

We'll be considering the following two sets of data:

$$S_1 = \{1, 7, 8, 8, 12, 14, 20\} \quad \text{and} \quad S_2 = \{6, 7, 9, 9, 10, 12, 13, 14\}$$

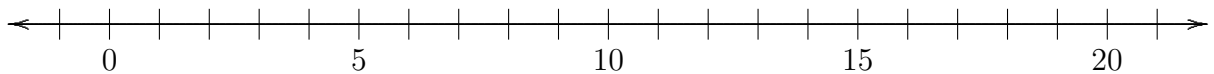
1 Mean and Median

- Compute the mean (the average, often called μ or mu) and the median of both sets. (The median is the middle number (if the set has an odd number of elements) or the average of the two middle numbers (if the set has an even number of elements).)
- Let N be the number of elements in the each set. (So for S_1 , $N = 7$.) Think of S as a random variable X , and assign to each outcome a probability $\frac{1}{N}$. (Since each outcome has the same probability, this is often called a *uniform probability distribution*.) For both S_1 and S_2 , compute $E(X)$ with this uniform probability.
- Explain why the answers you got in (a) and (b) agree.
- Calculate $E(X - \mu)$, $E((X - \mu)^2)$, and $E(X^2)$.

(Let me explain what is meant by these. If you have data $S = \{1, 2, 3\}$ thought of as a random variable X with the uniform probability distribution, then $\mu = E(X) = 2$. Just as $E(X)$ is the expected value of the data $\{x_1, x_2, x_3\} = \{1, 2, 3\}$, we get $E(X - \mu)$ is the expected value of the data $\{x_1 - \mu, x_2 - \mu, x_3 - \mu\} = \{-1, 0, 1\}$ and $E(X^2)$ is the expected value of the data $\{x_1^2, x_2^2, x_3^2\} = \{1^2, 2^2, 3^2\} = \{1, 4, 9\}$.)

2 The Variance

- Plot the data from S_1 and S_2 on the number line below. Put dots above the line for S_1 and dots below the line for S_2 .



- The *variance*, called $\text{Var}(X)$ or σ^2 , is defined to be

$$\sigma^2 = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_N - \mu)^2}{N} = E((X - \mu)^2) = E(X^2) - \mu^2.$$

This gives you a way to measure the spread (or dispersion) of the data, in contrast to the mean and median which describe only the center of the data. Compute the variances for both sets of data.

(Notice that the formula above is really *two* formulas: you can calculate $E((X - \mu)^2)$ or you can calculate $E(X^2) - \mu^2$. I think the second one is easier, plus you can write it as $E(X^2) - E(X)^2$, which looks kind of cool.)

3 Chebychev's Theorem Chebychev's Theorem says the following:

Let X be a random variable with mean μ and variance σ^2 . Then

$$P(\mu - k \leq X \leq \mu + k) \geq 1 - \frac{\sigma^2}{k^2}.$$

or

$$P(|X - \mu| \leq k) \geq 1 - \frac{\sigma^2}{k^2}.$$

Put another way: the probability that a randomly selected number lies between $\mu - k$ and $\mu + k$ is at least $1 - \frac{\sigma^2}{k^2}$.

(Note: This is always the case, but if we know more about the distribution of the random variable X , we can often say much more about this probability. In particular, $1 - \frac{\sigma^2}{k^2}$ is often a bad estimate for $P(|X - \mu| \leq k)$.)

Let's do some examples:

- (a) Suppose a candy company W&W's sells small candies by the bag (each imprinted with a small W). Experiments show that there are an average of $\mu = 55$ W's in each bag, with a variance of about $\sigma^2 = 6$. Use Chebychev's theorem to estimate the probability that a bag of candies will have between 50 and 60 (inclusive) W's.

- (b) Suppose the mean on an exam is 75 and the standard deviation is $\sigma = 10$. Estimate the probability that a randomly chosen student received a score between 60 and 90.

4 Sample versus Population There is a difference between *population data* (where we have information on the entire population) and *sample data* (where we have randomly selected a hopefully representative subset). If we assume S_1 and S_2 are *sample data* rather than population data, this changes the computations slightly. The mean is still computed the same:

$$\bar{X} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

but the variance is now

$$S^2 = \frac{(x_1 - \bar{X})^2 + (x_2 - \bar{X})^2 + \cdots + (x_n - \bar{X})^2}{n - 1}.$$

(Note the $n - 1$ in the denominator, not N .) Recompute the means and variances for S_1 and S_2 , now assuming they are sample data.

- 1 (a) For S_1 , $\mu = 10$ and the median is 8.
 For S_2 , $\mu = 10$ and the median is 9.5.
 (b) In both cases, $E(X) = \mu = 10$.
 (c) How did we calculate $E(X)$? If we write $S = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7\}$ (that is, if we write the outcomes of S as x_j), then

$$E(X) = x_1 \cdot \frac{1}{7} + x_2 \cdot \frac{1}{7} + \cdots + x_7 \cdot \frac{1}{7} = \frac{x_1 + x_2 + \cdots + x_7}{7}.$$

This is, of course, the average μ .

- (d) For both S_1 and S_2 , $E(X - \mu) = 0$. In fact, this is always the case. For S_1 :

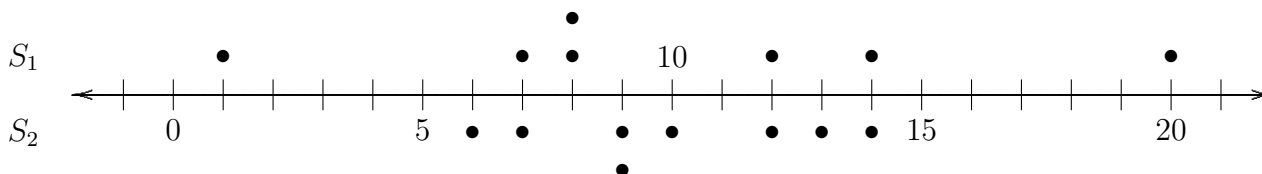
$$E((X - \mu)^2) = \frac{218}{7} = 31 \frac{1}{7} \approx 31.143 \quad \text{and} \quad E(X^2) = \frac{918}{7} = 131 \frac{1}{7} \approx 131.143.$$

For S_2 :

$$E((X - \mu)^2) = \frac{56}{8} = 7 \quad \text{and} \quad E(X^2) = \frac{856}{8} = 107$$

Note in both cases $E((X - \mu)^2) = E(X^2) - \mu^2$.

- 2 (a) Here are $S_1 = \{1, 7, 8, 8, 12, 14, 20\}$ (above the line) and $S_2 = \{6, 7, 9, 9, 10, 12, 13, 14\}$ (below the line):



- (b) For S_1 , $\sigma^2 = \text{Var}(X) = 31 \frac{1}{7} \approx 31.143$. For S_2 , $\sigma^2 = \text{Var}(X) = 7$. Notice that this somehow measures how widely each data set is dispersed around the mean. For S_1 , the dispersion about $\mu = 10$ is much greater than S_2 .

- 3 (a) $P(50 \leq X \leq 60) = P(|X - 55| \leq 5) \geq 1 - \frac{\sigma^2}{5^2} = 1 - \frac{6}{25} = 0.76$ or 76%.

(b) $P(60 \leq X \leq 90) = P(|X - 75| \leq 15) \geq 1 - \frac{\sigma^2}{15^2} = 1 - \frac{100}{225} = \frac{5}{9} \approx 55.56\%$.

- 4 For both S_1 and S_2 , $\bar{X} = E(X) = \mu = 10$.

For S_1 , $S^2 = \frac{218}{6} = 36 \frac{1}{3}$. For S_2 , $S^2 = \frac{56}{7} = 8$.