3770 HW6

Darnell Chen

2024-10-23

Problem 6.1.6

The sample mean of our data set can be calculated by adding all our values and dividing by the sample size:\

$\bar{x} = \sum_{i=1}^{n} x_i \approx 14.35895$

On the other hand, our standard deviance is simply the squareroot of our variance s^2:

$s^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n-1} \approx 356.4716$
$s = \sqrt{s^2} = \sqrt{356.4716} \approx 18.88045$

Thus, we get a mean of approx. 14.35895 and a standard deviation of approx. 18.88045.

```
times = read.table("6-8.txt", header = TRUE)
times = times$time

mean(times)
```

```
## [1] 14.35895
```

```
var(times)
```

```
## [1] 356.4716
```

```
sqrt(var(times))
```

```
## [1] 18.88045
```

Problem 6.2.4

The two middle numbers in our data set are both 90.4. Since $\frac{90.4+90.4}{2} = 90.4$, our median is thus, 90.4.

```
rating = read.table("6-30.txt", header = TRUE)
rating = rating$Rating

median(rating)
```

```
## [1] 90.4
```

Our Quartiles:

```
quantiles = quantile(rating,type=6)
quantiles
```

```
##      0%     25%     50%     75%    100%
##  83.400  88.575  90.400  92.200 100.300
```

Our Stem and Leaf Plot:

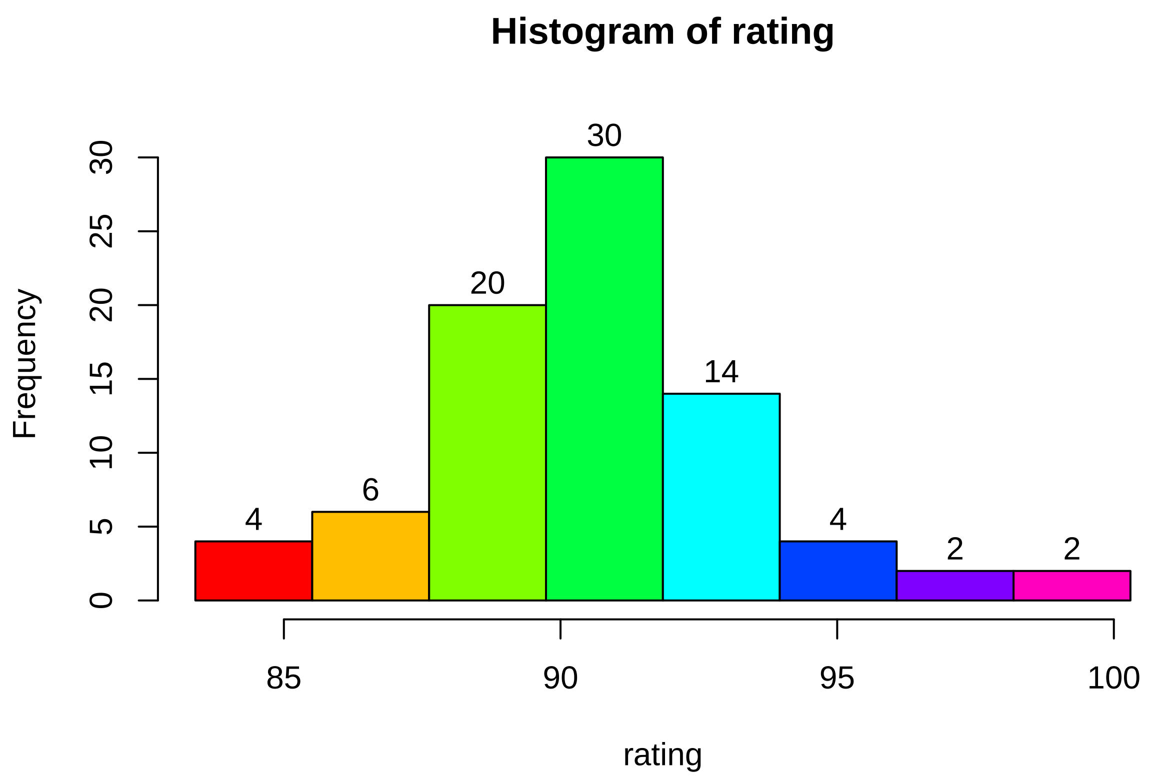
```
stem(rating, scale = 2)
```

```
##
##   The decimal point is at the |
##
##    83 | 4
##    84 | 33
##    85 | 3
##    86 | 777
##    87 | 456789
##    88 | 23334556679
##    89 | 0233678899
##    90 | 0111344456789
##    91 | 0001112256688
##    92 | 22236777
##    93 | 023347
##    94 | 2247
##    95 |
##    96 | 15
##    97 |
##    98 | 8
##    99 |
##   100 | 3
```

Problem 6.3.2

Our Histogram:

```
bin=seq(min(rating),max(rating),by=(max(rating)-min(rating))/8)
freqs = hist(rating, breaks=bin, label=TRUE, right=FALSE, col=rainbow(8), ylim=c(0, 32))
```




```
freqs
```

```
## $breaks
## [1]  83.4000  85.5125  87.6250  89.7375  91.8500  93.9625  96.0750  98.1875
## [9] 100.3000
##
## $counts
## [1]  4  6 20 30 14  4  2  2
##
## $density
## [1] 0.02309136 0.03463703 0.11545678 0.17318516 0.08081974 0.02309136 0.01154568
## [8] 0.01154568
##
## $mids
## [1] 84.45625 86.56875 88.68125 90.79375 92.90625 95.01875 97.13125 99.24375
##
## $xname
## [1] "rating"
```

```
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
```

Our Frequency Table:

```r
str=NULL
for (i in 1:7) {
  str = c(str, paste(freqs$breaks[i], "$\\le x < $", freqs$breaks[i+1]))
}

# Handle the last interval with inclusive bounds
str = c(str, paste(freqs$breaks[8], "$\\le x \\le $", freqs$breaks[9]))

df = data.frame(Class=str, Index=freqs$counts)
library(knitr)
library(kableExtra)
kable(df, "latex", align="c", escape = F, caption="Hint") %>%
kable_styling(latex_options = "hold_position")
```

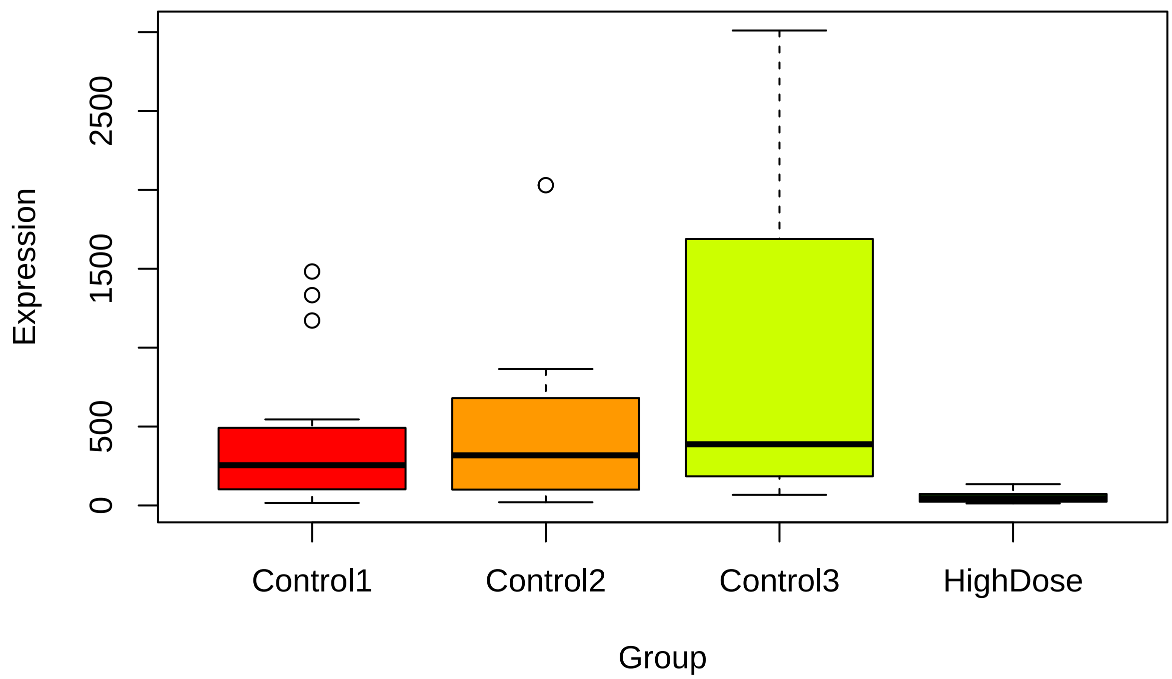
Table 1: Hint

Class	Index
$83.4 \le x < 85.5125$	4
$85.5125 \le x < 87.625$	6
$87.625 \le x < 89.7375$	20
$89.7375 \le x < 91.85$	30
$91.85 \le x < 93.9625$	14
$93.9625 \le x < 96.075$	4
$96.075 \le x < 98.1875$	2
$98.1875 \le x \le 100.3$	2

Problem 6.4.9

Based on my boxplot below, I think it's hard to definitively tell if the treatment is effective or not in gene expression. Because while it does have the smallest range - and practically no outliers - there's still a ton of difference in variance between the control groups. However, the variance and outliers in the High Dosage group itself definitely displayed a much smaller range compared to the other groups. So if I had to choose a definitive answer, I'd say that the high dosage treatment definitely helped in minimizing gene expression.

```
treatmentData = read.table("6-81.txt", header=TRUE)
attach(treatmentData)
boxplot(Expression~Group, col=rainbow(10))
```



Problem 6.7.2

Based on the normal distribution line below, it would be reasonable to assume that the octane rating follows a normal distribution. This is because most of our ratings/points follow the expected normal distribution line.

```r
qqnorm(rating)
qqline(rating, col="red")
```