Problem Set 6

cdurso

March 07, 2019

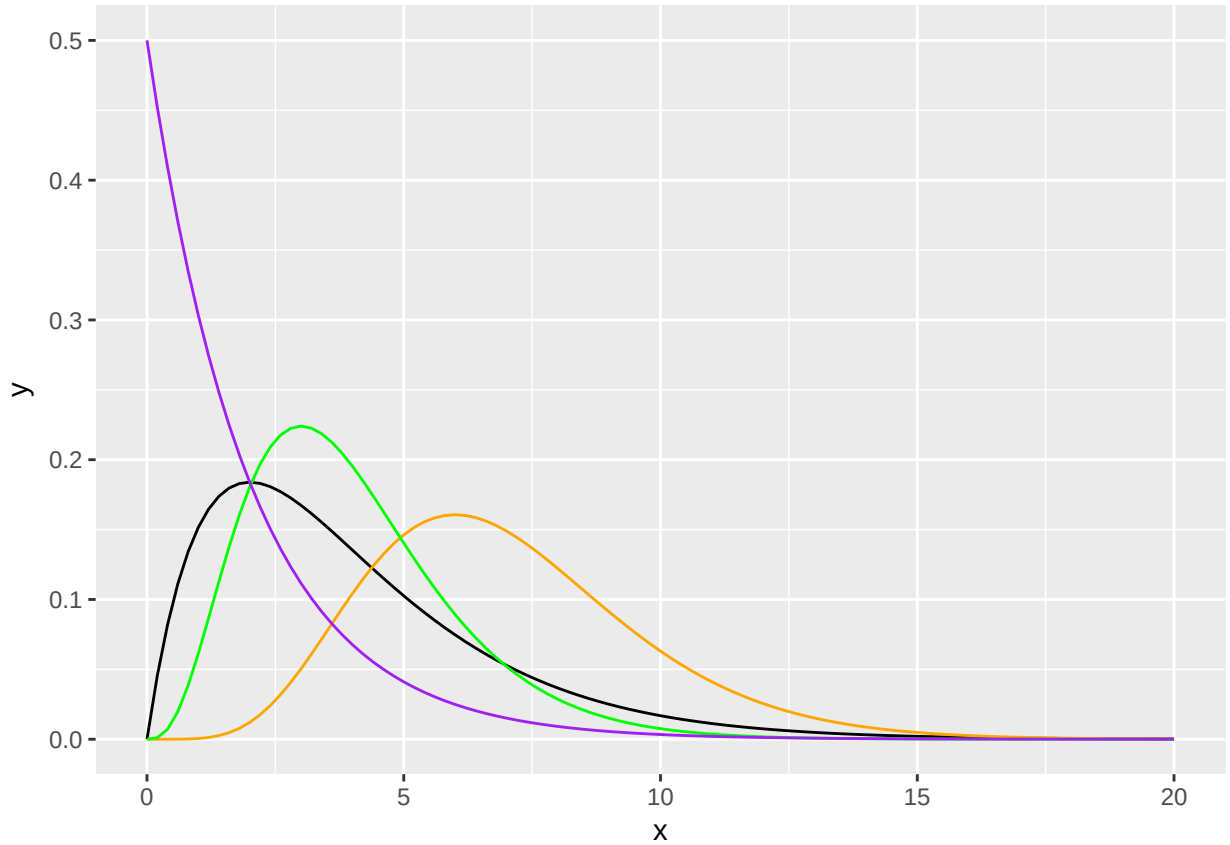
These questions were rendered in R markdown through RStudio (https://www.rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf, http://rmarkdown.rstudio.com).

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")
```



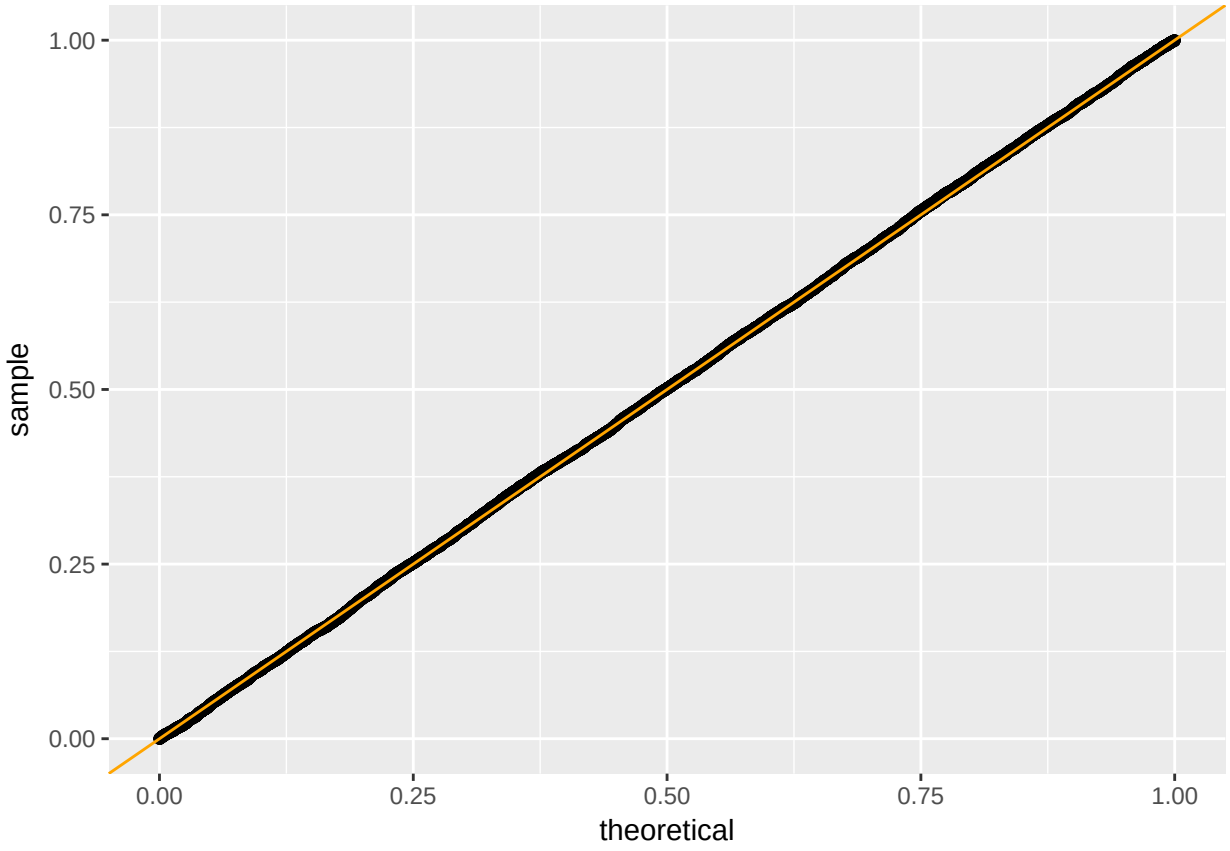
- 1.a. The function below generates two Normal samples of size 10, applies Welch's test, and reports the

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)

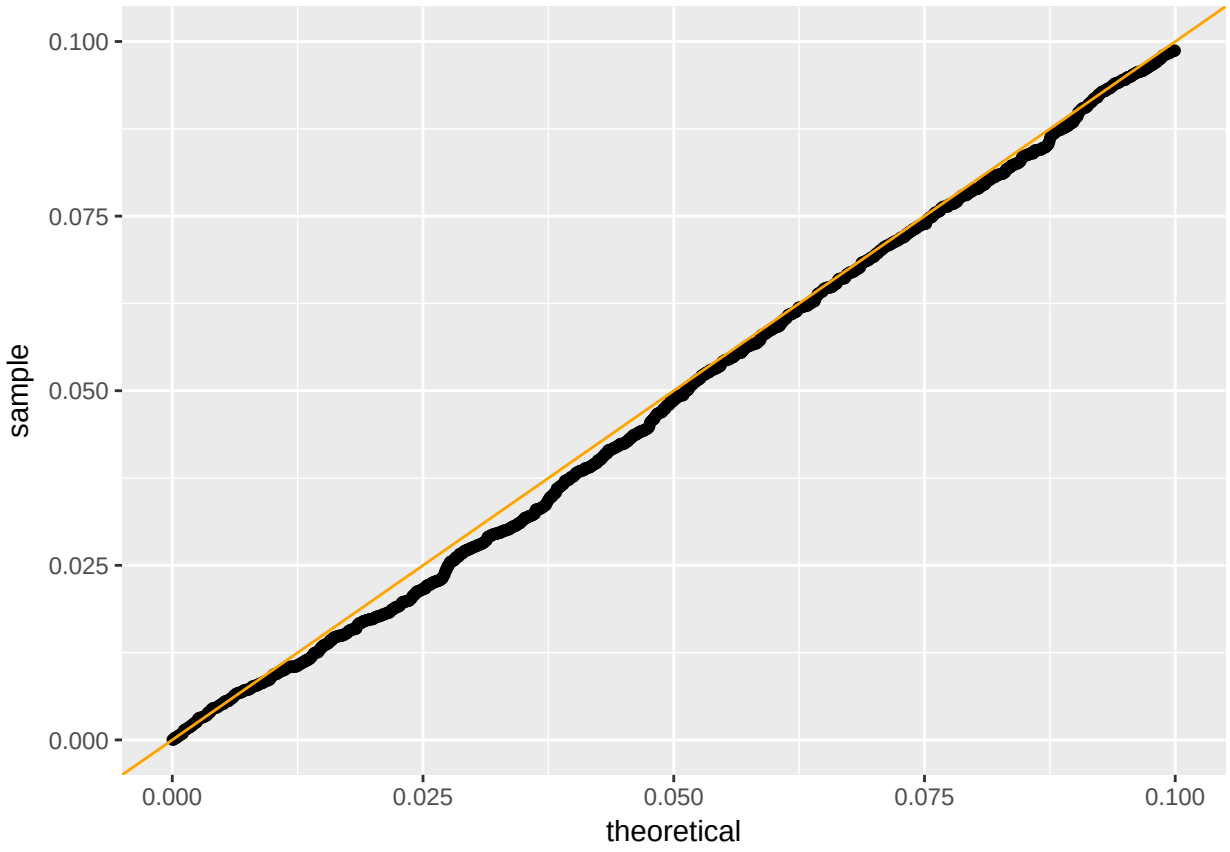
```
normal.p<-function(){
  x<-rnorm(10)
  y<-rnorm(10,sd=2)
  t.test(x,y)$p.value
}

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")
```



- 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.2.4" has 10,000 p-values generated by this process. Please use "samp.2.4" 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())
```

- 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.7.1" has 10,000 p-values generated by this process. Please use "samp.7.1" 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())
```

- 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.1.2" has 10,000 p-values generated by this process. Please use "samp.1.2" 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())
```

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

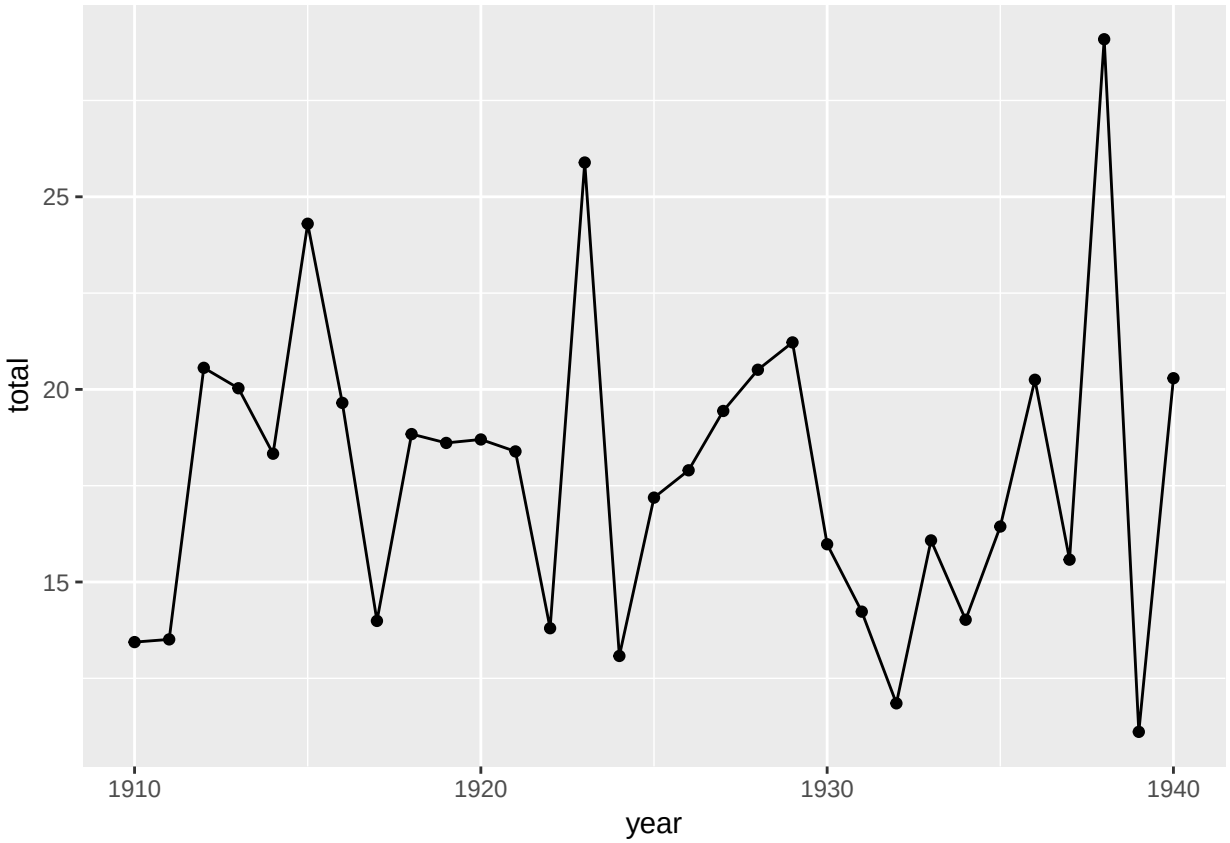
load("precipitation.RData")

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 1910-1920.

- 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()
```



- 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 relative to the earler period? Please comment on the applicability of the test, based on your observations in 2.a. (5 points)
- 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)
- 2.d. Please plot the densities of the total precipitation in 1930-1940 and 1910-1920. (5 points)
- 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)

3. Application of the F statistic to data of unknown means 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 $\bar{X} = \frac{\sum_{j=1}^{n} X_i}{n}$.

Create a visualization to compare the density function of data created by sampling 100,000 times from the distributions of $\frac{\sum_{j=1}^{n}\left(X_i-\bar{X}\right)^2}{\sqrt{n-1}}$ and $\frac{\sum_{j=1}^{n-1}(Y_i-\mu)^2}{\sqrt{n-1}}$ with the X's and Y's all independent samples from the standard Normal distribution. Please use $n = 6$. The geom_density() option in ggplot2 may come in handy. (10 points)