Statistical methods: Homework 8

Cameron McIntyre

October 22, 2018

1 5.2.4

Suppose a random sample of size n is drawn from the probability model

$$p_X(k;\theta) = \frac{\theta^{2k}e^{-\theta^2}}{k!},\ k = 0, 1, 2, ...$$

Find a formula for the maximum likelihood estimator, $\hat{\theta}$ **Answer:**

$$L_p(\theta) = \Pi\frac{\theta^{2k}e^{-\theta^2}}{k!} = \frac{\theta^{2n\sum k}e^{-n\theta^2}}{k!}$$
$$ln(L_p(\theta)) = 2n\sum kln(\theta) - n\theta^2 ln(e)$$

We differentiate and set it to 0,

$$\frac{d}{d\theta}ln(L_p(\theta)) = \frac{2n\sum k}{\theta} - 2n\theta^2 ln(e) = 0 \leftrightarrow n\theta^2 = \sum k$$

$$\hat{\theta} = \sqrt{\bar{K}}$$

2 5.2.8

The following data show the number of occupants in passenger cars observed during one hour at a busy intersection in Los Angeles (75). Suppose it can be assumed that these data follow a geometric distribution $p_X(k;p) = (1-p)^{k-1}p,\ k = 1, 2, ...$ Estimate p and compare the observed and expected frequencies for each value of X.

Number of Occupants	Frequency
1	678
2	227
3	56
4	28
5	8
6+	14
	1011

Answer:

$$L_p(p) = \Pi_{i=1}^{k}(1-p)^{k-1}p = (1-p)^{\sum k - n}p^n$$

$$\frac{d}{d\theta}ln(L_p(p)) = \frac{d}{d\theta}(\sum k_i - n)ln(1-p) + nln(p) = \frac{-\sum k_i + n}{1-p} + \frac{n}{p}$$

Set it to 0,

$$0 = \frac{-\sum k_i + n}{1-p} + \frac{n}{p} \leftrightarrow \hat{p} = \frac{\sum k_i}{n}$$

$$\hat{p} = \frac{1011}{1*678 + 2*227 + 3*56 + 4*28 + 5*8 + 6*14} = \frac{1011}{1536} = .658$$

Comparing to the data.

Number of Occupants	Frequency	Predicted Amount
1	678	665
2	227	227
3	56	78
4	28	27
5	8	9
6+	14	3
	1011	

3 5.2.12

A random sample of size n is taken from the pdf

$$f_Y(y;\theta) = \frac{2y}{\theta^2},\ 0 \leq y \leq \theta$$

Find an expression for $\hat{\theta}$, the maximum likelihood estimator for θ.
Answer:

$$L_p(\theta) = \Pi_{i=0}^{n}\frac{2y}{\theta^2} = \frac{2^n y^n}{\theta^{2n}} \leftrightarrow \frac{d}{d\theta}L_p(\theta) = -2n\frac{2^n y^n}{\theta^{2n-1}}$$

Setting this to 0,

$$2n\frac{2^n y^n}{\theta^{2n-1}} = 0$$

this expression is maximized when $\theta \to \infty$. Therefore our maximum likelihood estimator is $\hat{\theta} = Y_{max}$.

4 5.2.22

Find a formula for the method of moments estimate for the parameter θ in the Pareto pdf,

$$f_Y(y;\theta) = \theta k^\theta \left(\frac{1}{y}\right)^{\theta+1}, \; y \geq k; \; \theta \geq 1$$

Assume that k is known and that the data consist of a random sample of size n. Compare your answer to the maximum likelihood estimator found in Question 5.2.13.

Answer:

$$E[Y] = \int_k^\infty y\theta k^\theta \frac{1}{y}^{\theta+1} = \theta k^\theta \int_k^\infty \left(\frac{1}{y}\right)^\theta = \theta k^\theta \left[0 + \frac{k\frac{1}{k}^\theta}{\theta - 1}\right]$$

$$E[Y] = \frac{\theta k}{\theta - 1}$$

Using Method of Moments:

$$\bar{Y} = \frac{\theta k}{\theta - 1} \leftrightarrow \hat{\theta} = \frac{\bar{y}}{\bar{y} - k}$$

Now we need to find the M.L.E.:

$$L_p(\theta) = \Pi\theta k^\theta \frac{1}{y_i}^{\theta+1} = \theta^n k^{n\theta} (\Pi y_i)^{-(\theta+1}$$

Taking log and derivative:

$$\frac{d}{d\theta} ln(L_p(\theta)) = \frac{n}{\theta} + nln(k) - nln(\Pi y_i)$$

Set it to 0,

$$\frac{n}{\theta} + nln(k) - nln(\Pi y_i) = 0 \leftrightarrow \hat{\theta} = \frac{n}{ln(\Pi y_i) - nln(k)}$$

The Method of moments estimator $\hat{\theta} = \frac{\bar{y}}{\bar{y}-k}$ is not equal to the maximum likelihood estimator $\hat{\theta} = \frac{n}{ln(\Pi y_i) - nln(k)}$.

5 5.3.10

In 1927, the year he hit sixty home runs, Babe Ruth batted .356, having collected 192 hits in 540 official at-bats (150). Based on his performance that season, construct a 95% confidence interval for Ruth?s probability of getting a hit in a future at-bat.

Answer:

$$\hat{\mu} = .356$$

$$\hat{\sigma} = sqrt540 * .356 * .644 = 8.25$$

Therfore our confidence interval is,

$$.356 + -1.96 * \sqrt{\frac{.356(1 - .356)}{540}}$$

$$(.3156, .3964)$$

6 5.3.14

If $(0.57, 0.63)$ is a 50% confidence interval for p, what does $\frac{k}{n}$ equal, and how many observations were taken?

Answer:

$$(1)\ \frac{k}{n} + .67\sqrt{\frac{\frac{k}{n}(1-\frac{k}{n})}{n}} = .63$$

$$(2)\ \frac{k}{n} - .67\sqrt{\frac{\frac{k}{n}(1-\frac{k}{n})}{n}} = .57$$

Add 1 to 2

$$\frac{k}{n} * 2 = 1.2 \leftrightarrow = \frac{k}{n} = \frac{1.2}{2} = .6$$

Now we substitute to find n.

$$.6 + .67\sqrt{\frac{.6(.4)}{n}} = .63$$

$$n = 10.88^2 = 119$$

7 5.3.26

Suppose that p is to be estimated by $\frac{X}{n}$ and we are willinng to assume that the true p will not be greater than .4. What is the smallest n for which $\frac{X}{n}$ will have a 99% probability of being within 0.05 of p?

Answer: The formula for n is:

$$n = \frac{z_{frac\alpha 2}}{d^2} r_1(1-r_1) = \frac{2.58^2}{.05^2} .4 * .6 = 640$$

8 5.4.7

Let Y be the random variable described in Example 5.2.4, where $f_Y(y;\theta) = e^{-(y-\theta)}, y \geq \theta, \theta > 0$. Show that $Y_{min} - \frac{1}{n}$ is an unbiased estimator of θ.

Answer:

We can find the distrubution of $Y_{min} = n(1 - F_Y(y)^n) f_y(y)$

$$F_Y(y) = 1 - e^{-(y-\theta)} \leftrightarrow\ P(Y > y) = e^{-(y-\theta)}$$

So,

$$f_{Y_{min}}(y) = n(e^{-(y-\theta)^n})$$

And,

$$E[Y_{min}] = \int_\theta^\infty yne^{-(y-\theta)^n}$$

Substitute $u = y - \theta, du = dy, y = u + \theta$.

$$E[y_{min}] = n\int_0^\infty (u+\theta)e^{nu}du = n\Big[u\frac{e^{-nu}}{-n}\Big|_0^\infty - \int_0^\infty \frac{e^{-nu}}{n}du\Big] + n\theta\frac{e^{-nu}}{n}{}_0^\infty$$

$$= e_n^{-nu}\theta() + \frac{1}{n} = 0 + \frac{1}{n} + \theta = \frac{1}{n} + \theta$$

So,

$$E[Y_{min} - \frac{1}{2}] = E[Y_{min}] - -E[\frac{1}{2}] = \frac{1}{2} + \theta - \frac{1}{2} = \theta$$

Thus, $[Y_{min} - \frac{1}{2}]$ is an unbiased estimator for θ.

9 5.4.20

Given a random sample of size n from a Poisson distribution, $\hat{\lambda_1} = X_1$ and $\hat{\lambda_2} = \bar{X}$ are two unbiased estimators for λ. Calculate the relative efficiency of $\hat{\lambda_1}$ to $\hat{\lambda_2}$. **Answer:**

$$E[\hat{\lambda_1}] = E[X_1] = \lambda$$

Therefore $\hat{\lambda_1}$ is unbiased.

$$E[\hat{\lambda_2}] = E[\bar{X}] = E[\frac{\sum X_i}{n}] = \frac{1}{n} \cdot nE[X] = \lambda$$

Therefore $\hat{\lambda_2}$ is unbiased.

$$Var[\hat{\lambda_1}] = Var[X_1] = \lambda$$

$$Var[\hat{\lambda_2}] = Var[\bar{X}] = \frac{1}{n^2} nVar[X] = \frac{\lambda}{n}$$

The efficiency of the two estimators is the ratio of the variance.

$$Efficiency\ ratio\ \hat{\lambda_1}\ to\ \hat{\lambda_2} = \frac{\frac{\lambda}{n}}{\lambda} = \frac{1}{n}$$

10 5.6.6

Let $Y_1, Y_2, ..., Y_n$ be a random sample of size n from the pdf

$$f(y;\theta) = \theta y^{\theta-1}, 0 \leq 1 \leq 1$$

Use theorem 5.6.1 to show that $W = \Pi_{i=1}^n Y_i$ is a sufficient statistic for θ. Is the maximum likelihood estimator of θ a function of W.

Answer:

$$\Pi_{i=1}^n \theta y^{\theta-1} = \theta^n (\Pi y_i)^{\theta-1}$$

By theorem 5.6, we know that $\hat{\theta}$ is a sufficient statistics if and only if there exists functions $g(h(x_1, x_2...x_n, \theta))b(x_1, x_2...x_n) = L_p(\theta)$. For our situation set $g(h(x_1, x_2...x_n)) = \theta^n(\Pi y_i)^{\theta-1}$ and we can use the constant function and set it to $b(x_1, x_2...x_n) = 1$. Therefore Πy_i is a sufficient statistic.

Now we find the M.L.E.

$$\frac{d}{d\theta} ln(L_p(\theta)) = \frac{d}{d\theta} ln(\theta^n(\Pi y_i)^{\theta-1}) = \frac{n}{\theta} + ln(\Pi i y_i)$$

Setting this to 0.

$$\frac{n}{\theta} + \Pi y_i = 0 \leftrightarrow \hat{\theta} = \frac{n}{ln(\Pi y_i)} \leftrightarrow \hat{\theta} = \frac{n}{ln(W)}$$

So, yes the MLE is a function of W.

11 5.7.3

Suppose $Y_1, Y_2, ..., Y_n$ is a random sample from the exponential pdf, $fY(y; \lambda) = \lambda e^{-\lambda y}, y > 0$.

(a) Show that $\hat{\lambda}_n = Y_1$ is not consistent for λ.

(b) Show that $\hat{\lambda}_n = \sum_{i=1}^{n} Y_i$ is not consistent for θ.

Answer:

(a) We evaluate $\hat{\lambda}_n = \sum Y_1$

$$P(Y_1 > 2\lambda) = \int_{2\lambda}^{\infty} \lambda e^{-\lambda y} dy = \left[\frac{\lambda e^{-\lambda y}}{-\lambda}\right] = e^{-2\lambda^2}$$

We can use the probability inequality,

$$P(|y_1 - 2\lambda| < \frac{\lambda}{2}) < 1 - e^{-2\lambda^2} \rightarrow lim_{n\rightarrow\infty} P(|y_1 - 2\lambda| < \frac{\lambda}{2}) < 1$$

. This is less than one. Therefore the estimator is not consistent. If it were consistent it would converge to 1 in the limit. $hat\lambda_n = Y_1$ is not a consistent estimator.

(b) We evaluate $\hat{\lambda}_n = \sum Y_i$
We are going to have to use an inequality. It is fairly obvious that $P(\sum Y > 2\lambda) \geq P(Y - 1 > 2\lambda)$ And we can then recycle some of the math from last part of this question.

$$P(Y_i > 2\lambda) = \int_{2\lambda}^{\infty} \lambda e^{-\lambda y} dy = \left[\frac{\lambda e^{-\lambda y}}{-\lambda}\right] = e^{-2\lambda^2}$$

$$P(|\sum Y_i - 2\lambda| < \frac{\lambda}{2}) < P(|y_1 - 2\lambda| < \frac{\lambda}{2}) < 1 - e^{-2\lambda^2} \rightarrow lim_{n\rightarrow\infty} P(|\sum y_i - 2\lambda|, \frac{\lambda}{2}) < 1$$

Therefore $\hat{\lambda_n} = \sum Y$ is not consistent.

12 MONTE CARLO SECTION

We are going to expand on the section in the discussion board and simulate a normal distribution and display it on top of the histogram of returns generated from S and P 500 data.

13 Looking at the distribution of 1 Year returns in the SandP 500

We are going to look at the distribution of returns of the s and p 500 for a 1 year period form 1950 to 2018. We will see if they are log normally distributed.

```
In [2]: import pandas as pd
        import numpy as np

        sp500 = pd.read_csv('^GSPC.csv')
```

```
In [3]: sp500['1yYield'] = sp500['Adj Close'] / sp500.shift(252)['Adj Close']
        sp500[253:].head()
```

```
Out[3]:            Date       Open       High        Low      Close  Adj Close  \
        253  1951-01-08  21.000000  21.000000  21.000000  21.000000  21.000000
        254  1951-01-09  21.120001  21.120001  21.120001  21.120001  21.120001
        255  1951-01-10  20.850000  20.850000  20.850000  20.850000  20.850000
        256  1951-01-11  21.190001  21.190001  21.190001  21.190001  21.190001
        257  1951-01-12  21.110001  21.110001  21.110001  21.110001  21.110001

              Volume   1yYield
        253  2780000  1.246291
        254  3800000  1.247490
        255  3270000  1.227915
        256  3490000  1.240632
        257  2950000  1.239577
```

We take the log of the 1 year yields.

```
In [4]: sp500['log1yYield'] = np.log(sp500['1yYield'])
```

13.0.1 Lets look at some Summary Statistics

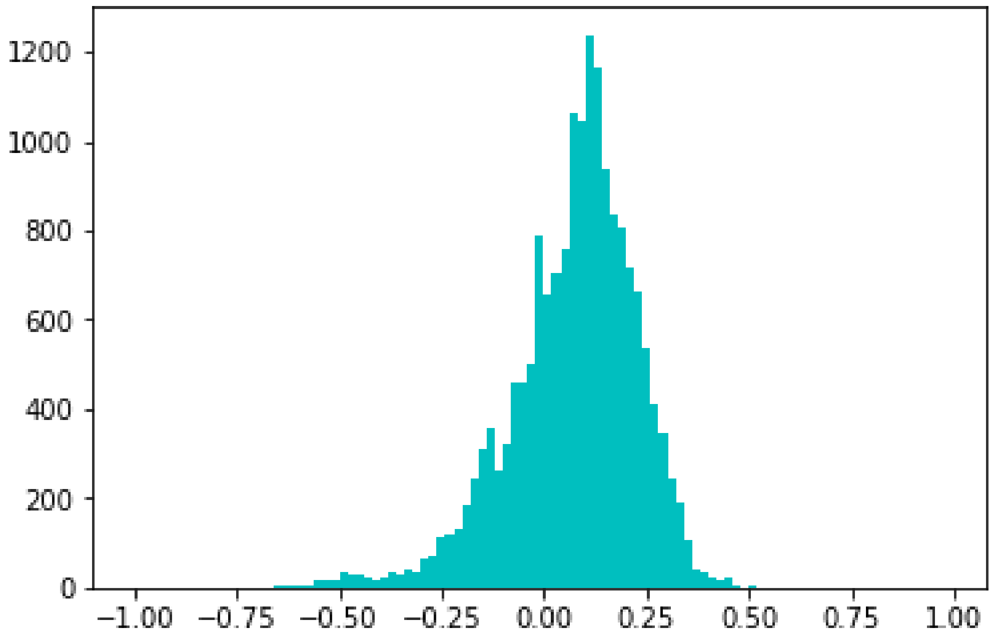
```
In [5]: sp500['log1yYield'].describe()
```

```
Out[5]: count    17060.000000
        mean         0.073889
        std          0.154376
        min         -0.669877
        25%         -0.008933
        50%          0.096759
        75%          0.177204
        max          0.522201
        Name: log1yYield, dtype: float64
```

We see that the mean of the log returns is .07. This means the average return is $e^{.07} = 1.0725081812542165$

Lets look at a histogram of the data.

```
In [7]: from matplotlib import pyplot as plt
        plt.hist(sp500['log1yYield'].replace(np.nan,0), bins = np.arange(-1,1,.02), color = 'c'
        plt.show()
```



13.0.2 Is it normally distributed?

This looks like it might be normally distributed. However lets do a few checks. We will assume the estimates of $\mu = 1.0725$ and $\sigma = 0.154376$

We know that if it is normally distributed, then the empiracle rule should apply (https://www.investopedia.com/terms/e/empirical-rule.asp)

68% of data should be within 1 standard deviation of the mean (-0.08048700000000002,0.228265).

95% of data should be within 2 standard deviation of the mean (-0.234863000,0.382641).

and 99.7 of data should be within 2 standard deviation of the mean (-0.38923900,0.53701700).

```
In [8]: print(sum((sp500['log1yYield']>-0.0804870000) & (sp500['log1yYield']< 0.228265))/sp500.s
        print(sum((sp500['log1yYield']>-0.2348635) & (sp500['log1yYield']< 0.382641))/sp500.shap
        print(sum((sp500['log1yYield']>-0.389239) & (sp500['log1yYield']< 0.5370170))/sp500.shap
```

```
0.7029228280961183
0.9423521256931608
0.9730244916820703
```

14 Summary

In the interval (-0.08048700000000002,0.228265) the data holds 70% of the probability mass. we would expect 68%.

In the interval (-0.234863000,0.382641) the data holds 94.235% of the probability mass. we would expect 95%.

In the interval (-0.38923900,0.53701700) the data holds 97.3024% of the probability mass. we would expect 99%.

Using the empiricle rule as a yardstick, we can see that assuming the S and P 500 is normally distributed is a bad idea. The actual distribution carries more weight in the tails than the normal distribution exhibits. I would expect this distribution to have a higher kurtosis number than what is expected for the normal distribution also.

14.1 Simulating the Parametric Distribution

This distribution has parameters of mean = 0.073889, and standard deviation of 0.154376. We will superimpose this on the histogram above.

```
In [11]: randoms = np.random.normal(loc =  0.073889, scale =0.154376, size = 17060)
```

```
In [17]: plt.hist(sp500['log1yYield'].replace(np.nan,0), bins = np.arange(-1,1,.02), color = 'c'
         plt.hist(randoms, color = 'r', bins = np.arange(-1,1,.02))
         plt.show()
```