Math 224 **Fall 2017**
Homework 6 **Drew Armstrong**

Problems from 9th edition of *Probability and Statistical Inference* by Hogg, Tanis and Zimmerman:

- Section 7.1, Exercises 2, 4, 7
- Section 7.3, Exercises 1, 6, 8(a,b)

Solutions to Book Problems.

7.1-2. A random sample of size 8 from $N(\mu, \sigma^2 = 72)$ yielded the sample mean $\overline{X} = 85$. Since this is an unrealistic textbook problem, the exact value of the population standard deviation is given to us:

$$\sigma = \sqrt{72} = 6\sqrt{2} \approx 8.485.$$

Thus for any probability value $0 < \alpha < 1$ we obtain an exact $(1-\alpha)100\%$ confidence interval for the population mean μ:

$$P\left(\overline{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha,$$

$$P\left(85 - z_{\alpha/2} \cdot \frac{\sqrt{72}}{\sqrt{8}} < \mu < 85 - z_{\alpha/2} \cdot \frac{\sqrt{72}}{\sqrt{8}}\right) = 1 - \alpha,$$

$$P\left(85 - z_{\alpha/2} \cdot 3 < \mu < 85 - z_{\alpha/2} \cdot 3\right) = 1 - \alpha,$$

Use this to find the following confidence intervals:

(a) $(1-\alpha)100\% = 99\%$. *Answer:*

$$P(85 - 2.575 \cdot 3 < \mu < 85 - 2.575 \cdot 3) = 99\%,$$
$$P(77.275 < \mu < 92.725) = 99\%.$$

(b) $(1-\alpha)100\% = 95\%$. *Answer:*

$$P(85 - 1.96 \cdot 3 < \mu < 85 - 1.96 \cdot 3) = 95\%,$$
$$P(79.12 < \mu < 90.88) = 95\%.$$

(c) $(1-\alpha)100\% = 90\%$. *Answer:*

$$P(85 - 1.645 \cdot 3 < \mu < 85 - 1.645 \cdot 3) = 90\%,$$
$$P(80.065 < \mu < 89.935) = 90\%.$$

(d) $(1-\alpha)100\% = 80\%$. *Answer:*

$$P(85 - 1.28 \cdot 3 < \mu < 85 - 1.28 \cdot 3) = 80\%,$$
$$P(81.16 < \mu < 88.84) = 80\%.$$

7.1-4. Let X be the weight in grams of a "52-gram" snack pack of candies. Assume that the distribution of X is $N(\mu, \sigma^2 = 4)$. A random sample of $n = 10$ observations of X yielded the following samples $X_1, \ldots, X_{10}$:

55.95	56.54	57.58	55.13	57.48
56.06	59.93	58.30	52.57	58.46

(a) Give a point estimate for μ. *Answer:* We will use the sample mean

$$\overline{X} = \frac{1}{10} \sum_i X_i = 56.8.$$

(b) Find the endpoints for a 95% confidence interval for μ. *Solution:* Since the standard deviation is **known**[1] ($\sigma = 2$) we have the exact confidence interval

$$\begin{aligned} P\left(\overline{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right) &= 95\%, \\ P\left(56.8 - 1.96 \cdot \frac{2}{\sqrt{10}} < \mu < 56.8 + 1.96 \cdot \frac{2}{\sqrt{10}}\right) &= 95\%, \\ P\left(56.8 - 1.24 < \mu < 56.8 + 1.24\right) &= 95\%, \\ P\left(55.56 < \mu < 58.04\right) &= 95\%. \end{aligned}$$

(c) On the basis of these very limited data, estimate $P(X < 52)$. *Solution:* Using our estimate for μ tells us that $X \approx N(56.8, 4)$, and hence

$$\frac{X - 56.8}{2} \approx N(0, 1).$$

Therefore we have

$$\begin{aligned} P(X < 52) &= P(X - 56.8 < -4.8) \\ &= P\left(\frac{X - 56.8}{2} < -2.4\right) \\ &\approx \Phi(-2.4) = 1 - \Phi(2.4) = 1 - 0.9918 = 0.82\%. \end{aligned}$$

Not very likely.

7.1-7. Thirteen tons of cheese,[2] including "22-pound" wheels (label weight), is stored in some old gypsum mines. A random sample of $n = 9$ of these wheels was weighed yielding the results $X_1, X_2, \ldots, X_9$ as shown in the following table. Assuming that the distribution of weights is $N(\mu, \sigma^2)$, use these data to find a 98% confidence interval for μ.

Solution: We use the following table to compute the sample mean $\overline{X} = \frac{1}{9}\sum_{i=1}^{9} X_i$ as well as the values $(X_i - \overline{X})^2$:

X_i	21.50	18.95	18.55	19.40	19.15	22.35	22.90	22.20	23.10
$\overline{X}$	20.9	20.9	20.9	20.9	20.9	20.9	20.9	20.9	20.9
$(X_i - \overline{X})^2$	0.36	3.8025	5.5225	2.25	3.0625	2.1025	4	1.69	4.84

[1] If the standard deviation were **unknown** then since the number of samples is small we would need to use a t distribution instead of a normal distribution.

[2] whatever

Then we compute the sample standard deviation

$$S = \sqrt{\frac{1}{9-1}\sum_{i=1}^{9}(X_i - \overline{X})^2} = 1.858.$$

Since the underlying distribution of the weights X_i is normal we know that the random variable $(\overline{X} - \mu)/(S/\sqrt{9})$ has a t distribution with $9 - 1 = 8$ degrees of freedom. To find a $(1-\alpha)100\% = 95\%$ confidence interval for μ we look in the table on page 496 to find the critical value

$$t_{\alpha/2}(8) = t_{0.025}(8) = 2.306.$$

Then we obtain the confidence interval

$$P\left(\overline{X} - t_{\alpha/2}(n-1)\cdot\frac{S}{\sqrt{n}} < \mu < \overline{X} + t_{\alpha/2}(n-1)\cdot\frac{S}{\sqrt{n}}\right) = 1 - \alpha,$$
$$P\left(20.9 - 2.306\cdot\frac{1.858}{3} < \mu < 20.9 + 2.306\cdot\frac{1.858}{3}\right) = 95\%,$$
$$P\left(20.9 - 1.428 < \mu < 20.9 + 1.428\right) = 95\%,$$
$$P\left(19.47 < \mu < 22.33\right) = 95\%.$$

This agrees with the answer in the back of the book.

7.3-1. Let p be the proportion of flawed toggle levers[3] that a certain machine shop manufactures. In order to estimate p a random sample of 642 levers was selected and it was found that 24 of them were flawed.

(a) Give a point estimate for p. *Solution:* We will use the sample mean

$$\hat{p} = \overline{X} = \frac{X}{n} = \frac{24}{642} = 3.74\%.$$

In parts (b), (c) and (d) we will use three different formulas to compute 95% intervals for p.

(b) Since $n = 642$ is relatively large we can use the simple formula

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

with $\alpha = 0.05$. By substituting $\hat{p} = 0.0374$, $n = 642$ and $z_{\alpha/2} = 1.96$ we obtain

$$0.0374 \pm 1.96\cdot\sqrt{\frac{(0.0374)(1-0.0374)}{642}} = \boxed{3.74\% \pm 1.47\%.}$$

(c) We get a more accurate answer by using the following formula from page 319:

$$\frac{\hat{p} + z_{\alpha/2}^2/(2n) \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n + z_{\alpha/2}^2/(4n^2)}}{1 + z_{\alpha/2}^2/n}$$

By substituting $\hat{p} = 0.0374$, $n = 642$ and $z_{\alpha/2} = 1.96$ we obtain

$$\frac{0.0374 + (1.96)^2/(2\cdot 642) \pm 1.96\cdot\sqrt{(0.0374)(1-0.0374)/642 + (1.96)^2/(4\cdot(642)^2)}}{1 + (1.96)^2/642}$$
$$= \boxed{4.01\% \pm 1.49\%.}$$

[3] whatever

(d) Since 3.74% is rather close to 0% we should also try the formula from page 321 which works when p is close to 0 or 1. For this we use the biased estimator

$$\tilde{p} = \frac{X+2}{n+4} = \frac{24+2}{642+4} = 4.02\%.$$

Then we will use the confidence interval $\tilde{p} \pm z_{\alpha/2}\sqrt{\tilde{p}(1-\tilde{p})/(n+4)}$. By substituting $\tilde{p} = 0.0402$, $n = 642$ and $z_{\alpha/2} = 1.96$ we obtain

$$0.0402 \pm 1.96 \cdot \sqrt{\frac{(0.0402)(1-0.0402)}{642+4}} = \boxed{4.02\% \pm 1.52\%.}$$

We observe that the result is closer to the more accurate formula in part (c), which confirms that the strange estimator $\tilde{p}$ is good for extreme values of p.

(e) Finally, since p is very small, we might be interested in a one-sided confidence interval for p. To compute a $(1-\alpha)100\%$ upper bound for p we can use any of the above three formulas to obtain

$$P(p < \text{old upper bound with } z_{\alpha/2} \text{ replaced by } z_\alpha) \approx 1-\alpha.$$

To compute a 95% upper bound for p we will substitute $z_{0.05} = 1.645$ in the place of $z_{0.025} = 1.96$. By doing this in all three formulas we obtain upper bounds

$$4.97\%, \quad 5.18\% \quad \text{and} \quad 5.29\%,$$

respectively. I see that the back of the textbook reports the value 4.97%, which means that they used the dumbest formula.

7.3-6. Let p equal the proportion of Americans who select jogging as one of their recreational activities. If 1497 out of a random sample of 5757 selected jogging, find an approximate 98% confidence interval for p.

Solution: Since the sample size $n = 5757$ is large we will use the most basic version of the confidence interval:

$$P\left(\hat{p} - z_{\alpha/2}\cdot\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{\alpha/2}\cdot\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) \approx 1-\alpha.$$

By substituting $\hat{p} = 1497/5757 = 0.26$, $n = 5757$ and $1-\alpha = 98\%$ we obtain

$$P\left(0.26 - 2.33\cdot\sqrt{\frac{(0.26)(0.74)}{5757}} < p < 0.26 + 2.33\cdot\sqrt{\frac{(0.26)(0.74)}{5757}}\right) \approx 95\%,$$

$$P\left(26\% - 1.35\% < p < 26\% + 1.35\%\right) \approx 95\%,$$

$$P\left(24.66\% < p < 27.35\%\right) \approx 95\%.$$

7.3-8(a,b). A proportion, p, that many opinion polls estimate is the number of Americans who sould say yes to the question, "If something were to happen to the president of the United States, do you think that the vice president would be qualified to take over as president?" In one such random sample of 1022 adults, 388 said yes.

(a) On the basis of the given data, find a point estimate of p. *Answer:*

$$\hat{p} = \frac{388}{1022} = 37.8\%.$$

(b) Find an approximate 90% confidence interval for p. *Answer:* Again, since $n = 1022$ is largre we use the most basic confidence interval. By substituting $\hat{p} = 0.3796$, $n = 1022$ and $1 - \alpha = 90\%$ we obtain

$$P\left(\hat{p} - z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) \approx 1 - \alpha$$

$$P\left(0.378 - 1.645 \cdot \sqrt{\frac{(0.378)(0.622))}{1022}} < p < 0.378 + 1.645 \cdot \sqrt{\frac{(0.378)(0.622)}{1022}}\right) \approx 90\%,$$

$$P\left(37.8\% - 3.54\% < p < 37.8\% + 3.54\%\right) \approx 90\%,$$

$$P\left(34.43\% < p < 41.5\%\right) \approx 90\%.$$

Additional Problems.

1. Sample Standard Deviation. Let $X_1, X_2, \ldots, X_n$ be independent samples from an underlying population with mean μ and variance σ^2. We have seen that the sample mean $\overline{X} = (X_1 + X_2 + \cdots + X_n)/n$ is an *unbiased estimator* for the population mean μ because

$$E[\overline{X}] = \mu.$$

The most obvious way to estimate the population variance σ^2 is to use the formula

$$V = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Unfortunately, you will show that this estimator is **biased**.

(a) Explain why $E[X_i^2] = \mu^2 + \sigma^2$ for each i.

(b) Use the linearity of expectation together with part (a) and the fact that $\sum X_i = n\overline{X}$ to show that

$$\begin{aligned} E[V] &= \frac{1}{n}\left(E[\textstyle\sum X_i^2] - 2E[\overline{X}\textstyle\sum X_i] + E[n\overline{X}^2]\right) \\ &= \frac{1}{n}\left(n(\mu^2 + \sigma^2) - nE[\overline{X}^2]\right) \\ &= \mu^2 + \sigma^2 - E[\overline{X}^2] \end{aligned}$$

(c) Use the formula $\text{Var}(\overline{X}) = E[\overline{X}^2] - E[\overline{X}]^2$ to show that

$$E[\overline{X}^2] = \mu^2 + \sigma^2/n.$$

(d) Put everything together to show that

$$E[V] = \frac{n-1}{n} \cdot \sigma^2 \neq \sigma^2,$$

hence V is a **biased** estimator for σ^2.

It follows that the weird formula

$$S^2 = \frac{n}{n-1} \cdot V = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

satisfies

$$E[S^2] = E\left[\frac{n}{n-1} \cdot V\right] = \frac{n}{n-1} \cdot E[V] = \frac{\cancel{n}}{\cancel{n-1}} \cdot \frac{\cancel{n-1}}{\cancel{n}} \cdot \sigma^2 = \sigma^2$$

and hence S^2 is an **unbiased** estimator for σ^2. We call it the *sample variance* and we call its square root S the *sample standard deviation*. It is a sad fact that S is a **biased** estimator for σ but you will have to take more statistics courses if you want to learn about that.

Proof: For part (a) note that $E[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$ are given to us. Then we obtain

$$\begin{aligned} \mathrm{Var}(X_i) &= E[X_i^2] - E[X_i]^2 \\ E[X_i^2] &= E[X_i]^2 + \mathrm{Var}(X_i) = \mu^2 + \sigma^2. \end{aligned}$$

For part (c) we first note that $E[\overline{X}] = \mu$ and $\mathrm{Var}(\overline{X}) = \sigma^2/n$. (You'd better remember why this is true for Exam3.) Then we obtain

$$\begin{aligned} \mathrm{Var}(\overline{X}) &= E[\overline{X}^2] - E[\overline{X}]^2 \\ E[\overline{X}^2] &= E[\overline{X}]^2 + \mathrm{Var}(\overline{X}) = \mu^2 + \sigma^2/n. \end{aligned}$$

Then for parts (b) and (d) we first note that $\sum_{i=1}^n X_i = n\overline{X}$. Then we have

$$\begin{aligned} E[V] &= E\left[\frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2\right] \\ &= \frac{1}{n}\sum_{i=1}^n E[X_i^2 - 2\overline{X}X_i + \overline{X}^2] \\ &= \frac{1}{n}\left(\sum_{i=1}^n E[X_i^2] - 2E[\overline{X}\sum_{i=1}^n X_i] + \sum_{i=1}^n \overline{X}^2\right) \\ &= \frac{1}{n}\left(nE[X_i^2] - 2E[\overline{X}n\overline{X}] + nE[\overline{X}^2]\right) \\ &= \frac{1}{n}\left(nE[X_i^2] - 2nE[\overline{X}^2] + nE[\overline{X}^2]\right) \\ &= \frac{1}{n}\left(nE[X_i^2] - nE[\overline{X}^2]\right) \\ &= E[X_i^2] - E[\overline{X}^2] \end{aligned}$$

and from parts (a) and (c) we obtain

$$E[V] = E[X_i^2] - E[\overline{X}^2] = (\mu^2 + \sigma^2) - (\mu^2 + \sigma^2/n) = \sigma^2 - \sigma^2/n = \frac{n-1}{n}\cdot\sigma^2.$$

[Remark: We have shown that the sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$$

is an unbiased estimator for the population variance, but we have said nothing more about its distribution. If you go further in statistics you will learn the following fact: If the underlying distribution is **normal** then the random variable

$$\frac{n-1}{\sigma^2}S^2 = \frac{n}{\sigma^2}V$$

has a "chi squared distribution with $n-1$ degrees of freedom," whatever that means.]