Homework3_Hwang

Charles Hwang

2/18/2022

Charles Hwang

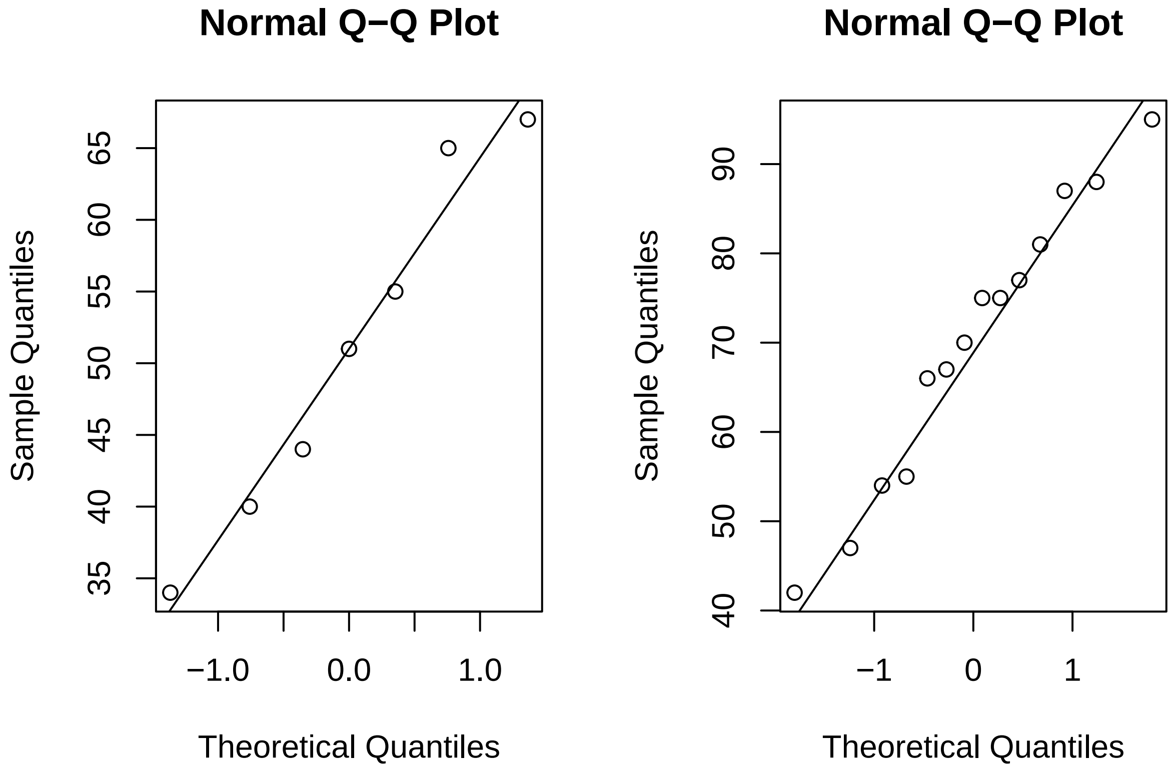
Dr. Perry

STAT 451-001

18 February 2022

Problem 1

```
rm(list=ls())
pr<-c(67,55,51,40,65,34,44)
nr<-c(95,87,47,55,70,42,75,88,67,54,66,75,77,81)
par(mfrow=c(1,2))                                                                     # Problem 1(a)
qqnorm(pr)
qqline(pr)
qqnorm(nr)
qqline(nr)
```



```
# It appears both datasets are approximately normal.
shapiro.test(pr)                                                                # Problem 1(b)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  pr
## W = 0.95109, p-value = 0.7396
```

```
# We fail to reject H0 at the alpha = 0.05 level. There is insufficient
# evidence (p = 0.7396) that the scores of "poor readers" is not normally distributed.
shapiro.test(nr)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  nr
## W = 0.96725, p-value = 0.8379
```

```
# We fail to reject H0 at the alpha = 0.05 level. There is insufficient
# evidence (p = 0.8379) that the scores of "normal readers" is not normally distributed.
# H0: mu_(nr) - mu_(pr) = 0                                                     # Problem 1(c)
# HA: mu_(nr) - mu_(pr) =/= 0
t.test(nr,pr,alternative="two.sided",conf.level=0.99)
```

```
##
##  Welch Two Sample t-test
##
## data:  nr and pr
## t = 3.0127, df = 15.135, p-value = 0.008674
## alternative hypothesis: true difference in means is not equal to 0
## 99 percent confidence interval:
##   0.4412999 37.7015573
## sample estimates:
## mean of x mean of y
##  69.92857  50.85714
```

```
# We reject H0 at the alpha = 0.01 level. There is sufficient evidence (p = 0.008674) that
# the mean score of "normal readers" is different than the mean score of "poor readers".
# H0: m_(nr) - m_(pr) = 0                                                       # Problem 1(d)
# HA: m_(nr) - m_(pr) =/= 0
wilcox.test(nr,pr,exact=TRUE)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  nr and pr
## W = 82, p-value = 0.01522
## alternative hypothesis: true location shift is not equal to 0
```

```
# We fail to reject H0 at the alpha = 0.01 level. There is insufficient evidence (p = 0.01522)
# that the median score of "normal readers" is different than the median score of "poor readers".
# I believe the t-test is better. It tends to be more robust                    # Problem 1(e)
# and exact and we are able to use it because the normality assumption is satisfied.
sqrt(3*length(pr)*length(nr)/pi/(length(pr)+length(nr)+1))*(mean(nr)-mean(pr))/14-qnorm(1-.01) #(f)
```

```
## [1] 0.4832397
```

$\Delta = \mu_{nr} - \mu_{pr} =$ 69.9285714 - 50.8571429 $= 19.0714286$

$\sigma = 14$

$m = 7$

$n = 14$

$N = m + n = 7 + 14 = 21$

$\alpha = 0.01$

$z_{1-\alpha} = z_{1-0.01} = z_{0.99} = 2.3263479$

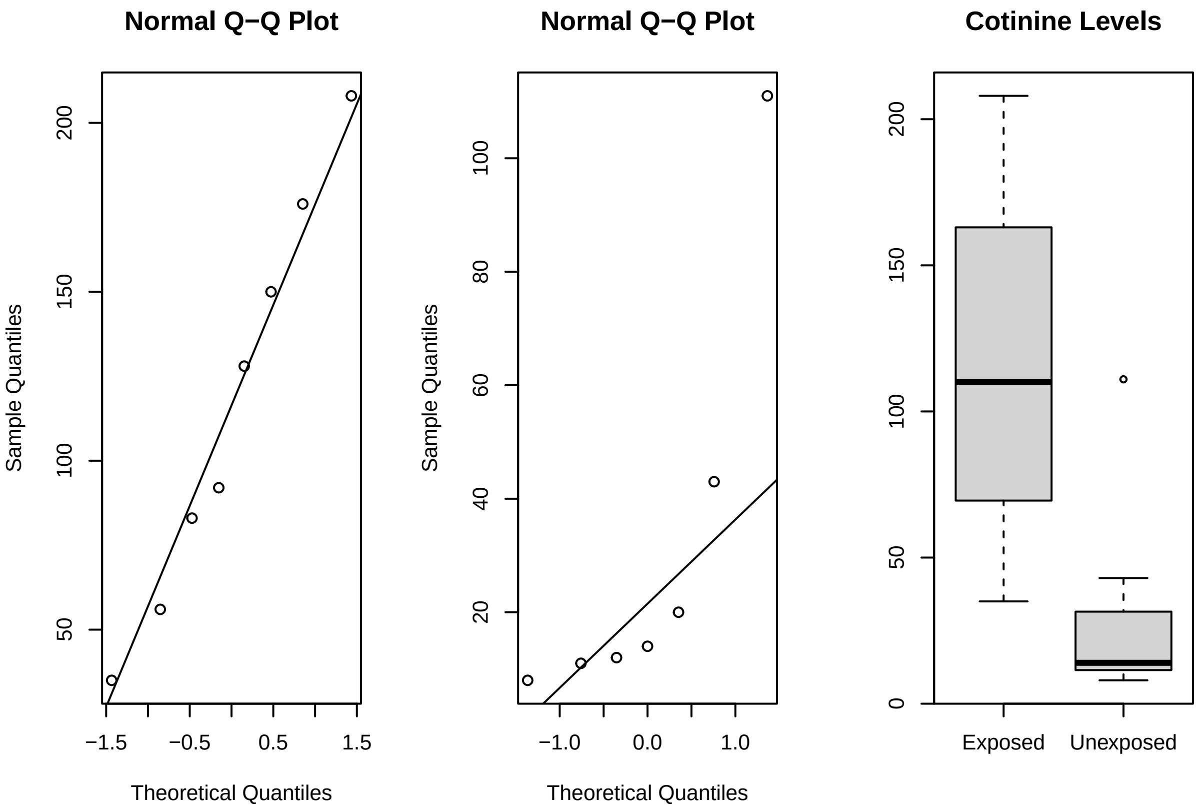
$A_{normal} = \frac{\Delta}{\sigma}\sqrt{\frac{3mn}{(N+1)\pi}} - z_{1-\alpha} = \frac{19.07143}{14}\sqrt{\frac{3(7)(14)}{((21)+1)\pi}} - z_{1-0.01} \approx \frac{19.0714286}{14}\sqrt{\frac{294}{22\pi}} - 2.3263479 = 0.4832397$

Problem 2

```
e<-c(35,56,83,92,128,150,176,208)
u<-c(8,11,12,14,20,43,111)
# H0: mu_e - mu_u = 25                                  # Problem 2(a)
# HA: mu_e - mu_u > 25
t.test(e,u,mu=25,alternative="greater") # Set mu to 25 and alternative to "greater"
```

```
##
##  Welch Two Sample t-test
##
## data:  e and u
## t = 2.3493, df = 11.817, p-value = 0.01853
## alternative hypothesis: true difference in means is greater than 25
## 95 percent confidence interval:
##  39.35281      Inf
## sample estimates:
## mean of x mean of y
## 116.00000  31.28571
```

```
# We reject H0 at the alpha = 0.05 level. There is sufficient evidence (p = 0.01853) that
# the mean cotinine concentration in urine of exposed infants exceeds the mean cotinine
# concentration in urine of unexposed infants by at least 25 units.
par(mfrow=c(1,3))                                       # Problem 2(b)
qqnorm(e)
qqline(e)
qqnorm(u)
qqline(u)
boxplot(e,u,names=c("Exposed","Unexposed"),main="Cotinine Levels")
```



```
# We can see the data for unexposed infants clearly has an outlier and is not normally
# distributed. This violates the normality assumption for the two-sample t-test.
# H0: T_e - T_u = 25                                         # Problem 2(c)
# HA: T_e - T_u > 25
c<-data.frame(c(e,u),c(rep("Exposed",length(e)),rep("Unexposed",length(u))))
names(c)<-c("Cotinine","Smoking")
P<-matrix(NA,nrow=length(c(e,u)),ncol=5000)
T<-rep(0,5000)
set.seed(1802,sample.kind="Rounding")
for(i in 1:5000){P[,i]<-sample(c$Cotinine,size=length(c(e,u)),replace=FALSE)
T[i]<-abs(mean(P[c$Smoking=="Exposed",i])-mean(P[c$Smoking=="Unexposed",i]))}
mean(T+25>mean(e)-mean(u)) # Add 25 to sample data
```

```
## [1] 0.0762
```

```
# We fail to reject H0 at the alpha = 0.05 level. There is insufficient
# evidence (p = 0.0762) that the cotinine concentration in urine of exposed infants exceeds
# the cotinine concentration in urine of unexposed infants by at least 25 units.
# H0: m_e - m_u = 25                                         # Problem 2(d)
# HA: m_e - m_u > 25
wilcox.test(e,u,mu=25,alternative="greater",exact=TRUE) # Set mu to 25
```

```
##
##  Wilcoxon rank sum exact test
##
## data:  e and u
## W = 45, p-value = 0.02704
## alternative hypothesis: true location shift is greater than 25
```

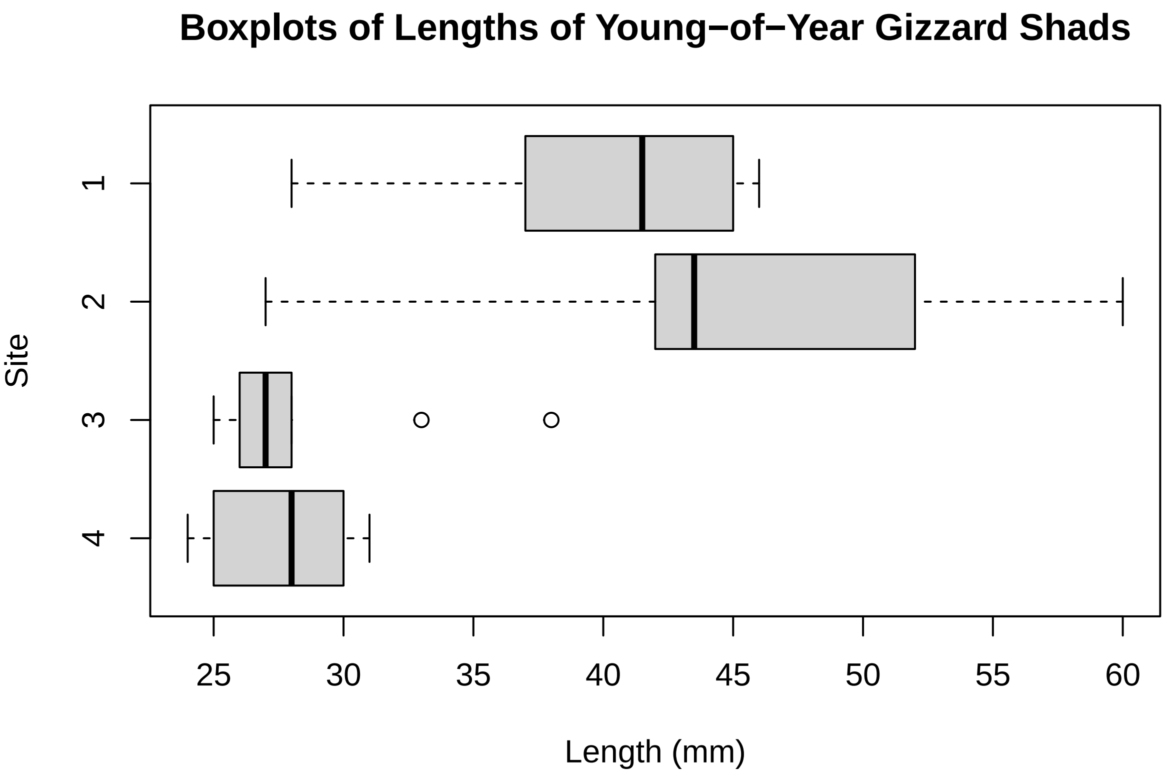
```
# We reject H0 at the alpha = 0.05 level. There is sufficient evidence (p = 0.02704)
# that the median cotinine concentration in urine of exposed infants exceeds the mean
# cotinine concentration in urine of unexposed infants by at least 25 units.
# H0: The scale parameters of "e" and "u" are the same  # Problem 2(e)
# HA: The scale parameters of "e" and "u" are different
ansari.test(e,u,exact=TRUE)
```

```
##
##  Ansari-Bradley test
##
## data:  e and u
## AB = 37, p-value = 0.5958
## alternative hypothesis: true ratio of scales is not equal to 1
```

We fail to reject H_0 at the $\alpha = 0.05$ level. There is insufficient evidence ($p = 0.5958042$) that the scale parameters of the distributions of cotinine concentrations in urine of exposed infants and unexposed infants are different.

Problem 3

```
s1<-c(46,28,46,37,32,41,42,45,38,44)
s2<-c(42,60,32,42,45,58,27,51,42,52)
s3<-c(38,33,26,25,28,28,26,27,27,27)
s4<-c(31,30,27,29,30,25,25,24,27,30)      # Problem 3(a)
boxplot(s4,s3,s2,s1,xlab="Length (mm)",ylab="Site",names=4:1,horizontal=TRUE,main="Boxplots of Lengths
```


Boxplots of Lengths of Young−of−Year Gizzard Shads

```
# The distributions of site 1 appears to be left skew, the distributions of sites 2 and
# 3 appear to be right skew, and the distribution of site 4 appears to be approximately
# normal. There are two visible outliers in the data for site 3.
l<-c(s1,s2,s3,s4)                             # Problem 3(b)
```

```
s<-rep(c("1","2","3","4"),each=length(s1))
# H0: mu_(s1) = mu_(s2) = mu_(s3) = mu_(s4)
# HA: At least one mu_(si) for i = 1, 2, 3, 4 is different
anova(lm(l~s))
```

```
## Analysis of Variance Table
##
## Response: l
##           Df Sum Sq Mean Sq F value    Pr(>F)
## s          3 2196.9  732.29  17.142 4.421e-07 ***
## Residuals 36 1537.9   42.72
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# We reject H0 at the alpha = 0.05 level. There is sufficient evidence (p < 0.000001)
# that at least one of the sites' mean lengths is different.
# H0: T_(s1) = T_(s2) = T_(s3) = T_(s4)  # Problem 3(c)
# HA: At least one T_(si) for i = 1, 2, 3, 4 is different
Y<-data.frame(l,s)
F<-rep(NA,20000)
set.seed(1802)
for (i in 1:20000){Y$l=l[sample(1:length(l),length(l))]
F[i]=anova(lm(l~s,data=Y))["s","F value"]}
mean(F>anova(lm(l~s))["s","F value"])
```

```
## [1] 0
```

```
# We reject H0 at the alpha = 0.05 level. There is sufficient evidence (p < 0.0001) that
# at least one of the sites' distributions is different.
# H0: m_(s1) = m_(s2) = m_(s3) = m_(s4)  # Problem 3(d)
# HA: At least one m_(si) for i = 1, 2, 3, 4 is different
kruskal.test(l~s)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  l by s
## Kruskal-Wallis chi-squared = 22.852, df = 3, p-value = 4.335e-05
```

We reject H_0 at the $\alpha = 0.05$ level. There is sufficient evidence ($p = 0.00004$) that at least one of the sites' median lengths is different.

Problem 3(e)

I believe the Kruskal-Wallis test is the most appropriate for this data. We saw the data were not all normal, which violated the normality assumption required for an ANOVA F-test. We also saw that because the F-value was so high ($F = 17.1418818$), there were no observations out of the 20,000 that were even close to being greater (the closest one being $F = 10.5450475$) and thus $p = \frac{0}{20000} = 0$. The Kruskal-Wallis test appeared to provide the most accurate p-value.