HW7 Nonparametric Statistics

Fan Heng fh2294
Columbia University

13 April, 2014

Question 1

(a)

The independent hypothesis is:

H_0: P(X=a,Y=b)=P(X=a)P(Y=b) a, b $\in -1, 0, 1 \quad vs \quad H_1 : P(X = a, Y = b) \neq P(X = a)P(Y = b)$ for some a and b.

(b)

The χ^2 statistic for this test is

$nQ_n = n \sum_{a,b \in (-1,0,1)} \frac{(P(X=a,Y=b)-P(X=a)P(Y=b))^2}{P(X=a)P(Y=b)}$

(c)

The χ^2 test in R is

```
XY <- array(c(1, 2, 4, 4, 3, 10, 10, 16, 50), c(3, 3), )
rownames(XY) <- c("X=-1", "X=0", "X=1")
colnames(XY) <- c("Y=-1", "Y=0", "Y=1")
XY
```

```
##      Y=-1 Y=0 Y=1
## X=-1    1   4  10
## X=0     2   3  16
## X=1     4  10  50
```

```
chisq.test(XY)
```

```
## Warning:  Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  XY
## X-squared = 1.442, df = 4, p-value = 0.8369
```

P value is 0.8369 which is very large, so we accept H_0.

Question 2

For $y_1 \leq y_2 \leq ... \leq y_n$
P$(Y_1 \leq y_1, Y_1 \leq y_1, ..., Y_n \leq y_n)$
=P$(F_x(X_{(1)}) \leq y_1, F_x(X_{(2)}) \leq y_2, ..., F_x(X_{(n)}) \leq y_n)$
=P$(X_{(1)} \leq F_x^{-1}(y_1), X_{(2)} \leq F_x^{-1}(y_2), ..., X_{(n)} \leq F_x^{-1}(y_n))$

Each X_i has the same probability to be chosen as $X_{(1)}, ...orX_{(n)}$. So there are n! permutation from $(X_1, X_2, ..., X_n)$ to $(X_{(1)}, X_{(2)}, ..., X_{(n))}$ with each having the same probability of $\frac{1}{n!}$. Therefore,

P$(X_{(1)} \leq F_x^{-1}(y_1), X_{(2)} \leq F_x^{-1}(y_2), ..., X_{(n)} \leq F_x^{-1}(y_n)) =$ n! P$(X_1 \leq F_x^{-1}(y_1), X_2 \leq F_x^{-1}(y_2), ..., X_n \leq F_x^{-1}(y_n))$
=n! P$(X_1 \leq F_x^{-1}(y_1))P(X_2 \leq F_x^{-1}(y_2)), ..., P(X_n \leq F_x^{-1}(y_n))$
=n! $F_x(F_x^{-1}(y_1))F_x(F_x^{-1}(y_2))$,...,$F_x(F_x^{-1}(y_n))$
=n! $y_1 y_2, .., y_n$

In other conditions that $y_1 \leq y_2 \leq ... \leq y_n$ doesn't hold, P$(Y_1 \leq y_1, Y_1 \leq y_1, ..., Y_n \leq y_n)$=0.

The joint probability density function of $Y_1, Y_2, ..., Y_n$ is

$f_{Y_1,Y_2,..,Y_n} = \frac{\partial^n P(Y_1 \leq y_1, Y_1 \leq y_1, ..., Y_n \leq y_n)}{\partial y_1 \partial y_2 ... \partial y_n}$
=n!

$f_{Y_1,Y_2,..,Y_n} =$

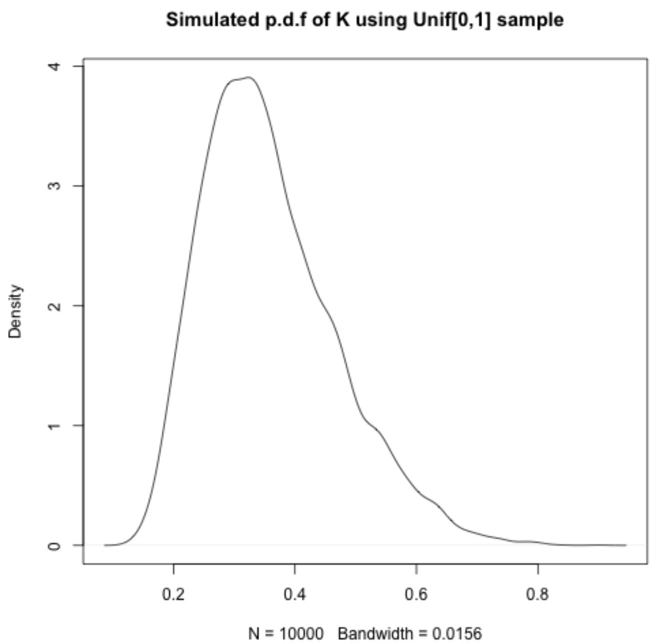
Question 3

Characterizing the distribution of K through Monte Carlo simulation.

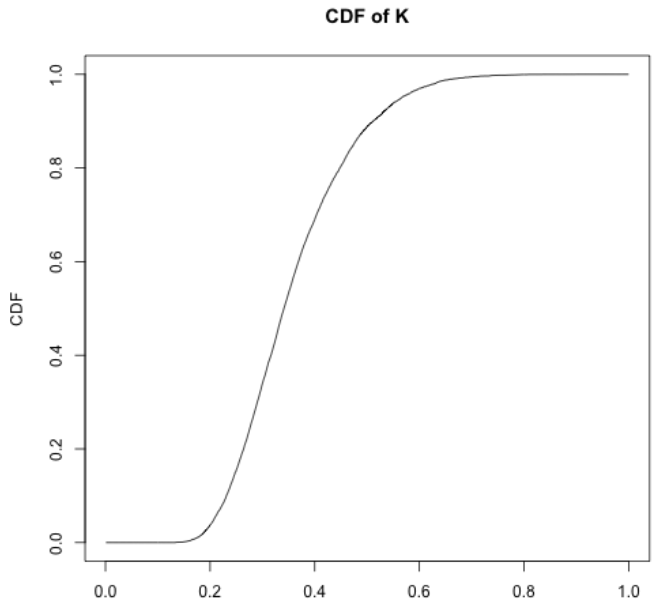
Since the distribution of K is free of F_0, we can generate sample data from any distribution. In this question, we generate data from uniform(0,1).

```
K <- rep(0, 10000)
Ind <- seq(1, 10000, 1)
for (i in 1:10000) {
    x <- runif(5, 0, 1)  # generate 5 data from Uniform(0,1)
    x.ord <- x[order(x)]

    K[i] <- max(abs((order(x.ord) - 1)/5 - x.ord), abs(order(x.ord)/5 - x.ord),
```

```r
        abs(x.ord[1]), abs(1 - x.ord[5]))
}

# the density function of K
plot(density(K), main = "Simulated p.d.f of K using Unif[0,1] sample")
```




```r
# the cumulative distribution function of K
quantileK <- seq(0.001, 1, 0.001)
cdfK <- rep(0, 1000)
for (i in 1:1000) {
    cdfK[i] <- sum(K <= quantileK[i])/10000
}
plot(quantileK, cdfK, xlab = "", ylab = "CDF", main = "CDF of K", type = "l")
```




```r
mean(K)

## [1] 0.3583

var(K)

## [1] 0.01206
```

The empirical distribution of K is N(0.427, 0.018).

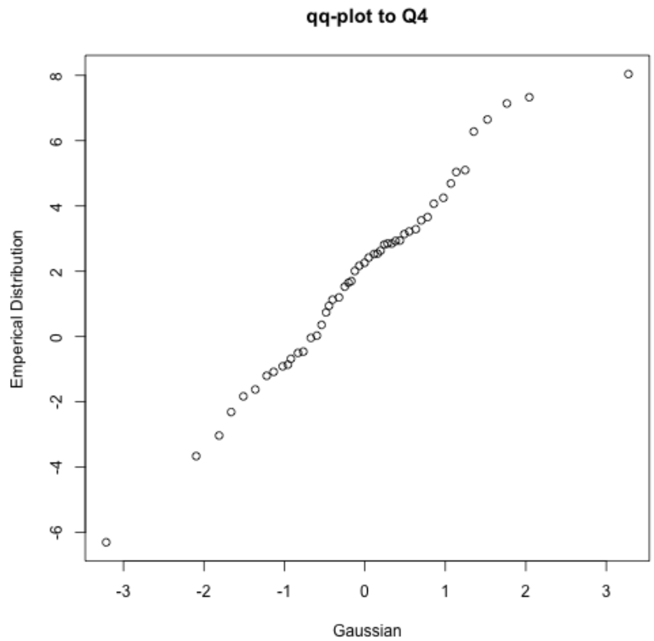
Question 4

```r
x <- c(-6.3, 2.94, 2.53, -0.86, 5.04, 3.22, -1.62, 3.56, 1.13, 2.63, -1.08,
    3.66, 4.07, -3.66, 0.74, 2.85, 2.85, 1.7, 1.53, 7.33, 2.82, -2.31, 0.94,
    -0.04, -1.2, 1.2, 5.1, 4.69, -0.46, 2.17, 2.01, 0.36, 3.14, 8.04, 7.14,
    2.54, -3.03, 4.25, -0.91, 1.65, 2.26, -1.83, -0.68, 6.28, 2.93, -0.5, 2.42,
    3.29, 0.03, 6.65)
```

(a)

Our assumption is that data $X_1, ..., X_{50}$ is drawm from a Gaussian distribution.

```r
# Generate a new data set y from Gaussian N(0,1)
yy <- rnorm(1000, 0, 1)
qq <- qqplot(yy, x, xlab = "Gaussian", ylab = "Emperical Distribution", main = "qq-plot to Q4")
```

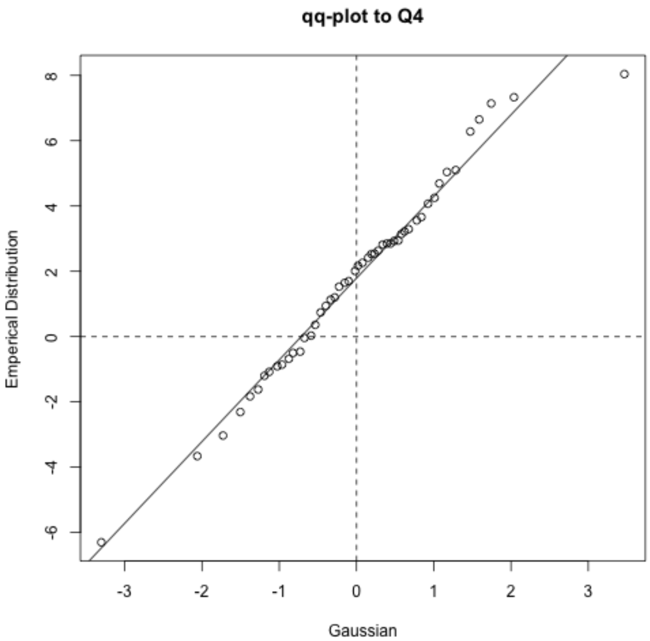


The qqplot seems like a straight line except a few points. It supports our assumption.

(b)

Fit a line to the qq-plot data and estimate mean and variance of distribution F.

```
yy <- rnorm(1000, 0, 1)
qq <- qqplot(yy, x, xlab = "Gaussian", ylab = "Emperical Distribution", main = "qq-plot to Q4")
qqfit <- lm(qq$y ~ qq$x)
abline(qqfit, lty = 1)
abline(v = 0, h = 0, lty = 2)
```




```
qqfit$coef

## (Intercept)        qq$x
##       1.790       2.502
```

QQ-plot shows the linear relationship of $t_0(a) \equiv F_0^{-1}(\alpha) and t(a) \equiv F^{-1}(\alpha). The linear equation is t(a) = \sigma t_0(a) + \mu where \mu is the mean of distribution F and \sigma is the variance.$

$From the fitted line to the qq-plot data, \hat{\sigma} = 2.637498, and \hat{\mu} = 1.823786.$

Question 5 , 6 on a separate page

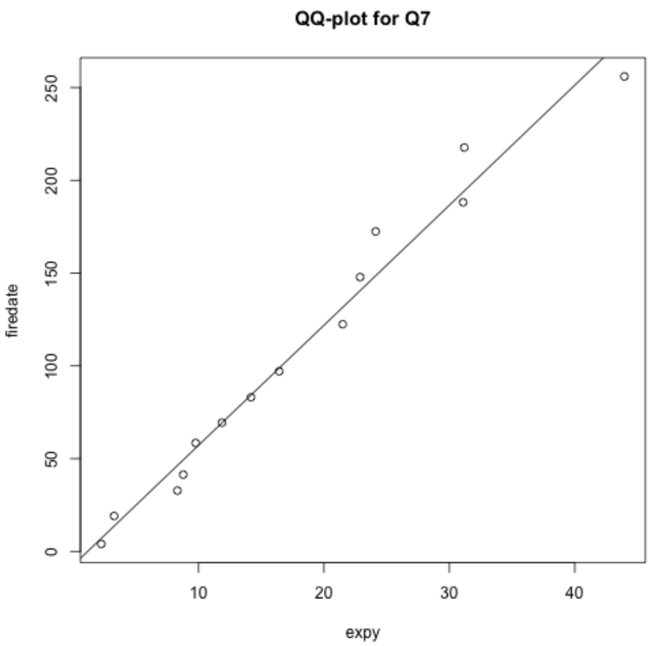
Question 7

The Null hypothesis for this question is that $H_0 : the time between the occurrence of fire in the reserve follows Exp(1/15) \quad v$
$the time between the occurence of the fire in the reserve does not follow Exp(1/15).$

(a)

```
firedate <- c(4, 18, 32, 37, 56, 64, 78, 89, 104, 134, 154, 178, 190, 220, 256)
fireinter <- firedate[-1] - firedate[-15]
# Generate dataset from Exp(1/15)
expy <- rexp(length(fireinter), rate = 1/15)
```

```r
# qqplot of the data
qqfire <- qqplot(expy, firedate, main = "QQ-plot for Q7")
abline(lm(qqfire$y ~ qqfire$x))
```



The qqplot doesn't seem to be a straight line so the clain is not justified.

(b)

Kolmogorov-Smirnov test

```r
# install.packages('exptest')
require(exptest)

## Loading required package:  exptest

ks.exp.test(fireinter)

##
##  Kolmogorov-Smirnov test for exponentiality
##
## data:  fireinter
## KSn = 0.3144, p-value = 0.0195
```

The P-value of Kolmogorov-Smirnov test is 0.0225 which is significant under 5% significant level. We reject the Null.

(c)

The Anderson-Darling test

```
# install.packages('ADGofTest')
require(ADGofTest)

## Loading required package:  ADGofTest

ad.test(fireinter, pexp)

##
##  Anderson-Darling GoF Test
##
## data:  fireinter  and  pexp
## AD = 170.4, p-value = 4.286e-05
## alternative hypothesis: NA
```

The P-value of Anderson-Darling test is 4.286e-05 which is significant under 5% significant level. We reject the Null.