Normal differences

33 marks

In this question, we revisit a study published in *Frontiers in Psychology* from 2011 on gender differences.

To quote from the paper:

> "Men and women have obviously different biological roles when it comes to propagation of the species, but how much they differ psychologically is a more controversial question, one that requires empirical research to answer adequately. Whether the underlying causes of psychological gender differences are evolutionary or socio-cultural, understanding how men and women differ in the ways in which they think, feel, and behave can shed light on the human condition.
>
> "The study of personality is particularly useful in attempting to examine psychological differences between genders. Personality is often conceptualized as the extent to which someone displays high or low levels of specific traits. Traits are the consistent patterns of thoughts, feelings, motives, and behaviors that a person exhibits across situations (Fleeson and Gallagher, 2009). That is, someone who scores high on a trait will exhibit psychological states related to that trait more often and to a greater extent than individuals who score low on that trait.
>
> "Gender differences in personality traits are often characterized in terms of which gender has higher scores on that trait, on average. For example, women are often found to be more agreeable than men (Feingold, 1994; Costa et al., 2001). This means that women, on average, are more nurturing, tender-minded, and altruistic more often and to a greater extent than men. However, such a finding does not preclude the fact that men may also experience nurturing, tender-minded, and altruistic states, and that some men may even score higher in these traits than some women. The goal of investigating gender differences in personality, therefore, is to elucidate the differences among general patterns of behavior in men and women on average, with the understanding that both men and women can experience states across the full range of most traits. Gender differences in terms of mean differences do not imply that men and women only experience states on opposing ends of the trait spectrum; on the contrary, significant differences can exist along with a high degree of overlap between the distributions of men and women (Hyde, 2005)."

The entire paper can be found at https://doi.org/10.3389/fpsyg.2011.00178 and the results and brief description can be found in the three files

- `males.csv`
- `females.csv` and
- `source.txt`

Assuming you have stored the above csv files in a directory/file folder called "data" within the folder containing this question file, you can read the data into R as follows:

```
# Assume the directory is the same as that of this file
# Otherwise set it to wherever it is ... note you may need
# to change the / to \\ or some other separator for file paths
# See .Platform$file.sep on your machine.
dataDirectory <- "./data"

# helper function
path_concat <- function(path, ...,
                        fsep = .Platform$file.sep) {
```

```
                        paste(path, ..., sep = fsep)}
# read in the data
femaleTraits <- read.csv(path_concat(dataDirectory, "females.csv"),
                         header = TRUE)
maleTraits <- read.csv(path_concat(dataDirectory, "males.csv"),
                       header = TRUE)
traits <- maleTraits$Trait
set.seed(123454321)
```

You now have two data frames `femaleTraits` and `maleTraits`. The variates Trait, Mean, SD are the summary results of females and males, respectively, for the 14 different psychological traits: Enthusiasm, Assertiveness, Compassion, Politeness, Industriousness, Orderliness, Volatility, Withdrawal, Intellect, Openness, Extraversion, Agreeableness, Conscientiousness, Neuroticism, stored in that order in `traits`, with each one measured on a five point scale (larger the value the stronger is that trait).

Based on the information provided, we will explore these differences in R. We begin with a few queries via some simple R functions. **All answers should be expressed using R code** (even though many could be answered simply by looking at the values since there are so few).

Show your code for every part.

a. **(2 marks)** For which traits is

i. the mean value for males greater than that for females?

```
test1 <- maleTraits$Mean > femaleTraits$Mean
n <- 1
test1List <- list()
for(i in 1:14)  {
  if(test1[i] == TRUE){
     test1List[n] <- maleTraits$Trait[i]
     n <- n+1
  }
}
test1unlist <- unlist(test1List)
test1unlist
```

```
## [1] "Assertiveness"   "Industriousness" "Intellect"
```

ii. the mean value for females greater than that for males?

```
#library(dplyr)
test2 <- maleTraits$Mean < femaleTraits$Mean
n <- 1
test2List <- list()
for(i in 1:14)  {
  if(test2[i] == TRUE){
     test2List[n] <- maleTraits$Trait[i]
     n <- n+1
  }
}
test2unlist <- unlist(test2List)
test2unlist
```

```
##  [1] "Enthusiasm"        "Compassion"        "Politeness"
##  [4] "Orderliness"       "Volatility"        "Withdrawal"
##  [7] "Openness"          "Extraversion"      "Agreeableness"
## [10] "Conscientiousness" "Neuroticism"
```

b. **(2 marks)** There are two handy functions in R called `which.min()` and `which.max()` which determine the index of the (first) minimum or maximum of a numeric vector. Use these as appropriate to identify those traits having the *highest* and those having the *lowest* mean score for

   i. females

```
index_max <- which.max(femaleTraits[,2])
femaleTraits[index_max,]$Trait
```

```
## [1] "Compassion"
```

```
index_min <- which.min(femaleTraits[,2])
femaleTraits[index_min,]$Trait
```

```
## [1] "Volatility"
```

   ii. males

```
index_max2 <- which.max(maleTraits[,2])
maleTraits[index_max2,]$Trait
```

```
## [1] "Compassion"
```

```
index_min2 <- which.min(maleTraits[,2])
maleTraits[index_min2,]$Trait
```

```
## [1] "Volatility"
```

c. **(3 marks)** Consider the **magnitude** of the difference in mean values between males and females. For which trait is this difference

i. the **greatest** with the **male** mean being **greater than** the **female** mean

```
allTraits <- merge(maleTraits,femaleTraits, by = 'Trait') #Mean.x , SD.x, Mean.y, SD.y
allTraits_diff <- with(allTraits,
                       data.frame(Trait, Mean.x,SD.x,Mean.y,SD.y,
                                  MaleFemale = Mean.x - Mean.y,
                                  FemaleMale = Mean.y - Mean.x,
                                  absDiff = abs(Mean.x - Mean.y)
                       ))
index_max3 <- which.max(allTraits_diff[,6])
allTraits_diff[index_max3,]$Trait
```

```
## [1] "Intellect"
```

ii. the **greatest** with the **female** mean being **greater than** the **male** mean

```
index_max4 <- which.max(allTraits_diff[,7])
allTraits_diff[index_max4,]$Trait
```

```
## [1] "Withdrawal"
```

iii. the **least** in absolute magnitude between the two sexes

```
index_abs <- which.min(allTraits_diff[,8])
allTraits_diff[index_abs,]$Trait
```

```
## [1] "Conscientiousness"
```

INTERLUDE

For every trait, we have the mean and the standard deviation of the scores for females and for males. Since that is all we have, we might **model** the distribution of the scores for each trait on the study population as being normal, or Gaussian, with parameter values given by the mean and standard deviation for that `trait` on that sex. To that end, we could generate a sample of size `n` from any one of these sampling distributions!

To do this, suppose a single psychological trait of interest is given by the value of the variable `trait`. Then a sample of size `n` could be generated using `rnorm()` using the `Mean` and `SD` for that sex and that `trait`.

Suppose you did this, and then saved the `n` values as a numeric vector on `xmale` (the sample generated from the distribution of `trait` for males) and as a numeric vector on `xfemale` (the sample generated from the distribution of `trait` for females). You could put these together as a `data.frame` as follows:

For each of the following parts, it is assumed that you have done the above (using the appropriate `Means` and `SDs`) and that you have stored the values as the variables

```
# sample size
n
# psychological trait name
trait
# parameters for the males on that trait
maleParams
# and for females
femaleParams
# So that the trait, mean, and standard deviation are
maleParams$Trait
maleParams$Mean
maleParams$SD
# and similarly for femaleParams
#
# The generated samples are
xmale
xfemale
# and combined in a data.frame as
testdata
```

Each of the remaining parts assumes these are the variable names used (same names, but different samples, for each part).

d. **(11 marks)** In this question, you will be comparing the sexes based on the trait you identified in part (c)(i).

i. *(4 marks)* Write the code to produce a sample of size `n = 100` from the normal distribution for each sex for this trait.

```
# sample size
n <- 100
# psychological trait name
trait <- "Intellect"
# parameters for the males on that trait
maleParams <- maleTraits[maleTraits$Trait == trait,]
# and for females
femaleParams <- femaleTraits[femaleTraits$Trait == trait,]
# The generated samples are generated using rnorm()
xmale <- rnorm(n, maleParams$Mean, maleParams$SD)
xfemale <- rnorm(n, femaleParams$Mean, femaleParams$SD)
# and combined in a data.frame as
testdata <- data.frame(sex = rep(c("M", "F"), each = n),
                       x = c(xmale, xfemale))
```

NOTE: You will be reusing this code with small variations (difference in sample size `n` and psychological `trait`) but this is the **only** time you need to show this code. Do **not** show it again (after this question, use `echo = FALSE` in any R snippet in RMarkdownthat contains the sampling code; use `echo = TRUE` for all other snippets).

ii. *(3 marks)* In R, if we have two samples `x` and `y` from different normal distributions, we can test the hypothesis $H : \mu_x = \mu_y$ using the function `t.test()`. Use this function to test the equality of means for the two sexes (two-sided test, do not assume equal variances) for the samples of size `n = 100` from each sex that you just generated. In writing up your answer

- show your code of course
- report the p-value
- write a one sentence conclusion about the hypothesis based on your findings

```
t.test(xmale,xfemale)
```

```
##
##     Welch Two Sample t-test
##
## data:  xmale and xfemale
## t = 0.95236, df = 189.94, p-value = 0.3421
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.0940607  0.2696778
## sample estimates:
## mean of x mean of y
##  3.607966  3.520158
```

```
t.test(xmale,xfemale)$p.value
```

```
## [1] 0.3421242
```

In the t.test, the p-value is 0.3421242, so we reject the null hypothesis.

iii. *(2 marks)* Repeat part (ii) above, but this time use samples of size `n = 10,000`.

```r
t.test(xmale,xfemale)
```

```
##
##     Welch Two Sample t-test
##
## data:  xmale and xfemale
## t = 16.6, df = 19990, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1284711 0.1628717
## sample estimates:
## mean of x mean of y
##  3.626650  3.480979
```

```r
t.test(xmale,xfemale)$p.value
```

```
## [1] 1.793326e-61
```

In the t.test, the p-value is 1.793326e-61 which is a strong evidence of rejecting the null hypothesis.

iv. *(2 marks)* Did you observe any difference in conclusions between parts (ii) and (iii) above? Explain why that would be the case.

The difference between the conclusions between part (ii) and (iii) is the p-values. In part ii, we obtained a pvalue of 0.3421242 and in part iii, we obtained a pvalue of 1.793326e-61. The reason why the p-value in part iii is so small that is close to 0 is because the p-values are affected by the sample size. When we have larger sample size, we will have smaller p-values. Increasing the sample size will tend to result in a smaller P-value only when the null hypothesis is false.

e. **(9 marks) Tail values**. Consider the data frame `testdata` you created above when $n = 10,000$. Suppose we focus attention on those individuals with a score of `x` greater than some value, say `x > cutoff`. Of these individuals, we want to determine the proportion of each sex.

i. *(3 marks)* Suppose that the variable `cutoff` has some numeric value (yet to be determined). Write R code that will determine the proportion of males in this group, and hence the proportion of females. (Since `cutoff` does not yet have a value, make sure that you have `eval = FALSE` in your code snippet.)

```
prop <- function(cutoff, samp){
   samp[samp$x>cutoff,]
   }
```

ii. *(3 marks)* Select all observations where `x` is greater than the median value of the 20,000 data points `x`.

- Of these values, what proportion are male? What proportion are female?
- What do you conclude about people in the top half of this distribution?

```
cutoff <- median(testdata$x)
proportion <- prop(cutoff, testdata)
denom <- nrow(proportion)
maleProportion <- proportion[proportion$sex == "M",]
nrow(maleProportion)/denom
```

```
## [1] 0.544
```

```
femaleProportion <- proportion[proportion$sex == "F",]
nrow(femaleProportion)/denom
```

```
## [1] 0.456
```

The people in the top half of this distribution contains more males (5440) compare to female (4560)

iii. *(3 marks)* Repeat part (ii) above but this time select only those people with `x` value in the top 1% of the `x` values in `testdata`.

```
cutoff = quantile(testdata$x, prob=c(.99)) # top 1 percent should have 200 values
proportion <- prop(cutoff, testdata)
denom <- nrow(proportion)
maleProportion <- proportion[proportion$sex == "M",]
nrow(maleProportion)/denom
```

```
## [1] 0.68
```

```
femaleProportion <- proportion[proportion$sex == "F",]
nrow(femaleProportion)/denom
```

```
## [1] 0.32
```

The number of males in the top 1 percent of x-values in this distribution is significantly more than females.

f. **(3 marks) Tail values**. Repeat part (e)(iii) above **but** now on the `trait` determined in part c(ii). Show all your code.

```
maleProportion <- proportion[proportion$sex == "M",]
nrow(maleProportion)/denom
```

```
## [1] 0.245
```

```
femaleProportion <- proportion[proportion$sex == "F",]
nrow(femaleProportion)/denom
```

```
## [1] 0.755
```

The number of females in the top 1 percent of x-values in this distribution is significantly more than males

g. **(3 marks) Tail values**. Repeat part (e)(iii) above **but** now on the `trait` determined in part c(iii). Show all your code.

```
maleProportion <- proportion[proportion$sex == "M",]
nrow(maleProportion)/denom
```

```
## [1] 0.385
```

```
femaleProportion <- proportion[proportion$sex == "F",]
nrow(femaleProportion)/denom
```

```
## [1] 0.615
```

The number of females in the top 1 percent of x-values in this distribution is significantly more than males.