Worksheet10

STAT414

2024-11-27

Module10 deals with a wellknown topic, but more in depth than you may have studied in your previous stat course (like Stat 350 or 351 or 355). The worksheet will require you to use the tTestPower function of EnvStats. Read the help file in the EnvStats carefully–will help you in completing the workseet.

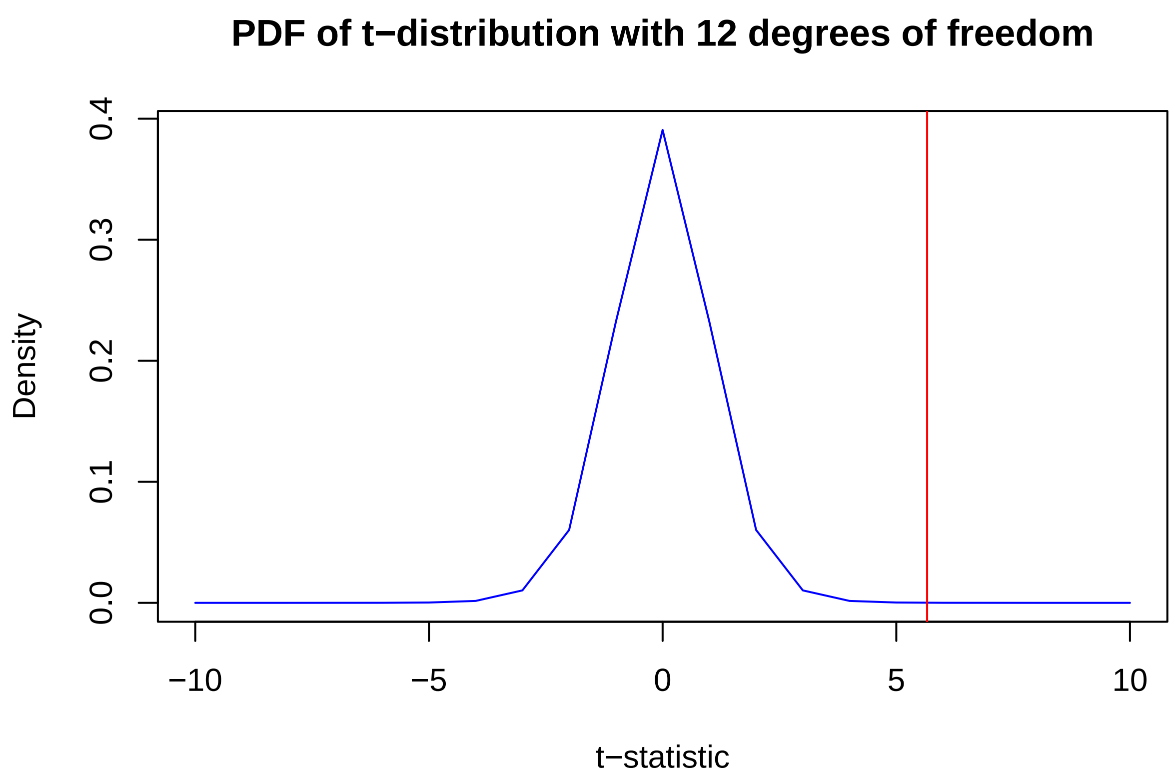
```
library(EnvStats)
```

```
##
## Attaching package: 'EnvStats'

## The following objects are masked from 'package:stats':
##
##     predict, predict.lm
```

```
# 1. we use the two-sample t-test to compare sulfate concentrations (EPA.09.Ex.16.1.sulfate.df)
# at a background and downgradient well. The resulting t-statistic is 5.66 with
# 12 degrees of freedom.

data <- EPA.09.Ex.16.1.sulfate.df
data
```

```
##    Month Year    Well.type Sulfate.ppm
## 1    Jan 1995   Background         560
## 2    Apr 1995   Background         530
## 3    Jul 1995   Background         570
## 4    Oct 1995   Background         490
## 5    Jan 1996   Background         510
## 6    Apr 1996   Background         550
## 7    Jul 1996   Background         550
## 8    Oct 1996   Background         530
## 9    Jan 1995 Downgradient          NA
## 10   Apr 1995 Downgradient          NA
## 11   Jul 1995 Downgradient         600
## 12   Oct 1995 Downgradient         590
## 13   Jan 1996 Downgradient         590
## 14   Apr 1996 Downgradient         630
## 15   Jul 1996 Downgradient         610
## 16   Oct 1996 Downgradient         630
```

```r
t_statistic <- 5.66
dof <- 12


# (a) Plot the pdf of a t-distribution with 12 degrees of freedom and add a
#vertical line at x = 5.66.

#Generate some set of values x to plot the pdf
x <- seq(-10, 10, by=1)
#dt(x, degrees_of_freedom) is the density of t-distribution
plot(x, dt(x,dof), type = "l", col="blue", xlab = "t-statistic", ylab = "Density",
     main = "PDF of t-distribution with 12 degrees of freedom")

#Superimpose line at x=5.66, v is the x-values for vertical lines
abline(v = t_statistic, col = "red")
```




```r
# (b) Explain what part of this plot represents the p-value for the test of the
# null hypothesis that the average sulfate concentrations at the two wells are the
# same against the alternative hypothesis that the average concentration of sulfate
# at the downgradient well is larger than the average concentration at the back-ground
# well.

cat(" mu1 = average concentration of sulfate at downgradient well, mu2 = background well
      H0: mu1 = mu2
      HA: mu1 > mu2")
```

```
##  mu1 = average concentration of sulfate at downgradient well, mu2 = background well
##       H0: mu1 = mu2
##       HA: mu1 > mu2
```

```
cat("The p-value in the plot is the area under the curve, since p-value can be
    represented by the area under the curve of the PDF. Therefore for the given
    t-statistic and x-line 5.66, we can analyze the p-value (area under the curve)
    above x=5.66 to determine a result for the hypothesis test. We use the upper-tail
    test P(T>5.66) which is to the right of x=5.66, because the t-statistic is
    calculated by  (mu1 - mu2 / SE) and our alternative hypothesis is HA: mu1 > mu2.")
```

```
## The p-value in the plot is the area under the curve, since p-value can be
##     represented by the area under the curve of the PDF. Therefore for the given
##     t-statistic and x-line 5.66, we can analyze the p-value (area under the curve)
##     above x=5.66 to determine a result for the hypothesis test. We use the upper-tail
##     test P(T>5.66) which is to the right of x=5.66, because the t-statistic is
##     calculated by  (mu1 - mu2 / SE) and our alternative hypothesis is HA: mu1 > mu2.
```

```
p1 <- pt(t_statistic, df=12, lower.tail=FALSE)
p1
```

```
## [1] 5.279756e-05
```

```
cat("The p-value of the upper-tail test is extremeley small and less than .05,
    therefore we reject H0 in support of HA.")
```

```
## The p-value of the upper-tail test is extremeley small and less than .05,
##     therefore we reject H0 in support of HA.
```

```
# 2. Consider the copper concentrations stored in the data frame EPA.09.Ex.16.4.copper.df
# in EnvStats package. Use the t-test and Wilcoxon rank sum test to compare the data
# from the two background wells.

data_downgradient <- data[data$Well.type == "Downgradient",]
data_background <- data[data$Well.type == "Background",]


cat("It turns out that, in estimating the variance of the difference of two
means, the assumption of equality of variances of the population is
crucial. This assumption is popularly known as homoscadasticity.
If we can assume that the variances of the two population are equal, in
other words the two populations are homoscedastic, then a more efficient
estimator of the variance of the two means can be obtained.
Therefore use var.equal = FALSE in t.test because we are testing the means
of the 2 independent and unpaired samples. ")
```

```
## It turns out that, in estimating the variance of the difference of two
## means, the assumption of equality of variances of the population is
## crucial. This assumption is popularly known as homoscadasticity.
## If we can assume that the variances of the two population are equal, in
## other words the two populations are homoscedastic, then a more efficient
```

```
## estimator of the variance of the two means can be obtained.
## Therefore use var.equal = FALSE in t.test because we are testing the means
## of the 2 independent and unpaired samples.
```

```r
t.test(data_downgradient$Sulfate.ppm, data_background$Sulfate.ppm,
       paired=FALSE, var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  data_downgradient$Sulfate.ppm and data_background$Sulfate.ppm
## t = 5.9826, df = 11.955, p-value = 6.485e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  45.82035 98.34631
## sample estimates:
## mean of x mean of y
##  608.3333  536.2500
```

```r
cat("The wilcoxon rank-sum test is the non-parametric equivalent of the independent
t-test. Used to test differences between two conditions in which different
participants have been used. Samples must have equal variance; homoscadasticity.")
```

```
## The wilcoxon rank-sum test is the non-parametric equivalent of the independent
## t-test. Used to test differences between two conditions in which different
## participants have been used. Samples must have equal variance; homoscadasticity.
```

```r
wilcox.test(data_downgradient$Sulfate.ppm, data_background$Sulfate.ppm,
       paired=FALSE, var.equal=TRUE)
```

```
## Warning in wilcox.test.default(data_downgradient$Sulfate.ppm,
## data_background$Sulfate.ppm, : cannot compute exact p-value with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  data_downgradient$Sulfate.ppm and data_background$Sulfate.ppm
## W = 48, p-value = 0.002309
## alternative hypothesis: true location shift is not equal to 0
```

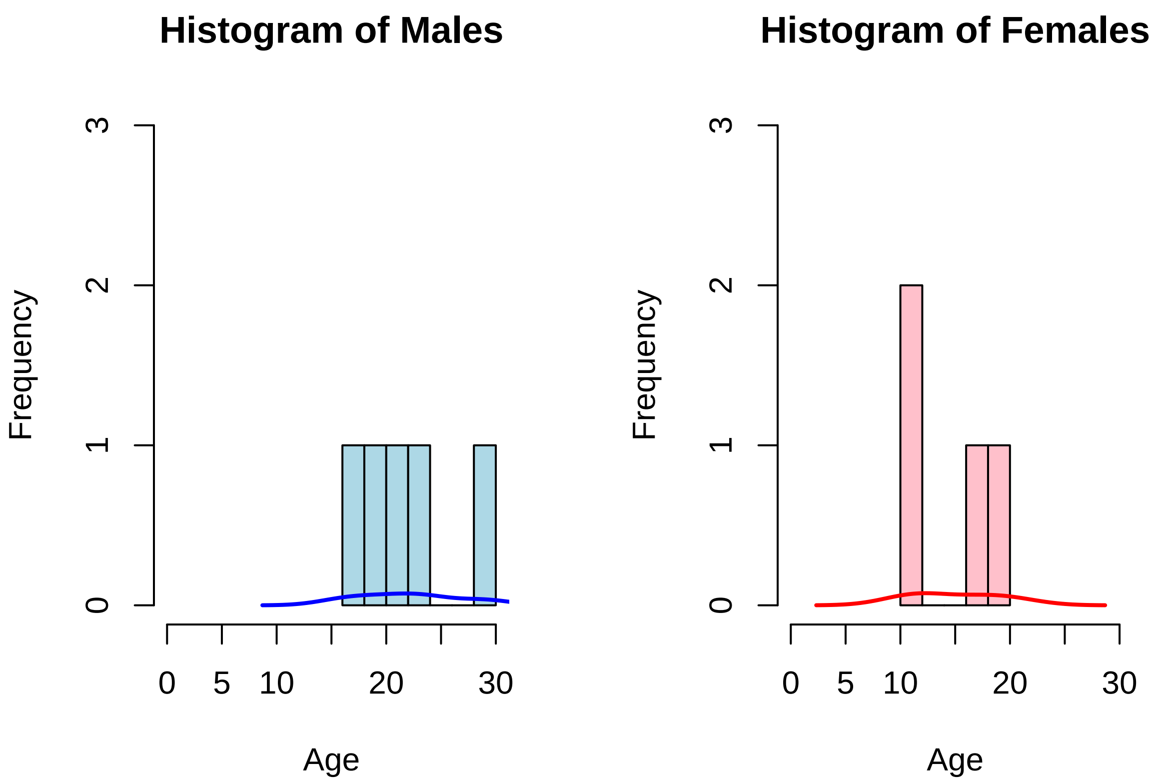
```r
cat("Both p-values for the wilcoxon rank sum test and and independent t-test are
    less than .05, meaning that there is significant statistical support for the
    alternative hypothesis, which is that the average sulfate concentrations for
    background wells and downgradient wells differ in value.")
```

```
## Both p-values for the wilcoxon rank sum test and and independent t-test are
##     less than .05, meaning that there is significant statistical support for the
##     alternative hypothesis, which is that the average sulfate concentrations for
##     background wells and downgradient wells differ in value.
```

```r
# 3. The following data shows age at diagnosis of Type II diabetes among young adults.
# Is the age at diagnosis different for males and females.
# Males 19, 22,16,29,24
# Females 20,11,17,12

males <- c(19,22,16,29,24)
females <- c(20,11,17,12)

# (a) What test procedures are applicable to this data structure? Write down the
# assumptions of for each of the candidate procedure.

cat("1. Independent t-test: Independent samples, follow normal distribution, equal variances
     2. Wilcoxon rank sum test: Independent samples, data is ordinal (can be ranked),
     distributions are similar in shape
     3. Welch's t-test: independent samples, follow normal distribution, does not
     need equal variances.
     Note: Wilxocon's signed-rank test is used for paired data, this is not paired
    ")
```

```
## 1. Independent t-test: Independent samples, follow normal distribution, equal variances
##      2. Wilcoxon rank sum test: Independent samples, data is ordinal (can be ranked),
##      distributions are similar in shape
##      3. Welch’s t-test: independent samples, follow normal distribution, does not
##      need equal variances.
##      Note: Wilxocon’s signed-rank test is used for paired data, this is not paired
##
```

```r
# (b) Which procedure would you recommend and why?

par(mfrow = c(1, 2), mar = c(5, 5, 4, 2))
# Histogram for males
hist(males, breaks = 5, col = "lightblue", main = "Histogram of Males",
     xlab = "Age", ylab = "Frequency", xlim=c(0,30), ylim=c(0,3))
lines(density(males), col = "blue", lwd = 2)

# Histogram for females
hist(females, breaks = 5, col = "pink", main = "Histogram of Females",
     xlab = "Age", ylab = "Frequency", xlim=c(0,30), ylim=c(0,3))
lines(density(females), col = "red", lwd = 2)
```




```r
var(males)
```

```
## [1] 24.5
```

```r
var(females)
```

```
## [1] 18
```

```r
cat("The distributions are too small and not curved to follow normality. The
    variances between males and females are not equal: 24.5 vs 18, therefore
    the best choice is the nonparametric approach of Wilcoxon rank sum test. The
    distribution are also somewhat similar in shape.")
```

```
## The distributions are too small and not curved to follow normality. The
##     variances between males and females are not equal: 24.5 vs 18, therefore
##     the best choice is the nonparametric approach of Wilcoxon rank sum test. The
##     distribution are also somewhat similar in shape.
```

```r
# (c) Regardless of your recommendation above, apply both the parametric and
# nonparametric methods to this example and compare the results.

# 1. Independent t-test, equal variances
t.test(males, females, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  males and females
## t = 2.2393, df = 7, p-value = 0.06014
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.3916462 14.3916462
## sample estimates:
## mean of x mean of y
##        22        15
```

```r
# 2. Wilcoxon rank sum test, nonparametric
wilcox.test(males, females)
```

```
##
##  Wilcoxon rank sum exact test
##
## data:  males and females
## W = 17, p-value = 0.1111
## alternative hypothesis: true location shift is not equal to 0
```

```r
# 3. Welch's t-test, unequal variances
t.test(males, females, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  males and females
## t = 2.2831, df = 6.9288, p-value = 0.05675
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2649222 14.2649222
## sample estimates:
## mean of x mean of y
##        22        15
```

```r
cat("Indepdent t-test:        p-value = 0.06014
     Wilcoxon rank sum test:  p-value = 0.1111
     Welch's t-test:          p-value = 0.05675

    All 3 tests fail to reject the null hypothesis at alpha = .05. Thus we conclude
    that there is not enough evidence to support the hypothesis that there is a
    statistically significant difference in the age of diagnosis for diabetes
    between males and females.")
```

```
## Indepdent t-test:        p-value = 0.06014
##      Wilcoxon rank sum test:  p-value = 0.1111
##      Welch's t-test:          p-value = 0.05675
##
##     All 3 tests fail to reject the null hypothesis at alpha = .05. Thus we conclude
##     that there is not enough evidence to support the hypothesis that there is a
##     statistically significant difference in the age of diagnosis for diabetes
##     between males and females.
```