STAT 5703 Homework 1

Shijie He(sh3975), Yunjun Xia(yx2569), Shuyu Huang(sh3967)

2/6/2020

Exercise 2

(1)

$$E[\bar{X}^2] = E[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} X_i X_j] = E[\frac{1}{n^2}\sum_{i=1}^{n} {X_i}^2] + E[\frac{1}{n^2}\sum_{i\neq j} X_i X_j]$$

We know:

$$E[{X_i}^2] = var(X_i) + (E[X_i])^2 = \lambda + \lambda^2$$

Hence,

$$E[\frac{1}{n^2}\sum_{i=1}^{n} {X_i}^2] = \frac{1}{n}(\lambda + \lambda^2)$$

Since X_is are independent, we can get:

$$E[\frac{1}{n^2}\sum_{i\neq j} X_i X_j] = \frac{2}{n^2}\binom{n}{2} E[X_1]E[X_2] = \frac{n-1}{n}\lambda^2$$

Thus,

$$E[\bar{X}^2] = \frac{1}{n}(\lambda + \lambda^2) + \frac{n-1}{n}\lambda^2 = \lambda^2 + \frac{1}{n}\lambda$$

(2)

$$E[s^2] = \frac{1}{n-1}E[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2] = \frac{1}{n-1}(nE[X_i^2] - nE[\bar{X}^2]) = \frac{n}{n-1}(\lambda + \lambda^2 - \lambda^2 - \frac{1}{n}\lambda) = \lambda$$

Thus, s^2 is an unbiased estimator of λ.

(3)

$$E[Y_i] = E[(X_i - \lambda)^2 - X_i] = var(X_i - \lambda) + (E[X_i - \lambda])^2 - E[X_i] = \lambda - \lambda = 0$$

$$var(Y_i) = E[Y_i^2] - (E[Y_i])^2 = E[(X_i-\lambda)^4 - 2X_i(X_i-\lambda)^2 + X_i^2] = E[X_i^4 - (4\lambda+2)X_i^3 + (6\lambda^2+4\lambda+1)X_i^2 - (4\lambda^3+2\lambda^2)X_i + \lambda^4]$$

To find the expected value of X^n, we use moment generating function:

$$M_X(t) = e^{\lambda(e^t-1)}$$

Hence,

$$E[X^3] = \frac{d^3}{dt^3}M_X(t)|_{t=0} = e^{\lambda(e^t-1)}(\lambda^3 e^{3t} + 3\lambda^2 e^{2t} + \lambda e^t)|_{t=0} = \lambda^3 + 3\lambda^2 + \lambda$$

$$E[X^4] = \frac{d^4}{dt^4}M_X(t)|_{t=0} = \lambda^4 + 6\lambda^3 + 7\lambda^2 + \lambda$$

Putting things together, we can get:

$$var(Y_i) = \lambda^4 + 6\lambda^3 + 7\lambda^2 + \lambda - (4\lambda+2)(\lambda^3 + 3\lambda^2 + \lambda) + (6\lambda^2 + 4\lambda + 1)(\lambda^2 + \lambda) - (4\lambda^3 + 2\lambda^2)\lambda + \lambda^4 = 2\lambda^2$$

Thus, $E[Y_i] = 0$, $var(Y_i) = 2\lambda^2$

(4)

$$\begin{aligned}
s^2 - \bar{X} &= \frac{1}{n-1}(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2) - \bar{X} \\
&= \frac{1}{n-1}(\sum_{i=1}^{n} Y_i + (2\lambda n + n)\bar{X} - n\lambda^2 - n\bar{X}^2) - \bar{X} \\
&= \frac{1}{n-1}\sum_{i=1}^{n} Y_i + \frac{2\lambda n + 1}{n-1}\bar{X} - \frac{n}{n-1}(\lambda^2 + \bar{X}^2) \\
&= \frac{1}{n-1}(\sum_{i=1}^{n} Y_i - n(\bar{X} - \lambda)^2 + \bar{X})
\end{aligned}$$

(5)

As $var(\bar{X}) \xrightarrow[n\to\infty]{p} 0$, $P(\bar{X} = \lambda) \xrightarrow[n\to\infty]{p} 1$.

Hence, as n approximates infinity, the distribution of $s^2 - \bar{X}$ can approximately equal to the distribution of $\frac{1}{n-1}\sum_{i=1}^{n} Y_i$.

Since,

$$\frac{1}{n-1}\sum_{i=1}^{n} Y_i \xrightarrow[n\to\infty]{D} N(0, \frac{2\lambda^2}{n})$$

we can get:

$$\sqrt{n}(s^2 - \bar{X}) \xrightarrow[n\to\infty]{D} N(0, 2\lambda^2)$$

As $\bar{X} \xrightarrow[n\to\infty]{P} \lambda$,

$$\sqrt{\frac{n}{2}}\frac{(s^2 - \bar{X})}{\bar{X}} \xrightarrow[n\to\infty]{D} N(0,1)$$

(6)

Calculate the test statistic $T = \sqrt{\frac{n}{2}}\frac{(s^2-\bar{X})}{\bar{X}}$. If the $T < -z_{1-\frac{\alpha}{2}}$ or $T > z_{1-\frac{\alpha}{2}}$, reject the null hypothesis. Otherwise, fail to reject the null hypothesis.

(7)

```r
samplea = c()
s_a = c()
for (i in 1:500){
  x = rpois(50,5)
  s_a = c(s_a, var(x))
  samplea = c(samplea, mean(x))
}

mean_a = mean(samplea)
s_a = mean(s_a)

sampleb = c()
s_b = c()
for (i in 1:500){
  theta = rgamma(50, shape = 2.5, rate = 0.5)
  y = c()
  for (j in 1:50){
    y = c(y, rpois(1, theta[j]))
  }
  sampleb = c(sampleb, mean(y))
  s_b = c(s_b, var(y))
}

mean_b = mean(sampleb)
s_b = mean(s_b)

cat("Test statistic for model a is: ", sqrt(500/2)*(s_a-mean_a)/mean_a, "\n")
```

```
## Test statistic for model a is:  0.132675
```

```r
cat("Test statistic for model b is: ", sqrt(500/2)*(s_b-mean_b)/mean_b)
```

```
## Test statistic for model b is:  31.85603
```

For significance level $\alpha = 0.05$, $z_{0.975} = 1.96$. Thus, we do not reject null hypothesis for model a, but reject null hypothesis for model b.

(8)

```r
dpd = c(rep(0,1), rep(1,4), rep(2,15), rep(3,31), rep(4,39), rep(5,55), rep(6,54),
        rep(7,49), rep(8,47), rep(9,31), rep(10,16), rep(11, 9), rep(12, 8), rep(13, 4), rep(14, 3))

n = length(dpd)
s = var(dpd)
m = mean(dpd)

cat("Test statistic is: ", sqrt(n/2)*(s-m)/m, "\n")
```

```
## Test statistic is:  0.9786746
```

We do not reject null hypothesis. Thus, there is no overdispersion for this data set.