Seemingly simple things are often not simple at all: Equal sample means are not sufficient to infer equal groups in a population.

As published in Benchmarks; Research Matters, June 2017

https://benchmarks.it.unt.edu/researchmatters

Jon Starkweather, PhD

Jon Starkweather, PhD
`jonathan.starkweather@unt.edu`
Consultant
Research and **S**tatistical **S**upport

https://www.unt.edu

http://it.unt.edu/research

R&SS hosts a number of “Short Courses”.
A list of them is available at:
http://it.unt.edu/researchshortcourses

Those interested in learning more about R, or how to use it, can find information here:
`http://bayes.acs.unt.edu:8083/BayesContent/class/Jon/R_SC/`

Seemingly simple things are often not simple at all: Equal sample means are not sufficient to infer equal groups in a population.

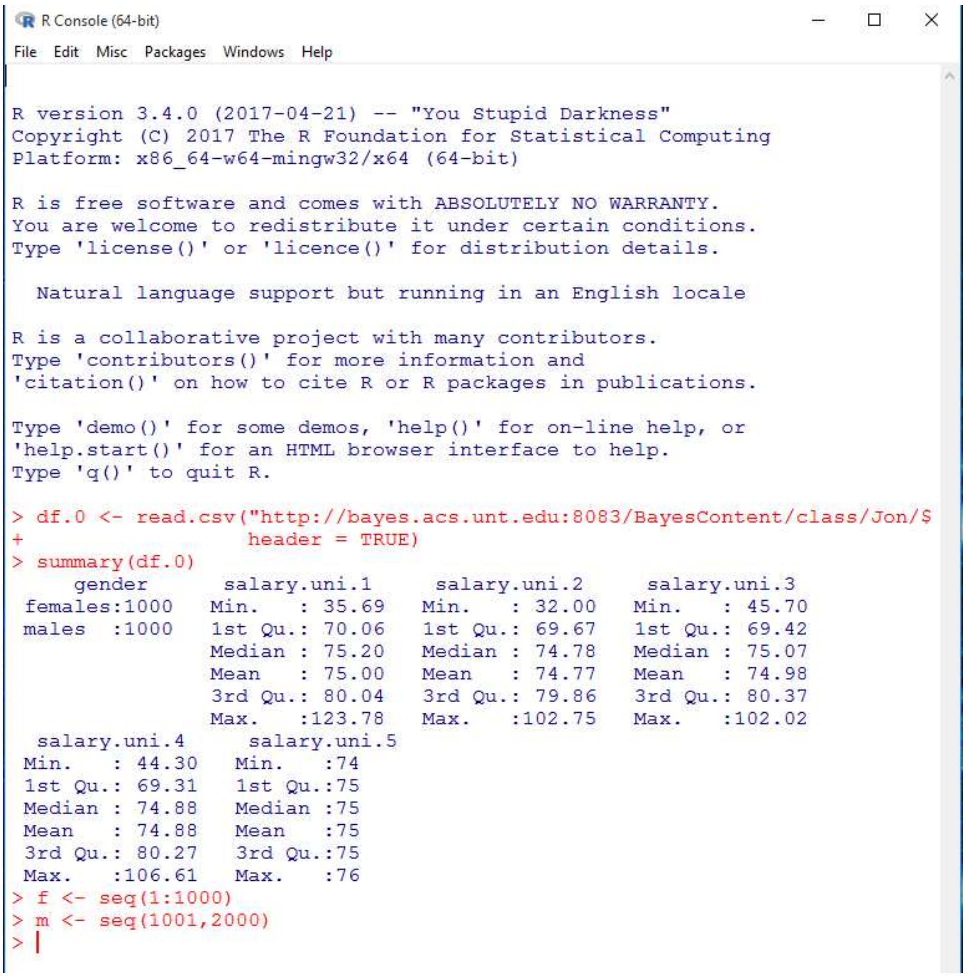
This month's article was motivated after overhearing a conversation in which one person claimed two groups were equal "because the means of each group were the same to the first decimal place and a *t*-test was non-significant." Even if the sample was randomly selected from a well-defined population and the cases were randomly assigned to the groups, there are still likely to be problems with such a simplistic modeling strategy. Bryk and Raudenbush (1988) have illustrated problems with homogeneity assumptions. Even if it were safe to assume no other variables were influencing the two being analyzed (i.e. strict experimental control; which is itself highly unlikely), it would still be very irresponsible to claim equality based solely on a *t*-test and equality of means. Below, we demonstrate a few examples why.

Let's pretend we want to explore salary equity among female and male university professors at five different universities. So, we collected a sample (*n* = 1000) of females and a sample of males (also; *n* = 1000) each from five different universities. Ignoring any other variables which might affect the relationship between gender and salary (which would be extremely unwise); what might we find at each university? Which university(-ies) would you consider male and female salaries equal enough to not raise suspicions of wage discrimination? Note we are not addressing the question of salaries at all universities, but instead at each of the five universities selected.

First, we import the simulated data into R, of course; get a summary of it and define the two groups. The URL for the data is here[1]. The salaries are listed in thousands of United States dollars (USD). A version of the R script used in this article can be found on the R&SS Do-It-Yourself Introduction to R website[2], particularly in the Module 5 section.

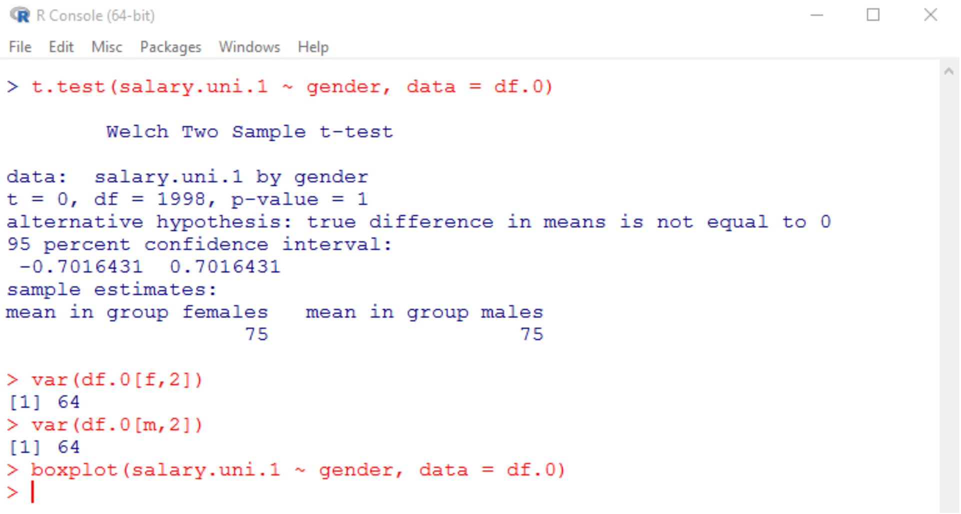
[1] http://bayes.acs.unt.edu:8083/BayesContent/class/Jon/ExampleData/SameMeanButNotEqual02.

[2] http://bayes.acs.unt.edu:8083/BayesContent/class/Jon/R_SC/



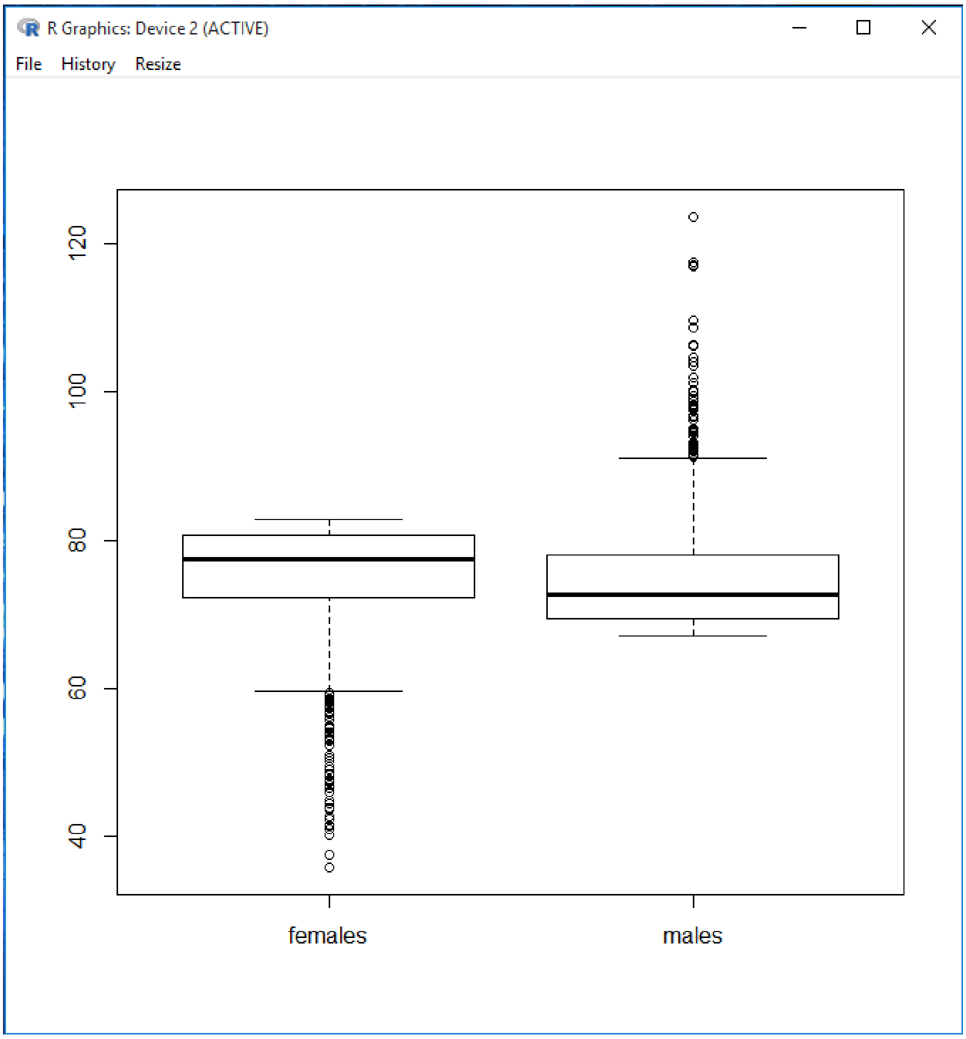
```
R version 3.4.0 (2017-04-21) -- "You Stupid Darkness"
Copyright (C) 2017 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> df.0 <- read.csv("http://bayes.acs.unt.edu:8083/BayesContent/class/Jon/$
+                    header = TRUE)
> summary(df.0)
     gender      salary.uni.1     salary.uni.2     salary.uni.3
 females:1000   Min.   : 35.69   Min.   : 32.00   Min.   : 45.70
 males  :1000   1st Qu.: 70.06   1st Qu.: 69.67   1st Qu.: 69.42
                Median : 75.20   Median : 74.78   Median : 75.07
                Mean   : 75.00   Mean   : 74.77   Mean   : 74.98
                3rd Qu.: 80.04   3rd Qu.: 79.86   3rd Qu.: 80.37
                Max.   :123.78   Max.   :102.75   Max.   :102.02
  salary.uni.4     salary.uni.5
 Min.   : 44.30   Min.   :74
 1st Qu.: 69.31   1st Qu.:75
 Median : 74.88   Median :75
 Mean   : 74.88   Mean   :75
 3rd Qu.: 80.27   3rd Qu.:75
 Max.   :106.61   Max.   :76
> f <- seq(1:1000)
> m <- seq(1001,2000)
> 
```

So, when we take a cursory look at university one, we find females and males have the same salary mean and same salary variance.

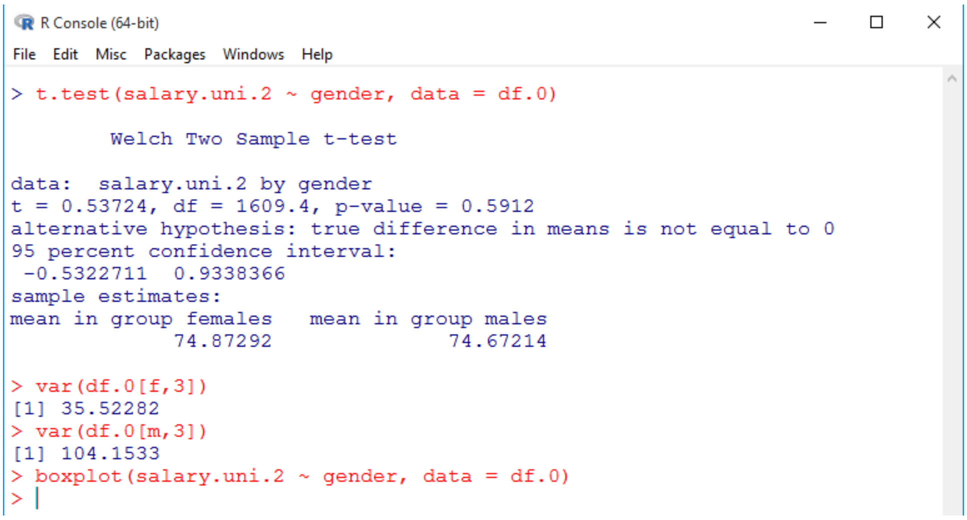


```
> t.test(salary.uni.1 ~ gender, data = df.0)

        Welch Two Sample t-test

data:  salary.uni.1 by gender
t = 0, df = 1998, p-value = 1
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.7016431  0.7016431
sample estimates:
mean in group females   mean in group males
                   75                    75

> var(df.0[f,2])
[1] 64
> var(df.0[m,2])
[1] 64
> boxplot(salary.uni.1 ~ gender, data = df.0)
> 
```

However, they do not appear to be *the same* when we visualize the data.

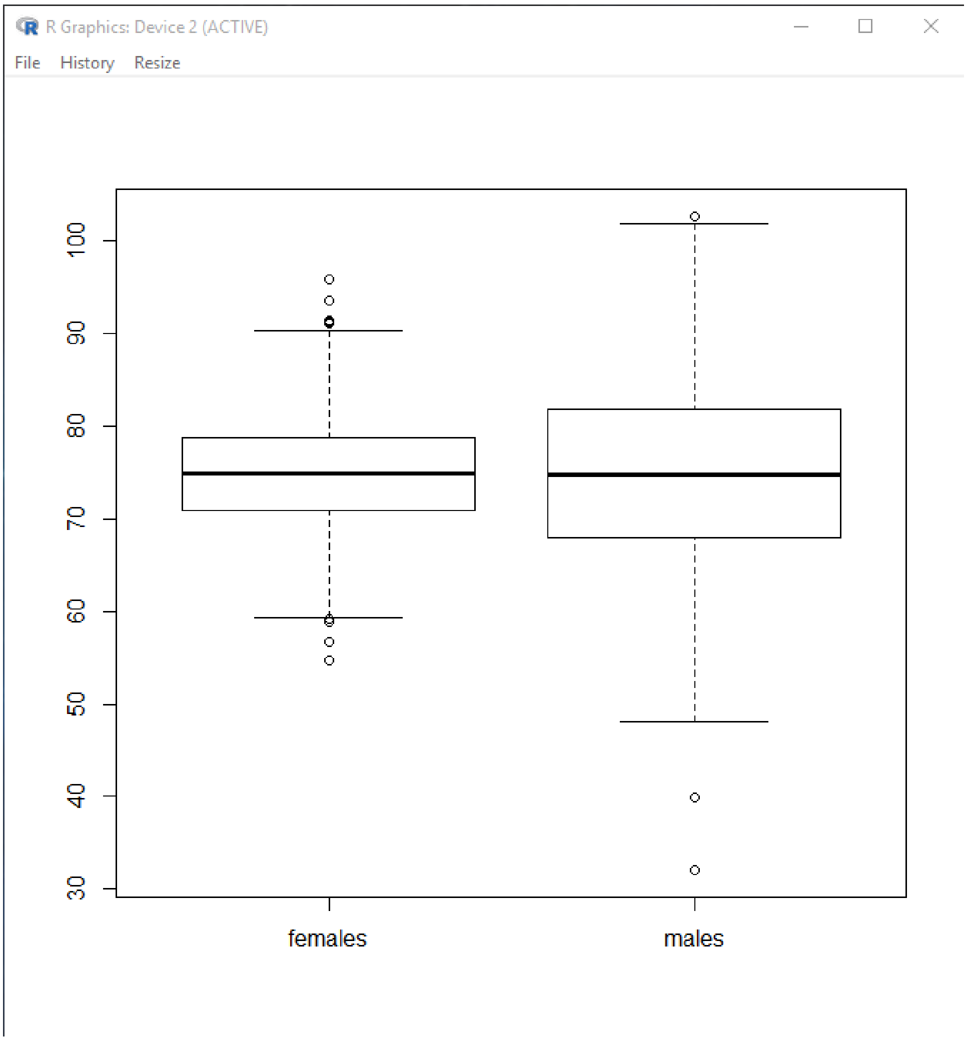


When we take a cursory look at university two we find the means are nearly the same, but the variances are drastically different.

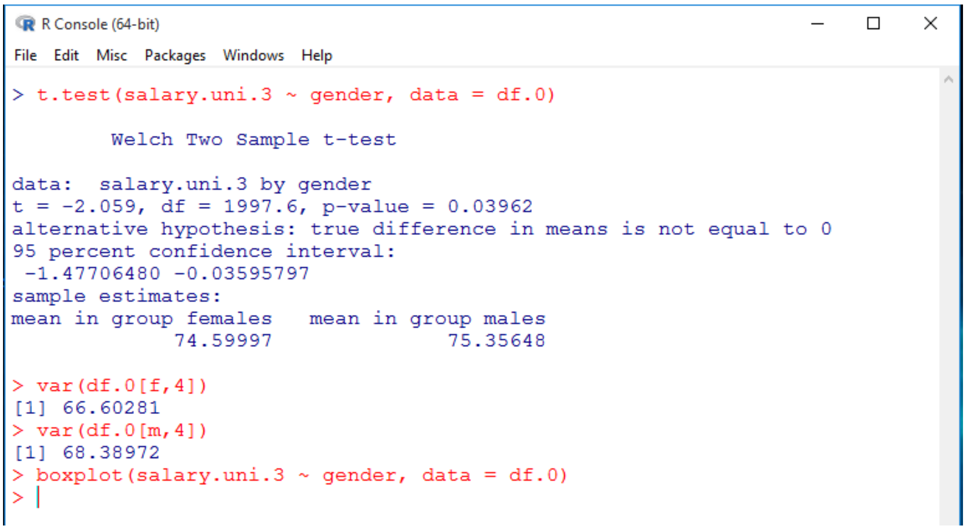


```
> t.test(salary.uni.2 ~ gender, data = df.0)

        Welch Two Sample t-test

data:  salary.uni.2 by gender
t = 0.53724, df = 1609.4, p-value = 0.5912
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.5322711  0.9338366
sample estimates:
mean in group females   mean in group males 
             74.87292              74.67214 

> var(df.0[f,3])
[1] 35.52282
> var(df.0[m,3])
[1] 104.1533
> boxplot(salary.uni.2 ~ gender, data = df.0)
> 
```

Again, visualization clarifies how the groups are indeed *different*.

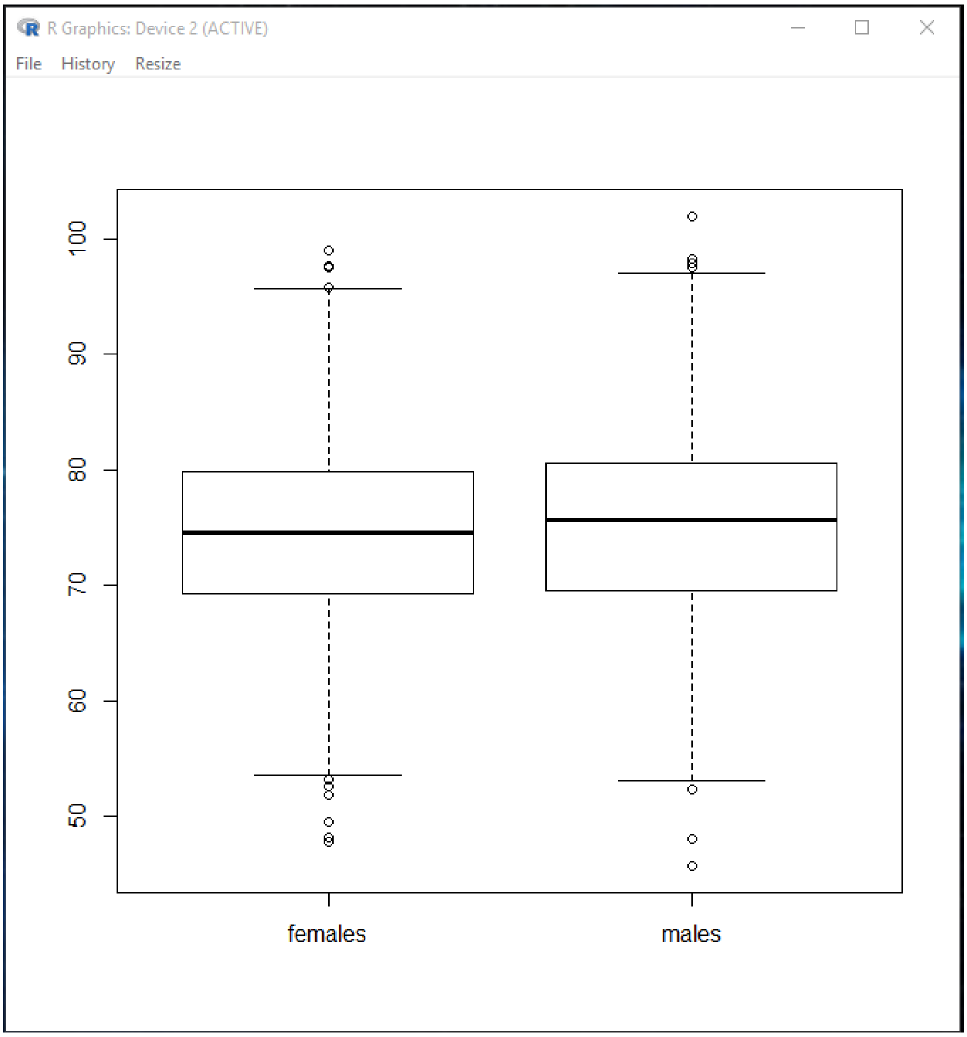


Data from university three displays virtually the same means, somewhat different variances.

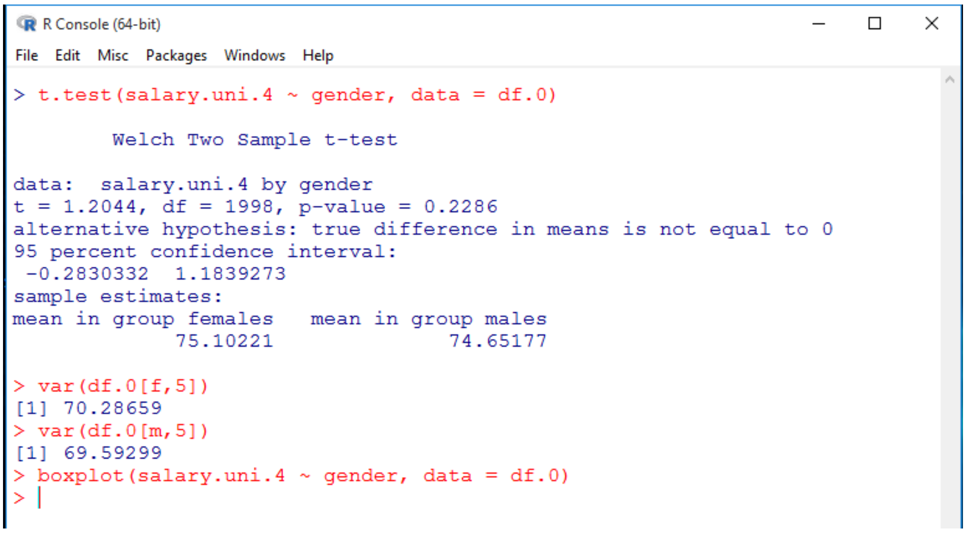


```
> t.test(salary.uni.3 ~ gender, data = df.0)

        Welch Two Sample t-test

data:  salary.uni.3 by gender
t = -2.059, df = 1997.6, p-value = 0.03962
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.47706480 -0.03595797
sample estimates:
mean in group females   mean in group males
             74.59997              75.35648

> var(df.0[f,4])
[1] 66.60281
> var(df.0[m,4])
[1] 68.38972
> boxplot(salary.uni.3 ~ gender, data = df.0)
> |
```

And again, the visualization clarifies.

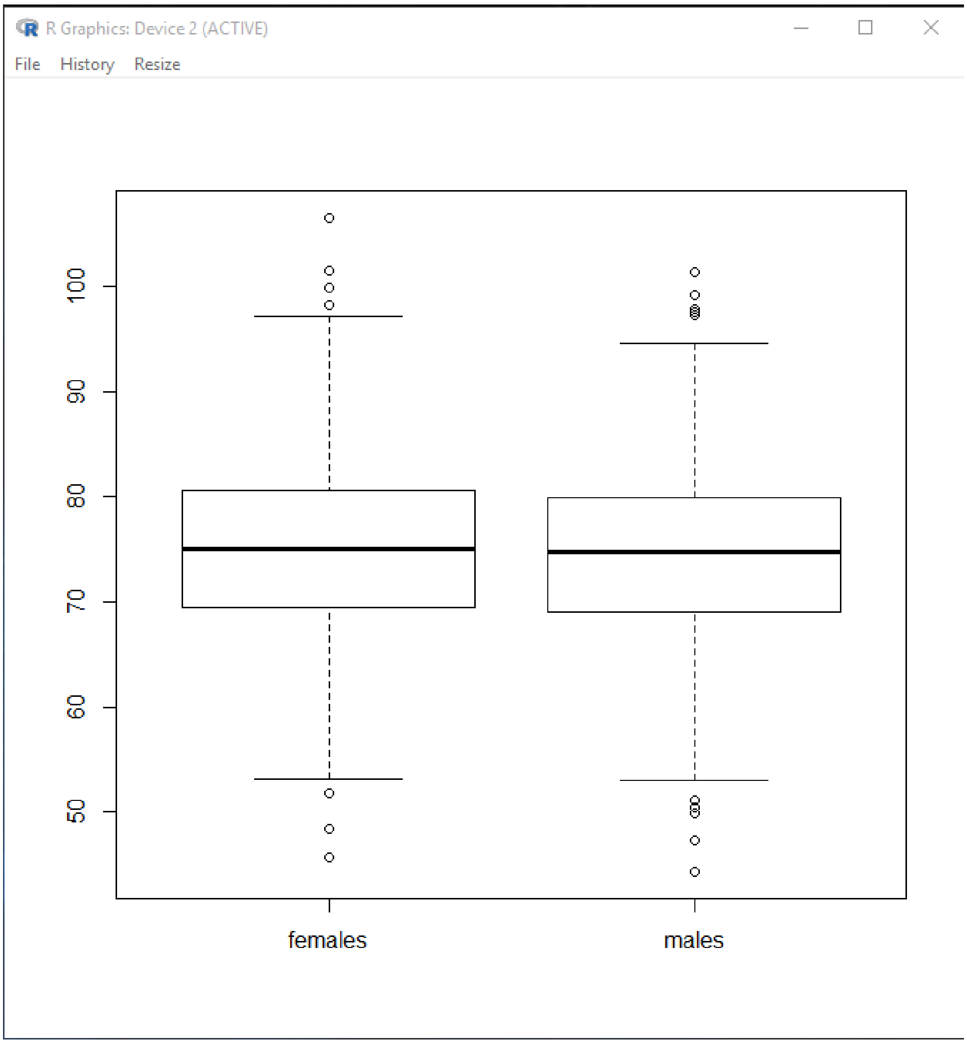


University four displays virtually the same means and virtually the same variances.

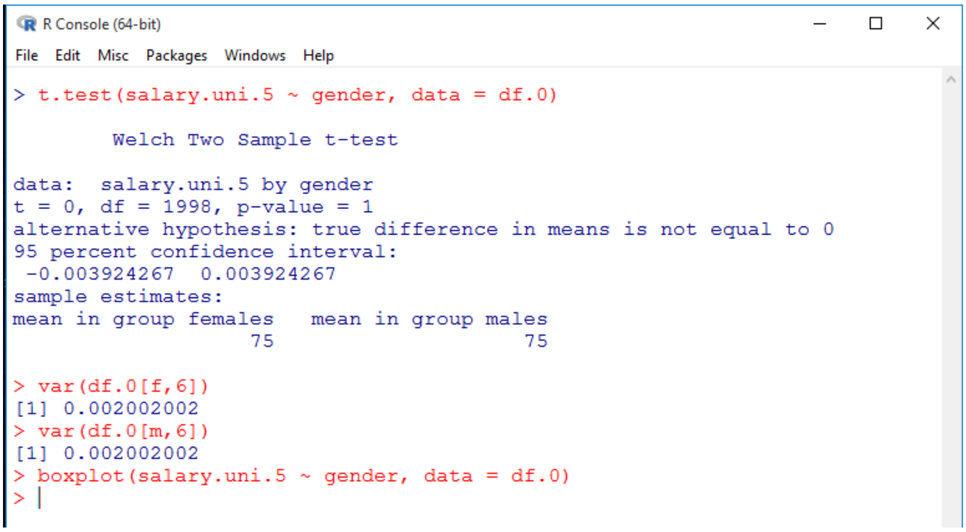


```
> t.test(salary.uni.4 ~ gender, data = df.0)

        Welch Two Sample t-test

data:  salary.uni.4 by gender
t = 1.2044, df = 1998, p-value = 0.2286
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.2830332  1.1839273
sample estimates:
mean in group females   mean in group males
             75.10221              74.65177

> var(df.0[f,5])
[1] 70.28659
> var(df.0[m,5])
[1] 69.59299
> boxplot(salary.uni.4 ~ gender, data = df.0)
> |
```

Are the groups *equal*?

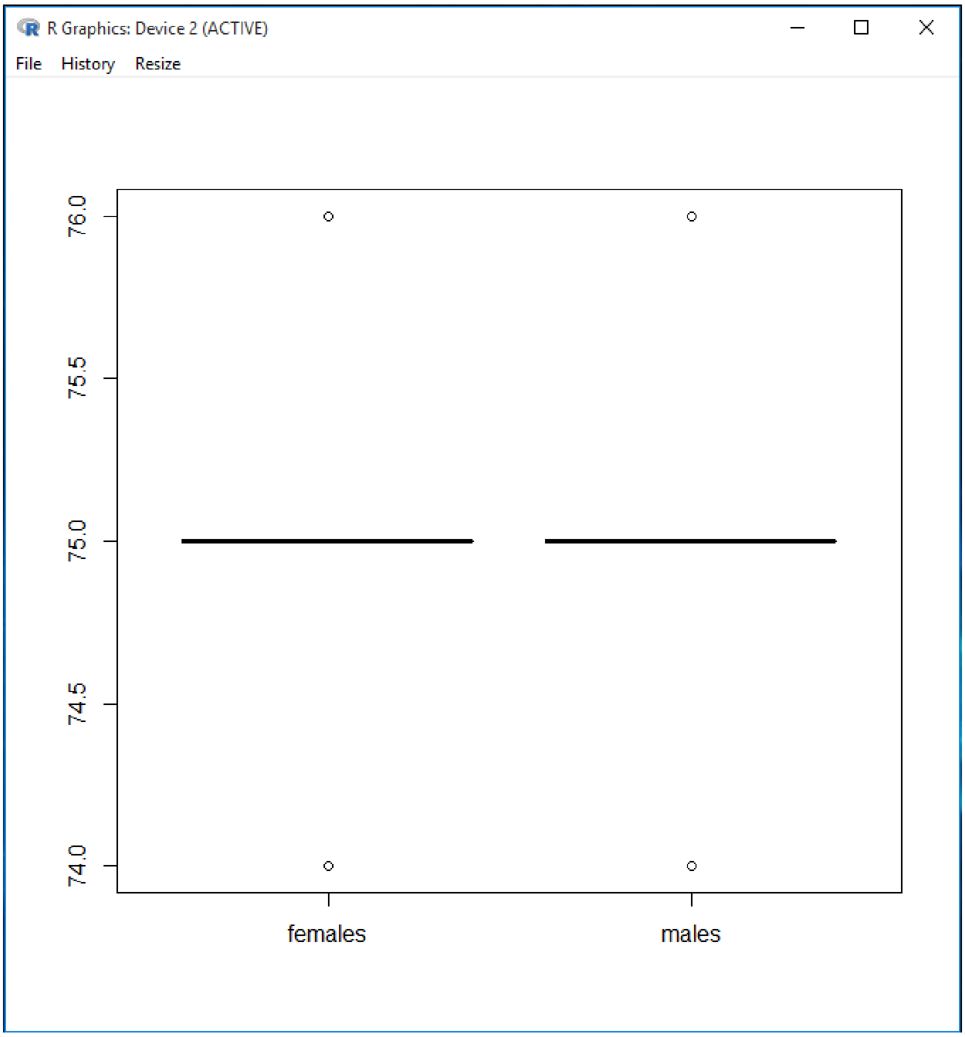


University five is the most extreme of those used in these examples, it is only included to encourage thinking (of assumptions and limitations of modeling strategies). University five's data shows exactly the same means and exactly the same variances; which are virtually zero.



```
> t.test(salary.uni.5 ~ gender, data = df.0)

        Welch Two Sample t-test

data:  salary.uni.5 by gender
t = 0, df = 1998, p-value = 1
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.003924267  0.003924267
sample estimates:
mean in group females   mean in group males 
                   75                    75 

> var(df.0[f,6])
[1] 0.002002002
> var(df.0[m,6])
[1] 0.002002002
> boxplot(salary.uni.5 ~ gender, data = df.0)
> |
```

University 5 pays everyone $75000; except 1 female earns $74000, 1 female earns $76000, 1 male earns $74000, and 1 male earns $76000.



To be clear, *t*-tests are generally used when seeking differences; not equality. For those interested in fitting simple modes to check for differences, equivalency, and indeterminacy, Tryon (2001) has provided an overview.

Do the above examples indicate simplistic models, such as the *t*-test model, should be relegated to the history books? No, but the utility of simplistic models needs to be recognized as severely limited. Those models are appropriate in only a *very* limited set of circumstances. The current focus in quantitative data analysis is on predominantly two aspects. Collecting large data arrays (i.e. rows and columns) which capture the complexity of most serious research endeavors. And using available hardware (i.e. supercomputers) with freely available software (e.g. R) to discover, model, and evaluate the complexity in those data to better inform decisions with meaningful consequences. It is incumbent upon all of us involved with data analysis to challenge ourselves to use modern technology (hardware & software) to analyze large data and fit models which better represent the complexity of reality. UNT has a goal to maintain Tier 1 status and that virtually mandates advanced, cutting edge research. Research and

Statistical Support[3] (R&SS) and High-Performance Computing[4] (HPC) services are available to help facilitate such research.

References & Resources

Bryk, A., S., & Raudenbush, S. W. (1988). Homogeneity of variance in experimental studies: A challenge to conventional interpretations. *Psychological Bulletin, 104*(3), 396 - 404.

Tryon, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods, 6*(4), 371 - 386.

This article was last updated on June 2, 2017.

This document was created using LaTeX

[3] https://it.unt.edu/research
[4] https://hpc.unt.edu/