Assignment 1
MATH 324 - Statistics
Prof. Masoud Asgharian
Winter 2019

LE, Nhat Hung
McGill ID: 260793376
Date: January 13, 2019
Due date: January 23, 2019

8.9 p. 395

Suppose that $Y_1, Y_2, ..., Y_n$ constitute a random sample from a population with probability density function

$$f(y) = \begin{cases} \left(\frac{1}{\theta+1}\right) e^{-y/(\theta+1)}, & y > 0, \theta > -1, \\ 0, & \text{elsewhere.} \end{cases}$$

Suggest a suitable statistic to use as an unbiased estimator for θ. [*Hint:* Consider $\overline{Y}$.]

$$Y \sim \text{Exponential with } \beta = \theta + 1$$

In an exponential distribution, β is the mean.
Therefore,

$$E(\overline{Y}) = \theta + 1 \Rightarrow E(\overline{Y} - 1) = \theta \Rightarrow \overline{Y} - 1 \text{ unbiased estimator of } \theta$$

8.10 p. 395

The number of breakdowns per week for a type of minicomputer is a random variable Y with a Poisson distribution and mean λ. A random sample $Y_1, Y_2, ..., Y_n$ of observations on the weekly number of breakdowns is available.

a. Suggest an unbiased estimator for λ.

$$E(\overline{Y}) = \lambda \Rightarrow \overline{Y} \text{ unbiased estimator of } \lambda$$

b. The weekly cost of repairing these breakdowns is $C = 3Y + Y^2$. Show that $E(C) = 4\lambda + \lambda^2$.

$$E(C) = 3E(Y) + E(Y^2) = 3\lambda + \text{Var}(Y) + E(Y)^2 = 3\lambda + \lambda + \lambda^2 = 4\lambda + \lambda^2 \quad \square$$

c. Find a function of $Y_1, Y_2, ..., Y_n$ that is an unbiased estimator of $E(C)$. [*Hint:* Use what you know about $\overline{Y}$ and $(\overline{Y})^2$.]

Let $\widehat{\theta}$ unbiased estimator of $E(C) = 4\lambda + \lambda^2$.

Want

$$E(\hat{\theta}) = 4\lambda + \lambda^2.$$

Consider $\overline{Y}$ and $(\overline{Y})^2$,

$$E(\overline{Y}) = \lambda$$

$$E(\overline{Y}^2) = \text{Var}(\overline{Y}) + E(\overline{Y})^2 = \text{Var}\left(\frac{1}{n}\sum Y_i\right) + \lambda^2 = \frac{1}{n^2}\text{Var}\left(\sum Y_i\right) + \lambda^2 = \frac{1}{n^2}\left(\sum \text{Var}(Y_i) + 2\sum\sum_{1\le i<j\le n} \text{Cov}(Y_i, Y_j)\right) + \lambda^2 = \frac{1}{n^2} n\sigma^2 + \lambda^2 = \frac{\lambda^2}{n} + \lambda^2 = (1 + 1/n)\lambda^2$$

$$\Rightarrow E(\overline{Y}) + E(\overline{Y}^2) = \lambda + (1+1/n)\lambda^2 \Rightarrow 4E(\overline{Y}) + \frac{1}{1+1/n}E(\overline{Y}^2) = 4\lambda + \lambda^2$$

$$\Rightarrow E\left(\overline{Y} + \frac{n}{n+1}\overline{Y}^2\right) = 4\lambda + \lambda^2 = E(C)$$

$$\Rightarrow \hat{\theta} = \overline{Y} + \frac{n}{n+1}\overline{Y}^2 \quad \text{is an unbiased estimator of } E(C)$$

8.13 p. 395

We have seen that if Y has a binomial distribution with parameters n and p, then Y/n is an unbiased estimator of p. To estimate the variance of Y, we generally use $n(Y/n)(1 - Y/n)$.

a. Show that the suggested estimator is a biased estimator of $V(Y)$.

$$\text{Var}(Y) = np(1-p)$$

$$E(n(Y/n)(1-Y/n)) = E(Y - Y^2/n) = E(Y) - \frac{1}{n}E(Y^2) = np - \frac{1}{n}(\text{Var}(Y) + E(Y)^2) = np - \frac{1}{n}(np(1-p) + n^2p^2) = np - p(1-p) - np^2 = np(1-p) - p(1-p) = (n-1)p(1-p) \neq \text{Var}(Y)$$

$$\Rightarrow n(Y/n)(1-Y/n) \quad \text{biased estimator of Var}(Y) \quad \square$$

b. Modify $n(Y/n)(1 - Y/n)$ slightly to form an unbiased estimator of $V(Y)$.

$$E(n(Y/n)(1-Y/n)) = (n-1)p(1-p) \Rightarrow \frac{n}{n-1}E(n(Y/n)(1-Y/n)) = np(1-p) = \text{Var}(Y) \Rightarrow E(\frac{n}{n-1}n(Y/n)(1-Y/n)) = \text{Var}(Y)$$

$$\Rightarrow \frac{n^2}{n-1}(Y/n)(1-Y/n) \quad \text{unbiased estimator of Var}(Y)$$

8.14 p. 395

Let $Y_1, Y_2, ..., Y_n$ denote a random sample of size n from a population whose density is given by

$$f(y) = \begin{cases} \alpha y^{\alpha-1}/\theta^\alpha, & 0 \le y \le \theta, \\ 0, & \text{elsewhere.} \end{cases}$$

where $\alpha > 0$ is a known, fixed value, but θ is unknown. (This is the power family distribution

introduced in Exercise 6.17.) Consider the estimator $\widehat{\theta} = \max(Y_1, Y_2, ..., Y_n)$.

a. Show that $\widehat{\theta}$ is a biased estimator for θ.

Want

$$E(\hat{\theta}) \neq \theta$$

Let $Y_{(n)}$ denote $\widehat{\theta} = \max(Y_1, Y_2, ..., Y_n)$.

$$F_{Y_{(n)}}(y) = P(Y_{(n)} \leq y) = P(Y_i \leq y, i = 1, ..., n)$$

Because $Y_1, Y_2, ..., Y_n$ are independently distributed,

$$F_{Y_{(n)}}(y) = P(Y_i \leq y, i = 1, ..., n) = \prod_{i=1}^{n} P(Y_i \leq y) = \prod_{i=1}^{n} F(y) = F(y)^n$$

Let $g_{(n)}(y)$ denote the density function of $Y_{(n)}$.

$$g_{(n)}(y) = nF(y)^{n-1} f(y)$$

$$F(t) = \int_0^t f(y)\mathrm{d}y = \int_0^t \alpha y^{\alpha-1}/\theta^\alpha \mathrm{d}y = \frac{t^\alpha}{\theta^\alpha}$$

$$\Rightarrow F(y) = \frac{y^\alpha}{\theta^\alpha}$$

$$\Rightarrow g_{(n)}(y) = n\left(\frac{y^\alpha}{\theta^\alpha}\right)^{n-1} \alpha y^{\alpha-1}/\theta^\alpha = n\alpha y^{n\alpha-1}/\theta^{n\alpha}$$

$$\Rightarrow Y_{(n)} \sim \text{Power family with parameters } n\alpha \text{ and } \theta$$

$$E(Y_{(n)}) = \int_0^\theta y g_{(n)}(y)\mathrm{d}y = \int_0^\theta y n\alpha y^{n\alpha-1}/\theta^{n\alpha}\mathrm{d}y = \frac{n\alpha}{\theta^{n\alpha}}\int_0^\theta y^{n\alpha}\mathrm{d}y = \frac{n\alpha}{n\alpha+1}\theta$$

Therefore,

$$E(\hat{\theta}) = \frac{n\alpha}{n\alpha+1}\theta \neq \theta \quad \square$$

b. Find a multiple of $\widehat{\theta}$ that is an unbiased estimator of θ.

$$\frac{n\alpha+1}{n\alpha}\hat{\theta} \text{ unbiased estimator of } \theta$$

c. Derive MSE($\widehat{\theta}$).

$$\mathrm{MSE}(\hat{\theta}) = E((\hat{\theta}-\theta)^2) = E(\hat{\theta}^2) - 2\theta E(\hat{\theta}) + \theta^2 = E(Y_{(n)}^2) - \frac{2n\alpha}{n\alpha+1}\theta^2 + \theta^2$$

$$E(Y_{(n)}^2) = \int_0^\theta y^2 g_{(n)}(y)\mathrm{d}y = \frac{n\alpha}{\theta^{n\alpha}}\int_0^\theta y^{n\alpha+1}\mathrm{d}y = \frac{n\alpha}{n\alpha+2}\theta^2$$

$$\Rightarrow \text{MSE}(\hat{\theta}) = \frac{n\alpha}{n\alpha+2}\theta^2 - \frac{2n\alpha}{n\alpha+1}\theta^2 + \theta^2 = \frac{2}{(n\alpha+1)(n\alpha+2)}\theta^2$$

8.24 p. 403

Results of a public opinion poll reported on the Internet indicated that 69% of respondents rated the cost of gasoline as a crisis or major problem. The article states that 1001 adults, age 18 years or older, were interviewed and that the results have a sampling error of 3%. How was the 3% calculated, and how should it be interpreted? Can we conclude that a majority of the individuals in the 18+ age group felt that cost of gasoline was a crisis or major problem?

The 2-standard-error is

$$2\sqrt{\frac{(0.69)(0.31)}{1001}} \approx 0.029 = 0.03 = 3\%,$$

meaning we can be confident to 95% that 66% to 72% of the 18+ age group in the population felt gasoline cost was a major problem.

Therefore, because 66%-72% is still a majority, we can conclude a majority of that age group felt gasoline cost was a major problem.

8.32 p.405

An auditor randomly samples 20 accounts receivable from among the 500 such accounts of a client's firm. The auditor lists the amount of each account and checks to see if the underlying documents comply with stated procedures. The data are recorded in the accompanying table (amounts are in dollars, Y = yes, and N = no).

Account	Amount	Compliance	Account	Amount	Compliance
1	278	Y	11	188	N
2	192	Y	12	212	N
3	310	Y	13	92	Y
4	94	N	14	56	Y
5	86	Y	15	142	Y
6	335	Y	16	37	Y
7	310	N	17	186	N
8	290	Y	18	221	Y
9	221	Y	19	219	N
10	168	Y	20	305	Y

Estimate the total accounts receivable for the 500 accounts of the firm and place a bound on the error of estimation. Do you think that the average account receivable for the firm exceeds $250? Why?

Let X_i denote an account receivable out of the 20.

The sample mean is

$$\overline{X} = \frac{1}{20}\sum X_i = 197.1,$$

with standard deviation

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum(X - \overline{X})^2}{19}} \approx 90.86.$$

The standard error is therefore

$$\epsilon = \frac{\sigma}{\sqrt{20}} = \frac{90.86}{\sqrt{20}}.$$

An estimate of the total accounts receivable is

$$500\overline{X} = 98550,$$

with error

$$500\epsilon = (500)\frac{90.86}{\sqrt{20}} \approx 10158.46,$$

and therefore error bound

$$2(10158.46) = 20316.92.$$

The error bound of the sample mean is

$$2\epsilon = (2)\frac{90.86}{\sqrt{20}} \approx 40.63.$$

$$\overline{X} + \ 40.63 = 197.1 + 40.63 = 237.73 < 250$$

Therefore, the average account receivable for the firm likely doesn't exceed $250.

8.39 p. 409

Suppose that the random variable Y has a gamma distribution with parameters $\alpha = 2$ and an unknown β. In Exercise 6.46, you used the method of moment-generating functions to prove a general result implying that $2Y/\beta$ has a χ^2 distribution with 4 degrees of freedom (df). Using $2Y/\beta$ as a pivotal quantity, derive a 90% confidence interval for β.

Let $U = 2Y/\beta$

$$U \sim \chi^2(4)$$

Want a, b such that

$$P(a \leq U \leq b) = 0.90,$$

in other words

$$P(U \leq a) = \int_0^a f_U(u)du = 0.05,$$
$$P(U \geq b) = \int_b^\infty f_U(u)du = 0.05.$$

From Table 6 of the textbook, we find

$$a = 0.710721,$$
$$b = 9.48773.$$

Therefore,

$$P(0.710721 \leq U \leq 9.48773) = P(0.710721 \leq 2Y/\beta \leq 9.48773) = 0.90$$

$$\Rightarrow P\left(\frac{2Y}{0.710721} \le \beta \le \frac{2Y}{9.48773}\right) = 0.90$$

Therefore,

$$\left(\frac{2Y}{0.710721}, \frac{2Y}{9.48773}\right)$$

is a 90% confidence interval for β.

8.44 p. 410

Let *Y* have density function

$$f_Y(y) = \begin{cases} \frac{2(\theta-y)}{\theta^2}, & 0 \le y \le \theta, \\ 0, & \text{elsewhere.} \end{cases}$$

a. Show that *Y* has distribution function

$$F_Y(y) = \begin{cases} 0, & y \le 0, \\ \frac{2y}{\theta} - \frac{y^2}{\theta^2}, & 0 < y < \theta, \\ 1, & y \ge \theta. \end{cases}$$

$$F_Y(y) = \begin{cases} 0, & y \le 0, \\ \int_0^y f_Y(t)\mathrm{d}t, & 0 < y < \theta, \\ 1, & y \ge \theta. \end{cases}$$

$$\int_0^y f_Y(t)\mathrm{d}t = \int_0^y \frac{2(\theta-t)}{\theta^2}\mathrm{d}t = \frac{2}{\theta^2}\left(\theta y - \frac{y^2}{\theta^2}\right) = \frac{2y}{\theta} - \frac{y^2}{\theta^2} \quad \square$$

b. Show that Y/θ is a pivotal quantity.

Let $U = Y/\theta$.

$$F_U(u) = P(U \le u) = P(Y \le \theta u) = F_Y(\theta u) = 2u - u^2, \quad 0 < u < 1$$

which is independent of θ.

Therefore $U = Y/\theta$ is a pivotal quantity. □

c. Use the pivotal quantity from part (b) to find a 90% lower confidence limit for θ.

$$F_U(a) = 0.05 \Rightarrow 2a - a^2 = 0.05 \Rightarrow a^2 - 2a + 0.05 = 0$$
$$\Rightarrow a \approx 0.0253206 \approx 0.03$$

Therefore, the 90% lower confidence limit for θ is

$$Y/0.03.$$

8.62 p. 418

The following statistics are the result of an experiment conducted by P. I. Ward to investigate a theory concerning the molting behavior of the male Gammarus pulex, a small crustacean. If a male needs to molt while paired with a female, he must release her, and so loses her. The theory is that the male G.

pulex is able to postpone molting, thereby reducing the possibility of losing his mate. Ward randomly assigned 100 pairs of males and females to two groups of 50 each. Pairs in the first group were maintained together (normal); those in the second group were separated (split). The length of time to molt was recorded for both males and females, and the means, standard deviations, and sample sizes are shown in the accompanying table. (The number of crustaceans in each of the four samples is less than 50 because some in each group did not survive until molting time.)

	Time to Molt (days)		
	Mean	s	n
Males			
Normal	24.8	7.1	34
Split	21.3	8.1	41
Females			
Normal	8.6	4.8	45
Split	11.6	5.6	48

a. Find a 99% confidence interval for the difference in mean molt time for "normal" males versus those "split" from their mates.

The difference in mean molt time is

$$\hat{\theta} = \mu_{normal} - \mu_{split} = 24.8 - 21.3 = 3.5$$

with standard error

$$\sigma_{\hat{\theta}} = \sqrt{\frac{s^2_{normal}}{n_{normal}} + \frac{s^2_{split}}{n_{split}}} = \sqrt{\frac{7.1^2}{34} + \frac{8.1^2}{41}}.$$

The critical value $z_{0.005}$ is 2.575829.
Therefore, the 99% confidence interval is

$$(3.5 - z_{0.005}\sigma_{\hat{\theta}}, 3.5 + z_{0.005}\sigma_{\hat{\theta}}) \approx (-1.02, 8.02)$$

b. Interpret the interval.

We are 99% confident that the mean molt time for "normal" males versus those "split" from their mates range from -1.02 to 8.02 days.

8.74 p. 424
Suppose that you want to estimate the mean pH of rainfalls in an area that suffers from heavy pollution due to the discharge of smoke from a power plant. Assume that σ is in the neighborhood of .5 pH and that you want your estimate to lie within .1 of μ with probability near .95. Approximately how many rainfalls must be included in your sample (one pH reading per rainfall)? Would it be valid to select all of your water specimens from a single rainfall? Explain.

Want

$$\frac{z_{0.025}\sigma}{\sqrt{n}} = 0.1$$

$$\Rightarrow \frac{(1.96)(0.5)}{\sqrt{n}} = 0.1 \Rightarrow n = \frac{(1.96)^2(0.25)}{0.01} \Rightarrow n = 96.04 \approx 97$$

Therefore, 97 rainfalls must be included in the sample.

Selecting all specimens from a single rainfall is invalid, as each pH reading would not be independent from one another.

8.85 p. 432

Two new drugs were given to patients with hypertension. The first drug lowered the blood pressure of 16 patients an average of 11 points, with a standard deviation of 6 points. The second drug lowered the blood pressure of 20 other patients an average of 12 points, with a standard deviation of 8 points. Determine a 95% confidence interval for the difference in the mean reductions in blood pressure, assuming that the measurements are normally distributed with equal variances.

$$\hat{\theta} = \mu_1 - \mu_2 = 11 - 12 = -1$$

The pooled sample estimator/variance is

$$S_p^2 = \frac{(n_1 - 1)\sigma_1^2 + (n_2 - 1)\sigma_2^2}{n_1 + n_2 - 2} = \frac{(15)(6)^2 + (19)(8)^2}{34} = \frac{878}{17}$$

and

$$t_{0.025} = 2.032$$

determined from the *t* distribution with $n_1 + n_2 - 2 = 34$ degrees of freedom.
(Source: https://www.medcalc.org/manual/t-distribution.php)

The 95% confidence interval is therefore

$$\hat{\theta} \pm t_{0.025} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = -1 \pm 2.032 \sqrt{\frac{878}{17}\left(\frac{1}{16} + \frac{1}{20}\right)} \approx (-5.898, 3.898)$$

8.101 p. 436

In laboratory work, it is desirable to run careful checks on the variability of readings produced on standard samples. In a study of the amount of calcium in drinking water undertaken as part of a water quality assessment, the same standard sample was run through the laboratory six times at random intervals. The six readings, in parts per million, were 9.54, 9.61, 9.32, 9.48, 9.70, and 9.26. Estimate the population variance σ^2 for readings on this standard, using a 90% confidence interval.

Assume the readings are normal.

$$\overline{Y} = 9.485$$

The sample variance is

$$\sigma_n^2 = \sum \frac{(Y_i - \overline{Y})^2}{5} = 0.02855$$

and

$$\chi_{0.05}^2 = 11.0705, \quad \chi_{0.95}^2 = 1.145476$$

determined from the χ^2 distribution with 5 degrees of freedom.

Therefore, the 90% confidence interval for σ^2 is

$$\left(\frac{(n-1)\sigma_n^2}{\chi^2_{0.05}}, \frac{(n-1)\sigma_n^2}{\chi^2_{0.95}}\right) = \left(\frac{(5)(0.02855)}{11.0705}, \frac{(5)(0.02855)}{1.145476}\right) \approx (0.013, 0.125)$$