Homework 7

Stat 345 - Spring 2020

Name: _____

Problem 1 (20 pts)

Let $X_1, X_2, ..., X_n$ be a random sample of size from a distribution with probability density function

$$f(x) = \lambda x^{\lambda-1},\ 0 < x < 1,\ \lambda > 0$$

Note that the distribution of X is $Beta(\lambda, 1)$, you can use the derived expected value of Beta and not to calculate it as I did in a).

a) Get the method of moments estimator of λ. Calculate the estimate when $x_1 = 0.1, x_2 = 0.2, x_3 = 0.3$.

$$E(X) = \int_0^1 x\lambda x^{\lambda-1}dx = \int_0^1 \lambda x^{\lambda}dx = \frac{\lambda}{\lambda+1}$$

To find the MOM estimator, we need to equate population moment to the sample moment:

$$E(X) = \bar{X}$$

Therefore, we need to solve

$$\frac{\lambda}{\lambda+1} = \bar{X}$$

$$\hat{\lambda} = \frac{\bar{X}}{1-\bar{X}}$$

The MOM estimate when $x_1 = 0.1, x_2 = 0.2, x_3 = 0.3$ is

$$\hat{\lambda} = \frac{\bar{x}}{1-\bar{x}} = \frac{0.2}{1-0.2} = 0.25$$

b) Get the maximum likelihood estimator of λ. Calculate the estimate when $x_1 = 0.1, x_2 = 0.2, x_3 = 0.3$.

The likelihood is given by

$$L(\lambda) = p(x_1)p(x_2)...p(x_n) = \prod_{i=1}^{n} \lambda x_i^{\lambda-1}$$

The log likelihood can be obtained by taking the natural logarithm of $L(\lambda)$:

$$logL(\lambda) = log\left(\prod_{i=1}^{n} \lambda x_i^{\lambda-1}\right) = \sum_{i=1}^{n} log\left(\lambda x_i^{\lambda-1}\right) = \sum_{i=1}^{n}\left(log(\lambda) + log(x_i^{\lambda-1})\right) =$$

$$= \sum_{i=1}^{n} \Big(log(\lambda) + (\lambda - 1) log(x_i) \Big) = nlog(\lambda) + (\lambda - 1) \sum_{i=1}^{n} log(x_i)$$

Taking the derivative with respect to λ

$$\frac{d \log L(\lambda)}{d\lambda} = \frac{n}{\lambda} + \sum_{i=1}^{n} log(x_i)$$

Equating to zero

$$\frac{n}{\lambda} + \sum_{i=1}^{n} log(x_i) = 0$$

$$\hat{\lambda} = -\frac{n}{\sum_{i=1}^{n} log(X_i)}$$

The MLE when $x_1 = 0.1, x_2 = 0.2, x_3 = 0.3$ is

$$\hat{\lambda} = -\frac{n}{\sum_{i=1}^{n} log(x_i)} = -\frac{3}{\sum_{i=1}^{3} log(x_i)} = -\frac{3}{log(0.1) + log(0.2) + log(0.3)} = 0.586$$

Note that MLE is different from the MOM here.

Problem 2 (30 pts)

Suppose that compressive strength is normally distributed with $\sigma^2 = 1000\ (psi)^2$. A random sample of 12 specimens has a mean compressive strength of $\bar{x} = 3250$ psi.

a) Construct a 95% and 99% two-sided confidence intervals on mean compressive strength. Compare their widths.

The distribution is Normal with σ^2 known, that is

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim Normal(0, 1)$$

The $100(1-\alpha)\%$ CI is

$$P(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}) = 1 - \alpha, \quad 0 \leq \alpha \leq 1$$

$$P(\bar{X} - z_{\alpha/2}\sigma/\sqrt{n} \leq \mu \leq \bar{X} + z_{\alpha/2}\sigma/\sqrt{n}) = 1 - \alpha$$

$$P(3250 - z_{\alpha/2}\sqrt{1000}/\sqrt{12} \leq \mu \leq 3250 + z_{\alpha/2}\sqrt{1000}/\sqrt{12}) = 1 - \alpha$$

The 95% CI is

$$P(3250 - z_{0.05/2}\sqrt{1000}/\sqrt{12} \leq \mu \leq 3250 + z_{0.05/2}\sqrt{1000}/\sqrt{12}) = 1 - 0.05$$

$$P(3250 - 1.96\sqrt{1000}/\sqrt{12} \leq \mu \leq 3250 + 1.96\sqrt{1000}/\sqrt{12}) = 0.95$$

$$\mu \in [3232.11, 3267.89]$$

The 99% CI is

$$P(3250 - z_{0.01/2}\sqrt{1000}/\sqrt{12} \leq \mu \leq 3250 + z_{0.01/2}\sqrt{1000}/\sqrt{12}) = 1 - 0.01$$

$$P(3250 - 2.576\sqrt{1000}/\sqrt{12} \leq \mu \leq 3250 + 2.576\sqrt{1000}/\sqrt{12}) = 0.99$$

$$\mu \in [3226.48, 3273.52]$$

The 99% CI is wider (the interval covers more values so we are more confident) than the 95% CI.

b) Suppose it is desired to estimate the compressive strength with an error that is less than 15 psi at 99% confidence. What sample size is required?

If $\bar{x}$ is used as an estimate of μ, we can be $100(1-\alpha)\%$ confident that the error $|\bar{x} - \mu|$ will not exceed a specified amount E when the sample size is

$$n = \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2 = \left(\frac{z_{0.01/2}\sqrt{1000}}{15}\right)^2 = \left(\frac{2.576\sqrt{1000}}{15}\right)^2 = 29.49$$

if n is not an integer, it must be rounded up, so we require $n = 30$.

Problem 3 (15 pts + bonus 5 pts)

A healthcare provider monitors the number of CAT scans performed each month in each of its clinics. The most recent year of data for a particular clinic follows (the reported variable is the number of CAT scans each month expressed as the number of CAT scans per thousand members of the health plan):

$$2.31, 2.09, 2.36, 1.95, 1.98, 2.25, 2.16, 2.07, 1.88, 1.94, 1.97, 2.02$$

a) Find a 95% two-sided CI on the mean number μ of CAT scans performed each month at this clinic.

The distribution is approximately normal but σ^2 is unknown. The $100(1-\alpha)\%$ CI is

$$P(-t_{\alpha/2,n-1} \leq \frac{\bar{X} - \mu}{s/\sqrt{n}} \leq t_{\alpha/2,n-1}) = 1 - \alpha, \quad 0 \leq \alpha \leq 1$$

$$P(\bar{X} - t_{\alpha/2,n-1}s/\sqrt{n} \leq \mu \leq \bar{X} + t_{\alpha/2,n-1}s/\sqrt{n}) = 1 - \alpha$$

Since

$$\bar{x} = \frac{\sum_{i=1}^{12} x_i}{12} = 2.08, \; s = \frac{\sum_{i=1}^{12}(x_i - \bar{x})^2}{n-1} = 0.156$$

The 95% CI is

$$P(2.08 - t_{0.05/2,12-1}0.156/\sqrt{12} \leq \mu \leq 2.08 + t_{0.05/2,12-1}0.156/\sqrt{12}) = 1 - 0.05$$

R: $t_{0.05/2,12-1}$ is $qt(1 - .05/2, 12 - 1) = 2.201$

$$2.08 - 2.201(0.156)/\sqrt{12} \leq \mu \leq 2.08 + 2.201(0.156)/\sqrt{12}$$

$$1.98 \leq \mu \leq 2.18$$

The 95% CI for the population mean number μ of CAT scans lies between 1.98 and 2.18.

b) (bonus) Historically, the mean number of scans performed by all clinics in the system has been $\mu_0 = 1.95$. If there any evidence that this particular clinic performs more CAT scans on average than the overall system average?

Conduct hypothesis testing

$$H_0 : \mu = 1.95, \quad H_1 : \mu > 1.95$$

Test the hypothesis at significance level $\alpha = 0.05$. You can calculate the test statistic and based on its value draw a conclusion. The 95% CI can also be used to make a desicion. Since the 95% CI

$$1.98 \leq \mu \leq 2.18$$

does not contain the mean number $\mu = \mu_0 = 1.95$, we can reject the null hypothesis H_0 in favor of H_1. With 95% confidence, the particular clinic performs more CAT scans on average than the overall system average.

Problem 4 (30 pts)

Medical researchers have developed a new artificial heart constructed primarily of titanium and plastic. The heart will last and operate almost indefinitely once it is implanted in the patient's body, but the battery pack needs to be recharged about every four hours. A random sample of 50 battery packs is selected and subjected to a life test. The average life of these batteries is 4.05 hours. Assume that battery life is normally distributed with standard deviation $\sigma = 0.2$ hour.

We know $\bar{x} = 4.05, n = 50, \sigma = 0.2$.

a) Is there evidence to support the claim that mean battery life differs from 4 hours? Use $\alpha = 0.05$.

$$H_0 : \mu = 4, \quad H_1 : \mu \neq 4$$

The disribution is normal with known σ. We will reject H_0 if

$$\left| \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \right| > z_{\alpha/2}$$

$$\left| \frac{4.05 - 4}{0.2/\sqrt{50}} \right| > z_{0.05/2}$$

R: $z_{0.05/2} : qnorm(1 - 0.05/2) = 1.96$ Since

$$|1.77| < 1.96$$

we fail to reject H_0. There is no evidence to support the claim that mean battery life differs from 4 hours at $\alpha = 0.05$.

b) Compute the power of the test if the true mean battery life is 4.15 hours.

The true value of the mean is $\mu = 4.15$ so $\delta = \mu - \mu_0 = 4.15 - 4 = 0.15$.

We can find the type II error β as

$$\beta = \Phi\left(z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma}\right) - \Phi\left(-z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma}\right) = \beta = \Phi\left(1.96 - \frac{0.15\sqrt{50}}{0.2}\right) - \Phi\left(-1.96 - \frac{0.15\sqrt{50}}{0.2}\right) =$$

$$\Phi(-3.143) - \Phi(-7.263) = 0.0008$$

R: $pnorm(-3.143) - pnorm(-7.263) = 0.0008361292$. The power is $1 - \beta = 0.9992$ which is very good.

c) What sample size would be required to detect a true mean battery life of 4.15 hours if you wanted the power of the test to be at least 0.99?

The sample size for two-sided test ($H_1 : \mu \neq \mu_0$) on the mean, variance known:

$$n = \frac{(z_{\alpha/2} + z_\beta)^2 \sigma^2}{\delta^2}$$

Here, $\alpha = 0.05, \beta = 1 - 0.99 = 0.01, \sigma = 0.2, \delta = 4.15 - 4 = 0.15$, therefore

$$n = \frac{(z_{0.05/2} + z_{0.01})^2 0.2^2}{(0.15)^2} = (1.96 + 2.576)^2 1.78 = 36.62$$

R: $z_{0.05/2} : qnorm(1 - 0.05/2) = 1.96, z_{0.01/2} : qnorm(1 - 0.01/2) = 2.576$. We require the sample size $n = 37$ so the power of the test to be at least 0.99.

Problem 5 (15 pts)

In a random sample of 500 handwritten zip code digits, 466 were read correctly by an optical character recognition (OCR) system operated by the U.S. Postal Service (USPS). USPS would like to know whether the rate is at least 90% correct. Do the data provide evidence that the rate is at least 90% at $\alpha = 0.01$?

Hypotheses: $H_0 : p = 0.9$ and $H_1 : p > 0.9$

The test statistic is

$$Z = \frac{X - np_0}{\sqrt{np_0(1-p_0)}} = \frac{466 - (500)0.9}{\sqrt{(500)0.9(1-0.9)}} = 2.385$$

where $Z \sim N(0, 1)$

We reject H_0 if

$$\left| \frac{X - np_0}{\sqrt{np_0(1-p_0)}} \right| > z_{\alpha/2}$$

Since

$$|2.385 > 1.96|$$

we reject the H_0 and conclude that there is a sufficient evidence that the rate is at least 90% at 5% level of significance.

Problem 6 (10 pts + bonus 5 pts)

What is the difference between commuting patterns for students and professors? A study compares mean commuting distances (in miles) for students and professors. Assume not equal variances: $\sigma_1^2 \neq \sigma_2^2$. Summary statistics:

	n	$\bar{x}$	s
students	38	6.8	4.8
professors	40	11.2	7.2

a) Determine whether there is any difference between mean commuting distances at $\alpha = 5\%$ level of significance?

Case: $\sigma_1^2 \neq \sigma_2^2$

$$H_0 : \mu_1 - \mu_2 = 0 \quad H_1 : \mu_1 - \mu_2 \neq 0$$

The statistic T_0^* is distributed approximately as t with degrees of freedom v

$$T = \frac{\bar{X}_1 - \bar{X}_2 - 0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

$$T = \frac{6.8 - 11.2 - 0}{\sqrt{\frac{4.8^2}{38} + \frac{7.2^2}{40}}} = -3.19$$

$$v = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

$$v = \frac{\left(\frac{4.8^2}{38} + \frac{7.2^2}{40}\right)^2}{\frac{(4.8^2/38)^2}{38-1} + \frac{(7.2^2/40)^2}{40-1}} = 68.28$$

If v is not an integer, round down to the nearest integer, so $v = 68$. Since

$$|T| = 3.19 > t_{0.05/2,68} = 1.995$$

R: $qt(1 - 0.05/2, 68) = 1.995$

we reject the null hypothesis and conclude that the means of commuting distances for students and professors are different.

b) (bonus) Construct a 95% confidence interval for the difference between commuting patterns for students and professors. How the result you have observed is connected to (a)?

$$\bar{x}_1 - \bar{x}_2 - t_{\alpha/2,v}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{x}_1 - \bar{x}_2 + t_{\alpha/2,v}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$6.8 - 11.2 - 1.995\sqrt{\frac{4.8^2}{38} + \frac{7.2^2}{40}} \leq \mu_1 - \mu_2 \leq 6.8 - 11.2 + 1.995\sqrt{\frac{4.8^2}{38} + \frac{7.2^2}{40}}$$

$$-7.152 \leq \mu_1 - \mu_2 \leq -1.648$$

The confidence interval doesn't include 0 meaning that there is a difference in two means. In particular, the mean commuting distance of students is smaller than that of professors (on average, students tend to live closer to the campus than professors).

Problem 7 (Bonus 10 pts)

Let $X_1, X_2, ..., X_n$ be a random sample of size from a distribution with probability density function

$$f(x) = \lambda x^{\lambda - 1},\ 0 < x < 1,\ \lambda > 0$$

Note that the distribution of X is $Beta(\lambda, 1)$.

a) (bonus) Show that $T = X^\lambda$ is a pivotal quantity. (*Hint:* You need to find cdf and then pdf of T and show that the distribution doesn't depend on λ. The last slide in the lecture 17 might help.)

$$P(T \leq t) = 1 - P(X^\lambda \leq t) = P(X \leq t^{1/\lambda}) = \int_0^{t^{1/\lambda}} \lambda x^{\lambda-1} dx = t - 1$$

The pdf is then $f(t) = (t-1)' = 1$. The distribution doesn't depend on λ, $T = X^\lambda \sim Uniform(0,1)$.

b) (bonus) Construct 95% confidence interval for λ using the pivotal quantity in (a). You don't need to find optimal q_1, q_2. Just write down how you can approach the problem.

To find 95% CI, we need to choose q_1 and q_2 such that

$$P(q_1 \leq X^\lambda \leq q_2) = 0.95$$

$$P(q_1 \leq T = X^\lambda \leq q_2) = \int_{q_1}^{q_2} f(t)dt = \int_{q_1}^{q_2} 1dt = q_2 - q_1 = 0.95$$

$$P(log(q_1) \leq \lambda log(X) \leq log(q_2)) = q_2 - q_1 = 0.95$$

Since $log(X) < 0$

$$P\left(\frac{log(q_2)}{log(X)} \leq \lambda \leq \frac{log(q_1)}{log(X)}\right) = q_2 - q_1 = 0.95$$

So $q_2 - q_1 = 0.95$ must hold and $0 < q_1, q_2 < 1$ because $T \sim Uniform(0,1)$.

The length of the interval is $\left(\frac{log(q_1)}{log(X)} - \frac{log(q_2)}{log(X)}\right)$. We can do anything with $log(X)$ because it depends on the data but we can minimize $log(q_1) - log(q_2)$ (should be negative, because $\lambda > 0$), subject to the constraint $q_2 - q_1 = 0.95$. The solution is $q_2 = 1, q_1 = 0.05$.

$$\lambda \in \left[\frac{log(q_2)}{log(X)}, \frac{log(q_1)}{log(X)}\right]$$

Among 95% CIs the shortest one is $\lambda \in \left[0, \frac{-3}{log(X)}\right]$. For example, if $x = 0.5$, then $\lambda \in [0, 4.33]$.