Computing and reporting descriptive statistics

Measures of spread & kurtosis for the mean

Jenine Harris
Brown School

Defining and calculating spread

In addition to using central tendency to characterize a variable, reporting a corresponding measure of how spread out the values are around the central value is also important to understanding numeric variables. Each measure of central tendency has one or more corresponding measures of spread.

- mean: use **variance** or **standard deviation** to measure spread
- median: use **range** or **interquartile range (IQR)** to measure spread
- mode: use the **index of qualitative variation** to measure spread

Spread to report with the mean

- The **variance** is the average of the squared differences between each value of a variable and the mean of the variable.

$$s_x^2 = \frac{\sum_{i=1}^{n}(x_i - m_x)^2}{n-1}$$

- s_x^2 is the variance of x
- the $\sum$ symbol is sum
- the x_i is each individual value of x
- the m_x is the mean of x
- n is the sample size
- Overall, the variance is the sum of the squared differences between each value of x and the mean of x (or the sum of squared deviation scores) divided by the sample size minus one.

Importing & cleaning the data

```
# import brfss data
brfss.trans.2014 <- read.csv(file = "data/transgender_hc_ch2.csv")

# open tidyverse for data management
library(package = "tidyverse")
# recode 77, 88, 99 on PHYSHLTH
brfss.2014.cleaned <- brfss.trans.2014 %>%
  mutate(PHYSHLTH = na_if(PHYSHLTH, 77)) %>%
  mutate(PHYSHLTH = na_if(PHYSHLTH, 99)) %>%
  mutate(PHYSHLTH = as.numeric(recode(PHYSHLTH, `88` = 0L)))

# examine PHYSHLTH to check data management
summary(object = brfss.2014.cleaned$PHYSHLTH)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   0.000   0.000   4.224   3.000  30.000   10303
```

Using R to find the variance

```
# variance of unhealthy days
var(x = brfss.2014.cleaned$PHYSHLTH, na.rm = TRUE)
```

```
## [1] 77.00419
```

- There is no direct interpretation of the variance.
- It is a general measure of how much variation there is in the values of a variable.

The standard deviation

- A more useful measure of spread is the standard deviation, which is the square root of the variance.

$$s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - m_x)^2}{n-1}}$$

Finding spread in the tidyverse

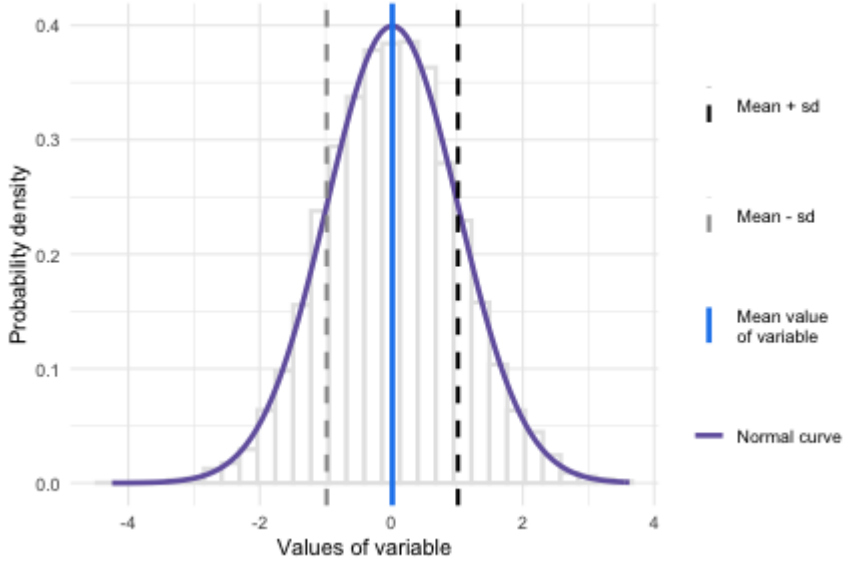
- Adding the variance and standard deviation to the **`tidyverse`** descriptive statistics code they had been accumulating.

```
# get mean, median, mode and spread
brfss.2014.cleaned %>%
  summarise(mean.days = mean(x = PHYSHLTH, na.rm = TRUE),
            sd.days = sd(x = PHYSHLTH, na.rm = TRUE),
            var.days = var(x = PHYSHLTH, na.rm = TRUE),
            med.days = median(x = PHYSHLTH, na.rm = TRUE),
            mode.days = names(x = sort(x = table(PHYSHLTH),
                                       decreasing = TRUE))[1])
```

```
##   mean.days  sd.days var.days med.days mode.days
## 1  4.224106 8.775203 77.00419        0         0
```

Interpreting the standard deviation

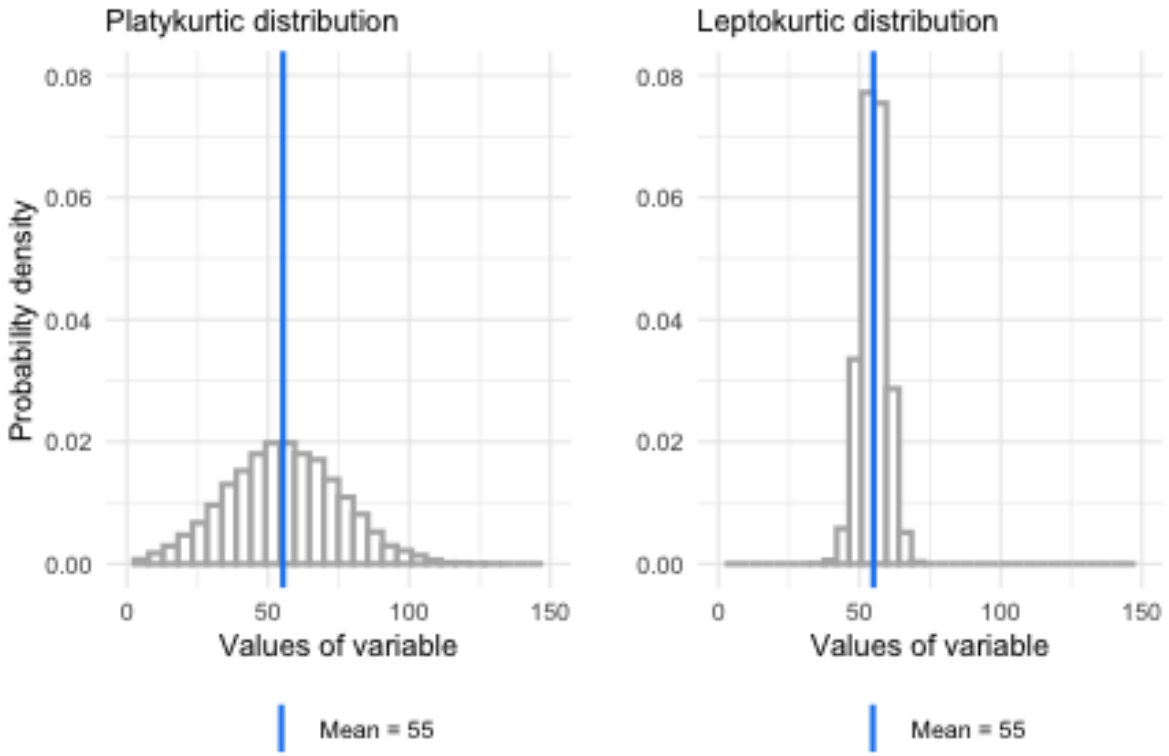
- The standard deviation was sometimes interpreted as the average amount an observation differs from the mean.
- This is conceptually close and a good way to think about it, but is not 100% accurate (see Equation).
- A visual representation can help interpretation:



Kurtosis

- In addition to central tendency and spread, **kurtosis** is another measure that helps determine if a continuous variable is truly normally distributed.
    - Kurtosis measures how many observations are in the tails of a distribution.
    - Some distributions look bell shaped, but have a lot of observations in the tails (platykurtic) or very few observations in the tails (leptokurtic).
- Sometimes kurtosis is described as how pointy a distribution is, with **platykurtic** distributions being more *flat* (as in a *platy*pus has a flat beak) and **leptokurtic** distributions being more *pointy*.
    - This is a common way of describing the shapes of these distributions, but that, technically, kurtosis measures whether there are many or few observations in the tails of the distribution.

Visualizing kurtosis



Catosis

- Kurtosis is also found in nature...

Why does kurtosis matter?

- Platykurtic and leptokurtic deviations from normality do not necessarily influence the mean, since it will still be a good representation of the middle of the data *if the distribution is symmetrical and not skewed*.
- However, a distribution with the same mean that is platykurtic or leptokurtic will usually have a different variance and standard deviation compared to a normal distribution.
- The variance and standard deviation are used not only to quantify spread, but are also used in many of the common statistical tests.
- Here is the formula for kurtosis where n is the sample size, s_x is the standard deviation of x, x_i is each value of x, and m_x is the mean of x. The $\sum_{i=1}^{n}$ symbol indicates the values from the first value of x (i = 1) to the last value of x (i = n) should be summed:

$$kurtosis_x = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - m_x}{s_x}\right)^4$$

Interpreting values of kurtosis

- A normal distribution will have a kurtosis value of three; distributions with kurtosis = 3 are described as **mesokurtic**.
- If kurtosis is above or below 3, there is excess kurtosis.
- Values of kurtosis above 3 indicate the distribution is *leptokurtic*, with fewer observations in the tails than a normal distribution (the fewer observations in the tails often give a distribution a pointy "look").
- Values of kurtosis below 3 indicate the distribution is *platykurtic*, with more observations in the tails than a normal distribution would have given the mean, standard deviation, and sample size (looks more flat).
- The same cutoff values from skew also apply for the z for small, medium, and large sample sizes in kurtosis:
    - If the sample size is small ($n < 50$), z values outside the -2 to 2 range are a problem.
    - If the sample size is between 50 and 300, z values outside the -3.29 to 3.29 range are a problem.
    - For large samples ($n > 300$), using a visual is recommended over the statistics, but generally z values outside the range of -7 to 7 can be considered problematic.

Computing kurtosis in R

- The semTools package includes the `kurtosis()` function to compute the kurtosis and a few related values.
- The `kurtosis()` function subtracts 3 from the kurtosis, so positive values will indicate a leptokurtic distribution and negative will indicate a platykurtic distribution.
- For example, computing kurtosis for the two variables used in the leptokurtic and platykurtic graph above, saved as `var1` in `lepto.plot` and `platy.plot` respectively:

```
# kurtosis of leptokurtic distribution variable
semTools::kurtosis(object = lepto.plot$var1)
```

```
## Excess Kur (g2)              se               z               p
##      0.05206405      0.04898979      1.06275295      0.28789400
```

```
# kurtosis of platykurtic distribution variable
semTools::kurtosis(object = platy.plot$var1)
```

```
## Excess Kur (g2)              se               z               p
##     -0.04920369      0.04898979     -1.00436604      0.31520221
```

- The values of z for the two variables used in the example graphs are small and so are not problematic regardless of sample size, so using statistics that rely on a normal distribution seems ok.