Statistics 135 – Lab Project

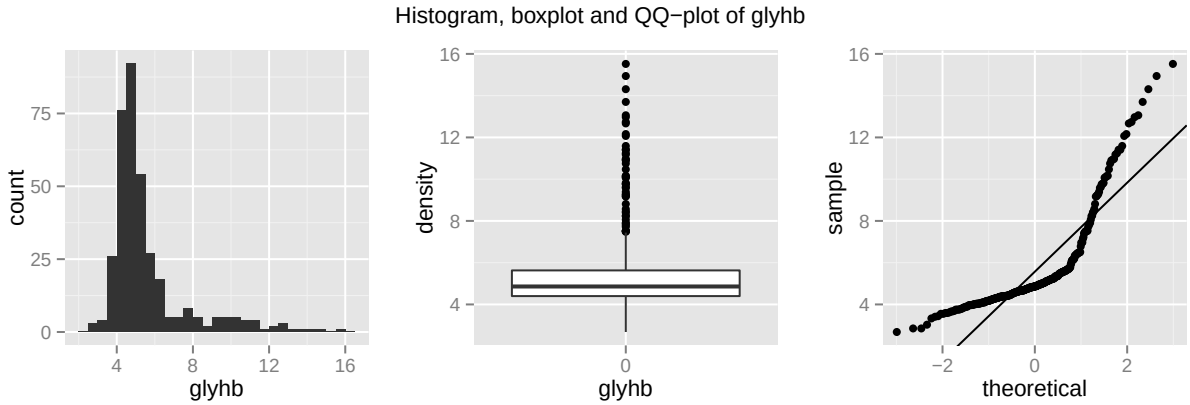
Lingtian Cheng, Yixuan Du, Ruijiao Song

May 1, 2015

1 Background

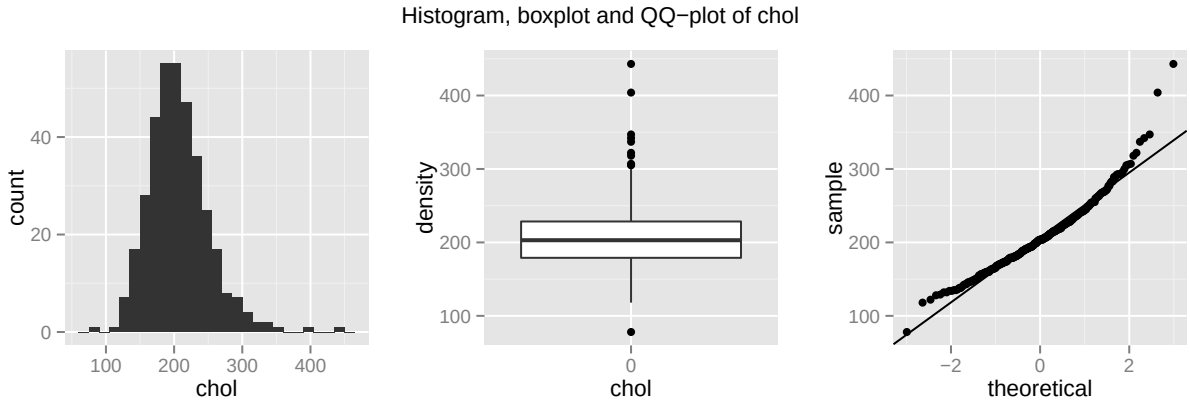
2 Accessing Data, Visualization and Summarization

1.



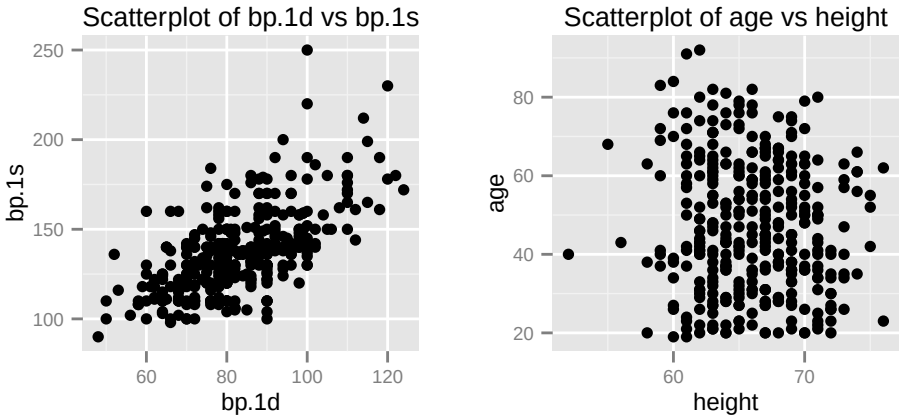
The mean, median and mode of `glyhb` are all approximately 5. The distribution of `glyhb` is left-skewed.

2.



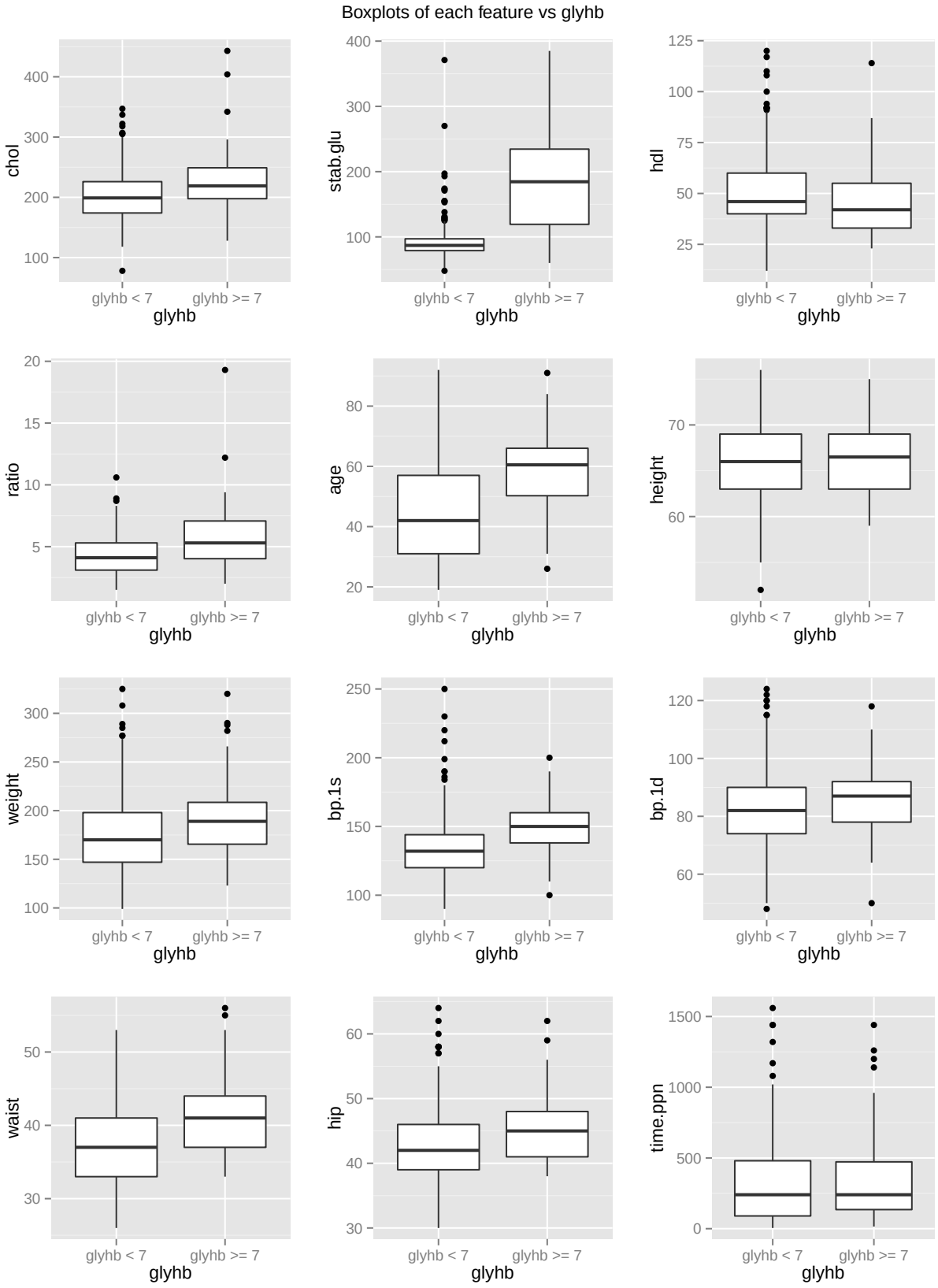
The mean, median and mode of `chol` are all approximately 200. The distribution of `chol` is better approximated with a Gaussian distribution.

3.

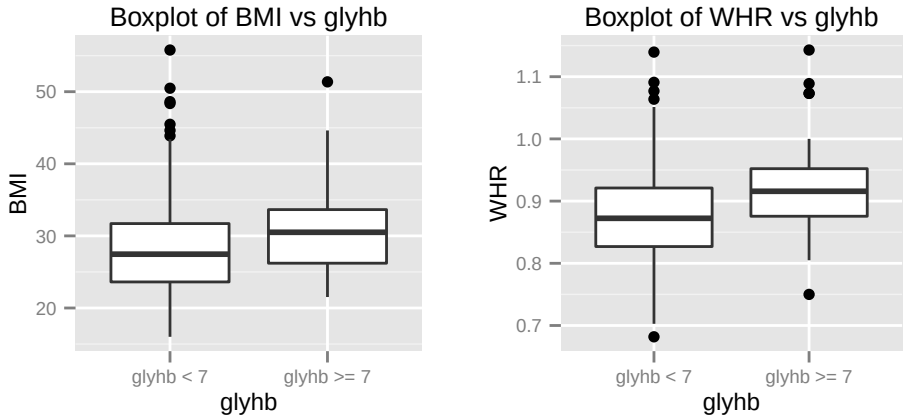


The scatterplot of `bp.1s` and `bp.1d` is near-linear, so they are approximately dependent. The scatterplot of `age` and `weight` is random, so they are approximately independent.

4.
- `chol`: The two distributions have small difference, so it MAY BE a relevant feature.
- `stab.glu`: The two distributions have substantial difference, so it SHOULD BE a relevant feature.
- `hdl`: The two distributions have small difference, so it MAY BE a relevant feature.
- `ratio`: The two distributions have small difference, so it MAY BE a relevant feature.
- `age`: The two distributions have substantial difference, so it SHOULD BE a relevant feature.
- `height`: The two distributions have little difference, so it MAY NOT BE a relevant feature.
- `weight`: The two distributions have small difference, so it MAY BE a relevant feature.
- `bp.1s`: The two distributions have small difference, so it MAY BE a relevant feature.
- `bp.1d`: The two distributions have small difference, so it MAY BE a relevant feature.
- `waist`: The two distributions have small difference, so it MAY BE a relevant feature.
- `hip`: The two distributions have small difference, so it MAY BE a relevant feature.
- `time.ppn`: The two distributions have small difference, so it MAY NOT BE a relevant feature.


Boxplots of each feature vs glyhb
chol
stab.glu
hdl
ratio
age
height
weight
bp.1s
bp.1d
waist
hip
time.ppn
glyhb < 7
glyhb >= 7
glyhb

5.



6. In light of these first experiments, `hdl`, `stab.glu`, `age`, `weight`, `bp.1s`, `bp.1d`, `waist` and `hip` seem related to the presence of type II diabetes; `chol`, `ratio`, `height` and `time.ppn` seem unrelated to the presence of type II diabetes.

3 Parametric Inference

1.

$$X \sim Gamma(\alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

$$E(X) = \frac{\alpha}{\beta}$$

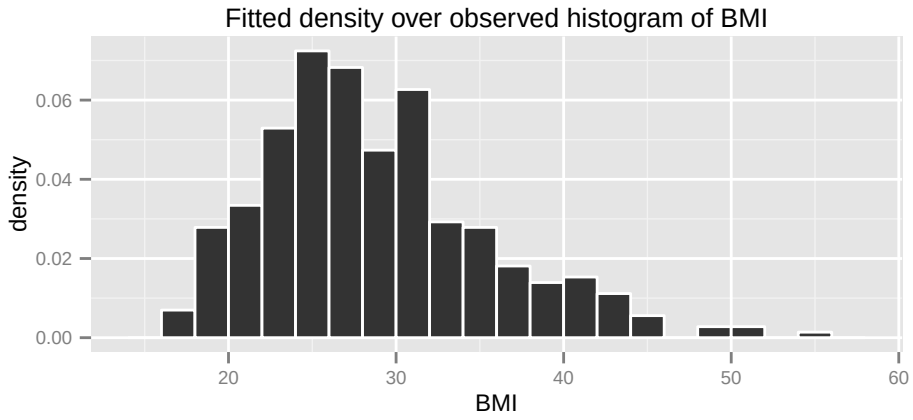
$$\begin{aligned} E(X^2) &= Var(X) + [E(X)]^2 \\ &= \frac{\alpha}{\beta^2} + \left(\frac{\alpha}{\beta}\right)^2 \\ &= \frac{\alpha(\alpha+1)}{\beta^2} \end{aligned}$$

$$\begin{cases} E(X) = \frac{\alpha}{\beta} \\ E(X^2) = \frac{\alpha(\alpha+1)}{\beta^2} \end{cases} \Longrightarrow \begin{cases} \alpha = \frac{[E(X)]^2}{Var(x)} \\ \beta = \frac{E(X)}{Var(x)} \end{cases} \Longrightarrow \begin{cases} \hat{\alpha}_{MOM} = \frac{\overline{X}_n}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2} \\ \hat{\beta}_{MOM} = \frac{\overline{X}_n^2}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2} \end{cases}$$

```
################################################################################
CI.BMI

##          alpha      beta
## 2.5%  15.87666 0.5460733
## 97.5% 21.62832 0.7591677

################################################################################
```



2.

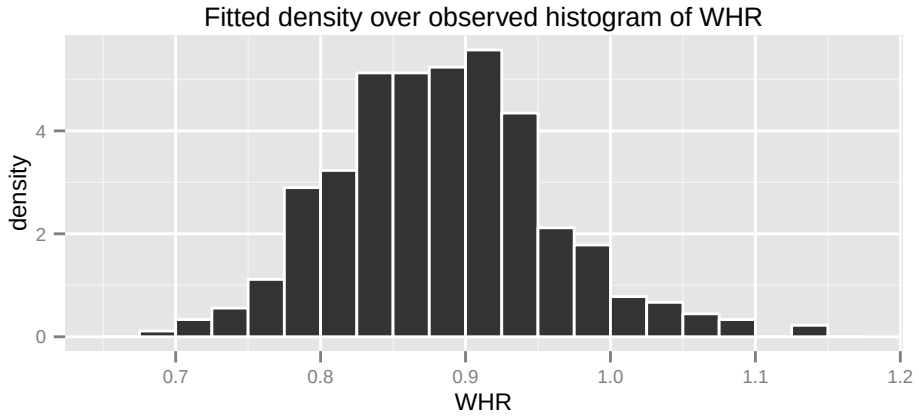
$$\hat{\mu}_{MLE} = \overline{X}_n$$

$$\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

```
################################################################################
CI.WHR

##               mu      sigma
## 2.5%  0.8747003 0.06788022
## 97.5% 0.8901891 0.08026878

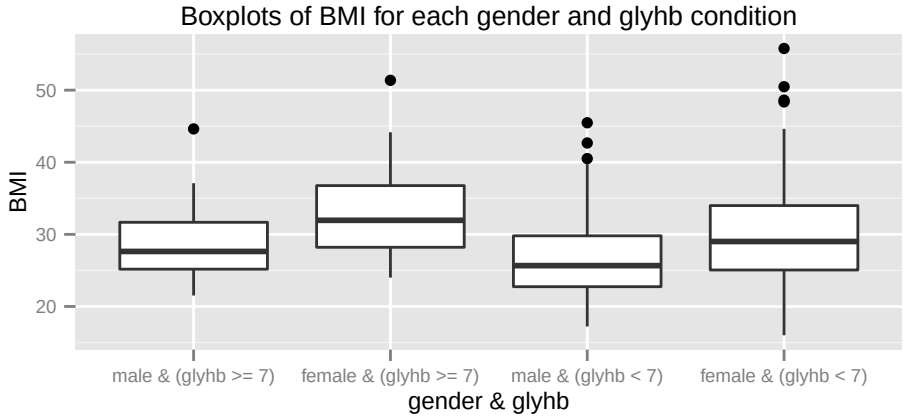
################################################################################
```



3.
```
##############################################################################
CIs.BMI

## $`male & (glyhb >= 7)`
##             mu    sigma
## 2.5%  26.89463 3.533570
## 97.5% 30.97481 7.164682
##
## $`female & (glyhb >= 7)`
##             mu    sigma
## 2.5%  30.90826 4.608429
## 97.5% 35.63231 8.207146
##
## $`male & (glyhb < 7)`
##             mu    sigma
## 2.5%  25.57434 4.711554
## 97.5% 27.47041 6.358453
##
## $`female & (glyhb < 7)`
##             mu    sigma
## 2.5%  28.84236 6.220170
## 97.5% 30.84639 7.862444

##############################################################################
```

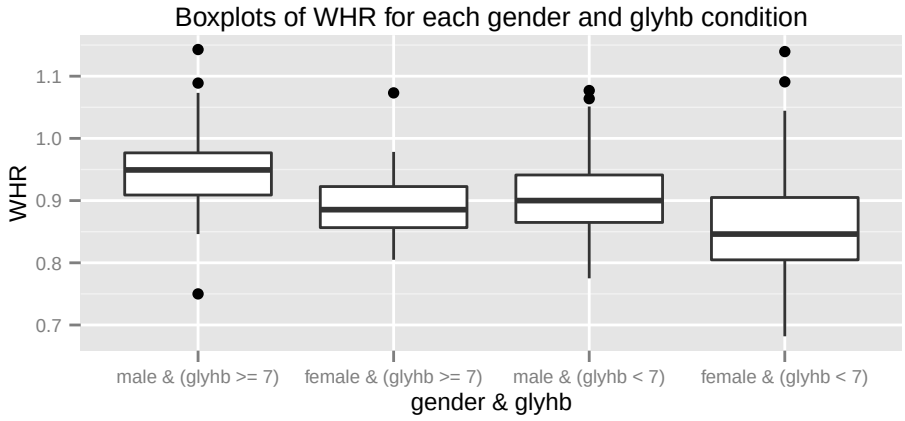


- On average, females have higher BMI than males.
- On average, people with type II diabetes (`glyhb >= 7`) have higher BMI than people without type II diabetes (`glyhb < 7`).
- People with type II diabetes (`glyhb >= 7`) have larger confidence intervals of both mean and standard deviation than people without type II diabetes (`glyhb < 7`), regardless of gender.

```
################################################################################
CIs.WHR

## $`male & (glyhb >= 7)`
##               mu      sigma
## 2.5%  0.9202216 0.04735021
## 97.5% 0.9791687 0.10652492
##
## $`female & (glyhb >= 7)`
##               mu      sigma
## 2.5%  0.8754801 0.04013498
## 97.5% 0.9139384 0.07537186
##
## $`male & (glyhb < 7)`
##               mu      sigma
## 2.5%  0.8938359 0.05568191
## 97.5% 0.9172246 0.07075628
##
## $`female & (glyhb < 7)`
##               mu      sigma
## 2.5%  0.8446725 0.06108226
## 97.5% 0.8644300 0.07829550

################################################################################
```



- On average, males have higher WHR than females.
- On average, people with type II diabetes (`glyhb >= 7`) have higher WHR than people without type II diabetes (`glyhb < 7`).
- People with type II diabetes (`glyhb >= 7`) have larger confidence intervals of both mean and standard deviation than people without type II diabetes (`glyhb < 7`), regardless of gender.

4 Testing

1.
```
##############################################################################
gender.glyhb.cond.table

##         glyhb >= 7 glyhb < 7
## male            24       125
## female          30       180

##############################################################################
fisher.test(gender.glyhb.cond.table)

##
##  Fisher's Exact Test for Count Data
##
## data:  gender.glyhb.cond.table
## p-value = 0.6552
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.6126316 2.1465820
## sample estimates:
## odds ratio
##   1.151538

##############################################################################
```

Since the p-value is 0.6552, which is greater than 0.05, we fail to reject the null hypothesis that males and females are equally exposed to type II diabetes, with 5% significance level.

2. We choose to the non-parametric Kruskal-Wallis test, because it does not rely on the assumed normal distribution and less affected by outliers.

```
################################################################################
kruskal.test(data.df$hdl, as.factor(data.df$glyhb.cond))

##
##  Kruskal-Wallis rank sum test
##
## data:  data.df$hdl and as.factor(data.df$glyhb.cond)
## Kruskal-Wallis chi-squared = 7.9732, df = 1, p-value = 0.004748

################################################################################
```

Since the p-value is 0.004748, which is smaller than 0.05, we reject the null hypothesis that `hdl` has equal means for those with type II diabetes and those without, with 5% significance interval.

```
################################################################################
kruskal.test(data.df$bp.1s, as.factor(data.df$glyhb.cond))

##
##  Kruskal-Wallis rank sum test
##
## data:  data.df$bp.1s and as.factor(data.df$glyhb.cond)
## Kruskal-Wallis chi-squared = 22.563, df = 1, p-value = 2.034e-06

################################################################################
```

Since the p-value is 2.034e-06, which is smaller than 0.05, we reject the null hypothesis that `bp.1s` has equal means for those with type II diabetes and those without, with 5% significance interval.

```
################################################################################
kruskal.test(data.df$bp.1d, as.factor(data.df$glyhb.cond))

##
##  Kruskal-Wallis rank sum test
##
## data:  data.df$bp.1d and as.factor(data.df$glyhb.cond)
## Kruskal-Wallis chi-squared = 1.9007, df = 1, p-value = 0.168

################################################################################
```

Since the p-value is 0.168, which is greater than 0.05, we fail to reject the null hypothesis that `bp.1d` has equal means for those with type II diabetes and those without, with 5% significance interval.

```
################################################################################
kruskal.test(data.df$BMI, as.factor(data.df$glyhb.cond))

##
##  Kruskal-Wallis rank sum test
##
## data:  data.df$BMI and as.factor(data.df$glyhb.cond)
## Kruskal-Wallis chi-squared = 9.5655, df = 1, p-value = 0.001983
```

```
################################################################################
```

Since the p-value is 0.001983, which is smaller than 0.05, we reject the null hypothesis that `BMI` has equal means for those with type II diabetes and those without, with 5% significance interval.

```
################################################################################
kruskal.test(data.df$WHR, as.factor(data.df$glyhb.cond))

##
##  Kruskal-Wallis rank sum test
##
## data:  data.df$WHR and as.factor(data.df$glyhb.cond)
## Kruskal-Wallis chi-squared = 15.146, df = 1, p-value = 9.95e-05

################################################################################
```

Since the p-value is 9.95e-05, which is smaller than 0.05, we reject the null hypothesis that `WHR` has equal means for those with type II diabetes and those without, with 5% significance interval.

3.
```
################################################################################
pi.male.BMI

## [1] 0.6326667

CI.pi.male.BMI

##      2.5%     97.5%
## 0.5222833 0.7420167

################################################################################
pi.male.WHR

## [1] 0.6853333

CI.pi.male.WHR

##      2.5%     97.5%
## 0.5716667 0.7950000

################################################################################
```

4. From 3.2, we know that `WHR` has a normal distribution. We assume all patients come from the same population, so the standard deviation is constant.

$$H_0 : N(\mu_0, \sigma^2)$$
$$H_1 : N(\mu_1, \sigma^2)$$

```
################################################################################
mean.male.glyhb.geq7.WHR

## [1] 0.9489365

mean.male.glyhb.l7.WHR

## [1] 0.9058419

sd.male.WHR

## [1] 0.06821165

################################################################################
```

$$\mu_0 = 0.9058419, \mu_1 = 0.9489365, \sigma = 0.06821165$$

$$\begin{aligned} lik(x) &= \frac{f_0(x)}{f_1(x)} \\ &= \frac{\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu_0)^2}{2}}}{\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu_1)^2}{2}}} \\ &= e^{-2(\mu_1-\mu_0)x+(\mu_1^2-\mu_0^2)} \end{aligned}$$

Let $T := X$.

$$\begin{aligned} \alpha &= P(T > t \mid H_0) \\ &= P(\frac{T-\mu_0}{\sigma} > \frac{t-\mu_0}{\sigma}) \\ &= 1 - \Phi\left(\frac{t-\mu_0}{\sigma}\right) \\ &\implies \\ t &= \Phi^{-1}(\alpha)\sigma + \mu_0 \end{aligned}$$

$$\begin{aligned} \beta &= P(T < t \mid H_1) \\ &= P(\frac{T-\mu_1}{\sigma} < \frac{t-\mu_1}{\sigma}) \\ &= \Phi\left(\frac{t-\mu_1}{\sigma}\right) \\ &= \Phi\left(\frac{\Phi^{-1}(\alpha)\sigma + \mu_0 - \mu_1}{\sigma}\right) \end{aligned}$$

```
################################################################################
t

## [1] 1.01804

power

## [1] 0.1555122

################################################################################
```

$$\alpha \leqslant 5\% \implies 1 - \beta < 0.1555122$$

5.

6.

5 Regression

1.

2.

3.

4.

5.

6.

7.

A Appendix

A.1 Background

```
library("ggplot2")
library("grid")
library("gridExtra")

data.df <- na.omit(read.csv("diabetes.csv"))
```

A.2 Accessing Data, Visualization and Summarization

1. ```
   glyhb.histogram <- ggplot(data.df) +
     geom_histogram(aes(x=glyhb), binwidth=0.5)

   glyhb.boxplot <- ggplot(data.df) +
     geom_boxplot(aes(x=factor(0), y=glyhb)) +
     labs(x="glyhb", y="density")

   glyhb.qqplot <- ggplot(data.df) +
     stat_qq(aes(sample=glyhb)) +
     geom_abline(aes(intercept=mean(glyhb), slope=sd(glyhb)))

   grid.arrange(glyhb.histogram, glyhb.boxplot, glyhb.qqplot, ncol=3,
                main="Histogram, boxplot and QQ-plot of glyhb")
   ```

2. ```
   chol.histogram <- ggplot(data.df) +
     geom_histogram(aes(x=chol), binwidth=15)

   chol.boxplot <- ggplot(data.df) +
     geom_boxplot(aes(x=factor(0), y=chol)) +
     labs(x="chol", y="density")

   chol.qqplot <- ggplot(data.df) +
     stat_qq(aes(sample=chol)) +
     geom_abline(aes(intercept=mean(chol), slope=sd(chol)))

   grid.arrange(chol.histogram, chol.boxplot, chol.qqplot, ncol=3,
                main="Histogram, boxplot and QQ-plot of chol")
   ```

3. ```
   bp.1d.bp.1s.scatterplot <- ggplot(data.df) +
     geom_point(aes(x=bp.1d, y=bp.1s)) +
     labs(title="Scatterplot of bp.1d vs bp.1s") +
     theme(text=element_text(size=8.5))

   height.age.scatterplot <- ggplot(data.df) +
     geom_point(aes(x=height, y=age)) +
     labs(title="Scatterplot of age vs height") +
     theme(text=element_text(size=8.5))

   grid.arrange(bp.1d.bp.1s.scatterplot, height.age.scatterplot, ncol=2)
   ```
   ```

4.
```
data.df$glyhb.cond <- NA
data.df[data.df$glyhb>=7, ]$glyhb.cond <- "glyhb >= 7"
data.df[data.df$glyhb<7, ]$glyhb.cond <- "glyhb < 7"

features <- c("chol", "stab.glu", "hdl", "ratio", "age", "height", "weight",
              "bp.1s", "bp.1d", "waist", "hip", "time.ppn")
feature.boxplots = list()
for (feature in features) {
  feature.boxplot <- ggplot(data.frame(glyhb.cond=data.df$glyhb.cond,
                                       feature=data.df[[feature]])) +
    geom_boxplot(aes(x=glyhb.cond, y=feature)) +
    labs(x="glyhb", y=feature)
  feature.boxplots[[feature]] <- feature.boxplot
}

do.call(grid.arrange, c(feature.boxplots, ncol=3,
                        main="Boxplots of each feature vs glyhb"))
```

5.
```
data.df$BMI <- 703*data.df$weight/data.df$height^2
data.df$WHR <- data.df$waist/data.df$hip

features <- c("BMI", "WHR")
feature.boxplots = list()
for (feature in features) {
  feature.boxplot <- ggplot(data.frame(glyhb.cond=data.df$glyhb.cond,
                                       feature=data.df[[feature]])) +
    geom_boxplot(aes(x=glyhb.cond, y=feature)) +
    labs(title=paste(c("Boxplot of", feature, "vs glyhb"), collapse=" "),
         x="glyhb", y=feature) +
    theme(text=element_text(size=8.5))
  feature.boxplots[[feature]] <- feature.boxplot
}

do.call(grid.arrange, c(feature.boxplots, ncol=2))
```

A.3 Parametric Inference

1.
```
gamma.boot <- function(x) {
  alpha.mom <- mean(x)^2/var(x)
  beta.mom <- mean(x)/var(x)

  bootstrap <- sapply(1:1000, function(i) {
    samples <- sample(x, length(x), replace=TRUE)
    alpha.sample <- mean(samples)^2/var(samples)
    beta.sample <- mean(samples)/var(samples)
    return(c(alpha.sample, beta.sample))
  })

  CIs <- sapply(1:2, function(i) {
    CI <- quantile(bootstrap[i, ], probs=c(0.025, 0.975))
    return(CI)
```

```
  })
  colnames(CIs) <- c("alpha", "beta")

  return(CIs)
}

CI.BMI <- gamma.boot(data.df$BMI)

ggplot(data.df) +
  geom_histogram(aes(x=BMI, y=..density..), binwidth=2, col="white") +
  stat_function(fun=function(x)
    dgamma(x, shape=alpha.mom, rate=beta.mom), col="white") +
  labs(title="Fitted density over observed histogram of BMI") +
  theme(text=element_text(size=8.5))
```

2.
```
normal.boot <- function(x) {
  mu.mle <- mean(x)
  sigma.mle <- sd(x)

  bootstrap <- sapply(1:1000, function(i) {
    samples <- sample(x, length(x), replace=TRUE)
    mu.sample <- mean(samples)
    sigma.sample <- sd(samples)
    return(c(mu.sample, sigma.sample))
  })

  CIs <- sapply(1:2, function(i) {
    CI <- quantile(bootstrap[i, ], probs=c(0.025, 0.975))
    return(CI)
  })
  colnames(CIs) <- c("mu", "sigma")

  return(CIs)
}

CI.WHR <- normal.boot(data.df$WHR)

ggplot(data.df) +
  geom_histogram(aes(x=WHR, y=..density..), binwidth=0.025, col="white") +
  stat_function(fun=function(x)
    dnorm(x, mean=mu.mle, sd=sigma.mle), col="white") +
  labs(title="Fitted density over observed histogram of WHR") +
  theme(text=element_text(size=8.5))
```

3.
```
data.df$gender.glyhb.cond <- NA
data.df[data.df$gender=="male" & data.df$glyhb>=7, ]$gender.glyhb.cond <-
  "male & (glyhb >= 7)"
data.df[data.df$gender=="female" & data.df$glyhb>=7, ]$gender.glyhb.cond <-
  "female & (glyhb >= 7)"
data.df[data.df$gender=="male" & data.df$glyhb<7, ]$gender.glyhb.cond <-
  "male & (glyhb < 7)"
```

```r
data.df[data.df$gender=="female" & data.df$glyhb<7, ]$gender.glyhb.cond <-
  "female & (glyhb < 7)"

conditions <- c("male & (glyhb >= 7)", "female & (glyhb >= 7)",
                "male & (glyhb < 7)", "female & (glyhb < 7)")

gamma.boot2 <- function(x) {
  alpha.mom <- mean(x)^2/var(x)
  beta.mom <- mean(x)/var(x)

  bootstrap <- sapply(1:1000, function(i) {
    samples <- sample(x, length(x), replace=TRUE)
    mu.sample <- mean(samples)
    sigma.sample <- sd(samples)
    return(c(mu.sample, sigma.sample))
  })

  CIs <- sapply(1:2, function(i) {
    CI <- quantile(bootstrap[i, ], probs=c(0.025, 0.975))
    return(CI)
  })
  colnames(CIs) <- c("mu", "sigma")

  return(CIs)
}

CIs.BMI <- lapply(conditions, function(x) {
  gender.glyhb.df <- subset(data.df, gender.glyhb.cond==x)
  CIs.BMI <- gamma.boot2(gender.glyhb.df$BMI)
  return(CIs.BMI)
})
names(CIs.BMI) <- conditions

ggplot(data.df) +
  geom_boxplot(aes(x=factor(gender.glyhb.cond, levels=conditions), y=BMI)) +
  labs(title="Boxplots of BMI for each gender and glyhb condition",
       x="gender & glyhb") +
  theme(text=element_text(size=8.5))

CIs.WHR <- lapply(conditions, function(x) {
  gender.glyhb.df <- subset(data.df, gender.glyhb.cond==x)
  CIs.WHR <- normal.boot(gender.glyhb.df$WHR)
  return(CIs.WHR)
})
names(CIs.WHR) <- conditions

ggplot(data.df) +
  geom_boxplot(aes(x=factor(gender.glyhb.cond, levels=conditions), y=WHR)) +
  labs(title="Boxplots of WHR for each gender and glyhb condition",
       x="gender & glyhb") +
  theme(text=element_text(size=8.5))
```

A.4 Testing