Homework 9

Bill Ang Li

March 27th, 2019

9.1.1

Suppose the following sample is assumed to be from an $N(\theta,4)$ distribution with $\theta \in \mathbb{R}^1$ unknown.

```
data = c(1.8, 2.1, -3.8, -1.7, -1.3, 1.1, 1.0, 0.0, 3.3, 1.0, -0.4, -0.1, 2.3, -1.6,
         1.1, -1.3, 3.3, -4.9, -1.1, 1.9)
length(data)
```

```
## [1] 20
```

```
mean(data)
```

```
## [1] 0.135
```

```
var(data)
```

```
## [1] 4.791868
```

```
residuals = data - mean(data)
residuals
```

```
##  [1]  1.665  1.965 -3.935 -1.835 -1.435  0.965  0.865 -0.135  3.165  0.865
## [11] -0.535 -0.235  2.165 -1.735  0.965 -1.435  3.165 -5.035 -1.235  1.765
```

```
var = 4
discrepancy = sum((data ^ 2) / var)
discrepancy
```

```
## [1] 22.8525
```

```
alpha = 0.95
qchisq(alpha, df = length(data) - 1)
```

```
## [1] 30.14353
```

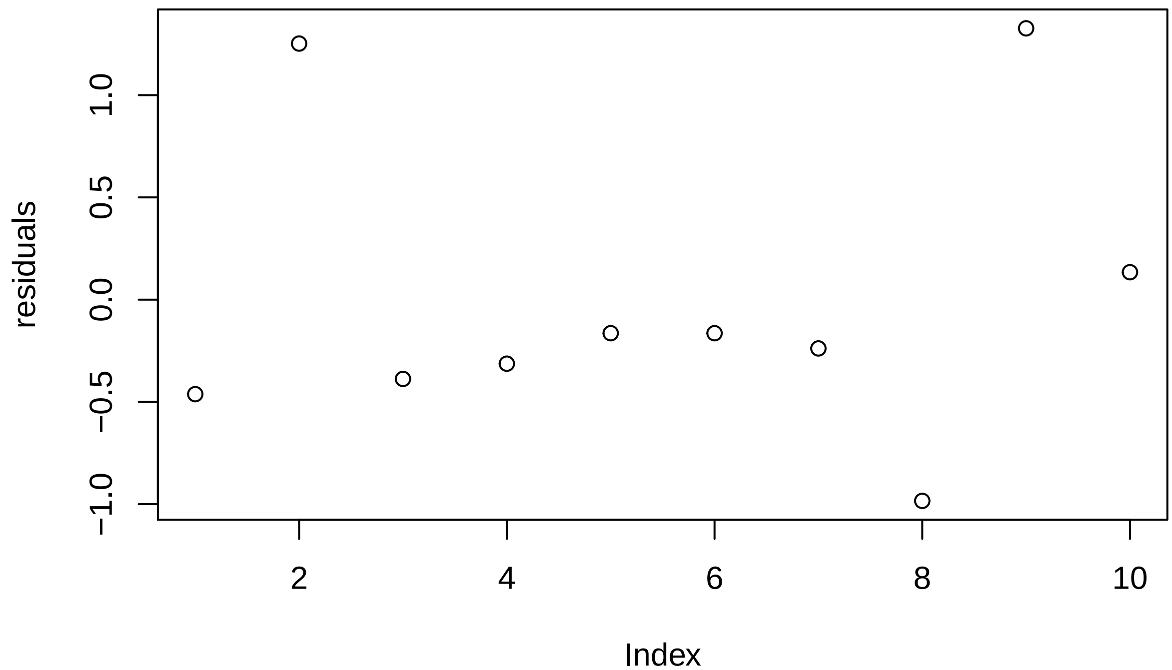
Since the discrepancy is lower than the chi-square value for alpha = 0.95 and df = 19, we fail to reject that $N(\theta, 4)$ is a good model for this set of data.

9.1.2

Suppose the following sample is assumed to be from an $N(\theta, 2)$ distribution with θ unknown.

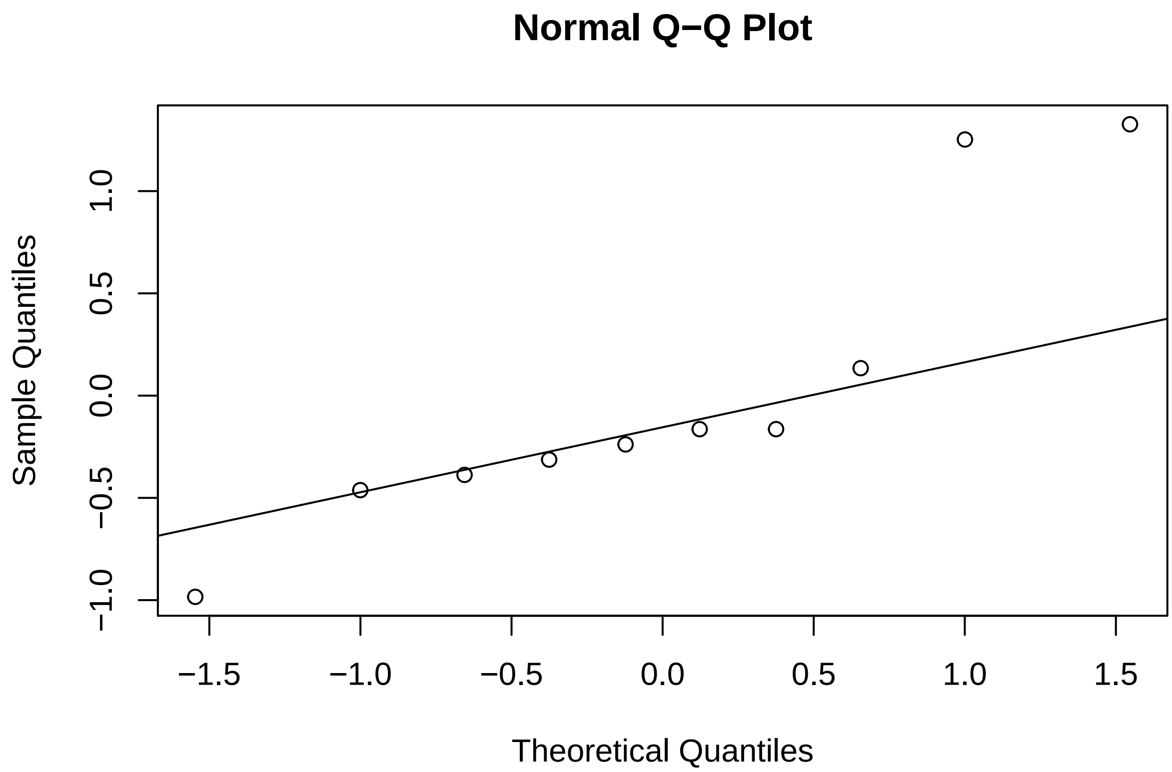
a. Plot the standardized residuals.

```
data = c(-0.4, 1.9, -0.3, -0.2, 0.0, 0.0, -0.1, -1.1, 2.0, 0.4)
var = 2
residuals = (data - mean(data)) / sqrt(var * (1 - 1 / length(data)))
plot(residuals)
```



b. Construct a normal probability plot of the standardized residuals.

```
qqnorm(residuals)
qqline(residuals)
```



c. What conclusions do you draw based on the results of parts (a) and (b)?

There may be outliers in this data, i.e. the two points with very large residuals. The data seems to have a larger tail than the model.

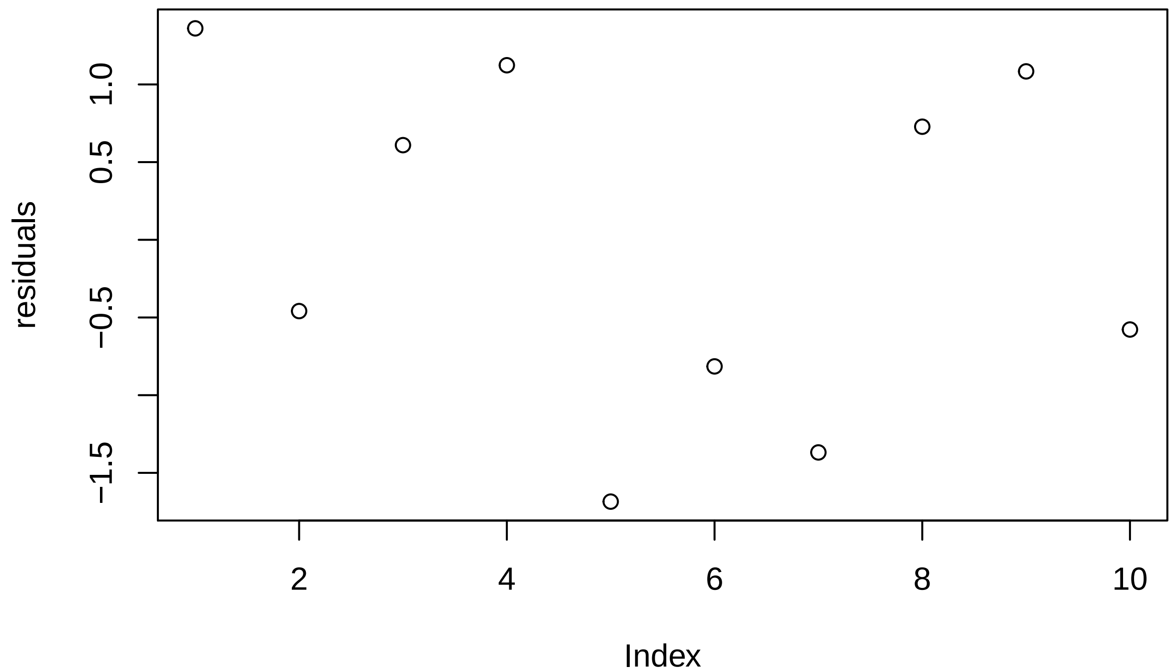
9.1.3

Suppose the following sample is assumed to be from an $N(\mu, \sigma^2)$ distribution, where $\sigma \in \mathbb{R}^1$ and $\sigma > 0$ are unknown.

```
data = c(14.0, 9.4, 12.1, 13.4, 6.3, 8.5, 7.1, 12.4, 13.3, 9.1)
```

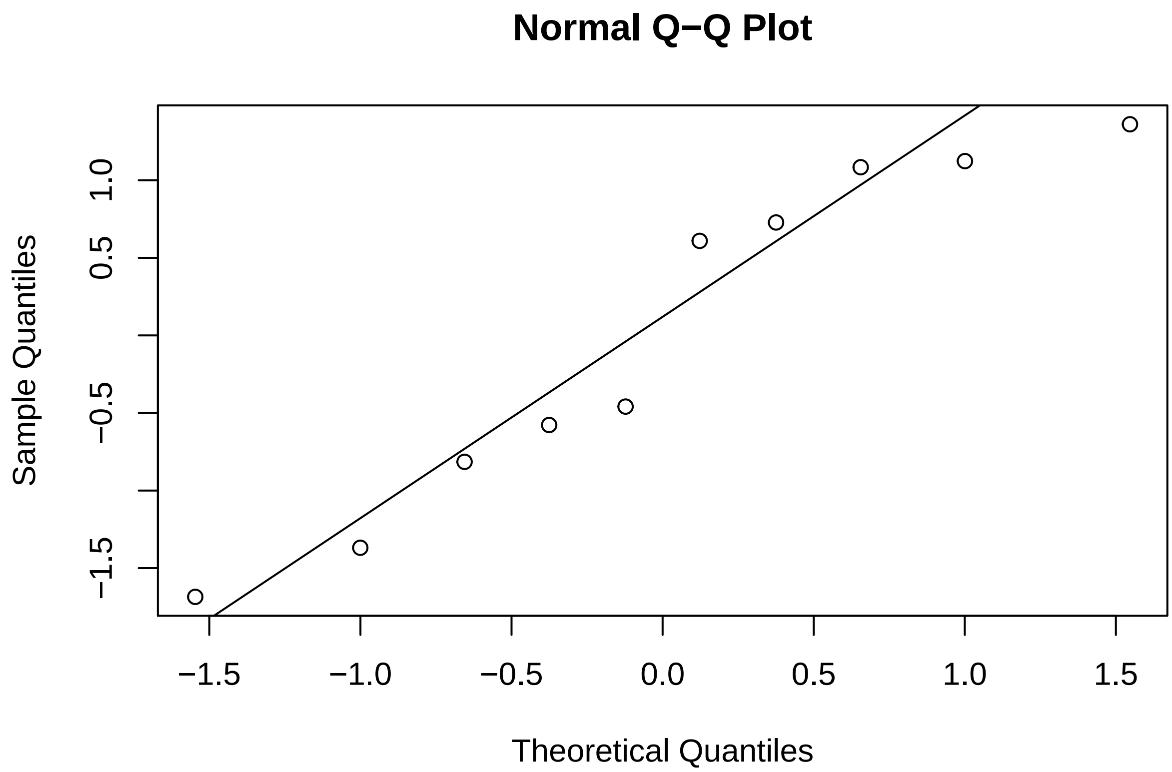
a. Plot the standardized residuals.

```
residuals = (data - mean(data)) / (sqrt(var(data)) * (1 - 1 / length(data)))
plot(residuals)
```



b. Construct a normal probability plot of the standardized residuals.

```r
qqnorm(residuals)
qqline(residuals)
```



c. What conclusions do you draw based on the results of parts (a) and (b)?

This normal model seems like a good fit. Perhaps the rightmost point in the normal probability plot is an outier. It's sample quantile is too small for its theoretical value.

9.1.5

The following sample of n=20 is supposed to be from a Uniform[0,1] distribution. After grouping the data, using a partition offive equal-length intervals, carry out thechi-squared goodness offit test to assess whether or not we have evidence against thisassumption. Record the standardized residuals.

```
data = c(0.11, 0.56, 0.72, 0.18, 0.26, 0.32, 0.42, 0.22, 0.96, 0.04, 0.45, 0.22, 0.08,
         0.65, 0.32, 0.88, 0.76, 0.32, 0.21, 0.80)
sort(data)
```

```
##  [1] 0.04 0.08 0.11 0.18 0.21 0.22 0.22 0.26 0.32 0.32 0.32 0.42 0.45 0.56
## [15] 0.65 0.72 0.76 0.80 0.88 0.96
```

```
alpha = 0.95
cutoff = qchisq(alpha, df = 4)
p = 1 - pchisq(3.5, df = 4)
p
```

```
## [1] 0.4778783
```

The p-value is very high, so we fail to reject that Uniform[0, 1] is a good model.

9.1.15

Using software, generate a sample of n=1000 from the Binomial(10, 0.2) distribution. Then, using the chi-squared goodness of fit test, check that this sample is indeed from this distribution. Use grouping to ensure $E(X_i) = np_i \geq 1$. What would you conclude if you got a P-value close to 0?

```
n = 1000
data = rbinom(n, 10, 0.2)
tbl = table(data)
tbl
```

```
## data
##   0   1   2   3   4   5   6   7
## 113 278 306 195  75  25   7   1
```

```
binom_exp = function(num) {
  if (num == 0) {
    return(pbinom(0, 10, 0.2))
  } else {
    return(pbinom(num, 10, 0.2) - pbinom(num - 1, 10, 0.2))
  }
}

max = max(data)
count = 0
sum = 0
while (count <= max) {
  exp = binom_exp(count)
  sum = sum + (tbl[names(tbl) == count] - exp * n) ^ 2 / (exp * n)
  count = count + 1
}
p = 1 - pchisq(sum, df=length(unique(data)))
p
```

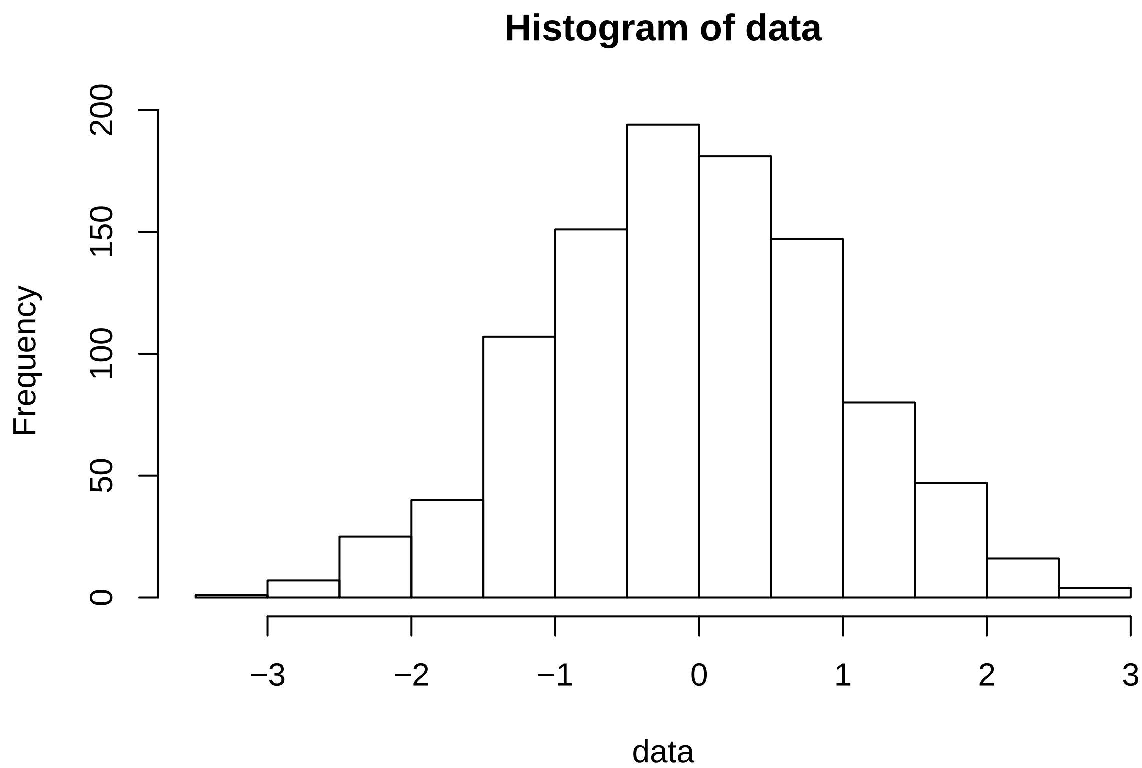
```
##         0
## 0.9089785
```

The p-value shouldn't be too low because we are drawing from a binomial distribution.

9.1.17

Using a statistical package, generate a sample of n=1000 from the N(0, 1) distribution. Then, using the chi-squared goodness of fit test based on grouping the observations in to five cells chosen to ensure $E(X_i) = np_i \geq 1$, check that this sampleis indeed from this distribution. What would you conclude if you got a P-value close to 0?

```
n = 1000
data = rnorm(n, 0, 1)
hist(data)
```



Find the quintiles to find the observed counts per bin.

```
q0 = qnorm(0)
q1 = qnorm(0.2)
q2 = qnorm(0.4)
q3 = qnorm(0.6)
q4 = qnorm(0.8)
q5 = qnorm(1)

obs = c(
  sum(data > q0 & data <= q1),
  sum(data > q1 & data <= q2),
  sum(data > q2 & data <= q3),
  sum(data > q3 & data <= q4),
  sum(data > q4 & data <= q5)
)
obs
```

```
## [1] 220 200 200 194 186
```

Find the chi-square test statistic value.

```
exp = 0.2 * n
test_statistic = sum((obs - exp) ^ 2) / exp
p = 1 - pchisq(test_statistic, df = 4)
p
```

```
## [1] 0.5314158
```

The p-value is very high, as expected.

Likelihood Ratio Test Example

The two distributions are $N(\mu_x, 3)$ and $N(\mu_y, 4)$

```
x = c(16.27, 11.66, 14.05, 15.43, 18.74, 13.42, 17.39, 18.71, 11.18, 13.52, 16.74,
      5.43, 16.45, 10.75, 19.06)
y = c(10.89, 7.57, 15.39, 8.43, 12.33, 7.43, 5.56, 18.07, 0.35, 7.62)

x_bar = mean(x)
y_bar = mean(y)
mu_bar = mean(c(x, y))
mu_bar
```

```
## [1] 12.4976
```

Rice 9.36

The National Center for Health Statistics (1970) gives the following data on distribution of suicides in the United States by month in 1970. Is there any evidence that the suicide rate varies seasonally, or are the data consistent with the hypothesis that the rate is constant? (Hint: Under the latter hypothesis, model the number of suicides in each month as a multinomial random variable with the appropriate probabilities and conduct a goodness-of-fit test. Look at the signs of the deviations, $O_i - E_i$, and see if there is a pattern.)

```
suicides = c(1867,1789,1944,2094,2097,1981,1887,2024,1928,2032,1978,1859)
days = c(31,28,31,30,31,30,31,31,30,31,30,31)
suicides_per_day = sum(suicides) / sum(days)
exp = days * suicides_per_day
suicides - exp
```

```
##  [1] -127.191781  -12.205479  -50.191781  164.136986  102.808219
##  [6]   51.136986 -107.191781   29.808219   -1.863014   37.808219
## [11]   48.136986 -135.191781
```

```
test_statistic = sum((suicides - exp) ^ 2 / exp)
p = 1 - pchisq(test_statistic, df=11)
p
```

```
## [1] 1.852011e-06
```

We have evidence to reject that the suicide rate is not constant. There are some months like April and May that have higher rates than the other months.

Rice 9.37

The following table gives the number of deaths due to accidental falls for each month during 1970. Is there any evidence for a departure from uniformity in the rate over time? That is, is there a seasonal pattern to this death rate? If so, describe its pattern and speculate as to causes.

```
deaths = c(1668,1407,1370,1309,1341,1338,1406,1446,1332,1363,1410,1526)
days = c(31,28,31,30,31,30,31,31,30,31,30,31)
deaths_per_day = sum(deaths) / sum(days)
exp = days * deaths_per_day
deaths - exp
```

```
##  [1] 231.29863 109.33425 -66.70137 -81.35616 -95.70137 -52.35616 -30.70137
##  [8]   9.29863 -58.35616 -73.70137  19.64384  89.29863
```

```
test_statistic = sum((deaths - exp) ^ 2 / exp)
p = 1 - pchisq(test_statistic, df=11)
p
```

```
## [1] 1.122225e-11
```

We have evidence to reject that the death rates does not have a seasonal pattern. The data shows that from Dec. to Feb., there are a lot more deaths than on average.

Rice 9.43

a. In 1965, a newspaper carried a story about a high school student who reported getting 9207 heads and 8743 tails in 17,950 coin tosses. Is this a significant discrepancy from the null hypothesis $H_0 : p = 12$?

```
tosses = c(9207, 8743)
total = 17950
exp = c(total * 0.5, total * 0.5)
test_statistic = sum((tosses - exp) ^ 2 / exp)
p = 1 - pchisq(test_statistic, df=1)
p
```

```
## [1] 0.000533662
```

Yes, this is a significant discrepancy from the null hypothesis.

b. Jack Youden, a statistician at the National Bureau of Standards, contacted the student and asked him exactly how he had performed the experiment (Youden 1974). To save time, the student had tossed groups of five coins at a time, and a younger brother had recorded the results, shown in the following table:

The index is the number of coins that came up heads.

```
freq = c(100,524,1080,1126,655,105)
freq
```

```
## [1]  100  524 1080 1126  655  105
```

```
exp = c(
  0.5 ^ 6,
  choose(5, 1) * 0.5 ^ 6,
  choose(5, 2) * 0.5 ^ 6,
  choose(5, 3) * 0.5 ^ 6,
  choose(5, 4) * 0.5 ^ 6,
  0.5 ^ 6
)
exp = exp * total / 5 * 2
test_statistic = sum((freq - exp) ^ 2 / exp)
p = 1 - pchisq(test_statistic, df=5)
p
```

```
## [1] 0.0006323943
```

The data rejects the hypothesis that the coins are fair.