

STAT2201 Assignment 5

Question 1. P-values with t

For the hypothesis test $H_0 : \mu = 7$ against $H_1 : \mu \neq 7$ with variance unknown and $n = 20$ approximate the P -value for each of the following test statistics:

- (a) $t_0 = 2.05$
- (b) $t_0 = -1.84$
- (c) $t_0 = 0.4$

First the `Distributions` package needs to be loaded and a distribution with 19 degrees of freedom needs to be created. As the question asks for a two sided test, the probability greater than the absolute value is calculated and then multiplied by two.

```
In [12]: 2.*ccdf(TDist(19),abs.([2.05,-1.84,0.4]))
```

```
Out [12]: 3-element Array{Float64,1}:  
 0.0544223  
 0.0814447  
 0.693618
```

Question 2. Syrup Prediction

A postmix beverage machine is adjusted to release a certain amount of syrup into a chamber where it is mixed with carbonated water. A random sample of 25 beverages was found to have a mean syrup content of $\bar{x} = 1.10$ fluid ounce and a standard deviation of $s = 0.015$ fluid ounce. Compute a 95% prediction interval on the syrup volume in the next beverage dispensed. Compare the length of the prediction interval with the length of a 95% CI on the population mean.

First calculated the confidence interval in Julia using the following code

```
In [13]: confint(OneSampleTTest(1.10,0.015,25))
```

```
Out [13]: (1.093808304315116,1.106191695684884)
```

The length of the confidence interval is 0.01238339136976796. To calculate the prediction interval for this sample the following formula is used:

$$\bar{x} \pm t_{df}^* s \sqrt{1 + \frac{1}{n}}.$$

Substituting the numbers from the question and coding in Julia the prediction interval can be calculated.

```
In [14]: 1.10.+quantile(TDist(24),[0.025,0.975]).*0.015.*sqrt(1+1/25)
```

```
Out [14]: 2-element Array{Float64,1}:  
 1.06843  
 1.13157
```

The length of the prediction interval is 0.06314315423890449. As can be seen this interval is larger than the confidence interval. Remember the confidence interval predicts where the population mean will lie, while the prediction interval predicts where the next observation will lie.

Question 3. Steel Rods

The diameter of steel rods manufactured on two different extrusion machines is being investigated. Two random samples of sizes $n_1 = 15$ and $n_2 = 17$ are selected, and the sample means and sample variances are $\bar{x}_1 = 8.73$, $s_1^2 = 0.35$, $\bar{x}_2 = 8.68$ and $s_2^2 = 0.40$, respectively. Assume that $\sigma_1^2 = \sigma_2^2$ and that the data are drawn from a normal distribution.

- (a) Is there evidence to support the claim that the two machines produce rods with different mean diameters? Use $\alpha = 0.05$ in arriving at this conclusion. Find the P -value.

First define the variables given in the question as follows

```
In [15]: n=[15,17]  
         means=[8.73,8.68]  
         vars=[0.35,0.40]
```

```
Out [15]: 2-element Array{Float64,1}:  
 0.35  
 0.4
```

Now calculate the pooled standard deviation using the following formula

$$s_{pooled} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Once this is calculated, the t -statistic can be calculated with the following formula

$$t = \frac{\bar{x}_1 - \bar{x}_2 - 0}{s_{pooled} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

```
In [16]: stdPoolQ3= sqrt(((n[1]-1)*vars[1]+(n[2]-1)*vars[2])/(n[1]+n[2]-2))  
         tStatisticQ3= (means[1]-means[2])/(stdPoolQ3*sqrt(1/n[1]+1/n[2]))
```

```
Out [16]: 0.22997811554215344
```

This t -value is then used to find the two-sided p -value using a t -distribution with 30 degrees of freedom.

```
In [18]: 2*ccdf(TDist(n[1]+n[2]-2),abs(tStatisticQ3))
```

```
Out [18]: 0.8196697157339579
```

As the p -value is greater than the α level of 0.05 the conclusion from this hypothesis test is that the machines do not produce rods with significantly different mean diameters.

- (b) Construct a 95% confidence interval for the difference in mean rod diameter. Interpret this interval.

To calculate the confidence interval the following formula is used

$$CI = \bar{x}_1 - \bar{x}_2 \pm t_{df}^* s_{pooled} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

In Julia this calculation is

```
In [19]: means[1]-means[2].+quantile(TDist(n[1]+n[2]-2), [0.0275,0.975])
        .*stdPoolQ3.*sqrt(1/n[1]+1/n[2])
```

```
Out[19]: 2-element Array{Float64,1}:
        -0.384109
         0.494015
```

So the confidence interval is $(-0.384109, 0.494015)$. As this interval contains zero it can be concluded that there is no significant difference in the mean diameter of rods produced by the two machines.

Question 4. Wet Chemical Etching

In semiconductor manufacturing, wet chemical etching is often used to remove silicon from the backs of wafers prior to metallization. The etch rate is an important characteristic in this process and known to follow a normal distribution. Two different etching solutions have been compared using two random samples of 10 wafers for each solution. The observed etch rates are as follows (in mils per minute):

Solution 1		Solution 2	
9.9	10.63	10.2	10.0
9.4	10.3	10.6	10.2
9.3	10.0	10.7	10.7
9.6	10.3	10.4	10.4
10.2	10.1	10.5	10.3

- (a) Construct normal probability plots for the two samples. Do these plots provide support for the assumptions of normality and equal variances? Write a practical interpretation for these plots.

Using the code presented in Assignment 3 for Normal Probability Plots

```
In [43]: using PyPlot
        function NormalProbabilityPlot(data)
            mu = mean(data)
            sig = std(data)
            n = length(data)
            p = [(i - 0.5)/n for i in 1:n]
            x = quantile(Normal(),p)
            y = sort([(i-mu)/sig for i in data])
            #y = quantile([(i-mu)/sig for i in data],p)
```

```

PyPlot.scatter(x,y)
xRange = maximum(x) - minimum(x)
PyPlot.plot([ minimum(x)- xRange/8,maximum(x) + xRange /8],
[minimum(x)- xRange/8,maximum(x)+ xRange /8],
color="red",linewidth =0.5)
xlabel("Theoretical quantiles")
ylabel("Quantiles of data");
return
end

```

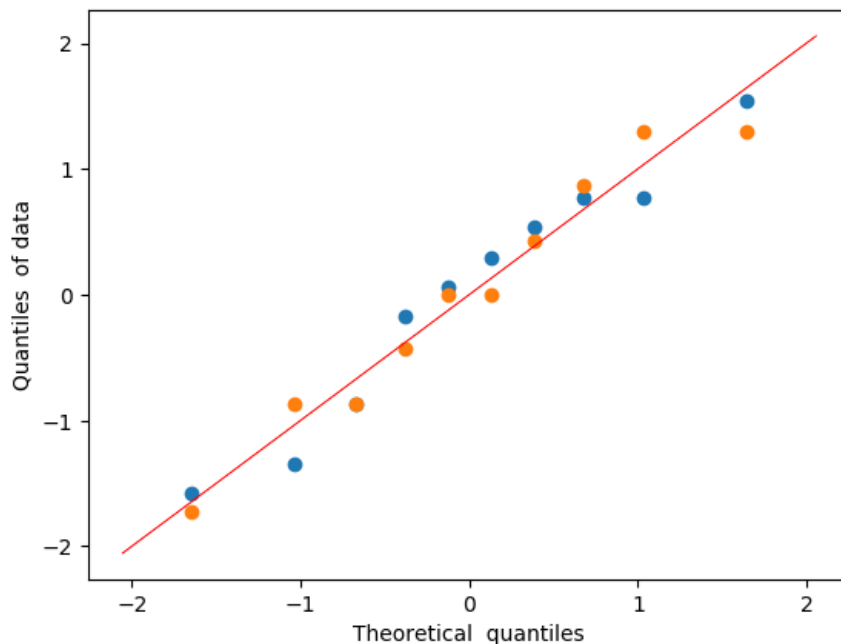
Out[43]: NormalProbabilityPlot (generic function with 1 method)

Now using this function to plot the Normal Probability Plots of the two solutions, the following plot is obtained.

```

In [44]: NormalProbabilityPlot(Solution_1)
NormalProbabilityPlot(Solution_2)

```



Looking at this plot it can be seen that both data sets follow the line closely and have about the same spread around the line.

- (b) Does the data support the claim that the mean etch rate is the same for both solutions? In reaching your conclusions, use $\alpha = 0.05$ and assume that both population variances are equal. Calculate a P-value.

Defining the two variables and then performing an equal variance *t*-test, the following output is obtained.

```

In [20]: Solution_1=[9.9,9.4,9.3,9.6,10.2,10.63,10.3,10.0,10.3,10.1]
Solution_2=[10.2,10.6,10.7,10.4,10.5,10.0,10.2,10.7,10.4,10.3]

testQ4=EqualVarianceTTest(Solution_1,Solution_2)

```

```

Out[20]: Two sample t-test (equal variance)
-----
Population details:
parameter of interest: Mean difference
value under h_0:      0
point estimate:      -0.42699999999999996
95% confidence interval: (-0.7494160569134942,-0.10458394308650493)

Test summary:
outcome with 95% confidence: reject h_0
two-sided p-value:      0.012291046308808899

Details:
number of observations: [10,10]
t-statistic:           -2.782410155904908
degrees of freedom:    18
empirical standard error: 0.15346407469574833

```

As the p -value is 0.0123 which is below the $\alpha = 0.05$, there is moderate evidence to reject the null hypothesis. Therefore the mean etch rate is significantly different between the solutions.

- (c) Find a 95% confidence interval on the difference in mean etch rates

Using the confidence interval from the test performed above

```
In [21]: confint(testQ4)
```

```
Out[21]: (-0.7494160569134942,-0.10458394308650493)
```

This means that Solution 1's mean etch rate is between 0.7494 and 0.1046 less than Solution 2's mean etch rate. Therefore it can be said that the mean etch rate of Solution 1 is less than the mean etch rate for Solution 2.

Question 5. Gold Ball Distance

The overall distance travelled by a golf ball is tested by hitting the ball with Iron Byron, a mechanical golfer with a swing that is said to emulate the distance hit by the legendary champion Byron Nelson. Ten randomly selected balls of two different brands are tested and the overall distance measured. The data is as follows:

Brand 1: 275, 286, 287, 271, 283, 271, 279, 275, 263, 267

Brand 2: 258, 244, 260, 265, 273, 281, 271, 270, 263, 268

- (a) Is there evidence that overall distance is approximately normally distributed? Is there an assumption of equal variances justified?

First define the variables and calculate the variances.

```
In [22]: Brand_1=[275,286,287,271,283,271,279,275,263,267]
```

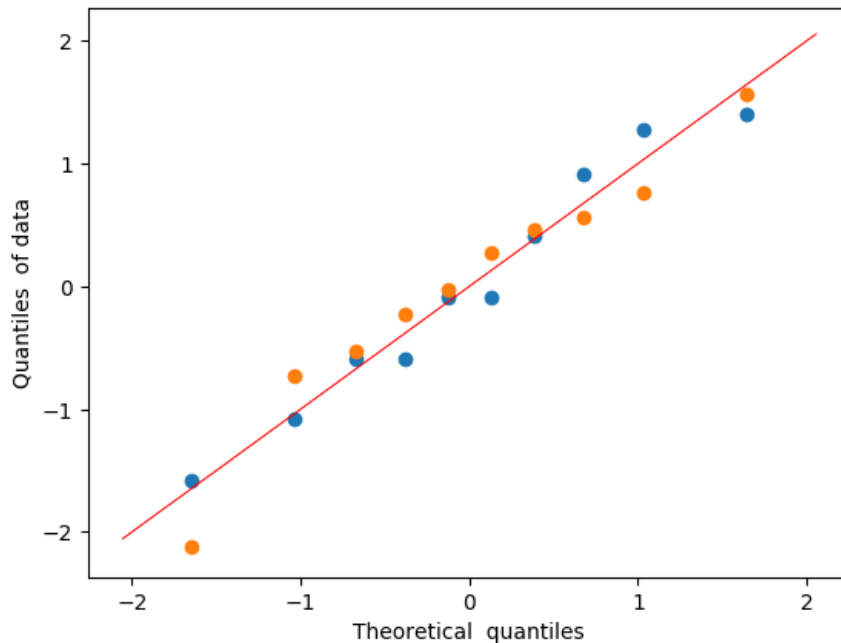
```
Brand_2=[258,244,260,265,273,281,271,270,263,268]
```

```
var(Brand_1),var(Brand_2)
```

```
Out [22]: (64.45555555555558, 100.90000000000003)
```

While these variances are different from each other, the variance of Brand 2 is less than two times the variance of Brand 1. As a factor of four is expected for significantly different variances it can be concluded that the assumptions of equal variances is appropriate here. Alternatively as with Question 4 the normal probability plots can be compared.

```
In [31]: NormalProbabilityPlot(Brand_1)
         NormalProbabilityPlot(Brand_2)
```



Looking at these plots it can be seen that both Brands' data follows a normal distribution and the variance is similar in both groups.

- (b) Test the hypothesis that both brands of ball have equal mean overall distance. Use $\alpha = 0.05$. What is the P-value?

Running an equal variance two sample *t*-test as before, the following output is obtained.

```
In [39]: testQ5=EqualVarianceTTest(Brand_1,Brand_2)
```

```
Out [39]: Two sample t-test (equal variance)
```

```
-----
Population details:
parameter of interest: Mean difference
value under h_0:      0
point estimate:      10.399999999999977
95% confidence interval: (1.8568244113862864, 18.94317558861367)
```

```
Test summary:
outcome with 95% confidence: reject h_0
two-sided p-value:          0.019784887263473237
```

Details:

```
number of observations: [10,10]
t-statistic:           2.5575488870470826
degrees of freedom:   18
empirical standard error: 4.066393433443887
```

Here the p -value is 0.0198 ($T_{19} = 2.558$) and so there is moderate evidence to reject the null hypothesis. Therefore there is moderate evidence of a significant difference in the mean overall distance between the brands.

- (c) Construct a 95% two-sided CI on the mean difference in overall distance for the two brands of golf balls.

As before, the confidence interval can be obtained from the test above.

```
In [40]: confint(testQ5)
```

```
Out[40]: (1.8568244113862864, 18.94317558861367)
```

Looking at this confidence interval it is noted that it is entirely positive. This means that the mean overall distance of Brand 1 will be between 1.857 and 18.943 greater than the mean overall distance of Brand 2.

- (d) What is the power of the statistical test in part(b) to detect a true difference in mean overall distance of 5 yards?

While a similar method to that taken in the past assignment can be taken, here a function to calculate power is presented.

```
In [25]: function powerTTest(n=NULL,delta=NULL,sd=1,sig_level=0.05,power=NULL
,tttype="two.sample",alternative="two.sided")
#Have to be using Distributions package. Can be used in two ways
#1. Enter n (size of sample), delta, sd, sig_level, 0, type and
alternative to find power of test (post-hoc)
#2. Enter 0, delta, sd, sig_level, power, type and alternative to find
sample size for power (pre-hoc)
tsample=if(tttype=="one.sample" || tttype=="paired") 1
elseif(tttype=="two.sample") 2 end
tside=if(alternative=="one.sided") 1 elseif(
alternative=="two.sided") 2 end
if(tside == 2 && !isnull(delta)) delta=abs(delta) end
if(n!=0)
  nu= (n-1)*tsample
  powerDist=TDist(nu)
  ncp = sqrt(n/tsample)*delta/sd
  qu = cquantile(powerDist,sig_level/tside)
  power = ccdf(powerDist,qu+ncp) + cdf(powerDist,ncp-qu)
  return round(power,5)
elseif(power!=0)
  normDist=Normal()
  nu= tsample*(sd^2*(quantile(normDist,power)+
cquantile(normDist,sig_level/tside))^2)/delta^2
```

```

n = 0
nv=0
while(powerTTest(nu,delta,sd,sig_level,0,ttype,alternative
) < power)
    nu += 10.0^-5
    n = n+1
end
return (round(nu,5),n)
end
end

```

Out[25]: powerTTest (generic function with 8 methods)

Using this function, the power can be computed using the following command.

```
In [38]: powerTTest(10,5,sqrt(mean([var(Brand_1),var(Brand_2)])),0.05,
0,"two.sample","two.sided")
```

Out[38]: 0.19938

Therefore the power of the test in part (b) to detect a true difference in mean overall distance of 5 yards is 19.93%.

- (e) What sample size would be required to detect a true difference in mean overall distance of 3 yards with power of approximately 0.75?

Again using the function defined above, the command to obtain the answer is below

```
In [34]: powerTTest(0,3,sqrt(mean([var(Brand_1),var(Brand_2)])),0.05,
0.75,"two.sample","two.sided")
```

Out[34]: (128.5113,99743)

The output above gives an estimate of the size of one group. This needs to be taken to the nearest integer greater than or equal to the value, then multiplied by two to get the total sample size. The required sample size to detect a true difference in mean overall distance of 3 yards with a power of approximately 0.75 is $2 \times 129 = 258$.

Question 6. Computer Output

Consider the following computer output:

```

Two-Sample T-Test and CI
Sample N      Mean      StDev      Se Mean
1      12         16       1.26       0.36
2      16      12.15     1.99       0.50
Difference = mu(1) - mu(2)
Estimate for difference: -1.210
95% CI for difference: (-2.560, 0.140)
T-test of difference = 0 (vs not =):
T-value = ?
P-value = ?
DF = ?
Both used Pooled StDev = ?

```


- (a) Fill in the missing values. Is this a one-sided or a two-sided test? Use lower and upper bounds for the P-value.

```
Two-Sample T-Test and CI
Sample N      Mean      StDev      Se Mean
1         12         16         1.26         0.36
2         16        12.15        1.99         0.50
Difference = mu(1) - mu(2)
Estimate for difference:  -1.210 [3.85]
95% CI for difference:  (-2.560, 0.140) [(2.500, 5.200)]
T-test of difference = 0 (vs not =):
T-value = -1.842804 [5.863466]
P-value = 0.05<p<0.1 [p<0.0002]
DF = 12+16-2=26
Both used Pooled StDev = 1.719404
```

(Values in [] from table rather than estimate for difference.)

- (b) What are your conclusions if $\alpha = 0.05$ What if $\alpha = 0.01$?

As the p -value is between 0.05 and 0.1 there is weak evidence to reject the null hypothesis. Therefore it can be concluded that there is weak evidence of a significant difference between the samples.

[As the p value is less than 0.0002 there is very strong evidence to reject the null hypothesis. Therefore there is strong evidence to support a significant difference between the samples.]

- (c) This test was done assuming that the two population variances were equal. Does this seem reasonable?

Given that the ratio ($\frac{s_1}{s_2}$) of standard deviations is less than 2 it seems reasonable that the population variances are equal.

- (d) Suppose that the hypothesis had been $H_0 : \mu_1 = \mu_2$ vs. $H_1 : \mu_1 < \mu_2$. What would your conclusions be if $\alpha = 0.05$?

As the t -value is negative the p -value for the test $\mu_1 < \mu_2$ is half of the p -value stated for the two sided test. Therefore the p -value is between 0.025 and 0.05, given moderate evidence to reject the null hypothesis. It can be concluded that there is moderate evidence that the mean of sample 1 is less than the mean of sample 2

[As the t -value is positive the p -value for the test $\mu_1 < \mu_2$ will be $p > 0.25$. Therefore there is inconclusive evidence to reject the null hypothesis. It can be concluded that the mean of sample 1 is not less than the mean of sample 2.]

Question 7. Melting Point

The melting points of two alloys used in formulating a solder were investigated by melting 21 samples of each material. The sample mean and standard deviation for alloy 1 was $\bar{x}_1 = 420^\circ$ F and $s_1 = 4^\circ$ F. For alloy 2, they were $\bar{x}_2 = 426^\circ$ F and $s_2 = 3^\circ$ F.

- (a) Does the sample data support the claim that both alloys have the same melting point? Use $\alpha = 0.05$ and assume that both populations are normally distributed and have the same standard deviation. Find the P-value for the test.

First define the variables and calculate the t -statistic using Julia as follows.

```
In [28]: n=[21,21]
          means=[420,426]
          vars=[4^2,3^2]

          sPooled7 = sqrt(mean(vars))
          tStatistic7= (means[1]-means[2])/(sPooled7*sqrt(2/21))
```

Out [28]: -5.499090833947007

Now as a test of difference is being conducted, the two sided p -value is calculated. This is done by taking the absolute value of the t -statistic and finding the probability greater than it, followed by multiplying this probability by two.

```
In [29]: 2*ccdf(TDist(40),abs(tStatistic7))
```

Out [29]: 2.3874112324086204e-6

The p -value here is very small and so there is strong evidence to reject the null hypothesis. Therefore there is strong evidence to support that the alloys have different melting points.

- (b) Suppose that the true mean difference in melting points is 3° F. How large a sample would be required to detect this difference using an $\alpha = 0.05$ level test with probability at least 0.9? Use $\sigma_1 = \sigma_2 = 4$ as an initial estimate of the common standard deviation.

Using the `powerTTest` defined in Question 5, the sample size can be obtained using the following command.

```
In [34]: powerTTest(0,3,4,0.05,0.9,"two.sample","two.sided")
```

Out [34]: (38.37258,101285)

From this the size of one group would need to be 39 samples. Therefore the sample size would need to be 78 samples between the two alloys.

Question 8. Experiment - validity of Approximation when Variance Unequal

The two sample t-test is based on the t-distribution and is exact when $\sigma_1 = \sigma_2$. However, when $\sigma_1 \neq \sigma_2$ the use of a t-distribution with v degrees of freedom is an approximation. Here,

$$v = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2}}$$

In this question, you are to carry out simulation experiments that explore the validity of this approximation.

Take $n_1 = 10$, $n_2 = 10$, and assume (under H_0) that $X_i \sim N(0, 0.7^2)$ and $Y_i \sim N(0, 1.3^2)$.

(a) Calculate,

$$v = \frac{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)^2}{\frac{(\sigma_1^2/n_1)^2}{n_1-1} + \frac{(\sigma_2^2/n_2)^2}{n_2-1}}$$

This can be performed in Julia by substituting the values given in the question ($n_1 = n_2 = 10, \sigma_1 = 0.7, \sigma_2 = 1.3$) into the equation given above.

```
In [30]: n8=[10,10]
          var8= [0.7^2,1.3^2]
          xDist= Normal(0,0.7)
          yDist= Normal(0,1.3)

          v8a = (var8[1]/n8[1]+var8[2]/n8[2])^2/((var8[1]/n8[1])^2/(n8[1]-1)+
          (var8[2]/n8[2])^2/(n8[2]-1))
```

Out [30]: 13.814223887345776

(b) Generate 10^6 replicates of the random variable

$$T = \frac{X - Y}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Note that each generation of T requires 10 X 's and 10 Y 's.

The code to generate these T is given below.

```
In [31]: xTest=rand(xDist,10,10^6)
          yTest=rand(yDist,10,10^6)

          T=(mean(xTest,1).-mean(yTest,1))./sqrt.(var(xTest,1)/10.+
          var(yTest,1)/10)
```

Out [31]: 1×1000000 Array{Float64,2}:
 -0.189168 0.235178 1.76341 0.324764 1.65242 -0.0206177
 -0.0561863

(c) Compare the 0.8, 0.9 and 0.95 quantiles of a t -distribution with v degrees of freedom and the empirical quantiles obtained from the simulation.

To find the quantiles of the t -distribution with v degrees of freedom and those from the vector T the following code is used

```
In [32]: quantile(T[1,:],[0.8,0.9,0.95]),quantile(TDist(v8a),[0.8,0.9,0.95])
```

Out [32]: ([0.869779,1.34458,1.762],[0.868421,1.34593,1.76299])

Looking at the two sets of quantiles, they appear to be within 0.001 of each other. Given the variability possible in such a simulation this would strongly support the validity of the approximation.

- (d) Explain your results and comment about this simulation experiment. Are there other ways to carry it out?

One possible way to conduct the simulation experiment is to use the bisection method to find where the degrees of freedom lie between the conservative estimate of 9 and the equal variance value of 18. In this the quantiles would have to be compared at each iteration to the ideal result.