

Statistics 135 – Lab Project

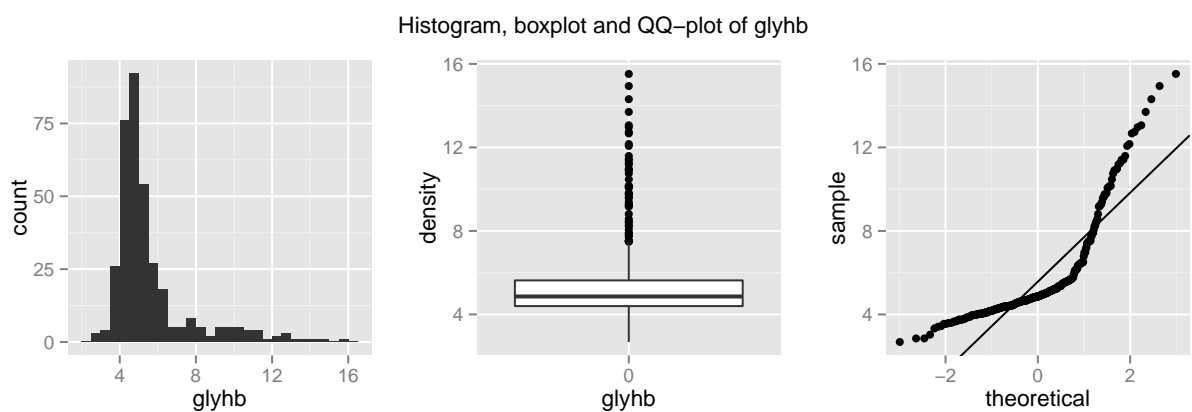
Lingtian Cheng, Yixuan Du, Ruijiao Song

May 1, 2015

1 Background

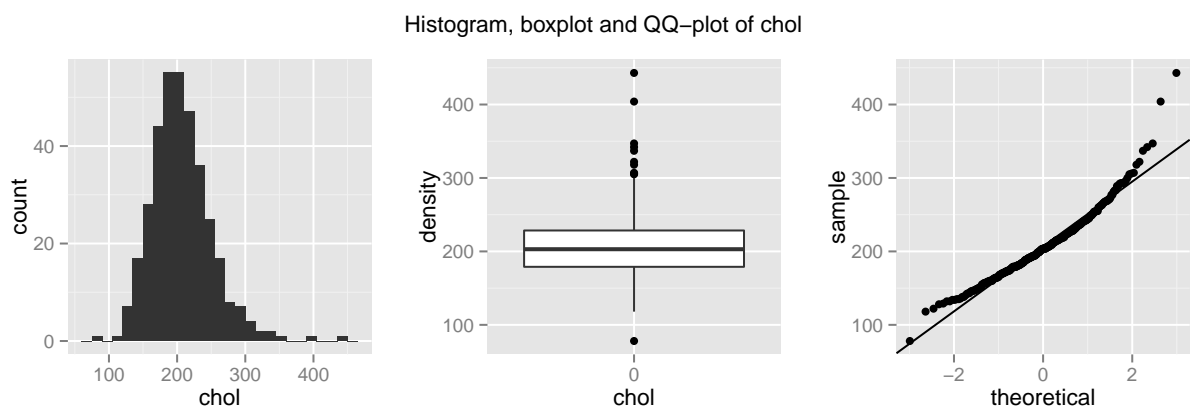
2 Accessing Data, Visualization and Summarization

1.



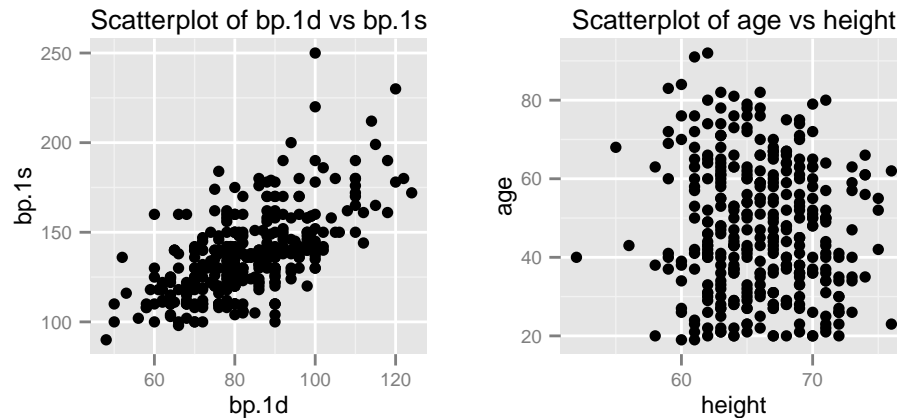
The mean, median and mode of `glyhb` are all approximately 5. The distribution of `glyhb` is left-skewed.

2.



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

3.

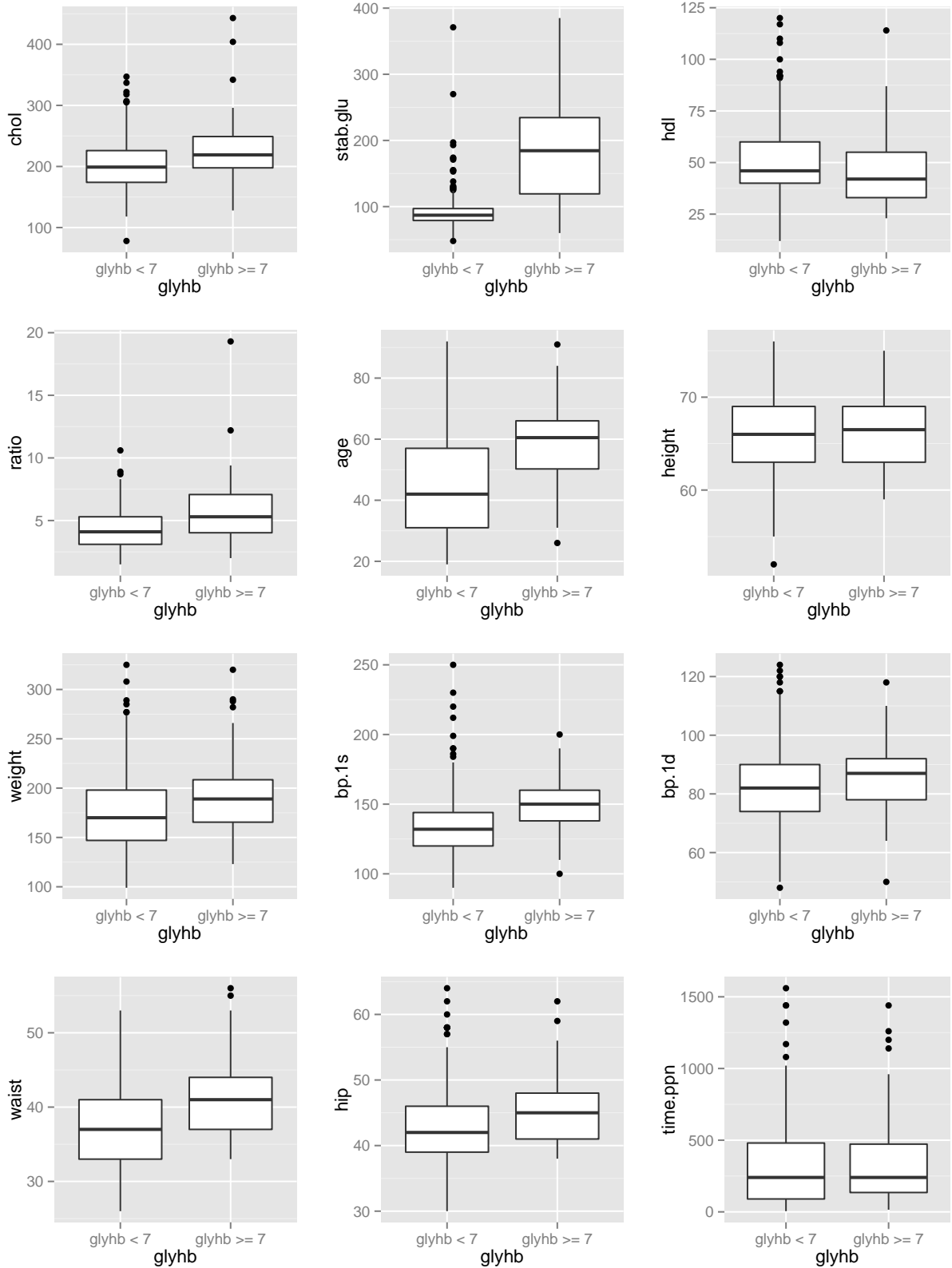


The scatterplot of `bp.1s` and `bp.1d` is near-linear, so they are approximately dependent. The scatterplot of `age` and `weight` is random, so they are approximately independent.

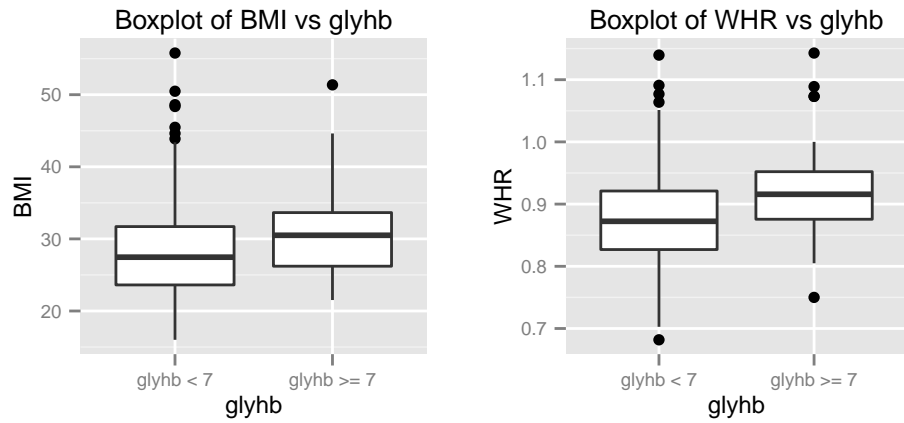
4.

- `chol`: The two distributions have small difference, so it may BE a relevant feature.
- `stab.glu`: The two distributions have substantial difference, so it may 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 may 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



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.

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

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

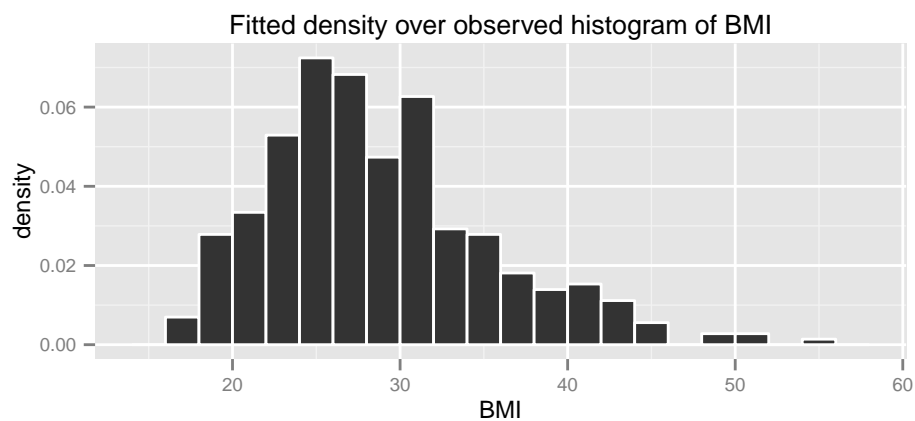
$$\begin{aligned} E(X^2) &= \text{Var}(X) + [E(X)]^2 \\ &= \frac{\alpha}{\beta^2} + \left(\frac{\alpha}{\beta}\right)^2 \\ &= \frac{\alpha(\alpha+1)}{\beta^2} \end{aligned}$$

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

```
#####
CI.BMI

##          alpha      beta
## 2.5%    15.77434 0.5420062
## 97.5%   21.52204 0.7547185

#####
```



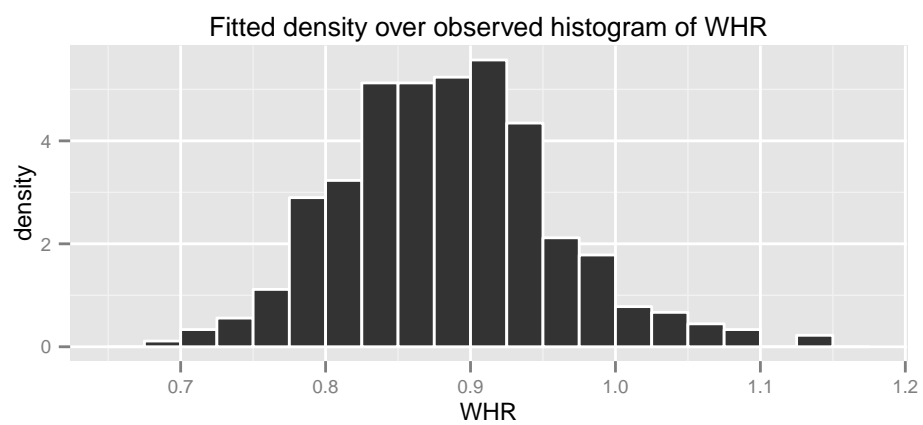
2.

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

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

```
#####
CI.WHR

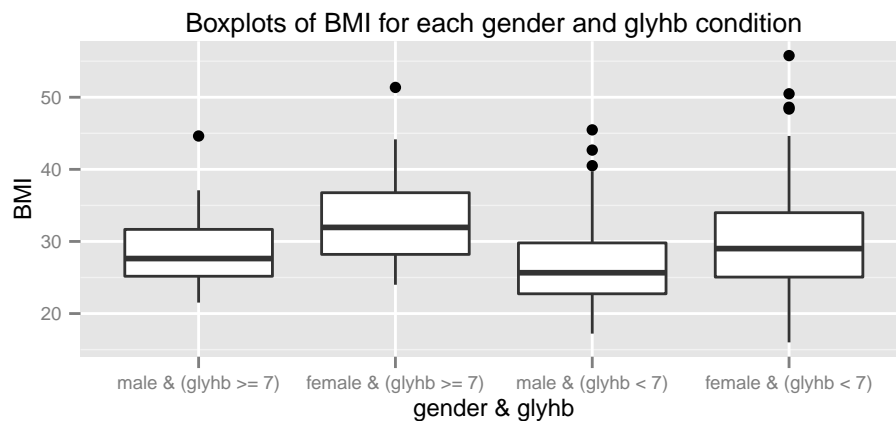
##          mu      sigma
## 2.5%  0.8742748 0.06756228
## 97.5% 0.8903982 0.07952507
#####
```



3.

```
#####
CIs.BMI

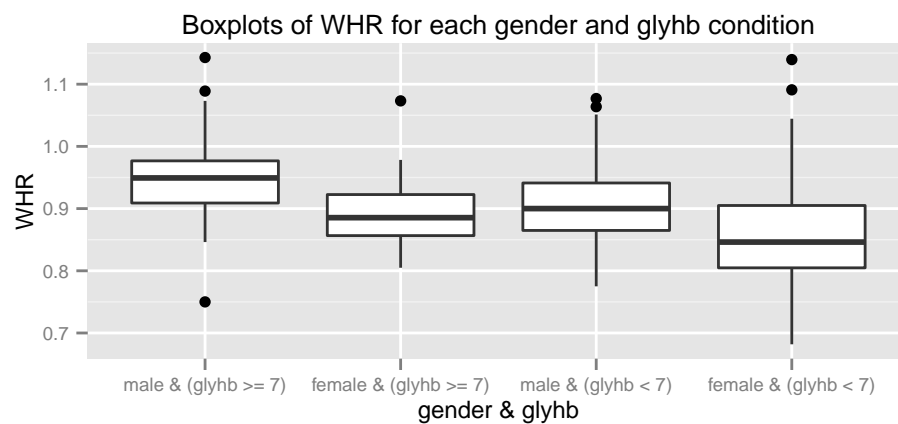
## $`male & (glyhb >= 7)`
##           mu      sigma
## 2.5%  26.84307 3.427979
## 97.5% 31.01954 7.085508
##
## $`female & (glyhb >= 7)`
##           mu      sigma
## 2.5%  31.05788 4.846527
## 97.5% 35.72685 8.152714
##
## $`male & (glyhb < 7)`
##           mu      sigma
## 2.5%  25.46761 4.762291
## 97.5% 27.43883 6.386971
##
## $`female & (glyhb < 7)`
##           mu      sigma
## 2.5%  28.78206 6.225949
## 97.5% 30.85951 7.839278
#####
```



- 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%  0.9188920 0.04389852
## 97.5% 0.9800054 0.10382869
##
## $`female & (glyhb >= 7)`
##           mu           sigma
## 2.5%  0.8728465 0.04107824
## 97.5% 0.9149814 0.07667731
##
## $`male & (glyhb < 7)`
##           mu           sigma
## 2.5%  0.8948607 0.05534934
## 97.5% 0.9170900 0.07072515
##
## $`female & (glyhb < 7)`
##           mu           sigma
## 2.5%  0.8445763 0.06156524
## 97.5% 0.8650507 0.07899223
#####
```



- 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. We assume that under null hypothesis, males and females are equally exposed to type II diabetes.

```
#####
gender.glyhb.table

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

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

##
## Fisher's Exact Test for Count Data
##
## data:  gender.glyhb.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

#####
```

The p -value is 0.6552, which is greater than 0.05, so we failed to reject the null hypothesis with 95% significance level. Therefore, males and females are equally exposed to type II diabetes.

- 2.
- 3.
- 4.
- 5.
- 6.

5 Regression

- 1.
- 2.
- 3.
- 4.
- 5.
- 6.
- 7.