Chapter 9 - Quantifying Scatter

Laura Saba

July 11, 2017

What this chapter covers

- Interpreting a standard deviation
- How it works: Calculating SD
- Why *n-1*?
- Situations in which *n* can seem ambiguous
- SD and sample size
- Other ways to quantify and display variability

Example Data Set - Sleep Time

One characteristic related to an individual's likelihood of becoming alcohol dependent is how sensitive they are to alcohol.

In the Radcliffe lab, they measure the how sensitive a particular mouse is to the hypnotic effects of alcohol by giving them a large enough dose of alcohol to cause them to 'fall asleep' and measure the number of minutes that pass before they wake up.

Sleep time or Loss of Righting Reflex (LORR) is the number of minutes between when the mouse first loses the ability to right themselves when placed on their back to when they can right themselves again.

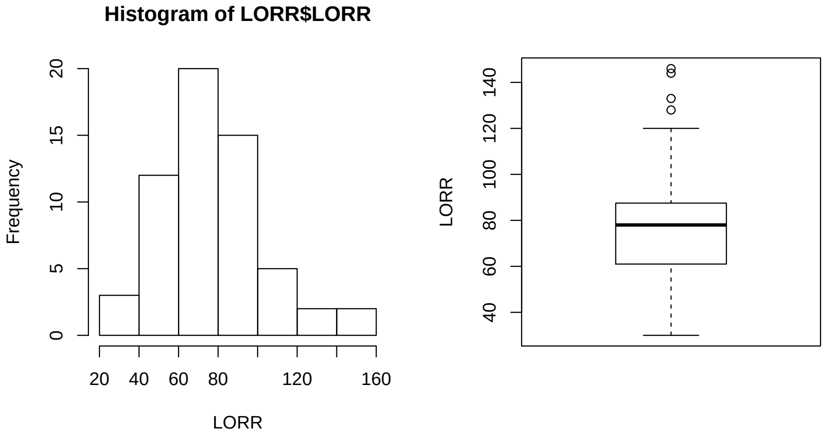
Sleep time

```
rm(list=ls())
options(stringsAsFactors = FALSE)

input = "/Volumes/sabal/Teaching/CoSIBS/sleepTime.txt"

LORR = read.table(file=input,sep="\t",header=TRUE)
summary(LORR)
```

```
##     Strain                  LORR       
##  Length:59          Min.   : 30.00  
##  Class :character   1st Qu.: 61.00  
##  Mode  :character   Median : 78.00  
##                     Mean   : 79.25  
##                     3rd Qu.: 87.50  
##                     Max.   :146.00  
```

Sleep time

```
par(mfrow=c(1,2))
hist(LORR$LORR,xlab="LORR")
boxplot(LORR$LORR,ylab="LORR")
```

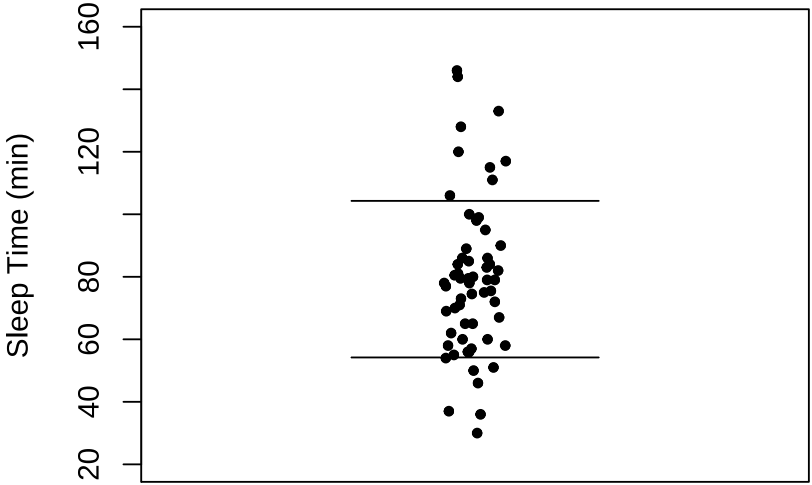


INTERPRETING A STANDARD DEVIATION

Standard deviation (SD) - is a measure of the variation among values that has the same units as the original data

Rule of Thumb: About two-thirds of the observations in a population *usually* lie within the range defined by the mean minus 1 SD to the mean plus 1 SD

Scatter and SD in Sleep Time



Two-Thirds Rule of Thumb

```r
mean_LORR = mean(LORR$LORR)
sd_LORR = sd(LORR$LORR)
n_LORR = length(LORR$LORR)
numValues = sum(LORR$LORR<(mean_LORR + sd_LORR)
                & LORR$LORR>(mean_LORR - sd_LORR))
pctValues = numValues/n_LORR
pctValues
```

```
## [1] 0.7288136
```

HOW IT WORKS: CALCULATING SD

Standard deviation is a summary of the 'deviation' of each value from the mean.

$$\text{SD} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{n-1}}$$

Step-by-step calculation of SD

```
## Step 1 - Calculate mean
mean.st = mean(LORR$LORR)

## Step 2 - Calculate the difference between
##          each value and the mean
dev.st = LORR$LORR - mean(LORR$LORR)

## Step 3 - Square each difference
sqDev.st = dev.st^2

## Step 4 - Add up the squared differences
totalSqDev.st = sum(sqDev.st)

## Step 5 - Divide by n-1
var.st = totalSqDev.st / (length(LORR$LORR) - 1)

## Step 6 - Take square root
SD.st = sqrt(var.st)
```

Step-by-step calculation of SD

```
## Step 1 - Calculate mean
mean.st
```

```
## [1] 79.24576
```

```
## Step 2 - Calculate the difference between each value anc
LORR$LORR[1:5]
```

```
## [1] 30 36 37 46 50
```

```
dev.st[1:5]
```

```
## [1] -49.24576 -43.24576 -42.24576 -33.24576 -29.24576
```

```
## Step 3 - Square each difference
sqDev.st[1:5]
```

```
## [1] 2425.1451 1870.1960 1784.7045 1105.2807  855.3146
```

Step-by-step calculation of SD

```
## Step 4 - Add up the squared differences
totalSqDev.st
```

```
## [1] 36381.69
```

```
## Step 5 - Divide by n-1
var.st
```

```
## [1] 627.2705
```

Step-by-step calculation of SD

```
## Step 6 - Take square root
SD.st
```

```
## [1] 25.04537
```

```
## DOUBLE CHECK WITH FUNCTION
sd(LORR$LORR)
```

```
## [1] 25.04537
```

WHY n-1?

Population SD vs. Sample SD

Population SD:

- assumes that your values represent the entire population
- cannot be extrapolated to other populations (rarely true in biostatistics)
- population mean is known (not estimated)
- denominator of SD formula is *n* instead of *n-1*

Sample SD:

- assumes your values only represent a subset of the entire population that you would like to make inference about
- CAN be extrapolated to other populations
- population mean is estimated from sample mean (loose one degree of freedom)
- denominator of SD formula is *n-1*

SITUATIONS IN WHICH n CAN SEEM AMBIGUOUS

- replicate measurements within subjects
- representative experiments
- trials with one subject

SD AND SAMPLE SIZE

SD estimates the variation within a population. Therefore:

- The estimate of SD does not differ based on sample size
- However, the estimate of SD will be more accurate with a larger sample size

OTHER WAYS TO QUANTIFY AND DISPLAY VARIABILITY

- Coefficient of variation
- Variance
- Interquartile range
- Five-number summary
- Median absolute deviation

Coefficient of variation

Coefficient of variation (CV) = SD/mean

- used to 'normalize' the standard deviation
- Example:
  - mean = 10 cm, sd = 2 cm
  - mean = 100 mm, sd = 20 mm
  - CV = 2/10 = 0.2 or CV = 20/100 = 0.2
- NOTICE: the CV is unitless
- often used to SD across different units of measurement

Coefficient of variation in R

```r
sd(LORR$LORR)/mean(LORR$LORR)
```

```
## [1] 0.3160468
```

Variance

Variance $= SD^2$

- squared units are hard to interpret, but most statistical theory is based on variance not standard deviation

```r
var(LORR$LORR)
```

```
## [1] 627.2705
```

Interquartile Range

Interquartile Range (IQR) = 75th percentile - 25th percentile

- See box plots from earlier

```
quantile(LORR$LORR,0.75) - quantile(LORR$LORR,0.25)
```

```
##  75%
## 26.5
```

Five-Number Summary

Five-Number Summary:

1. minimum
2. 25th percentile
3. 50th percentile (median)
4. 75th percentile
5. maximum

```
quantile(LORR$LORR,c(0,0.25,0.50,0.75,1))
```

```
##     0%   25%   50%   75%  100%
##   30.0  61.0  78.0  87.5 146.0
```

```
summary(LORR$LORR)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   30.00   61.00   78.00   79.25   87.50  146.00
```

Median absolute deviation

The median absolute deviation (MAD) is often used when there are outliers or the data distribution is non-normal (more to come on that in the next chapter).

Calculation of MAD:

1. Find the median value
2. Calculate the deviation of each value from the median
3. Take the absolute value of each deviation
4. Find the median value of the absolute deviations

Median absolute deviation - in R

Step-by-step

```
# Step 1 - find median value
median_value = median(LORR$LORR)
median_value
```

```
## [1] 78
```

```
# Step 2 - calculate deviations from median
dev = LORR$LORR - median_value
dev[1:5]
```

```
## [1] -48 -42 -41 -32 -28
```

Median absolute deviation - in R

Step-by-step (cont.)

```
#Step 3 - take absolute value of deviations
abs_dev = abs(dev)
abs_dev[1:5]
```

```
## [1] 48 42 41 32 28
```

```
#Step 4 - Find the median value of the absolute deviations
MAD = median(abs_dev)
MAD
```

```
## [1] 13
```

Median absolute deviation - using R function

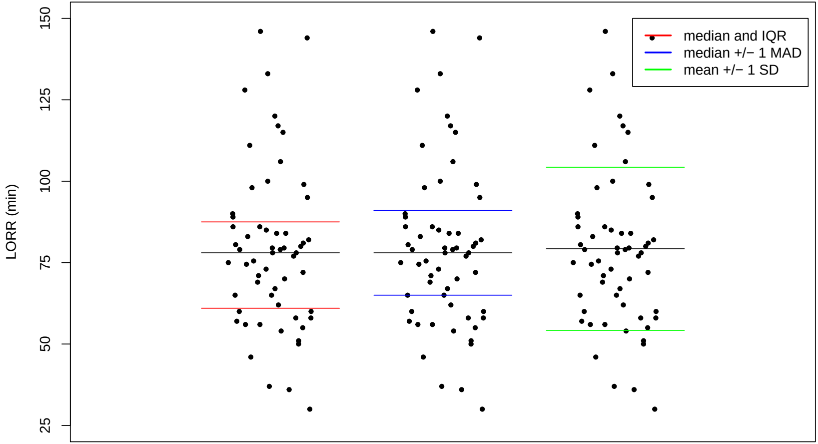
```r
mad(LORR$LORR,constant=1)
```

```
## [1] 13
```

MAD vs. IQR vs. SD

- half the values are within 1 MAD of the median
- the IQR contains half the values
- the exact number of values that fall within 1 SD of the mean varies
- the region $+/$- 1 MAD around the median is symmetric, but more values may follow above the median then below the median or vice versa
- the IQR is not symmetric around the median, but will have the same number of values above the median as there are below the median

MAD vs. IQR vs. SD



What did we learned

- The most common way to quantify scatter is with a standard deviation
- A useful rule of thumb is that about two thirds of the observations in a population usually lie within the range defined by the mean minus 1 SD to the mean plus 1 SD
- Other methods used to quantify scatter are:
  - variance (SD squared)
  - coefficient of variation (SD/mean)
  - interquartile range
  - median absolute deviation