第二次作业

赵映辉

Question #1: BigBangTheory. (Attached Data: BigBangTheory)

The Big Bang Theory, a situation comedy featuring Johnny Galecki, Jim Parsons, and Kaley Cuoco-Sweeting, is one of the most-watched programs on network television. The first two episodes for the 2011–2012 season premiered on September 22, 2011; the first episode attracted 14.1 million viewers and the second episode attracted 14.7 million viewers. The attached data file BigBangTheory shows the number of viewers in millions for the first 21 episodes of the 2011–2012 season (the Big Bang theory website, April 17, 2012).

a. Compute the minimum and the maximum number of viewers.

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become
```

```
library(readxl)
library(janitor)
```

```
## Warning: 程序包'janitor'是用R版本4.4.2 来建造的
```

```
##
## 载入程序包: 'janitor'
```

```
##
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```r
table1 = read.csv("./data/BigBangTheory.csv")

table1 = clean_names(table1)

# 获取最大值
max(table1$viewers_millions)
```

```
## [1] 16.5
```

```r
# 获取最小值
min(table1$viewers_millions)
```

```
## [1] 13.3
```

b. Compute the mean, median, and mode.

```r
# 平均数
mean(table1$viewers_millions)
```

```
## [1] 15.04286
```

```r
# 中位数
median(table1$viewers_millions)
```

```
## [1] 15
```

```r
# 众数
result = table(table1$viewers_millions, useNA = "no")

as.numeric(names(result[result == max(result)]))
```

```
## [1] 13.6 14.0 16.1 16.2
```

c. Compute the first and third quartiles

```r
# 四分
quantile(table1$viewers_millions, probs = c(0.25, 0.75))
```

```
##  25%  75%
## 14.1 16.0
```

d. has viewership grown or declined over the 2011–2012 season? Discuss.

还没做

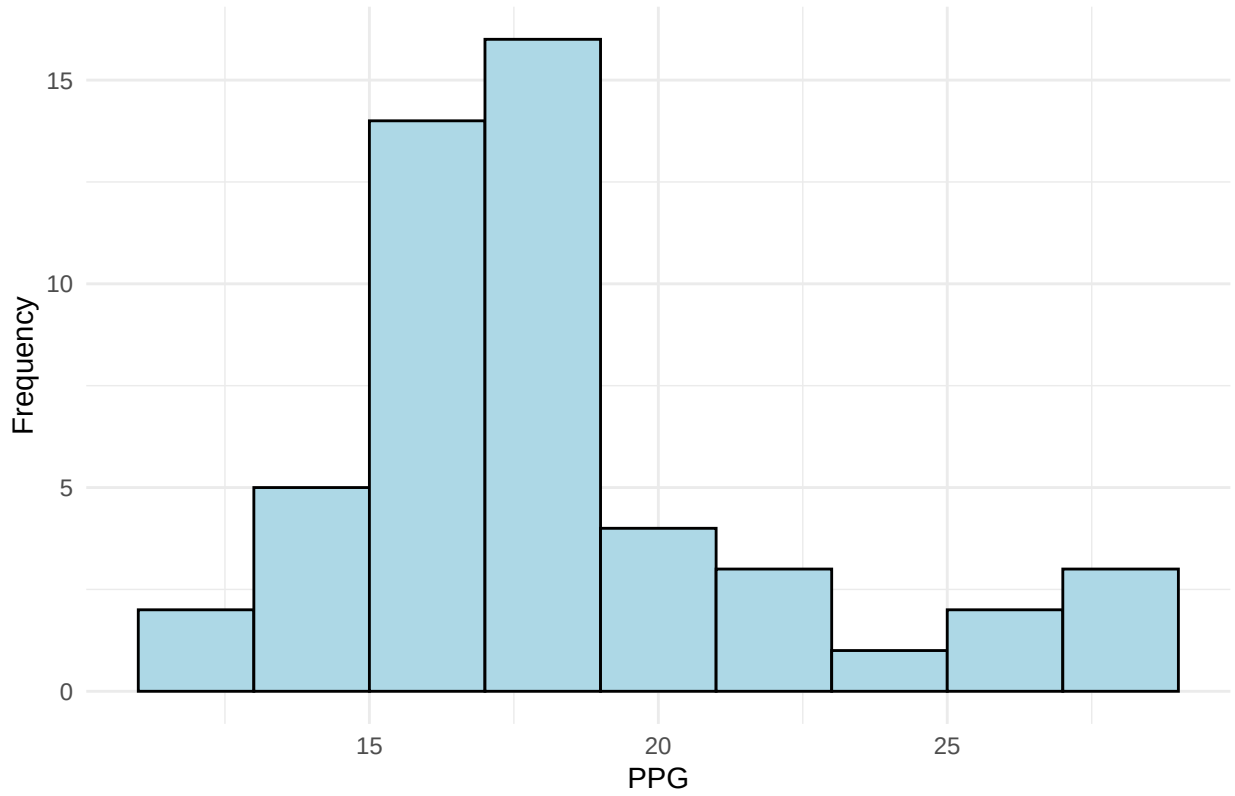
Question #2: NBAPlayerPts. (Attached Data: NBAPlayerPts)

CbSSports.com developed the Total Player Rating system to rate players in the National Basketball Association (NBA) based on various offensive and defensive statistics. The attached data file NBAPlayerPts shows the average number of points scored per game (PPG) for 50 players with the highest ratings for a portion of the 2012–2013 NBA season (CbSSports.com website, February 25, 2013). Use classes starting at 10 and ending at 30 in increments of 2 for PPG in the following.

a.Show the frequency distribution.

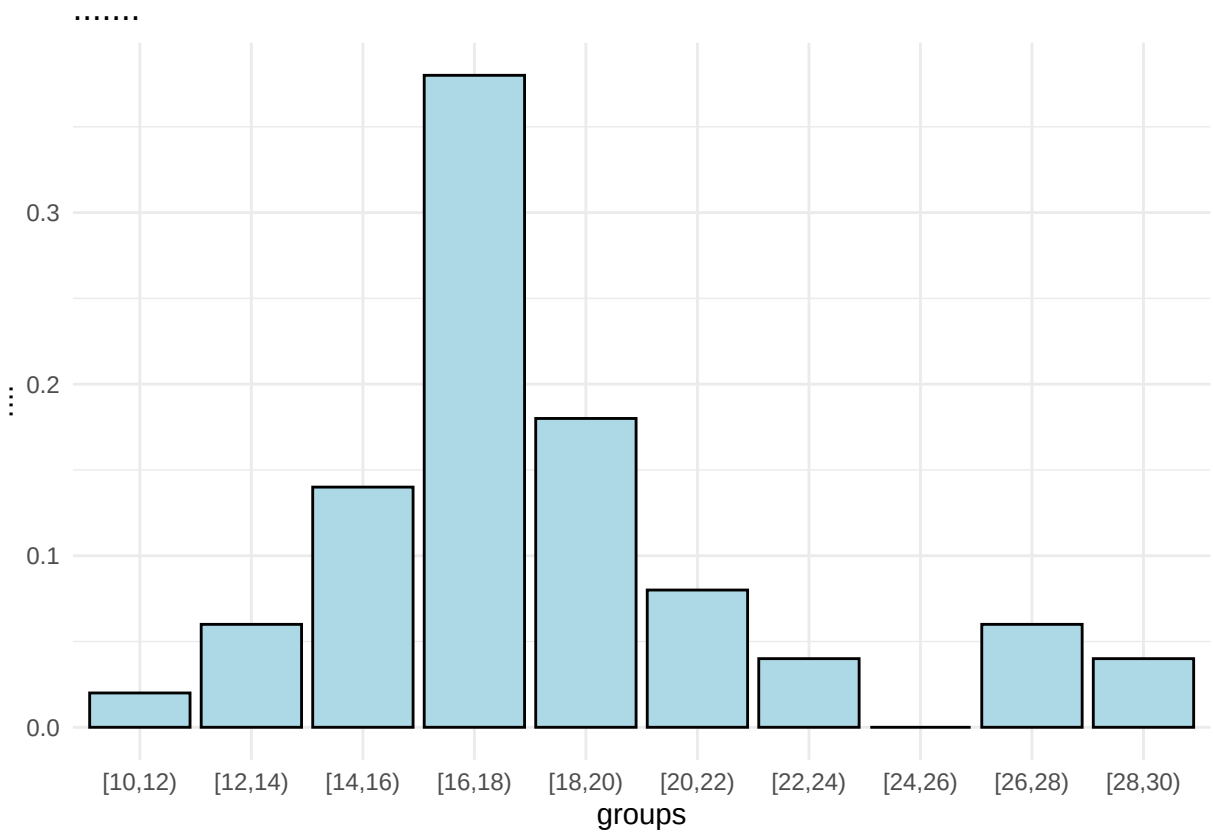
```r
# 实现频率直方图
library(tidyverse)
table2 = read.csv("./data/NBAPlayerPts.csv")

ggplot(table2, aes(x = PPG)) + geom_histogram(binwidth = 2, fill = "lightblue", color = "black") +
```

.....



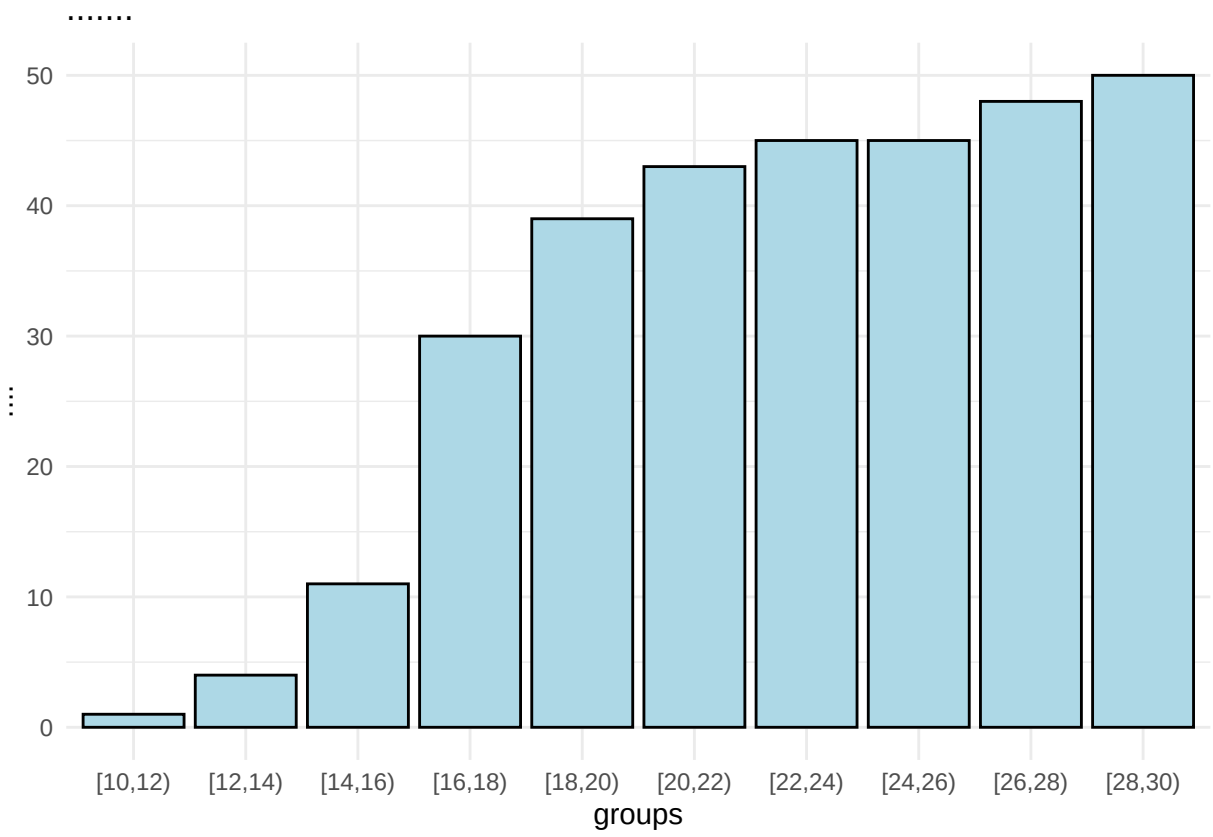
b. Show the relative frequency distribution.

```r
# 相对频率分布直方图
library(tidyverse)

breaks = seq(10, 30, by = 2)

groups = cut(table2$PPG, breaks, right = FALSE)

freq_table = as.data.frame(table(groups))

freq_table = freq_table %>% mutate(relativeFrequency = Freq / sum(Freq))

ggplot(freq_table, aes(x = groups, y = relativeFrequency)) + geom_bar(stat = "identity", fill = "l
```



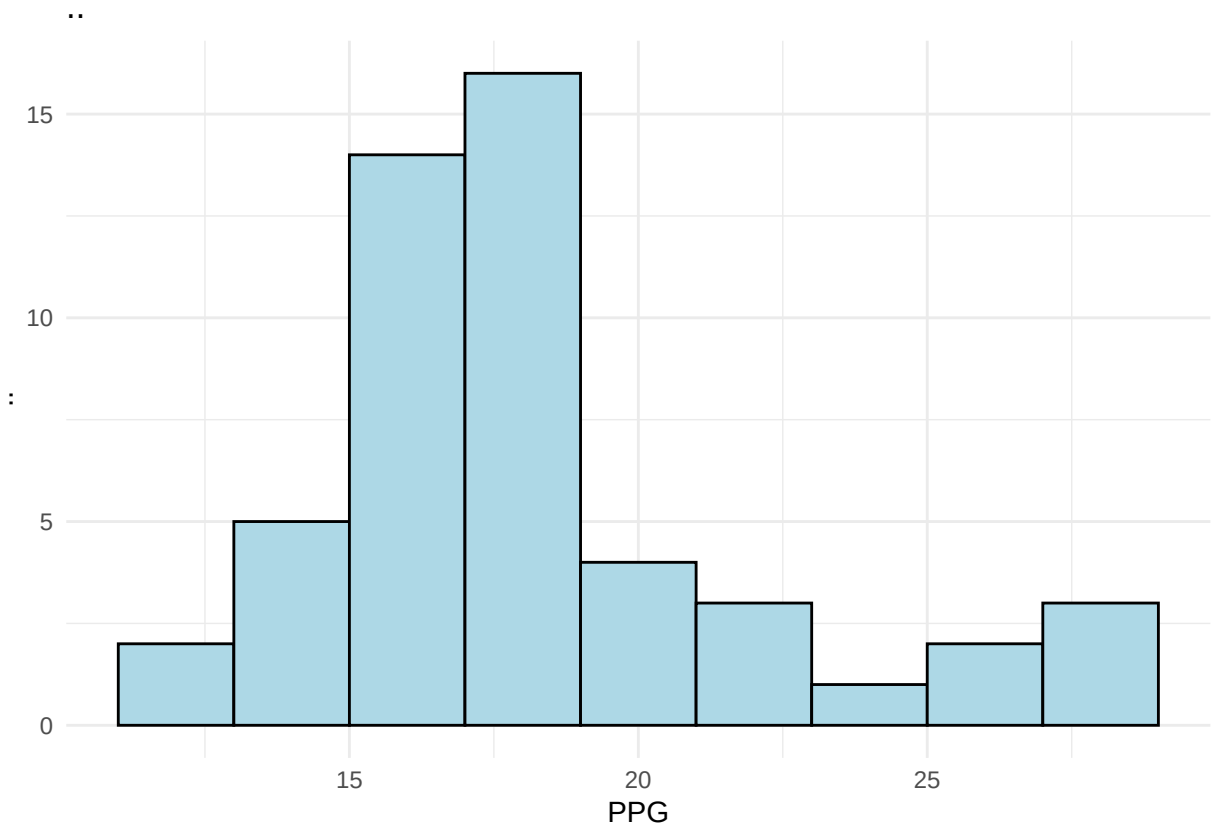
c. Show the cumulative percent frequency distribution.

```
# 累积频率分布直方图
library(tidyverse)

freq_table = freq_table %>% mutate(cumulativeFrequency = cumsum(Freq))

ggplot(freq_table, aes(x = groups, y = cumulativeFrequency)) + geom_bar(stat = "identity", fill =
```



d. Develop a histogram for the average number of points scored per game.

```
# 分数分布直方图
library(tidyverse)

ggplot(table2, aes(x = PPG)) + geom_histogram(binwidth = 2, fill = "lightblue", color = "black") +
```



e. Do the data appear to be skewed? Explain.

是的，可以发现频率分布直方图左偏，在累计分布的时候可以发现在 [14, 16), (16, 18) 这个区间是快速上升的，数据分布不均匀的

f. What percentage of the players averaged at least 20 points per game?

```
# 分数分布直方图
library(tidyverse)

nrow(table2 %>% filter(PPG >= 20)) / nrow(table2)
```

```
## [1] 0.22
```

Question #3: A researcher reports survey results by stating that the standard error of the mean is 20. The population standard deviation is 500.

a. How large was the sample used in this survey?

```r
(500 / 20) ^ 2
```

```
## [1] 625
```

b. What is the probability that the point estimate was within ±25 of the population mean?

```r
pnorm(25 / 20) - pnorm(- (25 / 20))
```

```
## [1] 0.7887005
```

如果用总体均值在 +-25 的时候，78% 的点会落到这个区间，证明这个波动是可靠的

Question #4: Young Professional Magazine (Attached Data: Professional)

Young Professional magazine was developed for a target audience of recent college graduates who are in their first 10 years in a business/professional career. In its two years of publication, the magazine has been fairly successful. Now the publisher is interested in expanding the magazine's advertising base. Potential advertisers continually ask about the demographics and interests of subscribers to young Professionals. To collect this information, the magazine commissioned a survey to develop a profile of its subscribers. The survey results will be used to help the magazine choose articles of interest and provide advertisers with a profile of subscribers. As a new employee of the magazine, you have been asked to help analyze the survey results.

Some of the survey questions follow:

What is your age?

Are you: Male_____ Female_____

Do you plan to make any real estate purchases in the next two years?

Yes_____ No_____

What is the approximate total value of financial investments, exclusive of your

home, owned by you or members of your household?

How many stock/bond/mutual fund transactions have you made in the past year?

Do you have broadband access to the Internet at home? Yes_____ No_____

Please indicate your total household income last year. _____

Do you have children? Yes_____ No_____

The file entitled Professional contains the responses to these questions.

Managerial Report:

Prepare a managerial report summarizing the results of the survey. In addition to statistical summaries, discuss how the magazine might use these results to attract advertisers. You might also comment on how the survey results could be used by the magazine's editors to identify topics that would be of interest to readers. Your report should address the following issues, but do not limit your analysis to just these areas.

a. Develop appropriate descriptive statistics to summarize the data.

b. Develop 95% confidence intervals for the mean age and household income of subscribers.

c. Develop 95% confidence intervals for the proportion of subscribers who have broadband access at home and the proportion of subscribers who have children.

d. Would Young Professional be a good advertising outlet for online brokers? Justify your conclusion with statistical data.

e. Would this magazine be a good place to advertise for companies selling educational software and computer games for young children?

f. Comment on the types of articles you believe would be of interest to readers of Young Professional.