ECON 580 **Shabab Ahmed**
Due Date: Saturday, December 05

Homework 7

PROBLEM 1:

Sampling Distributions:

A researcher plans to carry out an opinion survey regarding a yes/no question. Suppose that (unknown to the researcher) 89% of people in the target population answer 'yes'. We will use 1 to denote 'yes' and 0 to denote 'no'.

(a) Let X_1 be the response from the first person the researcher asks. Before this person is asked, what is the distribution of X_1. What is $E[X_1]$ and $V(X_1)$?

The researcher plans to obtain a random sample of size 25: $X_1, \ldots, X_{25}$.

(Suppose either that the researcher is sampling with replacement or that the population is so large that we may regard the samples as being taken with replacement.)

(b) Write down the distribution of X_i for $i = 1, \ldots, 25$; also write down $E[X_i]$, $V(X_i)$ and $\text{Cov}(X_i, X_j)$ for $i \neq j$.

(c) What is the distribution of $\bar{X}$? *Hint: see lecture notes.*

What is the mean and variance of this distribution?

(d) Use `R` or Python, replicate the experiment $10,000$ times.

Hint: Recall that a Bernoulli(p) random variable is just a Binomial$(1, p)$ random variable; thus to obtain a sample of size 25 *from a Bernoulli*(0.89) *random variable, one might use:*

```
p <- 0.89

n <- 25

x <- rbinom(n,1,p)
```

The vector `x` *will contain the* 25 *Bernoulli(p) observations, from which you can compute the sample mean. This should then be put in a loop which repeats this process, 10K times, storing the result.* Construct a histogram with the distribution of $\bar{X}$ for each sample. Does the distribution of sample means appear to be normal? Explain your answer.

(e) Find the mean and variance of $\bar{X}$, based on your 10K simulations. Does this agree with your calculation in (c)?

(f) Compute the mean squared error of $\bar{X}$ as an estimate of the unknown proportion $\theta = 0.89$ (i.e. find the squared error of the estimate, which is $(\bar{X} - \theta)^2$, in each repetition and then average over all repetitions). What do you notice? Give a simple explanation by referring to a result from the course.

SOLUTION:

Since we have an 'yes/no' question, X_1 is going to be a Bernoulli random variable. The probability that $X_1 = 1$, that is, the response from the first person is going to be 'yes' is simply 0.89. Hence, our Bernoulli parameter $p = 0.89$. Thus:

$$\boxed{X_1 \backsim \text{Bernoulli}(0.89)}.$$

Then,

$$\boxed{E[X_1] = p = 0.89}$$

and

$$\boxed{V[X_1] = p(1-p) = 0.89 \times 0.11 = 0.0979}.$$

---

b) Since, we are sampling with replacement:

$$\boxed{X_i \backsim \text{Bernoulli}(0.89) \text{ for } i = 1, \ldots, 25}$$

Hence,

$$\boxed{E[X_i] = 0.89 \text{ and } V[X_i] = 0.0979 \text{ for } i = 1, \ldots, 25}$$

Since, we are taking a random sample we can think of each of the responses to be i.i.d. Thus, each of the responses are independent and so each of the X_i as independent. Hence,

$$\boxed{\text{Cov}(X_i, X_j) = 0 \text{ for } i \neq j}$$

Since, X_i are i.i.d we have by the independence:

$$f(x_1, x_2, \ldots, x_{25}) = f(x_1) \times f(x_2) \cdots \times f(x_{25})$$

$$= p^{x_1}(1-p)^{(1-x_1)} \times p^{x_2}(1-p)^{(1-x_2)} \times p^{x_{25}}(1-p)^{(1-x_{25})}$$

$$= p^{(x_1+x_2+\cdots+x_{25})}(1-p)^{(1-x_1)+(1-x_2)+\cdots+(1-x_{25})}$$

$$= p^{\sum_{i=1}^{25} x_i}(1-p)^{n-\sum_{i=1}^{25} x_i}$$

$$= \boxed{(0.89)^{\sum_{i=1}^{25} x_i}(1-0.89)^{25-\sum_{i=1}^{25} x_i}}$$

Then:

$$\text{E}\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_{25} \end{bmatrix} = \begin{bmatrix} E[X_1] \\ E[X_2] \\ \vdots \\ E[X_{25}] \end{bmatrix}$$

$$\boxed{= \begin{bmatrix} 0.89 \\ 0.89 \\ \vdots \\ 0.89 \end{bmatrix}}$$

$$\text{V}\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_{25} \end{bmatrix} = \begin{bmatrix} V[X_1] & Cov(X_1, X_2) \ldots & & \\ Cov(X_1, X_2) & V[X_2] & & \\ & & \ddots & \\ & & & V[X_{25}] \end{bmatrix}$$

Since, $\text{Cov}(X_i, X_j) = 0$, we have that:

$$= \begin{bmatrix} V[X_1] & 0 & \dots & 0 \\ 0 & V[X_2] & & \\ & & \ddots & \\ 0 & & & V[X_{25}] \end{bmatrix}$$

$$= \begin{bmatrix} 0.0979 & 0 & \dots & 0 \\ 0 & 0.0979 & & \\ & & \ddots & \\ 0 & & & 0.0979 \end{bmatrix}$$

$$= \boxed{0.0979 \times I_{25\times 25}}$$

where $I_{25\times 25}$ is 25×25 identity matrix.

---

c) Since each $X_i \backsim$ Bernoulli(0.89) for $i = 1, \dots, 25$ we have from the lecture notes:

$$25\bar{X} \backsim \text{Binomial}(25, 0.89)$$

Let $Z = 25\bar{X}$ and so we want to find the distribution of $Y = \frac{Z}{25} = \bar{X}$.

Then, $f(y) = P(Y = y) = P(\frac{Z}{25} = y) = P(Z = 25y) = f(25y) = \frac{25!}{(25y!)(25-25y)!} p^{25y}(1-p)^{25-25y}$ where $p = 0.89$.

Notice: The support for $\bar{X} = \{0, \frac{1}{25}, \frac{2}{25}, \dots, \frac{25}{25}\}$. Therefore, the distribution of $\bar{X}$ is given by the following mass function:

$$\boxed{f(\bar{x}) = \begin{cases} \frac{25!}{(25\bar{x}!)(25-25\bar{x})!} 0.89^{25\bar{x}}(0.11)^{25-25\bar{x}} & \text{if } \bar{x} \in \{0, \frac{1}{25}, \frac{2}{25}, \dots, \frac{25}{25}\} \\ 0 & \text{otherwise} \end{cases}}$$

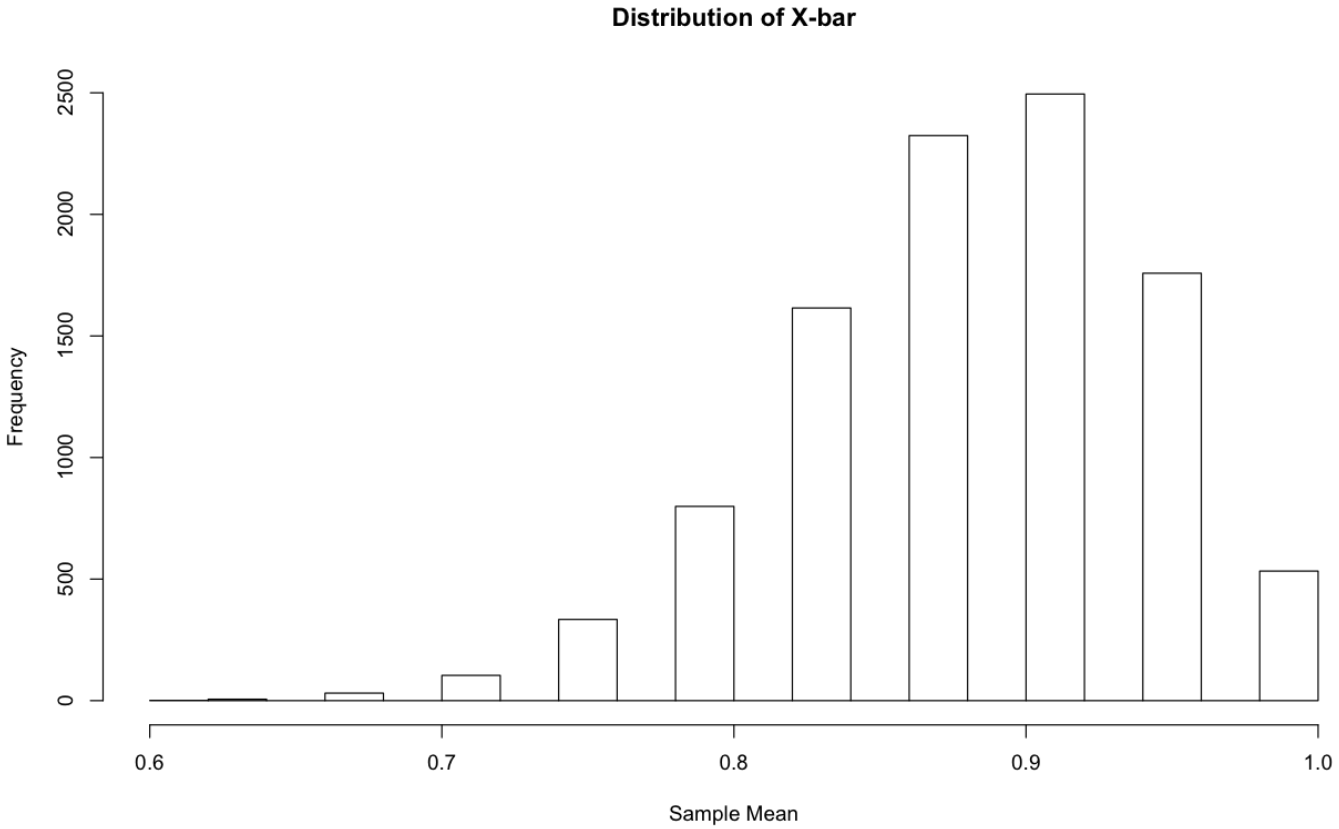
Notice: the population mean $\mu = p = 0.89$ and the population variance $\sigma^2 = p(1-p) = 0.0979$.

By the Sample Mean Theorem,

$$\boxed{E(\bar{X}) = \mu = p = 0.89}$$

and

$$\boxed{V(\bar{X}) = \frac{\sigma^2}{25} = \frac{0.0979}{25} = 0.003916}.$$



d) The distribution does not appear to be normal. The normal distribution is symmetric about its mean which is also its mode. It can be seen from the histogram that the distribution of $\bar{X}$ is not quite symmetric across its mode. The distribution appears skewed and we have more possible values to the left of the mode compared to the right of the mode. The distribution has a slightly longer tail on one side. This is also given by the fact that p is bounded on (0,1) and the center here is at around 0.9. Also, the fact that the distribution of $\bar{X}$ is a discrete random variable is clearly depicted here as can be seen in the gaps between the bars on the histogram.

---

e) Based on our 10,000 simulations:

$$\boxed{E[\bar{X}] = 0.891124}$$

and

$$\boxed{V[\bar{X}] = 0.00385}.$$

Yes, this agrees with our calculation in part c) which stated that $E[\bar{x}] = 0.89$ and $V[\bar{X}] = 0.003916$.The answers match exactly if we round to three decimal places for the variance and round to

two decimal places for the mean.

---

f) The mean squared error of $\bar{X}$ as an estimate of the unknown proportion $\theta = 0.89$ is equal $\boxed{0.00385272}$. We notice that is equal to the variance of the distribution of $\bar{X}$ we calculated in part c) and also the variance we got from our $10,000$ simulations. For an explanation regarding this we will refer back to T3 on the Theorem of Expectations:

$$E((X-c)^2) = V(X) + (c-\mu_x)^2$$

and this is minimized when $c = \mu_x$ and then $E((X-c)^2) = V(X)$. Let $X = \bar{X}$ and $c = \theta = p = 0.89$ and notice that $E(\bar{X}) = p = 0.89$. Therefore, $c = \theta = E(\bar{X})$ and from the result mentioned above we have that $E((\bar{X}-\theta)^2) = V(\bar{X})$.

Code:

```
set.seed(10000)

nsims <- 10000 # number of simulations
seq <-1:nsims

smean <- c() # initiliazing list of sample means
mse <- c() # initializing list to calculate mean squared error

for (i in seq){
p <- 0.89  # Bernoulli parameter
n <- 25 # sample size
x <- rbinom(n,1,p)
avg <- mean(x)
smean[i] <- avg
se <- (avg-p)^2
mse[i] <- se
}

hist(smean, main = "Distribution of X-bar", xlab = "Sample Mean")
ex <- mean(smean)
vx <- var(smean)

# Mean squared error

MSE <- mean(mse)
```

PROBLEM 2:

Given a sample $X_1, \ldots, X_n$ of independent Poisson(λ) random variables, a researcher intends to use $\bar{X}$, the sample mean, as an estimate of λ. Suppose that $n = 20$, and $\lambda = 4$.

(a) Write down the mean and variance of this distribution.

(b) Using R or Python, replicate the experiment $10,000$ times. Construct a histogram with the distribution of $\bar{X}$ for each sample. Does the distribution of sample means appear to be normal? Explain your answer.

(c) Find the mean and variance of $\bar{X}$, based on your 10K simulations. Does this agree with your calculation in (a)?

(d) Compute the mean squared error of $\bar{X}$ as an estimate of λ. Again what do you notice?

SOLUTION:

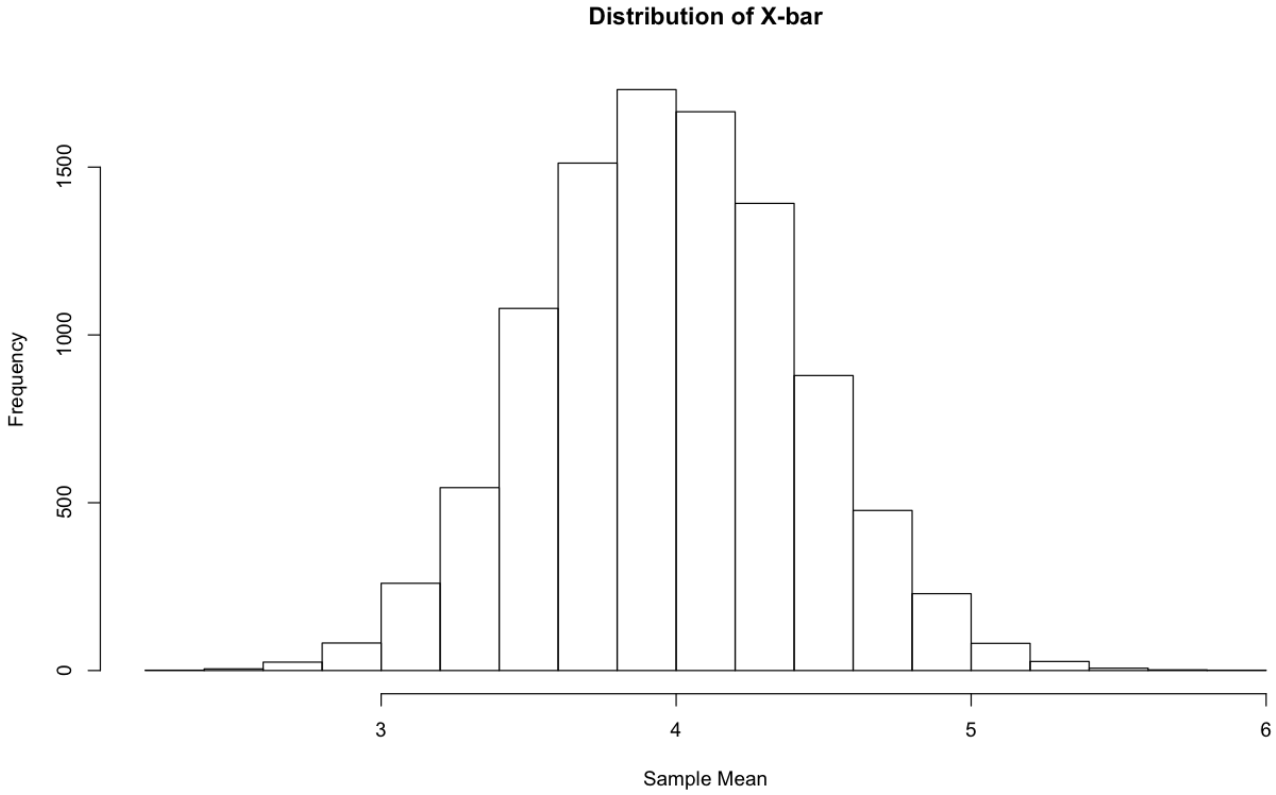
a) We have Poisson random variables and using Table 3.1 from Goldberger we have that the population mean $\mu = \lambda = 4$ and the population variance $\sigma^2 = \lambda = 4$.

By the Sample Mean Theorem, the mean of the distribution is given by,

$$\boxed{\mu = \lambda = 4}$$

and the variance of the distribution is given by,

$$\boxed{\frac{\sigma^2}{20} = 0.2}.$$



The distribution of the sample mean in this case appears to be normal. One key characteristic that I was looking for was being symmetric around the mode. The histogram in this case does seem to be symmetric about the mode. It may be slightly off but looks symmetric to the naked eye. Another key thing to note is that there are no gaps between the bars telling us that in fact we can think of this as a normal distribution as the the sample mean distribution has the appearance of being continuous.

---

c) Based on the $10,000$ simulations:

$$\boxed{\text{mean of } \bar{X} = 4.00381}$$

$$\boxed{\text{variance of } \bar{X} = 0.1980133 \approx 0.20}$$

This agrees with the calculation in (a) and our answers would exactly match up if we rounded to two decimal places.

---

d) The mean squared error of $\bar{X}$ as an estimate of λ is given by:

0.198008

which is the same (approximately) as the variance of $\bar{X}$ we calculated in part a) and also the variance we got in part c). This, again, follows from T3 and T4 of the theorem on expectations because in this case $c = \lambda$ and $E(\bar{X}) = \lambda$ and therefore we get $E((\bar{X} - \lambda)^2) = V(\bar{X})$.

Code:

```
set.seed(10000)

nsims <- 10000 # number of simulations
seq <-1:nsims

smean <- c() # initiliazing list of sample means
#mse <- c() # initializing list to calculate mean squared error

for (i in seq){
lambda <- 4 # Bernoulli parameter
n <- 20 # sample size
x <- rpois(n,lambda)
avg <- mean(x)
smean[i] <- avg
se <- (avg-lambda)^2
mse[i] <- se
}

hist(smean, main = "Distribution of X-bar", xlab = "Sample Mean")
ex <- mean(smean)
vx <- var(smean)

# Mean squared error

MSE <- mean(mse)
```

PROBLEM 3:

Suppose that $Y_1, \ldots, Y_n$ are i.i.d samples from a Poisson distribution with parameter λ.

(a) Find the log of the likelihood: $\log f(y_1, \ldots, y_n \mid \lambda)$;

(b) By differentiating the log likelihood, find the value $\hat{\lambda}$ of λ that maximizes the likelihood; confirm that $\hat{\lambda}$ is a maximum. The estimate $\hat{\lambda}$ is called the *maximum likelihood estimator (MLE)*.

Suppose that we wish to test H_0: $\lambda = \lambda_0$ vs. H_1: $\lambda \neq \lambda_0$.

(c) Using your answer to (b) write down the generalized likelihood ratio test statistic (LRT).

Hint: Your answer should be a function of n, λ_0 and $\hat{\lambda}$.

(d) Consider Bortkiewicz's Prussian Cavalry horse-kick fatality data:

No. of fatalities	0	1	2	3	4
No. of years	109	65	22	3	1

Find the value of $\sum_{i=1}^{n} y_i$, and use your answer to (b) to find the value of the MLE $\hat{\lambda}$.

Hint: The table contains data on $n = 200$ observations, $y_1, \ldots, y_{200}$, of which 109 were 0; there were 65 which were 1 and so on.

(e) Use your answer to (d) to find the LRT statistic for the hypothesis test with $\lambda_0 = 1$.

(f) Report the approximate p-value.

Hint: calculate $-2\log$(LRT), and compare to the appropriate χ^2 distribution. See Lecture 10, slide 43. Recall that small values of the LRT correspond to evidence against H_0.

SOLUTION:

a) $Y_1, \ldots, Y_n$ are i.i.d samples from a Poisson distribution with parameter λ. The independence gives us the following:

$$f(y_1, \ldots, y_n|\lambda) = f(y_1|\lambda) \times \cdots \times f(y_n|\lambda) = \prod_{i=1}^{n} f(y_i|\lambda)$$

Now, notice the following:

$$f(y_i|\lambda) = \frac{\lambda^{y_i} e^{-\lambda}}{(y_i)!} \qquad \text{for } i = 1, \ldots, n$$

Therefore:

$$f(y_1, \ldots y_n|\lambda) = \frac{\lambda^{y_1} e^{-\lambda}}{(y_1)!} \times \cdots \times \frac{\lambda^{y_n} e^{-\lambda}}{(y_n)!} = \frac{\lambda^{y_1+\cdots+y_n} e^{-n\lambda}}{(y_1)! \times \cdots \times (y_n)!}$$

Hence, the log of the likelihood is given by:

$$\log f(y_1, \ldots y_n|\lambda) = \log\left(\frac{\lambda^{y_1+\cdots+y_n} e^{-n\lambda}}{(y_1)! \times \cdots \times (y_n)!}\right)$$

$$= \log(\lambda^{y_1+\cdots+y_n}) + \log(e^{-n\lambda}) - \log((y_1)! \ldots (y_n)!)$$

$$= \boxed{(y_1 + \cdots + y_n)\log(\lambda) - n\lambda - \log((y_1)! \ldots (y_n)!)}$$

---

b) Differentiating the log likelihood function and setting it equal to zero:

$$\frac{y_1 + \cdots + y_n}{\lambda} - n = 0$$

$$\implies \boxed{\hat{\lambda} = \frac{y_1 + \cdots + y_n}{n} = \frac{\sum_{i=1}^{n} y_i}{n}} = \bar{y}$$

So, the MLE is just the sample mean.

Now, checking this is the maximum by considering the second derivative:

$$\frac{d^2(\log f(y_1, \ldots, y_n|\lambda))}{d\lambda^2} = -\frac{y_1 + \cdots + y_n}{\lambda^2}$$

Notice: The second derivative must be less than zero because $\lambda^2 > 0$ and sum of the y_is are going to be greater than zero since they are Poisson variables and the Poisson distribution only has support for non-negative values. We do not consider the case when all of the y_is

are equal to zero because then the log likelihood function will simply be $n\lambda$ and will not give us a solution for the maximum. Thus, at least one of the y_is must be strictly positive and then the sum of all the y_is will also be strictly positive. Therefore:

$$\frac{d^2(\log f(y_1, \ldots, y_n|\lambda))}{d\lambda^2} < 0$$

and we indeed do have a maximum.

---

c) Recall: When evaluating a (possibly) composite null vs a (possibly) composite alternative:

$$H_0 : \theta \in \Omega_0$$
$$H_1 : \theta \in \Omega_1$$

the generalized likelihood ratio test is given by:

$$\Lambda = \frac{\max_{\theta \in \Omega_0} f(x|\theta)}{\max_{\theta \in \Omega_0 \cup \Omega_1} f(x|\theta)}$$

In our case:

$$H_0 : \lambda = \lambda_0$$
$$H_1 : \lambda \neq \lambda_0$$

So, our $\Omega_0 = \lambda_0$ and $\Omega_1 = \{\text{all possible values of } \lambda \text{ except } \lambda_0\} = \{\lambda : \lambda > 0, \lambda \neq \lambda_0\}$.

Therefore, $\Omega_0 \cup \Omega_1 = \{\lambda : \lambda > 0\} = \{\text{all possible values of } \lambda\}$. Using our results from above:

$$\Lambda = \frac{f(y_1, \ldots, y_n|\lambda_0)}{f(y_1, \ldots, y_n|\hat{\lambda})}$$

Recall: From part a)

$$f(y_1, \ldots y_n|\lambda) = \frac{\lambda^{y_1 + \cdots + y_n} e^{-n\lambda}}{(y_1)! \times \cdots \times (y_n)!}$$

Hence,

$$f(y_1, \ldots y_n|\lambda_0) = \frac{\lambda_0^{y_1+\cdots+y_n} e^{-n\lambda_0}}{(y_1)! \times \cdots \times (y_n)!}$$

Now, we will calculate $f(y_1, \ldots, y_n|\hat{\lambda})$.

$$f(y_1, \ldots, y_n|\hat{\hat{\lambda}}) = \frac{\hat{\lambda}^{y_1+\cdots+y_n} e^{-n\hat{\lambda}}}{(y_1)! \times \cdots \times (y_n)!}$$

Thus,

$$\Lambda = \frac{\frac{\lambda_0^{y_1+\cdots+y_n} e^{-n\lambda_0}}{(y_1)!\times\cdots\times(y_n)!}}{\frac{\hat{\lambda}^{y_1+\cdots+y_n} e^{-n\hat{\lambda}}}{(y_1)!\times\cdots\times(y_n)!}}$$

$$= \frac{\lambda_0^{y_1+\cdots+y_n} e^{-n\lambda_0}}{\hat{\lambda}^{y_1+\cdots+y_n} e^{-n\hat{\lambda}}}$$

$$= \boxed{\left(\frac{\lambda_0}{\hat{\lambda}}\right)^{\sum_{i=1}^n y_i} e^{-n(\lambda_0-\hat{\lambda})}}$$

---

d) We will calculate $\sum_{i=1}^n y_i$:

$$\sum_{i=1}^{n} y_i = 1 \times 65 + 2 \times 22 + 3 \times 3 + 4 \times 1$$

$$= 65 + 44 + 9 + 4 = \boxed{122}$$

From part b),

$$\hat{y} = \frac{\sum_{i=1}^n y_i}{n} = \frac{122}{200} = \boxed{0.61}.$$

---

e) In this case, $\lambda_0 = 1$.

Therefore,

$$\Lambda = \left(\frac{1}{0.61}\right)^{122} e^{-200(1-0.61)}$$

$$= \boxed{2.06438 \times 10^{-8}}.$$

---

f)

Let $T = -2\ln(\Lambda) = -2\ln(2.06438 \times 10^{-8}) = 35.3917$.

Notice: $\dim(\Omega_0 \cup \Omega_1) = 1$ because the set corresponds to positive real numbers. It is simply a line and so has dimension 1.

Also, $\dim(\Omega_0) = 0$ because it simply a point. Then, $d = \dim(\Omega_0 \cup \Omega_1) - \dim(\Omega_0) = 1$.

We will find the p-value by computing:

$$1 - F_d(T) = 1 - pchisq(T, d) = 1 - pchisq(35.3917, 1)$$

Using R, the p-value is $\boxed{2.696335 \times 10^{-9}}$.

PROBLEM 4:

Suppose that we are planning an experiment to test hypotheses about the mean of a population that is known to be normal with standard deviation $\sigma = 4$. We wish to test the null hypothesis $H_0 : \mu = 0$ vs. the alternative $H_A : \mu > 0$. We intend to use a likelihood ratio test with significance level $\alpha = 0.05$.

(a) For which values of $\bar{X}$, the sample mean, will we reject the null hypothesis. Express your answer as a function of sample size, n.

(b) Suppose that we plan to obtain a sample of size 36. The researcher thinks that if the alternative is true then perhaps $\mu = 0.3$. Calculate the power of the test to reject the null hypothesis under this particular alternative hypothesis.

(c) Continuing from (b), the scientist who is planning the experiment wishes to have power at least 90%. (Thus your calculation in (b) shows that more than 36 observations are required.) Find approximately the smallest sample size at which this power can be achieved, against the specific alternative $\mu = 0.3$.

Hint: repeat the calculation performed in (b) *at different sample sizes using trial and error: you may wish to use* `R` *or a spreadsheet to speed up this calculation.*

(d) Suppose that the researcher obtains 200 samples, and observes $\bar{x} = 0.65$. Compute the p-value for this hypothesis test.

SOLUTION:

We know that $\sigma = 4$. We want to test the null hypothesis $H_0 : \mu = 0$ vs. the alternative $H_A : \mu > 0$. We will use a likelihood ratio test with significance level $\alpha = 0.005$.

---

a) Notice: The value of μ in the null hypothesis is less than the values of μ for the alternative hypothesis.

From the lecture notes, we will reject the null if:

$$\bar{X} > c$$

where

$$c = \mu_0 + \frac{\sigma}{\sqrt{n}} \times z_{1-\alpha}$$

Plugging in our vales:

$$c = 0 + \frac{4}{\sqrt{n}} \times z_{0.95}$$

Using R code:

```
qnorm(0.95)
```

we find that $z_{0.95} = 1.64$

Therefore,

$$c = \frac{4 \times 1.64}{\sqrt{n}} = \frac{6.56}{\sqrt{n}}.$$

Thus, we will reject the null hypothesis for the following values of $\bar{X}$:

$$\boxed{\bar{X} > c = \frac{6.56}{\sqrt{n}}}.$$

---

b) Suppose we plan to obtain a sample of size 36 and we have a set value of the alternative hypothesis as $\mu = 0.3 = \mu_1$.Recall: $\mu_0 = 0$.

$$\text{Power} = P(\text{Rej.}H_0|H_1)$$

$$= P\Big(\frac{\bar{X} - \mu_1)}{\sigma/\sqrt{n}} > \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}} + z_{1-\alpha}|\mu_1\Big)$$

$$= P\Big(Z > \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}} + z_{0.95}\Big)$$

$$= 1 - \text{pnorm}(\frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}} + z_{0.95})$$

where the fourth equality follows since $\frac{(\bar{X}-\mu_1)}{\sigma/\sqrt{n}}|\mu_1 \sim N(0,1)$.

Thus,

$$\text{Power} = 1 - \text{pnorm}(\frac{-0.3}{4/\sqrt{36}} + 1.64)$$

$$= 1 - \text{pnorm}(\frac{-0.3}{4/6} + 1.64)$$

$$= 1 - \text{pnorm}(1.19)$$

$$= \boxed{0.12} \quad \text{using R.}$$

---

c) We want to find approximately the smallest sample size at which power is at least 90%.

Using R we found that the sample size has to be at least $\boxed{1523}$. The following is the R code used:

```
u0 <- 0       # null hypothesis
u1 <- 0.3     # alternative hypothesis
alpha <- 0.05 # significance level
sigma <-4    # standard deviation

d_p <- 0.9  # desired power

z <- qnorm(1-alpha)

s<- 10000 # maximum sample size considered

seq <- 36:s

val <-c()

for (i in seq){
x <- (u0-u1)/(sigma/sqrt(i))+z
power <- 1-pnorm(x)
val[i] <- power
if (power > d_p){
n <- i # sample size for desire dpower
break
}
```

```
}
```

---

d) Suppose that the researcher obtains 200 samples and observes $\bar{x} = 0.65$. We will compute the p-value for this hypothesis test.

Recall: In a level α significant test we reject $H_0 : \mu = 0$ if:

$$\bar{x} > \frac{4}{\sqrt{n}} \times z_{1-\alpha} = \frac{4}{\sqrt{200}} \times z_{1-\alpha}$$

Hence, we will reject H_0 for an value $\alpha > \alpha^*$ where:

$$\bar{x} = \frac{4}{\sqrt{200}} \times z_{1-\alpha^*}$$

$$\iff \frac{\bar{x}\sqrt{200}}{4} = z_{1-\alpha^*}$$

$$\iff P(Z \leq \frac{\bar{x}\sqrt{200}}{4}) = 1 - \alpha^*$$

$$\iff \alpha^* = 1 - P(Z \leq \frac{\bar{x}\sqrt{200}}{4})$$

Plugging in $\bar{x} = 0.65$:

$$\alpha^* = 1 - P(Z \leq \frac{0.65\sqrt{200}}{4})$$

$$= 1 - P(Z \leq 2.298097)$$

$$= 1 - \text{pnorm}(2.298097)$$

$$= 0.01077813$$

using R. Thus, the p-value is $\boxed{\alpha^* = 0.01077813}$.

PROBLEM 5:

A researcher performs a sequence of independent experiments, up to and including the first 'success', after which the researcher stops. Each experiment has the same probability p of success. Let T be the number of experiments performed (including the first observed success).

The researcher wishes to test the null hypothesis

H_0: $p = 0.25$,

against the alternative hypothesis

H_1: $p > 0.25$.

The researcher proposes to reject the null hypothesis if $T < 4$.

(a) What is the significance level α of the test proposed by the researcher?

(b) What is the power of the researcher's test against the specific alternative hypothesis that $p = 0.5$.

(c) Re-express the researcher's rule for rejecting the null hypothesis in terms of the likelihood ratio:

$$\text{LRT} = p(t \mid p = 0.25)/p(t \mid p = 0.5).$$

Specifically, find the value ℓ such that the researcher will reject H_0 $p = 0.25$ in favor of $p = 0.5$ if LRT$< \ell$.

(Note: There will be a range of values for ℓ that will give the same test.)

Hint: (For all parts) Geometric distribution!

SOLUTION:

a) Recall:

$$\alpha = P(\text{reject } H_0 | H_0)$$

$$= P(T < 4 | H_0) = P(T \leq 3 | p = 0.25)$$

Now, we will use the fact that this is a Geometric distribution. The mass function for a Geometric distribution with parameter p is the following:

$$f(x) = \begin{cases} p(1-p)^{x-1} & \text{for } x \in \{1, 2, \dots\} \\ 0 & \text{otherwise} \end{cases}$$

Therefore,

$$\alpha = P(T = 3|p = 0.25) + P(T = 2|p = 0.25) + P(T = 1|p = 0.25)$$

$$= p(1-p)^2 + p(1-p) + p$$

$$= (0.25)(0.75)^2 + (0.25)(0.75) + 0.25$$

$$= \boxed{0.578125}$$

So the researcher proposes a significant level of $\alpha = 0.578125$.

---

b) $H_1 : p = 0.5$

We will calculate

$$\beta = P(\text{Fail to reject } H_0|H_1)$$

$$= P(T \geq 4|p = 0.5)$$

$$= 1 - (P(T = 3|p = 0.5) + P(T = 2|p = 0.5) + P(T = 1|p = 0.5))$$

$$= 1 - (0.5(0.5)^2 + (0.5 \times 0.5) + 0.5)$$

$$= 1 - 0.875 = 0.125$$

Then:

$$\text{Power} = 1 - \beta = 0.875.$$

We could also simply have calculated $P(T < 4|H_1) = P(T < 4|p = 0.5)$.

---

c)

$$\text{LRT} = \frac{p(t|p = 0.25)}{p(t|p = 0.5)}$$

$$= \frac{0.25(0.75)^{t-1}}{0.5(0.5)^{t-1}}$$

$$= 0.5(\frac{0.75}{0.5})^{t-1}$$

$$= \frac{1}{2}(\frac{3}{2})^{t-1}$$

Now, we reject the null when $T < 4 \implies T \leq 3$. So, let us consider $t = 3$.

Then,

$$\frac{p(3|p = 0.25)}{p(3|p = 0.5)} = \frac{1}{2}(\frac{3}{2})^2 = 1.125$$

Now, let us consider $t = 4$.

$$\frac{p(4|p = 0.25)}{p(4|p = 0.5)} = 1.6875$$

Therefore, for any $\boxed{l \in (1.125, 1.6875)}$ we will reject H_0 if LRT $< l$.

PROBLEM 6:

Consider a 95% confidence interval for the mean height μ in a population. Which of the following are true or false:

(a) Before taking our sample, the probability of the resulting 95% confidence interval containing μ is 0.95.

(b) If we take a sample and compute a 95% confidence interval for μ to be $[1.2, 3.7]$ then $P(\mu \in [1.2, 3.7]) = 0.95$.

(c) Before taking our sample, the center of a 95% confidence interval for the population mean is a random variable.

(d) 95% of individuals in the population have heights that lie in the 95% confidence interval for μ.

(e) Over hypothetical replications out of one hundred 95% confidence intervals for μ, on average 95 will contain μ.

(f) After obtaining our sample, the resulting confidence interval either does or does not contain μ.

SOLUTION:

a) True. μ will be in the 95% confidence interval in 95% of the samples and so the probability of the resulting 95% confidence interval containing μ is 95% $= 0.95$.

---

b) False. The 95% interval either does or does not contain μ.

---

c) True. The confidence interval is symmetric about $\bar{X}$ and so the center is $\bar{X}$. $\bar{X}$ is a random variable before we take the sample and will depend on the particular sample.

---

d) False. The 95% confidence interval for μ only tells that the interval was constructed by a procedure which will output an interval containing μ in 95% of samples.The 95% confidence interval does not give the range that contains 95% of the population values, which in this case are heights.

---

e) True. This is because μ will be in the 95% confidence interval for 95% of the samples. So, out of one hundred 95% confidence intervals over hypothetical replications will contain μ.

f) True. However, we do not whether it does or does not.

PROBLEM 7:

Goldberger Qu. 11.6 (Assume that the observations are drawn from a Normal Distribution and see Lecture 10, slide 46.)

A random sample from a population has $n = 30$, $\sum_i x_i = 120$ and $\sum_i x_i^2 = 8310$.

a) Calculate unbiased estimates of the population mean, the population variance, and the variance of the sample mean.

b) Provide an approximate 95% confidence interval for the population mean.

SOLUTION:
a) Using Goldberger Section 11.3, the unbiased estimator of the population mean is given by the following:

$$\bar{X} = \frac{1}{n}\sum_i x_i = \frac{120}{30} = \boxed{4}$$

Again, using Goldberger, the unbiased estimator of the population variance is given by:

$$S^{*2} = \frac{\sum_i (x_i - \bar{X})^2}{n-1} = \frac{n}{n-1}S^2$$

Notice:

$$\sum_i (x_i - \bar{X})^2 = \sum_i x_i^2 - 2x_i\bar{X} + \bar{X}^2 = \sum_i x_i^2 - 2\bar{X}\sum_i x_i + \sum_i \bar{X}^2$$

$$= 8310 - (8 \times 120) + (30 \times 16) = 7830$$

Therefore,

$$S^{*2} = \frac{7830}{29} = 270$$

Thus, the unbiased estimator of the population variance is $\boxed{270}$.

By the Sampling Mean Theorem, we have that the variance of the sample mean is given by:

$$\frac{S^{*2}}{n} = \frac{270}{30} = 9$$

where S^{*2} is the unbiased estimator of the population variance. Therefore, the variance of the sample mean is 9.

---

b) We will assume that the observations are drawn from a Normal distribution.

We have an unbiased estimate for σ^2 and we know $n = 30$. So, we know that:

$$P(|\bar{X} - \mu| \leq 1.96\frac{\sigma}{\sqrt{n}}) = 0.95$$

$$\implies P(-1.96\frac{\sigma}{\sqrt{n}} \leq \bar{X} - \mu \leq 1.96\frac{\sigma}{\sqrt{n}}) = 0.95$$

$$\implies P(\mu - 1.96\frac{\sigma}{\sqrt{n}} \leq \bar{X} \leq \mu + 1.96\frac{\sigma}{\sqrt{n}}) = 0.95$$

$$\implies P(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}) = 0.95$$

Then,

$$\mu \in \left[\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right]$$

is the 95% confidence interval for μ, the population mean.

Plugging in our values:

$$\mu \in \left[4 - 1.96\frac{\sqrt{270}}{\sqrt{30}}, 4 + 1.96\frac{\sqrt{270}}{\sqrt{30}}\right]$$

$$\implies \mu \in \Big[4 - (1.96 \times 3), 4 + (1.96 \times 3)\Big]$$

$$\implies \boxed{\mu \in \Big[-1.88, 9.88\Big]}$$

is the approximate 95% confidence interval for the population mean μ.