Data 8 Tutoring: Week 9 – SOLUTIONS

Normal Distributions, Sample Means, Experiment Design, Correlation

April 2, 2017

1 Normal Distributions and Sample Means

1. State the central limit theorem (CLT) and the 2 conditions needed to use it.

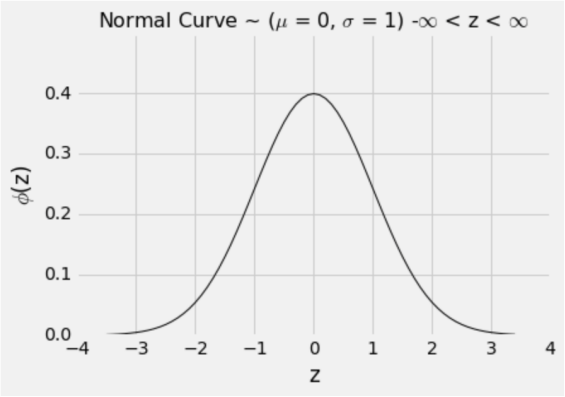
   CLT says that given the sample is (1) large and (2) drawn at random with replacement, then regardless of the shape of the population distribution, the the probability distribution of the sample average (or sample sum) is roughly bell-shaped.

2. For a normal distribution, what percentage of values are within 1 SD of the mean? 2 SDs? 3 SDs?

   68%, 95%, 99.73%

3. We have a (made-up) dataset containing the heights of all UC Berkeley students. The mean of the heights is 168 cm with a standard deviation of 3 cm. We now convert all the data into standard units for statistical purposes. What is the mean and standard deviation of the converted dataset?

   The mean is 0 and the SD is 1 by the standard normal curve shown:



4. Using the population from question 3, we take a sample of 900 students. What is the probability that the sample average is above 168.1 cm?
First we need to know the sample mean and sample SD:
Sample mean = population mean = 168cm
Sample SD = population SD / $\sqrt{samplesize}$
$= 3/\sqrt{900}$
$= 3/30$
$= 1/10$

Now we using the CLT, we can say the sample means are roughly normally distributed and calculate the proportion of values that should be 1 SD above the mean (168.1 is 1 SD above 168)
(100% - 68%) / 2 = 16%

5. We want to estimate the average weight of cats. We magically know that the standard deviation of cat weights is 2kg. How big a sample from the cat population do we need to take if we want a 99.73% confidence interval with a width that is less than 1kg?
For a 99.73% confidence interval, we want the sample mean ± 3 sample SDs. The width of the interval would then be 6 sample SDs. So we want

6 * sample SD < 1
6 * population SD $/\sqrt{samplesize} < 1$
$6 * 2/\sqrt{samplesize} < 1$
$12 < \sqrt{samplesize}$
$144 < samplesize$
$samplesize > 144$

2 Design of Experiments

1. For the following statements indicate if they are true for...
   - (A) any uniform random sample with replacement
   - (B) large uniform random samples with replacement only
   - (C) neither

   (a) For quantitative data, the sample SD is an unbiased estimate for the population SD.
   (C) We know the SD for a sample is scaled by 1 / $\sqrt{}$(sample size).

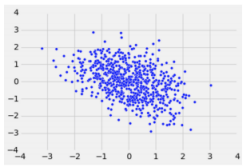
   (b) For quantitative data, the sample mean is an unbiased estimate for the population mean.
   (A) (which includes (B))

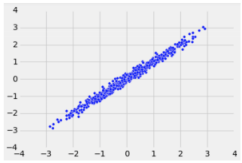
   (c) For categorical data with 5 categories, the total variation distance between the sample and the population will be small.
   (B) Law of averages; you can imagine if only a small sample of, let's say 10, were chosen, and by chance all were from the category with the largest proportion, then the TVD will be quite large.

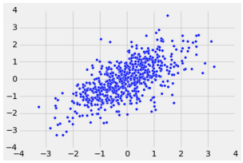
(d) For quantitative data, the histogram of the values in the sample will be bell-shaped.

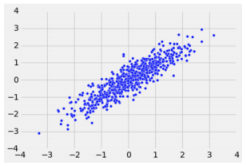
(C) The central limit theorem only applies to sums and means, not the original dataset, which can be any shape.

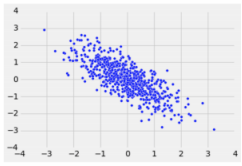
3 Correlation

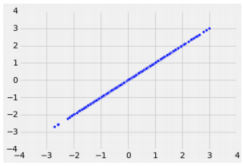
a) 

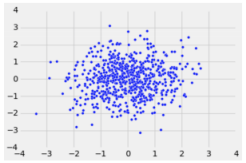
b) 

c) 

d) 

e) 

f) 

g) 

1. Given the scatter plots above, which of these would have the largest correlation coefficient (r)? The smallest r?

   Largest r: (f)
   Smallest r: (e)

2. Which of the above shows that the axes of the data are not as correlated? Does this necessarily correspond to the smallest correlation coefficient?

   (g) shows very little correlation between the axes. Though the correlation coefficient is around 0 and there is no association between the two axes, this is not necessarily the smallest correlation coefficient. Scatters like (e) have a stronger negative association with a smaller correlation coefficient.

3. Order the scatter plots above in terms of smallest to largest correlation coefficients.

   (e), (a), (g), (c), (d), (b), (f)

4. Identify whether each of the following statements is true or false. If false, explain why.

   - The correlation coefficient r is a number between 0 and 1.
     False. It is a number between -1 and 1.

- r measures the extent to which the scatter plot clusters around a straight line.

  True.

- r=1 if the scatter diagram is a perfect straight line sloping upwards, and r=-1 if the scatter diagram is a perfect straight line sloping downwards.

  True.

- Switching the axes of our scatter plot changes the value of r.

  False. Changing the axes will not impact r since r is measured as the mean of the product of x and y in standard units.

- r is unaffected by changing the units on either axis.

  True, this is why we convert to standard units.

- Correlation helps us establish conclusions about causality in our data.

  False. Correlation only tells us about association.

- r is the average of the products of the two variables, when both variables are measured in standard units.

  True.

- Outliers can have a significant impact on the correlation coefficient.

  True.

- Scatter plots that resemble parabolas (from quadratic functions) can either have an r of 1 or -1.

  False. Notice that correlation is only a measure of **linear** association. Because shapes like parabolas are not linear, it'll have a correlation coefficient of 0.

5. Fill in the blanks for the following function. You may assume there is a function called 'standard units' that has already been defined which converts any array of numbers to standard units. t represents the table of data where x and y are column names corresponding to the respective axes.

```
def correlation(t, x, y):
    x_in_su = standard_units(t.column(x))
    y_in_su = standard_units(t.column(y))
    return np.mean(x_in_su * y_in_su)
```