Data 8 Tutoring: Week 9

Normal Distributions, Sample Means, Experiment Design, Correlation

April 1, 2017

1 Normal Distributions and Sample Means

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

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

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?

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?

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?

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

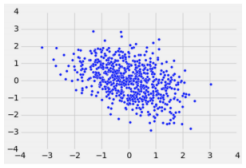
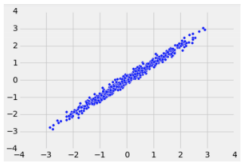
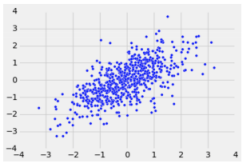
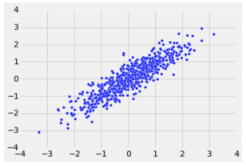
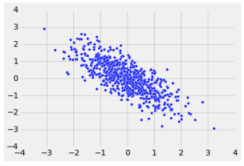
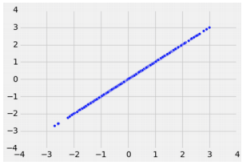
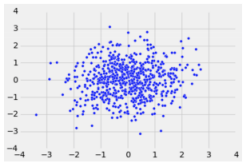
   (a) For quantitative data, the sample SD is an unbiased estimate for the population SD.

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

   (c) For categorical data with 5 categories, the total variation distance between the sample and the population will be small.

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

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?

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?

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

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

- Switching the axes of our scatter plot changes the value of r.
- r is unaffected by changing the units on either axis.
- Correlation helps us establish conclusions about causality in our data.
- r is the average of the products of the two variables, when both variables are measured in standard units.
- Outliers can have a significant impact on the correlation coefficient.
- Scatter plots that resemble parabolas (from quadratic functions) can either have an r of 1 or -1.

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 = ___________________
    y_in_su = ___________________
    return ___________(_____________ * _____________)
```