Chapter I

Sampling Distribution

Solutions

Part I:

1.

1.
The population consists of all students in the class.

2.
The sample includes the 10 students sitting in the front row.

3.
The sample is made up of just the 10 students sitting in the front row. The sample is not likely to be representative. Those who sit in the front row tend to be more interested in the class and to perform higher on the test. Hence, the sample may perform at a higher level than the population.

2.

1.

$$E(\bar{x}) = \mu = 9000$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{500}{\sqrt{25}} = 100.$$

2.
These calculations indicate that in the long run the mean of a large group of sample means, each based on a sample size of 25, will be equal to 9000 hr. Further, the variability of these sample means with respect to the expected value of 9000 hr is expressed by a standard deviation of 100 hr.

3.

The numbers of samples of size 25 that could be obtained theoretically from a group pf 3000 students with and without replacement are 3000^{25} and C_{3000}^{25}, which are much larger than 80. Hence, we do not a true sampling distribution of means but only an *experimental* sampling distribution. Nevertheless, since the number of samples is large, there should be close agreement between the two sampling

distributions. Hence, the mean and standard deviation of the 80 sample means would be close to those of the theoretical distribution. Therefore, we have:

1.

$$\mu_{\bar{x}} = \mu = 68.0 \text{ inches} \qquad \text{and} \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{25}} = 0.6 \text{ inches.}$$

2.

$$\mu_{\bar{x}} = \mu = 68.0 \text{ inches and } \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} = \frac{3}{\sqrt{25}} \cdot \sqrt{\frac{3000-25}{3000-1}} = 0.597594377 \text{ inches}$$

which is only very slightly less than 0.6 inches and can for all practical purposes be considered the same as in sampling with replacement.
Thus we would expect the experimental sampling distribution of means to be approximately normally distributed with mean 68.0 inches and standard deviation 0.6 inches.

4.

Since the sample size is greater than 30, the central limit theorem can be used. Therefore, the distribution of sample mean is approximately normal with

$$\mu_{\bar{x}} = 5, \qquad \sigma_{\bar{x}} = \frac{2.34}{\sqrt{36}} = 0.39 .$$

$$P(\bar{x} < 6) = F(6) = \Phi\left(\frac{6-5}{0.39}\right) = \Phi(2.56) = 0.994766.$$

5.

1

$$P(\bar{x} > 160.00) = 1 - P(\bar{x} \le 160) = 1 - F(160)$$

$$= 1 - \Phi\left(\frac{160.00 - 150.00}{35.00}\right)$$

$$= 1 - \Phi(0.29) = 1 - 0.6141 = 0.3859 .$$

2.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{35.00}{\sqrt{40}} \approx 5.53$$

$$P(\bar{x} > 160.00) = 1 - P(\bar{x} \le 160) = 1 - F(160)$$

$$= 1 - \Phi\left(\frac{160.00 - 150.00}{5.53}\right) = 1 - \Phi(1.81)$$

$$= 1 - 0.964852 = 0.035148\,.$$

6.

$\mu = 3.02, \quad \sigma = 0.39, \quad n = 20.$

$$\sigma_{\bar{x}} = \frac{0.39}{\sqrt{20}} = 0.087206651 \approx 0.0872.$$

1.

$$P(\bar{x} \ge 3.10) = 1 - P(\bar{x} < 3.10)$$

$$= 1 - \Phi\left(\frac{3.10 - 3.02}{0.0872}\right) = 1 - \Phi\left(0.92\right) = 1 - 0.8212 = 0.1788.$$

2.

$$P(\bar{x} \le 2.90) \approx P(\bar{x} < 2.90) = \Phi\left(\frac{2.90 - 3.02}{0.0872}\right) = \Phi\left(-1.38\right)$$

$$= 1 - \Phi\left(1.38\right) = 1 - 0.9162 = 0.0838.$$

3.

$$P(2.95 \le \bar{x} < 3.11) = F\left(3.11\right) - F\left(2.95\right)$$

$$= \Phi\left(\frac{3.11 - 3.02}{0.0872}\right) - \Phi\left(\frac{2.95 - 3.02}{0.0872}\right)$$

$$= \Phi\left(1.032\right) - \Phi\left(-0.803\right) = \Phi\left(1.032\right) - (1 - \Phi(0.803))$$

$$= 0.8485 - 1 + 0.7881 = 0.6366.$$

7.

1.

$\mu = 29, \quad \sigma = 9, \quad n = 40$

a)

$$\mu_{\bar{x}} = 29, \quad \sigma_{\bar{x}} = \frac{9}{\sqrt{40}} \cdot \sqrt{\frac{6000-40}{6000-1}} = 1.418391803 \approx 1.4184.$$

$$P\left(\left|\bar{x} - 29\right| < 2\right) = 2\Phi\left(\frac{2}{1.4184}\right) - 1 = 2\Phi\left(1.41\right) - 1$$

$$= 2 \cdot 0.920730 - 1 = 0.8415.$$

b)

$$\mu_{\bar{x}} = 29, \quad \sigma_{\bar{x}} = \frac{9}{\sqrt{40}} = 1.4230249 \approx 1.4230.$$

$$P\left(\left|\bar{x} - 29\right| < 2\right) = 2\Phi\left(\frac{2}{1.4230}\right) - 1 = 2\Phi\left(1.41\right) - 1$$

$$= 2 \cdot 0.920730 - 1 = 0.8415.$$

2.

$\mu = 29, \quad \sigma = 9, \quad n = 100$

a)

$$\mu_{\bar{x}} = 29, \quad \sigma_{\bar{x}} = \frac{9}{\sqrt{100}} \cdot \sqrt{\frac{6000-100}{6000-1}} = 0.892542868 \approx 0.8925.$$

$$P\left(\left|\bar{x} - 29\right| < 2\right) = 2\Phi\left(\frac{2}{0.8925}\right) - 1 = 2\Phi\left(2.24\right) - 1$$

$$= 2 \cdot 0.987126 - 1 = 0.9742.$$

b)

$$\mu_{\bar{x}} = 29, \quad \sigma_{\bar{x}} = \frac{9}{\sqrt{100}} = 0.9.$$

$$P\left(\left|\bar{x} - 29\right| < 2\right) = 2\Phi\left(\frac{2}{0.9}\right) - 1 = 2\Phi\left(2.22\right) - 1$$

$$= 2 \cdot 0.986791 - 1 = 0.9736.$$

8.

1.

Samples	Number of businesses in the samples	Sample mean
A, B, C, D	42, 39, 36, 33	37.50
A, B, C, E	42, 39, 36, 30	36.75
A, B, D, E	42, 39, 33, 30	36.00
A, C, D, E	42, 36, 33, 30	35.25
B, C, D, E	39, 36, 33, 30	34.50

2.

$\bar{p}$	34.50	35.25	36.00	36.75	37.50
$P(\bar{x})$	0.2	0.2	0.2	0.2	0.2

3.

Samples	Sample mean	Sampling error
A, B, C, D	37.50	1.50
A, B, C, E	36.75	0.75
A, B, D, E	36.00	0.00
A, C, D, E	35.25	0.75
B, C, D, E	34.50	1.50

4.

$\bar{p}$	34.50	35.25	36.00	36.75	37.50
$P(\bar{x})$	0.2	0.2	0.2	0.2	0.2

$$\mu = \frac{42+39+36+33+30}{5} = 36$$

$$\sigma^2 = \frac{(42-36)^2+(39-36)^2+(36-36)^2+(33-36)^2+(30-36)^2}{5} = \frac{90}{5} = 18$$

$$\mu_{\bar{p}} = 34.50\cdot 0.2 + 35.25\cdot 0.2 + 36.0\cdot 0.2 + 36.75\cdot 0.2 + 37.50\cdot 0.2 = 36$$

$\therefore\ \mu_{\bar{p}} = \mu.$

$$\sigma^2_{\bar{p}} = 34.50^2\cdot 0.20 + 35.25^2\cdot 0.20 + 36.00^2\cdot 0.2 + 36.75^2\cdot 0.2 + 37.50^2\cdot 0.20 - 36.00^2 = 1.125$$

$$\frac{\sigma^2}{n}\cdot\frac{N-n}{N-1} = \frac{18}{4}\cdot\frac{5-4}{5-1} = 1.125$$

$\therefore\quad \sigma^2_{\bar{x}} = \dfrac{\sigma^2}{n}\cdot\dfrac{N-n}{N-1}.$

9.

1.

$$\mu_{\bar{p}} = 0.40$$

2.

$$\sigma_{\bar{p}} = \sqrt{\frac{0.40\cdot 0.60}{100}} = 0.0490, \qquad (\because\ n = 100 > 30)$$

3.

$$P(\bar{p} > 0.5) = 1 - P(\bar{p} \le 0.5) \approx 1 - P(\bar{p} \le 0.5)$$

$$= 1 - F(0.50) = 1 - \Phi\left(\frac{0.50-0.40}{0.0490}\right)$$

$$= 1 - \Phi(2.04) = 1 - 0.978822 = 0.021178\,.$$

Part II: *SPSS*

1.

a)
age is interval level, so it satisfies the level of measurement requirement. Since the distribution of sampling means is based on the normal distribution, the correct probability for any specific mean value assumes that the variable follows *approximately* a normal distribution.
To verify that our data satisfies this assumption, we will also compute the skewness and kurtosis of the distribution of all cases in the data file choose

- ***Analyze -> Descriptive Statistics -> Descriptives …***
- Move the variable *age* to the ***Variable(s)*** list box.

  Click on the ***Options*** button.

  Mark the check boxes ***Mean, Kurtosis, and Skewness.***

  Continue

- ***OK***

Output and Interpretation:

Descriptive Statistics

	N	Mean	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
AGE OF RESPONDENT	2041	47,97	,292	,054	-,765	,108
Valid N (listwise)	2041					

Obtaining accurate probabilities for the sampling distribution of means assumes that the variable is normally distributed. The skewness of the distribution (0.292) is between -1.0 and +1.0 and the kurtosis of the distribution (-0.765) is also between -1.0 and +1.0.
(If we did not satisfy the requirements of a normal distribution, we might still be able to compute a correct probability, using the Central Limit Theorem).
Statement a) is, therefore, true.

b)
Since we are treating all of the cases in the data file as the population, the population mean is 47.97 years.
Statement b) is, therefore, true.

2.

- ***Transform -> Random Number Generators…***

- Select

  Set Starting Point and ***Fixed Value***
  Set seed to: 1234567
  OK.
  (When you click on *OK*, the seed will be set, but you will get no feedback from *SPSS* telling you that this has been done.)

- ***Data -> Select Cases…***

- Click on the option button

  Random Sample of Cases
  Click on the button
  Sample
- To specify the size of the sample, type 10 in the text box ***Approximately*** in front of the % symbol to specify the size of the sample.

  Continue
- ***OK.***
- Go to the ***Data View***.

  SPSS indicateswhich cases are excluded from the sample drawing a diagonal line through the case number.
  The cases which are included in the random sample have a line across the case number.
  Having drawn a 10% sample of the cases in the data set, we calculate the statistics for the sample,
- ***Analyze -> Descriptive Statistics -> Descriptives…***

- Move the variable *age* to the ***Variable(s)*** box.

  OK.

Output:

Descriptive Statistics

	N	Mean	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
AGE OF RESPONDENT	205	49,73	,261	,170	-,621	,338
Valid N (listwise)	205					

The sample mean is equal to 47.93.

2.

- ***Analyze -> Descriptive Statistics -> Descriptives…***

- Move the variable *pop* to the ***Variable(s)*** list box.

  Options

- (The check boxes for Mean and Std. Deviation are already marked by default.)

  Mark the ***Kurtosis*** and ***Skewness*** check boxes. This will provide the statistics for assessing normality.

  Continue.

- Mark the check box ***Save standardized values as variables***.

  OK.

Output and Answers:

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
population	218	6928	1273111290	23481986,58	92414464,160	11,711	,165	155,824	,328
Valid N (listwise)	218								

1.
The ***'Descriptive Statistics'*** table in the SPSS output shows the number of cases for the variable "*population*" [*pop*] to be 218.

(SPSS does not tell us the number of cases that are missing data in this table. To get the number missing, we would have to compare the number of cases in the data set to the N for population.
If we had more than one variable in the table, the Valid N ***(listwise)*** row would tell us the number of cases that are not missing data for any of the variables in the table.)

2.
The distribution of *Population* is "nearly normal" if the following conditions are fulfilled:

- Skewness between -1.0 and +1.0
- Kurtosis between -1.0 and +1.0
- No outliers with standard scores less than or equal to -3.0 or greater than or equal to +3.0

"*Population*" [*pop*] does not satisfy the criteria for a normal distribution. Both the skewness (11.711) and kurtosis (155.824) fall outside the range from -1.0 to +1.0.

Though we know that we do not satisfy the "nearly normal condition", we will still do the check for outliers:
Click the right mouse button on the column header for *Zpop*, and select *Sort Ascending* from the pop-up menu. This will show any negative outliers at the top of the column:
At the top of the column, we do not see any negative values less than or equal to -3.0.
Click the right mouse button again on the column header for *Zpop*, and select *Sort Descending* from the pop-up menu. This will show any positive outliers at the top of the column.
At the top of the column, we see one positive value (13.52) greater than or equal to +3.0.
If we scroll back to the left, we see that the outlier for population was China, with a population of 1,273,111,290.

Therefore, "*Population*" [*pop*] does not satisfy the criteria for a normal distribution. Both the skewness (11.71) and kurtosis (155.82) fall outside the range from -1.0 to +1.0.
There was one outlier that had a standard score less than or equal to -3.0 or greater than or equal to +3.0: China with a value of 1,273,111,290 ($Z = 13.52$)

3.
In this problem, the skewness was 11.711, so we use the logarithmic transformation.
To compute the transformed variable:

- Select

  Transform -> Compute Variable

- In the Compute Variable dialog box, type the name *LG_pop* for the new variable in the ***Target Variable*** text box.

  In the ***Numeric Expression*** text box, type the formula LG10(pop) to compute base10 logarithms of the values of *pop*.

  OK.

Scroll the data editor window to the right to see the transformed variable, *LG_pop*.

- To calculate the descriptive statistics so we can check the normality conditions for the transformed variable, click on the ***Dialog Recall*** tool button, and select ***Descriptives***.

- Since we want the same statistics computed for the variable *pop*, we only need to replace the variable *pop* with *LG_pop*.

  (Be sure the check box for saving standardized values remains checked so that ***Descriptives*** will compute standard scores for *LG_pop*.)

  OK.

Output and Answers:

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
LG_pop	218	3,84	9,10	6,4577	1,07727	-,495	,165	-,407	,328
Valid N (listwise)	218								

The log transformation of "population" [*LG_pop*] satisfies the criteria for a normal distribution. The skewness of the distribution (-0.495) is between -1.0 and +1.0 and the kurtosis of the distribution (-0.407) is between -1.0 and +1.0.

- Next, we will check for outliers that had a standard score less than or equal to -3.0 or greater than or equal to +3.0:
  Click the right mouse button on the column header for *ZLG_pop*, and select *Sort Ascending* from the pop-up menu. We see that there are no outliers with standard scores less than or equal to -3.0.
  Click the right mouse button again on the column header for *ZLG_pop*, and select *Sort Descending* from the pop-up menu. We see that there are no outliers with standard scores greater than or equal to +3.0.

  Therefore, the log transformation of "*population*" [*LG_pop*] satisfies the criteria for a normal distribution. The skewness of the distribution (-0.50) is between -1.0 and +1.0 and the kurtosis of the distribution (-0.41) is between -1.0 and +1.0. There are no outliers that have a standard score less than or equal to -3.0 or greater than or equal to +3.0.
  The log distribution does, therefore, satisfy the nearly normal condition.

4.

- We will create a new variable that will have a value of 1 if the standard score is within 1 standard deviation of the mean, and 0 if it has a value outside this range:

  Transform ->Compute Variable

  In the ***Compute Variable*** dialog box, type the name *within1sd* for the new variable in the ***Target Variable*** text box.

  In the *Numeric Expression* text box, type the formula

  ZLG_pop>=-1.0 and ZLG_pop<=+1.0

  (The formula will assign *within1sd* a value of 1 if the standard score the log transformation of population is greater than or equal to -1.0 and less than or equal to +1.0. If the value is not between -1.0 and +1.0, *within1sd* will be assigned a 0.)

  OK.

- Scroll down in data view to see values of 0 and 1 for *within1sd.*

  When the standard scores for *LG_pop* are larger than 1.0, *within1sd* is assigned the value of 0.
  When the standard scores for *LG_pop* are less than or equal to 1.0, *within1sd* is assigned the value of 1.

  Analyze ->Descriptive Statistics ->Frequencies…

- Move the variable *within1sd* to the ***Variable(s)*** list box.

  OK.

Output and Answers:

Statistics

withi1sd

N	Valid	218
	Missing	0

withi1sd

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	,00	73	33,5	33,5	33,5
	1,00	145	66,5	66,5	100,0
	Total	218	100,0	100,0	

66.5% of the values fall within one standard deviation of the mean.
If we use 2% as the margin of error, 66.5% is within 2% of the 68% prescribed by the empirical rule.

- We will create a second new variable that will have a value of 1 if the standard score is within 2 standard deviations of the mean and 0 if it has a value outside this range.

  To compute the new variable, select the ***Compute Variable*** command from the ***Recall Dialog*** pop-up menu:

- Replace the variable name "*within1sd*" with the name "*within2sd*".

  Replace the criteria of -1.0 with -2.0 and replace +1.0 with +2.0.

  OK.

  Scroll down in data view to see values of 0 and 1 for within2sd.
  When the standard scores for *LG_pop* are less than or equal to 2.0, *within1sd* is assigned the value of 1.
  When the standard scores for *LG_pop* are larger than 2.0, *within2sd* is assigned the value of 0.

- We will request a second frequency distribution to tally *within2sd*:

  Select the ***Frequencies*** command from the ***Recall Dialog*** pop-up menu.

- Remove the variable *within1sd* from the ***Variable(s)*** list box and move the variable *within2sd* into the list box.

  OK.

Statistics

withi2sd

N	Valid	218
	Missing	0

withi2sd

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	,00	9	4,1	4,1	4,1
	1,00	209	95,9	95,9	100,0
	Total	218	100,0	100,0	

95.9% of the values fall within two standard deviations of the mean.
If we use 2% as the margin of error, 95.9% is within 2% of the 95% prescribed by the empirical rule.

The actual percentage of the values of *ZLG_pop* between -1.0 and +1.0 was 66.5%, which is within 2% of 68%. The actual percentage of the values of *ZLG_pop* between -2.0 and +2.0 was 95.9%, which is within 2% of 95%.

Similarly, the procedure can be repeated for the case of 3 times standard deviation with the following output:

Statistics

withi3sd

N	Valid	218
	Missing	0

withi3sd

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1,00	218	100,0	100,0	100,0

(Last revised: 20.01.2020)