Formulas

Stephanie Ranft S2459825

September 30, 2019

Statistics 1A
PSBE1-08

1 Week One

1. Explain what is meant by the term population.
2. Explain what is meant by the term sample.
3. Explain how a sample differs from a population.
4. Explain what is meant by the term sample data.
5. Explain what a parameter is.
6. Explain what a statistic is.
7. Give an example of a population and two different characteristics that may be of interest.
8. Describe the difference between descriptive statistics and inferential statistics. Illustrate with an example.
9. Identify each of the following data sets as either a population or a sample:
   (a) The grade point averages (GPAs) of all students at a college.
   (b) The GPAs of a randomly selected group of students on a college campus.
   (c) The ages of the nine Supreme Court Justices of the United States on January 1,1842 .
   (d) The gender of every second customer who enters a movie theater.
   (e) The lengths of Atlantic croakers caught on a fishing trip to the beach.
10. Identify the following measures as either quantitative or qualitative:
   (a) The 30 high-temperature readings of the last 30 days.
   (b) The scores of 40 students on an English test.
   (c) The blood types of 120 teachers in a middle school.
   (d) The last four digits of social security numbers of all students in a class.
   (e) The numbers on the jerseys of 53 football players on a team.
11. Identify the following measures as either quantitative or qualitative:
   (a) The genders of the first 40 newborns in a hospital one year.
   (b) The natural hair color of 20 randomly selected fashion models.
   (c) The ages of 20 randomly selected fashion models.
   (d) The fuel economy in miles per gallon of 20 new cars purchased last month.
   (e) The political affiliation of 500 randomly selected voters.
12. A researcher wishes to estimate the average amount spent per person by visitors to a theme park. He takes a random sample of forty visitors and obtains an average of $28 per person.
   (a) What is the population of interest?
   (b) What is the parameter of interest?

(c) Based on this sample, do we know the average amount spent per person by visitors to the park? Explain fully.

13. A researcher wishes to estimate the average weight of newborns in South America in the last five years. He takes a random sample of 235 newborns and obtains an average of 3.27 kilograms.

    (a) What is the population of interest?

    (b) What is the parameter of interest?

    (c) Based on this sample, do we know the average weight of newborns in South America? Explain fully.

14. A researcher wishes to estimate the proportion of all adults who own a cell phone. He takes a random sample of 1,572 adults; 1,298 of them own a cell phone, hence $1298/1572 \approx 0.83$ or about 83% own a cell phone.

    (a) What is the population of interest?

    (b) What is the parameter of interest?

    (c) What is the statistic involved?

    (d) Based on this sample, do we know the proportion of all adults who own a cell phone? Explain fully.

15. A sociologist wishes to estimate the proportion of all adults in a certain region who have never married. In a random sample of 1,320 adults, 145 have never married, hence $145/1320 \approx 0.11$ or about 11% have never married.

    (a) What is the population of interest?

    (b) What is the parameter of interest?

    (c) What is the statistic involved?

    (d) Based on this sample, do we know the proportion of all adults who have never married? Explain fully.

16. What must be true of a sample if it is to give a reliable estimate of the value of a particular population parameter?

1.1 Solutions

1. A population is the total collection of objects that are of interest in a statistical study.

2. From a population of interest, you draw a finite number of data values which form a sample of the population. The method of drawing the sample will dictate how representative it is of the population.

3. A sample, being a subset, is typically smaller than the population. In a statistical study, all elements of a sample are available for observation, which is not typically the case for a population.

4. The required information which is drawn from the population forms the sample data.

5. A parameter is a value describing a characteristic of a population. In a statistical study the value of a parameter is typically unknown.

6. A statistic is a computed quantity whose numerical value allows for inferences to be made about a parameter.

7. All currently registered students at a particular college form a population. Two population characteristics of interest could be the average GPA and the proportion of students over 23 years.

8. Descriptive statistics is the process of using and analysing sample statistics, in order to summarise a sample.

   Statistical inference is the process of using data analysis to deduce properties of an underlying probability distribution, which infers properties of a population, for example by testing hypotheses and deriving estimates.

   For example, “the sample of 50 people had a mean of 5 and a standard deviation of 3” tells the reader that the information about the sample, and “after testing for significant difference, it was found that the population mean is 5” tells the reader information about the population based on a test conducted on the descriptive statistics of the sample.

9. (a) Population.
   (b) Sample.
   (c) Population.
   (d) Sample.
   (e) Sample.

10. (a) Qualitative.
    (b) Qualitative.
    (c) Quantitative.
    (d) Quantitative.
    (e) Qualitative.

11. (a) Qualitative.
    (b) Qualitative.
    (c) Quantitative.
    (d) Quantitative.
    (e) Qualitative.

12. (a) All people who have ever visited a theme park.
    (b) The population mean amount of money spent by any person who has ever visited a theme park.
    (c) No, the sample size is too small to be representative of the population.

13. (a) All newborn babies in South America in the last five years.
    (b) The average birth weight of all newborn babies in South America in the last five years.
    (c) No, not exactly, but we know the approximate value of the average.

14. (a) All adults.
    (b) The population proportion of all adults who own a cell phone.
    (c) The sample proportion, which is the number of adults in the sample who own a cell phone divided by the total number of adults in the sample.
    (d) No, but we can infer that the population proportion is approximately 0.83, given the large sample size.

15. (a) All adults in the region.
    (b) The proportion of the adults in the region who have never married.
    (c) The proportion computed from the sample, 0.1 .
    (d) No, not exactly, but we know the approximate value of the proportion.

16. The selection procedure must be random and the sampling method suitable, e.g. whether stratified or simple random sampling is best depends on the data and question to be answered. The sample size must be large enough so that the Central Limit Theorem allows for reliable inferential statistics.

2 Week 2

1. The twins Caroline and James have created a table of their school marks, which they got throughout the whole semester in certain subjects.

	Mathematics	Physics	Chemistry	Geography
Caroline	1,2,3,1,1,5,2	3,3,1,1,1,2	1,1,1,3,4,1	2,2,3,1,5,4
James	4,4,1,2,2	2,2,2,2	5,5,4,4,3,4,3,3	1,1,2,1,1,1,2,1

   Calculate the final school mark (as a percentage) of the twins in all subjects, if the range of the school marks is from 1 to 5.

2. The following table contains measured heights of 63 students with the corresponding frequencies:

   Determine the mean, median, mode, variance and a standard deviation of the student's height.

Height	Frequency	Height	Frequency	Height	Frequency	Height	Frequency
159	1	165	2	170	5	175	2
161	1	166	3	171	6	177	1
162	2	167	2	172	7	178	4
163	1	168	4	173	9	179	2
164	2	169	3	174	5	181	1

3. While weighing twenty one-kilogram sugar bags we noted the measured values in kg: 1.00, 1.01, 1.05, 0.99, 0.95, 1.00, 0.98, 0.99, 1.04, 1.06, 0.93, 1.00, 1.03, 0.97, 1.00, 0.99, 1.05, 1.01, 0.94, 1.00. Determine the mean and the variance of the measured weight. Draw a boxplot to represent this data and justify the removal of any outliers (you might need to do this more than once).

4. We measured the height x_i and the weight y_i of ten students; the values are shown in the table below. Find the means of the measured heights and of the measured weights, fill in the table and determine the correlation coefficient between the measured height and weight of the students. There are two methods for calculating the correlation coefficient; describe how you calculate this using both methods and the table below. Using the table, determine the standard deviations of both the heights and weights. What can you determine about the population based on the standard deviations and correlation coefficient?

i	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	z_{x_i}	z_{y_i}	$z_{x_i} \times z_{y_i}$
1	51	43								
2	54	48								
3	65	56								
4	46	42								
5	58	55								
6	72	54								
7	42	44								
8	69	49								
9	76	55								
10	38	38								
Σ										

5. We measured the flat area of 30 apartments and we have measured the following values in metres squared: 82.6, 57.3, 70.4, 65, 48.4, 103.8, 73.6, 43.5, 66.1, 93, 52.6, 70, 84.2, 55, 81.3, 61.5, 75.1, 34.8, 62.4, 116, 70.1, 63.6, 93, 59.2, 65.9, 77.2, 52.8, 68.7, 79.2, 87.4.

   (a) Create a table of grouped frequency distribution for the 9 classes.

   (b) Construct a histogram of relative frequencies of the flat areas.

   (c) From the specified values estimate the sample mean and the variance.

   (d) From the middle values of the intervals and from their frequencies estimate the sample mean and the variance.

6. The electrical wiring requires cables with a high strength. We examined values for two types of cables:

   1st type: 302, 310, 312, 310, 313, 318, 305, 309, 301, 309, 310, 307, 313, 229, 315, 312, 310, 308, 314, 333, 305, 310, 309, 314

   2nd type: 300, 310, 320, 309, 312, 311, 315, 317, 309, 313, 315, 314, 307, 322, 313, 313, 311, 316, 31, 314, 308, 319, 313, 312

   What can you tell me about the mean strengths of each type, given the information you obtain from the mean, median and mode. Hint: draw a graph which displays the location of each of the three statistics and draw inferences based on the shape.

7. Let X be normally distributed with mean 3 and variance 4. What is $\mathbb{P}(X \geq 6)$? What about $\mathbb{P}(X > 6)$; what can you say about $\mathbb{P}(X = 6)$?

8. Let X be normally distributed with mean 5 and $\mathbb{P}(X \leq 0) = 10\%$. What is the standard deviation of X?

9. Let X be normally distributed with $\mathbb{P}(X \leq 0) = 0.40$ and $\mathbb{P}(X \geq 10) = 0.10$. What is the mean and standard deviation deviation of X?

10. Consider two independent normally distributed random variables X and Y with means 1 and 2, and variances 3 and 4, respectively. What is $\mathbb{P}(X+Y \leq 5)$? Draw a sketch of each distribution, and the distribution of the combined variables $X+Y$. What if the variables were not independent?

2.1 Solutions

1. The final grades are as follows:

	Mathematics	Physics	Chemistry	Geography
Caroline	2.14	1.83	1.83	2.83
James	2.6	2	3.88	1.25

Caroline's and James' final grades are 43.15% and 48.65%, respectively.

2. As n is an odd number, the median lies at the 32nd position when the values are ordered, meaning $Q_2 = 171$. The mode is the value with the highest frequency, i.e. 173. The mean is given by

$$\bar{y} = (159 + 161 + 2*162 + 163 + 2*164 + 2*165 + 3*166 + 2*167 + 4*168 + 3*169 + 5*170 + 6*171 + 7*172 + 9*173 + 5*174 + 2*175 + 177 + 4*178 + 2*179 + 181)/63 = \frac{10761}{63} = 170.81.$$

The variance is given by

$$\begin{aligned} s_y^2 = \Big(&(159-170.81)^2 + (161-170.81)^2 + 2*(162-170.81)^2 + (163-170.81)^2 + 2*(164-170.81)^2 \\ &+2*(165-170.81)^2 + 3*(166-170.81)^2 + 2*(167-170.81)^2 + 4*(168-170.81)^2 \\ &+3*(169-170.81)^2 + 5*(170-170.81)^2 + 6*(171-170.81)^2 + 7*(172-170.81)^2 \\ &+9*(173-170.81)^2 + 5*(174-170.81)^2 + 2*(175-170.81)^2 + (177-170.81)^2 \\ &+4*(178-170.81)^2 + 2*(179-170.81)^2 + (181-170.81)^2 \Big)/(63-1) = 22.22. \end{aligned}$$

3. For the data set as is, the mean is 1.00 with standard deviation 0.04, mode is 1.00, and the 5-number summary is (0.93, 0.99, 1.00, 1.02, 1.06) with IQR=0.03, which results in the lower and upper bounds (0.95, 1.06).

We omit data values which sit lower than 0.95 and above 1.06 resulting in the following reduced data set: 1, 1.01, 1.05, 0.99, 0.95, 1, 0.98, 0.99, 1.04, 1, 1.03, 0.97, 1, 0.99, 1.05, 1.01, 1. The mean remains the same with new standard deviation 0.03, same mode, and the 5-number summary is (0.95, 0.99, 1.00, 1.01, 1.05) with IQR=0.02, which results in lower and upper bounds (0.96, 1.04).

Again, we omit data values lower than 0.96 and above 1.04 resulting in the following reduced data set: 1, 1.01, 0.99, 1, 0.98, 0.99, 1.04, 1, 1.03, 0.97, 1, 0.99, 1.01, 1. The mean remains the same with standard deviation 0.02, same mode, and the 5-number summary (0.97, 0.99, 1, 1.01, 1.04) with IQR=0.02, which results in lower and upper bounds (0.96, 1.03).

You must remove 9 data values from the set (in total - not all steps above present) and the final data set is: 0.98, 0.99, 0.99, 0.99, 1, 1, 1, 1, 1, 1.01, 1.01. This data set has no outliers, mean 1.00, standard deviation 0.01, mode 1.00, and 5-number summary (0.98, 0.99, 1.00, 1.00, 1.01).

How you calculate your Q_1 and Q_3 will greatly affect your reduction of outliers and therefore your representation of the boxplot. The above was calculated with EXCEL which uses Method 2. If you use Method 1, you would get totally different results. See this EXCEL sheet to illustrate my meaning - click between the tabs on the bottom to see that using Method 1 results in fewer reductions than using Method 2.

There is no assertion given by the agresti book as to which is best, only that p% must lie below the pth percentile. You can think about this like if you were to divide the sample size into 4 evenly sized portions, what would they look like? 20 divided by 4 is an even 5, so we know the first reduction to 17 is valid. 17 divided by 4 is 4.25, so Q_1 lies at the 5th ordered data point, but there is 0.75 of that point left over which is used for the next group of 4.25 i.e. 4.25 + 0.75 + 3.5 = 8.5 so Q_2 lies on the 9th ordered data point with half left over. Then 8.5 + 0.5 + 3.75 = 12.75, so Q_3 lies on the 13th ordered data point and there is 4.25 = 17 - 12.75 left for the upper 75%.

4. Results are here.

5.

6.

7. $X \sim \mathcal{N}(3, 4^2)$ in formal notation, and X is a continuous variable. This means that $\mathbb{P}(X > 6) = \mathbb{P}(X \geq 6)$, and the reason is similar to cutting a cake, i.e. the few crumbles left on the knife after cutting is $\mathbb{P}(X = 6) = 0$. We can use the z-transformation to find this probability:

$$\mathbb{P}\left(Z = \frac{X-\mu}{\sigma} > \frac{6-3}{4} = 0.75\right) \approx 0.226627.$$

We can approximate this by using the 68-95-99.7 rule: 68% lie between -1 and 1, so 34% lies between 0 and 1, and as 50% lie below zero we have that (34+50) = 84% lie below 1 or equivalently 16% lies above 1. We calculated $z = 0.75$ so we know that the probability we are looking for is between 16% and 50%.

8. Draw a sketch of this marking the points $X = 0$, $\mu = 5$, and shade the lower 10%. Calculate the z-score which corresponds to the shaded area: $\mathbb{P}(Z < -1.28) \approx 0.10$. Then,

$$z = \frac{x-\mu}{\sigma} = \frac{0-5}{\sigma} = -1.28 \implies \sigma = \frac{5}{1.28} \approx 3.902.$$

9. Draw a sketch of this! You have the following:

$$\left.\begin{array}{l}\mathbb{P}\left(Z = \frac{X-\mu}{\sigma} \leq \frac{0-\mu}{\sigma} = -0.25\right) \approx 0.40 \\ \mathbb{P}\left(Z = \frac{X-\mu}{\sigma} \geq \frac{10-\mu}{\sigma} = 1.28\right) \approx 0.10\end{array}\right\} \implies \begin{cases}\mu = 0.25\sigma \\ 10 - \mu = 1.28\sigma\end{cases} \implies 10 - [0.25\sigma] = 1.28\sigma$$

$$\implies \sigma = \frac{10}{1.28 + .25} \approx 6.536; \quad \mu = 0.25\sigma \approx 1.634.$$

10. Let $W = X + Y$ be the new variable after combining X and Y, then

$$\mu_W = \mu_{X+Y} = \mu_X + \mu_Y = 3$$
$$\sigma_W^2 = \sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2 + 2\,\mathrm{Cov}(X, Y) = 7.$$

Now we can calculate the z-score for $w = 5$:

$$z = \frac{5-3}{\sqrt{7}} \approx 0.75593 \implies \mathbb{P}(W = X + Y \leq 5) = \mathbb{P}(Z \leq 0.76) \approx 0.224855.$$

Compare to Q7, which had a similar z-value and note that we could have also guessed an approximate value using the 68-95-99.7 rule.

If the variables were not independent, then the covariance would not be zero and the variance of W would be greater.

Think about it in terms of this example: say you want to purchase 2 second-hand books and you know the average prices and margins, i.e. you know the mean price of the book online, as well as how much that price varies at auction (standard deviation). It's second-hand, so the purchase price depends on the buyer/seller involved, e.g. whether they can swing the purchase price in their favour. You want to know if you can get both books for less than a certain combined purchase price. If these books are on the same subject matter, it might be that some sellers have both books and are willing to sell at a lower price if you buy both together, i.e. negative covariance. If these books are by the same author, it could be that the price of these books vary in accordance with the author's popularity, i.e. positive covariance. If these books are not really related by any normal means, then the prices vary independently of each other, i.e. zero covariance. Further to this example, you know that you want two books but you have randomly selected a seller for each, so you haven't done your research and have just selected the book randomly. Normally online shopping sites have some type of ordering of the items for sale, so suppose you were able to print out all ads of a book, fold them and put them in a hat and select one at random; you do this for both books. You don't know the price that the seller is willing to part with the book, for either book, but you accept their purchase price regardless. So, it could be that the purchase prices of the books is far apart, so the variance would be great. Or, perhaps the purchase prices are very close together, then there is little variance. This is why we must add the variances and incorporate any covariance.