Project 5 Solutions

Statistics 401: Fall 2016

Due: Tuesday, October 25

1. (2 pts) Let X be the number of SPAM emails received per day per employee at a large software engineering company. Suppose that the distribution of X is:

X	0	1	2
$P(X = x)$	0.60	0.30	0.10

   By definition, the population mean is

$$\mu_X = \sum_x xP(x) = 0(.6) + 1(.3) + 2(.1) = .5 \text{ SPAM emails}$$

   and the population variance is

$$\sigma_X^2 = \sum_x (x-\mu)^2 P(x) = (0-.5)^2(.6) + (1-.5)^2(.3) + (2-.5)^2(.1) = .45$$

   so the population standard deviation is $\sigma_X = \sqrt{.45} = 0.6708$ SPAM emails.

2. (8 pts) Consider the population of four textbooks. The variable being measured is the number of pages in a book. The 4 books have 212, 350, 379 and 575 pages respectively.

   (a) The population mean is

$$\mu = \sum_x xP(x) = 212\left(\frac{1}{4}\right) + 379\left(\frac{1}{4}\right) + 350\left(\frac{1}{4}\right) + 575\left(\frac{1}{4}\right) = 379 \text{ pages.}$$

   The population variance is

$$\begin{aligned}\sigma^2 &= \sum_x (x-\mu)^2 P(x) \\ &= (212-379)^2\left(\frac{1}{4}\right) + (379-379)^2\left(\frac{1}{4}\right) + (350-379)^2\left(\frac{1}{4}\right) + (575-379)^2\left(\frac{1}{4}\right) \\ &= 16786.5,\end{aligned}$$

   so the population standard deviation is $\sigma = \sqrt{16786.5} \approx 129.56$ pages.

   (b) Table 1 shows the 6 possible samples that can be drawn from the population and the corresponding values of $\overline{X}$. Table 2 shows the sampling distribution for $\overline{X}$.

   (c) The center of the sampling distribution can be calculated directly from the definition,

$$\mu_{\overline{X}} = \sum_{\bar{x}} \bar{x}P(\bar{x}) = \frac{281 + 295.5 + 364.5 + 393.5 + 462.5 + 477}{6} = 379 \text{ pages.}$$

   Or you can use the theory that says that, for any n, $\mu_{\overline{X}} = \mu = 379$ pages that you calculated in part (a). Of course, you get the same answer either way.

(d) The spread of the sampling distribution can be calculated by first using the definition for the population variance

$$\begin{aligned}\sigma^2_{\bar{X}} &= \sum_{\bar{x}}(\bar{x}-\mu)^2 P(\bar{x}) \\ &= \frac{(281-379)^2+(295.5-379)^2+(364.5-379)^2+(393.5-379)^2+(462.5-379)^2+(477-379)^2}{6} \\ &= 5595.5.\end{aligned}$$

so the population standard deviation is $\sigma_{\bar{X}} = \sqrt{5595.5} \approx 74.8$ pages. Or you can use the theory that says that, for any sample of size n from a small population of size N,

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{1-\frac{n-1}{N-1}} = \frac{129.56}{\sqrt{2}}\sqrt{1-\frac{2-1}{4-1}} = 74.8 \text{ pages}$$

where you are using the result from part (a) that $\sigma = 129.56$ pages. The above formula for $\sigma_{\bar{X}}$ includes the finite correction factor (i.e., $\sigma_{\bar{X}} \neq \frac{\sigma}{\sqrt{n}}$) because (1) the size of each sample is $n = 2 > 0.05N = 0.05(4) = 1/5$ and because (2) the books were sampled without replacement.

Table 1: Samples from the population of 4 books

Sample	$\bar{x}$
1. 212, 379	295.5
2. 212, 350	281
3. 212, 575	393.5
4. 379, 350	364.5
5. 379, 575	477
6. 350, 575	462.5

Table 2: Sampling Distribution of $\overline{X}$

Value of $\bar{x}$	$P(\bar{x})$
281	$\frac{1}{6}$
295.5	$\frac{1}{6}$
364.5	$\frac{1}{6}$
393.5	$\frac{1}{6}$
462.5	$\frac{1}{6}$
477	$\frac{1}{6}$
$\mu_{\bar{X}}$:	379
$\sigma^2_{\bar{X}}$:	5595.5
$\sigma_{\bar{X}}$	74.803

3. (Exercise 4.6 on page 205, 7 pts) Elijah and Tyler, two high school juniors, conducted a survey on 15 students at their school, asking the students whether they would like the school to offer an after-school art program, counted the number of "yes" answers, and recorded the sample proportion. 14 out of the 15 students responded "yes". They repeated this 100 times and built a distribution of sample proportions.

(a) The distribution of sample proportions is called the **sampling distribution** for the sample proportion $\hat{p}$.

(b) The variable being measured from each individual is:

$$X = \begin{cases} 0 & \text{not in support of an after-school art program} \\ 1 & \text{in support of an after-school art program} \end{cases}.$$

In the first sample of $n = 15$ students taken, Elijah and Tyler found that $\hat{p} = 14/15$ students answered that they would like an after-school art program. This suggests that there is overwhelming support of about 93.3% of student in favor of the art program; that is, $P(X = 0) \approx 0.07$ and $P(X = 1) \approx 0.93$, which suggests a left-skew in the distribution for X. The sample size $n = 15$ is not considered large because $n(1 - \hat{p}) = 15(1 - 14/15) = 1 < 10$ so we cannot apply the Central Limit Theorem (CLT). If the distribution for X has left skew, then the sampling distribution for $\hat{p}$ for such a small sample size will also have left skew (and not be approximtely normally distributed).

(c) The variability of the sampling distribution is the **sampling variability**. For a sample proportion $\hat{p}$, the sampling variability is $\sigma_{\hat{p}} = \sqrt{p(1-p)/n}$. Because p is estimated by $\hat{p}$, then we estimate the sampling variability to be

$$\sqrt{\hat{p}(1-\hat{p})/n} = \sqrt{\frac{14}{15}\left(1 - \frac{14}{15}\right)/15} = 0.064.$$

(d) Suppose that the students were able to recruit a few more friends to help them with sampling, and are now able to collect data from random samples each of size $n = 25$ students. Once again, the number of "yes" answers are recorded, and the sample proportion $\hat{p}$ is calculated. They repeat this 100 times to build a new distribution of sample proportions. The variability of this new distribution ($\sqrt{p(1-p)/25}$) is expected to be less than the variability of the original distribution ($\sqrt{p(1-p)/15}$) because:

$$\sqrt{p(1-p)/25} < \sqrt{p(1-p)/15}.$$

4. (3 pts) A simple random sample of $n = 100$ Twitter users were asked if they get some news from Twitter. Assuming that the true proportion of Twitter users who get some news from Twitter is $p = 0.52$:

(a) To determine the approximate sampling distribution of the sample proportion $\hat{p}$ for $n = 100$ Twitter users, consider the assumptions of the CLT. First, $n = 100$ is much less than 5% of the population of Twitter users, so the population is large enough to use CLT. Second, $n = 100$ is a large sample because $np > 52$ and $n(1 - p) = 48 > 10$. Thus, by CLT,

$$\hat{p} \dot{\sim} N(p, \sqrt{p(1-p)/n}) = N(0.52, 0.05).$$

(b) To calculate $P(\hat{p} < 0.5)$, we can use the normal sampling distribution from part (a),

$$\begin{aligned} P(\hat{p} < 0.5) &= P\left(Z < \frac{0.5 - 0.52}{0.05}\right) \\ &= P(Z < -0.4) \\ &= 0.345. \end{aligned}$$

In other words, there is a 34.5% chance that, if the true proportion is $p = 0.52$, that the sample proportion in a random sample of size $n = 100$ will be less than 0.5.

5. (Exercise 4.38 on page 216, 3 pts) The distribution of the original variable is left skewed with a mean $\mu = 60$ and standard deviation $\sigma = 18$.

   (a) The distribution of a single random sample of $n = 500$ values from this population is expected to reflect the true distribution of the population, and so is expected to have left skew with a sample mean $\bar{x}$ close to $\mu = 60$ and a sample standard deviation s close to $\sigma = 18$. This is Plot B.

   (b) The distribution of 500 sample means $\bar{X}$ from random samples of size $n = 18$ will have $\mu_{\bar{X}} = \mu = 60$ and standard deviation $\sigma_{\bar{X}} = \sigma/\sqrt{n} = 18/\sqrt{18} = 4.24$. This is Plot C.

   (c) The distribution of 500 sample means $\bar{X}$ from random samples of each size $n = 81$ will be approximately normal by CLT (because $n = 81 > 30$) with $\mu_{\bar{X}} = \mu = 60$ and standard deviation $\sigma_{\bar{X}} = \sigma/\sqrt{n} = 18/\sqrt{81} = 2$. This is Plot A.

6. (Exercise 4.40 on page 216, 7 pts) A manufacturer of compact fluorescent light bulbs advertises that the distribution of the lifespans of these light bulbs, X, is nearly normal with a mean of $\mu = 9,000$ hours and a standard deviation of $\sigma = 1,000$ hours.

   (a) Because the data are normal (i.e., $X \sim N(9000, 1000)$), the probability that a randomly chosen light bulb will last more than 10,500 hours is calculated by the normal distribution:

$$\begin{aligned} P(X > 10,500) &= P\left(Z > \frac{10500 - 9000}{1000}\right) \\ &= P(Z > 1.5) \\ &= 0.067. \end{aligned}$$

   (b) Because the distribution of lifespans, X, are normal, then the sampling distribution of the mean lifespan of $n = 15$ lightbulbs is also normal,

$$\bar{X} \sim N(9000, 1000/\sqrt{15}).$$

   Because the data are normal, we do not need to use CLT, and so it is not necessary to check the assumptions of the CLT.

   (c) Because the sampling distribution of $\bar{X}$ is normal (see part (b)), the probability that the mean lifespan of $n = 15$ randomly chosen light bulbs will be larger than than 10,500 hours is:

$$\begin{aligned} P(\bar{X} > 10,500) &= P\left(Z > \frac{10500 - 9000}{1000/\sqrt{15}}\right) \\ &= P(Z > 5.81) \\ &= 0.0000. \end{aligned}$$

   (d)

   (e) If the lifespans of light bulbs had a skewed distribution then we should not estimate the probabilities in parts (a) or (c)using the normal distribution. In part (a), we would need to use the skewed distribution to do the probability calculations. In part (c), because $n = 15 < 30$, we cannot use the CLT to assume that $\bar{X}$ is approximately normal.