A Glass Half Full: Analyzing a Dataset of Wine Review

Alexander Davies and Max Nadeau

5/9/2020

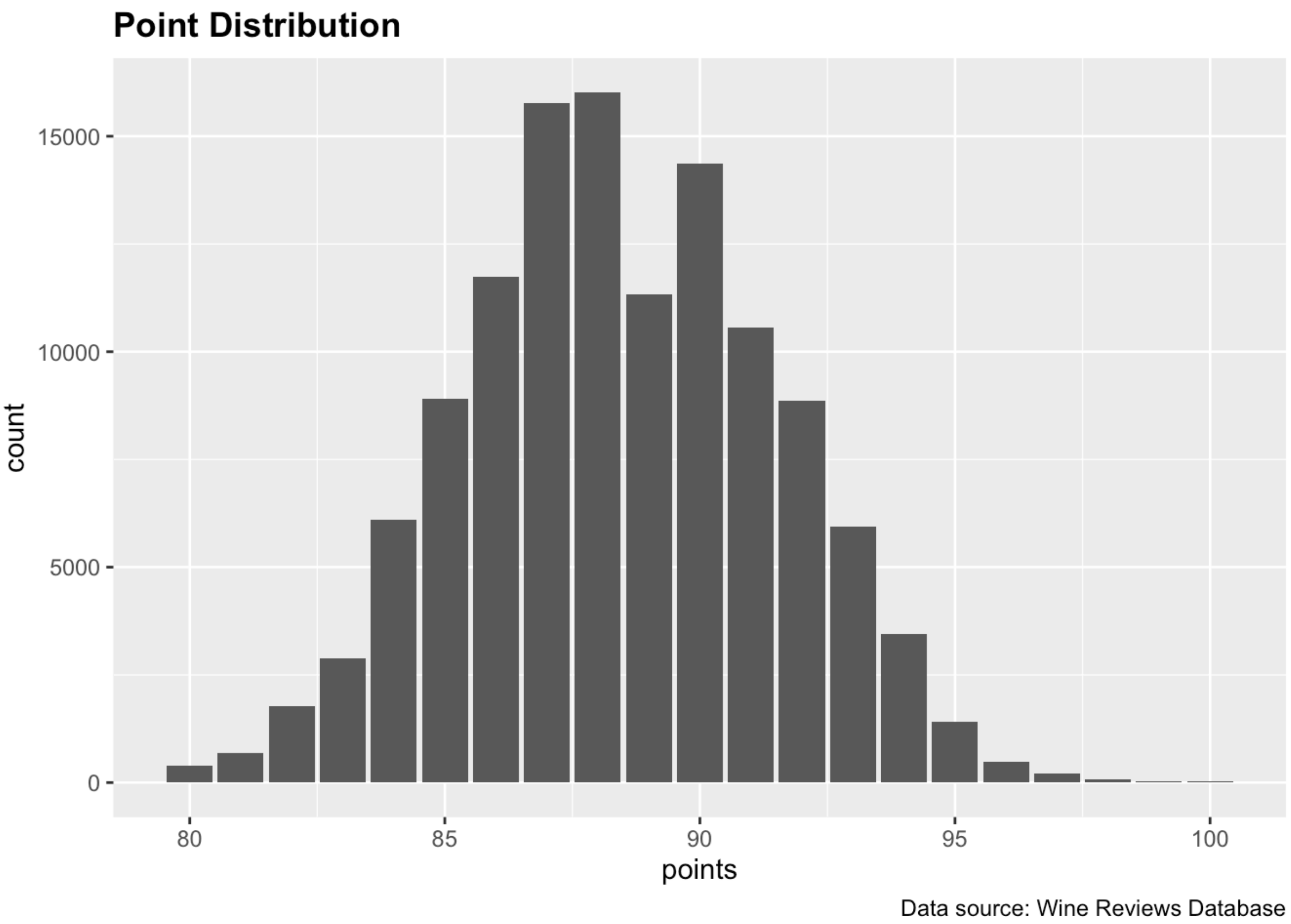
The Data

For our final project, we used the techniques learned in Math 23c to perform statistical analyses on a dataset of about 130,000 reviews of wine. The data came from Kaggle, where a user named zackthoutt had posted the results of scraping the website WineEnthusiast. Each row in the data frame has a points rating, ostensibly out of 100 but actually confined to between 80 and 100, a price in dollars, a region of origin, and the type of wine, among other columns. The points and price acted as our quantitative variables, while region of origin and variety functioned as categorical variables.

Point Distibutions

First, we made a barplot of the top 10 regions of origin for the wines in the dataset. California, with about 35,000 wines, was far ahead of all the rest. Washington State, in second place, had less than 10,000 wines. Outside of the United States, Bordeaux, Tuscany, and Northern Spain produced the most wine. Next, we generated a violin plot with the ggplot2 package to see how points rating varied by region of origin. The distributions were very similar, but the median rating for Bordeaux wines was, surprisingly, less than the other high-production regions. The overall distribution of wine ratings was roughly symmetrical, slightly skewed right and centered on around 88 points. We also interlayed a boxplot over the violin plots to perform some basic quartile analysis: The oscillateion of the density function is due to the integer-only rating system. The median of the Bordeaux province is was lower than the other four regions, whose medians were all 88, and its upper quartile was also significantly lower. California had the largests range of quartiles (especially lower quartile), suggesting a wider distribution of points (which is also visible in the density graph). Oregon, Tuscany, and Washington have nearly identical quartiles and medians, but the density function suggests Tuscany has fewer wines with very low ratings (a shorter bototm tail) than Oregon or Washington.

```r
library(ggplot2)
ggplot(wine, aes(x = points)) + geom_bar(stat = "count") +
  labs(title = "Point Distribution", caption = "Data source: Wine Reviews Database")
+
  theme(plot.title = element_text(face = "bold"))
```

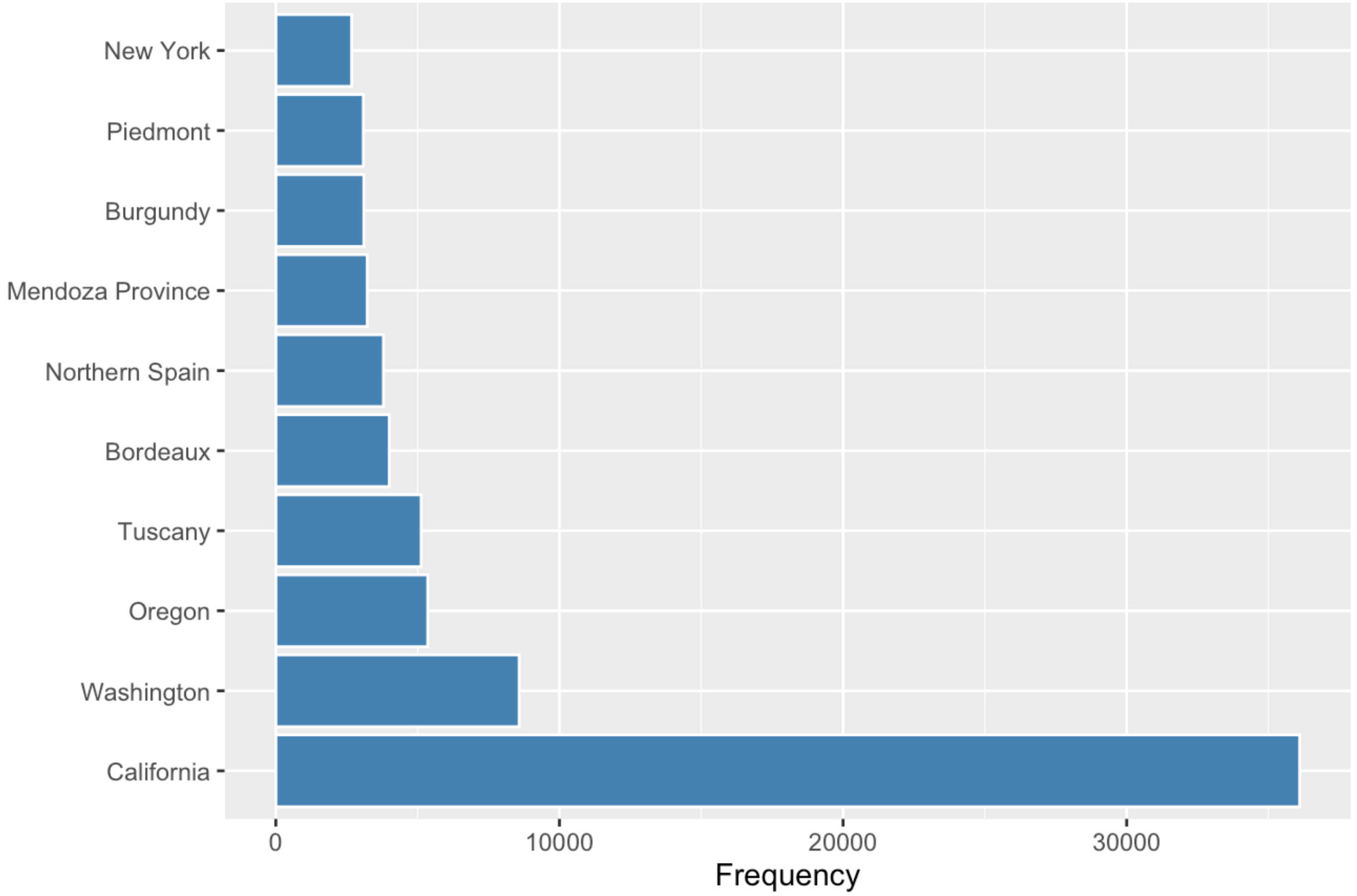


Point Distribution

Data source: Wine Reviews Database

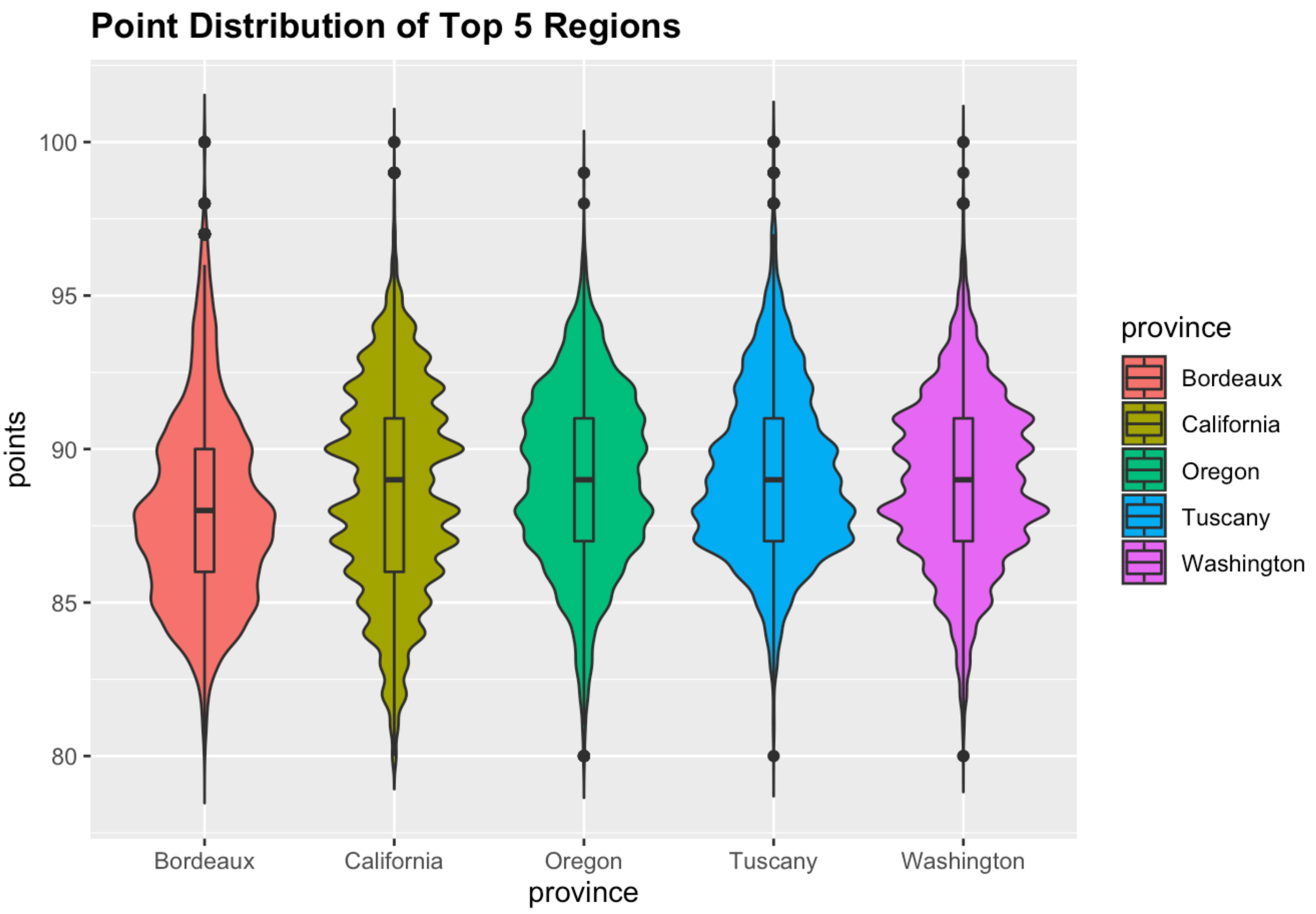
```
ten_wine_table <- sort(table(wine$province), decreasing = TRUE)[1:10]
ten_wine <- subset(wine, province %in% as.vector(as.list(as.data.frame(ten_wine_table
)["Var1"])$Var1))
ggplot(as.data.frame(ten_wine_table), aes(x = Var1, y = Freq)) +
  geom_bar(color = "white", fill = "steelblue", stat = "identity") +
  ylab("Frequency") + xlab("") +
  labs(title = "Top 10 Regions", caption = "Data source: Wine Reviews Database") +
  theme(plot.title = element_text(face = "bold")) + coord_flip()
```

Top 10 Regions



Data source: Wine Reviews Database

```
five_wine_table <- sort(table(wine$province), decreasing = TRUE)[1:5]
five_wine <- subset(wine, province %in% as.vector(as.list(as.data.frame(five_wine_tab
le)["Var1"])$Var1))
ggplot(five_wine, aes(x = province, y = points, fill=province)) +
  geom_violin(trim = FALSE) +
  stat_summary(fun.y=median, geom="point", shape=23, size=2) +
  geom_boxplot(width=0.1) +
  labs(title = "Point Distribution of Top 5 Regions", caption = "Data source: Wine Re
views Database") +
  theme(plot.title = element_text(face = "bold"))
```



Data source: Wine Reviews Database

Permutation Test

We then performed a permutation test to see if the difference in the average rating of wines from Southwest France, around 88.6 points, was statistically significant from the population mean, 88.4. Somewhat surprisingly, the permutation test produced a p-value of about 0.01, which made the difference statistically significant. As the CLT implies that the sampling distribution should be approximately normal, we confirmed this p-value with pnorm, the normal cdf, which also produced a p-value of around 0.01. Finally, we generated a 95% confidence interval for the average rating of wines from SW France, 88.5 to 88.8, which does not include the population mean.

```
# Permutation Test
# Testing the question: is wine from Southwest France statistically significantly
# different than the population of wines?
region = subset(wine, province == "Southwest France")

# The two have very close means, 88.60745 for SWF and 88.42188 for all wines
region_mean = mean(region$points); region_mean; mu
```

```
## [1] 88.63596
## [1] 88.42188
```

```
# However, the number of wines from SWF is large, 1503. Is that enough for
# such a small difference to be significant?
nrow(region)
```

```
## [1] 1335
```

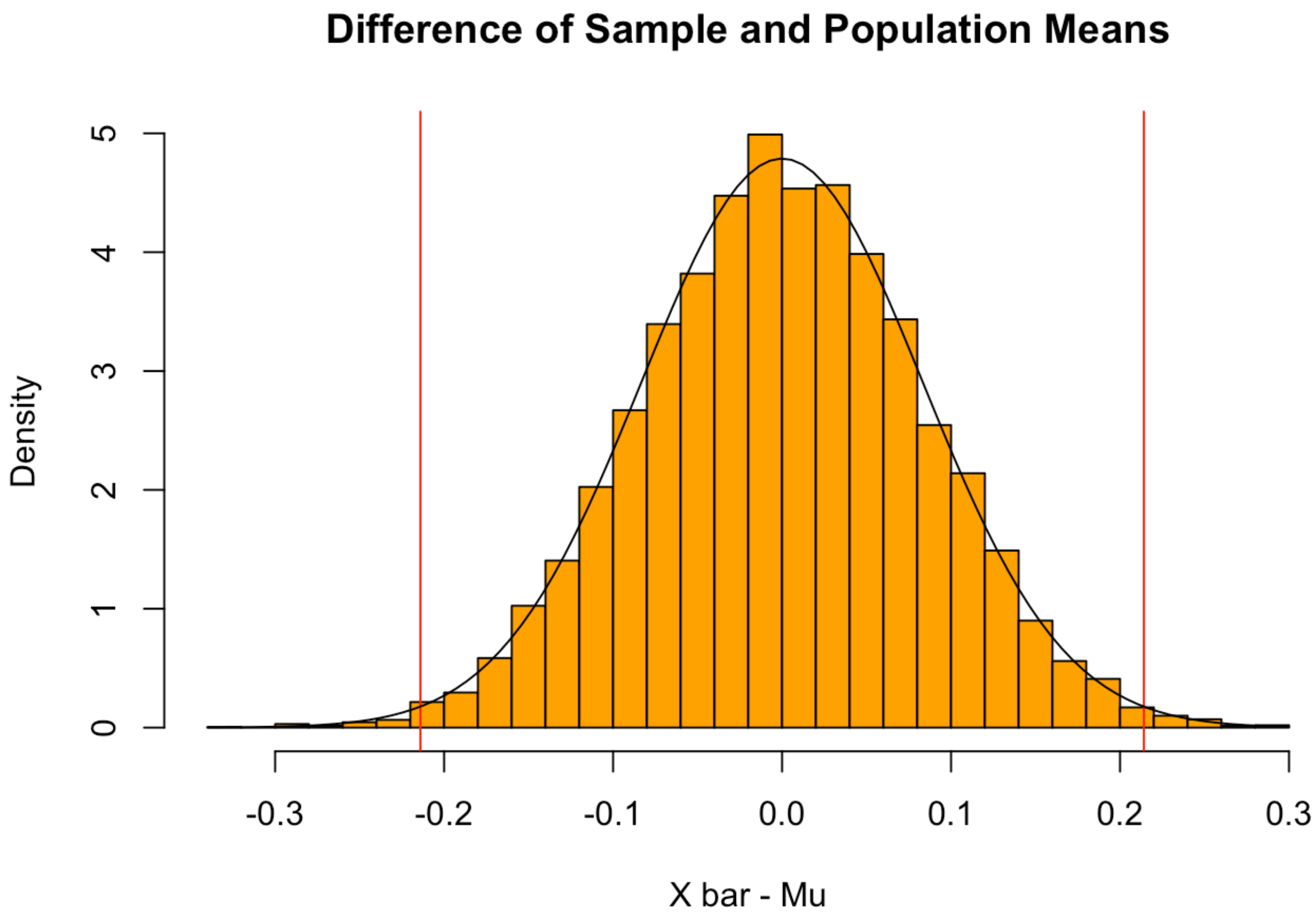
```
n = numeric(10000)
for(i in 1:length(n)){
  s = sample(1:nrow(wine), nrow(region))
  n[i] = mean(wine$points[s]) - mu
}
hist(n, breaks = 40, prob = TRUE, main = "Difference of Sample and Population Means",
xlab = "X bar - Mu", col = "orange")
abline(v = region_mean - mu, col = "red")
abline(v = mu - region_mean, col = "red")

p = mean(abs(n) > abs(region_mean - mu)); p # p = 0.0101 on the most recent run
```

```
## [1] 0.0092
```

```
# Significant at 0.05

# p-value based on a distribution function:
# By CLT, sample means should be normally distributed
sampdist_sd = sig/sqrt(nrow(region))
curve(dnorm(x, mean = 0, sd = sampdist_sd), add = TRUE)
```



```
p = pnorm(-abs(region_mean - mu), mean = 0, sd = sampdist_sd) + pnorm(abs(region_mean
- mu), mean = 0, sd = sampdist_sd, lower.tail = FALSE)

# The p-value calculations by the simulation method produced (on my specific run)
# a p-value of 0.0101, which is quite close to the classically-calculated p-value
# of 0.0102. Success!


# 95% Confidence interval of SWF points does not contain the population mean
z = abs(qnorm(0.025))
region_mean - z * sampdist_sd; region_mean + z * sampdist_sd
```

```
## [1] 88.47264
```

```
## [1] 88.79927
```

```
mu
```

```
## [1] 88.42188
```

Contingency Table

We then performed a permutation test to see if the difference in the average rating of wines from Southwest France, around 88.6 points, was statistically significant from the population mean, 88.4. Somewhat surprisingly, the permutation test produced a p-value of about 0.01, which made the difference statistically significant. As the CLT implies that the sampling distribution should be approximately normal, we confirmed this p-value with pnorm, the normal cdf, which also produced a p-value of around 0.01. Finally, we generated a 95% confidence interval for the average rating of wines from SW France, 88.5 to 88.8, which does not include the population mean.

```
# Contingency Table
cali = wine$province == "California"
chard = wine$variety =="Chardonnay"
mean(cali)
```

```
## [1] 0.2984418
```

```
mean(chard)
```

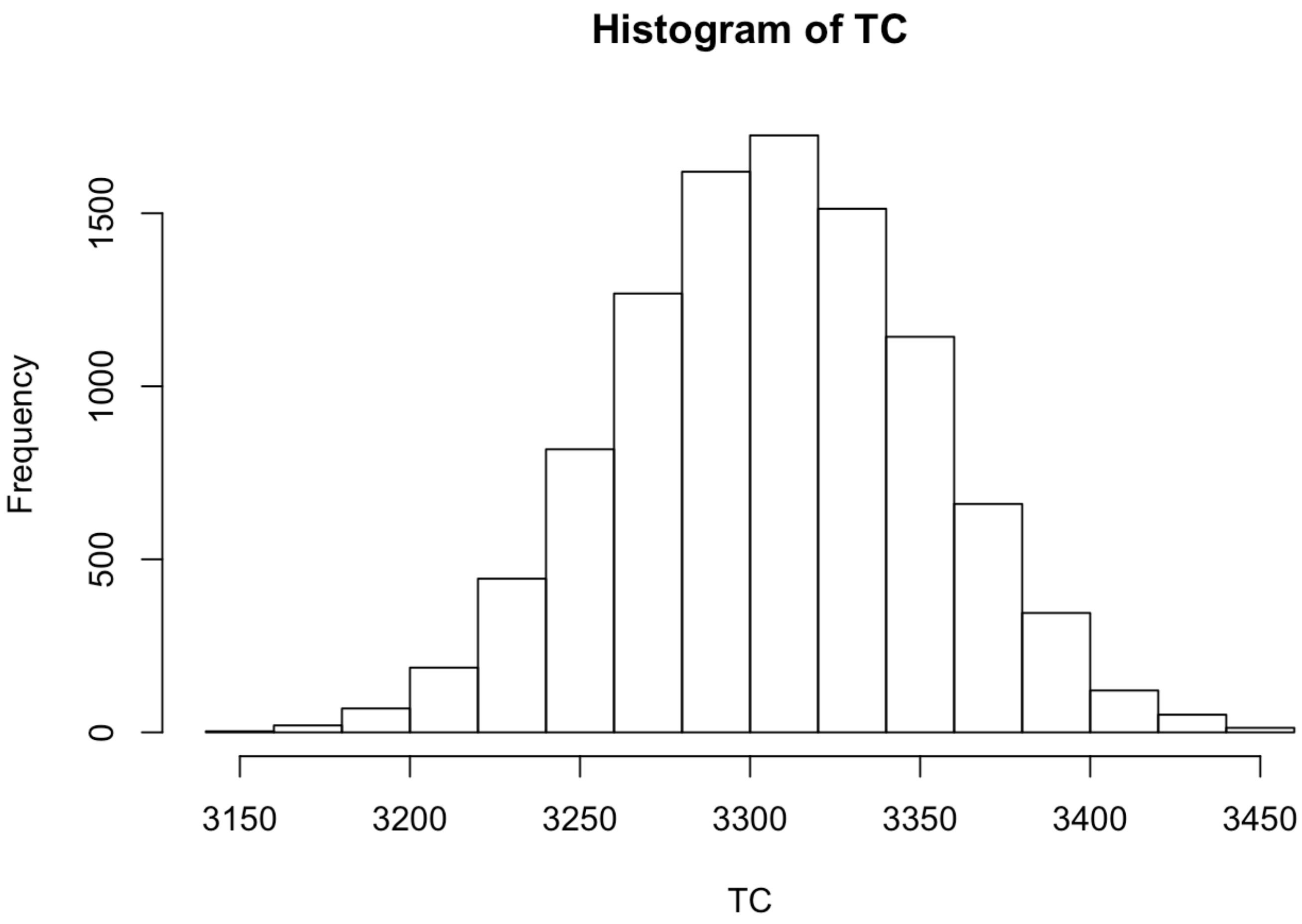
```
## [1] 0.09158917
```

```
table(chard, cali)
```

```
##        cali
## chard   FALSE  TRUE
##   FALSE 78948 30947
##   TRUE   5923  5157
```

```
#we do a simulation.
N <- 10000; TC <- numeric(N);
for (i in 1:N){
  scramble <- sample(chard, length(chard), replace = FALSE)
  TC[i] <- sum(cali&scramble)
}

hist(TC, breaks = 20)
```



```
sum(cali&chard) # off the charts, p ~=~ 0
```

```
## [1] 5157
```

```
fisher.test(chard,cali, alternative = "g") # p-value < 2.2 * 10 ^ (-16)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  chard and cali
## p-value < 2.2e-16
## alternative hypothesis: true odds ratio is greater than 1
## 95 percent confidence interval:
##  2.148155      Inf
## sample estimates:
## odds ratio
##   2.221155
```

```
# Very significant!
```

Regression

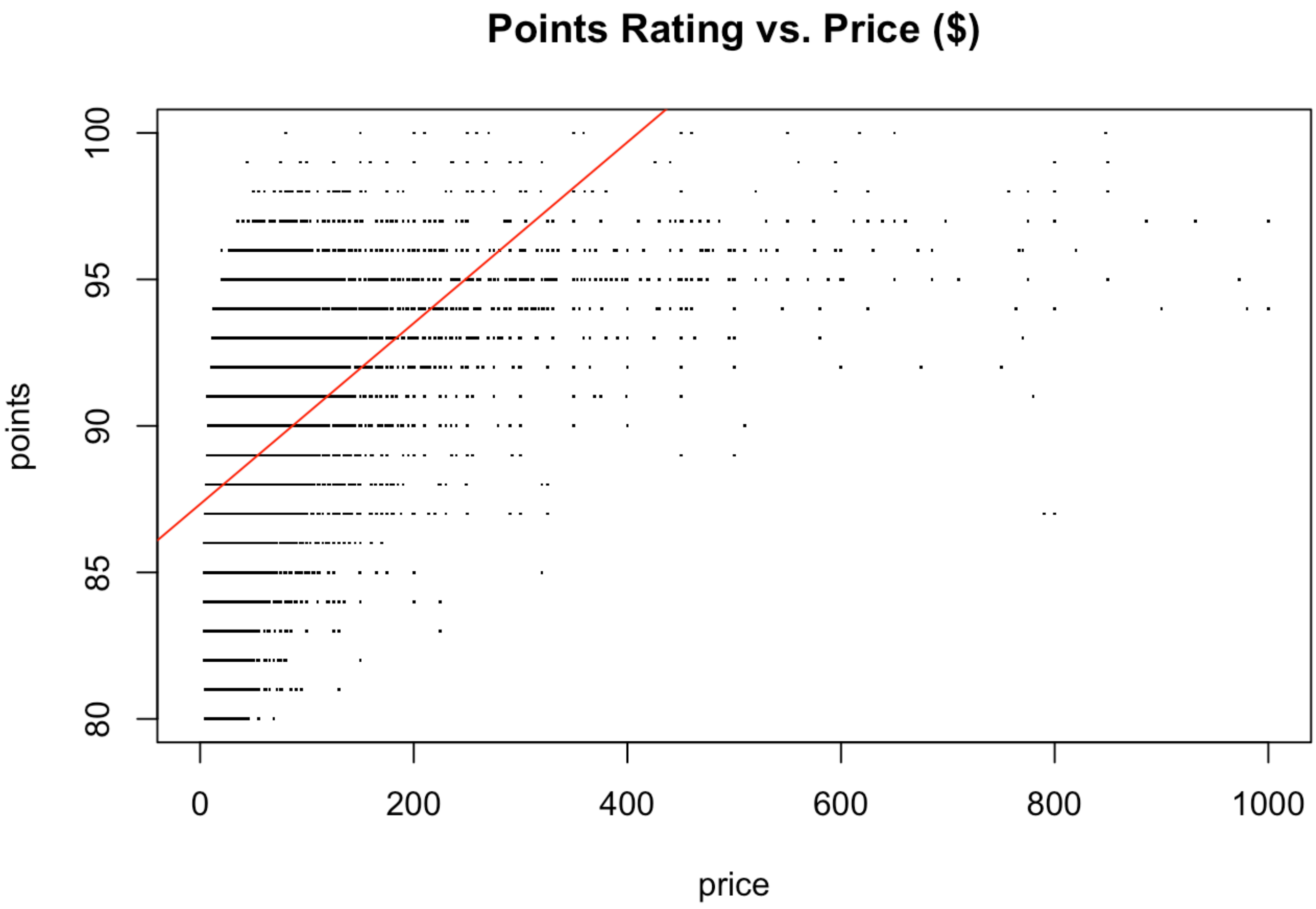
Finally, we regressed point ratings against prices to see if more expensive ratings were given higher prices. Using a linear model, the price of wine is positively correlated with the rating but only explains 17% of the variation in price, indicating that many factors besides price impact the rating of a wine.

```
# Linear regression
price = wine$price
plot(price, points, xlim = c(0,1000), pch = ".", main = "Points Rating vs. Price ($)"
)

lm <- lm(points~price);lm
```

```
##
## Call:
## lm(formula = points ~ price)
##
## Coefficients:
## (Intercept)        price
##    87.32964      0.03089
```

```
coef = lm$coefficients
abline(coef[1], coef[2], col = "red")
```



```
pred = coef[1] + coef[2] * price
r = sqrt(var(pred) / var(points))
r^2
```

```
## [1] 0.1731948
```

```
# r^2 = 0.173, so 17% of variation in points is explained by price
summary(lm) # R^2 = 0.173, confirmed
```

```
##
## Call:
## lm(formula = points ~ price)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -101.254   -1.793    0.053    2.003   10.311
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 87.329638   0.010509  8310.4   <2e-16 ***
## price        0.030886   0.000194   159.2   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.768 on 120973 degrees of freedom
## Multiple R-squared:  0.1732, Adjusted R-squared:  0.1732
## F-statistic: 2.534e+04 on 1 and 120973 DF,  p-value: < 2.2e-16
```

```
r # 0.416, a moderate correlation
```

```
## [1] 0.4161667
```