DSCI/MATH 530 RLab Six

Overview

In this assignment we will learn how to use R to

-generate a confidence interval for the difference of population means

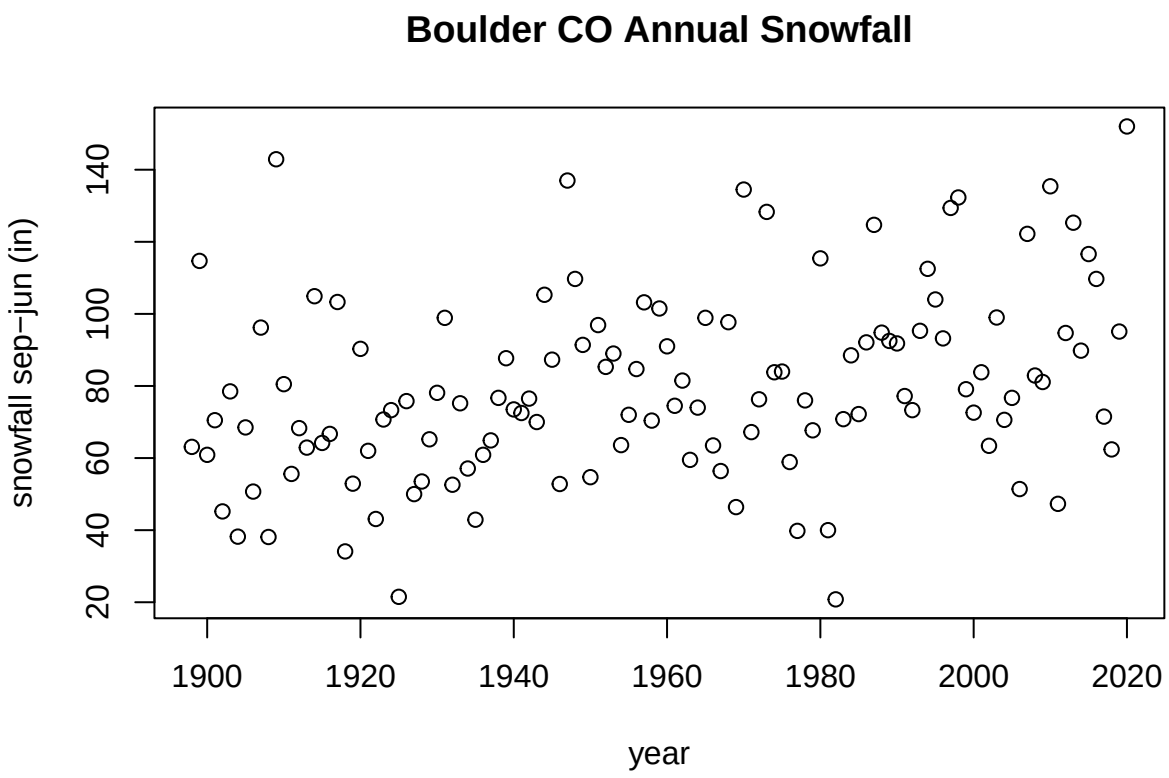
-perform a hypothesis test for the difference of means

-more interpretation of p-values.

1. Two independent samples

Read in the Boulder snowfall data and take a look at the times series plot for thse annual totals. In a previous R Lab we just selected the values after 1960 and created a confidence interval for the totals

```
BoulderSnowfall<-
       read.table("BoulderSnowfall.txt", header=TRUE )
# note subscripts based on the names to make this easier to
# follow year is also column 14 in the data object.
year<-    BoulderSnowfall[ ,'Year']
snow<- BoulderSnowfall[ ,'Sep.Jun']
plot( year, snow, ylab="snowfall sep-jun (in)", xlab="year")
title("Boulder CO Annual Snowfall")
```



1(a)

A simple way to find evidence for a trend over time in climate data is to divide the time period into two parts and consider a difference in population means between the first and second parts. Here we cut the data at 1960.

```
snowA<- snow[ year <= 1960]
snowB<- snow[ year > 1960]
```

Create a 95% confidence interval for the *difference* between the population means for these two periods. Based on this interval do you find any evidence for a change.

```r
sigmaA=sd(snowA)
sigmaB=sd(snowB)
sigma=sqrt(sigmaA^2/length(snowA)+sigmaB^2/length(snowB))
xbarA=mean(snowA)
xbarB=mean(snowB)
sigma
```

```
## [1] 4.574162
```

```r
xbar=xbarA-xbarB
xbar
```

```
## [1] -12.40183
```

```r
xbar+1.96*sigma
```

```
## [1] -3.436468
```

```r
xbar-1.96*sigma
```

```
## [1] -21.36718
```

1(b)

Referring to the data problem in (a) what is the p-value for the hypothesis test

$$H_0 : \mu_A = \mu_B \quad H_a : \mu_A \neq \mu_B$$

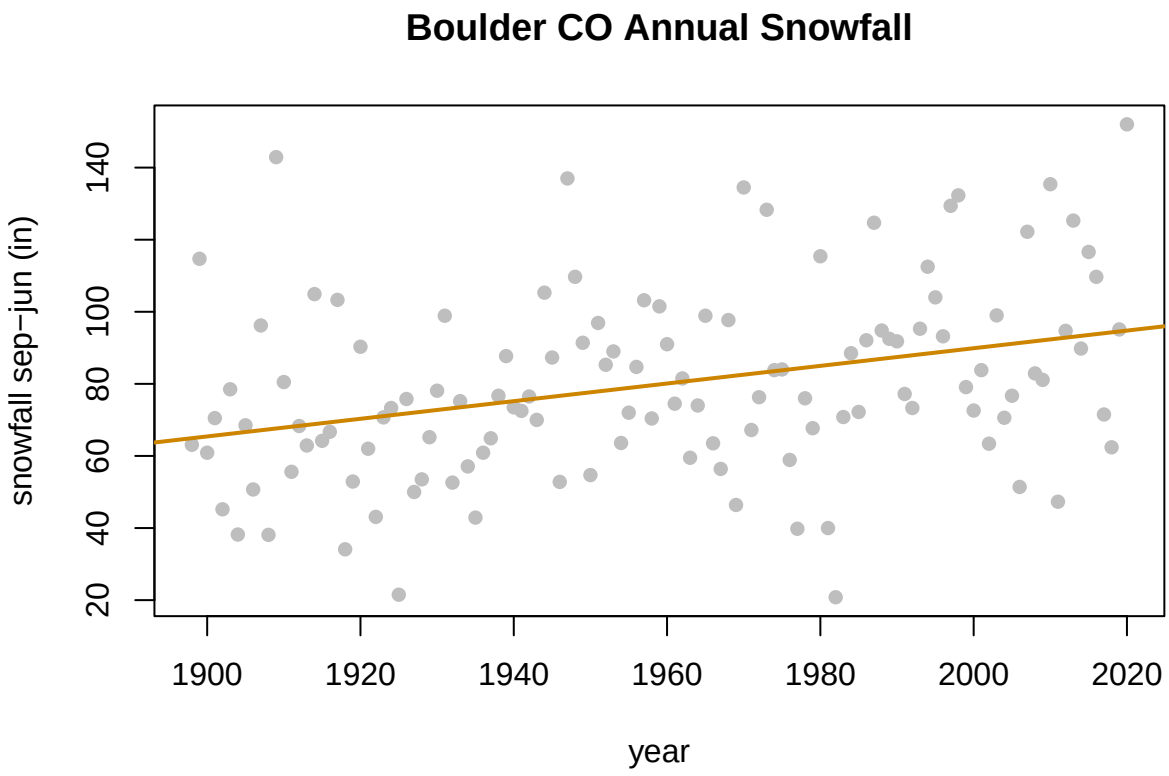
Explain how this result is consistent with your conclusion in (a).

```r
t=xbar/sigma
pnorm(t,0 ,1)
```

```
## [1] 0.00335122
```

WQe conclude ther eis a statisticallly significant difference between these two population means at the .05 level. Thi sis consistent with our condfidence interval since it doesn't contain zero. ## 1(c) EXTRA CREDIT Dividing up these data into two parts is a crude way to look for a change. Below is the result of fitting a line to these data and adding it to the time series plot.

```r
linearFit<- lm( snow ~ year)
beta1Hat<- linearFit$coefficients[1]
beta2Hat<- linearFit$coefficients[2]
plot( year, snow, ylab="snowfall sep-jun (in)", xlab="year",
      pch=16, col="grey")
title("Boulder CO Annual Snowfall")
abline( beta1Hat, beta2Hat, col="orange3",lwd=2)
```




```
look<- summary(linearFit)
look$coefficients
```

```
##                 Estimate   Std. Error   t value    Pr(>|t|)
## (Intercept) -399.5843009 121.97983599 -3.275823 0.001374680
## year           0.2447357   0.06225615  3.931108 0.000141346
```

```
pValue<- look$coefficients[2,4]
```

Specifically if the true slope is β_2 and true intercept β_1 the the trend line is $\beta_1 + \beta_2$Year. These two population parameters are estimated to be `-399.5843009` and `r beta2Hat`. Note that a positive slope that is significantly different from zero indicates a trend over time. The R code also gives a small table of some statistics from the fit – don't worry about the second and third columns. The last column are p-values. In particular, the p-value reported for the slope is based on the hypothesis test

$$H_0 : \beta_2 = 0 \quad H_a : \beta_2 \neq 0$$

Given this p-value is there evidence at the 99% level of confidence that there is a nonzero trend in in Boulder annual snowfall? Comparing this result with the test in part (b) which gives stronger statistical evidence for a change over time?

P value is less than .01 so there is a statistically significant slope. This is consistent with our difference in means from part B since we noted a statisitcal significant difference in means. Thus there is change over time.

2 Paired and independent samples. Let's compare two consecutive months – February and March over the period 1961-2020 and estimate the difference in snowfall between these two

```
feb<- BoulderSnowfall[ year>=1961,"Feb"]
mar<- BoulderSnowfall[  year>=1961,"Mar"]

mean( mar) - mean(feb)
```

```
## [1] 3.765
```

2(a)

Let μ_M be the population mean for March and μ_F the same for February. Find and interpret a 95% confidence interval for $\mu_M - \mu_F$.

```
sigmaA=sd(feb)
sigmaB=sd(mar)
sigma=sqrt(sigmaA^2/length(snowA)+sigmaB^2/length(snowB))
xbarA=mean(mar)
xbarB=mean(feb)
sigma
```

```
## [1] 1.864216
```

```
xbar=xbarA-xbarB
xbar
```

```
## [1] 3.765
```

```
xbar+1.96*sigma
```

```
## [1] 7.418863
```

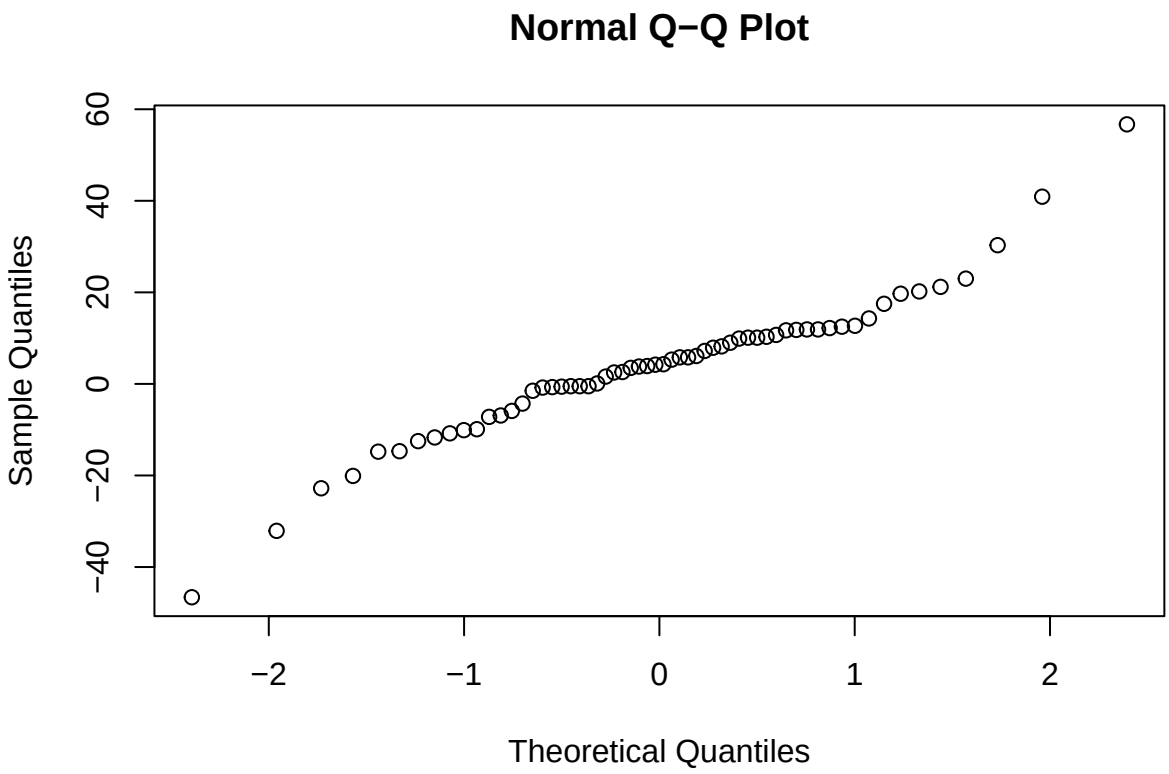
```
xbar-1.96*sigma
```

```
## [1] 0.1111368
```

2(b)

Make a Normal probability plot (**qqnorm**) of the (Mar- Feb) differences and comment on the normal distribution assumption.

```
qqnorm(mar-feb)
```

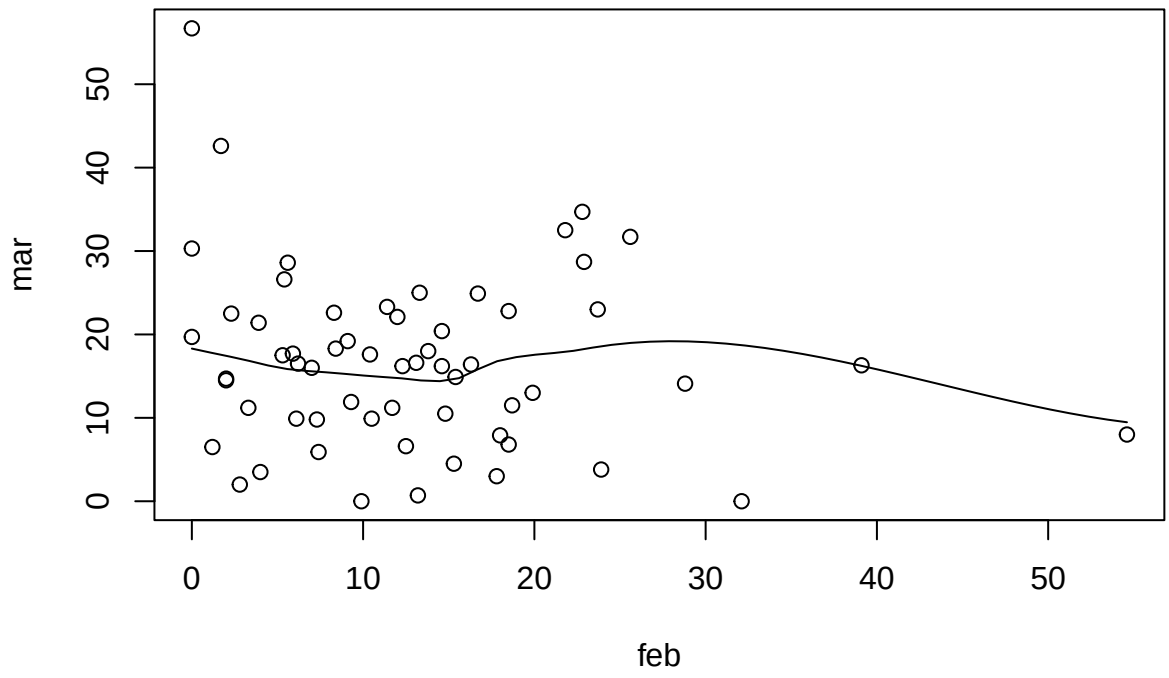


This is farily normally distributed with only a little tailing towards either end. Thus the dat ais normal or we can assume such.

3 Checking for matched pairs

Make a scatter plot of the february totals (X axis) against the march totals (Y axis). Based on this plot does the March snowfall appear to depend on the February value? Another way to think about this issue is: If we know February snowfall would it help us predict the snowfall for the next month? How would this judgement influence whether you use a matchedpaired or independent confidence interval method in 2(a)?

```
scatter.smooth(feb,mar)
```




```
linearFit<- lm(mar ~feb)
look<- summary(linearFit)
look$coefficients
```

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 18.8427422  2.2337537  8.435461 1.145717e-11
## feb         -0.1703293  0.1372606 -1.240919 2.196316e-01
```

```
pValue<- look$coefficients[2,4]
```

There is a statistically significant correlation between March and February at the .05 level. We can conclude we can predict March snow fall from February snowfall.