Computing and reporting descriptive statistics

Measures of spread for the median

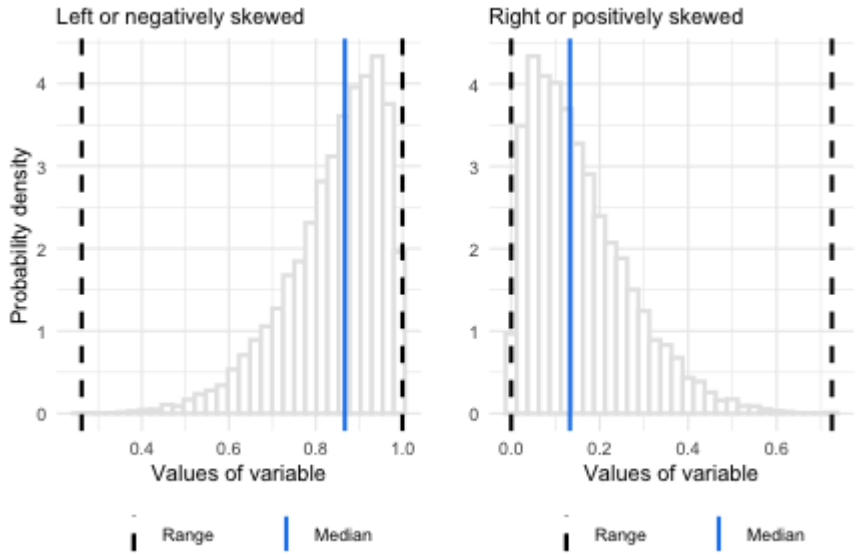
Jenine Harris
Brown School

Spread to report with the median

- When distributions are not normally distributed, the median is often a better choice than the mean.
- For medians, however, they need something different to report for spread.
- The variance and standard deviation will not work.
- Just like the very large values influence the mean, they also influence the standard deviation since the mean is part of the standard deviation formula.

Options for measuring spread with the median

- First, the **range** is the span between the largest and smallest values of a variable.



Computing the range in R

To compute the range for the `PHYSHLTH` variable:

```
# import brfss data
brfss.trans.2014 <- read.csv(file = "~/Box/teaching/Teaching/Fall2020/da

# 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)))

# range of days of physical health
range(brfss.2014.cleaned$PHYSHLTH, na.rm = TRUE)
```

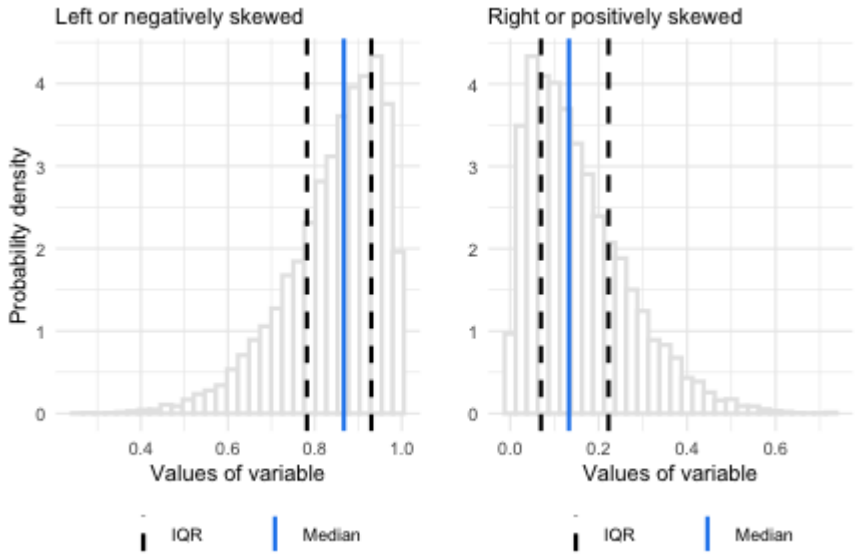
```
## [1]  0 30
```

Interpreting the range

- The range of unhealthy days in a month is 0 to 30.
- For the `PHYSHLTH` variable, the ends of the range are the highest and lowest possible values of the variable.
- The range does not provide any indication of how the data are distributed across the possible values.
- For example, maybe there is one person who has 30 days of poor physical health and everyone else is between 0 and 10.
- Or, maybe half the people have 0 and half the people have 30 and no people have a value in between.

Interquartile range (IQR)

- The **interquartile range** or **IQR** might be more appropriate for this sort of data, or for data that are highly skewed.
- The IQR is the difference between the first and third quartiles.
- A quartile is one-quarter of the data, so the difference between the first and third quartiles would be the boundaries around the middle 50% of the data.



Getting the IQR in R with tidyverse

- Use `IQR()` and add it to the tidyverse descriptive statistics list:

```
# get descriptive statistics for PHYSHLTH
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),
            iqr.days = IQR(x = PHYSHLTH, na.rm = TRUE),
            mode.days = names(x = sort(x = table(PHYSHLTH),
                                       decreasing = TRUE))[1])
```

```
##   mean.days  sd.days var.days med.days iqr.days mode.days
## 1  4.224106 8.775203 77.00419        0        3         0
```

Optimizing the code

- The `na.rm =` is repeated 5 times in the tidyverse code.
- When something is repeated often, there may be a more efficient option.
- The `drop_na()` code might work, like this:

```
# get descriptive statistics
brfss.2014.cleaned %>%
  drop_na(PHYSHLTH) %>%
  summarise(mean.days = mean(x = PHYSHLTH),
            sd.days = sd(x = PHYSHLTH),
            var.days = var(x = PHYSHLTH),
            med.days = median(x = PHYSHLTH),
            iqr.days = IQR(x = PHYSHLTH),
            mode.days = names(x = sort(x = table(PHYSHLTH),
                                       decreasing = TRUE))[1])
```

```
##   mean.days  sd.days var.days med.days iqr.days mode.days
## 1  4.224106 8.775203 77.00419        0        3         0
```

Upper and lower bounds of IQR

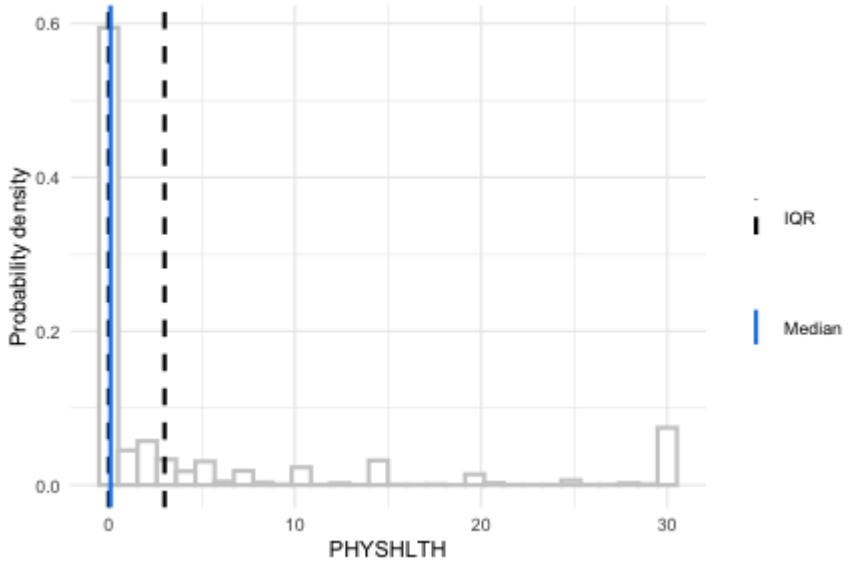
- There is no way to do this in the `IQR()` command, but the `quantile()` function can be used to find the bounds around the middle 50%, which are the IQR boundaries.

```
#interquartile range of unhealthy days
quantile(x = brfss.2014.cleaned$PHYSHLTH, na.rm = TRUE)
```

```
##   0%  25%  50%  75% 100%
##    0    0    0    3   30
```

- The middle 50% of the data is between the 25% and 75% quantiles.
- In this case, the bounds around the middle 50% of unhealthy days is 0 to 3.
- Fifty percent of observations (people) in this data set have between 0 and 3 physically unhealthy days per month.

Visualizing the IQR



- The lower boundary of the IQR is the same as the median in the graph, which is due to so many values being zero and this being the lowest possible value of the variable.