Univariate Data - 2. Numeric Summaries

Young W. Lim

2018-08-01 Mon

Outline

Based on

"Using R for Introductory Statistics" John Verzani

Functions

```
function(x) {
  sum(x) / length(x)
}

functions(x) {
  total <- sum(x)
  n <- length(x)
  total / n
}

my_mean <- function(x) {
  sum(x) / length(x)
}

my_mean( c(1,2,3,4) )
```

TOC

- Center
  - the sample 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

Equations (1)

- sample mean

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}\sum_i x_i$$

- weighted averages

$$\frac{1}{n}\sum_k n_k \cdot y_k = \sum_k \frac{n_k}{n} \cdot y_k = \sum_k w_k \cdot y_k$$

- sample variance

$$s^2 = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2$$

Equations (2)

- sample standard deviation

$$\sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_i (x_i - \bar{x})^2}$$

- sample skewness

$$\sqrt{n}\frac{\sum (x_i - \bar{x})^2}{(\sum (x_i - \bar{x})^2)^{3/2}} = \frac{1}{n}\sum_i z_i^3$$

- sample excess kurtosis

$$n\frac{\sum (x_i - \bar{x})^4}{(\sum (x_i - \bar{x})^2)^2} - 3 = \frac{1}{n}\sum z_i^4 - 3$$

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 trimmed mean

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

mean(exec.pay)
mean(exec.pay, trim=0.10)

the Macdonell (HistData) data set
w <- Macdonell$frequency / sum(Macdonell$frequency)
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)
```

Sample standard deviation

```
x <- c(100, 300, 900, NA, 200, 500, 700)
range(x, na.rm=TRUE)
sd(x, na.rm=TRUE)

z_score <- function(x) (x - mean(x)) / sd(x)
z_score(x)

scale(x)[,1]  # to extract the 1st column

z <- (x -mean(x)) / sd(x)
x[ z >= 1.28]

mean(x) + 1.28 * sd(x)

z <- (exec.pay- mean(exec.pay)) / sd(exec.pay)
out <- abs(z) > 3
sum(out) / length(z)
```

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)
```

mad (Median Absolute Deviation)

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

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

Dot plots

- a numer line for the range of data
- dots for each data point
- for repeated data values, use stacks or jitter
  - jitter gives a samll random variance
- mean : the balancing point
- median : the middle point
- IQR : four parts (top and bottom quarters)
- useful to identify the shape of data set
  - whether it is skewed or multimodal
- limitation
  - repeated data values
  - only for relatively small data set

```
jitter
stripchart
```

Stem-and-leaf plots (1)

- records a number of for each data point
- placing the number with the proper stem
- place the stems in a vertical column
- sort the leaves
- think a data set containing two digit decimal numbers
  - a 1s-digit and a 10s-digit
  - eg. 16 :
    - a 1 (stem)
    - a 6 (leaf)

```
0 | 0022344     ---> (0, 0, 0, 2, 2, 3, 4, 4)
1 | 2344        ---> (12, 13, 14, 14)
2 | 67          ---> (26, 27)
3 | 1           ---> (31)
```

Stem-and-leaf plots (2)

- a stem-and-leaf plot shows
  - the sorted data
  - the data range
  - the median (the middle)
  - a rough shape (skewedness)
  - good for relatively small size data set
- `stem`
  - `scale` argument - to adjust the meaning of the stem when too many leaves in a stem
- eg. `stem(bumper)`
  - `The decimal point is 3 digit(s) to the right of the |`
  - `21 ---> 2.1e^3 = 2100`

```
stem
stem(bumpers)
```

Histogram (1)

- group indivisual data points
- represent them with a bar of a given area
- bins : to break the number line into sub-intervals
- count the number of data points for each sub-interval
- draw a bar with a size proportional to the proportion of data points
- can identify
  - center (mean, median)
  - spread
  - shape

```
hist(x)
hist(x, probability=TRUE) # probability scaling
# scale the y-axis so the entire area is 1
```

Histogram (2)

```
bins <- seq(40, 100, by=5)
# create bins with a size of 5 each
# data ranges from 40 to 100

x <- faithful$waiting

out <- cut(x, breaks=bins)
# count the number of values in each bin
# categorize each value by the breaks(bins) specified
# represented bin example : (75, 80]

head(out)
table(out)
```

Density plots

```
plot( density(bumpers) )

b_hist <- hist(bumpers, plot=FALSE)
b_dens <- density(bumpers)

hist(bumpers, probability=TRUE,
     xlim=range(c(b_hist$breaks, b_dens$x)),
     ylim=range(c(b_hist$density, b_dens$y)) )
lines(b_dens, lwd=2)
```

Standard plotting arguments

xlim	set x coordinate range
ylim	set y coordinate range
xlab	set label for x axis
ylab	set label for y axis
main	set the main title
pch	adjust plot symbols (?pch)
cex	adjust size of text and symbols
col	adjust color of objects drawn (?colors)
lwd	adjust width of lines drawn
lty	adjust how line is drawn
	"blank", "solid", "dashed", dotdash"
bty	adjust bos type ("o", "l", "7",
	"c", "u", "]"

Functions to add layers

points	add points to a graphic
lines	add points connected by lines to a graphic
abkube	add a line of the form a + bx, y=h, or x=v
text	add text to a graphic
mtext	add text to margins of a graphic

Various plots

```
hist(bumpers, probability=TRUE,
     xlim = range( c(b_hist$breaks, b_dens$x)),
     ylim = range( c(b_hist$density, b_dens$y)))
lines(b_dens, lwd=2)

boxplot(bumpers, horizontal=TRUE, main="Bumpers")

x <- rep(macdonell$finger, Macdonell$frequency)
qqnorm(x)

x <- jitter(HistData::Galton$child, factor=5)
qqnorm(x)
```