Statistics 210 Assignment 4 Due: 11:59pm 9^{th} February, 2023

1. In question 1(c) of assignment 2 we examined the validity the Poisson distribution. Here we will do this more formally with a Poisson dispersion test by way of a likelihood ratio test.

Suppose we have observations $x_1, x_2, \ldots x_n$ from a Poisson(λ) distribution. For example, in this assignment 1, we observed for the last 13 years ($n = 13$) we observed

`Drownings=c(65,91,75,77,70,85,84,87,66,79,84,90,93).`

We know that the MLE for λ is given by the sample mean $\bar{x}$. The likelihood ration statistic is given by comparing $\hat{\lambda} = \bar{x}$ to the situation whereby each observation has a different λ_i with its estimate being $\hat{\lambda}_i = x_i$.

Note that, for example, $\prod_{i=1}^{n} x_i = x_1 \times x_2 \times \cdots \times x_n$ is the multiplicative (product) analog to summation.

the likelihood ratio statistic , λ is calculated as follows:

$$\Lambda = \frac{\prod_{i=1}^{n} \frac{\lambda e^{-\hat{\lambda}}}{x_i!}}{\prod_{i=1}^{n} \frac{\hat{\lambda}_i e^{-\hat{\lambda}_i}}{x_i!}} = \prod_{i=1}^{n} \left(\frac{\bar{x}}{x_i}\right)^{x_i} e^{x_i - \bar{x}} \ .$$

It turns out that $-2\log_e(\Lambda) = -2\sum_{i=1}^{n}\left(x_i \log_e\left(\frac{\bar{x}}{x_i}\right) + (x_i - \bar{x})\right)$ is approximately Chi-squared($df = n - 1$) when λ is deemed to be large — as it is for this example.

(a) The expression $-2\log_e(\Lambda) = -2\sum_{i=1}^{n}\left(x_i \log_e\left(\frac{\bar{x}}{x_i}\right) + (x_i - \bar{x})\right)$

can be simplified to $-2\log_e(\Lambda) = 2\sum_{i=1}^{n} x_i \log_e\left(\frac{x_i}{\bar{x}}\right)$.

How was this achieved? **(4 marks)**

(b) In the following we will produce repeated observations to produce many $-2\log_e(\Lambda)$ random statistics from a Poisson($\lambda = 80$) distribution. Reproduce this code and the display the resulting plot. Comment briefly on the how well the Chi-squared($df = 13-1$) approximates the distribution of the $-2\log_e(\Lambda)$.

```
xmat=matrix(rpois(13*1e4,lambda=80), nc=13)
xbars=apply(xmat,1, mean)

# produce -2log(Lambda)
sims=2*apply(xmat*log(xmat/xbars),1,sum)

# histogram
hist(sims, ylim=1.1*c(0,dchisq(13-2, df=13-1)),prob=T,
     xlab=expression(paste("-2log(",Lambda,")")), ylab="probability",
     main=expression(paste("Histogram of simulated -2log(",Lambda,")")))
xs=seq(min(sims), max(sims), length=1e3)
# superimpose Chi-squared pdf
lines(xs, dchisq(xs, df=13-1), col="blue")
```

(**4 marks**)

(c) In the above plot we formatted the limits of the probability by the command `ylim=1.1*c(0,dchisq(13-2, df=13-1))`. This allowed us the see the peak value of the density of the Chi-squared probability density function at $k = 13 - 2 = 11$. Here we will see how this number is computed for $k > 2$.

The pdf of a Chi-squared($df = k$) is given by:

$f_X(x) = \dfrac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)}\, x^{\frac{k}{2}-1}e^{-\frac{x}{2}}$ for $x > 0$.

Show that for $k > 2$ peaks at the value $k - 2$. Verify is it a maximal value.

Hints: As $f_X(x) > 0$ maximising $f_X(x)$ is equivalent to maximising $\log_e(f_X(x))$. Identify constants.

(**4 marks**)

(d) Compute $-2\log_e(\Lambda)$ for the data
`Drownings=c(65,91,75,77,70,85,84,87,66,79,84,90,93)`.
Compute the P-value to test the hypothesis that we have the data is consistent H_0: data comes from a Poisson distribution using the Chi-squared($df = 13 - 1 = 12$) distribution. Comment briefly on you conclusion.

Hint: This is a more formal test to the one done in Question 1 (c) of Assignment 2. (**4 marks**)

(e) Here we will produce a Taylor series approximation for

$-2\log_e(\Lambda) \approx \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\bar{x}} = \dfrac{(n-1)s^2}{\bar{x}}$ where s^2 is the sample variance.

We start by defining here $g(x) = 2x\log_e\left(\dfrac{x}{\bar{x}}\right)$ and from the above and not that $-2\log_e(\Lambda) = \sum_{i=1}^{n} g(x_i)$.

Your first task is to approximate a Taylor series for the function up to the quadratic term. Then use this approximation to derive the statement above.

(**6 marks**)

(f) Verify that approximation is, not surprisingly, is close to the same Chi-squared distribution. Produce a similar plot to the above.

Hint: Adapt the `R` code, above, to obtain a similar histograms with density plot. We the sample means, $\bar{x}$ via
`apply(xmat,1 mean)`. The sample variance, s^2, in `R` can be computed using the `var` command.

(**4 marks**)

(g) Test the null hypothesis above. How does the statistic above compare to the ad-hoc test we did in question 1(c) of assignment two. Comment briefly.

`Drownings=c(65,91,75,77,70,85,84,87,66,79,84,90,93)` and comment on the P-value obtained for testing the null hypotheisi: H_0: data comes from a Poisson distribution. (**4 marks**)

2. Suppose an insect lays N random eggs which can be described by a Poisson(λ) distribution. Furthermore, suppose of those laid, the the number that hatch (i.e. $Y|N = n$) can be described by a Binomial(n, p) probability distribution.

(a) Use the partition theorem for expectation and variances to obtain the expected number and variance of the total number of eggs that hatch Y. (**5 marks**)

(b) The relationship above suggests that Y has a certain distribution. Suggest a distribution and explain why you suspect this is so. (**2 marks**)

(c) Derive the distribution for Y and verify your suggestion above. (**7 marks**)

3. In this question we will examine the awesome bi-variate normal distribution.

(a) Generate random bi-variate Normal observations of size $n = 1000$ with correlation coefficient of $\rho = -.95, \rho = 0$, and $\rho = .95$, respectively, as described in example 7.3 of your notes. Plot the the resulting observations of Y versus X,

Note: Start with two $n = 1000$ observations using the `R` functions `rnorm`.

Here's the code for the first plot:

```
Z=rnorm(1e3); W=rnorm(1e3)
rho=-.95
X=Z; Y=rho*Z+sqrt(1-rho^2)*W
plot(Y~X, col="lightgrey",
main=expression(
paste("Bivariate Normal("
,mu[x],"=0,",sigma[X],"=1,"
,mu[Y],"=0," ,sigma[Y],"=1,"
,rho[XY],"=-0.95)")))

abline(0, rho, lty=2, col="blue" )
abline(h=0, col="red", lty=3)
abline(v=0, col="red", lty=3)
```

Adapt the code above for $\rho = 0$ and $\rho = .95$. Comment briefly on all three plots.

(6 marks)

(b) It can be shown that for the bi-variate normal distributions (with $\mu_X = \mu_Y = 0$ and $\sigma_X^2 = \sigma_Y^2 = 1$) that

$\mathbb{P}(X > 0, Y > 0) = \frac{1}{4} + \frac{\sin^{-1}(\rho)}{2\pi}$ when

Calculate this probability for each of the three cases above and explain how these 'make sense' when you compare these values with their respective plots.

Note: the inverse sine function gives the angle in radians fora given value $-1 \leq \rho \leq 1$, $\sin^{-1}(\rho)$ =arcsine(ρ)=`asin(rho)` in `R`. Also, π =`pi` in `R`.

(6 marks)

(c) Consider the model $Y_i = \beta x_i + \varepsilon_i$ where $\varepsilon_i \sim iid$ Normal$(0, \sigma^2)$. Let us assume σ^2. Compute the MLE estimate of β, $\widehat{\beta}$. Ensure you confirm that it is unbiased and that it maximises the likelihood.
(6 marks)

(d) Compute the variance of this MLE estimate.
(4 marks)

(e) Using the data generated above above when $\rho = 0.95$ - calculate the MLE estimate for these data. **(4 marks)**

(f) Using the distribution of $Y|X$ discussed in example 7.3. Explain why this result is not at all surprising. As a consequence tell us what the variance of of each ε_i would be. **(4 marks)**

Total: 74 marks.