Stats 100B: Homework #6

Professor Nicolas Christou
Assignment: 1-10

Eric Chuu
UID: 604406828

March 3, 2017

Exercise 1.
(a) Find the MLE estimates for the mean and variance, μ, σ^2.
(b) Give the 90%, 95%, 99% confidence intervals for μ, σ^2. In constructing the confidence interval for the population variance, use the unbiased estimate for σ^2, not the MLE.
(c) Using the results from (b), give 90%, 95%, 99% confidence intervals for σ.
(d) How much larger a sample do you think you would need to halve the length of the interval for μ?

Solution.
(a) We write the log-likelihood function and differentiate with respect to μ and σ^2 to find the MLEs. Since the samples are independent and drawn from a normal distribution, we can write

$$L = \prod_{i=1}^{n} (2\pi\sigma^2)^{-\frac{1}{2}} \cdot \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu)^2\right) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right)$$

$$\ln L = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

$$\frac{\partial \ln L}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) = 0, \quad \Rightarrow \hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{X}$$

$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2 = 0 \quad \Rightarrow \hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n}$$

(b) We can use `R` to generate the three confidence intervals for μ. The results are shown below

`t.test(x, conf.level = 0.90)` : $[2.262, 4.596]$
`t.test(x, conf.level = 0.95)` : $[2.801, 4.421]$
`t.test(x, conf.level = 0.99)` : $[2.249, 4.972]$

We can use the chi-square distribution to find the confidence interval for the population variance. In particular,

$$\mathbf{Pr}\left(\chi^2_{\frac{\alpha}{2};n-1} \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi^2_{1-\frac{\alpha}{2};n-1}\right) = 1 - \alpha$$

$$\mathbf{Pr}\left(\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2};n-1}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2};n-1}}\right) = 1 - \alpha$$

The the confidence intervals for confidence levels 90%, 95%, and 99% are as follows

$$\mathbf{Pr}\left(2.05 \leq \sigma^2 \leq 7.06\right) = 0.90$$
$$\mathbf{Pr}\left(1.87 \leq \sigma^2 \leq 8.192\right) = 0.95$$
$$\mathbf{Pr}\left(1.563 \leq \sigma^2 \leq 11.144\right) = 0.99$$

(c) Using the confidence intervals calculated from part (b), we can construct the confidence intervals for σ:

$$\mathbf{Pr}\left(1.43 \leq \sigma \leq 2.66\right) = 0.90$$
$$\mathbf{Pr}\left(1.37 \leq \sigma \leq 2.86\right) = 0.95$$
$$\mathbf{Pr}\left(1.25 \leq \sigma \leq 3.34\right) = 0.99$$

(d) Consider the confidence interval for μ with confidence level $1-\alpha$.

$$\mathbf{Pr}\left(\bar{X} - t_{1-\frac{\alpha}{2};n-1} \cdot \frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{1-\frac{\alpha}{2};n-1} \cdot \frac{s}{\sqrt{n}}\right) = 1-\alpha$$

In order to halve the interval, the margin of error needs to be halved. This can be accomplished by using a sample size of $4n$, which would scale the denominator by a factor of 2. In this case, we would need a sample size of $4 \cdot 16 = 64$. □

Exercise 2. Given $\bar{X} - \bar{Y}$ having mean $\mu_1 - \mu_2$ and variance $\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}$. Show the steps needed to construct a $1-\alpha$ confidence level for $\mu_1 - \mu_2$. Assume that σ_1, σ_2 are known.

Solution. We're given the distribution of $\bar{X} - \bar{Y}$, so we can standardize it and get following:

$$\mathbf{Pr}\left(-z_{1-\frac{\alpha}{2}} \leq \frac{(\bar{X}-\bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} \leq z_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$

Multiplying through by the standard deviation and isolating $\mu_1 - \mu_2$, we get the following confidence interval

$$\mathbf{Pr}\left(-z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}} + (\bar{X}-\bar{Y}) \leq \mu_1 - \mu_2 \leq z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}} + (\bar{X}-\bar{Y})\right) = 1-\alpha$$

□

Exercise 3. If X_1, X_2 are independent random variables having, respectively, binomial distributions with parameters n_1, p_1 and n_2, p_2, construct a $1-\alpha$ confidence level for $p_1 - p_2$.

Solution. We consider the distributions of $\frac{X_1}{n_1}$ and $\frac{X_2}{n_2}$.

$$\mathrm{E}\left(\frac{X_1}{n_1}\right) = \frac{1}{n_1} \cdot n_1 p_1 = p_1, \quad \mathrm{Var}\left(\frac{X_1}{n_1}\right) = \frac{1}{n_1^2} \cdot n_1 p_1(1-p_1) = \frac{p_1(1-p_1)}{n_1}$$
$$\mathrm{E}\left(\frac{X_2}{n_2}\right) = \frac{1}{n_2} \cdot n_2 p_2 = p_2, \quad \mathrm{Var}\left(\frac{X_2}{n_2}\right) = \frac{1}{n_2^2} \cdot n_2 p_2(1-p_2) = \frac{p_2(1-p_2)}{n_2}$$

By the Central Limit Theorem, we know that when n_1, n_2 are sufficiently large, the distribution $\frac{X_1}{n_1}, \frac{X_2}{n_2}$ can be approximated by the normal distribution, with their respective mean and variance. Thus, the distribution of $\frac{X_1}{n_1} - \frac{X_2}{n_2}$ is given by

$$\left(\frac{X_1}{n_1} - \frac{X_2}{n_2}\right) \sim N\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)$$

Then the confidence interval for $p_1 - p_2$ can be calculated as follows. Note that since p_1, p_2 are unknown, we use $\hat{p_1} = \frac{X_1}{n_1}, \hat{p_2} = \frac{X_2}{n_2}$

$$\mathbf{Pr}\left(-z_{1-\frac{\alpha}{2}} \leq \frac{\left(\frac{X_1}{n_1} - \frac{X_2}{n_2}\right) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \leq z_{1-\frac{\alpha}{2}} = 1 - \alpha\right)$$

$$\mathbf{Pr}\left(-z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} + \left(\frac{X_1}{n_1} - \frac{X_2}{n_2}\right) \leq p_1 - p_2 \leq z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} + \left(\frac{X_1}{n_1} - \frac{X_2}{n_2}\right)\right) = 1-\alpha$$

$$\mathbf{Pr}\left(-z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{X_1(1-\frac{X_1}{n_1})}{n_1^2} + \frac{X_2(1-\frac{X_2}{n_2})}{n_2^2}} + \left(\frac{X_1}{n_1} - \frac{X_2}{n_2}\right) \leq p_1 - p_2 \leq z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{X_1(1-\frac{X_1}{n_1})}{n_1^2} + \frac{X_2(1-\frac{X_2}{n_2})}{n_2^2}} + \left(\frac{X_1}{n_1} - \frac{X_2}{n_2}\right)\right) = 1-\alpha$$

Exercise 4. The manager of a supermarket would like to know the average time that a person checkout counter. Using a stopwatch, he observes 100 customers. He computed the sample mean $\bar{x} = 15.35$ minutes and the sample standard deviation to be $s = 6.1$ minutes.

(a) Construct a 95% confidence interval for the population mean μ.
(b) Suppose that the manager wants a smaller error in estimation. He wants his error to be ± 1 minute with 95% confidence. How many customers will he need? Assume $\sigma = 6.1$.

Solution.
(a) Since $\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t_{n-1}$, we can construct the following confidence interval

$$1 - \alpha = \mathbf{Pr}\left(-t_{1-\frac{\alpha}{2};n-1} \leq \frac{\bar{X} - \mu}{s/\sqrt{n}} \leq t_{1-\frac{\alpha}{2};n-1}\right) = \mathbf{Pr}\left(\bar{X} - t_{1-\frac{\alpha}{2};n-1} \cdot \frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{1-\frac{\alpha}{2};n-1} \cdot \frac{s}{\sqrt{n}}\right)$$

$$0.95 = \mathbf{Pr}\left(15.25 - 1.98 \cdot \frac{6.1}{\sqrt{100}} \leq \mu \leq 15.25 + 1.98 \cdot \frac{6.1}{\sqrt{100}}\right)$$
$$= \mathbf{Pr}\left(14.142 \leq \mu \leq 16.558\right)$$

(b) Assuming $\sigma = 6.1$, then we use the definition of the margin of error and find the necessary sample size:

$$E = z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \quad \Rightarrow \quad n = z_{1-\frac{\alpha}{2}}^2 \cdot \frac{\sigma^2}{E^2} = 1.96^2 \cdot \frac{6.1^2}{1} = 143$$

□

Exercise 5. A precision instrument is guaranteed to read accurately to within 2 units. A sample of 4 instrument readings on the same object yield the measurements 353, 351, 351, 355. Find the 90% confidence interval for the population variance. What assumptions are necessary. Does the guarantee seem reasonable?

Solution. Under assumptions of normality and independence of the samples, we can find the 90% confidence interval for the population variance using the chi-square distribution:

$$\mathbf{Pr}\left(\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2};n-1}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2};n-1}}\right) = 1-\alpha$$

$$\mathbf{Pr}\left(\frac{3 \cdot 3.667}{7.815} \leq \sigma^2 \leq \frac{3 \cdot 3.667}{0.352}\right) = 0.90$$

$$\mathbf{Pr}\left(1.41 \leq \sigma^2 \leq 31.25\right) = 0.90$$

Given the large confidence interval for the variance and the small sample size, the guarantee does not seem reasonable. □

Exercise 6. Recently there have been discussions about constructing a subway system that would run from Downtown Los Angeles to Santa Monica through Wilshire Boulevard. Suppose a random sample of 900 voters in Hollywood indicates that 600 support such an idea.
(a) Construct a 95% confidence interval for the Hollywood population proportion of residents who would support this idea.
(b) Suppose that the City of LA wants to estimate with 95% confidence the percentage of residents who would support this idea in Hollywood. The city wants the error of estimation to be ±2% of the population proportion. What is the minimum sample size required?
(c) Suppose that the city of LA wants to estimate with 95% confidence the percentage of residents who would support this idea in Westwood. The city wants the error of estimation to be ±2% of the population proportion. What is the minimum sample size required? Assume that there is no prior information about the population proportion.

Solution
(a) Let p be the of Hollywood population proportion of residents who support the idea and X be the number of people from the sample who support the idea. Since the population proportion is unknown, we use the estimate $\hat{p} = \frac{X}{n} = \frac{2}{3}$. Then the confidence interval for the population proportion is given by

$$\mathbf{Pr}\left(\hat{p} - z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = 1-\alpha$$

$$\mathbf{Pr}\left(\frac{2}{3} - 1.96 \cdot \sqrt{\frac{\frac{2}{3}(\frac{1}{3})}{900}} \leq p \leq \frac{2}{3} + 1.96 \cdot \sqrt{\frac{\frac{2}{3}(\frac{1}{3})}{900}}\right) = 0.95$$

$$\mathbf{Pr}\left(0.6359 \leq p \leq 0.6975\right) = 0.95$$

(b) If we want the margin of error to be within ±2% of he population proportion, then we can solve for the sample size as follows

$$E^2 = z^2_{1-\frac{\alpha}{2}} \cdot \frac{\hat{p}(1-\hat{p})}{n} \Rightarrow \quad n = z^2_{0.975} \cdot \frac{\frac{2}{3} \cdot \frac{1}{3}}{0.02^2} = 2134$$

(c) If we have no prior information about the population proportion, then $\hat{p} = \frac{1}{2}$, and the sample size is calculated as follows:

$$n = z^2_{0.975} \cdot \frac{\frac{1}{2} \cdot \frac{1}{2}}{0.02^2} = 2401$$

□

Exercise 7. Show that two independent random samples of n_1, n_2 observations are selected from normal populations with means μ_1, μ_2, and variances σ_1^2, σ_2^2 respectively. Find a confidence interval for the variance ratio $\frac{\sigma_1^2}{\sigma_2^2}$ with confidence level $1-\alpha$.

Solution. Since $\frac{(n_1-1)S_x^2}{\sigma_1^2} \sim \chi^2_{n_1-1}$ and $\frac{(n_2-1)S_y^2}{\sigma_2^2} \sim \chi^2_{n_2-1}$, then

$$\frac{\frac{(n_2-1)S_y^2}{\sigma_2^2}/(n_2-2)}{\frac{(n_1-1)S_x^2}{\sigma_1^2}/(n_1-1)} = \frac{\sigma_1^2}{\sigma_2^2} \cdot \frac{s_y^2}{s_x^2} \sim F_{n_2-1,n_1-1}$$

Then we con construct the following confidence interval

$$\mathbf{Pr}\left(F_{\frac{\alpha}{2};n_2-1;n_1-1} \leq \frac{\sigma_1^2}{\sigma_2^2} \cdot \frac{s_y^2}{s_x^2} \leq F_{1-\frac{\alpha}{2};n_2-1;n_1-1}\right) = 1-\alpha$$
$$\mathbf{Pr}\left(F_{\frac{\alpha}{2};n_2-1;n_1-1} \cdot \frac{s_x^2}{s_y^2} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq F_{1-\frac{\alpha}{2};n_2-1;n_1-1} \cdot \frac{s_x^2}{s_y^2}\right) = 1-\alpha$$

□

Exercise 8. The sample mean $\bar{X}$ is a good estimator of the population mean μ. It can also be used to predict a future value of X independently selected from the population. Assume that you have a sample mean $\bar{x}$ and sample variance s^2, based on a random sample of n measurements from a normal population. Construct a prediction interval for a new observation x, say x_p. Use $1-\alpha$ confidence level.

Solution. Let σ^2 be the variance of the population. We consider the quantity $X_p - \bar{X}$ and note that

$$X_p - \bar{X} \sim N\left(0, \sigma\sqrt{1+\frac{1}{n}}\right)$$

Then we can construct a variable that follows the t-distribution with $n-1$ degrees of freedom.

$$\frac{\frac{X_p-\bar{X}}{\sigma\sqrt{1+\frac{1}{n}}}}{\sqrt{\frac{\frac{(n-1)s^2}{\sigma^2}}{(n-1)}}} = \frac{X_p - \bar{X}}{s\sqrt{1+\frac{1}{n}}} \sim t_{n-1}$$

Then we can construct the prediction interval with $1-\alpha$

$$\mathbf{Pr}\left(-t_{1-\frac{\alpha}{2};n-1} \leq \frac{X_p - \bar{X}}{s\sqrt{1+\frac{1}{n}}} \leq t_{1-\frac{\alpha}{2};n-1}\right) = 1-\alpha$$

$$\mathbf{Pr}\left(-t_{1-\frac{\alpha}{2};n-1} \cdot s\sqrt{1+\frac{1}{n}} + \bar{X} \leq X_p \leq t_{1-\frac{\alpha}{2};n-1} \cdot s\sqrt{1+\frac{1}{n}} + \bar{X}\right) = 1-\alpha$$

□

Exercise 9. Assume that the Poisson Distribution with unknown parameter λ would be a plausible model for describing the variability from grid square to grid square in this situation.

(a) Use the method of maximum likelihood to estimate the parameter λ.
(b) Use the asymptotic properties of the MLE to construct a 95% confidence interval for λ.

Solution.
(a) Since each of the $X_i \sim (\lambda)$, then $f(x_i) = \frac{\lambda^{x_i}}{x_i!}e^{-\lambda}$. The X_i's are independent, so we can write the likelihood function

$$L = \prod_{i=1}^{n} f(x_i;\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!}e^{-\lambda} = \frac{\lambda^{x_1+x_2+\cdots x_n}e^{-n\lambda}}{\prod_{i=1}^{n} x_i!}$$

$$\ln L = \sum_{i=1}^{n} x_i \cdot \ln\lambda - n\lambda - \sum_{i=1}^{n} \ln(x_i!)$$

$$\frac{\partial \ln L}{\partial \lambda} = \frac{\sum_{i=1}^{n} x_i}{\lambda} - n = 0, \quad \Rightarrow \hat{\lambda} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{X}$$

In the context of the problem, we can estimate the parameter λ with

$$\hat{\lambda} = \frac{\sum_{i=1}^{23} x_i}{n} = 24.91$$

(b) We use the idea that for large samples, the distribution of $\sqrt{nI(\theta)}(\hat{\theta}-\theta)$ is approximately the standard normal. We calculate the second partial derivative of the log-pmf and calculate the Fisher Information for λ

$$\ln f(x) = x\ln(\lambda) - \lambda - \ln(x!)$$

$$\frac{\partial \ln f(x)}{\partial \lambda} = \frac{x}{\lambda} - 1$$

$$\frac{\partial^2 \ln f(x)}{\partial \lambda^2} = -\frac{x}{\lambda^2}$$

$$I(\lambda) = -\mathrm{E}\left(\frac{\partial^2 \ln f(x)}{\partial \lambda^2}\right) = \mathrm{E}\left(\frac{X}{\lambda^2}\right) = \frac{1}{\lambda}$$

Since λ is unknown, we use the maximum likelihood estimate, $\hat{\lambda}$ in the expression of the $I(\lambda)$. Thus,

$$\sqrt{nI(\hat{\lambda})} \cdot (\hat{\lambda} - \lambda) = \sqrt{\frac{n}{\hat{\lambda}}} \cdot (\hat{\lambda} - \lambda) \sim Z(0,1)$$

We can then use this to construct a confidence interval for with $1-\alpha$ confidence.

$$\mathbf{Pr}\left(-z_{1-\frac{\alpha}{2}} \le \sqrt{\frac{n}{\hat{\lambda}}} \cdot (\hat{\lambda}-\lambda) \le z_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$

$$\mathbf{Pr}\left(\hat{\lambda} - z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{\lambda}}{n}} \le \lambda \le \hat{\lambda} + z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{\lambda}}{n}}\right) = 1-\alpha$$

For 95% confidence, we have

$$\mathbf{Pr}\left(24.91 - 1.96 \cdot \sqrt{\frac{24.91}{23}} \le \lambda \le 24.91 + 1.96 \cdot \sqrt{\frac{24.91}{23}}\right) = 0.95$$

Then we can say the following:

$$\lambda \in [22.87, 26.95] \quad \text{with 95\% confidence.}$$

□

Exercise 10. Use `R` to access the data from the Maas river.
(a) Use `R` to compute the sample mean and sample standard deviation of lead.
(b) Construct a 95% confidence interval for the population mean of lead in this area.
(c) Based on the confidence level from (b), in which category does the soil of this area fall in terms of the ppm concentration of lead?
(d) Do you see any problems in these calculations (meaning by just using the averages)?

Solution.
(a) Using `R` the mean of lead is 153.36 and the sample standard deviation is 111.32.
(b) Using the function `t.test()`, we can construct a 95% confidence interval for the population mean of lead in this area:

$$\mu \in [135.6976, 171.0250], \quad \text{with 95\% confidence}$$

(c) Based on the confidence interval from (b), the soil falls in the categories lead-free, and lead-safe.
(d) By just using the averages, we fail to account for areas of high concentration, since the average is less informative about specific regions.

□