R Numeric Summaries

Young W. Lim

2018-07-27 Fri

Outline

Based on

"Using R for Introductory Statistics", 2nd ed John Verzani

TOC

- Center
  - the aample mean
  - the sample median
  - measures of position
  - other measures of center
- Spread
  - the variance and standard deviation
  - IQR
- Shape
  - viewing the shape of a data set

Center

- a sense of the center of a data set
- mean
  - the average value
- median
  - the middle value of in the sorted data set
- mode -the most common value

Spread

- measures the variability in a data set
- how far from the center something is
- a sense of scale
- if the variability is large,
  the mean informs much less
- without sense of variability,
  interpretation would not assure

Shape

- influences how much we can interprete
  from knowing both the center and spread
- are values larger than the mean equallly likely as for values less?
- are values very far from the mean really unlikely or not so unlikely?
- are there values where the measurements cluster?
- are the possible values spread out?
- the bell shape
  - the two sides are equally likely
  - large values are rather unlikely
  - values tend to cluster near the mean

TOC

- the sample mean (`mean`)
- the sample median (`median`)
- measure of position (`quantiles`)

The Sample Mean

```
x <- c(38, 43, ..., 27)
sort(x)
mean(x)
devs <- x - mean(x)
mean(devs)

mean(x, trim=0.10)   # trim 10% of both ends

w <- Macdonell$frequency / sum(Macdonell$frequency)  # n_k / n
y <- Macdonell$height
sum(w*y)
```

The Sample Median

```
median(x)

n <- length(x); trim=0.10
lo <- 1 + floor(n*trim)
hi <- n + 1 - lo
median(sort(x[lo:hi])
```

Measures of position

```
x <- 0:5
length(x)
mean(sort(x)[3:4])
median(x)
quantile(x, 0.25)
quantile(x, seq(0, 1, by=0.2)
quantile(x)

fivenum(x)
```

Other measures of center

```
income <- c("90"=110651, "95"=155193, ... , "99.99"=7969900)
income

table(x)
table(x) == max(table(x))
which(table(x) == max(table(x)))
as.numeric(names( which(talbe(x) == max(table(x))) ))
```

TOC

- `range`
- `diff`
- variance (`var`)
- standard deviation (`sd`)
- the z-score (`z_score`)
- `scale`
- IQR (InterQuantile Range, `IQR`)
- mad (median absolute deviation, `mad`)

The variance and standard deviation

```
range(x)  # min and max values
diff(range(x))

var(x)
sum( (x-mean(x)^2) / (length(x)-1)

x <- c(10500, ..., NA, 62000)
range(x, na.rm=TRUE)
sd(x, n.ram=TRUE)
```

The z-Score

```
z_score <- function(x) (x-mean(x))/sd(x)
z_score(x)
scale(x)[,1]

x<-c(54, 50, ..., 80)
z<-(x-mean(x))/sd(x)
x[z>=1.28]
mean(x)+1.28*sd(x)

z <- (x - mean(x))/ sd(x)
out <- abs(z) >
sum(out)/length(z)

sd(x)/mean(x)
```

IQR

```
median(x)

IQR(x)
IQR(x)/sd(x)

mad(x)/sd(x)

x<- kid.weights$height
mad(x)/sd(x)
```

TOC

- Symmetry
- Skew (`skew`)
- tail (`kurtosis`)
- viewing the shape
  - dot plots
  - stem-and-leaf
  - histogram
  - density plot
  - box plot
  - quantile graphs

Symmetry and skew

```
skew <- function(x) {
  n <- length(x)
  z <- (x - mean(x)) / sd(x)
  sum(z^3) / n
}

skew(x)
```

Tail

```
kurtosis <- function(x) {
  n <- length(x)
  z <- (x - mean(x)) / sd(x)
  sum(z^4)/n - 3
}
```

Modes

```
kurtosis <- function(x) {
  n <- length(x)
  z <- (x - mean(x)) / sd(x)
  sum(z^4)/n - 3
}
```

TOC

- Dot plots
- Stem and leaf plot
- Histogram
- Density plots
- Box plots
- Quantile graphs