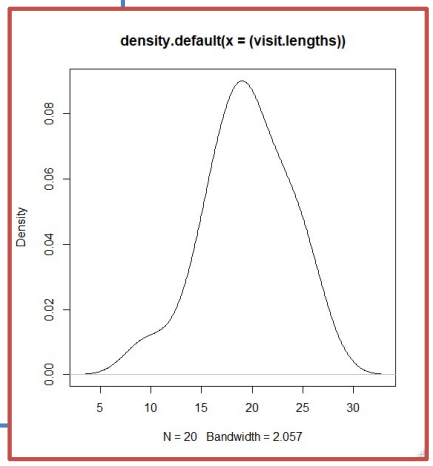
https://onlinecourses.science.psu.edu/stat414/node/318/

38

Teste do sinal para uma condição
Testes_Sinal&Mediana_1grupo.R

```
library(DescTools)
visit.lengths <- c(9.4, 13.4, 15.6, 16.2, 16.4, 16.8, 18.1, 18.7, 18.9, 19.1,
                   19.3, 20.1, 20.4, 21.6, 21.9, 23.4, 23.5, 24.8, 24.9, 26.8)
plot(density((visit.lengths)))
table(visit.lengths)
cat("N = ", sum(!is.na(visit.lengths)), sep="")
summary(visit.lengths)
DescTools::SignTest(visit.lengths, mu = 22, alternative = "less")
```

```
visit.lengths
 9.4 13.4 15.6 16.2 16.4 16.8 18.1 18.7 18.9 19.1 19.3 20.1 20.4 21.6 21.9
   1    1    1    1    1    1    1    1    1    1    1    1    1    1    1
23.4 23.5 24.8 24.9 26.8
   1    1    1    1    1
N = 20
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   9.40   16.70   19.20   19.46   22.27   26.80

   One-sample Sign-Test

data:  visit.lengths
S = 5, number of differences = 20, p-value = 0.02069
alternative hypothesis: true median is less than 22
97.9 percent confidence interval:
 -Inf 21.6
sample estimates:
median of the differences
                     19.2
```



39

Pressão arterial

- Um estudo é realizado para determinar os efeitos da remoção de um bloqueio renal em pacientes cuja função renal está comprometida devido a malignidade metastática avançada de causa não-urológica.
- A pressão arterial de uma amostra aleatória de 10 pacientes é medida antes e após a cirurgia para o tratamento do bloqueio, produzindo os seguintes dados:

Row	before	after	diff
1	150	90	60
2	132	102	30
3	130	80	50
4	116	82	34
5	107	90	17
6	100	94	6
7	101	84	17
8	96	93	3
9	90	90	0
10	78	80	-2

- Podemos concluir que a cirurgia tende a baixar a pressão arterial?

Resposta

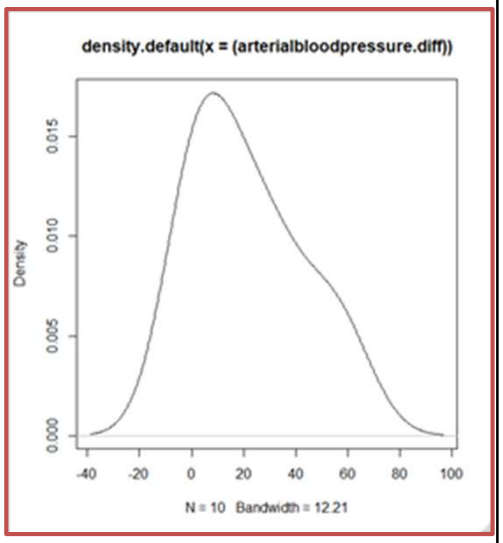
- Teste do sinal: Para alfa = 5%, sim. Para alfa = 1%, não.
- Teste de Wilcoxon: Para alfa = 5%, sim. Para alfa = 1%, sim.

40

Teste do sinal para uma condição
Testes_Sinal&Mediana_1grupo.R

```
library(DescTools)
arterialbloodpressure.diff <- c(60, 30, 50, 34, 17, 6, 17, 3, 0, -2)
plot(density((arterialbloodpressure.diff)))
table(arterialbloodpressure.diff)
cat("N = ", sum(!is.na(arterialbloodpressure.diff)), "\n", sep="")
summary(arterialbloodpressure.diff)
DescTools::SignTest(arterialbloodpressure.diff, mu = 0, alternative = "greater")
```

```
arterialbloodpressure.diff
-2  0  3  6 17 30 34 50 60
 1  1  1  1  2  1  1  1  1
N = 10
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  -2.00    3.75   17.00   21.50   33.00   60.00

   One-sample Sign-Test

data:  arterialbloodpressure.diff
S = 8, number of differences = 9, p-value = 0.01953
alternative hypothesis: true median is greater than 0
98.9 percent confidence interval:
   0 Inf
sample estimates:
median of the differences
                       17
```



41

Nonparametric tests

Nonparametric tests robustly compare skewed or ranked data.

NATURE METHODS | VOL.11 NO.5 | MAY 2014

Teste do sinal para uma mediana populacional

Martin Krzywinski & Naomi Altman

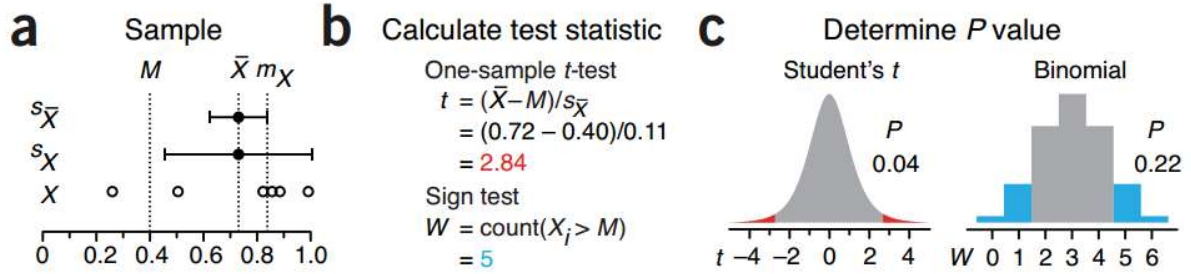


Figure 1 | A sample can be easily tested against a reference value using the sign test without any assumptions about the population distribution. (**a**) Sample X ($n = 6$) is tested against a reference $M = 0.4$. Sample mean $\bar{X}$ is shown with s.d. (s_X) and s.e.m. error bars ($s_{\bar{X}}$). m_x is sample median. (**b**) The t-statistic compares $\bar{X}$ to M in units of s.e.m. The sign test's W is the number of sample values larger than M. (**c**) Under the null, t follows Student's t-distribution with five degrees of freedom, whereas W is described by the binomial with 6 trials and $P = 0.5$. Two-tailed P values are shown.

Prob. de sucesso

Effect Size Statistics for the McNemar, Sign, and Wilcoxon Tests

SPSS does not report an effect size index for these three tests. However, simple indices can be reported for each of these tests to communicate the size of the effect.

For the McNemar test, a good effect index is the difference in the proportion of individuals who fall into one of the two categories on the one occasion (or condition) versus the proportion of individuals who fall into the same category on the other occasion (or other condition). For our data, the two proportions being compared are the proportion of employees who are extremely concerned about job pay, which is .50 = (7 + 8)/30, and the proportion of employees who are extremely concerned about job security, which is .37 = (4 + 7)/30. Consequently, the difference in proportions is .50 − .37 = .13.

For the sign test, the proportion of individuals who had a positive (or negative) difference score in comparison with negative (or positive) differences can be reported. For example, the proportion of employees who showed greater concern for job pay in comparison with greater concern for job security of .77 = 20/(20 + 6) could be reported. As an alternative, you could report the proportion of employees who showed greater concern for job security in comparison with greater concern for job pay of .23 = 6/(20 + 6).

For the Wilcoxon test, the mean positive ranked difference score and the mean negative ranked difference score could be reported.

Using SPSS for Windows and Macintosh - analyzing and understanding - 7e - Green & Salkind - 2014

43

Let's close up our discussion of the sign test by taking a note of the following:

(1) The sign test was presented here as a test for the median H_0: $m = m_0$, but if we were to make the additional assumption that the distribution of X is symmetric, then the sign test is also a valid test for the mean H_0: $\mu = \mu_0$.

(2) A primary advantage of the sign test is that by using only the signs of the $X_i - m_0$ quantities, the test completely obliterates the negative effect of outliers.

(3) A primary disadvantage of the sign test is that by using only the signs of the $X_i - m_0$ quantities, we potentially lose useful information about the magnitudes of the $X_i - m_0$ quantities. For example, which data have more evidence against the null hypothesis H_0: $m = m_0$? (1, 1, −1) versus (5, 6, −1)? Or (1, 1, −1) versus (10000, 6, −1)?

That last point suggests that we might want to also consider a test that takes into account the magntitudes of the $X_i - m_0$ quantities. That's exactly what the Wilcoxon signed rank test does. Let's go check it out.

Teste de Wilcoxon para uma mediana populacional

- Se a VD é quantitativa com distribuição normal, o teste z ou t para uma condição é o mais poderoso para comparar a média populacional com uma constante. Nesse caso, a média é igual à mediana
- Porém, se a variável é intervalar com distribuição populacional desconhecida e tamanho da amostra menor que 30, a medida de tendência central mais adequada é a mediana
- Se a distribuição da VD é simétrica, o teste é também de média populacional
- O teste usa o sinal e a magnitude do posto da diferença não nula entre o valor da variável e a mediana populacional hipotetizada

45

Teste de Wilcoxon para *mediana* populacional

The Wilcoxon–Mann–Whitney test under scrutiny

Morten W. Fagerland*, † and Leiv Sandvik

STATISTICS IN MEDICINE
Statist. Med. 2009; **28**:1487–1497

(i) H_0: equal population means *versus* H_1: unequal population means.
(ii) H_0: equal population medians *versus* H_1: unequal population medians.

For symmetric distributions, (i) and (ii) are equivalent, but for asymmetric distributions, different results from using (i) and (ii) were expected.

46

Teste de Wilcoxon para uma mediana populacional
Testes_Sinal&Mediana_1grupo.R

```
library(DescTools)
visit.lengths <- c(9.4, 13.4, 15.6, 16.2, 16.4, 16.8, 18.1, 18.7, 18.9, 19.1,
                   19.3, 20.1, 20.4, 21.6, 21.9, 23.4, 23.5, 24.8, 24.9, 26.8)
table(visit.lengths)
cat("N = ", sum(!is.na(visit.lengths)), sep="")
summary(visit.lengths)
wilcox.test(visit.lengths, mu = 22, alternative = "less",
            conf.int = TRUE, exact=TRUE, na.rm=TRUE)
```

```
        Wilcoxon signed rank test with continuity correction

data:  visit.lengths
V = 38.5, p-value = 0.006867
alternative hypothesis: true location is less than 22
95 percent confidence interval:
     -Inf 21.20002
sample estimates:
(pseudo)median
      19.55002
```

47

Teste de Wilcoxon para uma mediana populacional
Testes_Sinal&Mediana_1grupo.R

```
library(DescTools)
arterialbloodpressure.diff <- c(60, 30, 50, 34, 17, 6, 17, 3, 0, -2)
table(arterialbloodpressure.diff)
cat("N = ", sum(!is.na(arterialbloodpressure.diff)), "\n", sep="")
summary(arterialbloodpressure.diff)
wilcox.test(arterialbloodpressure.diff, mu = 0, alternative = "greater",
            conf.int = TRUE, exact=TRUE, na.rm=TRUE)
```

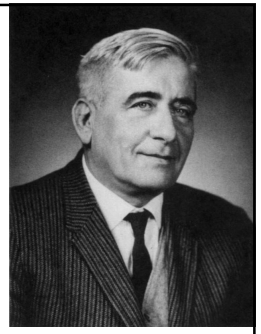
```
   Wilcoxon signed rank test with continuity correction

data:  arterialbloodpressure.diff
V = 44, p-value = 0.006386
alternative hypothesis: true location is greater than 0
95 percent confidence interval:
 9.999971      Inf
sample estimates:
(pseudo)median
      23.50004
```

48

Alternativas ao teste t para duas condições: U de Mann-Whitney (Generalizado) e W de Wilcoxon

- Os testes Mann-Whitney e Wilcoxon avaliam se existe uma diferença estatisticamente significante entre as suas condições
- O teste U de Mann-Whitney é usado se há condições independentes (Mann-Whitney-Wilcoxon (MWW ou WMW))
- O teste W de Wilcoxon é usado se há os mesmos participantes emparelhados nas duas condições
- Ambos os testes precisam que os escores de duas condições sejam pelo menos ordinais a fim de que o teste estatístico seja calculado a partir dos postos
- Se não existe diferença entre as duas condições, esperaríamos encontrar um número semelhante de postos altos e baixos em cada condição; especificamente, se somarmos os postos, esperaríamos que a soma total dos postos em cada condição fosse aproximadamente a mesma

Teste U de Mann-Whitney



Henry B. Mann (1870-1940)

- Teste não-paramétrico adequado para comparar as distribuições populacionais da VD pelo menos ordinal em duas condições independentes

50

Teste de Wilcoxon para duas *medianas* populacionais

The Wilcoxon–Mann–Whitney test under scrutiny

Morten W. Fagerland*, † and Leiv Sandvik

STATISTICS IN MEDICINE
Statist. Med. 2009; **28**:1487–1497

(i) H_0: equal population means *versus* H_1: unequal population means.
(ii) H_0: equal population medians *versus* H_1: unequal population medians.

For symmetric distributions, (i) and (ii) are equivalent, but for asymmetric distributions, different results from using (i) and (ii) were expected.

51

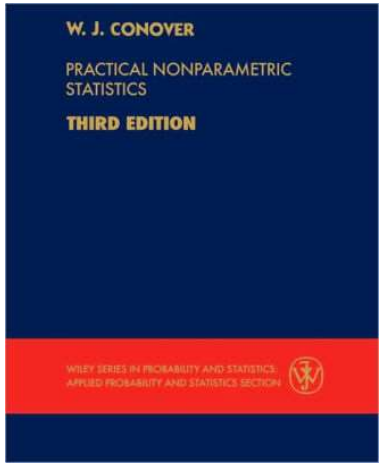
The Mann-Whitney Test

Data The data consist of two random samples. Let $X_1, X_2, \ldots, X_n$ denote the random sample of size n from population 1 and let $Y_1, Y_2, \ldots, Y_m$ denote the random sample of size m from population 2. Assign the ranks 1 to $n + m$ to the observations from smallest to largest. Let $R(X_i)$ and $R(Y_j)$ denote the rank assigned to X_i and Y_j for all i and j. For convenience, let $N = n + m$.

If several sample values are exactly equal to each other (tied), assign to each the average of the ranks that would have been assigned to them had there been no ties (see Example 1).

Assumptions

1. Both samples are random samples from their respective populations.
2. In addition to independence within each sample, there is mutual independence between the two samples.
3. The measurement scale is at least ordinal.



Comment

The Mann-Whitney test is unbiased and consistent when testing the preceding hypotheses involving $P(X > Y)$. However, the same is not always true for the hypotheses involving $E(X)$ and $E(Y)$. To insure that the test remains consistent and unbiased for hypotheses involving $E(X)$ it is sufficient to add another assumption to the previous model.

Assumption 4. If there is a difference between population distribution functions, that difference is a difference in the location of the distributions. That is, if $F(x)$ is not identical with $G(x)$, then $F(x)$ is identical with $G(x + c)$, where c is some constant.

52

Hypotheses

A. (Two-Tailed Test) Let $F(x)$ and $G(x)$ be the distribution functions corresponding to X and Y, respectively. Then the hypotheses may be stated as follows.

$$H_0: F(x) = G(x) \qquad \text{for all } x$$
$$H_1: F(x) \neq G(x) \qquad \text{for some } x$$

The test is sensitive for $H_1: E(X) \neq E(Y)$, and can be used as a test for means. In many real situations any difference between distributions implies that $P(X > Y)$ is no longer equal to $1/2$. Therefore $H_1: P(X > Y) \neq P(X < Y)$ is often used instead of the above.

Reject H_0 at the level of significance α if T is less than its $\alpha/2$ quantile or greater than its $1 - \alpha/2$ quantile obtained from Table A7 and Equation 3 for $n \leq 20$ and $m \leq 20$, or obtained from Table A1 and Equation 5 in the large sample approximation. If T_1 is used instead of T the quantiles are obtained directly from Table A1.

The approximate two-tailed p-value may be found using Table A1. In the case of T, substitute the smaller of T or T' into the following

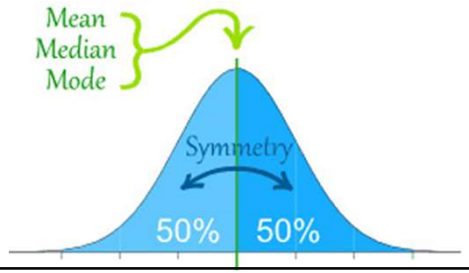
$$p\text{-value} = 2 \cdot P\left(Z \leq \frac{T \text{ (or } T') + \frac{1}{2} - n\frac{N+1}{2}}{\sqrt{\frac{nm(N+1)}{12}}} \right) \tag{6}$$

where Z is a standard normal random variable. In the case of T_1 the p-value is twice the smaller of $P(Z \leq T_1)$ or $P(Z \geq T_1)$.

53

Teste U de Mann-Whitney

- As hipóteses nula e alternativa podem ser enunciadas em termos de medianas populacionais se as funções de distribuição da VD nas duas condições têm formatos idênticos:
  - Se elas têm formatos idênticos, as medianas populacionais são iguais, mas se as medianas populacionais são iguais, não implica necessariamente que as distribuições têm formatos idênticos
- Na prática, é possível ter a VD em duas condições independentes com a mesma mediana amostral e rejeitar H_0 com o teste U de MW



54

Mann-Whitney test is not just a test of medians: differences in spread can be important

Anna Hart BMJ VOLUME 323 18 AUGUST 2001 British Medical Journal

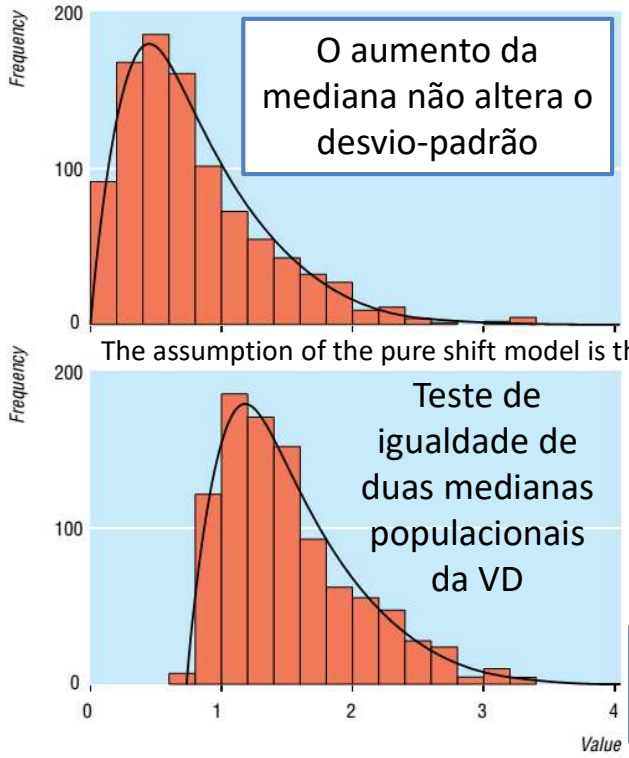


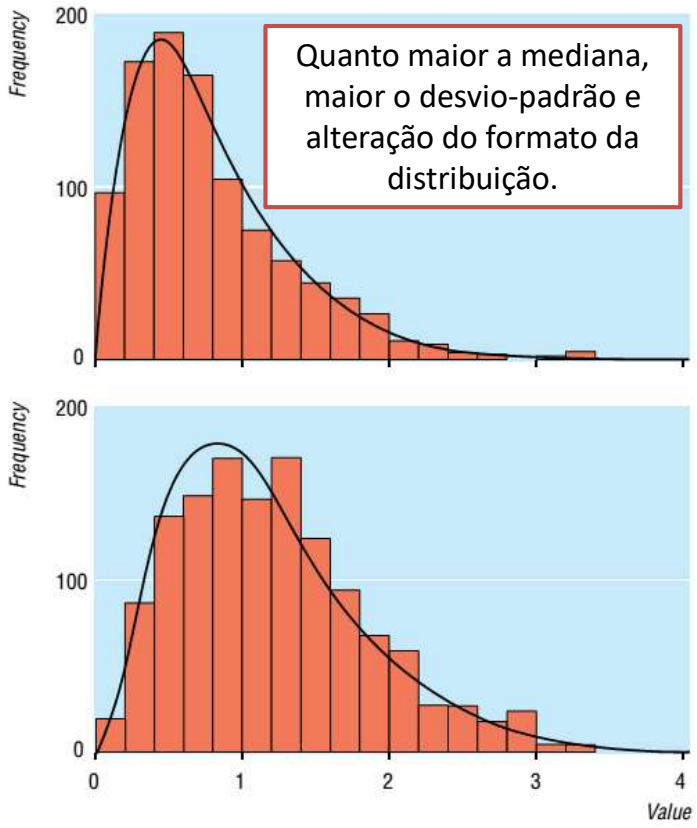
Fig 1 Two distributions with a difference in median but no difference in shape and spread

The Mann-Whitney (or Wilcoxon-Mann-Whitney) test is sometimes used for comparing the efficacy of two treatments in clinical trials. It is often presented as an alternative to a *t* test when the data are not normally distributed. Whereas a *t* test is a test of population means, the Mann-Whitney test is commonly regarded as a test of population medians. This is not strictly true, and treating it as such can lead to inadequate analysis of data.

The assumption of the pure shift model is thus often unrealistic. Fagerland & Sandvik (2009)

the Mann-Whitney test is "a two-sample rank test for the difference between two population medians … It assumes that the data are independent random samples from two populations that have the same shape." Figure 1 shows two distributions for which this is the case. One distribution is shifted 0.75 units to the right: the medians differ by 0.75 units but the shapes are identical.

55



Quanto maior a mediana, maior o desvio-padrão e alteração do formato da distribuição.

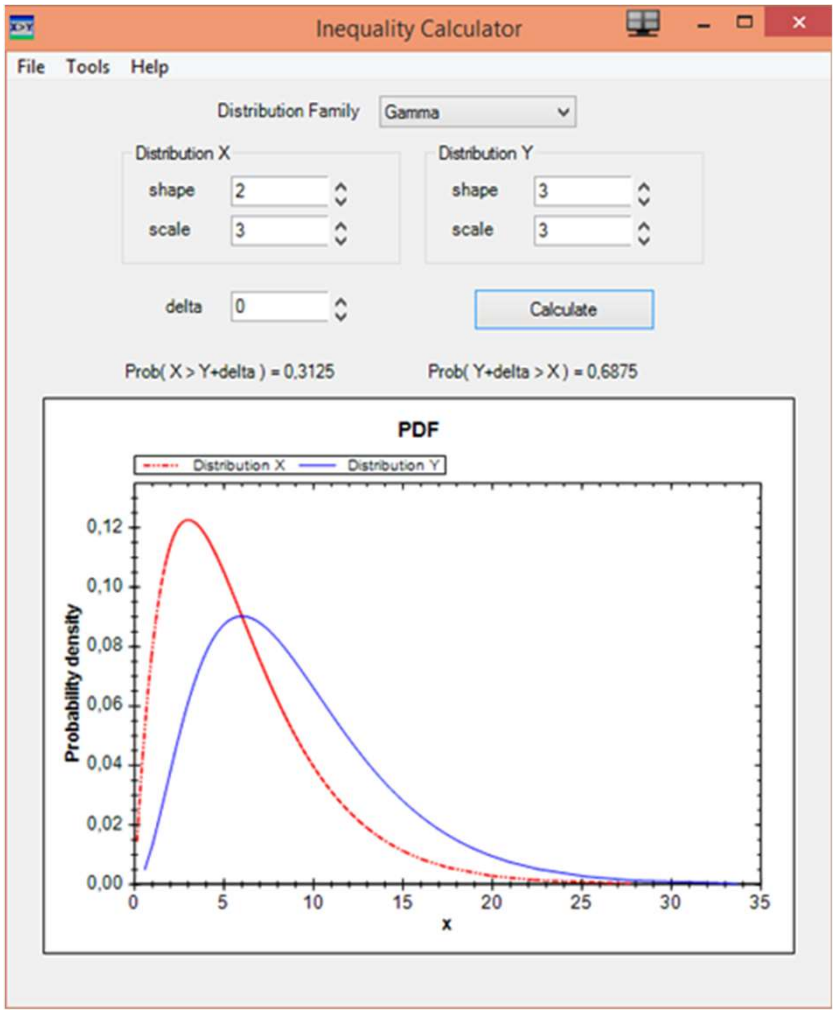
Fig 2 Two distributions with different medians and different shapes. The distribution with the larger median also has a greater spread

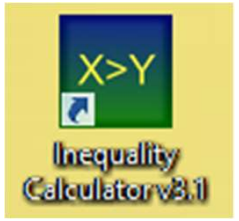
Differences in population medians are often accompanied by other differences in spread and shape. Moreover, the difference in medians may not be the most striking or the most clinically important difference. It is important to look at distributional differences and discuss them. Figure 2 shows an example in which the median values are 0.65 and 1.14 units. The distribution with the larger median also has larger spread. The spread is shown clearly in figure 3 , which shows box plots of samples of 25 drawn from these two distributions. (The P value from the Mann-Whitney test is 0.02.) If the difference is assumed to be merely a difference in medians other clinically important information could be ignored.

Recommendations

Researchers should take care to describe their data and to be clear about the features that are most clinically important. They should use the statistical test that is most relevant for their hypotheses, and describe the features of the data that are likely to have caused a hypothesis to be rejected. As is always the case, it is not sufficient merely to report a P value. In the case of the Mann-Whitney test, differences in spread may sometimes be as clinically important as differences in medians, and these need to be made clear to the reader.

56





Gama	X	Y
a=	2	3
b=	3	3
media=	6,0000	9,0000
mediana=	5,0323	8,0217
variancia=	18,0000	27,0000
assimetria=	1,4142	1,1547

57

Summary points

The Mann-Whitney test is used as an alternative to a *t* test when the data are not normally distributed

The test can detect differences in shape and spread as well as just differences in medians

Differences in population medians are often accompanied by equally important differences in shape

Researchers should describe the clinically important features of data and not just quote a P value

58

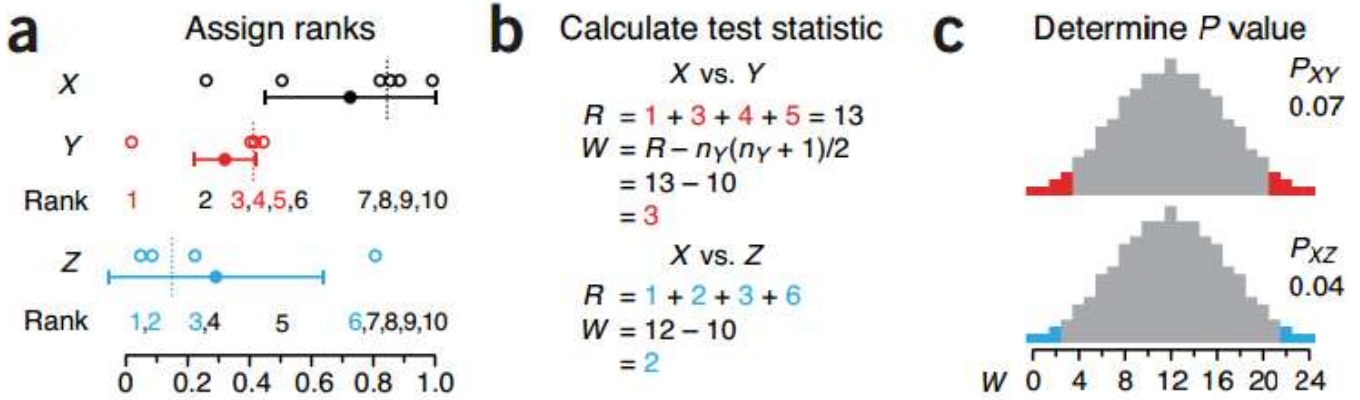


Figure 2 | Many nonparametric tests are based on ranks. (**a**) Sample comparisons of *X* vs. *Y* and *X* vs. *Z* start with ranking pooled values and identifying the ranks in the smaller-sized sample (e.g., 1, 3, 4, 5 for *Y*; 1, 2, 3, 6 for *Z*). Error bars show sample mean and s.d., and sample medians are shown by vertical dotted lines. (**b**) The Wilcoxon rank-sum test statistic *W* is the difference between the sum of ranks and the smallest possible observed sum. (**c**) For small sample sizes the exact distribution of *W* can be calculated. For samples of size (6, 4), there are only 210 different rank combinations corresponding to 25 distinct values of *W*.

Nonparametric tests - Krzywinski & Altman - 2014

59

Because there is a finite number (210) of combinations of rank-ordering for X ($n_x = 6$) and Y ($n_Y = 4$), we can enumerate all outcomes of the test and explicitly construct the distribution of W (**Fig. 2c**) to assign a P value to W. The smallest value of $W = 0$ occurs when all values in one sample are smaller than those in the other. When they are all larger, the statistic reaches a maximum, $W = n_X n_Y = 24$. For X versus Y, $W = 3$, and there are 14 of 210 test outcomes with $W \leq 3$ or $W \geq 21$. Thus, $P_{XY} = 14/210 = 0.067$. For X versus Z, $W = 2$, and $P_{XZ} = 8/210 = 0.038$. For cases in which both samples are larger than 10, W is approximately normal, and we can obtain the P value from a z-test of $(W - \mu_W)/\sigma_W$, where $\mu_W = n_1(n_1 + n_2 + 1)/2$ and $\sigma_W = \sqrt{(\mu_W n_2/6)}$.

Nonparametric tests - Krzywinski & Altman - 2014

60

Should we always choose a nonparametric test when comparing two apparently nonnormal distributions?

Eva Skovlund[a,*], Grete U. Fenstad[b]

[a]*Norwegian Cancer Society and Section of Medical Statistics, University of Oslo, Oslo, Norway*
[b]*Department of Mathematics, University of Oslo, Oslo, Norway*

Journal of Clinical Epidemiology 54 (2001) 86–92

Abstract

When clinical data are subjected to statistical analysis, a common question is how to choose an appropriate significance test. Comparing two independent groups with observations measured on a continuous scale, the question is typically whether to choose the two-sample-*t* test or the Wilcoxon–Mann–Whitney test (WMW test). Similar results are often obtained, but which conclusion can be drawn if significance tests give highly different *P*-values? The *t* test is optimal for normally distributed observations with common variance and robust to deviations from normality if sample sizes are not very small. The WMW test makes no distributional assumptions, but depends heavily on equal shape and variance of the two distributions (homoscedasticity). We have compared the properties of the traditional two-sample *t* test, a modified *t* test allowing unequal variance, and the WMW test by stochastic simulation. All show acceptable behaviour when the two distributions have similar variance. When variances differ, the modified *t* test is superior to the other two.

Keywords: *P*-Values; Two-sample *t* test; Welch's test; Wilcoxon–Mann–Whitney test; Heteroscedasticity; Stochastic simulation

61

The Wilcoxon–Mann–Whitney test condemned

J. Ludbrook

British Journal of Surgery 1996, **83**, 132–138

The Wilcoxon and Mann–Whitney tests are exactly equivalent. In both, continuous data are transformed into rank order. It is widely supposed that this makes the Wilcoxon–Mann–Whitney test superior to Student's *t* test if the samples come from non-normally distributed populations. But this and other advantages sometimes claimed for the test are illusory.

The conditions under which Student's *t* and the Wilcoxon–Mann–Whitney tests are, or are not, inaccurate are practically identical. The tests are in fact equally robust against excessive Type I error when the sampled populations are non-normally but symmetrically distributed, although only when certain conditions are fulfilled[2]. Neither test is robust when the population variances are unequal, especially if the sample sizes are also unequal[2]. Specifically, if the smaller sample comes from the population with the greater variance, the risk of Type I error is inflated[2–4]. In the articles reviewed in *The British Journal of Surgery*, the sample variances were as often grossly unequal as they were approximately equal. In the former case the statistical inferences are likely to have been flawed.

65

The Wilcoxon–Mann–Whitney and *t* tests are inaccurate if the assumptions on which they depend are breached and because the inferential model on which they are based assumes random sampling of defined populations. Random sampling is rarely employed in biomedical research. Instead, the experimental units (patients, animals, tissues or cells) are randomly divided into groups that are then exposed to different treatments or conditions. In this circumstance the proper inferential model is the randomization one, and the valid tests of significance are exact randomization or permutation tests for differences between means or other indices of central tendency[5].

I urge surgical researchers to abandon the Wilcoxon–Mann–Whitney test, unless they have measured up their results on a rank-order scale. Instead, I suggest they use one of the options mentioned above: a randomization (permutation) test if they are dealing with randomized groups or the Welch modification of Student's *t* test if they employ random sampling.

67

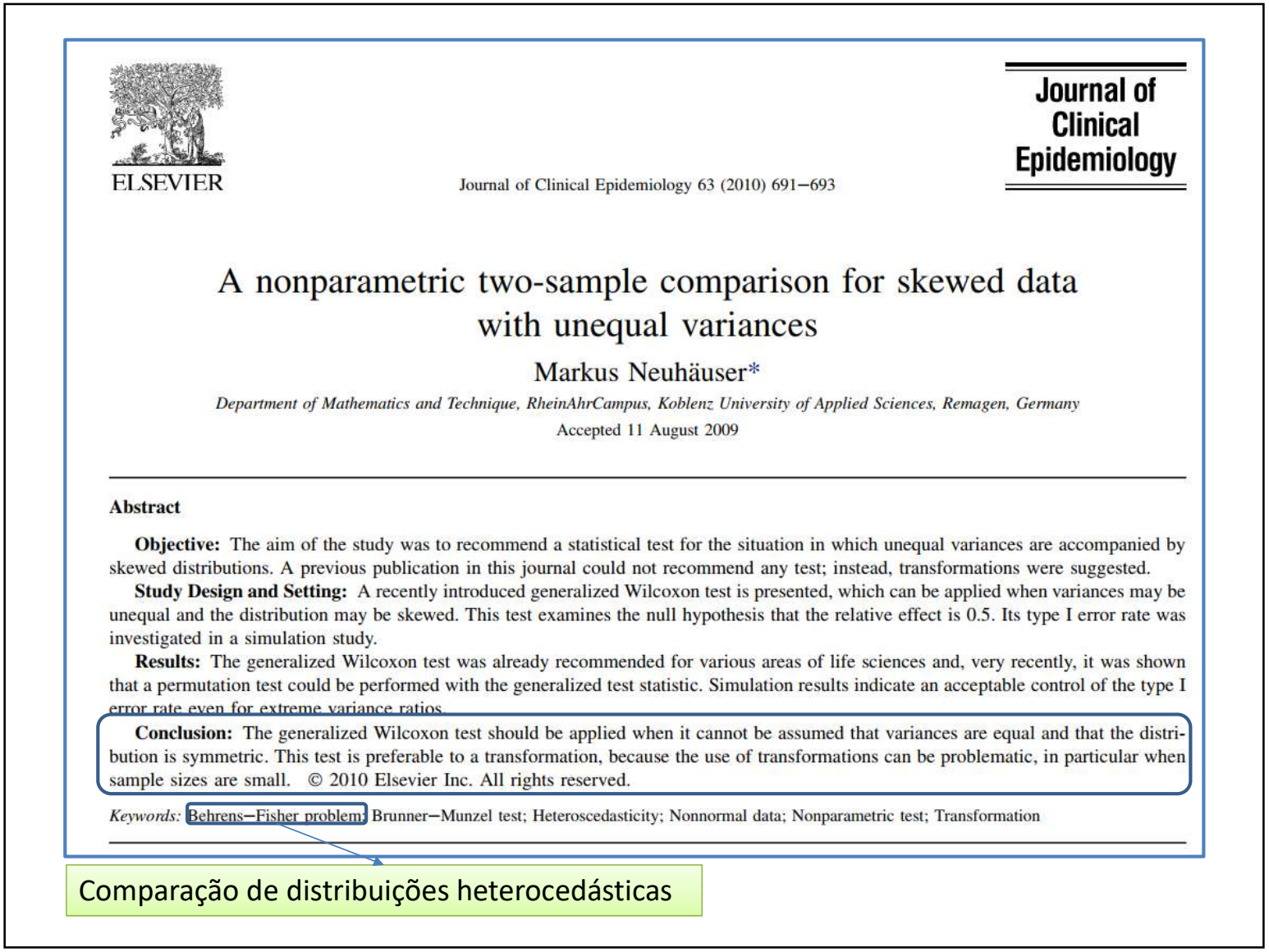
Invalidation of Parametric and Nonparametric Statistical Tests by Concurrent Violation of Two Assumptions

DONALD W. ZIMMERMAN
Carleton University

To cite this article: Donald W. Zimmerman (1998) Invalidation of Parametric and Nonparametric Statistical Tests by Concurrent Violation of Two Assumptions, The Journal of Experimental Education, 67:1, 55-68, DOI: 10.1080/00220979809598344

ABSTRACT. To provide counterexamples to some commonly held generalizations about the benefits of nonparametric tests, the author concurrently violated in a simulation study 2 assumptions of parametric statistical significance tests—normality and homogeneity of variance. For various combinations of nonnormal distribution shapes and degrees of variance heterogeneity, the Type I error probability of a nonparametric rank test, the Wilcoxon–Mann–Whitney test, was found to be biased to a far greater extent than that of its parametric counterpart, the Student *t* test. The Welch–Satterthwaite separate-variances version of the *t* test, together with a preliminary outlier detection and downweighting procedure, protected the statistical significance level more consistently than the nonparametric test did. Those findings reveal that nonparametric methods are not always acceptable substitutes for parametric methods such as the *t* test and the *F* test in research studies when parametric assumptions are not satisfied. They also indicate that multiple violations of assumptions can produce anomalous effects not observed in separate violations.

69



ELSEVIER

Journal of Clinical Epidemiology 63 (2010) 691–693

Journal of Clinical Epidemiology

A nonparametric two-sample comparison for skewed data with unequal variances

Markus Neuhäuser*

Department of Mathematics and Technique, RheinAhrCampus, Koblenz University of Applied Sciences, Remagen, Germany

Accepted 11 August 2009

Abstract

Objective: The aim of the study was to recommend a statistical test for the situation in which unequal variances are accompanied by skewed distributions. A previous publication in this journal could not recommend any test; instead, transformations were suggested.

Study Design and Setting: A recently introduced generalized Wilcoxon test is presented, which can be applied when variances may be unequal and the distribution may be skewed. This test examines the null hypothesis that the relative effect is 0.5. Its type I error rate was investigated in a simulation study.

Results: The generalized Wilcoxon test was already recommended for various areas of life sciences and, very recently, it was shown that a permutation test could be performed with the generalized test statistic. Simulation results indicate an acceptable control of the type I error rate even for extreme variance ratios.

Conclusion: The generalized Wilcoxon test should be applied when it cannot be assumed that variances are equal and that the distribution is symmetric. This test is preferable to a transformation, because the use of transformations can be problematic, in particular when sample sizes are small.

Keywords: Behrens–Fisher problem; Brunner–Munzel test; Heteroscedasticity; Nonnormal data; Nonparametric test; Transformation

Comparação de distribuições heterocedásticas

70

Tamanho de efeito não-paramétrico
Probabilidade de melhores respostas

- Se as distribuições apresentam assimetrias acentuadas ou com outliers ou amostra pequena, as médias e desvios-padrão não são tão úteis e o resultado do tamanho de efeito $(m_1 – m_2)/s$ pode não ser apropriado.
- Uma medida do tamanho de efeito não-paramétrica é a proporção dos pares de observações (uma para cada condição) para a qual a observação da primeira condição é maior.
- Se X representa uma observação aleatoriamente selecionada do Grupo 1 e Y representa uma observação aleatoriamente selecionada do Grupo 2, então essa medida estima a probabilidade de X ser maior que Y, i.e., $P(X > Y)$.

Adaptado de AGRESTI, A & FINLAY, B (2012) *Métodos estatísticos para as Ciências Sociais*. 4ª ed. Porto Alegre: PENSO, p. 234-5

73

Tamanho de efeito
Probabilidade de melhores respostas

- Um estudo de anorexia tem 4 meninas, 2 usando a nova terapia (X) e 2 no grupo de controle (Y).
- As mudanças de peso são:
  - Y: 2, 6
  - X: 4, 10
- Existem quatro pares de observações com uma em cada grupo:
  - $X = 4 > Y = 2$
  - $X = 4 < Y = 6$
  - $X = 10 > Y = 2$
  - $X = 10 > Y = 6$
- A observação do grupo terapia X é maior em 3 de 4 pares.
- Assim, a estimativa da probabilidade de melhores respostas é $P(X > Y) = ¾ = 0,75$.
- Se duas observações têm o mesmo valor, $P(X = Y)$, adota-se ½ para o par.
- Sob H_0: ausência de efeito, então $P(X > Y) = 0,5$.
- Quanto mais distante $P(X > Y) + 0,5\ P(X = Y)$ está de 0,5, mais forte o efeito.

74

Teste de Brunner-Munzel de igualdade probabilística entre duas condições independentes

Generalized WMW test

```
library(lawstat)
Y <- c(2,6)
X <- c(4,10)
lawstat::brunner.munzel.test(Y, X)
```

```
        Brunner-Munzel Test

data:  Y and X
Brunner-Munzel Test Statistic = 0.70711, df = 2, p-value = 0.5528
95 percent confidence interval:
 -0.7712175  2.2712175
sample estimates:
P(X<Y)+.5*P(X=Y)
            0.75
```

75

“Teste t de Welch não-paramétrico” = Teste de Brunnel Munzel = Teste de WMW generalizado

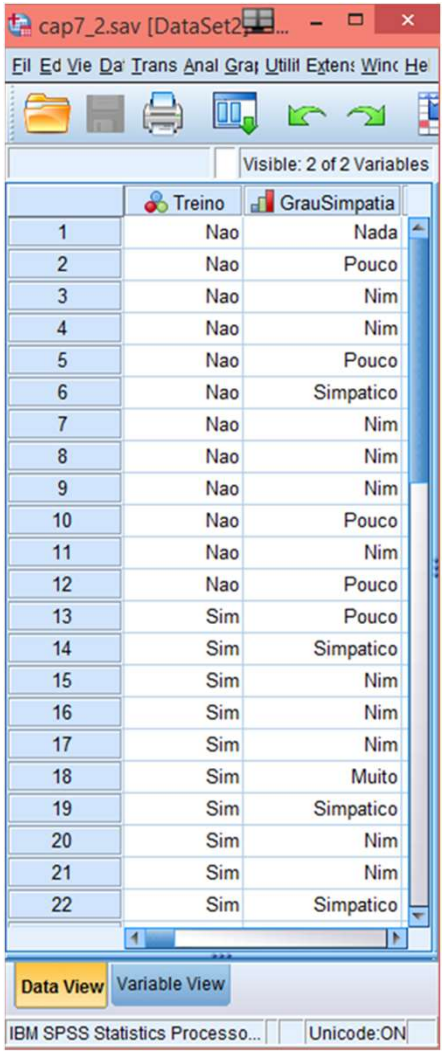
t-tests, non-parametric tests, and large studies—a paradox of statistical practice?

Morten W Fagerland

BMC Medical Research Methodology 2012, **12**:78

used. The Brunner-Munzel test, a non-parametric test that adjusts for unequal variances, may be used as an alternative to the WMW test. It is not widely available in

76



Treino	GrauSimpatia	#(T2>T1)	#(T2=T1)
1	1		
1	2		
1	2		
1	2		
1	2		
1	3		
1	3		
1	3		
1	3		
1	3		
1	3		
1	4		
2	2	1	4
2	3	5	6
2	3	5	6
2	3	5	6
2	3	5	6
2	3	5	6
2	4	11	1
2	4	11	1
2	4	11	1
2	5	12	0
	#Total(T2>T1)=	71	
	#Total(T2=T1)=		37
	Total de comparações=	120	
	P(T2>T1) + 0,5 P(T2=T1)=	0,7458	

77

Teste de Brunner-Munzel de igualdade probabilística entre duas condições independentes

Generalized WMW test
WMWGeneralizado.R

```
library(lawstat)
A <- c(1,2,2,2,2,3,3,3,3,3,3,4)
B <- c(2,3,3,3,3,3,4,4,4,5)
lawstat::brunner.munzel.test(A, B)
wilcox.test(A, B, exact = FALSE)
dif <- median(B) - median(A)
cat("Diferenca das medianas (Sim-Nao) = ",dif,sep="")
```

IBM SPSS Statistics 24

Test Statistics[a]

	Grau de simpatia
Mann-Whitney U	30,500
Wilcoxon W	108,500
Z	-2,099
Asymp. Sig. (2-tailed)	,036
Exact Sig. [2*(1-tailed Sig.)]	,050[b]
Exact Sig. (2-tailed)	,049
Exact Sig. (1-tailed)	,025
Point Probability	,014

a. Grouping Variable: Treino de competências sociais

b. Not corrected for ties.

```
## Brunner-Munzel Test
Brunner-Munzel Test Statistic = 2.5443, df = 18.934, p-value = 0.01983
95 percent confidence interval:
 0.5435583 0.9481084
sample estimates:
P(X<Y)+.5*P(X=Y)
      0.7458333

            Wilcoxon rank sum test with continuity correction

data:  A and B
W = 30.5, p-value = 0.03908
alternative hypothesis: true location shift is not equal to 0
Diferenca das medianas (Sim-Nao) = 0
```

Tamanho de efeito não-paramétrico
Probabilidade de melhores respostas

- Se os dois grupos têm distribuições normais com o mesmo desvio-padrão, existe uma conexão entre esse tamanho de efeito não-paramétrico e o do paramétrico, $(m_X - m_Y)/s$.
- Se $(m_X - m_Y)/s = 0$, $P(X > Y) = 0,5$
- Se $(m_X - m_Y)/s = 0,5$, $P(X > Y) = 0,64$
- Se $(m_X - m_Y)/s = 0,8$, $P(X > Y) = 0,71$
- Se $(m_X - m_Y)/s = 1$, $P(X > Y) = 0,76$
- Se $(m_X - m_Y)/s = -1$, $P(X > Y) = 0,24$
- Se $(m_X - m_Y)/s = 2$, $P(X > Y) = 0,92$
- Tamanho de efeito não-paramétrico forte
  - $P(X > Y) > 0,7$ ou $P(X > Y) < 0,3$

WolframAlpha computational intelligence.

```
NIntegrate[PDF[NormalDistribution[0, 1], x]*CDF[NormalDistribution[0, 1], x], {x, -\[Infinity], \[Infinity]}]
```

Mathematica 11

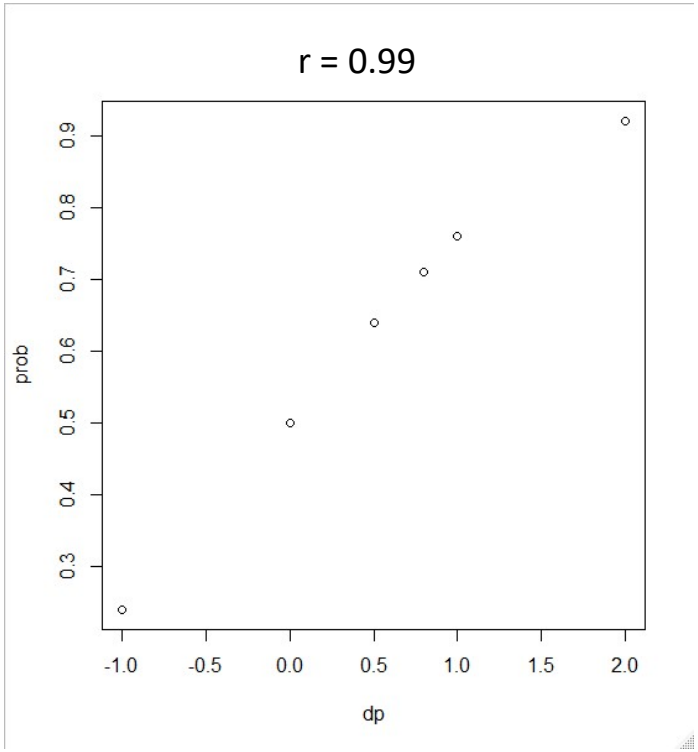
$$P(X > Y) = \int_{-\infty}^{\infty} f_X(t) F_Y(t)\, dt \quad = \int F_Y \, dF_X$$

```
In[19]:= NIntegrate[PDF[NormalDistribution[0, 1], x] * CDF[NormalDistribution[0, 1], x], {x, -∞, ∞}]
Out[19]= 0.5

In[21]:= NIntegrate[PDF[NormalDistribution[1, 2], x] * CDF[NormalDistribution[0, 2], x], {x, -∞, ∞}]
Out[21]= 0.638163

In[20]:= NIntegrate[PDF[NormalDistribution[1, 1], x] * CDF[NormalDistribution[0, 1], x], {x, -∞, ∞}]
Out[20]= 0.76025

In[23]:= NIntegrate[PDF[NormalDistribution[2, 1], x] * CDF[NormalDistribution[0, 1], x], {x, -∞, ∞}]
Out[23]= 0.92135
```

79

Tamanho de efeito não-paramétrico
Probabilidade de melhores respostas

```
dp <- c(0, 0.5, 0.8, 1, -1, 2)
prob <- c(0.5, 0.64, 0.71, 0.76, 0.24, 0.92)
plot(dp, prob)
cor(dp, prob)
```



80

University of Texas MD Anderson Cancer Center
Department of Biostatistics

Inequality Calculator, Version 3.0
November 25, 2013
User's Guide

The purpose of the software is to calculate the probability that one random variable is greater than another. The two random variables are assumed to follow the same standard distribution family, with different parameter values. The distributions currently supported are: beta, gamma, inverse gamma, normal, log normal and Weibull.

For two general continuous random variables X and Y, the probability that $X > Y$ is given by

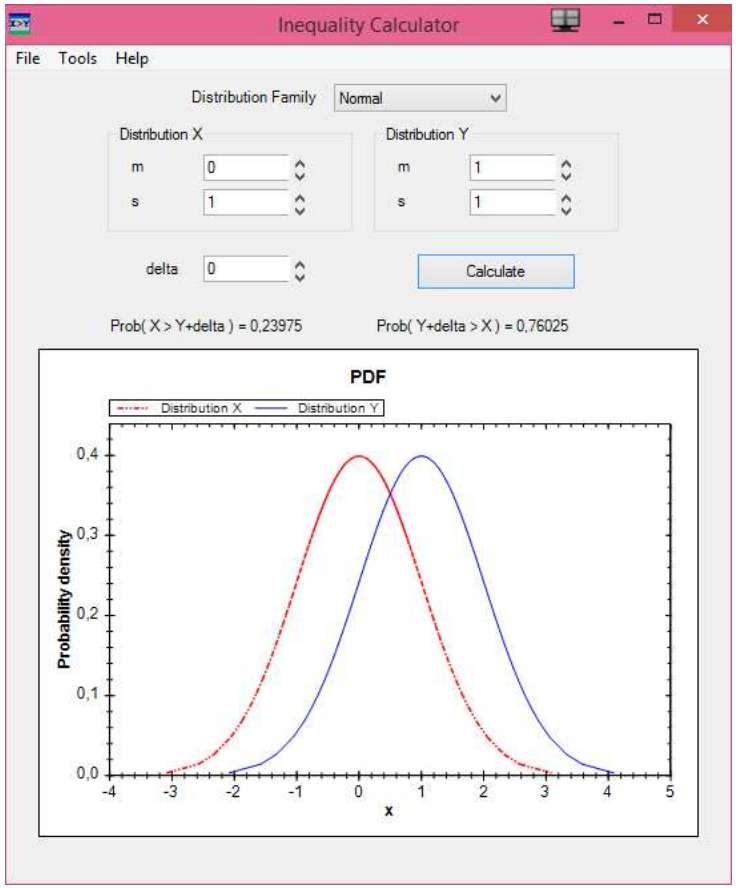
$$P(X > Y) = \int_{-\infty}^{\infty} f_X(t)\, F_Y(t)\, dt \quad = \int F_Y \, dF_X$$

where f_X is the PDF (probability density function) of X and F_Y is the CDF (cumulative distribution function) of Y. The program computes this probability for the case of X and Y following the same standard distribution family.

In the example of Figure 1, we let $X \sim$ Normal(0, 1) and $Y \sim$ Normal(1, 1) and click on "Calculate" to obtain the result Prob($X > Y$) = 0.23975, along with the complementary value Prob($Y > X$) = 0.76025.

https://biostatistics.mdanderson.org/softwaredownload/SingleSoftware.aspx?Software_Id=9

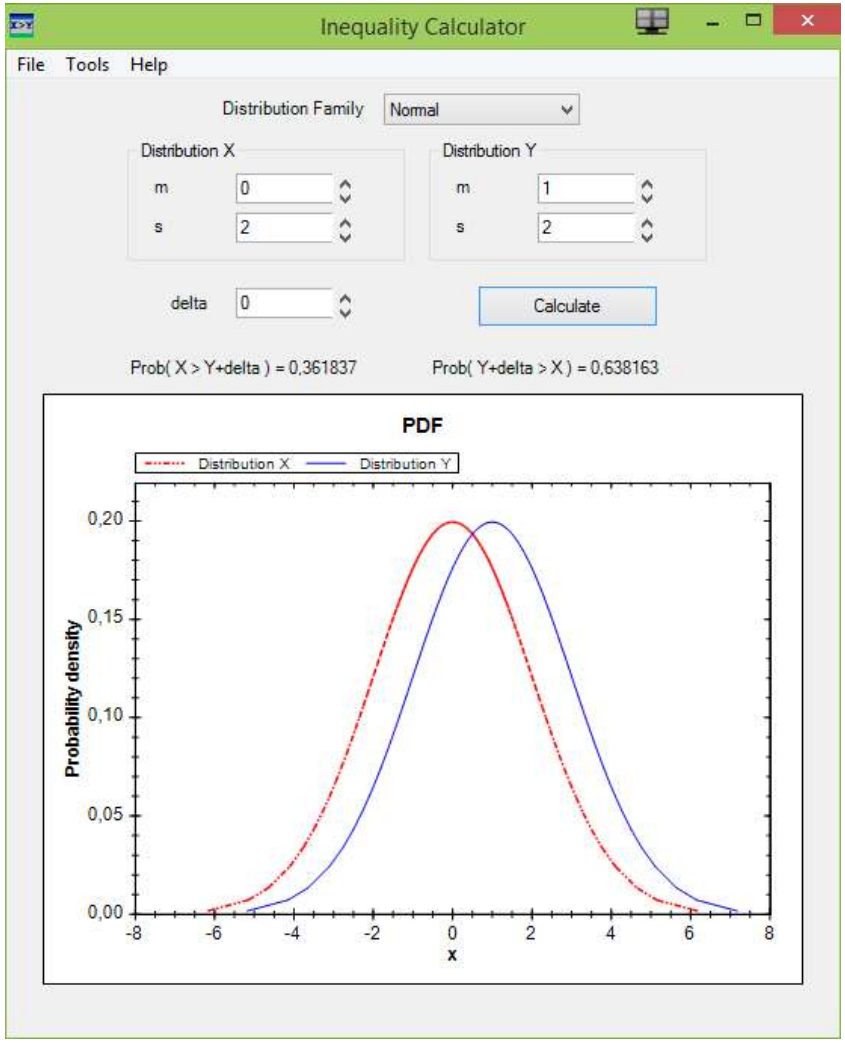
81



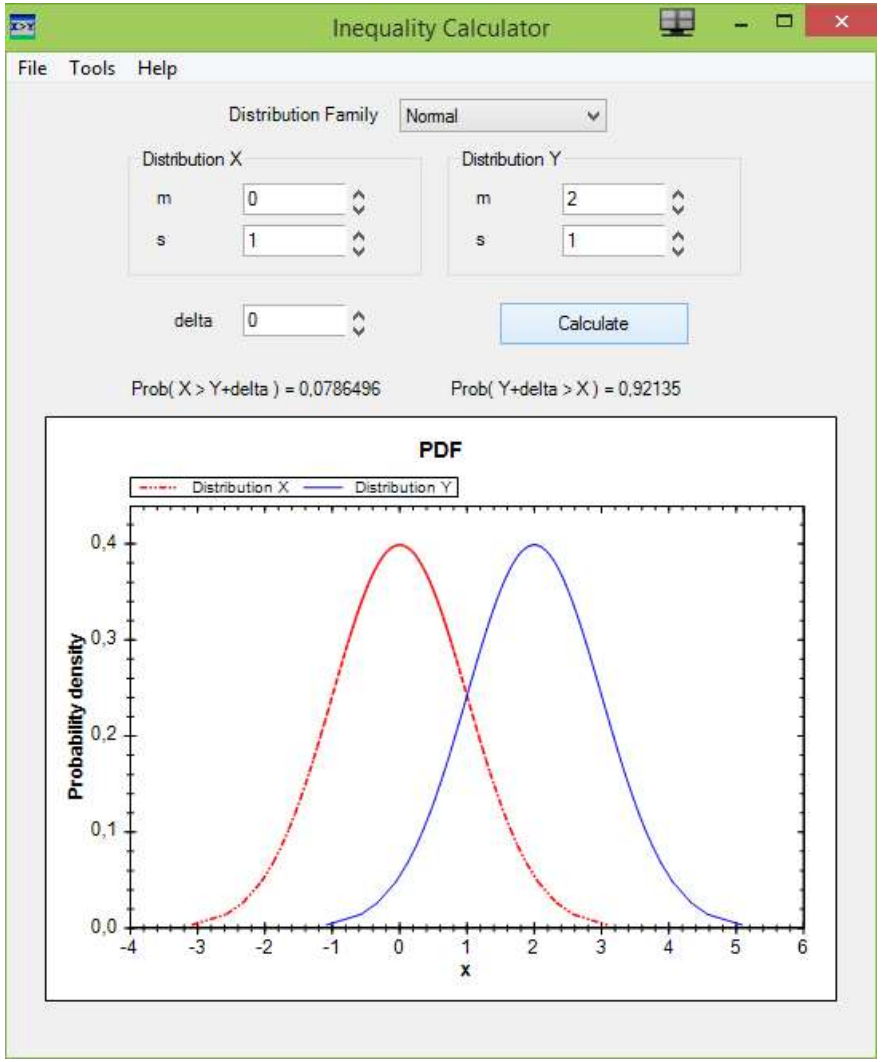
- First, suppose X and Y have normal distributions with standard deviation 1.
- If X has mean 0 and Y has mean 1, what is P(X > Y)? Some would say 1, because X is bigger than Y.
- But that's not true.
- X has a larger mean than Y, but fairly often a sample from Y will be larger than a sample from X.
- P(X > Y) = 0.76 in this case
- https://www.johndcook.com/blog/2008/07/26/random-inequalities-i/

?

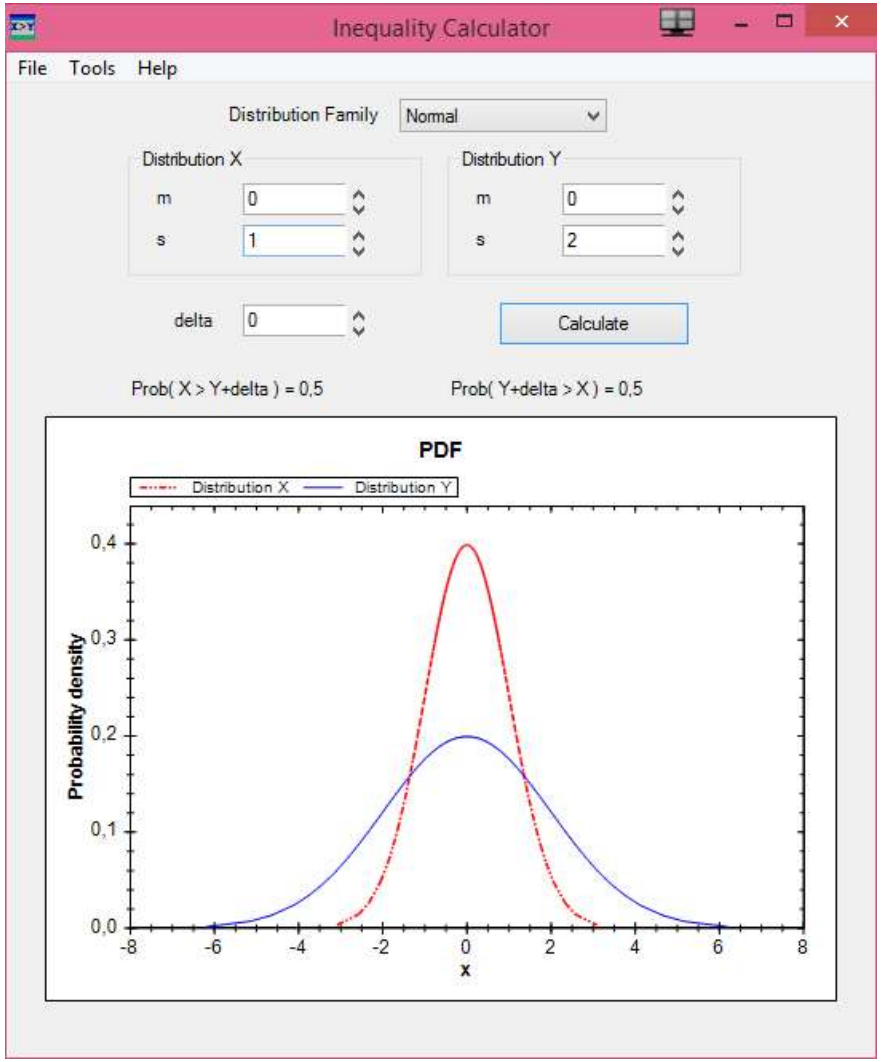
82


Inequality Calculator
File Tools Help
Distribution Family
Normal
Distribution X
m
0
s
2
Distribution Y
m
1
s
2
delta
0
Calculate
Prob(X > Y+delta) = 0,361837
Prob(Y+delta > X) = 0,638163
PDF
Distribution X
Distribution Y
Probability density
0,20
0,15
0,10
0,05
0,00
-8
-6
-4
-2
0
2
4
6
8
x

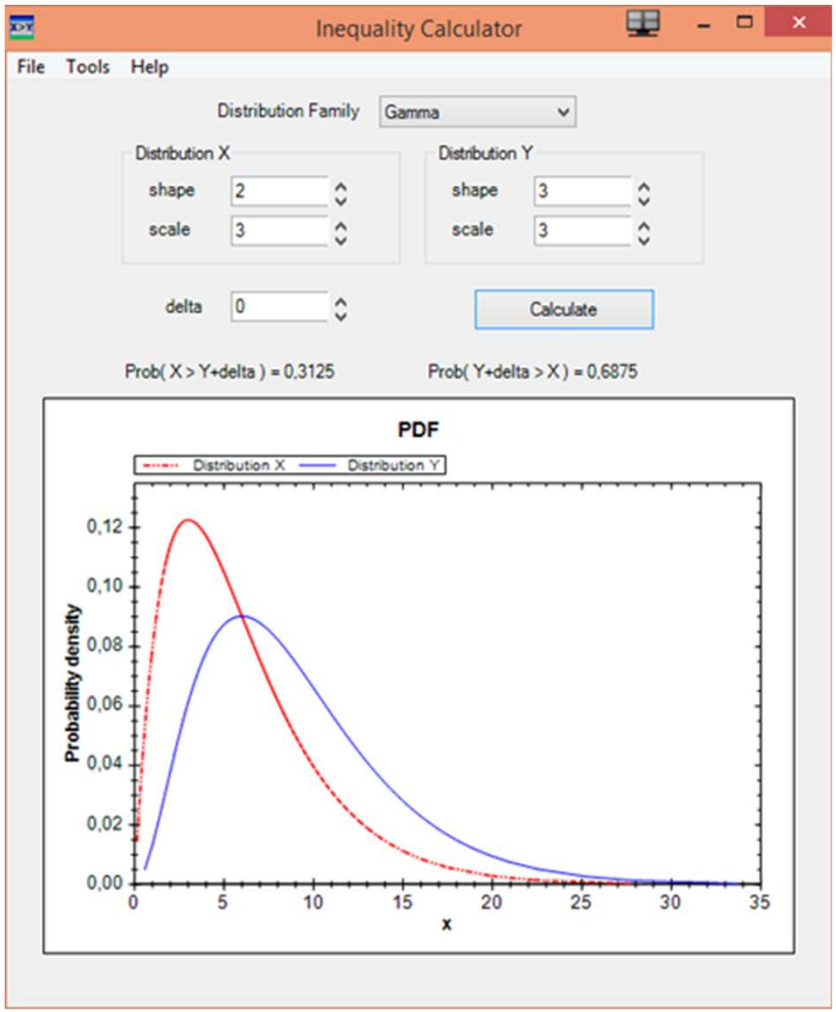
83


Inequality Calculator
File Tools Help
Distribution Family
Normal
Distribution X
m
0
s
1
Distribution Y
m
2
s
1
delta
0
Calculate
Prob(X > Y+delta) = 0,0786496
Prob(Y+delta > X) = 0,92135
PDF
Distribution X
Distribution Y
Probability density
0,0
0,1
0,2
0,3
0,4
-4
-3
-2
-1
0
1
2
3
4
5
6
x

84


Inequality Calculator
File
Tools
Help
Distribution Family
Normal
Distribution X
m
0
s
1
Distribution Y
m
0
s
2
delta
0
Calculate
Prob(X > Y+delta) = 0,5
Prob(Y+delta > X) = 0,5
PDF
Distribution X
Distribution Y
Probability density
0,4
0,3
0,2
0,1
0,0
-8
-6
-4
-2
0
2
4
6
8
x

85



Gama	X	Y
a=	2	3
b=	3	3
media=	6,0000	9,0000
mediana=	5,0323	8,0217
variancia=	18,0000	27,0000
assimetria=	1,4142	1,1547

86

POINTS OF SIGNIFICANCE

martin Krzywinski & naomi Altman

Nonparametric tests

Nonparametric tests robustly compare skewed or ranked data.

| VOL.11 NO.5 | MAY 2014 | NATURE METHODS

The fact that the output of a rank test is driven by the probability that a value drawn from distribution *A* will be smaller (or larger) than one drawn from *B* without regard to their absolute difference has an interesting consequence: we cannot use this probability (pairwise preferences, in general) to impose an order on distributions. Consider a case of three equally prevalent diseases for which treatment *A* has cure times of 2, 2 and 5 days for the three diseases, and treatment *B* has 1, 4 and 4. Without treatment, each disease requires 3 days to cure—let's call this control *C*. Treatment *A* is better than *C* for the first two diseases but not the third, and treatment *B* is better only for the first. Can we determine which of the three options (*A*, *B*, *C*) is better? If we try to answer this using the probability of observing a shorter time to cure, we find $P(A < C) = 67\%$ and $P(C < B) = 67\%$ but also that $P(B < A) = 56\%$—a rock-paper-scissors scenario.

Confusão entre medida de tamanho de efeito e valor-p

87

Teste de Brunner-Munzel de igualdade probabilística entre duas condições independentes

Generalized Wilcoxon test

```
# Krzywinski, M & Altman, N (2014) Nonparametric tests
# Nature Methods 11(5)
library(lawstat)
A <- c(2,2,5)
B <- c(1,4,4)
C <- c(3,3,3)
# P(X<Y)+.5*P(X=Y) = probabilidade de X ter tempo de cura menor que Y
lawstat::brunner.munzel.test(A, C)
lawstat::brunner.munzel.test(C, B)
lawstat::brunner.munzel.test(B, A)
```

88

Teste de Brunner-Munzel de igualdade probabilística entre duas condições independentes

Generalized Wilcoxon test

```
> brunner.munzel.test(A, C)
          Brunner-Munzel Test
data:  A and C
Brunner-Munzel Test Statistic = 0.5, df = 2, p-value = 0.6667
95 percent confidence interval:
 -0.7675509  2.1008842
sample estimates:
P(X<Y)+.5*P(X=Y)
 P(A < C) = 0.6666667
> brunner.munzel.test(C, B)
          Brunner-Munzel Test
data:  C and B
Brunner-Munzel Test Statistic = 0.5, df = 2, p-value = 0.6667
95 percent confidence interval:
 -0.7675509  2.1008842
sample estimates:
P(X<Y)+.5*P(X=Y)
 P(C < B) = 0.6666667
> brunner.munzel.test(B, A)
          Brunner-Munzel Test
data:  B and A
Brunner-Munzel Test Statistic = 0.17678, df = 4, p-value = 0.8683
95 percent confidence interval:
 -0.316997  1.428108
sample estimates:
P(X<Y)+.5*P(X=Y)
 P(B < A) = 0.5555556
```

89

ORIGINAL ARTICLES Epidemiology Biostatistics and Public Health - 2016, Volume 13, Number 2

Parametric and nonparametric two-sample tests for feature screening in class comparison: a simulation study

Elena Landoni [1], Federico Ambrogi [2], Luigi Mariani [1], Rosalba Miceli [1]

ABSTRACT

Background: The identification of a location-, scale- and shape-sensitive test to detect differentially expressed features between two comparison groups represents a key point in high dimensional studies. The most commonly used tests refer to differences in location, but general distributional discrepancies might be important to reveal differential biological processes.

Methods: A simulation study was conducted to compare the performance of a set of two-sample tests, *i.e.* Student's t, Welch's t, Wilcoxon-Mann-Whitney (WMW), Podgor-Gastwirth PG2, Cucconi, Kolmogorov-Smirnov (KS), Cramer-von Mises (CvM), Anderson-Darling (AD) and Zhang tests (Z_K, Z_C and Z_A) which were investigated under different distributional patterns. We applied the same tests to a real data example.

Results: AD, CvM, Z_A and Z_C tests proved to be the most sensitive tests in mixture distribution patterns, while still maintaining a high power in normal distribution patterns. At best, the AD test showed a power loss of ~ 2% in the comparison of two normal distributions, but a gain of ~ 32% with mixture distributions with respect to the parametric tests. Accordingly, the AD test detected the greatest number of differentially expressed features in the real data application.

Conclusion: The tests for the general two-sample problem introduce a more general concept of 'differential expression', thus overcoming the limitations of the other tests restricted to specific moments of the feature distributions. In particular, the AD test should be considered as a powerful alternative to the parametric tests for feature screening in order to keep as many discriminative features as possible for the class prediction analysis.

Key words: high-dimensional data; class comparison; location-scale problem; general two-sample problem; mixtures.

90

Teste W de Wilcoxon

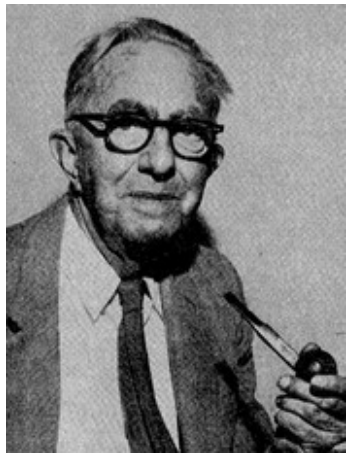
Delineamento intraparticipantes

- O teste W de Wilcoxon (Wilcoxon *signed-rank test*) testa a hipótese nula de igualdade das distribuições populacionais da VD intervalar simétrica em duas condições dependentes

Individual Comparisons by Ranking Methods

Frank Wilcoxon

Biometrics Bulletin, Vol. 1, No. 6. (Dec., 1945), pp. 80-83.



Frank Wilcoxon (1892-1965)

92

Statistica Neerlandica (1999) Vol. 53, nr. 3, pp. 277–286

Nonparametric methods for paired samples

U. Munzel*

Department of Medical Statistics, University of Göttingen, Humboldtallee 32, 37073 Göttingen, Germany

The small sample and asymptotic properties of nonparametric tests for paired sampled are examined. Linear rank statistics are compared with the paired t-test and the Wilcoxon-signed-rank test in simulation studies. From a minimax point of view the linear rank statistics turn out to be the best. Moreover, it is illustrated that the Wilcoxon-signed-rank test should not be used if it is not clear that the differences of the pairs have a symmetric distribution.

Key Words & Phrases: Asymmetry, Behrens-Fisher problem, paired t-test, rank transform, ties, Wilcoxon-signed-rank test.

93

1 Introduction

When data from a paired two-sample design are analysed without assuming the normal distribution, commonly the Wilcoxon-signed-rank test (WSR) is used. It is well known, however, that the theoretical results considering the WSR depend on the rather restrictive assumption that the distribution of the differences of the paired observations is symmetric. Moreover, the WSR cannot be used for ordered categorical data because the differences of the paired observations which are needed to compute the statistic, are meaningless in the context of categorical data.

almost arbitrary ties, in particular for ordered categorical data. In this setup, the class of distributions is more general than in the semiparametric location model, where $F_j(x) = F(x - \mu_j)$. No further assumptions about the shape of the distribution functions are made while the WSR assumes the differences $X_{i1} - X_{i2}; i = 1, \ldots, n$ to be distributed symmetrically. Since the model does not contain any parameters, it is reasonable to formulate the hypothesis of no treatment effect in terms of the marginal distribution functions, namely

$$H_0^F : F_1 = F_2$$

The use of this so called *marginal model* for the paired sample design dates back to HOLLANDER, PLEDGER and LIN (1974) and GOVINDARAJULU (1975). Note that, for the location model $F_j(x) = F(x - \mu_j)$, and hypothesis H_0^F is equivalent to H_0^μ: $\mu_1 = \mu_2$.

94

7 Summary

It turned out that the WSR should only be used in practice if one can be sure that the distribution of the differences of the pairs X_{i1}, X_{i2} is symmetric. This restrictive assumption may not even be violated slightly. The simulation studies showed that even the t-test is preferable if the distribution of the differences is skew, because the WSR does not maintain the preassigned level in this case which becomes worse.

95

Enfermeiros simpáticos

- Os enfermeiros gerais receberam um questionário que media o nível de simpatia com pacientes que sofrem de esclerose múltipla (EM).
- Para cada enfermeiro, um escore total **INTERVALAR** que varia entre 1 e 10 foi observado.
- Os enfermeiros então participaram de um grupo de discussão (uma hora), que incluía pacientes com EM.
- Mais tarde, um questionário parecido foi dado novamente a eles.
- Obviamente, esse é um delineamento dentre participantes, pois os mesmos participantes estão sendo medidos nas condições "antes" e "depois".
- Aqui estabeleceremos uma hipótese direcional.
- Ela deve ser definida quando existem provas que apoiam a direção, por exemplo, de estudos já feitos.
- Nossa hipótese é de que haverá uma diferença significativa entre os escores antes de depois da discussão, de modo que estes sejam maiores após a discussão.
- Observe que essa é uma hipótese unilateral, pois especificamos a direção da diferença.

Wilcoxon signed rank test with continuity correction
Wilcoxon.R

```
library(haven)
Dados <- haven::read_sav("Simpatia de enfermeiros.sav")
with(Dados, wilcox.test(Depois, Antes, mu=0,
                        paired = TRUE, alternative = "greater",
                        conf.int=TRUE))
medianDepois <- median(Dados$Depois)
medianAntes <- median(Dados$Antes)
dif <- medianDepois - medianAntes
cat("Diferenca das medianas (Depois-Antes) = ",dif,sep="")

        Wilcoxon signed rank test with continuity correction

data:  Depois and Antes
V = 34, p-value = 0.01439
alternative hypothesis: true location shift is greater than 0
95 percent confidence interval:
 0.9999487       Inf
sample estimates:
(pseudo)median
      2.170837
Diferenca das medianas (Depois-Antes) = 2
```

97

Tamanho de efeito

- TE = |medianaDepois – medianaAntes|
  = |7 – 5|
  TE = 2

- Z é a estatística de teste e por isso não pode ser medida de tamanho de efeito, pois quanto maior o tamanho da amostra, maior seu valor

98

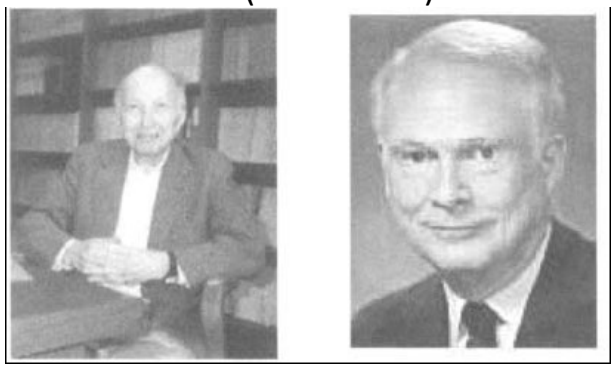
Conclusão

- Como o número de participantes é pequeno (N = 10) e apesar do teste K-S de normalidade da diferença entre os escores de simpatia não ser estatisticamente significante com valor-p igual a 89%, o teste estatístico escolhido foi o Wilcoxon.
- O teste Wilcoxon apresenta com valor-p unilateral exato de 1,4%.
- As medianas de simpatia dos enfermeiros nas condições antes e depois do grupo de discussão são, respectivamente, 5 e 7.
- Portanto, pode ser concluído que a atitude dos enfermeiros com doentes que sofrem de EM é mais simpática após terem participado do grupo de discussão.

Teste H de Kruskal-Wallis

- O teste H de Kruskal-Wallis testa a hipótese nula de igualdade das distribuições populacionais da VD pelo menos ordinal em três ou mais condições independentes.

William Kruskal (1919-2005)



Wilson Wallis (1912-1998)

100

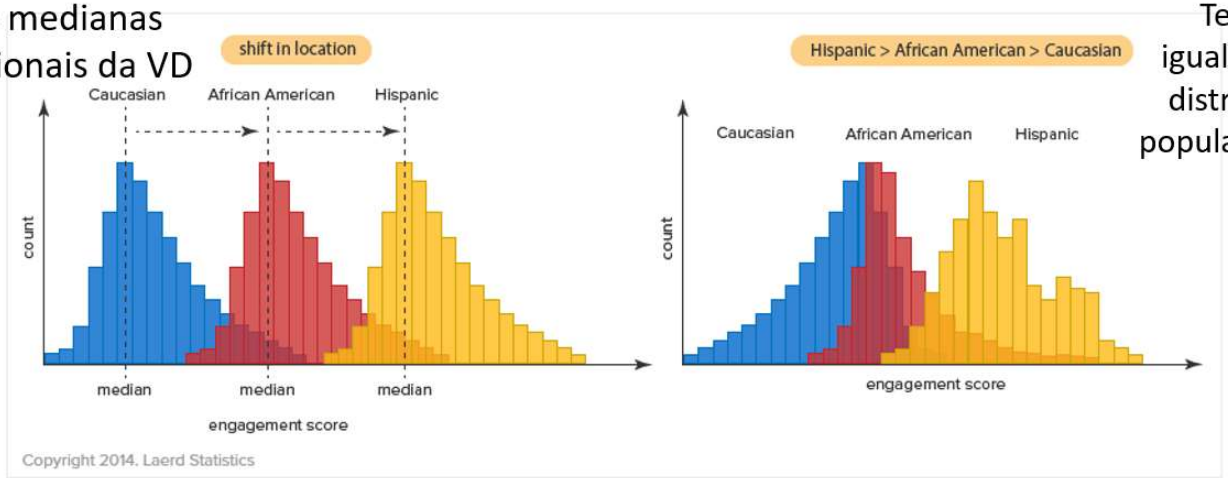
Kruskal-Wallis Test

For k independent samples from a continuous field, this tests:

H_0: The distributions of the k samples are the same

H_A: At least one sample is different

Teste de medianas populacionais da VD



Teste de igualdade das distribuições populacionais da VD

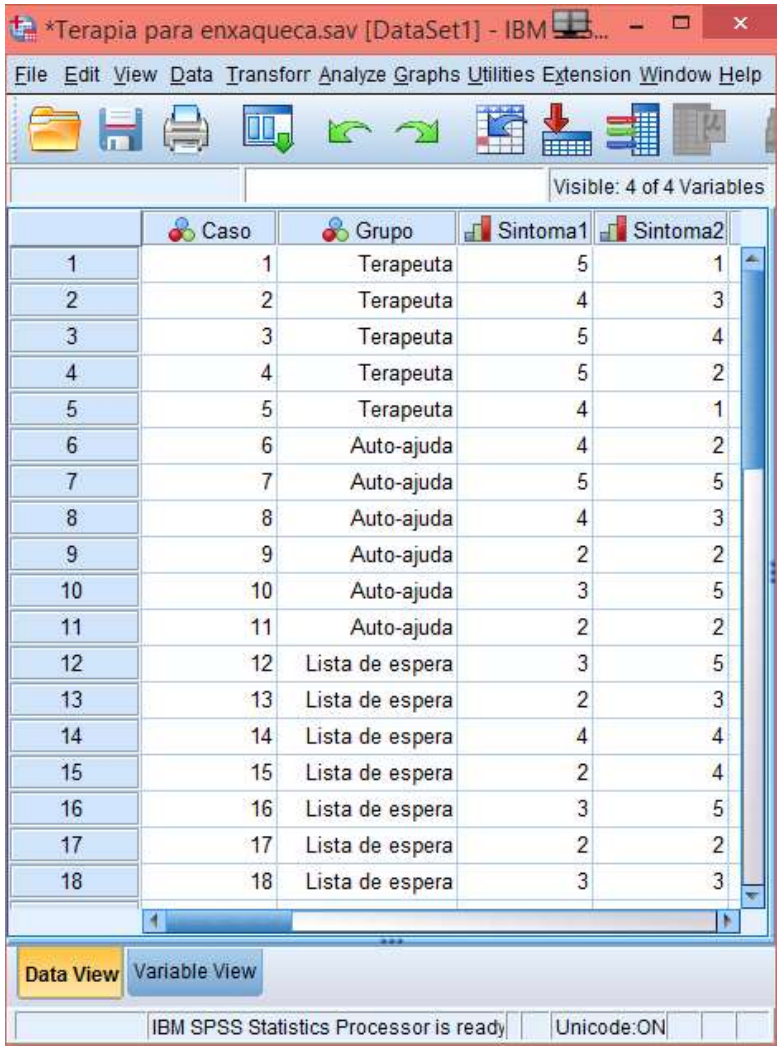
IBM SPSS Statistics Algorithms 25

101

Terapias para enxaqueca

- Como parte de seu projeto conjunto do ano sobre a utilidade da terapia para pessoas que sofrem de enxaqueca.
- Os pesquisadores distribuíram aleatoriamente 18 pessoas que sofrem de enxaqueca em três grupos.
- O grupo 1 tem seis sessões de uma hora de terapia com um terapeuta estagiário; o grupo 2 tem seis sessões de autoajuda de uma hora (que não são lideradas por um facilitador - a agenda é determinada pelos próprios membros do grupo), e o grupo 3 consiste em pessoas que sofrem de enxaqueca que gostariam de participar de terapia ou de autoajuda, mas têm que esperar.
- Os pesquisadores preveem que os grupos de terapia e de autoajuda avaliarão que sofrem menos de enxaquecas do que o grupo na lista de espera quando avaliarem sua enxaqueca em um segundo ponto no tempo.
- No início do estudo, os participantes avaliam os seus sintomas no último mês, de 0 (sem sofrimento) a 5 (sofrimento terrível).
- Supor que as variáveis dependentes são ordinais.
- Quatorze dias mais tarde, avaliam os seus sintomas (no último mês) novamente.

105


*Terapia para enxaqueca.sav [DataSet1] - IBM
File Edit View Data Transforr Analyze Graphs Utilities Extension Window Help
Visible: 4 of 4 Variables
Caso Grupo Sintoma1 Sintoma2
1 1 Terapeuta 5 1
2 2 Terapeuta 4 3
3 3 Terapeuta 5 4
4 4 Terapeuta 5 2
5 5 Terapeuta 4 1
6 6 Auto-ajuda 4 2
7 7 Auto-ajuda 5 5
8 8 Auto-ajuda 4 3
9 9 Auto-ajuda 2 2
10 10 Auto-ajuda 3 5
11 11 Auto-ajuda 2 2
12 12 Lista de espera 3 5
13 13 Lista de espera 2 3
14 14 Lista de espera 4 4
15 15 Lista de espera 2 4
16 16 Lista de espera 3 5
17 17 Lista de espera 2 2
18 18 Lista de espera 3 3
Data View
Variable View
IBM SPSS Statistics Processor is ready
Unicode:ON

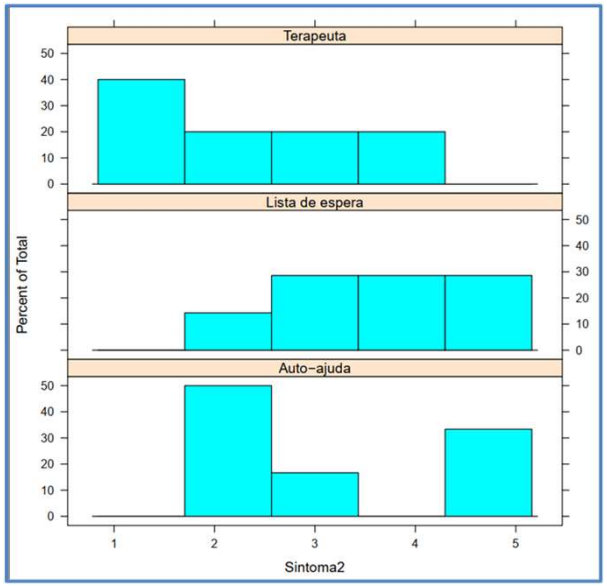
106

Teste H de Kruskal-Wallis

Kruskal-Wallis_enxaqueca.R

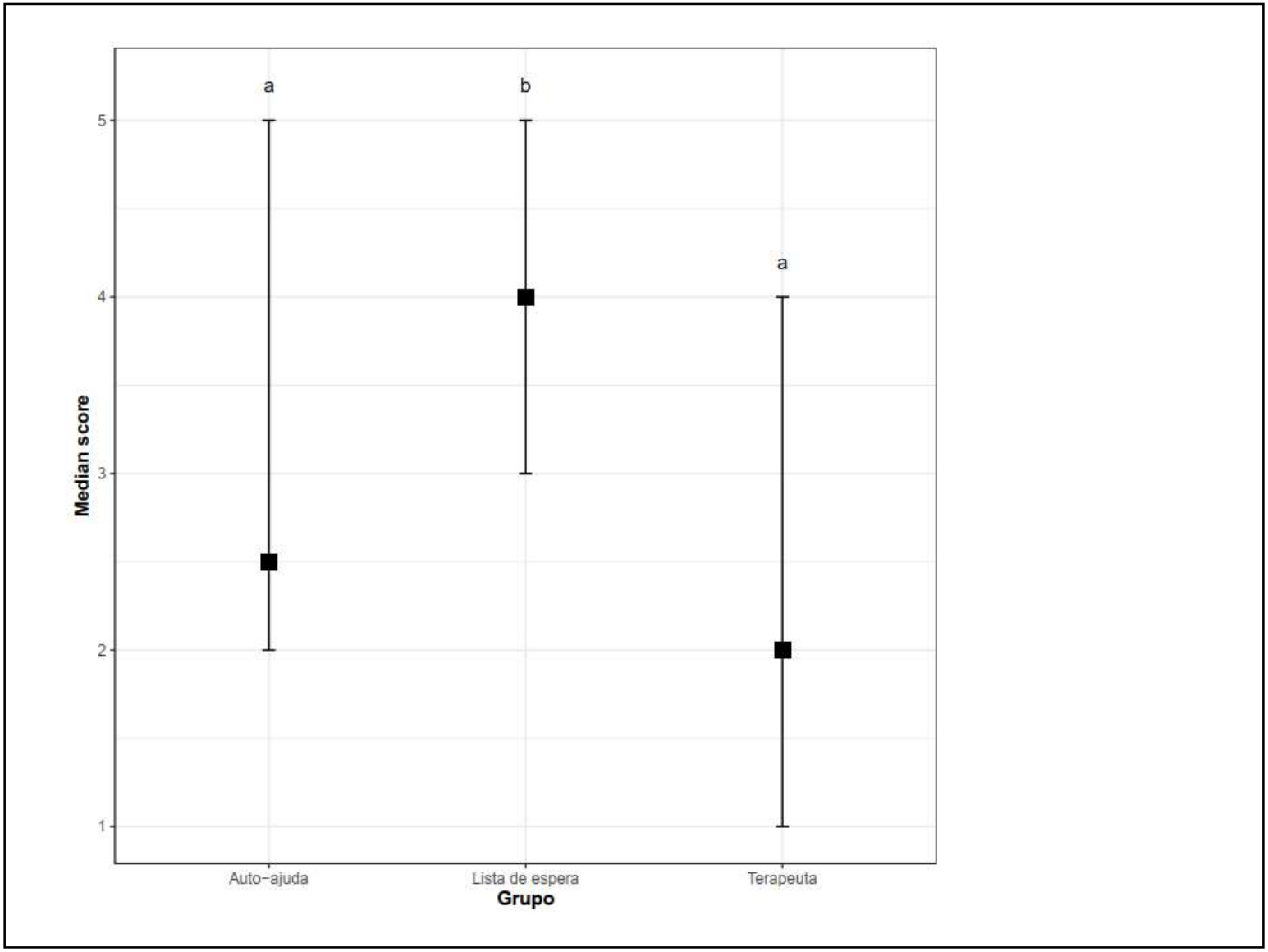
```
# https://rcompanion.org/handbook/F_08.html
library(readxl)
library(rcompanion)
library(FSA)
library(ggplot2)
library(lattice)
Dados <- readxl::read_excel("Terapia para enxaqueca.xlsx")
sink("TesteKruskal-Wallis_enxaqueca.txt")
pdf("TesteKruskal-Wallis_enxaqueca,pdf")
xtabs( ~ Grupo + Sintoma2,
       data = Dados)
lattice::histogram(~ Sintoma2 | Grupo,
          data=Dados,
          layout=c(1,3)      #  columns and rows of individual plots
)
with(Dados, kruskal.test(Sintoma2,Grupo))
with(Dados, rcompanion::epsilonSquared(Sintoma2,Grupo))
print(dt <- FSA::dunnTest(Sintoma2~Grupo,data=Dados,method="bh"))
pt <- dt$res
rcompanion::cldList(P.adj ~ Comparison, data = pt,threshold = 0.05)
Sum <- rcompanion::groupwiseMedian(Sintoma2~Grupo,
                      data       = Dados,
                      conf       = 0.95,
                      boot = TRUE,
                      R          = 1e4,
                      percentile = TRUE,
                      bca        = FALSE,
                      digits     = 3)
X <- 1:3
Y <- Sum$Percentile.upper + 0.2
Label <- c("a", "b", "a")
ggplot2::ggplot(Sum,                ### The data frame to use.
       aes(x = Grupo,
           y = Median)) +
  geom_errorbar(aes(ymin = Percentile.lower,
                    ymax = Percentile.upper),
                width = 0.05,
                size  = 0.5) +
  geom_point(shape = 15,
             size  = 4) +
  theme_bw() +
  theme(axis.title   = element_text(face  = "bold")) +
  ylab("Median score") +
  annotate("text",
           x = X,
           y = Y,
           label = Label)
dev.off()
sink()
```

Teste H de Kruskal-Wallis
Kruskal-Wallis_enxaqueca.R




```
 Sintoma2
Grupo              1 2 3 4 5
  Auto-ajuda       0 3 1 0 2
  Lista de espera  0 1 2 2 2
  Terapeuta        2 1 1 1 0

        Kruskal-Wallis rank sum test
data:  Sintoma2 and Grupo
Kruskal-Wallis chi-squared = 3.5595, df = 2, p-value = 0.1687

epsilon.squared
          0.209

 Dunn (1964) Kruskal-Wallis multiple comparison
  p-values adjusted with the Benjamini-Hochberg method.
Comparison          Z    P.unadj     P.adj
1 Auto-ajuda - Lista de espera -0.783990 0.43304598 0.4330460
2       Auto-ajuda - Terapeuta  1.103476 0.26982062 0.4047309
3  Lista de espera - Terapeuta  1.886053 0.05928782 0.1778635
Erro: No significant differences.
```

109


a
b
a
5
4
3
2
1
Median score
Auto−ajuda
Lista de espera
Terapeuta
Grupo

110

ANOVA de Friedman Delineamento intraparticipantes

- A ANOVA de Friedman é um equivalente não-paramétrico da ANOVA unifatorial para medidas repetidas e é uma generalização do teste de Wilcoxon para duas condições dependentes.

Milton Friedman (1912-2006)

111

ANOVA de Friedman

TesteFriedman.R

```
library(PMCMRplus)
## Sachs, 1997, p. 675
## Six persons (block) received six different diuretics
## (A to F, treatment).
## The responses are the Na-concentration (mval)
## in the urine measured 2 hours after each treatment.
## Assume A is the control.
Data <- matrix(c(
        3.88, 5.64, 5.76, 4.25, 5.91, 4.33, 30.58, 30.14, 16.92,
        23.19, 26.74, 10.91, 25.24, 33.52, 25.45, 18.85, 20.45,
        26.67, 4.44, 7.94, 4.04, 4.4, 4.23, 4.36, 29.41, 30.72,
        32.92, 28.23, 23.35, 12, 38.87, 33.12, 39.15, 28.06, 38.23,
        26.65),nrow=6, ncol=6,
        dimnames=list(1:6, LETTERS[1:6]))
print(Data)
## Global Friedman test
PMCMRplus::friedmanTest(Data)
## Exact many-one test
PMCMRplus::frdManyOneExactTest(y=Data, p.adjust = "bonferroni")
## Eisinga et al. 2017
PMCMRplus::frdAllPairsExactTest(y=Data, p.adjust = "bonferroni")
```

ANOVA de Friedman

TesteFriedman.R

```
     A     B     C    D     E     F
1 3.88 30.58 25.24 4.44 29.41 38.87
2 5.64 30.14 33.52 7.94 30.72 33.12
3 5.76 16.92 25.45 4.04 32.92 39.15
4 4.25 23.19 18.85 4.40 28.23 28.06
5 5.91 26.74 20.45 4.23 23.35 38.23
6 4.33 10.91 26.67 4.36 12.00 26.65

        Friedman rank sum test
Friedman chi-squared = 23.333, df = 5, p-value = 0.0002915

Pairwise comparisons using Eisinga-Heskes-Pelzer and Grotenhuis many-to-one test
for a two-way balanced complete block design
  A
B 0.114
C 0.043
D 1.000
E 0.014
F 8.4e-05
P value adjustment method: bonferroni
alternative hypothesis: two.sided

Pairwise comparisons using Eisinga, Heskes, Pelzer & Te Grotenhuis all-pairs test
with exact p-values for a two-way balanced complete block design
  A       B       C       D       E
B 0.34101 -       -       -       -
C 0.12897 1.00000 -       -       -
D 1.00000 0.78175 0.34101 -       -
E 0.04094 1.00000 1.00000 0.12897 -
F 0.00025 1.00000 1.00000 0.00197 1.00000

P value adjustment method: bonferroni
```

113

Effect Size Statistics for the Friedman and the Cochran Tests

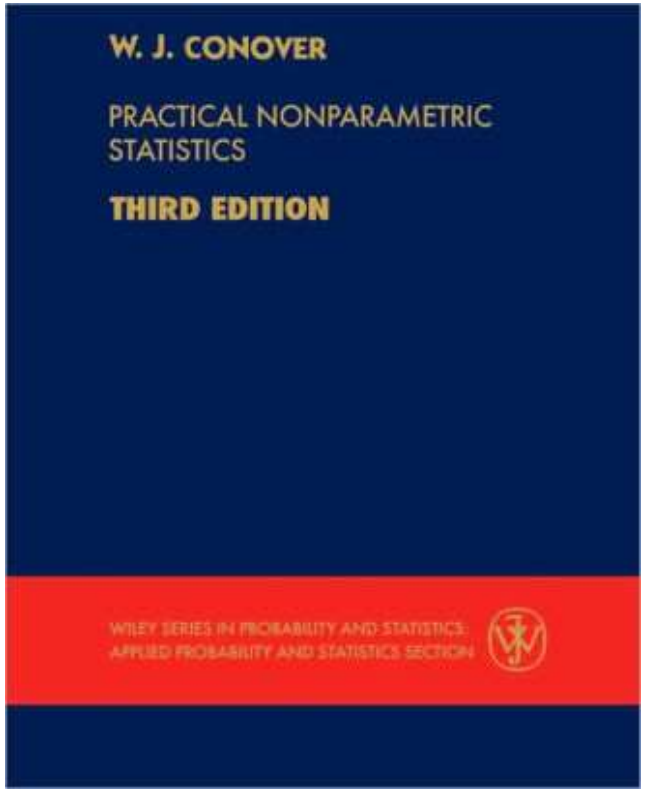
SPSS computes Kendall's coefficient of concordance (Kendall's *W*), a strength-of-relationship index. The coefficient of concordance ranges in value from 0 to 1, with higher values indicating a stronger relationship. This coefficient can be used as a strength-of-relationship index for both the Cochran and the Friedman tests. See Marascuilo and Serlin (1988) for further discussion about the relationship between the coefficient of concordance and these tests.

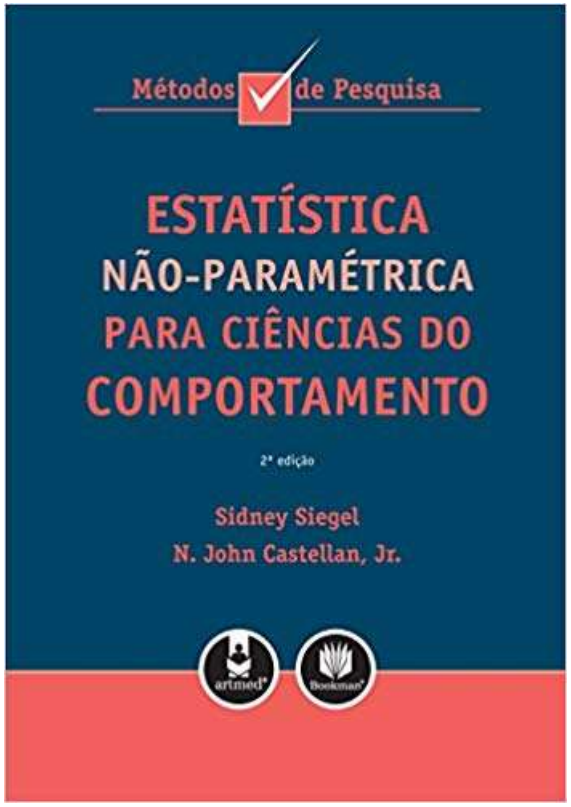
Using SPSS for Windows and Macintosh - analyzing and understanding - 7e - Green & Salkind - 2014

118

148

Livros


W. J. CONOVER
PRACTICAL NONPARAMETRIC STATISTICS
THIRD EDITION
WILEY SERIES IN PROBABILITY AND STATISTICS
APPLIED PROBABILITY AND STATISTICS SECTION


Métodos de Pesquisa
ESTATÍSTICA
NÃO-PARAMÉTRICA
PARA CIÊNCIAS DO
COMPORTAMENTO
2ª edição
Sidney Siegel
N. John Castellan, Jr.
artmed
Bookman

149

Introduction to Traditional Nonparametric Tests

- https://rcompanion.org/handbook/F_01.html

Traditional Nonparametric Tests

150

FUNDAMENTALS OF MATHEMATICAL STATISTICS

(A Modern Approach)

A Textbook written completely on modern lines for Degree, Honours, Post-graduate Students of all Indian Universities and Indian Civil Services, Indian Statistical Service Examinations.

(Contains, besides complete theory, more than 650 fully solved examples and more than 1,500 thought-provoking Problems with Answers, and Objective Type Questions)

S.C. GUPTA
Reader in Statistics
Hindu College,
University of Delhi
Delhi

V.K. KAPOOR
Reader in Mathematics
Shri Ram College of Commerce
University of Delhi
Delhi

Tenth Revised Edition
(Greatly Improved)

SULTAN CHAND & SONS
Educational Publishers
New Delhi

* First Edition : Sept. 1970
Tenth Revised Edition : August 2000
Reprint : 2002

* Price : **Rs. 210.00**

ISBN 81-7014-791-3

* *Published by* :
Sultan Chand & Sons
23, Darya Ganj, New Delhi-110002
Phones : 3277843, 3266105, 3281876

* *Laser typeset by* : T.P.
Printed at:- New A.S. Offset Press Laxmi Nagar Delhi-92

Five-Point Likert Items: t test versus Mann-Whitney-Wilcoxon

Practical Assessment, Research & Evaluation, Vol 15, No 11

Joost C. F. de Winter and Dimitra Dodou
Department of BioMechanical Engineering, Delft University of Technology

Likert questionnaires are widely used in survey research, but it is unclear whether the item data should be investigated by means of parametric or nonparametric procedures. This study compared the Type I and II error rates of the *t* test versus the Mann-Whitney-Wilcoxon (MWW) for five-point Likert items. Fourteen population distributions were defined and pairs of samples were drawn from the populations and submitted to the *t* test and the *t* test on ranks, which yields the same results as MWW. The results showed that the two tests had equivalent power for most of the pairs. MWW had a power advantage when one of the samples was drawn from a skewed or peaked distribution. Strong power differences between the *t* test and MWW occurred when one of the samples was drawn from a multimodal distribution. Notably, the Type I error rate of both methods was never more than 3% above the nominal rate of 5%, even not when sample sizes were highly unequal. In conclusion, for five-point Likert items, the *t* test and MWW generally have similar power, and researchers do not have to worry about finding a difference whilst there is none in the population.

156

Austral Ecology (2001) **26**, 32–46

A new method for non-parametric multivariate analysis of variance

MARTI J. ANDERSON
Centre for Research on Ecological Impacts of Coastal Cities, Marine Ecology Laboratories A11, University of Sydney, New South Wales 2006, Australia

Abstract Hypothesis-testing methods for multivariate data are needed to make rigorous probability statements about the effects of factors and their interactions in experiments. Analysis of variance is particularly powerful for the analysis of univariate data. The traditional multivariate analogues, however, are too stringent in their assumptions for most ecological multivariate data sets. Non-parametric methods, based on permutation tests, are preferable. This paper describes a new non-parametric method for multivariate analysis of variance, after McArdle and Anderson (in press). It is given here, with several applications in ecology, to provide an alternative and perhaps more intuitive formulation for ANOVA (based on sums of squared distances) to complement the description provided by McArdle and Anderson (in press) for the analysis of any linear model. It is an improvement on previous non-parametric methods because it allows a direct additive partitioning of variation for complex models. It does this while maintaining the flexibility and lack of formal assumptions of other non-parametric methods. The test-statistic is a multivariate analogue to Fisher's *F*-ratio and is calculated directly from any symmetric distance or dissimilarity matrix. *P*-values are then obtained using permutations. Some examples of the method are given for tests involving several factors, including factorial and hierarchical (nested) designs and tests of interactions.

Key words: ANOVA, distance measure, experimental design, linear model, multifactorial, multivariate dissimilarity, partitioning, permutation tests, statistics.

158

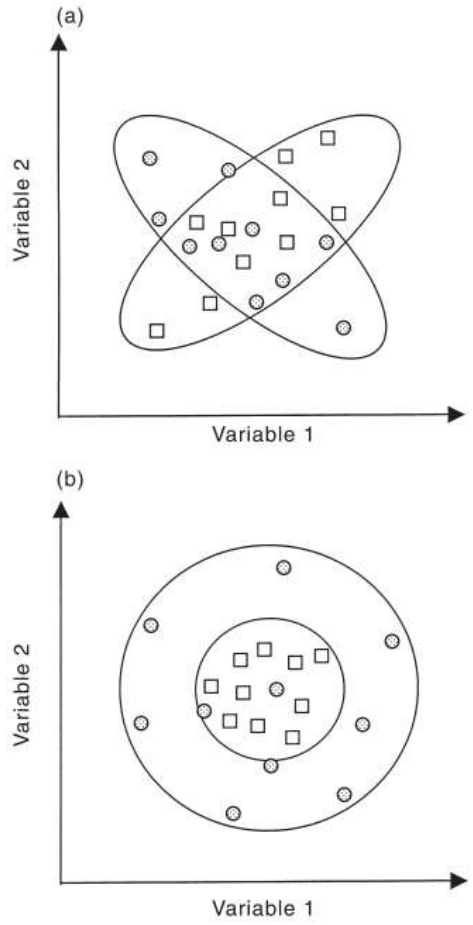


Fig. 4. Two variables in each of two groups of observations where (a) the groups differ in correlation between variables, but not in location or dispersion and (b) the groups differ in dispersion, but not in location or correlation between variables.

159

Available online at www.sciencedirect.com
ScienceDirect
Journal of Statistical Planning and Inference 137 (2007) 2706–2720
ELSEVIER
journal of statistical planning and inference
www.elsevier.com/locate/jspi

How robust are tests for two independent samples?

Dieter Rasch[a], Friedrich Teuscher[b], Volker Guiard[b,*]

[a] *Insitut für angewandte Statistik und EDV Universität für bodenkultur Wien, Österreich*
[b] *Research Institute for the Biology of Farm Animals Dummerstorf, Research Unit Genetics & Biometry, Germany*

Received 30 December 2005; accepted 14 April 2006
Available online 14 January 2007

174

Abstract

Non-parametric procedures are sometimes in use even in cases where the corresponding parametric procedure is preferable. This is mainly due to the fact that in practical applications of statistical methods too much attention is paid to any violation of the normality assumption–normal distribution is, however, primarily supposed in order to easily derive the exact distribution of the statistic used within parametric approaches.

As concerns the case of two independent samples and the comparison of (the expectations of) two populations the *t*-test and its non-parametric counterpart, the Wilcoxon (Mann–Whitney) test are of particular interest. These both serve as the illustrative example, given here.

The Wilcoxon test compares the two distributions, indeed, but may in cases where we are interested in comparing expectations lead to significance even if the expectations are equal; this because any higher moments in the two populations may differ. On the other hand, the *t*-test is so robust against non-normality that there is nearly no need to use the Wilcoxon test in comparing expectations.

Most results for continuous distributions have been obtained in a research group in Dummerstorf–Rostock some years ago and have been published by Herrendörfer [1980. Robustheit I, Arbeitsmaterial zum Forschungsthema Robustheit, Probleme der angewandten Statistik, vol. 4. Forschungszentrum Dummerstorf-Rostock, Heft], Herrendörfer et al. [1983. Robustness of statistical methods. II, Methods for the one-sample problem. Biometrical J. 25, 327–343], Rasch [1995. Mathematische Statistik. Johann Ambrosius Barth, Leipzig-Heidelberg], Rasch and Tiku [1984. Robustness of Statistical Methods and Nonparametric Statistics. VEB Deutscher Verlag der Wissenschaften, Berlin (D, Reidel Publ. Co., Dordrecht, Lancaster, Boston, Tokyo, 1985)], and Rasch and Guiard [2004. The robustness of parametric statistical methods. Psychol. Sci. 46(2), 175–208]. Most of the results are based on extensive simulation experiments with 10 000 runs each.

In the present paper we discuss the two-sample problem firstly by summarising the results for continuous distributions from some preprints in German language; however, secondly we offer new results. All above we investigate the *t*-test's and the Wilcoxon test's robustness in cases of the distribution not being continuous.

All the results are that in most practical cases the two-sample *t*-test is so robust that it can be recommended in nearly all applications.

Keywords: *t*-test; Wilcoxon test; Ordered categorical data; Robustness; Simulation

3. Robustness against ordinal distributions

In many applications—especially in the social sciences—ordered categorical data are observed. It is obvious that the assignment of natural numbers to the categories is purely arbitrary and any monotone transformation of given scores is equally justified but may lead to different results using parametric tests (see for instance Ivanova and Berger, 2001).

In textbooks on psychological statistics and in sociological research practice for ordered categorical data the Wilcoxon test is recommended and used, respectively. Though, most of the recommendations are not aware of the fact, that this test is based on the assumption of a continuous distribution. There are developments to use rank tests and permutation tests (Conover, 1973; Streitberg and Roehmel, 1986; Brunner and Puri, 1996) instead of the Wilcoxon test, but these tests are seldom used in practice up to now. Brunner and Puri (2001) summarise recent developments of the use of nonparametric methods in simple and high-order factorial experiments for independent observations as well as for repeated measurements. The Wilcoxon test comes out as a very special case in a simple design. Because many of the results are asymptotic, the practical behaviour of nonparametric tests has been investigated for small samples. For the paired sample case see Munzel (1999) and for more general repeated measurement designs in factorial experiments (Singer et al., 2004). Both compare parametric and nonparametric tests for repeated measurements.

We now concentrate on the one-sided independent two-sample problem and discuss in the following mainly the special case that there is some underlying hidden continuous variable which generates ordered categorical data. We investigate the properties of the likelihood ratio (LR) test for categorical data generated in this way and compare it with the properties of the Wilcoxon test and the two-sample t-test, if integers are used for k categories, that is the natural order of numbers $0, 1, \ldots, k-1$ or $1, 2, \ldots, k$—but suppose, as indicated—any plausible underlying hidden continuous variable. We consider distributions of $k = 5$ ordered categories.

3.3. The tests used

We consider the one-sided two-sample problem with the index $l = 1, 2$ for the two samples, respectively. Here $l = 1$ corresponds in experiments to the control group and $l = 2$ to the treatment group. In the sequel we use p_{li} instead of p_i and n_l instead of n. Written in terms of the logistic function (13) we have $F_1 = F(x)$ for $l = 1$ and $F_2 = F(x - \theta)$ for $l = 2$ and $\theta > 0$ (two-sided and the left-sided alternatives can be handled analogously).

With respect to the null hypothesis

$$H_0 : \theta = 0$$

and the alternative hypothesis

$$H_A : \theta > 0$$

we compared the power function of the following statistical 0.05-tests:

- normal approximation for the Wilcoxon test without correction for ties;
- normal approximation for the Wilcoxon test with correction for ties;
- chi-square test for $2 \times k$-tables;
- logistic procedure with the Wald test;
- logistic procedure with the LR test; and
- t-test using the values 1, 2, . . . for the categories.

The use of values 1, 2, . . . for the categories is a tool for the applicability of the t-test but nevertheless the hypotheses about θ have to be tested.

If the values 1, 2, . . . are considered as original variables, the t-test is related to test the hypothesis

$$H_0 : \mu_1 = \mu_2$$

against the alternative hypothesis

$$H_A : \mu_2 - \mu_1 = \delta > 0.$$

(Note that there is a difference between the shift parameter θ of the logistic distribution and $\mu_2 - \mu_1 = \delta$.)

177

4. Final remarks

We investigated ordinal random variables amongst them also those which can be thought to be the result of an underlying "hidden" logistic distribution. By ordinal variables as well the normality assumption of the t-test (after replacing the k categories by $1, 2, \ldots, k$) as also the continuity assumption of the Wilcoxon test is violated.

For the hidden variable we investigated the logistic distribution because it is used as a model in many applications (for instance in drug screening and for several psychological characters) but other distributions are still open to discussion.

We further limited our investigation on $k = 5$ categories. This because less than five categories relatively seldom occur in applications. If we have more than 5 we expect that the results will become better because we expect that with larger k the number of ties will become smaller which have to be corrected for the Wilcoxon test. But also in this direction further investigations are needed.

By this paper we do not recommend to use the numbers $1, 2, \ldots, k$ for the categories. We only show that if such a scoring is used then the t-test is applicable.

Robustness to Non-Normality of Common Tests for the Many-Sample Location Problem

AZMERI KHAN† azmeri@deakin.edu.au
School of Computing and Mathematics, Deakin University, Waurn Ponds, VIC3217, Australia

GLEN D. RAYNER*
National Australia Bank, Australia

JOURNAL OF APPLIED MATHEMATICS AND DECISION SCIENCES, 7(4), 187–206

Abstract. This paper studies the effect of deviating from the normal distribution assumption when considering the power of two many-sample location test procedures: ANOVA (parametric) and Kruskal-Wallis (non-parametric). Power functions for these tests under various conditions are produced using simulation, where the simulated data are produced using MacGillivray and Cannon's [10] recently suggested *g-and-k* distribution. This distribution can provide data with selected amounts of skewness and kurtosis by varying two nearly independent parameters.

Keywords: ANOVA, *g-and-k* distribution, Kruskal-Wallis test, Quantile function.

6. Conclusion

We distil our study into four observations/recommendations.

1. Both the ANOVA and Kruskal-Wallis tests are vastly more affected by the kurtosis of the error distribution rather than by its skewness, and the effect of skewness is unrelated to its direction.

2. Both the ANOVA and Kruskal-Wallis test sizes do not seem to be particularly affected by the shape of the error distribution.

3. The Kruskal-Wallis test does not seem to be an appropriate test for small samples (say $n < 5$). Even for non-normal data, the ANOVA test is a better option than the Kruskal-Wallis test for small sample sizes (say $n = 3$). This comment is made on the basis of the comparison between a Kruskal-Wallis test of nominal size 1.1% and an ANOVA test of size 0.5%. It is clearly understandable that a comparison between the Kruskal-Wallis and ANOVA tests of same size is a more reasonable procedure.

4. The Kruskal-Wallis tests clearly performs better than the ANOVA test if the sample sizes are large and kurtosis is high. Increasing sample size drastically improves the performance of the Kruskal-Wallis test, whereas the ANOVA test does not seem to improve as much or as quickly.

The first result above reflects commonly held wisdom as well as the results presented in [16].

While the simulation results included here are for only three treatment groups, it is not unreasonable to use these as a guide in the case of larger or more complex ANOVA-based models where normality would probably be a more rather than less critical assumption.

Stat Papers (2011) 52:219–231
DOI 10.1007/s00362-009-0224-x

REGULAR ARTICLE

The two-sample *t* test: pre-testing its assumptions does not pay off

Dieter Rasch · Klaus D. Kubinger · Karl Moder

Abstract Traditionally, when applying the two-sample t test, some pre-testing occurs. That is, the theory-based assumptions of normal distributions as well as of homogeneity of the variances are often tested in applied sciences in advance of the tried-for t test. But this paper shows that such pre-testing leads to unknown final type-I- and type-II-risks if the respective statistical tests are performed using the same set of observations. In order to get an impression of the extension of the resulting mis-interpreted risks, some theoretical deductions are given and, in particular, a systematic simulation study is done. As a result, we propose that it is preferable to apply no pre-tests for the t test and no t test at all, but instead to use the Welch-test as a standard test: its power comes close to that of the t test when the variances are homogeneous, and for unequal variances and skewness values $|\gamma_1| < 3$, it keeps the so called 20% robustness whereas the t test as well as Wilcoxon's U test cannot be recommended for most cases.

Keywords Pre-tests · Two-sample t test · Welch-test · Wilcoxon-U test

181

6 Conclusion

From Fig. 2 and in much more detail from Table 2 (but also for all the results left by the authors), we can conclude the following:

- The assumptions underlying the two-sample t test should not be pre-tested.
- The Welch-test should be introduced in text books and statistical program packages as the standard test for comparing expectations.
- The Wilcoxon-U test should not be used in the given context.

182

J Clin Epidemiol Vol. 52, No. 3, pp. 229–235, 1999

0895-4356/99/ $–see front matter
PII S0895-4356(98)00168-1

Increasing Physicians' Awareness of the Impact of Statistics on Research Outcomes: Comparative Power of the t-test and Wilcoxon Rank-Sum Test in Small Samples Applied Research

Patrick D. Bridge[1,*] *and Shlomo S. Sawilowsky*[2]

[1]Department of Family Medicine, Wayne State University School of Medicine and [2]Department of Theoretical and Behavioral Foundations, College of Education, Wayne State University, Detroit, Michigan

ABSTRACT. To effectively evaluate medical literature, practicing physicians and medical researchers must understand the impact of statistical tests on research outcomes. Applying inefficient statistics not only increases the need for resources, but more importantly increases the probability of committing a Type I or Type II error. The t-test is one of the most prevalent tests used in the medical field and is the uniformally most powerful unbiased test (UMPU) under normal curve theory. But does it maintain its UMPU properties when assumptions of normality are violated? A Monte Carlo investigation evaluates the comparative power of the independent samples t-test and its nonparametric counterpart, the Wilcoxon Rank-Sum (WRS) test, to violations from population normality, using three commonly occurring distributions and small sample sizes. The t-test was more powerful under relatively symmetric distributions, although the magnitude of the differences was moderate. Under distributions with extreme skews, the WRS held large power advantages. When distributions consist of heavier tails or extreme skews, the WRS should be the test of choice. In turn, when population characteristics are unknown, the WRS is recommended, based on the magnitude of these power differences in extreme skews, and the modest variation in symmetric distributions. J CLIN EPIDEMIOL 52;3:229–235, 1999. © 1999 Elsevier Science Inc.

KEY WORDS. Research methods; t-test; Wilcoxon Rank-Sum test; nonparametric statistics; parametric statistics; power

183

Should We Abandon the *t*-Test in the Analysis of Gene Expression Microarray Data: A Comparison of Variance Modeling Strategies

Marine Jeanmougin[1,2,3,4]*, Aurelien de Reynies[1], Laetitia Marisa[1], Caroline Paccard[2], Gregory Nuel[3], Mickael Guedj[1,2]

1 Programme Cartes d'Identité des Tumeurs (CIT), Ligue Nationale Contre le Cancer, Paris, France, **2** Department of Biostatistics, Pharnext, Paris, France, **3** Department of Applied Mathematics (MAP5) UMR CNRS 8145, Paris Descartes University, Paris, France, **4** Statistics and Genome Laboratory UMR CNRS 8071, University of Evry, Evry, France

Abstract

High-throughput post-genomic studies are now routinely and promisingly investigated in biological and biomedical research. The main statistical approach to select genes differentially expressed between two groups is to apply a *t*-test, which is subject of criticism in the literature. Numerous alternatives have been developed based on different and innovative variance modeling strategies. However, a critical issue is that selecting a different test usually leads to a different gene list. In this context and given the current tendency to apply the *t*-test, identifying the most efficient approach in practice remains crucial. To provide elements to answer, we conduct a comparison of eight tests representative of variance modeling strategies in gene expression data: Welch's *t*-test, ANOVA [1], Wilcoxon's test, SAM [2], RVM [3], limma [4], VarMixt [5] and SMVar [6]. Our comparison process relies on four steps (gene list analysis, simulations, spike-in data and re-sampling) to formulate comprehensive and robust conclusions about test performance, in terms of statistical power, false-positive rate, execution time and ease of use. Our results raise concerns about the ability of some methods to control the expected number of false positives at a desirable level. Besides, two tests (limma and VarMixt) show significant improvement compared to the *t*-test, in particular to deal with small sample sizes. In addition limma presents several practical advantages, so we advocate its application to analyze gene expression data.

PLoS ONE

September 2010 | Volume 5 | Issue 9

186

The type-I error-rate is often referred to as false-positive rate. It differs from the false-discovery rate (FDR) in the sense that it represents the rate that truly null features are called significant whereas the FDR is the rate that significant features are truly null [21].

Briefly, most of the eight tests are parametric and estimate a gene-by-gene variance: ANOVA (homoscedastic), Welch's *t*-test (heteroscedastic), RVM (homoscedastic), limma (homoscedastic and based on a Bayesian framework) and SMVar (heteroscedastic and based on structural model); we also select two non-parametric approaches with the Wilcoxon's test and the SAM test, which do not rely on assumptions that the data are drawn from a given probability distribution.

187

Table 2. False-positive rate study from simulations.

	M1		M2		M3		M4	
Sample size	**n = 5**	**n = 100**	**n = 5**	**n = 100**	**n = 5**	**n = 100**	**n = 5**	**n = 100**
t-test▼	3.8 – 4.6	4.5 – 5.4	4.0 – 4.8	4.6 – 5.5	3.8 – 4.6	4.7 – 5.6	3.9 – 4.7	4.4 – 5.3
ANOVA	4.5 – 5.2	4.5 – 5.4	4.7 – 5.6	4.6 – 5.5	4.5 – 5.4	4.7 – 5.6	4.5 – 5.3	4.4 – 5.3
Wilcoxon▼	2.8 – 3.5	4.6 – 5.5	2.6 – 3.3	4.5 – 5.4	2.8 – 3.5	4.7 – 5.6	2.7 – 3.4	4.5 – 5.4
SAM	4.6 – 5.5	4.5 – 5.3	4.2 – 5.1	4.5 – 5.4	4.7 – 5.6	4.7 – 5.6	4.3 – 5.2	4.4 – 5.3
RVM▲	5.7 – 6.7	4.5 – 5.4	5.6 – 6.5	4.5 – 5.4	5.4 – 6.3	4.7 – 5.6	5.3 – 6.2	4.7 – 5.5
limma	4.6 – 5.5	4.6 – 5.5	4.2 – 5.1	4.5 – 5.4	4.7 – 5.6	4.7 – 5.6	4.4 – 5.3	4.3 – 5.1
SMVar▲	7.0 – 8.1	4.7 – 5.6	–	–	5.9 – 6.8	4.8 – 5.7	4.6 – 5.5	4.5 – 5.3
VarMixt	4.7 – 5.5	4.6 – 5.5	4.3 – 5.2	4.6 – 5.5	4.8 – 5.6	4.6 – 5.5	4.5 – 5.4	4.5 – 5.3

For small and large samples, this table presents the 95% confidence-interval of false-positive rate obtained by applying a threshold of 0.05 to the *p*-values. Up triangles ▲ (resp. down triangles ▼) indicate an increase (resp. a decrease) of the false-positive rate compared to the expected level of 5%. Two triangles inform of a deviation in both small and large sample sizes.

doi:10.1371/journal.pone.0012336.t002

Table 3. Summary table.

	False-positive rate		Power		In practice	
	Small samples	Large samples	Small samples	Large samples	Ease of use	Execution time
t-test	+	+++	+	+++	+++	+++
ANOVA	+++	+++	+	+++	+++	+++
Wilcoxon	+	+	+	++	+++	++
SAM	+++	+++	+	++	++	++
RVM	+	++	+++	+++	++	+
limma	+++	+++	+++	+++	++	+++
VarMixt	+++	+++	+++	+++	+	+
SMVar	+	+	++	+++	++	+++

This table summarizes the results of our study in terms of false-positive rate, power and practical criteria. The number of "+" indicates the performance, from weak (+), to very good one (+++).

doi:10.1371/journal.pone.0012336.t003

189

Psychological Bulletin
1985, Vol. 97, No. 1, 119-128

0033-2909/85/$00.75

Comparison of the Power of the Paired Samples *t* Test to That of Wilcoxon's Signed-Ranks Test Under Various Population Shapes

R. Clifford Blair
Department of Educational Measurement and Research
University of South Florida

James J. Higgins
Department of Statistics
Kansas State University

Monte Carlo methods were used to assess the relative power of the paired samples *t* test and Wilcoxon's signed-ranks test under ten population shapes. Results of the study indicated that: (a) Each statistic was more powerful than the other in given situations; (b) the power advantages of the *t* test under normal theory were small; (c) in the nonnormal situation, the *t* test never attained more than modest advantages over the Wilcoxon test; (d) the Wilcoxon test was not only more often the more powerful test, but in those situations where it was more powerful, it was often vastly so; and (e) the magnitude of the Wilcoxon's power advantage often increased with increases in sample size. It was concluded that, insofar as these two statistics are concerned, the often-repeated claim that parametric tests are more powerful than nonparametric tests is not justified.

190

Rank Transformations and the Power of the Student *t* Test and Welch *t'* Test for Non-Normal Populations With Unequal Variances

DONALD W. ZIMMERMAN *Carleton University*
BRUNO D. ZUMBO *University of Ottawa*

Canadian Journal of Experimental Psychology, 1993, 47:3, 523-539

Abstract Classical studies have disclosed that parametric significance tests such as *t* and *F* are robust under violation of homogeneity of variance, provided sample sizes are equal. But relatively little is known about effects of unequal variances on nonparametric counterparts of the tests or about non-normality combined with unequal variances. In the present computer simulation study, the Student *t* test and the Welch version of the *t* test (the *t'* test) were performed first on the initial sample values and then on ranks of the sample values. Unequal variances together with unequal *N*'s markedly altered the probability of Type I and Type II errors for normal and for eight kinds of non-normal distributions, including mixed-normal, exponential, lognormal, and Cauchy distributions. Substitution of the Welch *t'* test for the Student *t* test eliminated effects of unequal variances, but not effects of non-normality. The *t* test on ranks, which is equivalent to the Mann-Whitney-Wilcoxon test, was more powerful than the Student *t* test for several non-normal distributions, but exhibited a substantial power loss when variances were unequal. The Welch *t'* test in conjunction with the rank transformation simultaneously counteracted effects of both non-normality and unequal variances.

neutralizado

191

Package 'snpar'

February 20, 2015

Type Package

Title Supplementary Non-parametric Statistics Methods

Version 1.0

Date 2014-08-11

Author Debin Qiu

Maintainer Debin Qiu <debinqiu@uga.edu>

Description contains several supplementary non-parametric statistics methods including quantile test, Cox-Stuart trend test, runs test, normal score test, kernel PDF and CDF estimation, kernel regression estimation and kernel Kolmogorov-Smirnov test.

License GPL (>= 2)

NeedsCompilation no

Repository CRAN

Date/Publication 2014-08-16 10:45:40

R topics documented:

Package 'randtests'

February 20, 2015

Type Package

Title Testing randomness in R

Version 1.0

Date 2014-11-16

Author Frederico Caeiro and Ayana Mateus

Maintainer Frederico Caeiro <fac@fct.unl.pt>

Description Several non parametric randomness tests for numeric sequences

License GPL (>= 2)

LazyLoad yes

NeedsCompilation no

Repository CRAN

Date/Publication 2014-11-17 00:02:16

R topics documented:

193

EXAMPLES OF CORRELATED OBSERVATIONS

Correlated data arise when pairs or clusters of observations are related and thus are more similar to each other than to other observations in the dataset. Observations may be related because they come from the same subject—for example, when subjects are measured at multiple time points (repeated measures) or when subjects contribute data on multiple body parts, such as both eyes, hands, arms, legs, or sides of the face. Observations from different subjects also may be related—for example, if the dataset contains siblings, twin pairs, husband-wife pairs, control subjects who have been matched to individual cases, or patients from the same physician practice, clinic, or hospital. Cluster randomized trials, which are performed to assign interventions to groups of people rather than to individual subjects (for example, schools, classrooms, cities, clinics, or communities), also are a source of correlated data because subjects within a cluster will likely have more similar outcomes than subjects in other clusters.

THE CONSEQUENCES OF IGNORING CORRELATIONS

Many statistical tests assume that observations are independent. The application of these tests to correlated observations will lead to the overestimation of *P* values in certain cases (when one considers within-subject or within-cluster effects) and underestimation in others (when one considers between-subject or between-cluster effects). These errors are illustrated in the following sections.

194

Example 1. The authors of a recent randomized, blinded trial compared the efficacy of 2 sunscreens by using a split-face design [1]. Fifty-six subjects applied sunscreen with a sun protective factor (SPF) of 85 to one side of their face and an SPF of 50 to the other side of their face (the application sides were randomly chosen, and the sunscreen types were concealed) before spending 5 hours participating in outdoor sports on a sunny day. Investigators determined the occurrence of sunburn on each side of the participants' faces at the end of the day. A person's tendency to burn on one side of his or her face is highly correlated with his or her tendency to burn on the other side. However, when the data were analyzed, these correlations were ignored: the authors reported that 1 of 56 participants were burned on the SPF 85 side of the face, whereas 8 of 56 were burned on the SPF 50 side ($P = .03$, Fisher exact test, Table 1a). This analysis treats all observations equally, as if there are 112 unrelated sides of the face. Table 1b shows the correct way to present and analyze the data.

Table 1a. *Original data table from Russak et al (1)*

Sun Protection Factor	Sunburned	Not Sunburned
85	1	55
50	8	48

$P = .03$, Fisher exact test.
Reprinted with permission [1].

Table 1b. *Correct presentation of the data from Russak et al (1)*

	SPF-50 Side	
SPF-85 Side	**Sunburned**	**Not Sunburned**
Sunburned	1	0
Not sunburned	7	48

$P = .0156$, McNemar exact test.
Reprinted with permission [1].

Volunteers who burned on both sides of their face (n = 1) or neither side (n = 48) do not help us to discriminate between the performance of SPF 85 and SPF 50; only the volunteers who burned on a single side (n = 7) are informative. The correct analysis—called the McNemar exact test [2]—focuses only on these discordant subjects. In all 7 cases, the sunburn occurred on the SPF 50 side. The 2-sided *P* value associated with this extreme outcome (a 7-0 split) is .0156 (determined by a binomial distribution with n = 7 and $P = .5$). Thus the difference between the sunscreens is actually more significant than the authors have reported. Although the *P* values (.03 vs .0156) do not differ enough to change the study's conclusions, they can differ markedly in many cases, as the next example illustrates.

Example 2. Consider a simple hypothetical dataset in which investigators conducted a study with twins to examine the association of exercise with blood pressure. Six pairs of twins reported their physical activity levels and had their blood pressures measured. Investigators hypothesized that the more active twins would have lower blood pressures than the less active twins. The results are presented in Table 2.

The mean blood pressure for the more active twins is 3.5 mm Hg lower than for the less active twins (76.5 vs 80.0). If we ignore the correlations and analyze the data as 2 independent groups, this difference is not statistically significant ($P = .41$, 2-sample t-test). However, if we correctly analyze these data by focusing on the differences within twin pairs, it is statistically significant ($P = .02$, paired t-test). The P value is reduced because the variation in blood pressure within twin pairs (standard deviation = 2.6) is considerably less than between unrelated twins (standard deviation = 7.0 or 7.1) and because the paired t-test only has to account for one source of variability (variability within pairs) rather than 2 sources (variability from two groups of twins).

Table 2. *A simple hypothetical dataset involving correlated data (twin pairs)*

Twin Pair	Diastolic Blood Pressure in the Less Active Twin, mm Hg	Diastolic Blood Pressure in the More Active Twin, Mm Hg	Difference (More Active – Less Active), Mm Hg
1	87	82	−5
2	88	83	−5
3	80	78	−2
4	79	80	+1
5	77	71	−6
6	69	65	−4
Mean (SD)	80.0 (7.0)	76.5 (7.1)	−3.5 (2.6)
Test statistic	Two-sample t-test (incorrect analysis): $T_{10} = \frac{-3.5}{\sqrt{\frac{7.0^2}{6} + \frac{7.0^2}{6}}} = -0.86$, $p = .41$		Paired t-test (correct analysis): $T_5 = \frac{-3.5}{\sqrt{\frac{2.6^2}{6}}} = -3.31$, $p = .02$

196

Table 3. *A simple hypothetical dataset from a trial in which 50 subjects were randomly assigned to receive active drug (n =25) or placebo (n = 25) in both eyes*

Analysis	N (%) of Eyes Improving in the Control Group	N (%) of Eyes Improving in the Treatment Group	*P* Value	Odds Ratio and 95% Confidence Interval
Assuming eyes are independent*	17/50 (34)	27/50 (54)	.046	2.28 (1.02–5.11)
Correcting for within-subject correlation†	17/50 (34)	27/50 (54)	.11	2.28 (0.83–6.28)

*Data were analyzed with unconditional logistic regression.
†Data were analyzed by the use of a generalized estimating equation, correcting for within-subject correlation.

Example 1. In a hypothetical trial, 50 patients with bilateral eye disease were randomly assigned to receive an active drug or a placebo solution in both eyes (sample size per group is 25 patients [50 eyes]). Treatment was considered a success if symptoms improved by more than 50% in a given eye. Table 3 shows hypothetical results from this trial.

Strong agreement between eyes was found—80% of the subjects had the same outcome in both eyes (κ coefficient = .60). Thus treating the data as if there are 100 independent eyes will overstate the evidence for the drug's effectiveness. The informative sample size is actually somewhere between 100 and 50 (if there were perfect agreement between eyes, a subject's second eye would contribute no independent evidence of the drug's effectiveness and the sample size would be 50). The incorrect analysis (a χ^2 test or logistic regression) yields an artificially low *P* value of .046, whereas the correct analysis (a generalized estimating equation, corrected for within-subject correlation) yields a nonsignificant result of $P = .11$.

197

Table 4. *A hypothetical cluster-randomized trial, from Calhoun et al (4)*

Analysis	Average Charting Errors From Control Physicians (n = 40 Patients, 4 Physicians)	Average Charting Errors From Treated Physicians (n = 40 Patients, 4 Physicians)	*P* Value
Assuming patients are independent*	2.75	1.7	<.0001
Correcting for within-physician correlation†	2.75	1.7	.273

*Data were analyzed with a 2-sample *t*-test.
†Data were analyzed by the use of hierarchical linear modeling.

Example 2. Cluster-randomized trials are a common source of correlated data, but researchers often neglect the correlations in their analyses [3,4]. Calhoun et al [4] present a hypothetical example that shows the consequence of this failure. In this hypothetical randomized trial of an intervention to reduce physician error, 8 physicians were randomly assigned to a reduced shift length (n = 4) or control condition (n = 4). The outcome was the average number of charting errors per patient; data were obtained on 10 patients per physician for a total of 80 patients. Table 4 shows results from this hypothetical trial.

Observations made by the same physician will be highly correlated. For example, 2 of the 4 physicians in the intervention group are highly conscientious individuals who made no charting errors during the study period; thus it is clear that these 2 physicians each contribute just 1 unit of evidence for the intervention's effectiveness, not 10. If the data are analyzed as 80 independent observations (with use of a 2-sample *t*-test), the *P* value is highly significant, but the correct analysis (a hierarchical linear model) yields a nonsignificant result of *P* = .273.

t-tests, non-parametric tests, and large studies—a paradox of statistical practice?

Morten W Fagerland

BMC Medical Research Methodology 2012, **12**:78

Abstract

Background: During the last 30 years, the median sample size of research studies published in high-impact medical journals has increased manyfold, while the use of non-parametric tests has increased at the expense of t-tests. This paper explores this paradoxical practice and illustrates its consequences.

Methods: A simulation study is used to compare the rejection rates of the Wilcoxon-Mann-Whitney (WMW) test and the two-sample t-test for increasing sample size. Samples are drawn from skewed distributions with equal means and medians but with a small difference in spread. A hypothetical case study is used for illustration and motivation.

Results: The WMW test produces, on average, smaller *p*-values than the t-test. This discrepancy increases with increasing sample size, skewness, and difference in spread. For heavily skewed data, the proportion of $p < 0.05$ with the WMW test can be greater than 90% if the standard deviations differ by 10% and the number of observations is 1000 in each group. The high rejection rates of the WMW test should be interpreted as the power to detect that the probability that a random sample from one of the distributions is less than a random sample from the other distribution is greater than 50%.

Conclusions: Non-parametric tests are most useful for small studies. Using non-parametric tests in large studies may provide answers to the wrong question, thus confusing readers. For studies with a large sample size, t-tests and their corresponding confidence intervals can and should be used even for heavily skewed data. TCL

Keywords: T-test, Non-parametric test, Wilcoxon-Mann-Whitney test, Welch test, Sample size, Statistical practice

199

Table 1 Trends in the use of t-tests and non-parametric tests in the NEJM

Statistical procedure	1978–1979	1989	2004–2005
t-tests*	44%	39%	26%
Non-parametric tests†	11%	21%	27%

*one-sample, two-sample, and matched-pair [2].
†Wilcoxon-Mann-Whitney, sign, and Wilcoxon signed rank sum [2].

used. The Brunner-Munzel test, a non-parametric test that adjusts for unequal variances, may be used as an alternative to the WMW test. It is not widely available in

200

Neuropsychologia. 2010 Jan;48(1):341-3. doi: 10.1016/j.neuropsychologia.2009.09.016.

Inappropriate usage of the Brunner-Munzel test in recent voxel-based lesion-symptom mapping studies.

Medina J[1], Kimberg DY, Chatterjee A, Coslett HB.

⊕ Author information

Abstract

Voxel-based lesion-symptom mapping (VLSM) techniques have been important in elucidating structure-function relationships in the human brain. Rorden, Karnath, and Bonilha (2007) introduced the non-parametric Brunner-Munzel rank order test as an alternative to parametric tests often used in VLSM analyses. However, the Brunner-Munzel statistic produces inflated z scores when used at any voxel where there are less than 10 subjects in either the lesion or no lesion groups. Unfortunately, a number of recently published VLSM studies using this statistic include relatively small patient populations, such that most (if not all) examined voxels do not meet the necessary criteria. We demonstrate the effects of inappropriate usage of the Brunner-Munzel test using a dataset included with MRIcron, and find large Type I errors. To correct for this we suggest that researchers use a permutation derived correction as implemented in current versions of MRIcron when using the Brunner-Munzel test.

201

© Journal of the American Statistical Association
December 1971, Volume 66, Number 336
Theory and Methods Section

On the Sign Test for Symmetry

JOSEPH L. GASTWIRTH*

This article studies the effect of estimating the center of symmetry by the sample mean on the sign test for symmetry. Under the null hypothesis, the modified sign test is not distribution-free. Indeed, the asymptotic variance of the modified sign test can be quite different from that of the standard sign test.

Testing symmetry with a procedure, combining the sign test and the signed rank test

by P. A. R. KOOPMAN *

Summary A procedure for testing symmetry is described which combines the ease of the sign test with the efficiency of the signed rank test.

Statistica Neerlandica

Year:1979
Month:
Day:
Volume:33
Issue:3
First page:137
Last page:142

203

A Distribution-Free Test for Symmetry Based on a Runs Statistic

THOMAS P. McWILLIAMS*

I present a simple test, based on a runs statistic, for symmetry of a continuous distribution about a known median. The statistic has a binomial sampling distribution and desirable invariance properties. Monte Carlo studies demonstrate that, for a wide variety of alternative asymmetric distributions, the test is more powerful than tests proposed by Butler (1969), Rothman and Woodroofe (1972), or Hill and Rao (1977).

KEY WORDS: Invariant test; Runs test; Test of symmetry.

Journal of the American Statistical Association
December 1990, Vol. 85, No. 412, Theory and Methods

Journal of Applied Statistics, Vol. 21, No. 5, 1994 409

Distribution-free test for symmetry based on Wilcoxon two-sample test

I. H. TAJUDDIN, *King Saud University, Riyadh*

SUMMARY *McWilliams introduced a test of symmetry based on a runs statistic for a continuous distribution about a known median. His test statistic R* performs better than other competitors considered in his study. In this paper, we present a conditional test for symmetry based on a Wilcoxon two-sample test. The proposed test turns out to be competitive with the test of McWilliams.*

205

An Asymptotically Distribution-Free Test for Symmetry Versus Asymmetry

RONALD H. RANDLES, MICHAEL A. FLIGNER,
GEORGE E. POLICELLO II, and DOUGLAS A. WOLFE*

An asymptotically distribution-free procedure is proposed for testing whether a univariate population is symmetric about some unknown value versus a broad class of asymmetric distribution alternatives. The consistency class of the test is discussed and two competing tests are described, one based on the sample skewness, and the other on Gupta's nonparametric procedure. A Monte Carlo study shows that the proposed test is superior to either competitor since it maintains the designated α levels fairly accurately even for sample sizes as small as 20, while displaying good power for detecting asymmetric distributions.

KEY WORDS: Symmetry; Asymmetry; Asymptotically distribution free; Nonparametric; U statistic.

March 1980, Volume 75, Number 369

206

The Canadian Journal of Statistics
Vol. 24, No. 3, 1996, Pages 349–361
La Revue Canadienne de Statistique

A simple test of symmetry about an unknown median*

Paul CABILIO and Joe MASARO

Acadia University

Key words and phrases: Symmetry, skewness, asymmetry, deviations from sample median, Monte Carlo study, robustness of significance level.
AMS 1991 subject classifications: 62G10, 62G35.

ABSTRACT

A simple test statistic for testing symmetry of a distribution function about an unknown value is presented. The asymptotic distributions under symmetry and asymmetry are derived. Using the normal as a "calibration" distribution, the critical values of the test are calculated by Monte Carlo methods. Comparisons with other tests indicate that this procedure performs well.