Group #5 @1pm
Mimi Morales, Yiyi Zhu, Vi Pham, Brian Fong May 14^{th}, 2019

Lecture Worksheet 16

Task 1

- How would you define a random variable by means of these data points?

  We can define the random variable to be

  $$X : S \longmapsto \mathbb{R}$$

  $$X(S) = \{50, 52, 53, 54, 55, 58, 59, 65\}$$

  where S = Source of randomness, we don't know what S is, but just that it exists.

- What probability would you assign to the values of this random variable?

  We can compute the probabilities by applying the principle of favorable to possible outcomes.

  $$P(X = 50) = \frac{5}{20}$$
  $$P(X = 52) = \frac{4}{20}$$
  $$P(X = 53) = \frac{1}{20}$$
  $$P(X = 54) = \frac{3}{20}$$
  $$P(X = 55) = \frac{2}{20}$$
  $$P(X = 58) = \frac{1}{20}$$
  $$P(X = 59) = \frac{1}{20}$$
  $$P(X = 65)\frac{3}{20}$$

- What is the average of the collected values? What is the expected value of the random variable?

  The average is obtained by:

$$\mu = \frac{1}{n} \cdot \sum_{j=1}^{n} x_j$$

$$= 54.75$$

  The expected value is obtained by:

$$E[X] = \sum_{x=j}^{n} x_j \; P(x_j)$$

$$= 54.75$$

  They are the same because the operation is identical, just rewritten differently.

Task 2

- In representation A, the dots indicate the collected values. Is this enough information to reconstruct the random variable corresponding to the data?

  No, we cannot reconstruct the random variable because we do not know the frequency of each data point.

- In representation B, each value is plotted above the number of the experiment which yielded it. Is this enough information to recover the corresponding random variable?

  Yes, because we know the frequency of each data point. Also, the order of the experiments is preserved in case the random experiment may be time-dependent, etc.

- Explain the meaning of the representation C. What do the axes represent? Is the random variable fully captured in this representation?

  C represents a bar graph that depicts the number of occurrences versus the values. The random variable may be fully captured because we know the frequency of each data point and somewhat about how the values are distributed.

- Discuss the pros and cons of representations B and C?

  Representation B preserves the order in case that matters to use. Also, it is easy to draw a line representing the mean through the data
  Whereas for Representation C, we can immediately see the number of frequencies for each value, however it is difficult to determine the mean value accurately.

Task 3

- State the definitions of expected value, of variance, and of median of a (discrete) random variable

  $$E(X) = \sum_{i=1}^{n} x_i \, P(x_i)$$

  $$Var(X) = E[X^2] - \mu^2$$

  $M(X)$ is any value s.t the number of values greater

  than M(X) is equal to the number of values less than M(X)

- Explain the meaning of expected value, variance, and median of a data set and of a random variable.

  The expected value turns out to be the weighted average of the random variable's values. It is essentially the mean and we may expect in general most of the random variable's values to lie somewhat near the expected value.

  The variance is mathematically the average of the squared distances from the mean for each of the random variable's values. This gives us an estimate of how far each value is from the mean and thus how distributed the data is.

  The median of a data set almost resembles the mid-point of the random variable's range, it divides the data into two equal parts.

- How would you use representations B and C to visualize and explain expected value, variance, and median?

  - *For representation B*
    - *I would draw a horizontal line across the graph indicating the mean. This graphical line is dependent on how far the data points are from this line, it must be such that the distances all sum up to zero, (those above the line are positive, and below are negative).*
    - *The variance is difficult to portray since there is no clear reference point for the mean.*
    - *The median would also be a horizontal line across the graph such that the number of points above equals the number of points below*
  - *For representation C*

* *It is difficult to determine how the mean should be portrayed for a histogram. One method could be viewing the area for each bar and drawing a vertical line such that the area to the left and to the right are equal.*
* *For the variance, it is easier to get an intuitive and visual idea of how much the variance ought to be based no how distributed the bars are on the graph. Still, it is difficult to find a precise value since we are working with a skewed data set that is somewhat widely distributed.*
* *For the median, we can draw a vertical line such that the occurrences to the left and to the right side of the line are the same.*