Lesson 12

Thursday 3/7/24

Interquartile Range

$$\mathrm{IQR} = Q_3 - Q_1 = P_{75} - P_{25}$$

where Q_3 = the value of x at the 3rd quartile (75th %ile) and Q_1 = the value of x at the 1st quartile (25th %ile). Key advantage: it reduces the influence of outliers or extreme observations; Key disadvantage: it ignores variation below Q_1 and above Q_3.

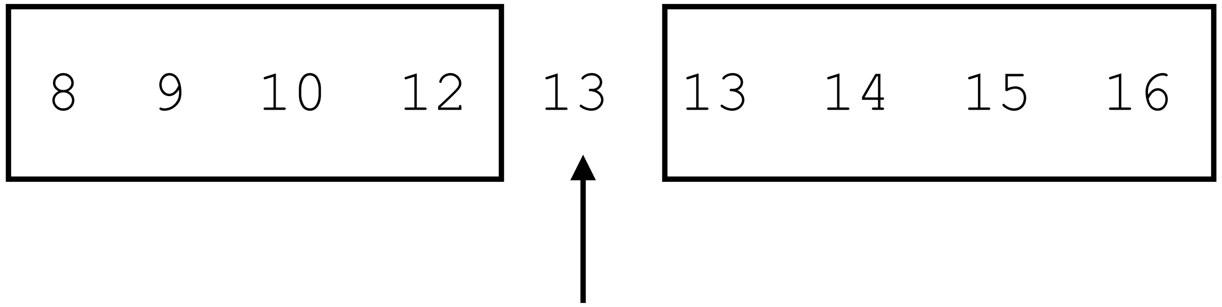
IQR Example

Ages of a sample of 9 kids appearing in juvenile court:

13 13 10 12 15 8 9 16 14

Step 1 - put the data in order: 8,9,10,12,13,13,14,15,16

Step 2 - let's divide up the distribution



8 9 10 12 13 13 14 15 16

Median, Q_2

Truncated Median Position (TMP) = 5

Step 3 - calculate the position of Q_1 and Q_3 by calculating (TMP+1)/2 = (5+1)/2 = 3

Step 4 - Q1 is in the third position from the left (10) and Q3 is in the third position from the right (14)

Step 5 - calculate IQR = 14-10 = 4

Another IQR Example

16 people sentenced to jail; these are the sentence lengths in months:

16 10 13 15 14 14 11 12 14 12 17 14 15 12 11 10

Step 1 - put the sentence lengths in order:

10 10 11 11 12 12 12 $\boxed{13\quad 14}$ 14 14 14 15 15 16 17

Median, Q_2 = 13.5
Median Position = 8.5

Truncated Median
Position (TMP) = 8

Step 3 - calculate the position of Q_1 and Q_3 by calculating (TMP+1)/2 = (8+1)/2 = 4.5

Step 4 - Q1 is in the 4.5th position (11.5) and Q3 is in the 12.5th position (14.5)

Step 5 - calculate IQR = 14.5-11.5 = 3

Mean Deviation

So far, we have used the variation ratio, the range and the inter-quartile range to get a single summary statistic that measures the dispersion of a distribution. The mean deviation gives us a measure of dispersion for each individual observation in the data set.

$$\text{Mean Deviation}(x_i) = x_i - \overline{X}$$

where x_i is an indivdual x score and $\overline{\mathrm{X}}$ is the sample mean of the variable, x.

Mean Deviation Example

Waiting times to rearrest for a sample of 8 recidivists:

11 11 10 14 12 9 8 12

Step 1: calculate the sample mean:

$$\overline{X} = \frac{\Sigma_{i=1}^{N} x_i}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{11+11+10+14+12+9+8+12}{8} = 10.875$$

Step 2: calculate the mean deviations:

$$x_{i=1} - \overline{X} = 11 - 10.875 = 0.125$$

$$x_{i=2} - \overline{X} = 11 - 10.875 = 0.125$$

$$x_{i=3} - \overline{X} = 10 - 10.875 = -0.875$$

$$\dots$$

$$x_{i=7} - \overline{X} = 8 - 10.875 = -2.875$$

$$x_{i=8} - \overline{X} = 12 - 10.875 = 1.125$$

Summarizing Mean Deviations

Interpreting a mean deviation statistic for each case rapidly becomes unwieldy. We would like to have a summary measure for the whole dataset. One intuitive idea is to sum up the deviations or calculate the "average" of the mean deviations.

$$x_{i=1} - \overline{X} = 11 - 10.875 = 0.125$$

$$x_{i=2} - \overline{X} = 11 - 10.875 = 0.125$$

$$x_{i=3} - \overline{X} = 10 - 10.875 = -0.875$$

$$x_{i=4} - \overline{X} = 14 - 10.875 = 3.125$$

$$x_{i=5} - \overline{X} = 12 - 10.875 = 1.125$$

$$x_{i=6} - \overline{X} = 9 - 10.875 = -1.875$$

$$x_{i=7} - \overline{X} = 8 - 10.875 = -2.875$$

$$x_{i=8} - \overline{X} = 12 - 10.875 = 1.125$$

Problem: the sum of these deviations is zero. Two things follow:

Mean Deviations Sum to Zero

$$\text{Sum of Mean Deviations} = \sum_{i=1}^{N} (x_i - \overline{X}) = 0$$

$$\text{Average of the Mean Deviations} = \frac{\text{Sum of Mean Deviations}}{\text{\# of Scores}} = \frac{\sum_{i=1}^{N} (x_i - \overline{X}) = 0}{8} = 0$$

Variance

We can avoid the "sum to zero" problem by squaring the deviations as follows:

$$\text{Population Variance} = \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{X})^2$$

$$\text{Sample Variance} = s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{X})^2$$

Variance Example

Among a sample of 7 people exiting prison, we count the number of prior arrests for each person:

11 7 5 8 10 9 7

Step 1: calculate the mean of the observations:

(11+7+5+8+10+9+7)/7 = 8.143

Step 2: calculate the mean deviation scores for each observation:

11-8.143 = 2.857
7-8.143 = -1.143
5-8.143 = -3.143
8-8.143 = -0.143
10-8.143 = 1.857
9-8.143 = 0.857
7-8.143 = -1.143

Variance Example (Cont'd)

Step 3: square each of the mean deviation scores:

$(11-8.143)^2 = 2.857^2 = 8.162$
$(7-8.143)^2 = -1.143^2 = 1.306$
$(5-8.143)^2 = -3.143^2 = 9.878$
$(8-8.143)^2 = -0.143^2 = 0.020$
$(10-8.143)^2 = 1.857^2 = 3.448$
$(9-8.143)^2 = 0.857^2 = 0.734$
$(7-8.143)^2 = -1.143^2 = 1.306$

Step 4: sum each of the squared mean deviation scores:

$8.162+1.306+9.878+0.020+3.448+0.734+1.306 = 24.854$

Step 5: multiply 1/(N-1) by the sum in step 4:

$1/(7-1) \times 24.854 = \boxed{4.142}$

The (n-1) Problem (advanced)

On p. 119, your book says, "[t]he reason we use n-1 in the denominator of the sample variance is that it is a biased estimator of the population value σ^2. To correct for this bias, we divide by n-1 rather than n in our sample formula." The following computer simulation demonstrates the bias referred to in your book:

```
> var1 <- vector()
> var2 <- vector()
>
> for(i in 1:1000000){
+   x <- rnorm(n=7,mean=0,sd=1)
+   xbar <- mean(x)
+   dsq <- (x-xbar)^2
+   var1[i] <- 1/7*sum(dsq)
+   var2[i] <- 1/6*sum(dsq)
+   }
>
> sum(var1)/i
[1] 0.8573886
> sum(var2)/i
[1] 1.000287
>
```

For this simulation, the true value of the variance is 1. I drew 1M samples of size N=7 from the population. I calculate the variance both ways; var1 multiplies the sum of the squared mean deviations by 1/7 while var2 multiplies by 1/6. As you can see, the var2 approach gives us the correct answer.

Standard Deviation

The simplest definition of the standard deviation is that it is the square root of the variance. Why do we use it? It is much easier to interpret in problems and applications we will be studying later in the semester. Let's return to our earlier example, beginning at step 4:

Step 4: sum each of the squared mean deviation scores:

8.162+1.306+9.878+0.020+3.448+0.734+1.306 = 24.854

Step 5: multiply 1/(N-1) by the sum in step 4:

1/(7-1) x 24.854 = 4.142

Step 6: take the square root of the variance:

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{X})^2} = \sqrt{4.142} = \boxed{2.035}$$