Descriptive Statistics

Statistics

- **Descriptive statistics**
  - the discipline of quantitatively describing the main features of a collection of information
  - aim to summarize a sample, rather than use the data to learn about the population
- **Inferential statistics**
  - the process of deducing properties of an underlying distribution by analysis of data
  - the observed data is assumed to be sampled from a larger population
  - based on probability theory

https://en.wikipedia.org/wiki/Descriptive_statistics

Descriptive statistics

- **Measures of central tendency**
  - Mean
  - Median
  - Mode
- **Measures of variability or dispersion**
  - Standard deviation
  - Variance
  - Min
  - Max
  - Kurtosis
  - Skewness

https://en.wikipedia.org/wiki/Descriptive_statistics

Central tendency

Type	Description	Example	Result
Arithmetic mean	Sum of values of a data set divided by number of values: $\bar{x}=\frac{1}{n}\sum_{i=1}^{n} x_i$	(1+2+2+3+4+7+9) / 7	**4**
Median	Middle value separating the greater and lesser halves of a data set	1, 2, 2, **3**, 4, 7, 9	**3**
Mode	Most frequent value in a data set	1, **2**, **2**, 3, 4, 7, 9	**2**

https://en.wikipedia.org/wiki/Central_tendency

Central tendency

```
v <- c(1,2,2,3,4,7,9)
mean(v)
median(v)
mode(v)
table(v)
sort(table(v))
rev(sort(table(v)))
names(rev(sort(table(v))))[1]

v <- c(1,2,2,3,4,7,9,NA)
mean(v,na.rm = TRUE)
```

Measures of variability

- Min
- Max
- Standard deviation

$$s_N = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2},$$

- Variance

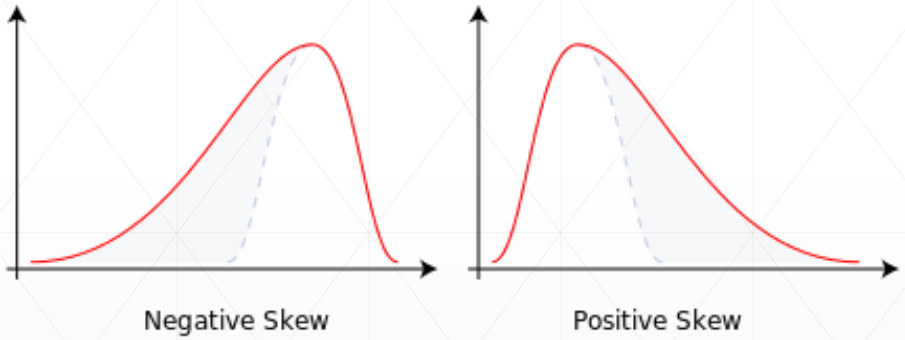
$$\operatorname{Var}(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2.$$

- Skewness
- Kurtosis

Skewness

- a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean

left-skewed, *left-tailed* *right-skewed*, *right-tailed*



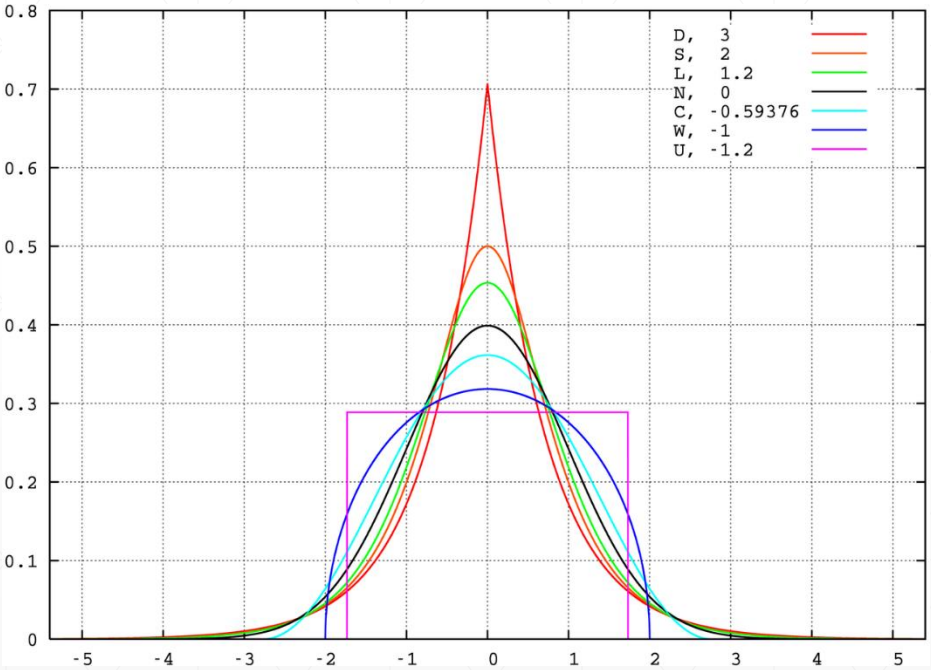
$$\gamma_1 = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\mu_3}{\sigma^3} = \frac{\mathrm{E}\left[(X-\mu)^3\right]}{(\mathrm{E}\left[(X-\mu)^2\right])^{3/2}} = \frac{\kappa_3}{\kappa_2^{3/2}},$$

https://en.wikipedia.org/wiki/Skewness

Kurtosis

- a descriptor of the shape of a probability distribution

$$\operatorname{Kurt}[X] = \frac{\mu_4}{\sigma^4} = \frac{\operatorname{E}[(X-\mu)^4]}{(\operatorname{E}[(X-\mu)^2])^2},$$



https://en.wikipedia.org/wiki/Kurtosis

Descriptive Statistics

```
v <- c(1,2,2,3,4,7,9)
min(v)
max(v)
range(v)
var(v)
sd(v)
```

```
library(moments)
skewness(v)
kurtosis(v)
quantile(v)
length(v)
which.max(v)
which.min(v)
```

Reading from CSV Files

- The `read.csv` function can read CSV files

```
csv.path = "http://spatial.binghamton.edu/geog533/data/students.csv"
df <- read.csv(csv.path)
df <- read.csv("students.csv")
str(df)
df$Last.Name <- as.character(df$Last.Name)
df$First.Name <- as.character(df$First.Name)
str(df)

df <- read.csv("students.csv",header = TRUE,as.is = TRUE)
str(df)
```

http://spatial.binghamton.edu/geog533/data/students.csv

Reading from Excel Files

- Two ways:
  - Convert Excel files to csv files, then use `read.csv` function
  - Use `readxl` to read Excel files

```
install.packages("readxl")
library(readxl)
# http://spatial.binghamton.edu/geog533/data/students.xls"
df.xls <- read_excel("students.xls")
View(df.xls)
```

Writing to CSV Files

- You want to save a matrix or data frame in a file using the comma-separated values format.

```
write.csv(x, file="filename", row.names=FALSE)


df <- read.csv("students.csv")
df2 <- df[,1:5]
write.csv(df2,file = "stu.csv")
write.csv(df2,file = "stu2.csv",row.names = FALSE)
```

Exploring Data

- `df <- read.csv("students.csv")`
- `summary(df) # Provides basic descriptive statistics and frequencies.`
- `edit(df) # Open data editor`
- `str(df) # Provides the structure of the dataset`
- `names(df) # Lists variables in the dataset`
- `head(df) # First 6 rows of dataset`
- `head(df, n=10)# First 10 rows of dataset`
- `head(df, n= -10) # All rows but the last 10`
- `tail(df) # Last 6 rows`
- `tail(df, n=10) # Last 10 rows`
- `df[1:10, ] # First 10 rows`
- `df[1:10,1:3] # First 10 rows of data of the first 3 variables`
- `df[c("Last.Name","First.Name","City","State")]`

http://spatial.binghamton.edu/geog533/data/students.csv

Descriptive Statistics using fBasics

```
install.packages("fBasics")

library(fBasics)

df <- read.csv("students.csv")

SAT <- df$SAT

basicStats(SAT)

summary(SAT)

hist(SAT)

hist(df$SAT,main = "Histogram of SAT Score",xlab =
"SAT Score",ylab = "Frequency",col="green")
```

Descriptive statistics by groups

- Descriptive statistics by groups using **tapply**
  - `tapply(X, INDEX, FUN = NULL, ..., simplify = TRUE)`

```
df <- read.csv("students.csv")
SAT.mean <- tapply(df$SAT, df$Gender, mean)
SAT.median <- tapply(df$SAT, df$Gender, median)
SAT.sd <- tapply(df$SAT, df$Gender, sd)
SAT.max <- tapply(df$SAT, df$Gender, max)

round(cbind(SAT.mean,SAT.median,SAT.sd,SAT.max),digits = 1)
t1 <- round(cbind(SAT.mean,SAT.median,SAT.sd,SAT.max),digits = 1)
```

Descriptive statistics by groups

- Descriptive statistics by groups using **aggregate**
  - `aggregate(x, by, FUN, ..., simplify = TRUE)`

```
df <- read.csv("students.csv")

aggregate(df[c("Age","SAT")], df["Gender"],mean,na.rm=TRUE)

aggregate(df[c("Age","SAT")],by=list(sex=df$Gender,major=df$Major,status=df$Student.Status),mean)
```