662midterm2018

Ty Darnell

7/17/2019

```
library(data.table)
library(tidyverse)
library(knitr)
```

```
mot=fread("data/Midterm_BWT.dat")
```

Problem 2)

Set impossible values to missing

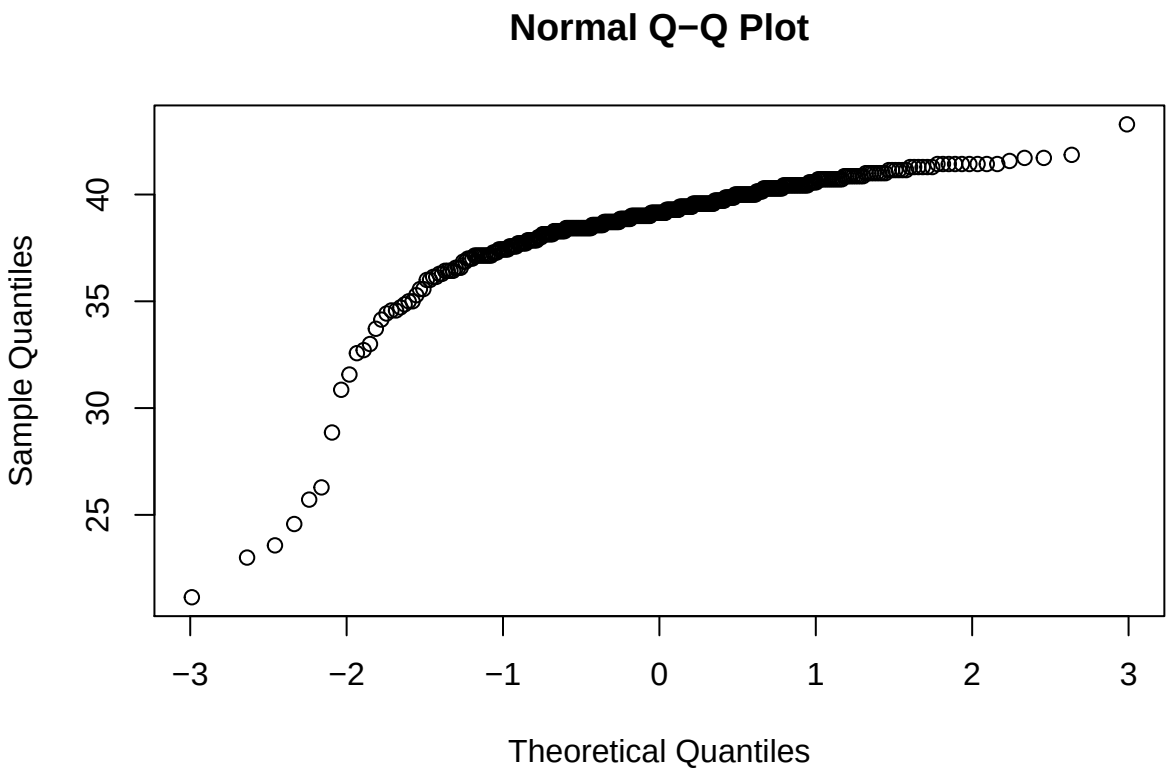
`ga_ultra` is the gestational age in weeks estimated by ultra sound

`ga_est` is the gestational age in weeks estimated at birth

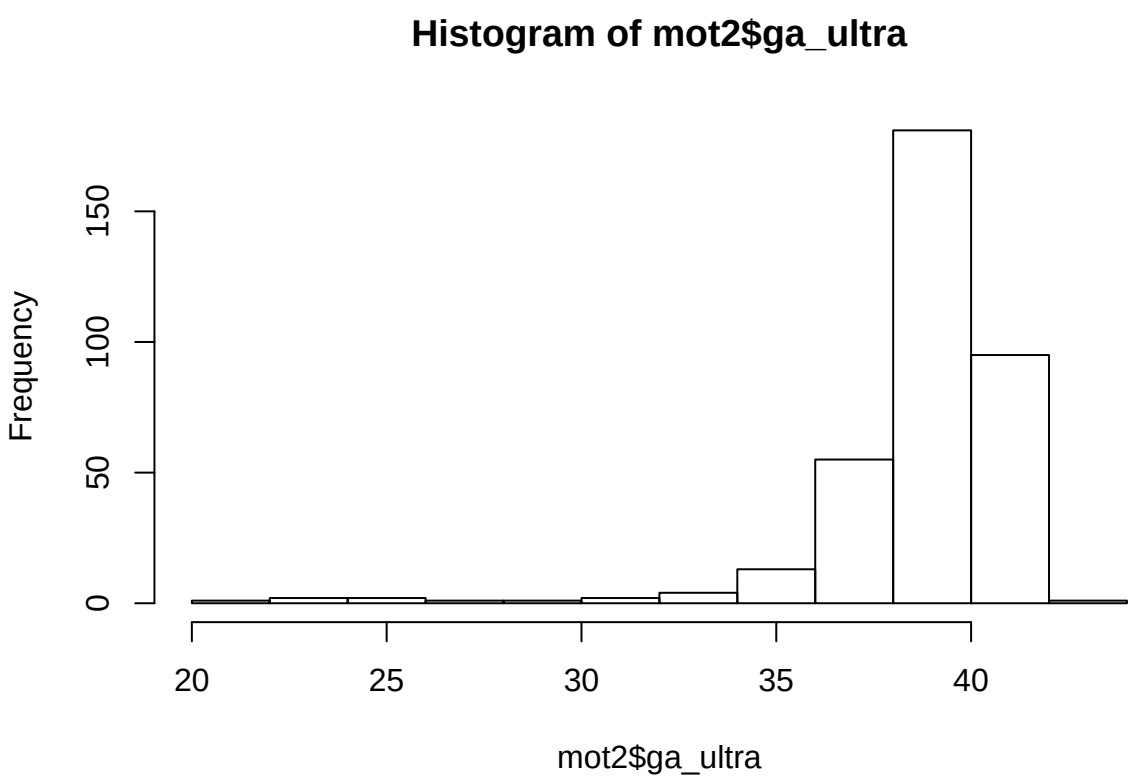
```
mot2=mot%>%mutate(ga_ultra=replace(ga_ultra,ga_ultra>70,NA))
mot2=mot2%>%mutate(rand_month=replace(rand_month,rand_month>12|rand_month<1,NA))
mot2=mot2%>%mutate(ppnum=replace(ppnum,ppnum<0,NA))
```

a) Is the ultrasound version of GA approximately normally distributed?

```
qqnorm(mot2$ga_ultra)
```




```
hist(mot2$ga_ultra,breaks=15)
```



The normal qq plot of ga_ultra suggests a departure from normalility. A histogram of ga_ultra suggests the data is left skewed. Using a lilliefors KS test to assess normality.

H_0: ga_ultra is normally distributed

H_A: ga_ultra is not normally distributed

```
library(nortest)
ult=mot2$ga_ultra
```

```
lillie.test(ult)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  ult
## D = 0.1859, p-value < 2.2e-16
```

p-value$= 2.2 * 10^{-16} \approx 0$

Using an α value of .05 Reject H_0 since p-value$< \alpha$

There is evidence that ga_ultra is not normally distributed.

b) Do the means of the two gestational age variables differ?

```
ult=mot2$ga_ultra
est=mot2$ga_est
diff=ult-est
```

```
a=mean(ult,na.rm=T)
b=mean(est,na.rm=T)
z=mean(diff,na.rm=T)/(sd(diff,na.rm=T)/sqrt(length(diff)))
```

The sample means are:

38.73 weeks for ga_ultra

38.41 weeks for ga_est

Since we have paired data, the two subsamples are not independent. Since we have a large sample (n=359), we will use the CLT/Slutsky's conduct a test of difference of means using a z statistic. Each observation of the difference of means vector is independent.

μ_{diff} = difference between means ga_ultra and ga_est

$\alpha = .05$

$H_0 : \mu_{diff} = 0$ (no difference in means)

$H_A : \mu_{diff} \neq 0$ (there is a difference in the means)

$Z = \dfrac{\bar{Y} - \mu_0}{s/\sqrt{n}} \sim N(0,1)$ (approximately normal)

$z_{stat} = \dfrac{\bar{y} - 0}{sd(y)\sqrt{359}} = 6.276453 \approx 6.28$

$C_{.05} = \{z : |z| > 1.96\}$

$|6.28| > 1.96$ Thus reject H_0, there is evidence that the difference in means is not 0.

Also since there is evidence that the means are different, and ga_ultra has a larger sample mean, the mean of ga_ultra appears larger.

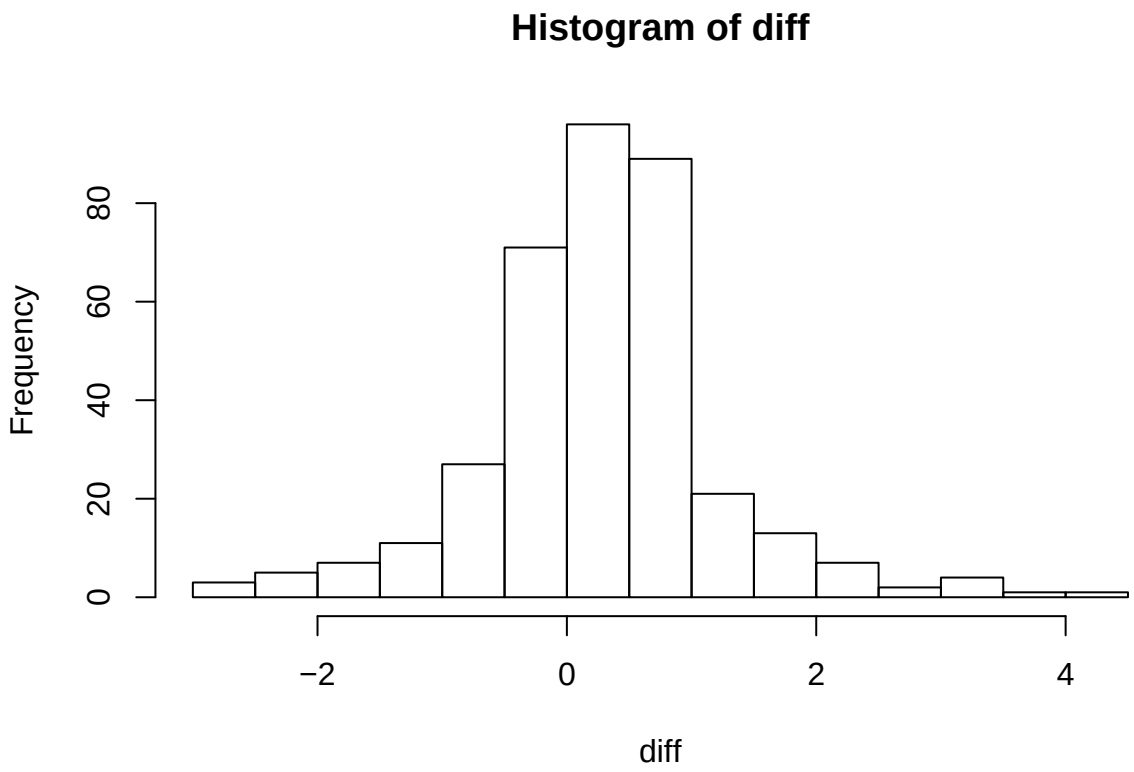
We could also use a one sample t-test using the same hypothesis:

```
t.test(diff)
```

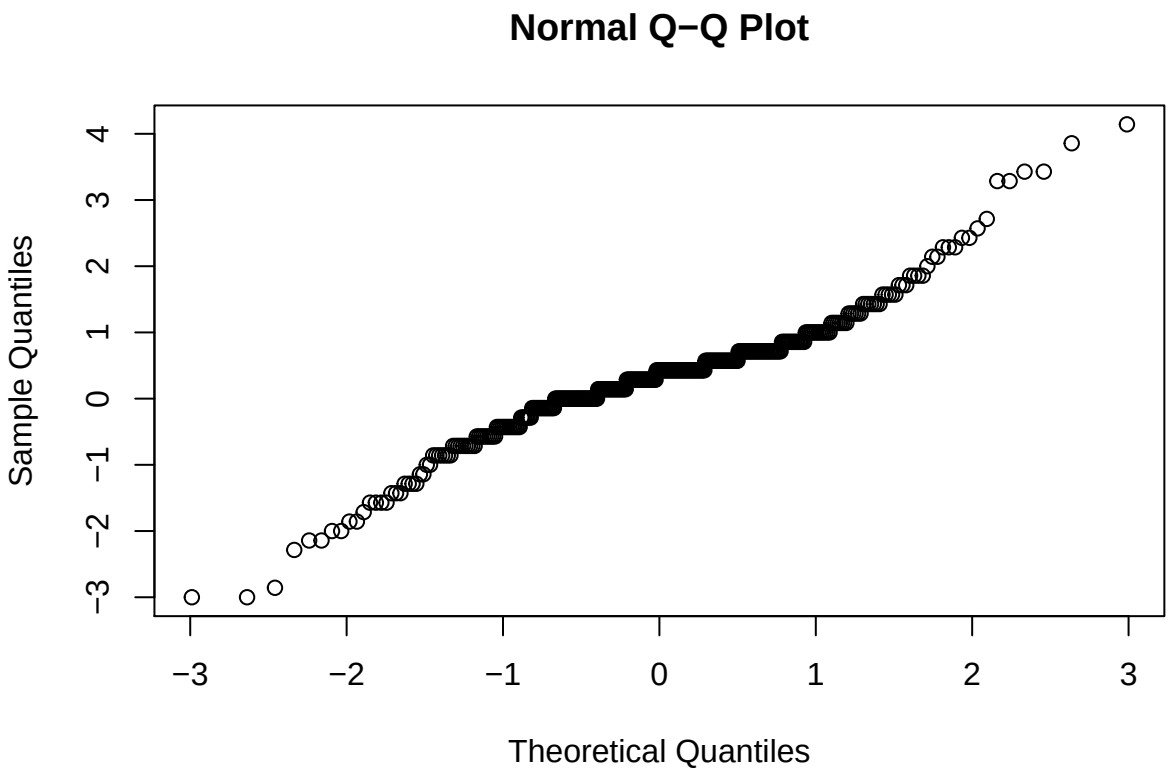
```
##
##  One Sample t-test
##
## data:  diff
## t = 6.2677, df = 357, p-value = 1.057e-09
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.2182471 0.4178305
## sample estimates:
## mean of x
## 0.3180388
```

p-value $< \alpha$ thus reject H_0 and conclude there is a difference in means between ga_ultra and ga_est.

```
hist(diff,breaks=15)
```




```
qqnorm(diff)
```



c) After taking into account any difference in the means (whether or not statistically significant), do the shapes of the distributions of the two gestational age variables differ?

```
mean(diff,na.rm=T)
```

```
## [1] 0.3180388
```

The mean of the differences is .318 weeks Since the means differ by .318 weeks and ga_ultra has a larger sample mean, we will add .318 to every observation of ga_est to account for the difference between the means.

```
estc=est+.318
```

One way to test whether there are other differences between the distributions after eliminating the difference between the means is to use the KS test. The KS test in this situation assumes independence of the two samples, but here the two measures are used for each infant and so the assumption of independence is violated. As we don't have a test that does not make the independence assumption we will use the KS test even though it is not ideal.

$H_0 : F_1(y) = F_2(y)$ for all y (the two distributions have the same shape)

$H_A : F_1(y) \neq F_2(y)$ for at least one y (distributions do not have the same shape)

F_1 cdf of ga_ultra

F_2 cdf of ga_est corrected

```
ks.test(estc,ult)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
```

```
##
## data:  estc and ult
## D = 0.13851, p-value = 0.00206
## alternative hypothesis: two-sided
```

p-value=.002<.05 thus reject H_0

We reject the hypothesis that the two distributions have the same shape. Keeping in mind we have violated the independence assumption, we must be cautious when drawing conclusions from this test.

d) Classify both versions of gestational age into 3 intervals:

$(0, 37)$, $[37, 40)$, and $[40, \infty)$

low= (0,37) med= [37,40) high = [4,Inf)

```
ga=mot2%>%select(ga_ultra,ga_est)
ga$ga_est=cut(ga$ga_est,c(0,37,40,Inf),right=F,labels=c("low","med","high"))
ga$ga_ultra=cut(ga$ga_ultra,c(0,37,40,Inf),right=F,labels=c("low","med","high"))
```

Determine how well the two versions agree and provide a 95% confidence interval for the true agreement

```
library(vcd)
```

Making a contingency table

```
gatab=table(ga)
gatab2=addmargins(gatab)
gatab2
```

```
##          ga_est
## ga_ultra low med high Sum
##     low   33   6    0  39
##     med    6 170   30 206
##     high   0  22   91 113
##     Sum   39 198  121 358
```

total=358 (1 value missing)

Observed proportion of agreement= pa

pa= (33+170+91)/358 =.82123 sum of diagonal divided by total

Expected proportion of agreement = pc

pc= $E_{11}+E_{22}+E_{33}=.4368

$$\kappa = \frac{pa - pc}{1 - pc} = .68258$$

```
(k=Kappa(gatab))
```

```
##             value     ASE     z   Pr(>|z|)
## Unweighted 0.6826 0.03663 18.64  1.619e-77
## Weighted   0.7182 0.03348 21.45 4.593e-102
```

```
confint(Kappa(gatab))
```

```
##
## Kappa              lwr       upr
##   Unweighted 0.6107953 0.7543650
```

```
##   Weighted   0.6525813 0.7838333
```

Looking at the simple kappa, not the weighted.

```
k$Unweighted
```

```
##     value       ASE
## 0.6825801 0.0366256
```

The chance-corrected measure of agreement is $\kappa = 0.68$

95% CI for κ : (.611,.754)

$\kappa = .68$ indicates moderate agreement

e) Is the number of women randomized in each month consistent with the number of days in each month?

Conducting a χ^2 test of goodness of fit to determine if the number of women randomized each month is consistent with the number of days in each month

```
days=c(31,28,31,30,31,30,31,31,30,31,30,31) #create a vector of the number of days in each month
wtab=table(mot2$rand_month) #make table of women randomized per month
```

Format for chisq GOF test is: chisq.test(frequencys, p=null.probs) in our case, frequencys is the number of women randomized each month and null probability is days/365

Rejecting the null implies the model does not provide an adequate fit to the data

$$H_0 : \pi_{\text{month i}} = \frac{\text{days in month i}}{365}, i = 1, 2, ..., 12$$

The null hypothesis is that the proportion of women randomized in any given month is equal to the number of days in the month divided by 365)

H_A: at least one of the equalities is false

$\alpha = .05$

```
chisq.test(wtab, p=days/365,correct = F)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  wtab
## X-squared = 33.434, df = 11, p-value = 0.0004474
```

p-value $= .0004 < \alpha$ Thus reject H_0, conclude that the proportion of women randomized each month is not consistent with the number of days in the month

f) Without doing any additional tests, comment on how the distribution of the number of births each month compares with that of the number of women randomized each month.

```
btab=table(mot2$birth_month)
a=as.data.frame(wtab)
colnames(a)=c("month","randomized")
b=as.data.frame(btab)
colnames(b)=c("month","births")
br=left_join(a,b)
month=br$month
randpct=prop.table(br$randomized)
birthpct=prop.table(br$births)
```

```
dat=cbind(month,randpct,birthpct)
dat #creating dataframe of percentage of births and women randomized by month
```

```
##       month    randpct   birthpct
##  [1,]     1 0.06162465 0.06128134
##  [2,]     2 0.04761905 0.06406685
##  [3,]     3 0.11764706 0.08635097
##  [4,]     4 0.06722689 0.07520891
##  [5,]     5 0.11484594 0.07520891
##  [6,]     6 0.09803922 0.09749304
##  [7,]     7 0.11204482 0.09470752
##  [8,]     8 0.06162465 0.09749304
##  [9,]     9 0.08963585 0.08635097
## [10,]    10 0.10924370 0.11420613
## [11,]    11 0.08123249 0.07242340
## [12,]    12 0.03921569 0.07520891
```

The table gives percentages of women randomized by month and live births by month. The percentage of birth each month is much more even than that of the number of women randomized.

Dichotomize ga_ultra by <37 weeks to define preterm delivery

preterm = 0 not preterm delivery

preterm = 1 preterm delivery

```
mot3=mot2%>%mutate(preterm=as.numeric(ga_ultra<37))
```

g) Does the risk of preterm delivery vary monotonically with the number of previous pregnancies?

Does the probability of disease vary monotically with the exposure level? Chi-square test for trend

prop.trend.test(x=number of events,n=number of trials,score=group number)

In our case we have a disease exposure model

x= number of diease+

n = column total

score = exposure level

Conducting a chi square test for trend

Let p_i denote the probability preterm=1 in ppnum category i. i= 0,1,2,3+

$H_0 : p_0 = p_1 = p_2 = p_3$

$H_A : p_0 \leq p_1 \leq p_2 \leq p_3$ or $p_0 \geq p_1 \geq p_2 \geq p_3$ with at least one of the inequalities being strict

```
tab1=table(mot3$preterm,mot3$ppnum) #Create 2x2 Disease exposure table
tab2=addmargins(tab1) #adding margins to table
tab2
```

```
##
##          0   1   2  3+ Sum
##    0    57 117  72  28 274
##    1    10  10   9   5  34
##    Sum  67 127  81  33 308
```

```r
x=c(10,10,9,5) #number of diease+
n=c(67,127,81,33) #column total
score=c(1,2,3,4) # exposure level(rank order)
prop.trend.test(x,n,score)
```

```
##
##  Chi-squared Test for Trend in Proportions
##
## data:  x out of n ,
##  using scores: 1 2 3 4
## X-squared = 0.0011197, df = 1, p-value = 0.9733
```

p-value=.9733 $>$ .05 So we do not reject H_0 and conclude there isn't much evidence for a monotonic trend in the risk of preterm delivery with the number of previous pregnancies.

h) Based on this study, is treating periodontal disease in pregnant women effective in terms of reducing the risk of prematurity?

chi square test of no association also called chi square test for independence

we want to test whether there is an association (dependence) between treatment group and preterm delivery

H_0: preterm delivery is independent of treatment group

H_A: preterm delivery is associated with treatment group

```r
group=mot3$group
preterm=mot3$preterm
gptab=table(preterm,group) #creating 2x2 table
gptab
```

```
##        group
## preterm   1   2
##       0 148 171
##       1  23  16
```

```r
chisq.test(gptab,correct = F) #chisquare test without continuity correction
```

```
##
##  Pearson's Chi-squared test
##
## data:  gptab
## X-squared = 2.204, df = 1, p-value = 0.1376
```

p-value = .1376 $<$ .05 thus we fail to reject the null hypothesis and there is not enough evidence of an association between treatment group and preterm delivery

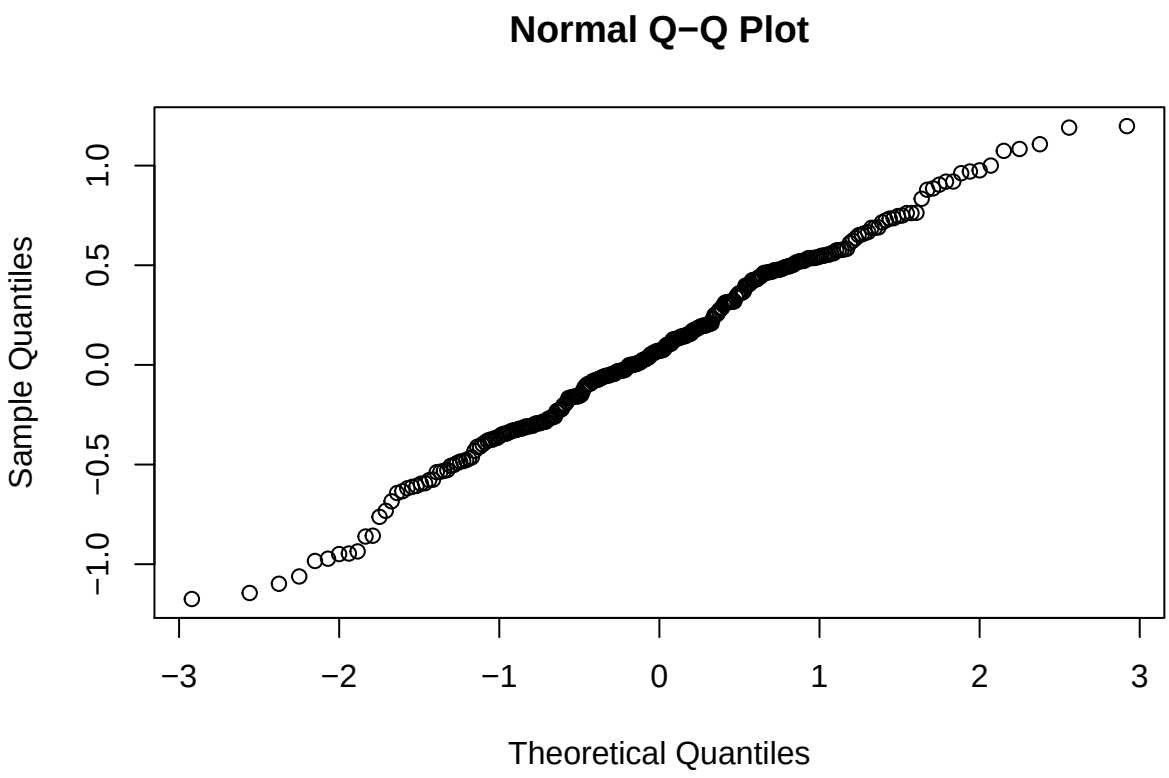
So there is no evidence that treatment of periodontal disease reduces the risk of preterm delivery.

i) Ignoring treatment group, is there a difference between the mean average pocket depth at randomization and the mean average pocket depth after delivery?

pd_pre and pd_post measurements on each woman are not independent thus we cannot use a two-sample test

we will create a vector of differences between the two measurements (remove NA values) and conduct a one sample test of difference in means

```
pre=mot3$pd_pre
post=mot3$pd_post
pdiff=post-pre #create vector of differences
pdiff=na.omit(pdiff) #removing na values
qqnorm(pdiff) #qqplot of the difference vector
```



Looking at the qqplot of the differnces, the data appears approximately normally distributed, with n=286. Since we have a large sample, we can use CLT/Slutsky's to give a test using the Z-statistic.

Taking the difference as the pd_post-pd_pre, where pd_pre is the pocket depth at baseline, the difference will be positive if pocket depth has increased (that is, periodontal disease has progressed) and negative if it has declined (that is, if there has been an improvement).

Let Y be the difference vector

```
ybar=mean(pdiff)
sdy=sd(pdiff)
n=286
z=ybar/(sdy/sqrt(n))
z
```

```
## [1] 3.029479
```

The mean of the differences is 0.08351273. Since this is positive, this suggests and that there has been an increase in pocket depth from baseline to delivery (peridontal disease has increased)

Want to test difference in means between pd_pre and pd_post using a Z-test.

$H_0 : \mu_{diff} = 0$ (no difference in means)

$H_A : \mu_{diff} \neq 0$ (there is a difference in means)

$Z = \dfrac{\bar{Y} - \mu_0}{s/\sqrt{n}} \sim N(0,1)$ (approximately normal)

$z_{stat} = \dfrac{\bar{y} - 0}{sd(y)\sqrt{286}} = 3.029479 \approx 3.03$

$C_{.05} = \{z : |z| > 1.96\}$

$|3.03| > 1.96$ Thus reject H_0, there is evidence that the means are different.

Conclude there is a difference in means between pd_pre and pd_post.

The estimate of the mean of the differences is 0.08351273. Since this is positive, this suggests and that there has been an increase in pocket depth from baseline to delivery (peridontal disease has increased)

j) Did the mean change in average pocket depth differ between the two treatment groups?

group1 is the prenatal treatment group

group2 is the postpartum treatment group

The two treatment groups are independent thus we have two independent sets of measurements. Create a vector of differences, pd_post - pd_pre, for each group and conduct a two-sample test.

```
pdiff2=mot3%>%mutate(diff=pd_post-pd_pre)%>%select(group,diff) #creating difference vector
pdiff2=na.omit(pdiff2) #removing NA values
g1=pdiff2%>%filter(group==1) #separting the two groups
g2=pdiff2%>%filter(group==2)
g1=g1$diff
g2=g2$diff
```

Z-test (using CLT/Slutsky's) for difference in means.

Now we do have two independent sets of measurements – the data from the two treatment groups (with the data within each group being the difference between the pocket depth measurements at the two time points, as in part (e)). The sample sizes in the two groups are still large (n 1 = 61 in the prenatal group and n 2 = 62 in the post-partum group), so we can again rely on the CLT and Slutsky's Theorem to give a test using the Z statistic.

```
y1=mean(g1)
n1=length(g1)
s1=sd(g1)
y2=mean(g2)
n2=length(g2)
s2=sd(g2)
zstat=(y1-y2)/sqrt(s1^2/n1+s2^2/n2)
pvalue=pnorm(-3.26)*2
paste("z stat =", round(zstat,digits = 3), "p-value =",round(pvalue,digits =4))
```

```
## [1] "z stat = -3.261 p-value = 0.0011"
```

$H_0 : \mu_{diff1} = \mu_{diff2}$ (the mean of the differences are the same for both treatment groups)

$H_A : \mu_{diff1} \neq \mu_{diff2}$ (the mean of the differences are not the same for both treatment groups)

$Z = \dfrac{(\bar{Y_1} - \bar{Y_2}) - 0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} = -3.26$

$C_{.05} = \{z : |z| > 1.96\}$

$|-3.26| > 1.96$ Thus reject H_0 and conclude there is a difference between the change in pocket depth in the prenatal treatment group compared with the post-partum treatment group.

We could also do a t-test

(make sure to use var.equal since they came from the same sample distribution this is a pooled t-test)

```
t.test(g1,g2,var.equal = T,alternative="two.sided")
```

```
##
##  Two Sample t-test
##
## data:  g1 and g2
## t = -3.2954, df = 284, p-value = 0.001108
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.28616062 -0.07214217
## sample estimates:
##   mean of x   mean of y
## -0.01295341  0.16619799
```

p-value $= .0011 < .05$ Thus reject H_0

Problem 3

First assume this was an unmatched case-control study

```
name=c("id","case","exposed","agegroup")
cc=fread("data/Midterm_CC.dat")
names(cc)=name
cc$case=factor(cc$case)
cc$exposed=factor(cc$exposed)
cc$agegroup=factor(cc$agegroup)
```

a) Determine whether premature birth case status is associated with being exposed to periodontal disease

First create a 2x2 contingency table with margins

```
exposed=cc$exposed
case=cc$case
taba=table(exposed,case)
taba1=addmargins(taba)
taba1
```

```
##        case
## exposed   0   1 Sum
##     0    69  51 120
##     1    18  36  54
##     Sum  87  87 174
```

```
etab=table(case,exposed)[2:1,2:1]
etab2=addmargins(etab)
etab2
```

```
##      exposed
## case    1   0 Sum
```

```
##   1    36  51  87
##   0    18  69  87
##   Sum  54 120 174
```

We have a large sample, n=174

We are comparing two proportions, proportion exposed in control group and proportion exposed in case group.

π_1 = prob of being exposed in the control group

π_2 = prob of being exposed in the case group

We want to conduct a large sample test comparing two proportions

We will conduct a chi square test of association

$H_0 : \pi_1 = \pi_2$ (probabilities of exposure are equal)

$H_A : \pi_1 \neq \pi_2$ (probabilities of exposure are not equal)

$\alpha = .05$

$X^2 = \frac{(N)(n_{11}n_{22} - n_{12}n_{21})^2}{n_1 n_2 m_1 m_2} \sim \chi^2_1$

$C_\alpha = \{x^2 : x^2 \geq \chi^2_{1,1-\alpha}\}$

```
x2=(174*((36*69)-(18*51))^2)/(54*120*87*87)
chicrit=qchisq(p=.95,df=1)
paste("Test stat=",x2, "critical value=", round(chicrit,digits = 3))
```

```
## [1] "Test stat= 8.7 critical value= 3.841"
```

$8.7 \geq 3.841$ Thus reject H_0 and conclude conclude that moderate to severe periodontal disease is associated with premature birth.

Running the test using chisq.test (make sure correct is False) also table needs to be in epid format. In our case our exposure is really our disease and vice versa.

```
chisq.test(etab,correct=F) #correct=F turns off continuity correction, which we dont want in this case
```

```
##
##  Pearson's Chi-squared test
##
## data:  etab
## X-squared = 8.7, df = 1, p-value = 0.003182
```

p-value= $.003 < .05$ Thus reject H_0

b) Provide an estimate for a measure of the association between exposure and case status and give a 95% confidence interval for the true measure.

Since we have a case-control study, the appropriate measure of association is the odds ratio.

$\hat{OR} = \frac{n_{11}n_{22}}{n_{21}n_{12}} = \frac{36 * 69}{18 * 51} = 2.705 \approx 2.71$

```
OR=(36*69)/(18*51)
OR
```

```
## [1] 2.705882
```

When using the package epitools i.e. oddsratio.wald(table) Rows should be the exposures, columns the case status Unexposed controls should be in top left cell

```r
library(epitools)
```

```r
oddsratio.wald(taba)
```

```
## $data
##        case
## exposed  0  1 Total
##   0     69 51   120
##   1     18 36    54
##   Total 87 87   174
##
## $measure
##        odds ratio with 95% C.I.
## exposed estimate    lower    upper
##       0 1.000000       NA       NA
##       1 2.705882 1.382336 5.296684
##
## $p.value
##        two-sided
## exposed  midp.exact fisher.exact  chi.square
##       0          NA           NA          NA
##       1 0.003401645  0.005090619 0.003182101
##
## $correction
## [1] FALSE
##
## attr(,"method")
## [1] "Unconditional MLE & normal approximation (Wald) CI"
```

$\hat{OR} = 2.71$

95% CI=$(1.38, 5.30)$

c) Repeat part (b) above, taking age group into account.

Using the Mantel-Haenszel Test to obtain an age adjusted OR estimate and a 95% CI

```r
age=cc$agegroup
tabage=etab=table(case,exposed,age) #creating an array of 2x2 tables stratfied by age group
mantelhaen.test(tabage) #running the test on the array
```

```
##
##  Mantel-Haenszel chi-squared test with continuity correction
##
## data:  tabage
## Mantel-Haenszel X-squared = 8.4597, df = 1, p-value = 0.003631
## alternative hypothesis: true common odds ratio is not equal to 1
## 95 percent confidence interval:
##  1.481862 6.280179
## sample estimates:
## common odds ratio
##          3.050633
```

age adjusted $\hat{OR}$=3.05 95% CI=$(1.48, 6.28)$

d) Does age group appear to be a confounder? Is the pooled estimate in part (c) a reasonable way to summarize the association here?

The adjusted odds ratio of 3.05 is reasonably similar to the unadjusted one of 2.71, so age group does not appear to be a substantial confounder of the association between periodontal disease and premature birth.

The number of cases and controls are equal within each age group. This is because cases and controls were matched on age and number of previous pregnancies, thus age and case status are not associated.

```
oddsratio.wald(tabage[,,1])
```

```
## $data
##        exposed
## case     0  1 Total
##   0     30 16    46
##   1     21 25    46
##   Total 51 41    92
##
## $measure
##     odds ratio with 95% C.I.
## case estimate     lower    upper
##    0 1.000000        NA       NA
##    1 2.232143 0.9641419 5.167768
##
## $p.value
##     two-sided
## case midp.exact fisher.exact chi.square
##    0         NA           NA         NA
##    1 0.06407049   0.09279014 0.05905081
##
## $correction
## [1] FALSE
##
## attr(,"method")
## [1] "Unconditional MLE & normal approximation (Wald) CI"
```

```
oddsratio.wald(tabage[,,2])
```

```
## $data
##        exposed
## case     0 1 Total
##   0     17 1    18
##   1     14 4    18
##   Total 31 5    36
##
## $measure
##     odds ratio with 95% C.I.
## case estimate     lower    upper
##    0 1.000000        NA       NA
##    1 4.857143 0.4856844 48.57442
##
## $p.value
##     two-sided
## case midp.exact fisher.exact chi.square
##    0         NA           NA         NA
##    1  0.1915584    0.3376623  0.1482348
```

```
##
## $correction
## [1] FALSE
##
## attr(,"method")
## [1] "Unconditional MLE & normal approximation (Wald) CI"
```

```
oddsratio.wald(tabage[,,3])
```

```
## $data
##        exposed
## case     0 1 Total
##   0     22 1    23
##   1     16 7    23
##   Total 38 8    46
##
## $measure
##     odds ratio with 95% C.I.
## case estimate    lower    upper
##    0    1.000       NA       NA
##    1    9.625 1.075027 86.17513
##
## $p.value
##     two-sided
## case midp.exact fisher.exact chi.square
##    0         NA           NA         NA
##    1  0.0253676   0.04697704 0.01959783
##
## $correction
## [1] FALSE
##
## attr(,"method")
## [1] "Unconditional MLE & normal approximation (Wald) CI"
```

From the 3 separate 2x2 tables above (one for each age group), we obtain estimated odds ratios of 2.23, 4.86 and 9.63, respectively. This suggests that the odds ratios are not homogeneous across the age groups and so it may not be appropriate to pool them using the Mantel-Haenszel estimator

Part ii)

e) Repeat parts (a) and (b) above assuming an individually-matched case-control design

Matching on age and ppnum, one control per case

The 2x2 table has to be in a different form, we need each pair to be a single observation

To do this:

1) rename the exposure variables so that those for cases and controls are distinct
2) split the dataset into two, one consisting of cases, the other of controls
3) merge the two on the part of the ID that is common to the members of a pair

Creating a matched pair variable, the 3rd and 4th character of ID uniquely identify a matched pair.

```
cc1=cc%>% mutate(mp=str_sub(id,3,4))%>%select(case,exposed,mp)#mp is matched pair variable
cases=cc1%>%filter(case==1)%>%rename(ecase=exposed)
cases=as_tibble(cases)
```

```
controls=cc1%>%filter(case==0)%>%rename(econtrol=exposed)
controls=as_tibble(controls)
cc2=full_join(cases,controls,by='mp')%>%select(c(mp,ecase,econtrol)) #joining the separated tables by m
```

Creating a 2x2 table

```
ecase=cc2$ecase
econtrol=cc2$econtrol
mptab=table(econtrol,ecase)[2:1,2:1]
mptab1=addmargins(mptab)
mptab1
```

```
##         ecase
## econtrol  1  0 Sum
##      1   11  7  18
##      0   25 44  69
##      Sum 36 51  87
```

Determine where birth case status is associated with being exposed to periodonal disease

Since we have matched pairs we will use McNemar's test statistic (similar to a one sample binomial test)

Keep in mind that this is a test of association with a risk factor, not a test for agreement between the members of a pair

n12+n21= 7+25=32>30 so we can use chi square approximation

$H_0 : \pi_1 = \pi_2$

$H_A : \pi_1 \neq \pi_2$

$M = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}} \sim \chi^2_1$

$C_\alpha = \{M : M > \chi^2_{1,1-\alpha}\}$

```
#takes a 2x2 table (in epid form with controls as rows and cases as columns)
#computes the mcnemar test stat
mstat=function(table){
 t2=table[2]
 t3=table[3]
 m=(t2-t3)^2/(t3+t2)
 m
}
```

```
m=mstat(mptab)
crit=qchisq(.95,1)
paste("test stat is",m, "critical value is",round(crit,digits=3))
```

```
## [1] "test stat is 10.125 critical value is 3.841"
```

$10.125 > 3.841$ Thus reject H_0

```
mcnemar.test(mptab,correct=F) #make sure correct=False so we dont have the continuity correction
```

```
##
##  McNemar's Chi-squared test
##
## data:  mptab
## McNemar's chi-squared = 10.125, df = 1, p-value = 0.001463
```

p-value= .001463 $< .05$ thus reject H_0 and conclude that moderate to severe periodontal disease is associated with premature birth.

Provide an estimate for the odds ratio between exposure and cases status and give 95% CI

```
orm= 25/7
v=1/7+1/25
c1=exp(log(orm)-1.96*sqrt(v))
c2=exp(log(orm)+1.96*sqrt(v))
ci=c(c1,c2)
paste("OR =", round(orm,digits = 2))
```

```
## [1] "OR = 3.57"
```

```
ci
```

```
## [1] 1.544707 8.257294
```

$\hat{OR}_M = n_{21}/n_{12} = 25/7 \approx 3.57$

$\hat{Var}(\ln(\hat{OR}_M)) \approx 1/n_{12} + 1/n_{21} \approx .183$

95% CI $exp(ln(\hat{OR}_M) \pm 1.96\sqrt{\hat{Var}(\ln(\hat{OR}_M))})$

95% CI $(1.54, 8.26)$

f) Which of the estimates of the measure of association in (b), (c) and (e) is most appropriate? Justify your choice

The estimate in part (e) is most appropriate because it takes into account the matched case-control design.

(part g is in solution pdf)