Psychology 350: Advanced statistics and programming in R
Student's t and Cohen's d

William Revelle

Department of Psychology
Northwestern University
Evanston, Illinois USA

April, 2022

Outline

Gosset and the t-test

1. Perhaps the most important inferential statistic ever developed, the `t.test` was originally published anonymously by Student (1908)
2. William Sealy Gosset, while working for Guinness Brewery, solved the problem of small sample inferential tests
3. Normal theory allowed one to test differences between means if the samples were large enough ($N > 30$ per group)
4. The problem was that that small samples, the standard deviation varied as a function of the mean.
5. Gosset solved the problem with the t-distribution.
6. His original data came from a paper on the effect of drugs on sleeping in psychotic patients (the `cushny` data set (Cushny & Peebles, 1905)).

The t-test is just a difference in means adjusted by standard errors

1. The differences of means

$$\bar{X}_1 - \bar{X}_2$$

2. The standard error of the Mean is just

$$\sqrt{\frac{\sigma^2_{pooled}}{N-1}}$$

(the standard deviation divided by $\sqrt{N}$)

3. The pooled variance is the weighted average of the two variances

$$\sigma^2_{pooled} = \frac{N_1\sigma_1^2 + N_2\sigma_2^2}{N_1 + N_2}$$

The original data set from Cushny used by Student

R code

```
cushny   #show the data
colMeans(cushny)
with(cushny, t.test(drug1,drug2L,paired=TRUE))   #do a t-test
```

```
   Control drug1 drug2L drug2R delta1 delta2L delta2R
1      0.6   1.3    2.5    2.1    0.7     1.9     1.5
2      3.0   1.4    3.8    4.4   -1.6     0.8     1.4
3      4.7   4.5    5.8    4.7   -0.2     1.1     0.0
4      5.5   4.3    5.6    4.8   -1.2     0.1    -0.7
5      6.2   6.1    6.1    6.7   -0.1    -0.1     0.5
6      3.2   6.6    7.6    8.3    3.4     4.4     5.1
7      2.5   6.2    8.0    8.2    3.7     5.5     5.7
8      2.8   3.6    4.4    4.3    0.8     1.6     1.5
9      1.1   1.1    5.7    5.8    0.0     4.6     4.7
10     2.9   4.9    6.3    6.4    2.0     3.4     3.5
>
Control   drug1  drug2L  drug2R  delta1 delta2L delta2R
   3.25    4.00    5.58    5.57    0.75    2.33    2.32

        Paired t-test

data:  drug1 and drug2L
t = -4.0621, df = 9, p-value = 0.002833
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 -2.4598858 -0.7001142
sample estimates:
mean difference
          -1.58
```

The data were paired (each person was repeated

1. The power of the Cushny design was that each subject was repeated.
2. This reduces the standard error by removing the effect of the individual differences
3. We can also do the t-test with out the paired design

R code

```
with(cushny, t.test(drug1,drug2L,paired=FALSE))
```

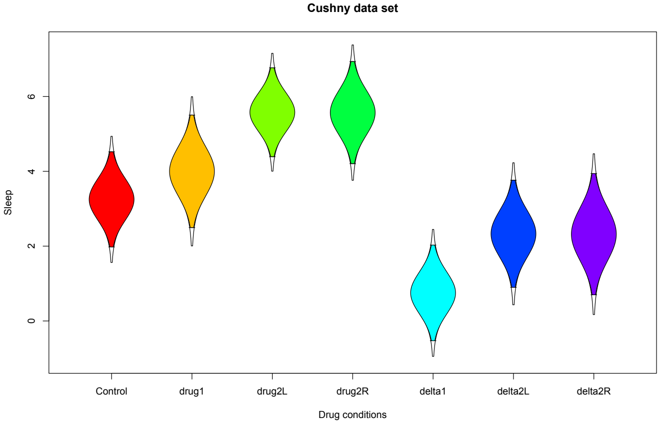
```
Welch Two Sample t-test

data:  drug1 and drug2L
t = -1.8647, df = 17.086, p-value = 0.0795
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.3670084  0.2070084
sample estimates:
mean of x mean of y
     4.00      5.58
```

Several ways to show the data

R code

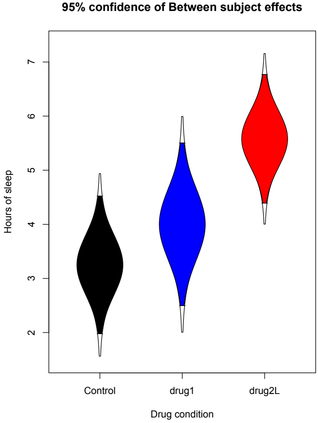
```
   #all of the data with rainbow colors

error.bars(cushny,col=rainbow(8), main="Cushny data set",
  ylab= "Sleep",  xlab="Drug conditions")
par(mfrow=c(1,2) )   # a two panel graph

 error.bars(cushny[1:3],within=FALSE,
    ylab="Hours of sleep", xlab="Drug condition",
      main="95% confidence of Between subject effects")
 error.bars(cushny[1:3],within=TRUE,ylab="Hours of sleep",
    xlab="Drug condition",
       main="95% confidence of Within subject effects")

par(mfrow=c(1,1) )   #set it back to the default
```

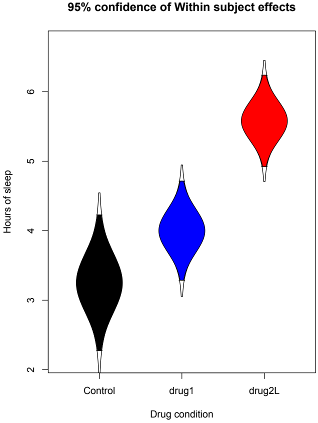
The Cushny data set used by Gosset



The Cushny data set used by Gosset

Can show the results controlling for the within subject nature of the data





Other ways of finding t

Find the descriptive statistics and then use m2t

R code

```
describe(cushny)
m2t(4,5.58, 2.1,1.66,n1=10,n2=10)
```

```
describe(cushny)
        vars  n mean   sd median trimmed  mad  min max range  skew kurtosis   se
Control    1 10 3.25 1.78   2.95    3.21 1.63  0.6 6.2   5.6  0.20    -1.23 0.56
drug1      2 10 4.00 2.10   4.40    4.04 2.59  1.1 6.6   5.5 -0.25    -1.67 0.66
drug2L     3 10 5.58 1.66   5.75    5.66 1.41  2.5 8.0   5.5 -0.30    -0.99 0.53
drug2R     4 10 5.57 1.91   5.30    5.66 1.56  2.1 8.3   6.2 -0.09    -1.07 0.60
delta1     5 10 0.75 1.79   0.35    0.67 1.56 -1.6 3.7   5.3  0.42    -1.30 0.57
delta2L    6 10 2.33 2.00   1.75    2.24 2.45 -0.1 5.5   5.6  0.28    -1.66 0.6

t =  -1.866512 df = 18  with probability =
 0.07834775
```

Probability values and NHST versus effect size

1. The t-test (and its generalization to the F test) allows us to test whether a difference between two means differs from 0, but does not tell us how big the difference is.
2. Null Hypothesis Significance Tests compare the hypothesis of No difference (the Nill hypothesis) versus there is a difference.
3. Cohen (1988, 1994) suggested we should compare differences in terms of standardized differences in order to generalize across studies.
4. if $d = \frac{\bar{X}_1 - \bar{X}_2}{\sigma}$ then $t = d * df$ (use `cohen.d` for raw data, `t2d` for conversions from a t-test result)
5. Funder & Ozer (2019) suggest that we should use the correlation coefficient, r, as our measure of effect size. ($r = \frac{d}{d^2+4}$) or use the `d2r` function.

Effect size examples using GERAS data set

1. Gruber, Distlberger, Scherndl, Ortner & Pletzer (2020) reported item scores for males and females on the Gender Related Attributes Survey
2. The GERAS has three broad item types to show gender differences: Personality, Cognitive interests, Interests and Activities
3. Eagly & Revelle (2022) use these data to show the power of aggregation.
4. The `GERAS` data set has 51 items, and several broad scales. It is included in the *psychTools*
5. We treat gender as a binary variable (because that is the way Gruber et al. report them)
6. Item content is available in the `GERAS.dictionary`, item scores in `GERAS.items`, scale scores in `GERAS.scales` .

Using the GERAS data set

Included in *psychTools*

R code

```
library{psychTools}
describe(GERAS.scales)
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
M.pers	1	471	4.19	0.62	4.20	4.20	0.59	2.40	5.80	3.40	-0.17	-0.15	0.03
F.pers	2	471	4.84	0.61	4.90	4.88	0.59	2.60	6.00	3.40	-0.75	0.65	0.03
M.cog	3	471	4.52	0.84	4.57	4.54	1.06	2.29	6.00	3.71	-0.28	-0.63	0.04
F.cog	4	471	4.80	0.74	4.86	4.86	0.85	2.29	6.00	3.71	-0.67	0.06	0.03
M.act	5	471	3.95	0.97	4.00	3.95	1.11	2.00	6.00	4.00	-0.01	-0.75	0.04
F.act	6	471	3.87	0.90	3.88	3.86	1.11	2.00	6.00	4.00	0.08	-0.76	0.04
Pers	7	471	3.68	0.49	3.65	3.67	0.44	2.40	5.60	3.20	0.31	0.52	0.02
Cog	8	471	3.86	0.57	3.86	3.85	0.53	2.43	5.64	3.21	0.05	-0.33	0.03
Act	9	471	4.04	0.71	4.00	4.03	0.83	2.19	5.88	3.69	0.10	-0.60	0.03
M	10	471	4.21	0.60	4.20	4.21	0.65	2.64	5.56	2.92	-0.01	-0.47	0.03
F	11	471	4.52	0.55	4.52	4.54	0.59	2.96	5.76	2.80	-0.28	-0.36	0.03
MF.all	12	471	3.84	0.46	3.82	3.84	0.47	2.76	5.24	2.48	0.10	-0.37	0.02
gender	13	471	1.49	0.50	1.00	1.49	0.00	1.00	2.00	1.00	0.05	-2.00	0.02

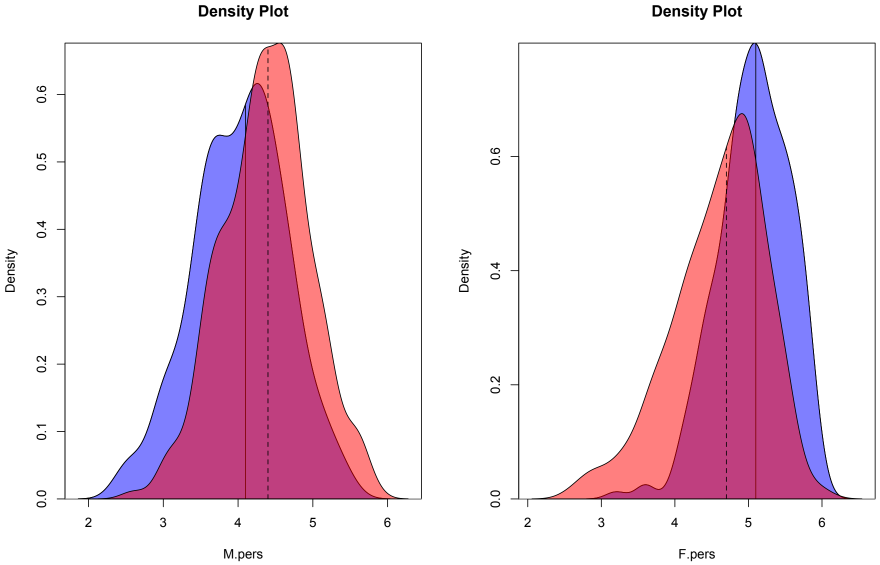
Show distributions by group

1. One way to show distributions by group is to plot density by score and group
2. This uses the densityBy function.

R code

```
par(mfrow=c(1,2))  #set up a two panel graph
densityBy(M.pers  ~ gender, data= GERAS.scales)
densityBy(F.pers  ~ gender, data= GERAS.scales)
par(mfrow=c(1,1)  #return to one panel displays
```

Density for two scales



Lets consider some effect sizes and graphical displays

```
cd.small <- cohen.d(GERAS.scales[1:100,1:6],
    group=GERAS.scales[1:100,"gender"])
Call: cohen.d(x = GERAS.scales[1:100, 1:6],
       group = GERAS.scales[1:100,
    "gender"])
Cohen d statistic of difference between two means
       lower effect upper
M.pers  0.08   0.51  0.94
F.pers -1.44  -1.00 -0.55
M.cog   0.54   0.99  1.43
F.cog  -0.44  -0.02  0.40
M.act   0.10   0.53  0.96
F.act  -2.13  -1.65 -1.17

Multivariate (Mahalanobis) distance
       between groups
[1] 2.1
r equivalent of difference between two means
M.pers F.pers  M.cog  F.cog  M.act  F.act
  0.23  -0.42   0.42  -0.01   0.24  -0.61
```

```
cd.big <- cohen.d(GERAS.scales[,1:6],
       group=GERAS.scales[,"gender"])
Call: cohen.d(x = GERAS.scales[, 1:6],
       group = GERAS.scales[, "gender"])
Cohen d statistic of difference between two mea
       lower effect upper
M.pers  0.39   0.57  0.76
F.pers -1.05  -0.86 -0.67
M.cog   0.46   0.65  0.84
F.cog  -0.59  -0.40 -0.22
M.act   0.64   0.83  1.01
F.act  -1.81  -1.61 -1.40

Multivariate (Mahalanobis) distance
      between groups
[1] 2
r equivalent of difference between two means
M.pers F.pers  M.cog  F.cog  M.act  F.act
  0.28  -0.39   0.31  -0.20   0.38  -0.63
```

Notice that the estimates are very similar, but the confidence intervals of the estimates are smaller for larger samples

the t and p values are available in the objects

R code

```
names(cd.big)
summary.stats.df <- data.frame(small.t = cd.small$t,big.t = cd.big$t,
round(summary.stats.df,3)
cd.small$Call   #what where the calling commands
cd.big$Call
```

```
names(cd.big)
 [1] "cohen.d"     "hedges.g"    "M.dist"      "r"           "t"           "n"
    "p"            "wt.d"        "descriptive"    "se"          "dict"         "Call"
> summary.stats.df <- data.frame(small.t = cd.small$t,
   big.t = cd.big$t, small.p = cd.small$p, big = cd.big$p)
> round(summary.stats.df,3)
       small.t   big.t small.p big
M.pers   2.342   6.217   0.021   0
F.pers  -4.558  -9.283   0.000   0
M.cog    4.535   7.041   0.000   0
F.cog   -0.096  -4.371   0.924   0
M.act    2.412   8.943   0.018   0
F.act   -7.576 -17.377   0.000   0

cd.small$Call   #what where the calling commands
cohen.d(x = GERAS.scales[1:100, 1:6], group = GERAS.scales[1:100,
    "gender"])
> cd.big$Call
cohen.d(x = GERAS.scales[, 1:6], group = GERAS.scales[, "gender"])
>
```

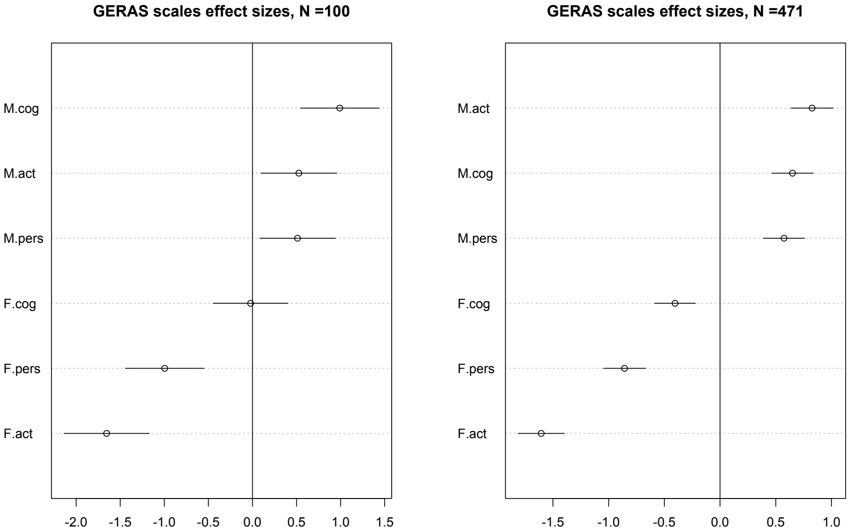
Graphic displays of effect sizes

1. the `error.dots` function will take the output of `cohen.d` and draw effect sizes

R code

```
error.dots(cd.small,main="GERAS scales effect sizes, N =100")
abline(v=0)    #make  a vertical line at 0


error.dots(cd.big,main="GERAS scales effect sizes, N =471")
abline(v=0)
```

GERAS effect sizes and sample size



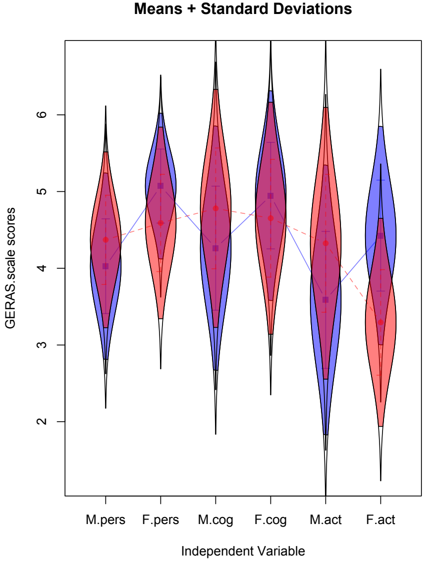
Effect sizes show standardized differences

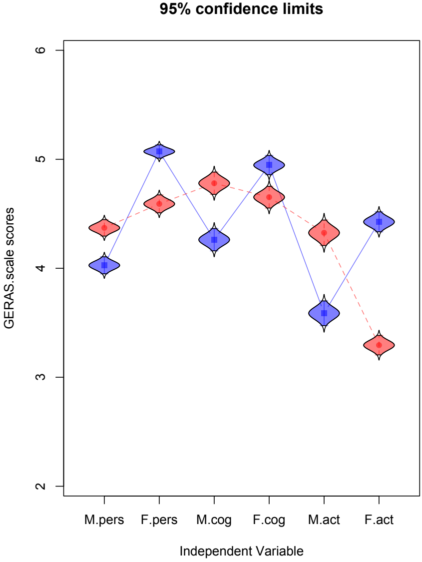
1. Another way to compare two groups is to plot their means and standard deviations
2. Or their means and standard errors. This is the more traditional way, but emphasizes statistically “significant: differences at the cost of over emphasizing small effects.
3. Consider two `error.bars.by` plots

R code

```
error.bars.by(GERAS.scales[,1:6],group=GERAS.scales[,"gender"],
       ylab = "GERAS.scale scores",by.var=FALSE,
              col=c("blue","red"),sd=TRUE,density=-.5)
 error.bars.by(GERAS.scales[,1:6],group=GERAS.scales[,"gender"],
 ylab = "GERAS.scale scores",by.var=FALSE,
      col=c("blue","red"),sd=FALSE,density=-.5)
```

error.bars with standard deviations or standard errors





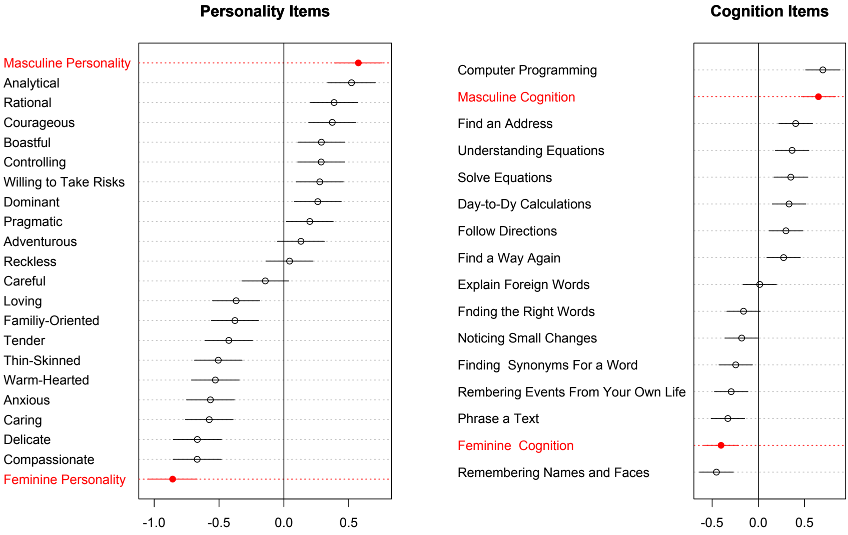
The power of aggregation

1. At the item level, most items do not show big effects
2. But, if we aggregate items within domains, the effects are bigger
3. Aggregating across domains makes these effects even bigger.

Creating multiple plots in one set

```
 cd.pers <- cohen.d(GERAS.combined[,p.items],GERAS.combined[,"gender"],
 dictionary=GERAS.dictionary["Content"])
cd.cog <-  cohen.d(GERAS.combined[,cog.items],GERAS.combined[,"gender"],
        dictionary=GERAS.dictionary["Content"])
cd.act <-  cohen.d(GERAS.combined[,act.items],GERAS.combined[,"gender"],
         dictionary=GERAS.dictionary["Content"])

error.dots(cd.pers ,main="Personality Items",select=c(2:21))
 error.dots(cd.pers,  main="Personality Items", select =c(1,22), col="red" , add=TRUE  ,
         bg="red",lcolor="red",fg="red")
abline(v=0)
 error.dots(cd.cog ,main="Cognition Items", select = c(1,3:14,16))
 error.dots(cd.cog ,main="Cognition Items", select =c(2,15), col="red" , add=TRUE  ,
          bg="red",lcolor="red",fg="red")
 abline(v=0)

error.dots(cd.act ,main="Interests and Activities Items", select =c(2:17))
error.dots(cd.act ,main="Interests and Activities Items", select =c(1,18),
          col="red" , add=TRUE  ,
          bg="red",lcolor="red",fg="red")
abline(v=0)
```

GERAS item level analyis

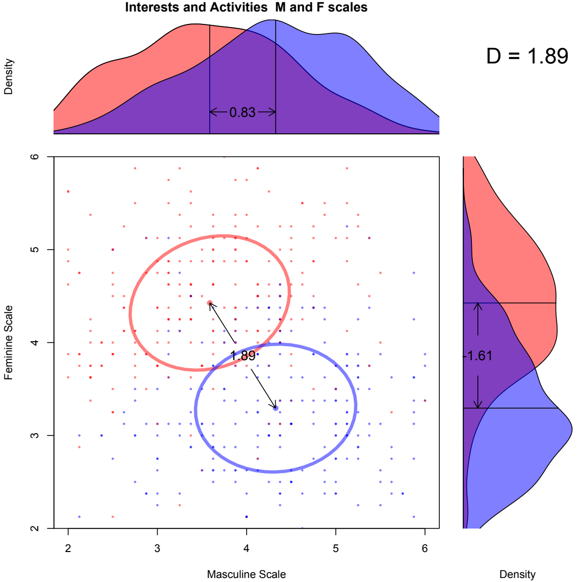


We can also show the bivariate scatter plot by groups

1. This uses the `scatterHist` function.
2. It can display the Mahalobinis distance as well.

R code
```
scatterHist(F.act ~ M.act+ gender, data=GERAS.scales, cex.point=.3,
    smooth=FALSE,xlab="Masculine Scale",ylab="Feminine Scale",
      correl=FALSE, d.arrow=TRUE, col=c("red","blue"),
      bg=c("red","blue"), lwd=4,
       title="Interests and Activities  M and F scales",
   cex.cor=2,cex.arrow=1.25)
```

Showing groups in scatter plot



Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed ed.). Hillsdale, N.J.: L. Erlbaum Associates.

Cohen, J. (1994). The earth is round ($p < .05$). *American psychologist*, *49*(12), 997–1003.

Cushny, A. R. & Peebles, A. R. (1905). The action of optical isomer. ii. hyoscines. *The Journal of Physiology*, *32*(501-510).

Eagly, A. H. & Revelle, W. (2022). Understanding the Magnitude of Psychological Differences between women and men requires seeing the forest and the trees (in press). *Perspectives in Psychological Science (in press)*.

Funder, D. C. & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, *2*(2), 156–168.

Gruber, F. M., Distlberger, E., Scherndl, T., Ortner, T. M., & Pletzer, B. (2020). Psychometric properties of the multifaceted gender-related attributes survey (GERAS). *European Journal of Psychological Assessment*, *36*(4), 612–623.

Student (1908). The probable error of a mean. *Biometrika*, *6*(1), 1–25.