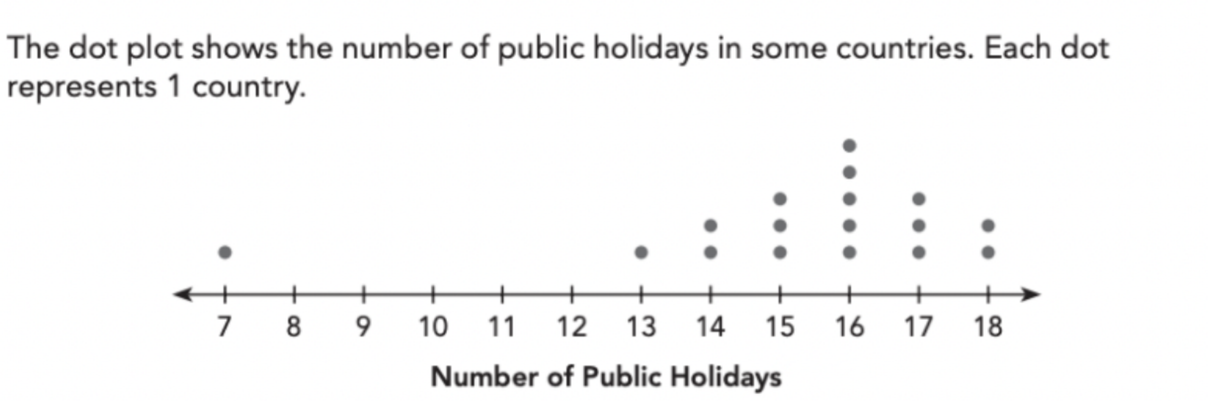
ICME Intro to Stats Summer Workshop

Section 1 and 2 Solutions

2023-07-24

Section 1

1.



The dot plot shows the number of public holidays in some countries. Each dot represents 1 country.

Figure 1: dot plot

Let's start by writing all of the data values in order from smallest to greatest:

$$7, 13, 14, 14, 15, 15, 15, 16, 16, 16, 16, 16, 17, 17, 17, 18, 18$$

a) Mean: We find the mean using this formula:

$$\bar{x} = \frac{1}{n} \sum_n x_n = \frac{1}{17} * 260 \approx 15.29$$

Median: we can find the median by finding the middle number. In this case, it's 16.

Mode: This is the data point that occurs most often. It's easy to see by looking at the dot plot that this number is 16 (with 5 occurrences).

b) The mean is less than the median because there is an outlier to the left of the mean at 7.

c) Median describes the dataset better than the mean because it is not affected by singular, extreme outliers.

d) The data is mostly symmetrical, with a clear outlier at 7. This causes it to be *left* skewed which occurs whenever the mean is < the median.

2.

a) Mean: Since mean is a linear function adding 5 to each data point should add 5 to the mean. The proof is as follows: Given a mean $\hat{x}$ we can calculate the new mean:

$$x^* = \frac{1}{n}\sum_n (x_n + 5) = \frac{1}{n}(\sum_n x_n + \sum_n 5) = \frac{1}{n}\sum_n x_n + \frac{1}{n}5n = \hat{x} + 5$$

b) Median: The median, once again, is a linear function. Adding 5 to each data point will also add 5 to the median.

c) Standard Deviation: Variance (and standard deviation) is not a linear function. The derivation of the effect of adding 5 to each data point of the standard deviation σ is as follows:

$$\sigma^2 = \frac{1}{n}\sum_n (X_n - \mu)^2$$

We already know that the mean μ will increase by 5 if we add 5 to each data point. Therefore, our formula can be rewritten as follows:

$$\sigma^2 = \frac{1}{n}\sum_n (X_n + 5 - (\mu + 5))^2 = \frac{1}{n}\sum_n (X_n - \mu)^2 = \sigma^2$$

As we can see, as long as we add the same amount to each data point, there is no effect on the variance (or standard deviation). Intuitively, this makes sense. If we shift a data sets mean but do not change its dispersion in any way, the measure of its dispersion, standard deviatoin, should also not change.

Section 2

1. Bayes Theorem

Let's define our events:

- A = Positive Test
- B = Infected Person

Let's apply Bayes Theorem:

$$P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{P(A|B)P(B)}{P(A|B)} = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|notB)P(notB)}$$

$$\frac{0.95 * 0.01}{0.95 * 0.01 + 0.02 * 0.99} = 32.4$$

$P(B = Infected\, Person) = 0.01,\ P(notB = Not\, Infected\, Person) = 0.99,\ P(A|B = Positive\, Test\, Given\, Infected\, Person$

2. Normal Distribution

For this question we should use a Z table.

a)

$$Z = \frac{x - \mu}{\sigma} = \frac{70 - 60}{4} = 2.5$$

$$P(Z > 2.5) = 1 - P(Z \leq 2.5) = 1 - 0.99379 = 0.0062$$

b)

$$Z = \frac{x - \mu}{\sigma} = \frac{55 - 60}{4} = -1.25$$

$$P(Z \leq -1.25) = 0.106$$

c)

$$P(X \leq 68) - P(X \leq 52) = P(Z \leq \frac{68 - 60}{4}) - P(Z \leq \frac{52 - 60}{4}$$

$$P(Z \leq 2) - P(Z \leq -2) = 0.97725 - 0.02275 = 0.9545$$

$$P(|Z| \leq 2) = 0.95$$