Problem Set 6

cdurso

March 12, 2019

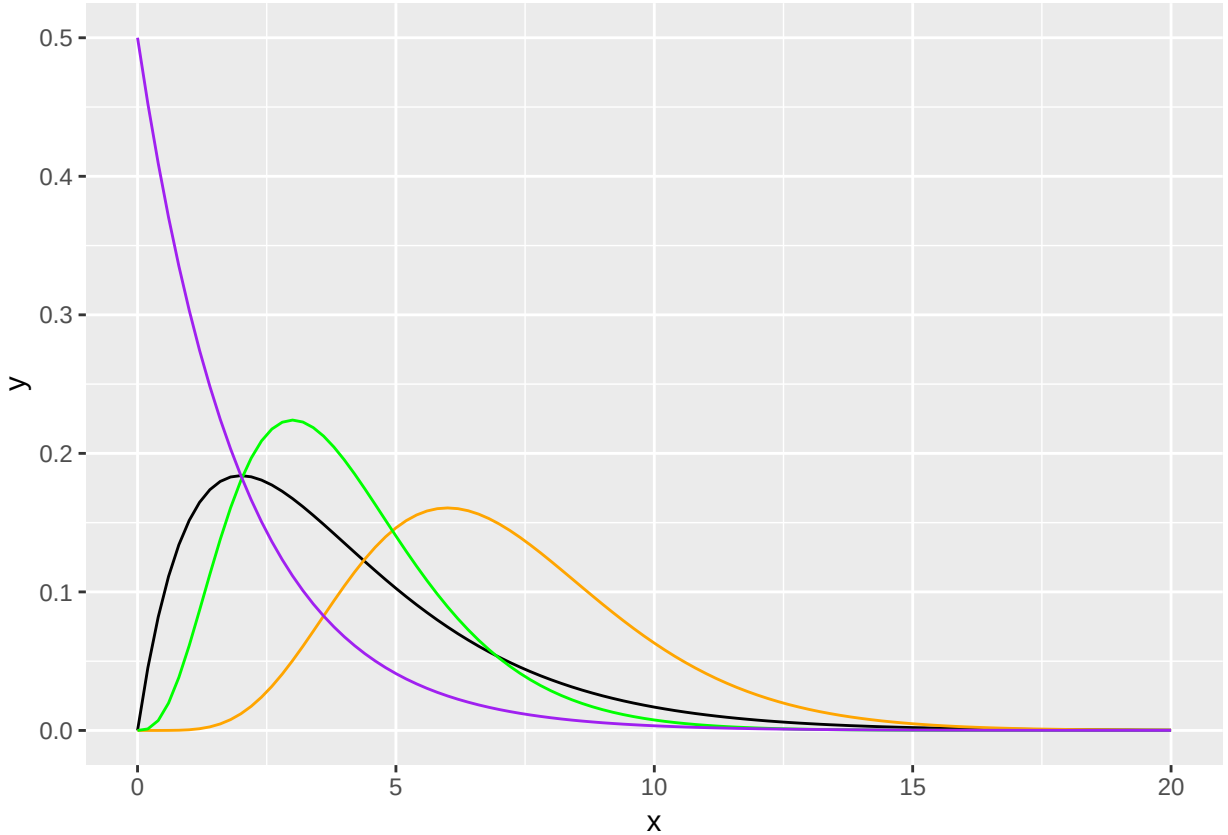
These questions were rendered in R markdown through RStudio (https://www.rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf, http://rmarkdown.rstudio.com).

Please complete the following tasks, in R where applicable. Please generate a solution document in R markdown and upload the rendered .doc, .docx, or .pdf document. You may add hand computatios to a .doc or .docx if you prefer. In the rendered document, please show your code.That is, don't use "echo=FALSE".

In either case, your work should be based on the data's being in the same folder as the R files. Please turn in your work on Canvas. Your solution document should have your answers to the questions and should display the requested plots.

1. (Robustness of Welch's test) The following code illustrates several members of the parametrized family of Gamma distributions.

```
theor<-data.frame(x=seq(0,20,by=.1))
ggplot(data=theor, aes(x=x))+stat_function(fun=dgamma,args=list(shape=2,scale=2))+
  stat_function(fun=dgamma,args=list(shape=7,scale=1) , color="orange")+
  stat_function(fun=dgamma,args=list(shape=4,scale=1) , color="green")+
  stat_function(fun=dgamma,args=list(shape=1,scale=2) , color="purple")
```



- 1.a. The function below generates two Normal samples of size 10, applies Welch's test, and reports the results. The vector "samp.norm" has 10,000 p-values generated by this process. Please use "samp.norm" to examine how well the approximate test performed. A plot of the quantiles of samp.norm against the

theoretical quantiles for the uniform distribution on [0,1] is provided. How is it related to this question? Also, what is the meaning the proportion of p-values less than or equal to 0.05 for these distributions? (4 points)

This graph shows that the distribution of p-values from 0 to 1 are evenly distributed/represented by this experiment. In the second graph (close-up) we also notice that the values lie very close to the line and we notice some small deviations for p-value < 0.05 showing that those p-values are not as evenly distributed/represented. From the set-up of the experiment we know that the samples used satisfy the conditions of Welch's test, i.e. normality of the sample's distribution.

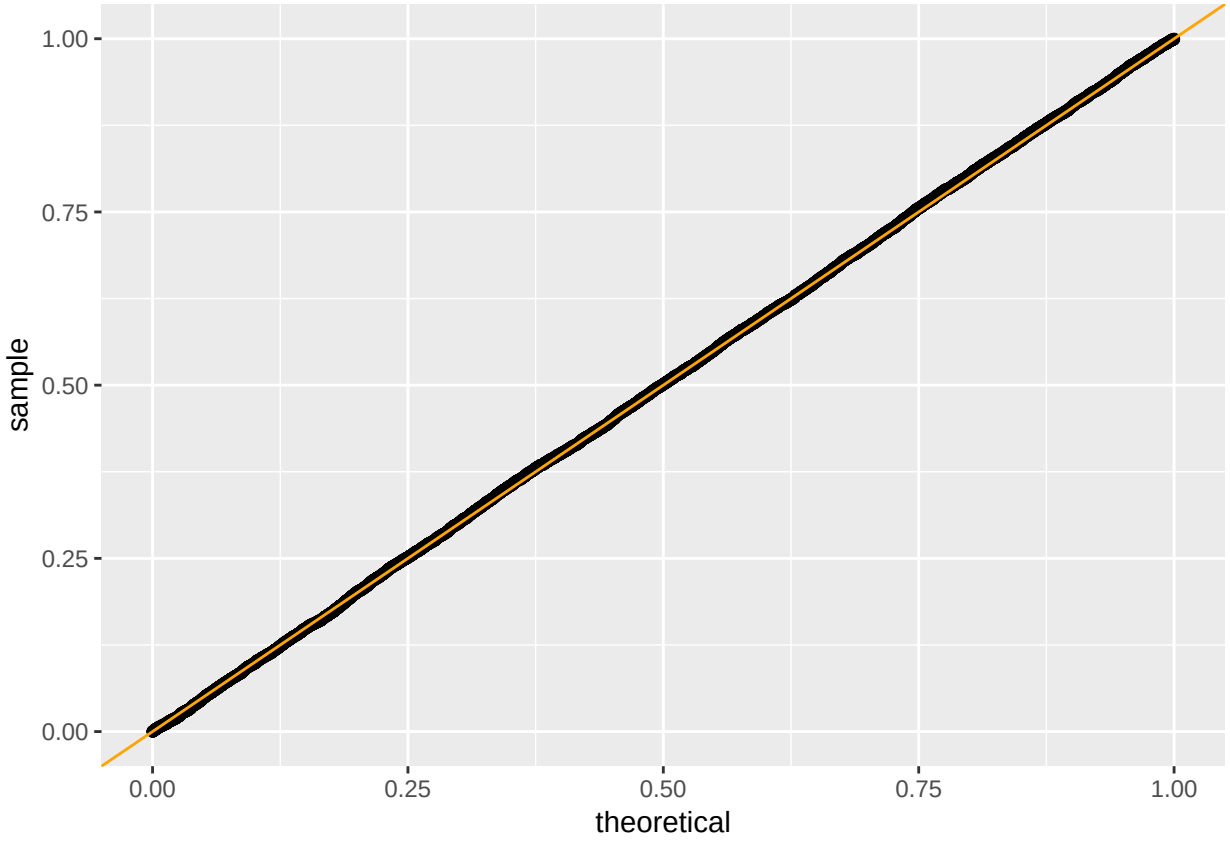
The proportion of p-values is 0.0513 (roughly 0.05). This corresponds to the value we expect to have outside of the confidence interval of 95%, which in this case corresponds to the proportion of false negatives, i.e. we know the data has the same mean but the test is telling us otherwise.

```
normal.p<-function(){
  x<-rnorm(10)
  y<-rnorm(10,sd=2)
  t.test(x,y)$p.value
}

set.seed(56789)
samp.norm<-replicate(10000,normal.p())
mean(samp.norm<.05)
```

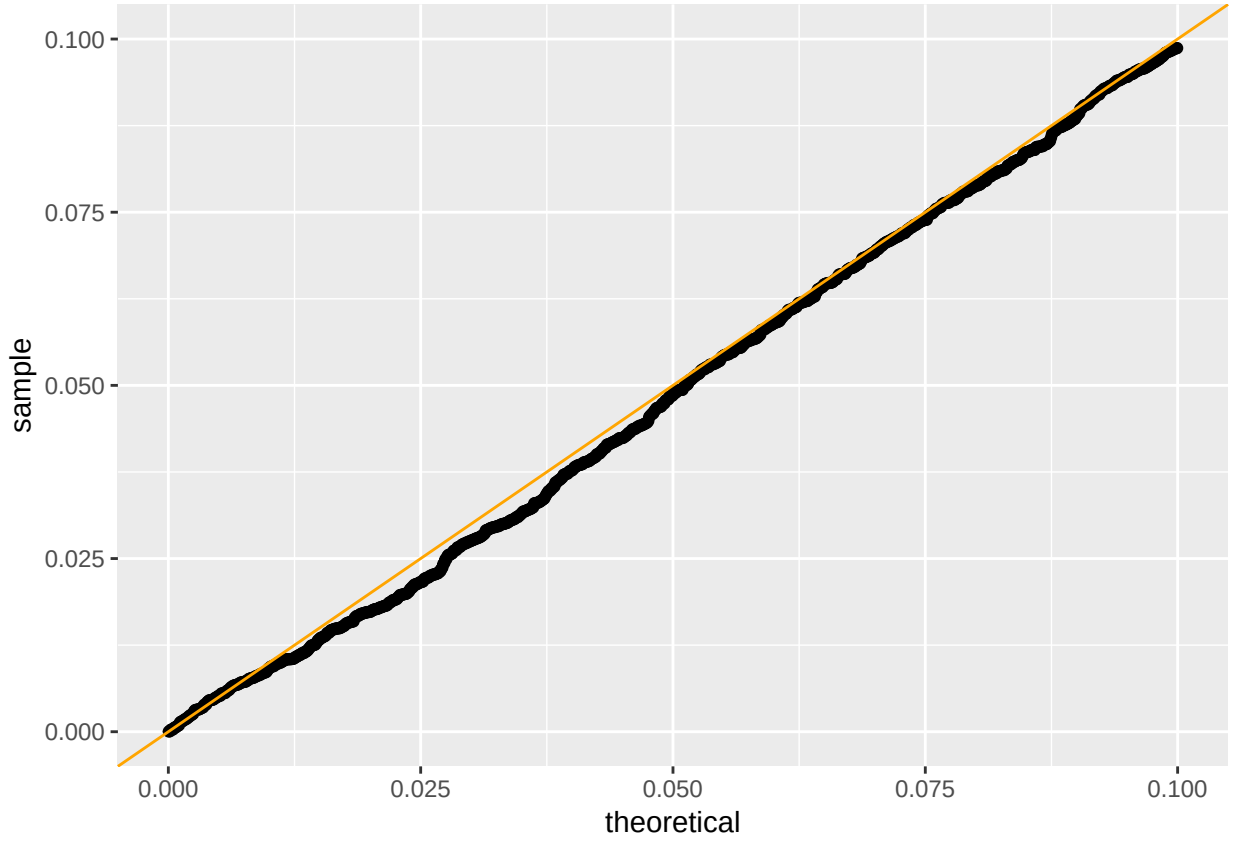
```
## [1] 0.0513
```

```
dat.norm<-data.frame(samp=samp.norm)
ggplot(data=dat.norm)+stat_qq(aes(sample=samp),distribution=stats::qunif)+
  geom_abline(slope=1,intercept=0,color="orange")
```



```
ggplot(data=dat.norm)+stat_qq(aes(sample=samp),distribution=stats::qunif)+xlim(0,.1)+ylim(0,.1)+
  geom_abline(slope=1,intercept=0,color="orange")
```

```
## Warning: Removed 9000 rows containing missing values (geom_point).
```



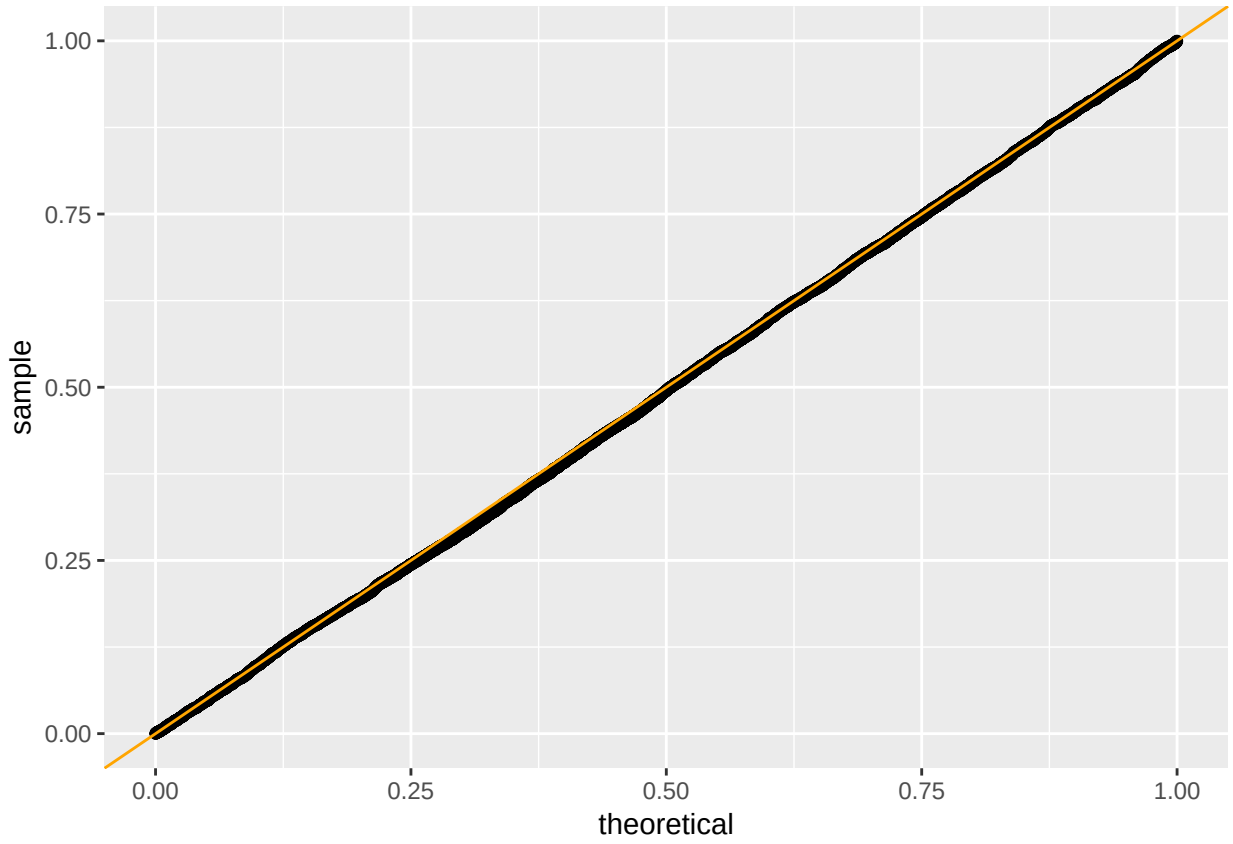
- 1.b. The function below generates two samples of size 10 from two members of the family of Gamma distributions, applies Welch's test, and reports the results. The vector "samp" has 10,000 p-values generated by this process. Please use "samp" to examine how well the approximate test performed under this violation of the hypotheses. Note both sampled distributions have the same mean=(shape)(scale). (4 points)

```
gamma.p.2.4<-function(){
  x<-rgamma(10,shape=2,scale=2)
  y<-rgamma(10,shape=4,scale=1)
  t.test(x,y)$p.value
}

set.seed(56789)
samp.2.4<-replicate(10000,gamma.p.2.4())
mean(samp.2.4<.05)
```

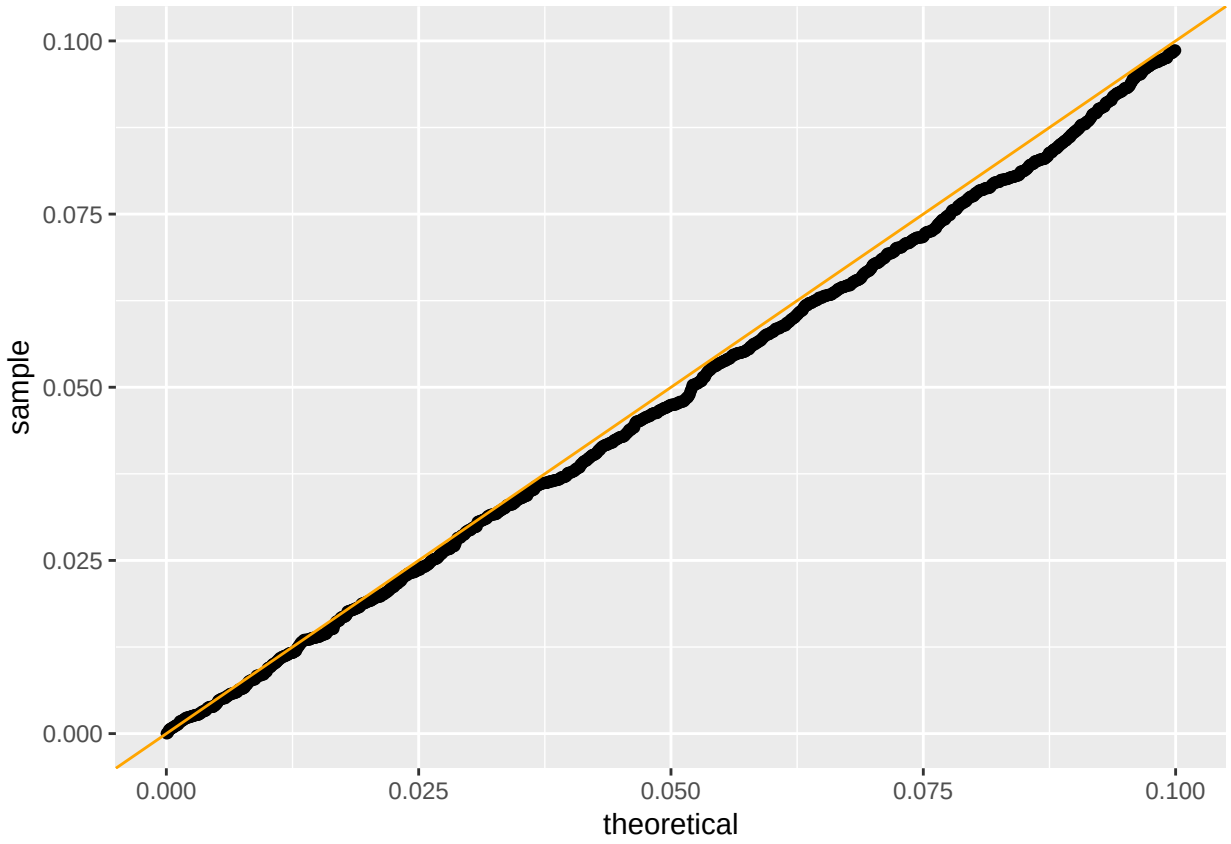
```
## [1] 0.0521
```

```
dat.2.4<-data.frame(samp=samp.2.4)
ggplot(data=dat.2.4)+stat_qq(aes(sample=samp),distribution=stats::qunif)+
  geom_abline(slope=1,intercept=0,color="orange")
```



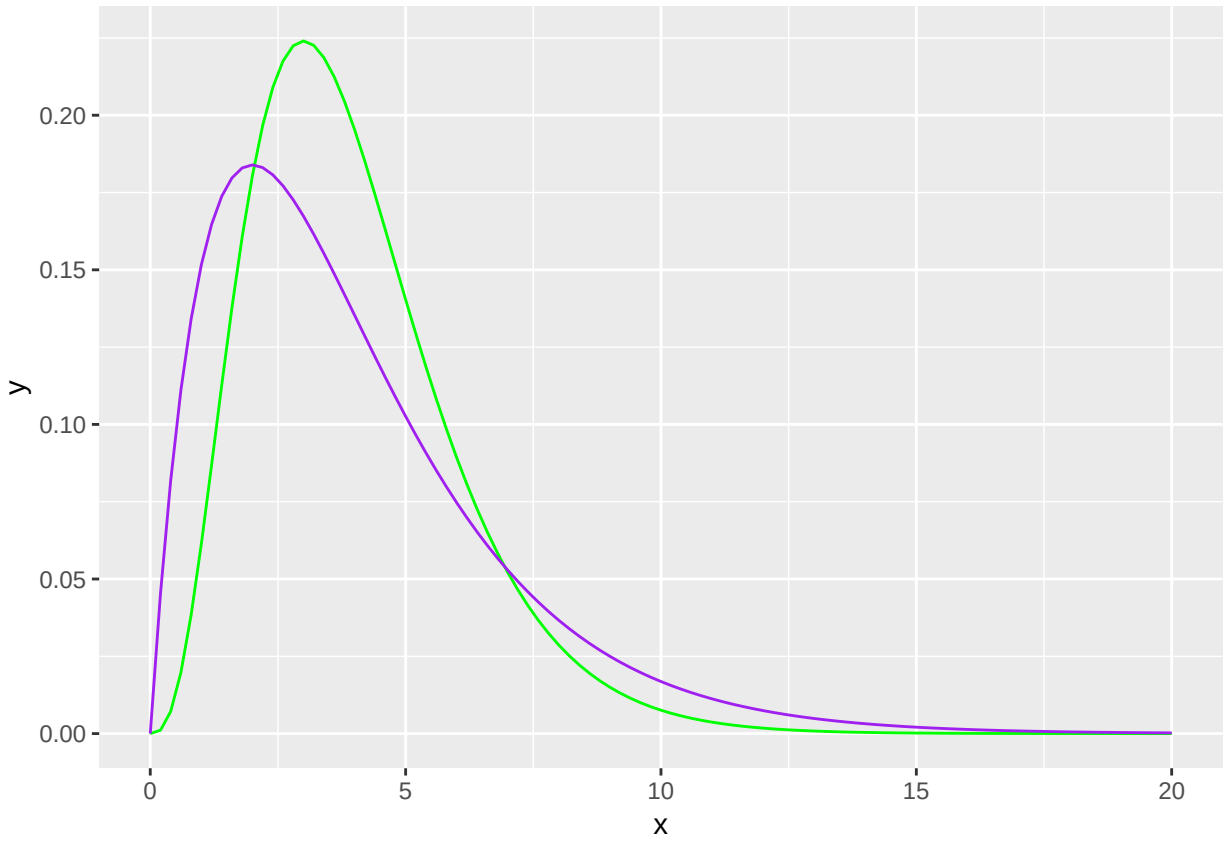

```
ggplot(data=dat.2.4)+stat_qq(aes(sample=samp),distribution=stats::qunif)+
  geom_abline(slope=1,intercept=0,color="orange")+xlim(0,.1)+ylim(0,.1)
```

```
## Warning: Removed 9000 rows containing missing values (geom_point).
```



This graph shows that the distribution of p-values from 0 to 1 are roughly evenly distributed/represented by this experiment. In the second graph (close-up) we notice that the values lie more or less on the line and we notice some deviations of the p-values, showing that those p-values are not as evenly distributed/represented. From the set-up of the experiment we know that the samples used do not satisfy the conditions of Welch's test, i.e. the samples come from gamma distributions and are strongly non-normal (c.f. image below).

```
ggplot(data=theor, aes(x=x))+
  stat_function(fun=dgamma,args=list(shape=4,scale=1) , color="green")+
  stat_function(fun=dgamma,args=list(shape=2,scale=2) , color="purple")
```



The proportion of p-values is 0.0521 (roughly 0.05), but higher than for question 1a). This increase corresponds to the difference and non-normality of our data and it corresponds to the proportion of false negatives, i.e. we know the data has the same mean but the test is telling us otherwise.

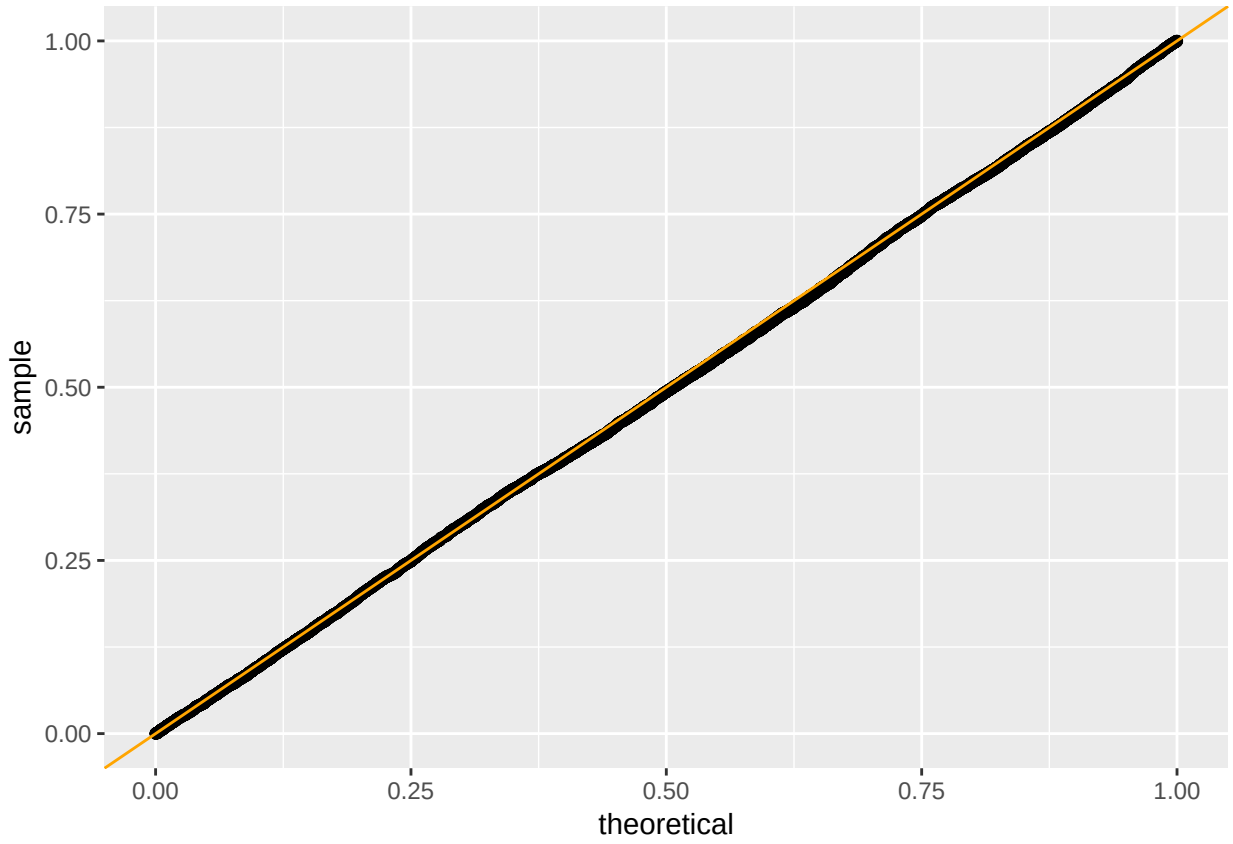
- 1.c. The function below generates two samples of size 10 from a member of the family of Gamma distributions, applies Welch's test, and reports the results. The vector "samp" has 10,000 p-values generated by this process. Please use "samp" to examine how well the approximate test performed under this violation of the hypotheses. (4 points)

```
gamma.p.7.1<-function(){
  x<-rgamma(10,shape=7,scale=1)
  y<-rgamma(10,shape=7,scale=1)
  t.test(x,y)$p.value
}

set.seed(56789)
samp.7.1<-replicate(10000,gamma.p.7.1())
mean(samp.7.1<.05)
```

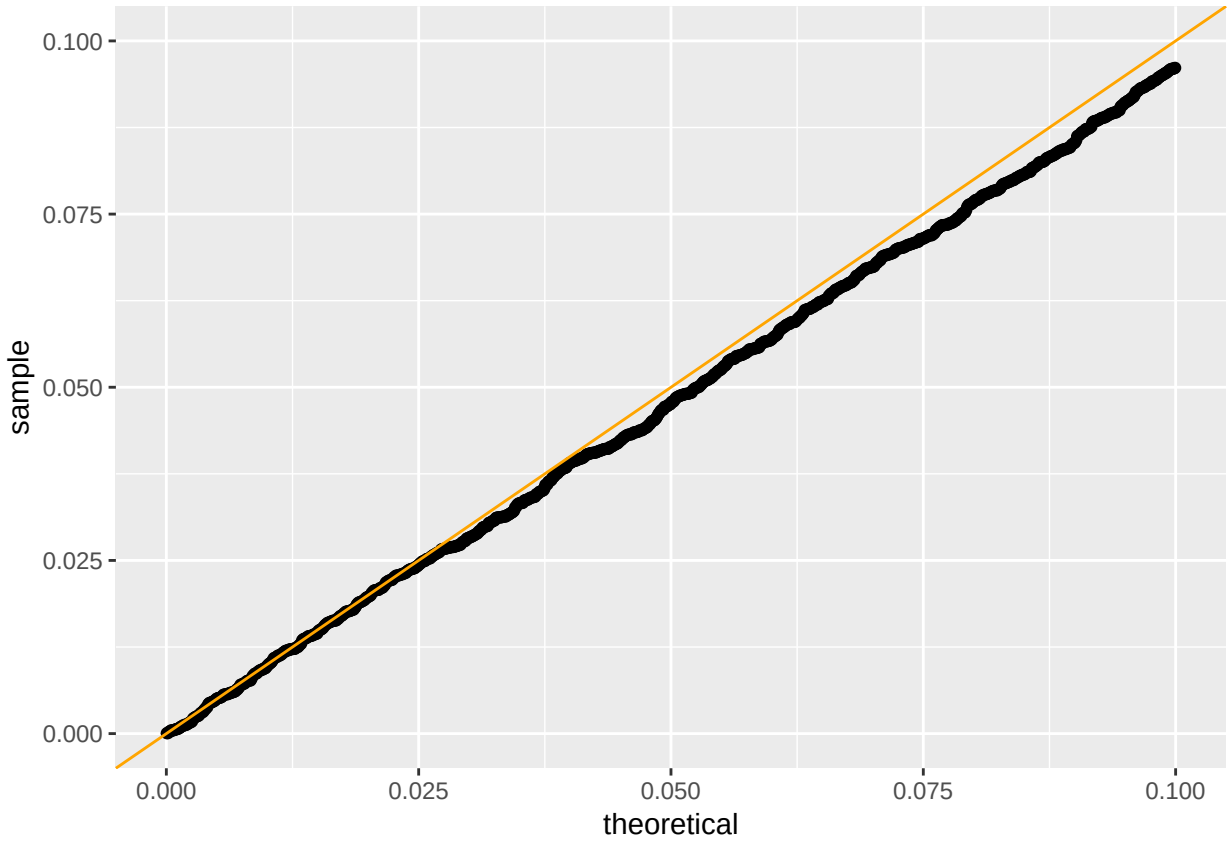
```
## [1] 0.0526
```

```
dat.7.1<-data.frame(samp=samp.7.1)
ggplot(data=dat.7.1)+stat_qq(aes(sample=samp),distribution=stats::qunif)+
  geom_abline(slope=1,intercept=0,color="orange")+xlim(0,1)+ylim(0,1)
```



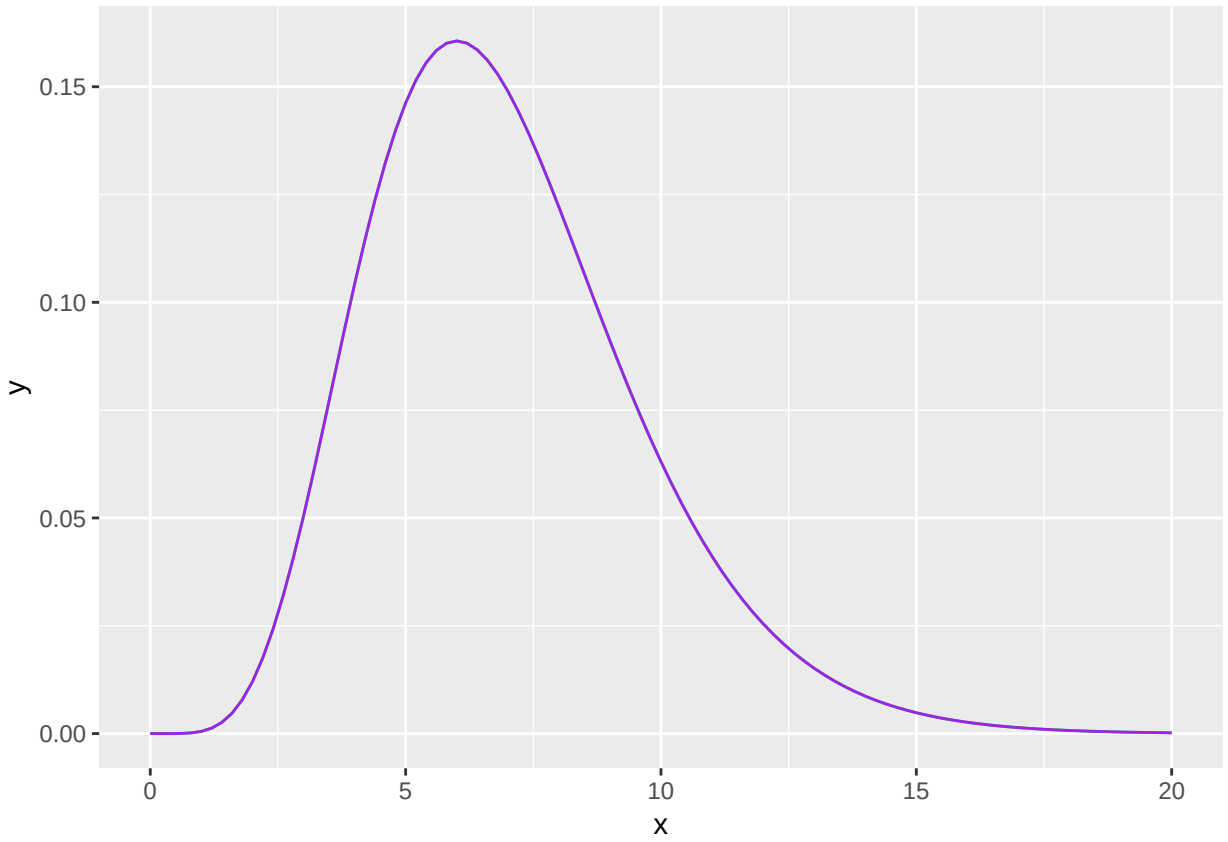

```
ggplot(data=dat.7.1)+stat_qq(aes(sample=samp),distribution=stats::qunif)+
  geom_abline(slope=1,intercept=0,color="orange")+xlim(0,.1)+ylim(0,.1)
```

```
## Warning: Removed 9000 rows containing missing values (geom_point).
```



This graph shows that the distribution of p-values from 0 to 1 are somewhat evenly distributed/represented by this experiment. In the second graph (close-up) we notice a significant deviation from the line of the p-values, showing that those p-values are not as evenly distributed/represented. From the set-up of the experiment we know that the samples used do not satisfy the conditions of Welch's test, i.e. the samples come from gamma distributions and are non-normal (c.f. image below).

```
ggplot(data=theor, aes(x=x))+
  stat_function(fun=dgamma,args=list(shape=7,scale=1) , color="green")+
  stat_function(fun=dgamma,args=list(shape=7,scale=1) , color="purple")
```



The proportion of p-values is 0.0526 (roughly 0.05), but higher than for question 1a) and 1b). This corresponds, roughly, to the value we expect to have outside of the confidence interval of 95%, which in this case corresponds to the proportion of false negatives, i.e. we know the data has the same mean but the test is telling us otherwise. We also take notice that the gamma distributions look more normal than in the previous question

- 1.d. The function below generates two samples of size 10 from a member of the family of Gamma distributions, applies Welch's test, and reports the results. The vector "samp" has 10,000 p-values generated by this process. Please use "samp" to examine how well the approximate test performed under this violation of the hypotheses.(3 points)

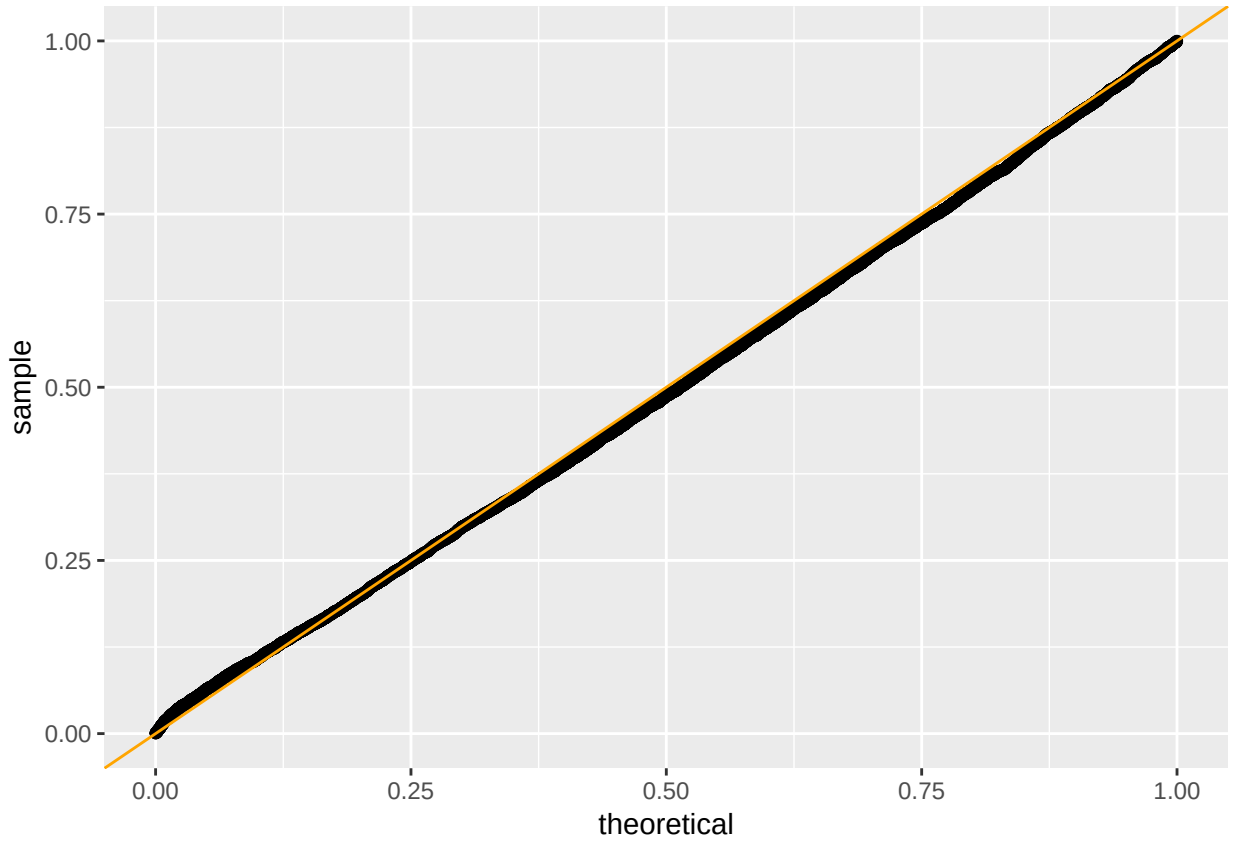
```
gamma.p.1.2<-function(){
  x<-rgamma(10,shape=1,scale=2)
  y<-rgamma(10,shape=1,scale=2)
  t.test(x,y)$p.value
}

set.seed(56789)
samp.1.2<-replicate(10000,gamma.p.1.2())
mean(samp.1.2<.05)
```

```
## [1] 0.0365
```

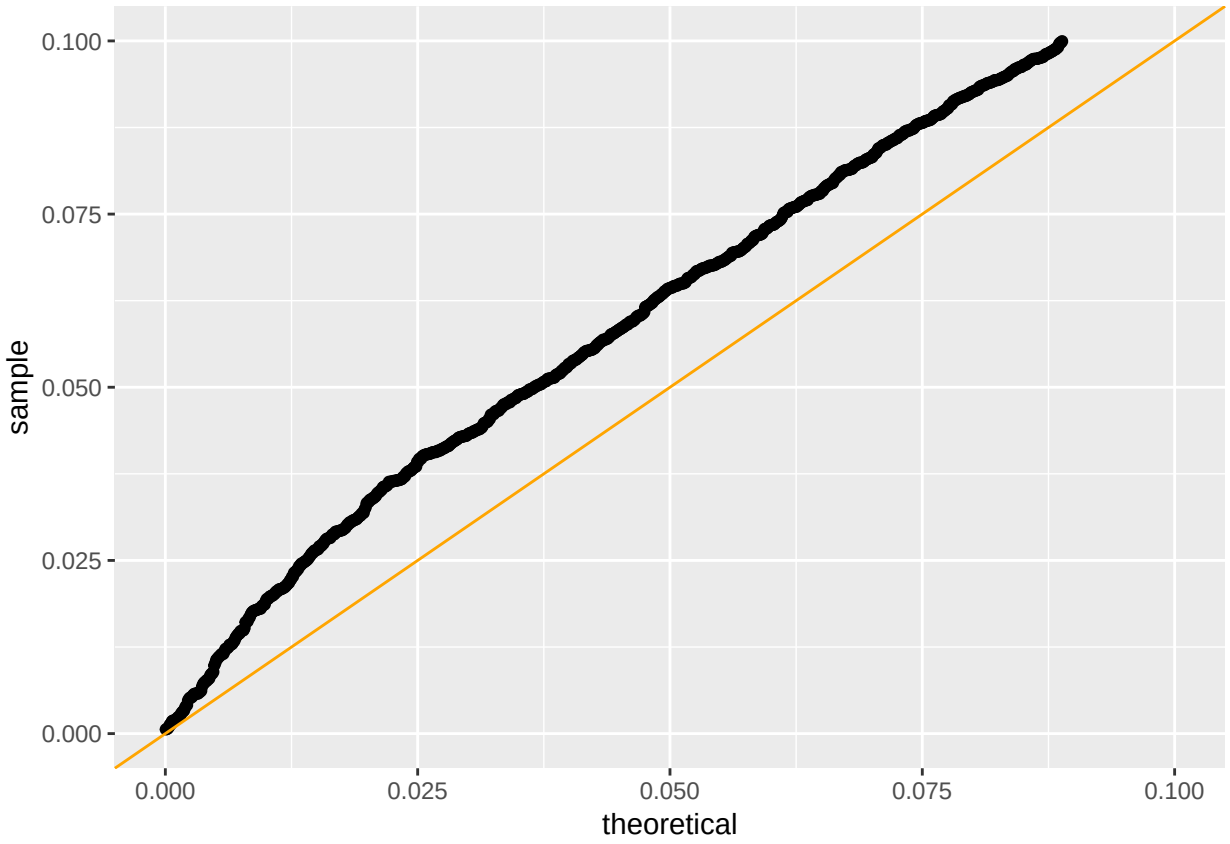
```
dat.1.2<-data.frame(samp=samp.1.2)

ggplot(data=dat.1.2)+stat_qq(aes(sample=samp),distribution=stats::qunif)+
  geom_abline(slope=1,intercept=0,color="orange")+xlim(0,1)+ylim(0,1)
```



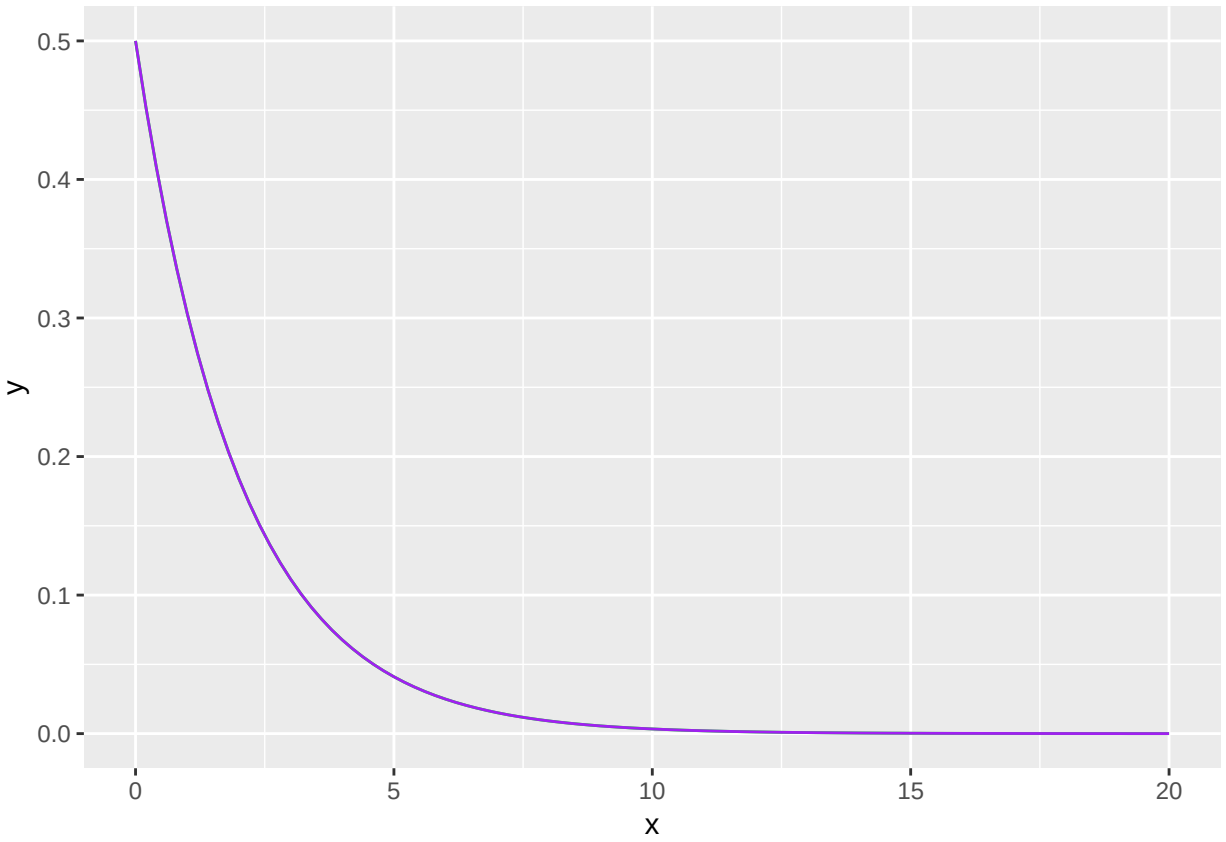

```
ggplot(data=dat.1.2)+stat_qq(aes(sample=samp),distribution=stats::qunif)+
  xlim(0,.1)+ylim(0,.1)+geom_abline(slope=1,intercept=0,color="orange")
```

```
## Warning: Removed 9111 rows containing missing values (geom_point).
```



This graph shows that the distribution of p-values from 0 to 1 are not very evenly distributed/represented by this experiment. In the second graph (close-up) we notice a very significant deviation from the line of the p-values, showing that those p-values are not as evenly distributed/represented. From the set-up of the experiment we know that the samples used do not satisfy the conditions of Welch's test, i.e. the samples come from gamma distributions and are strongly non-normal (c.f. image below).

```
ggplot(data=theor, aes(x=x))+
  stat_function(fun=dgamma,args=list(shape=1,scale=2) , color="green")+
  stat_function(fun=dgamma,args=list(shape=1,scale=2) , color="purple")
```



The proportion of p-values is 0.0365, fewer than for question 1a), 1b) and 1c). This decrease if due to the parity of both curves and that sampling only takes place on a "single" side of the gamma distribution, i.e. we do not see the tytpical gamma distribution graph given at the beginning of this problem set. The 0.0365 corresponds to the proportion of false negatives, i.e. we know the data has the same mean but the test is telling us otherwise.

2. Returning to the data from "precipitation_boulder.csv", the file "precipitation.RData" contains the data.frame dat.trim created from the code in ps5, followed by the command

save(dat.trim,file="precipitation.RData")

The command

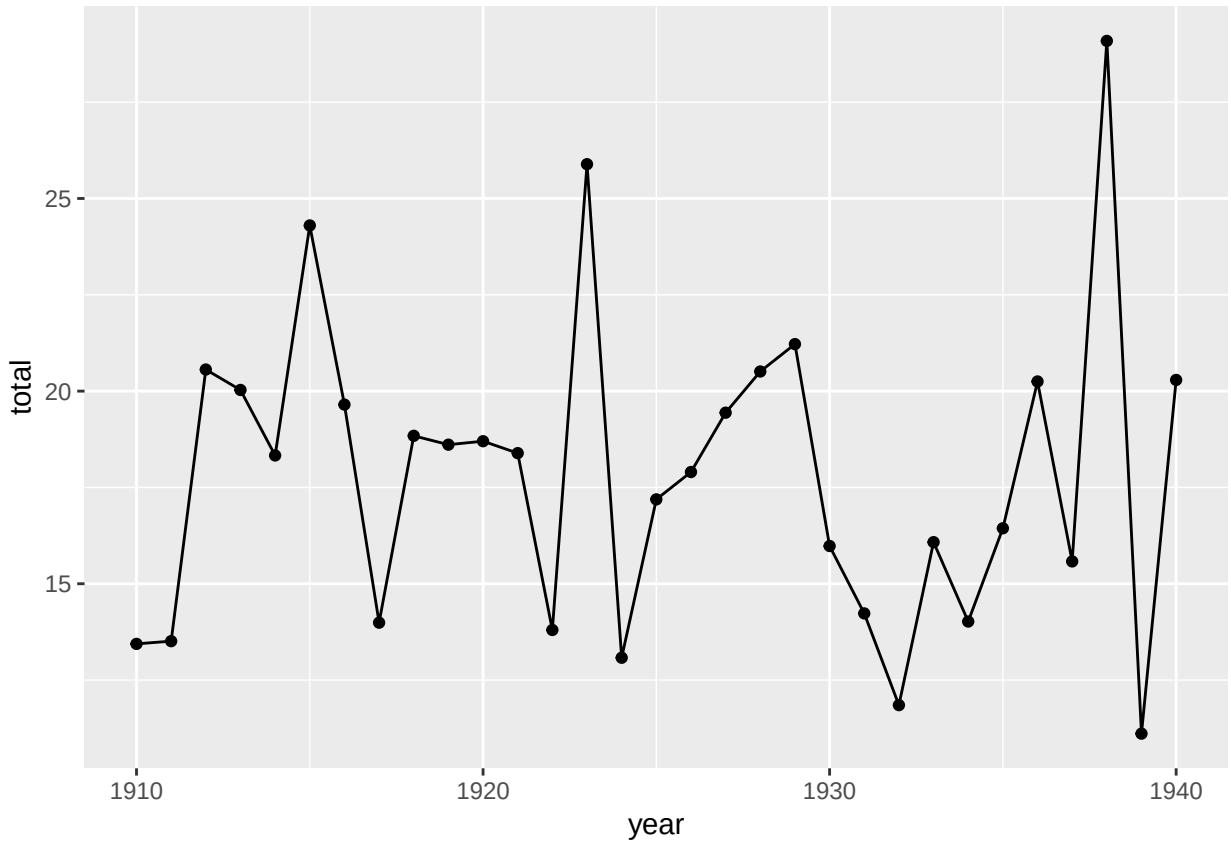
load("precipitation.RData")

imports dat.trim directly into the current environment.

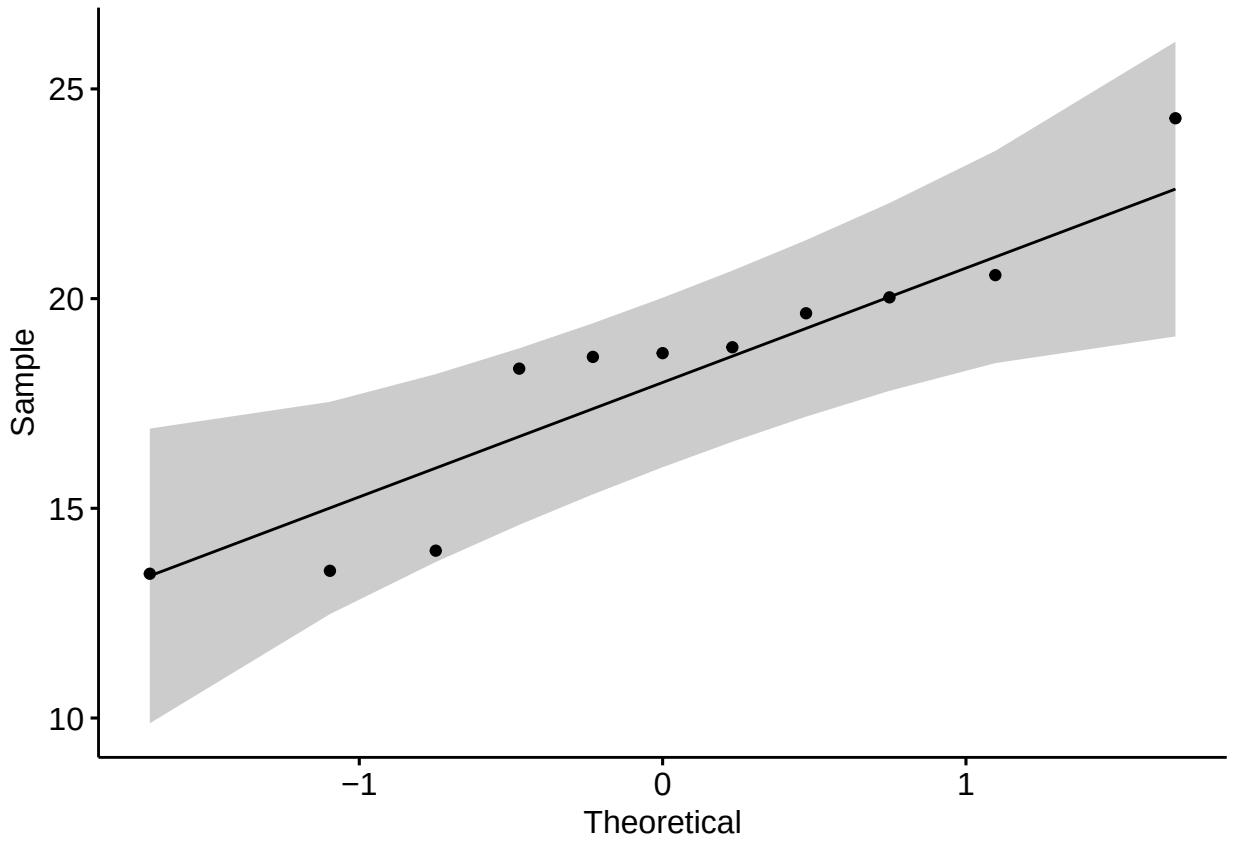
The period 1930-1940 in the Midwest was characterized by drought and dust storms. In this problem, you will examine whether precipitation in Boulder was also unusually low during this period, compared to the preceding period 1900-1920.

- 2.a. Are the annual precipitation totals in 1930-1940 approximately Normally distributed? Are the annual precipitation totals in 1910-1920 approximately Normally distributed? Please provide visual support for your conclusions. (5 points)

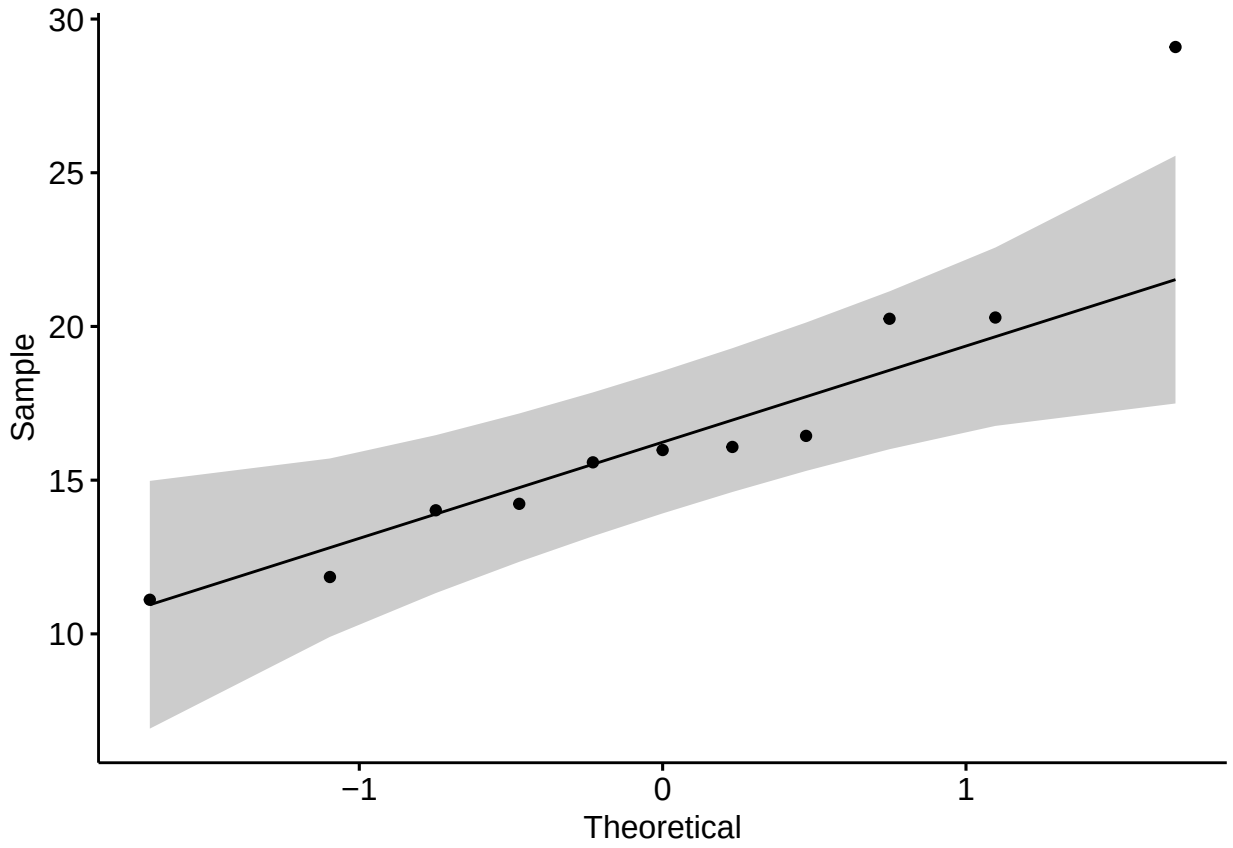
```
load("precipitation.RData")
ggplot(data=filter(dat.trim,year>=1910 & year<=1940),
       aes(x=year,y=total))+geom_line()+geom_point()
```




```
dat.old<-filter(dat.trim, year>=1910 & year<=1920)
dat.new<-filter(dat.trim, year>=1930 & year<=1940)
ggqqplot(dat.old$total)
```




```
ggqqplot(dat.new$total)
```



Both populations have qqplots that are comparable to qqplots for Normal samples of the same size.

- 2.b. Perform Welch's t test to test the hypothesis that the annual precipitation totals in 1930-1940 and the annual precipitation totals in 1910-1920 are drawn from populations with equal means. Does the result provide evidence for unusually low rainfall in the 1930-1940 time period? Please comment on the applicability of the test, based on your observations in 2.a. (5 points)

```
t.test(dat.old$total,dat.new$total)
```

```
##
##  Welch Two Sample t-test
##
## data:  dat.old$total and dat.new$total
## t = 0.75517, df = 17.455, p-value = 0.4602
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.445109  5.179655
## sample estimates:
## mean of x mean of y
##  18.17818  16.81091
```

```
mean(dat.new$total)
```

```
## [1] 16.81091
```

```
mean(dat.old$total)
```

```
## [1] 18.17818
```

The samples are from populations that can reasonably be viewed as Normal and independent from each other. The assumption that values for consecutive years are independent is less easily justified. One could argue

that for such a short period, long range climate effects aren't noticeable, and that short term effects are of shorter term than 1 year. Welch's test is an approximate test, and these data are on the small side for this test, but their satisfaction of the hypotheses makes this a reasonable test.

- 2.c. Are the variances of the annual precipitation totals in 1930-1940 and the annual precipitation totals in 1910-1920 approximately equal? (5 points)

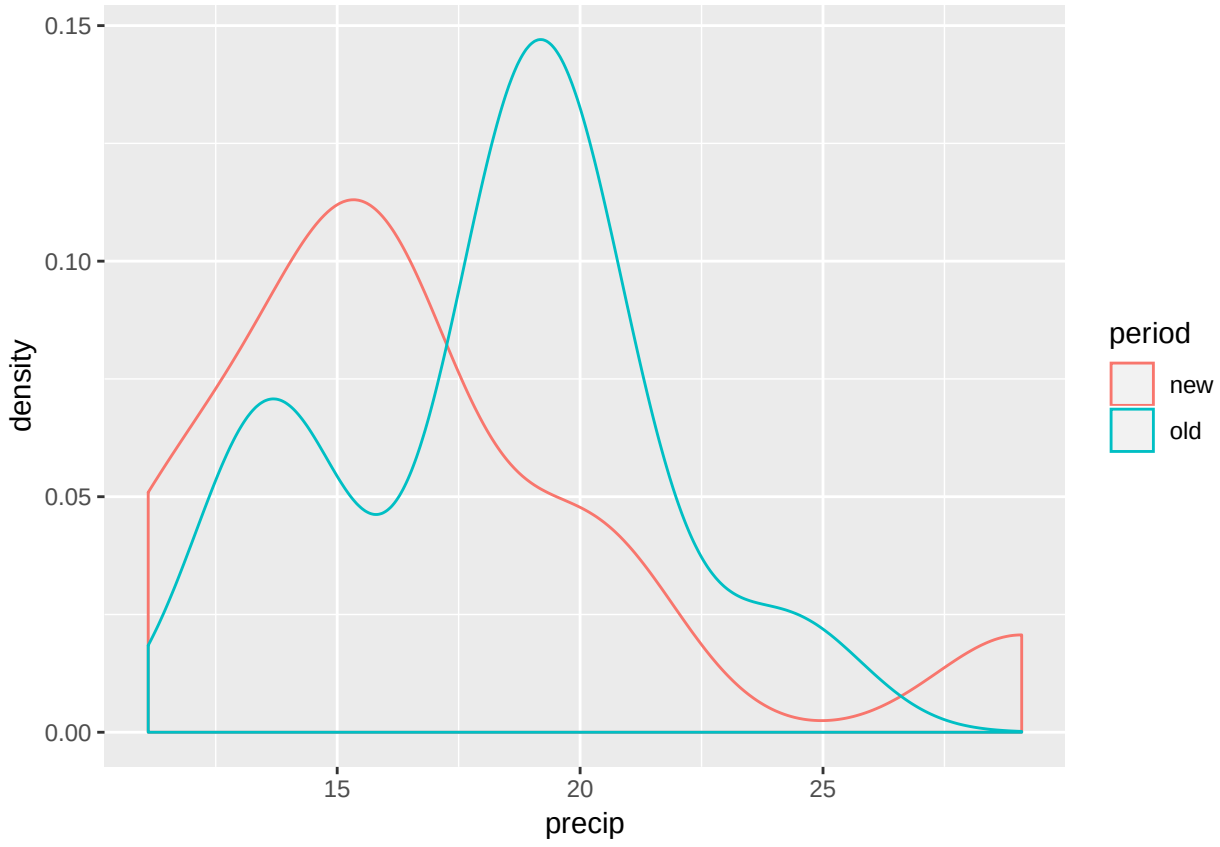
```
var.test(dat.old$total,dat.new$total)
```

```
##
##  F test to compare two variances
##
## data:  dat.old$total and dat.new$total
## F = 0.44732, num df = 10, denom df = 10, p-value = 0.2206
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.1203512 1.6625973
## sample estimates:
## ratio of variances
##          0.4473205
```

The variances are not statistically significantly different.

- 2.d. Please plot the densities of the total precipitation in 1930-1940 and 1910-1920. (5 points)

```
n.old<-nrow(dat.old)
n.new<-nrow(dat.new)
dat.both<-data.frame(period=rep(c("old","new"),c(n.old,n.new)),precip=c(dat.old$total,dat.new$total))
#ggplot(data=dat.both,aes(x=precip,fill=period))+geom_histogram(position="dodge",binwidth = 3)
ggplot(data=dat.both,aes(x=precip,color=period))+geom_density()
```



- 2.e. Please compare the total precipitation in the years 1930-1940 and 1910-1920 using the Mann-Whitney U-test, aka Wilcoxon rank sum test. Please interpret the results. (5 points)

```
wilcox.test(dat.old$total, dat.new$total)
```

```
##
##  Wilcoxon rank sum test
##
## data:  dat.old$total and dat.new$total
## W = 74, p-value = 0.4009
## alternative hypothesis: true location shift is not equal to 0
```

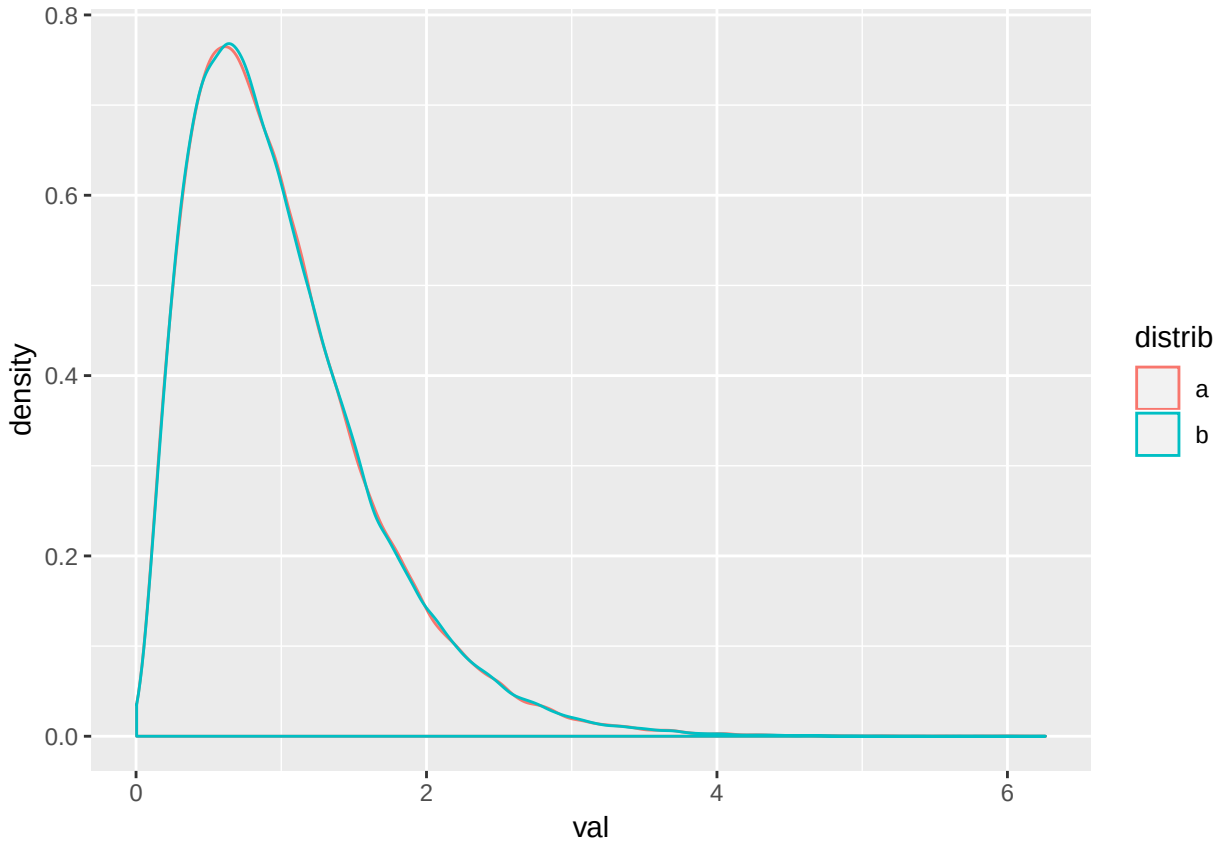
This test shows that we do not have a statistically different precipitation total for these two time periods.

3. Application of the F statistic to data of unknown mean depends on the claim the if $X_1, X_2, ...X_n$ and $Y_1, Y_2, ...Y_{n-1}$ are independent identically distributed samples from a Normal distribution with mean equal to μ then $\frac{\sum_{j=1}^{n}\left(X_i-\bar{X}\right)^2}{n-1}$ has the same distribution as $\frac{\sum_{j=1}^{n-1}(Y_i-\mu)^2}{n-1}$. Here $\bar{X} = \frac{\sum_{j=1}^{n} X_i}{n}$. (10 points)

Create a visualization to compare the density function of data created by sampling 100,000 times from the distributions of $\frac{\sum_{j=1}^{n}\left(X_i-\bar{X}\right)^2}{n-1}$ and $\frac{\sum_{j=1}^{n-1}(Y_i-\mu)^2}{n-1}$. The geom_density() option in ggplot2 may come in handy. (10 points)

```
mean.sq.diff<-function(){
  samp<-rnorm(6)
  return(sum((samp-mean(samp))^2)/5)
}
set.seed(8767)
n<-100000
m6<-replicate(n,mean.sq.diff())
m6<-sort(m6)

sum.sq<-function(){
  samp<-rnorm(5)
  return(sum((samp)^2)/5)
}
m5<-replicate(n,sum.sq())
m5<-sort(m5)
dat<-data.frame(distrib=rep(c("a","b"),each=n),val=c(m6,m5))
ggplot(data=dat, aes(x=val,color=distrib))+geom_density()
```


0.8
0.6
0.4
0.2
0.0
density
0
2
4
6
val
distrib
a
b