

Project 5 Solutions

Statistics 401: Fall 2006

Due: Monday, October 16

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, $\mu_X = \sum_x xP(x) = 0(.6) + 1(.3) + 2(.1) = .5$ and $\sigma_X^2 = \sum_x (x - \mu)^2 P(x) = (0 - .5)^2(.6) + (1 - .5)^2(.3) + (2 - .5)^2(.1) = .45$ so $\sigma_X = \sqrt{.45} = 0.6708$.

2. (6 pts) Consider the population of four textbooks from problem 8.10(a) on page 340.

- (a) The population mean is $\mu = \frac{212+379+350+575}{4} = 379$ pages. The population variance is $\sigma^2 = \frac{(212-379)^2+(379-379)^2+(350-379)^2+(575-379)^2}{4} = 16786.5$, so $\sigma = \sqrt{16786.5} \approx 129.56$.
- (b) and (c) Table 1 shows the 6 possible samples that can be drawn from the population and the corresponding values of \bar{X} . Table 2 shows the sampling distribution for \bar{X} , where $\mu_{\bar{X}} = \frac{281+295.5+364.5+393.5+462.5+477}{6}$ and $\sigma_{\bar{X}}^2 = \frac{(281-379)^2+(295.5-379)^2+(364.5-379)^2+(393.5-379)^2+(462.5-379)^2+(477-379)^2}{6}$.

For any type of sample (doesn't have to be random and the sample size doesn't have to be less than 5% of the size of the population!) that $\mu_{\bar{x}} = \mu_x$. However, since the size of each sample of $n = 2$ is 50% of the population of size $N = 4$, then $\sigma_{\bar{x}} \neq \frac{\sigma}{\sqrt{n}}$.

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 \bar{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
$\sigma_{\bar{X}}^2$:	5595.5
$\sigma_{\bar{X}}$	74.803

3. (Problem 8.16 on page 350, 1pt) The population mean μ is equal to the mean of the sampling distribution of \bar{X} , $\mu_{\bar{x}}$. On the other hand, the sampling variability of \bar{X} gets smaller as the sample size gets larger, $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.
4. (3 pts) Regarding wait times for an elevator in Problem 8.18 on page 350:
- The mean is $\mu_{\bar{x}} = \mu = .5$ minutes. The standard deviation is $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{.289}{\sqrt{16}} \approx .07225$ minutes.
 - The mean is $\mu_{\bar{x}} = \mu = .5$ minutes. The standard deviation is $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{.289}{\sqrt{50}} \approx .0409$ minutes.
 - The approximate distribution of \bar{X} when the sample size is $n = 50$ is $N(.5, .0409)$ because the Central Limit Theorem applies when $n = 50 > 30$.
 - The probability is $P(\bar{X} > \frac{25}{60}) = P(z > \frac{\frac{25}{60} - .5}{.0409} \approx -2.04) \approx 0.9792$. Thus, for 50 individuals, the average weight time is more than 25 seconds 98% of the time.
5. (4 pts) Regarding the “Should Women Move” in Problem 8.32 on page 357:
- The Central Limit Theorem applies when $n\pi \geq 10$ and $n(1 - \pi) \geq 10$. In this case, $n\pi = 10(.3) = 3$ (and $n(1 - \pi) = 7$), so CLT does not apply. Thus, we can not be certain that p has an approximate normal distribution.
 - The mean is $\mu_p = n\pi = .3$ and the standard deviation is $\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{.3(.7)}{400}} \approx .0229$
 - For a sample size of $n = 400$, $n\pi = 400(.3) = 120 > 10$ and $n(1 - \pi) = 400(.7) = 280 > 10$, so the Central Limit Theorem applies and shows that $p \sim N(.3, .0229)$. Thus, $P(.25 \leq p \leq .35) = P(\frac{.25 - .3}{.0229} \leq z \leq \frac{.35 - .3}{.0229}) \approx P(-2.18 \leq z \leq 2.18) \approx .9707$. Thus, when the sample size is 400, the sample proportion will be between .25 and .35 97% of the time.
 - When the sample size increases, the sampling variability of p decreases, and so the probability that p is between .25 and .35 will decrease.

6. (Problem 8.34 on page 3.57, 1pt) For a sample size of $n = 500$, $n\pi = 500(.48) = 240 > 10$ and $n(1 - \pi) = 400(.52) = 260 > 10$, so the Central Limit Theorem applies and shows that $p \sim N(.48, \sqrt{\frac{.48(.52)}{500}} \approx .0223)$. Thus, $P(p > .5) = P(z > \frac{.5-.48}{.0223} \approx .8969) \approx .1849$. Thus, the polling organization will incorrectly predict the election 18.5% of the time.

7. The simulations from the website

<http://www.maths.soton.ac.uk/teaching/units/ma1c6/links/samplingapplet/samplingapplet.html>

produced the following results:

- (a) (2 pts) Repeatedly taking samples of size $n = 1$ gives a histogram which approximates the density curve for each of the normal, exponential, and uniform population distributions. Table 3 gives statistics which describe the center and spread of each of these histograms, as well as a description of the shape of the histogram for each.

Table 3: Statistics from three different distributions

Distribution	μ	σ^2	σ	shape
Normal	0	1	1	symmetric and bell-shaped
Exponential	1	1	1	severely right skewed
Uniform	.5	.08 $\bar{3}$	$\sqrt{.08\bar{3}}$	symmetric and rectangular

- (b) After repeatedly taking samples of size $n = 5$ from each of the normal, exponential and uniform distributions, the applet displayed histograms of the sampling distribution of \bar{X} when $n = 5$ for each of these three cases.
- i. (2 pts) Table 4 shows statistics which describe the center and spread of the sampling distributions, as well as a description of the shape of the histogram for each of the normal, exponential and uniform distributions. The results agree with the theoretical fact that $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.

Table 4: Statistics from three different sampling distributions when $n = 5$

Distribution	μ	σ^2	σ	shape
Normal	0	$\frac{1}{5} = .2$	$\frac{1}{\sqrt{5}} \approx .4472$	symmetric and bell-shaped
Exponential	1	$\frac{1}{5} = .2$	$\frac{1}{\sqrt{5}} \approx .4472$	slightly right skewed
Uniform	.5	$\frac{.08\bar{2}}{5} = .01\bar{6}$	$\frac{\sqrt{.08\bar{3}}}{\sqrt{5}} \approx .1291$	symmetric and triangular

- ii. (1 pt) When the population distribution is normal, then the sampling distribution of \bar{X} is exactly normal and therefore the histogram for the normal data “looks the most normal.”
- iii. (1 pt) The sampling distribution of \bar{X} from the exponential distribution is the “least normal” since the distribution of the data is exponential and has a very strong skew.
- (c) (1 pt) Since we have a large sample size, the Central Limit Theorem assures that the sampling distribution of \bar{X} is approximately normal for even severely skewed data.
8. When rolling a six-sided die, let X be the side which faces up.
- (a) (1 pt) The distribution of the variable X is given in Table 5.

Table 5: The distribution of the up-side of a tossed six-sided die

X	1	2	3	4	5	6
$P(X)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

(b) (2 pts) The population mean is $\mu_x = \frac{1+2+3+4+5+6}{6} = 3.5$ and $\sigma_x^2 = \frac{(1-3.5)^2+(2-3.5)^2+(3-3.5)^2+(4-3.5)^2+(5-3.5)^2+(6-3.5)^2}{6} \approx 2.9166$ so $\sigma_x \approx 1.7078$.

9. Simulations from

<http://www.stat.sc.edu/~west/javahtml/CLT.html>

regarding dice rolls helped to answer the following:

- (a) (1 pt) The histogram is centered at $\mu = 3.5$ and the height of each bar is about $\frac{1}{6}$. This agrees with the distribution of X given in 8(a).
- (b) (3 pts)
- For $n = 2$ rolls, the histogram is centered at $n\bar{\mu} = 2(3.5) = 7$ agreeing with the given fact that $\mu_{\sum x_i} = n\mu_x$. The spread in the histogram increases as predicted by the fact that $\sigma_{\sum x_i} = \sqrt{n}\sigma$.
 - For $n = 5$ rolls, the histogram is now centered at $n\mu_x = 5(3.5) = 17.5$ Also, the standard deviation is again increasing by $\sqrt{n}\sigma_x$.
 - The histogram from $n = 5$ rolls is “more normal” as suggested by the Central Limit Theorem due to the larger sample size.