STAT021 Statistical Methods II

Lecture 2 Variables and Distributions

Lu Chen
Swarthmore College

9/6/2018

Outline

- Data structure, variables and relationships
- Distribution
- Normal distribution
- Population, sample, parameter, statistic
- Random sampling
- Sampling distribution
- Central Limit Theorem (CLT)

Data structure

```
horse <- read.table("../Datasets/HorsePrices.txt",sep="\t",header=T) # Input the data
dim(horse) # dimension of the dataset
```

```
## [1] 50  5
```

```
head(horse, 10) # first 10 rows of the dataset
```

```
##    HorseID Price Age Height Sex
## 1       97 38000   3  16.75   m
## 2      156 40000   5  17.00   m
## 3       56 10000   1     NA   m
## 4      139 12000   8  16.00   f
## 5       65 25000   4  16.25   m
## 6      184 35000   8  16.25   f
## 7       88 35000   5  16.50   m
## 8      182 12000  17  16.75   f
## 9      101 22000   4  17.25   m
## 10     135 25000   6  15.25   f
```

- Rows: cases/**observations**/objects/subjects
    - Each row is a unique object with a unique ID
- Columns: **variables**
    - Categorical variable: qualititative values, several categories
    - Quantitative variable: numerical values
    - Label/ID: usually the first column

Variables and relationships

> **Association**: Two variables measured on the same observation are **associated** if knowing the values of one of the variables tells you something about the values of the other variable.

```r
head(horse, 10) # first 10 rows of the dataset
```

```
##    HorseID Price Age Height Sex
## 1       97 38000   3  16.75   m
## 2      156 40000   5  17.00   m
## 3       56 10000   1     NA   m
## 4      139 12000   8  16.00   f
## 5       65 25000   4  16.25   m
## 6      184 35000   8  16.25   f
## 7       88 35000   5  16.50   m
## 8      182 12000  17  16.75   f
## 9      101 22000   4  17.25   m
## 10     135 25000   6  15.25   f
```

- Two associated variables
  - **Response** vs. **explanatory** variable
  - Dependent vs. independent variable
  - Outcome vs. predictor
- Types of relationships:
  - Association (does NOT imply causation)
  - Causation

Describe a categorical variable

```r
horse$Sex # View the variable Sex
```

```
##  [1] m m m f m f m f m f m f f m f m m f f f m m m f m m m m m f f m f
## [34] f m m m m m m f f f f f m m m m m
## Levels: f m
```

Table of counts/proportions

```r
table(horse$Sex)
```

```
## 
##  f  m 
## 20 30
```

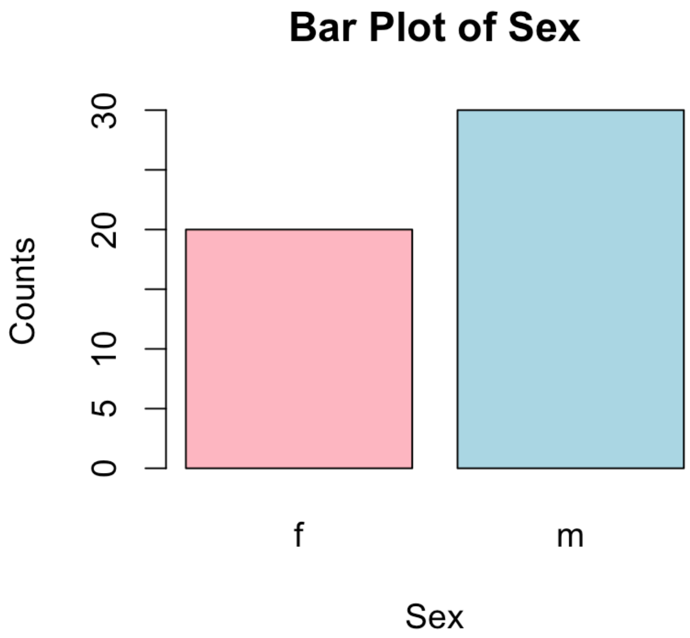
```r
tab.sex <- table(horse$Sex)
prop.table(tab.sex)
```

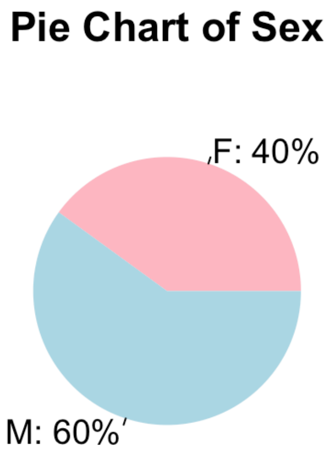
```
## 
##   f   m 
## 0.4 0.6
```

Describe a categorical variable

Bar plot and pie chart: display the *distribution* of a categorical variable

```
barplot(tab.sex,xlab="Sex",ylab="Counts",main="Bar Plot of Sex",
        col=c("lightpink","lightblue"))
pie(tab.sex, main="Pie Chart of Sex", labels=c("F: 40%","M: 60%"),
    col=c("lightpink","lightblue"))
```





Describe a quantitative variable

Summary statistics

```
horse$Price
```

```
##  [1] 38000 40000 10000 12000 25000 35000 35000 12000 22000 25000 40000
## [12] 25000  4500 19900 45000 45000 48000 15500  8500 22000 35000 16000
## [23] 16000 15000 33000 20000 25000 30000 50000  1100 15000 45000  2000
## [34] 20000 45000 20000 50000 50000 39000 20000 12000 15000 27500 12000
## [45]  6000 15000 60000 50000 30000 40000
```

```
mean(horse$Price); sd(horse$Price)
```

```
## [1] 26840
```

```
## [1] 14980.22
```

```
median(horse$Price); quantile(horse$Price, prob=c(0.25, 0.75))
```

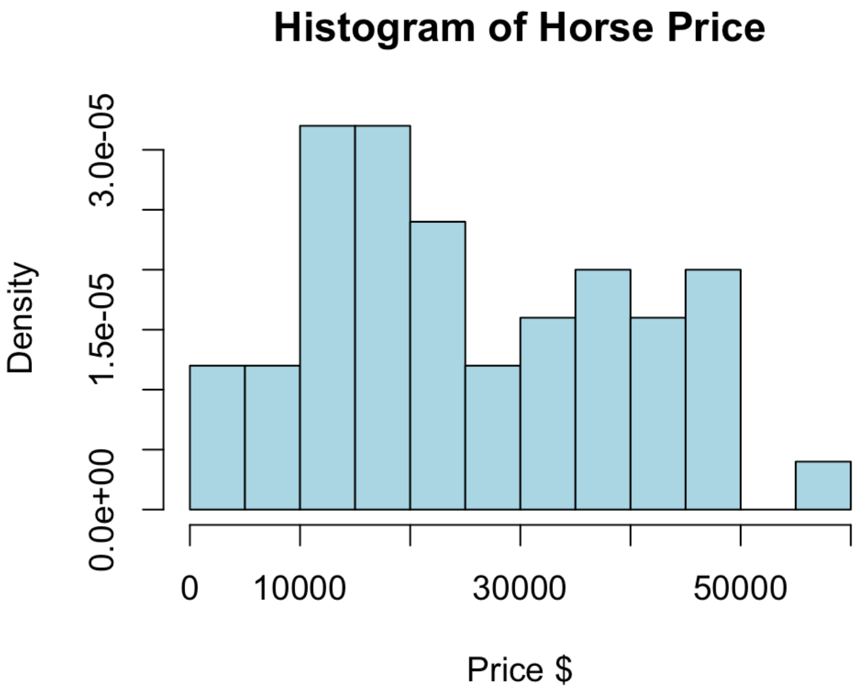
```
## [1] 25000
```

```
##   25%   75%
## 15000 39750
```

Describe a quantitative variable

Histogram: display the *distribution* of a quantitative variable

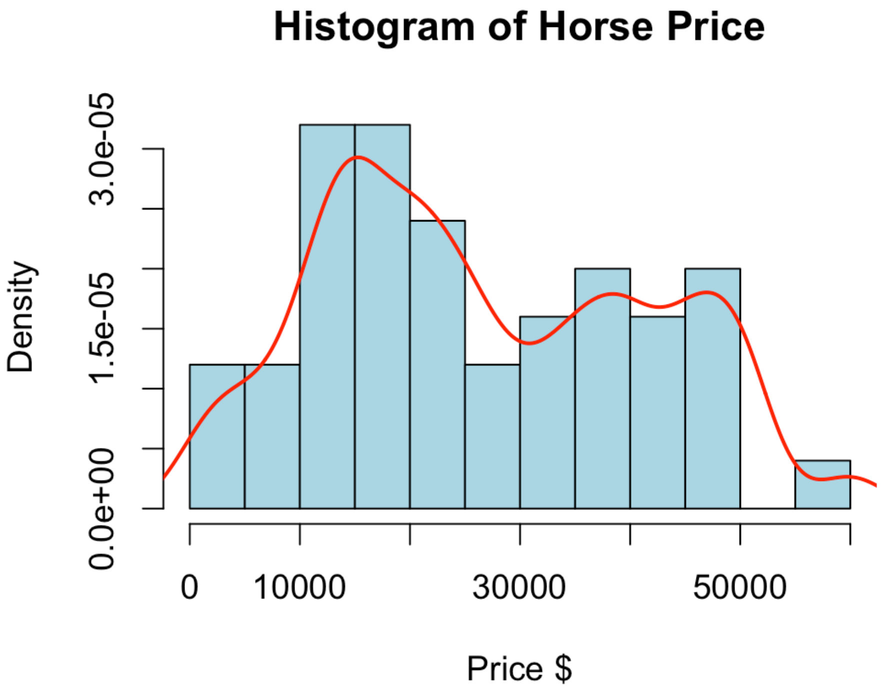
```
hist(horse$Price, breaks=15, col="lightblue", freq=F,
     xlab="Price $", main="Histogram of Horse Price")
```



- x-axis: values of *Price*
- y-axis: density (NOT probability/proportion)
- y-axis value of the 3rd bar: '3.2e-05' $= 3.2 \times 10^{-5} = 0.000032$
- Area of the 3rd bar: $5000 \times 0.000032 = 0.16$
- 16% of the 50 horses have prices between 10 and 15 thousand dollars.
- Total area of all the bars: 1

Distribution and density curve

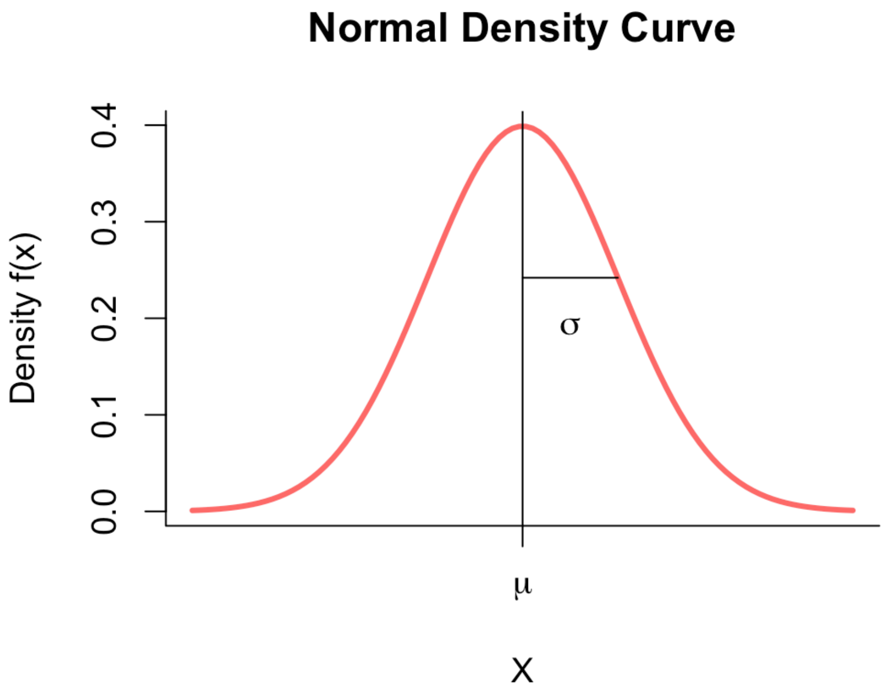
```r
hist(horse$Price, breaks=15, col="lightblue", freq=F,
     xlab="Price $", main="Histogram of Horse Price")
lines(density(horse$Price, adjust=0.5), col="red", lwd=2)
```



> A **density curve** describes the overall pattern of a distribution.

- The total area under the curve is 1.
- The **area** under the curve and above any range of values is the **proportion** of all observations that fall in that range.

Normal distribution

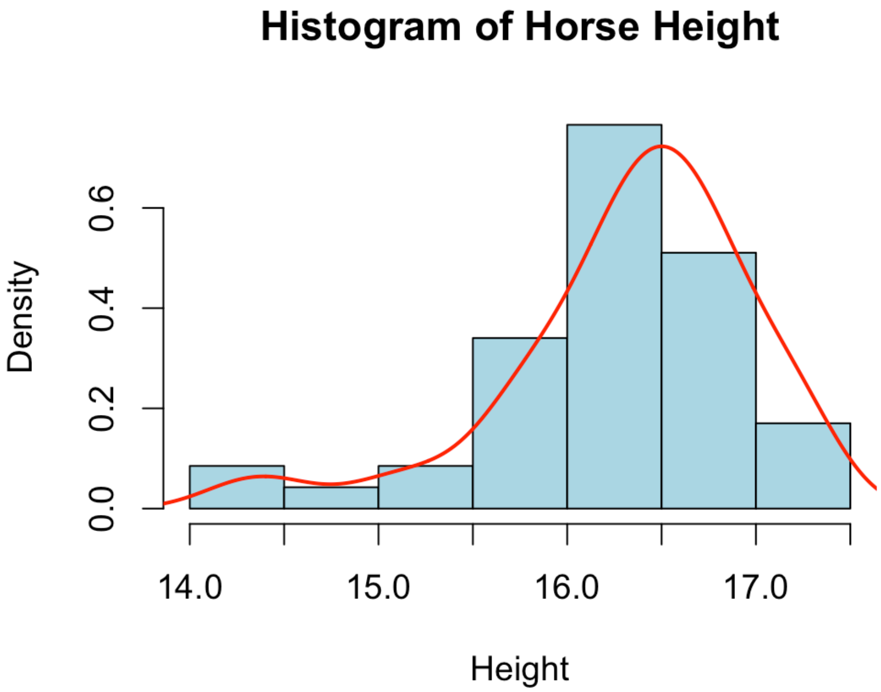


- **Normal density curves** are used to describe **Normal distributions**.
- Unimodal, symmetric and bell-shaped
- Normal density curve is characterized by its center and spread
  - Center: mean μ
  - Spread: standard deviation σ
- $X \sim N(\mu, \sigma)$: variable X follows a Normal distribution with mean μ and standard deviation σ.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Normal distribution

```r
hist(horse$Height, breaks=10, col="lightblue", freq=F,
     xlab="Height", main="Histogram of Horse Height")
lines(density(horse$Height, na.rm=T), col="red", lwd=2)
```

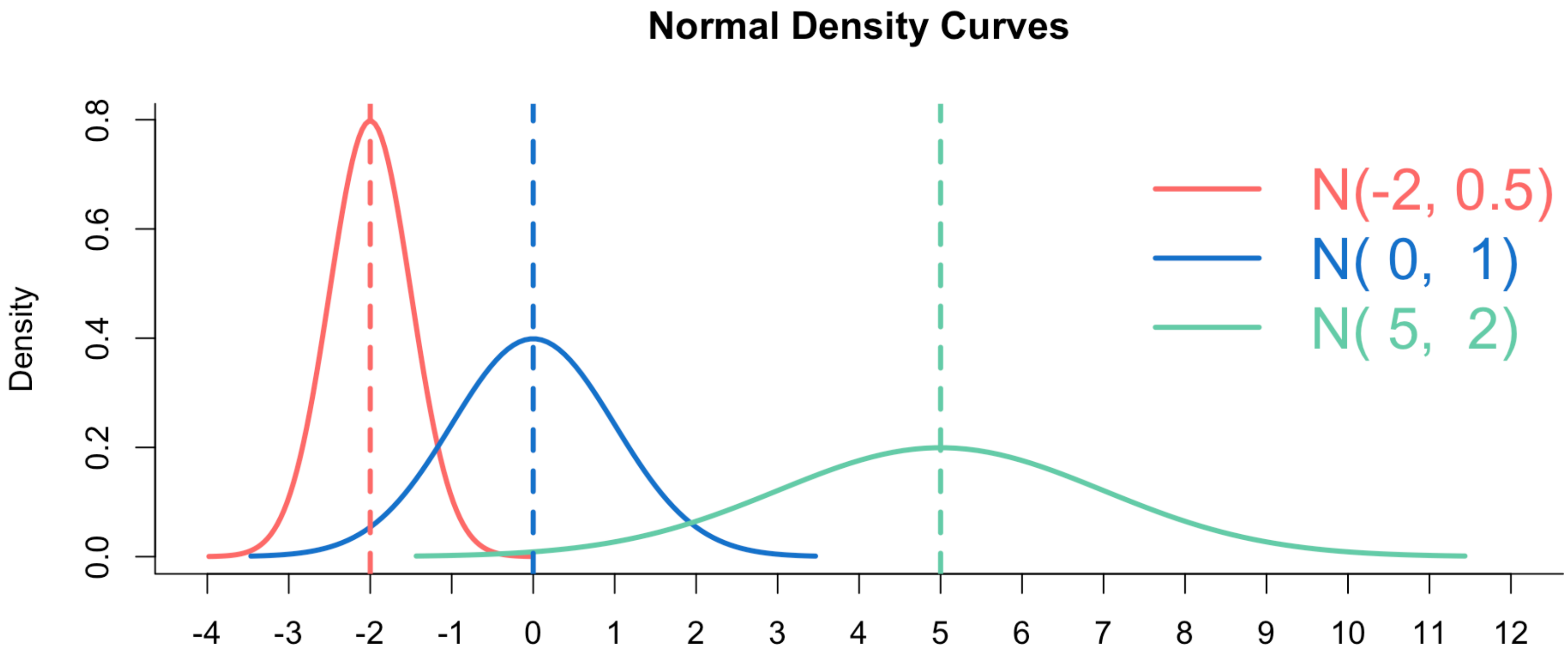


- Mean: 16.3 inches
- Standard deviation (SD): 0.7 inch
- Denote the height of the horses as X and suppose $X \sim N(\mu, \sigma)$, we may write

$$X \sim N(16.3, 0.7)$$

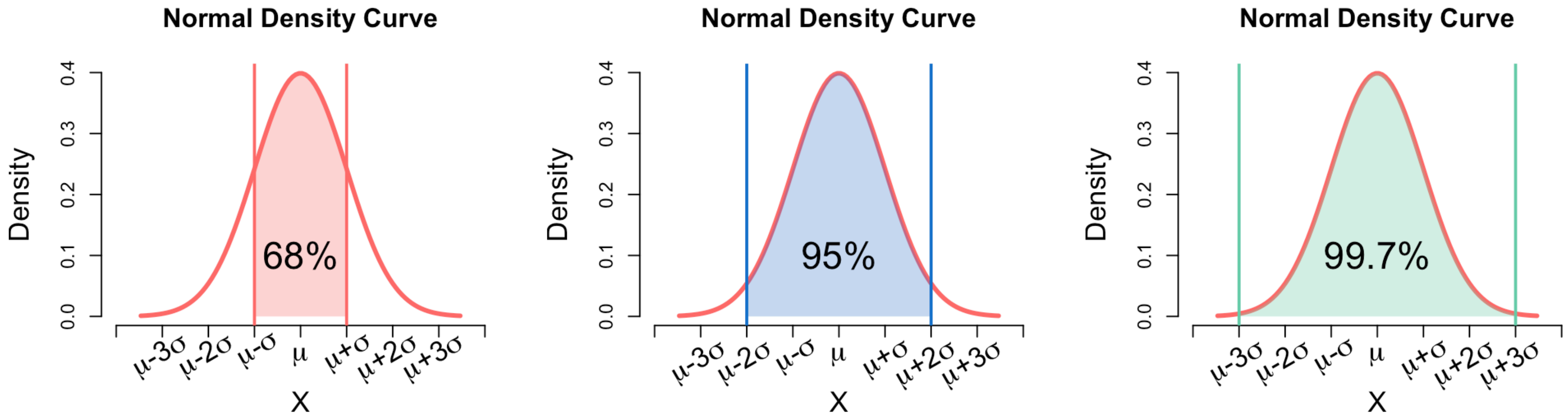
X follows an approximate Normal distribution with mean 16.3 inches and SD 0.7 inch.

Normal distribution



- μ: mean of X, determines the center of the curve, *location* parameter
- σ: SD of X, determines the width of the curve, *scale* parameter
- Note: the area under each curve is always 1.

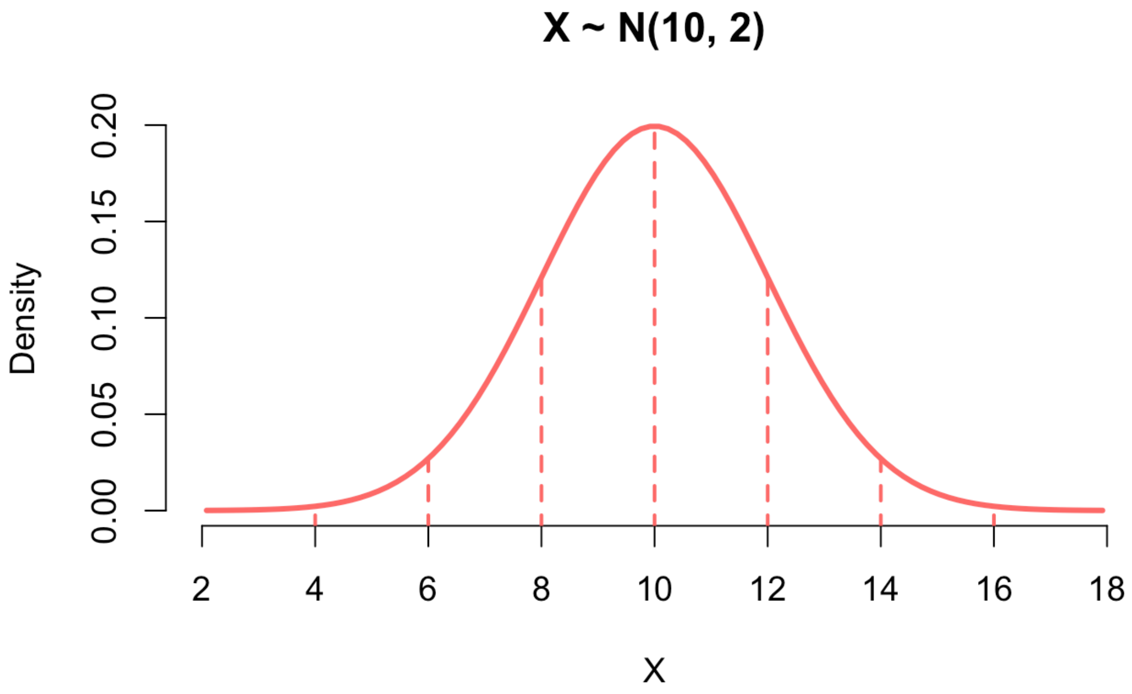
Normal distribution 68-95-99.7 rule



For any Normal distribution with mean μ and standard deviation σ:

- Approximately 68% of the observations fall within one σ of the mean μ.
- Approximately 95% of the observations fall within 2σ of μ.
- Approximately 99.7% of the observations fall within 3σ of μ.

Normal distribution 68-95-99.7 rule



For $X \sim N(10, 2)$, what is the proportion of observations that fall

- between 8 and 12?
- between 6 and 14?
- between 4 and 16?
- below 8?
- above 6?
- between 4 and 6?
- above 11.5?

Normal distribution

```r
# Functions for Normal distribution: dnorm(), pnorm() and qnorm()
dnorm(x=8, mean=10, sd=2) # input: x value; output: density value
```

```
## [1] 0.1209854
```

```r
pnorm(q=8, mean=10, sd=2) # input: q value; output: proportion *below* q
```

```
## [1] 0.1586553
```

```r
qnorm(p=0.5, mean=10, sd=2) # input: proportion *below* q; output: q value
```

```
## [1] 10
```

```r
1 - pnorm(q=0.43, mean=0, sd=1) # output: proportion above q for N(0,1)
```

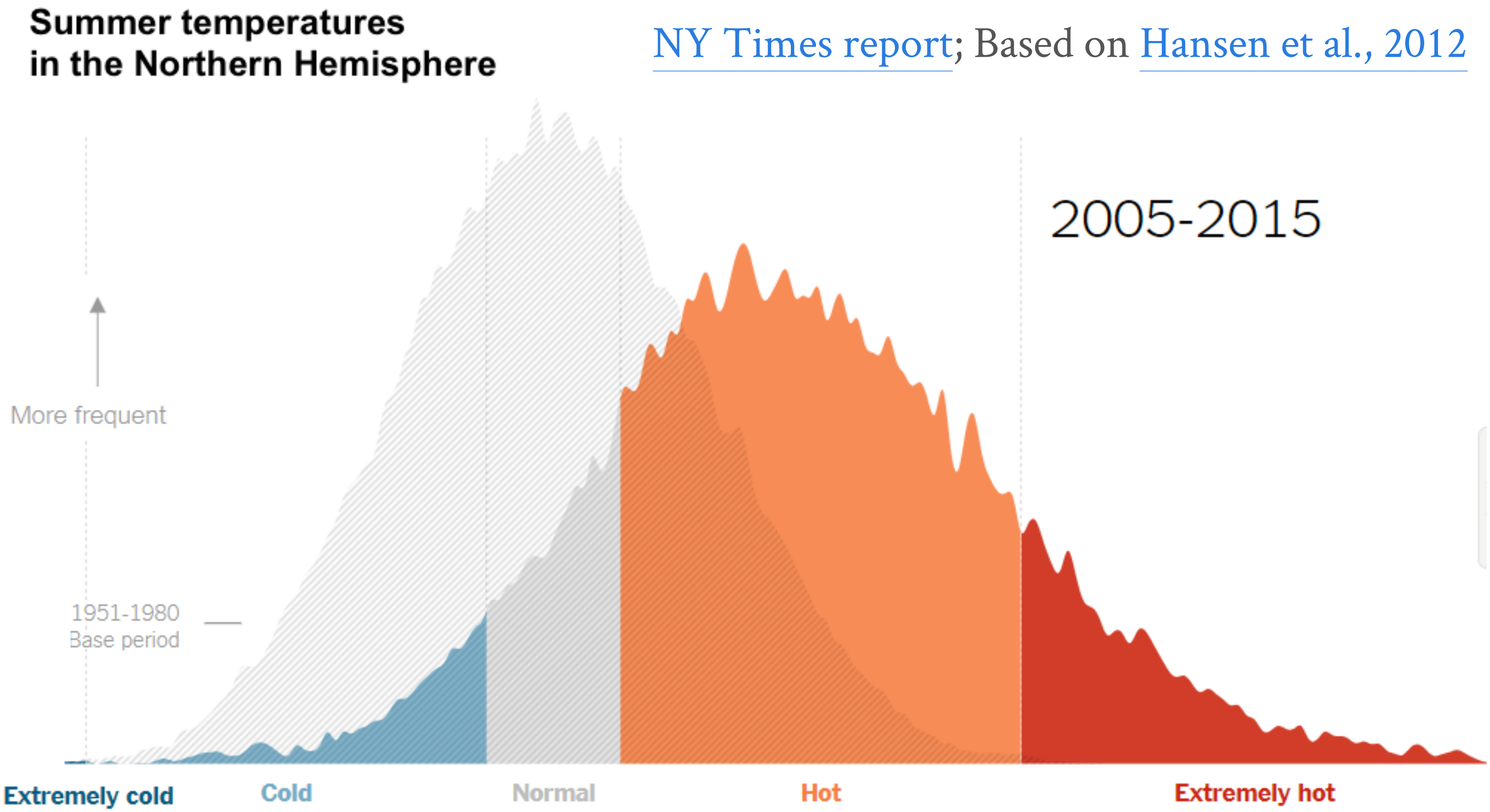
```
## [1] 0.3335978
```

```r
pnorm(0.43, 0, 1) - pnorm(-0.43, 0, 1) # output: proportion between -0.43 and 0.43
```

```
## [1] 0.3328044
```

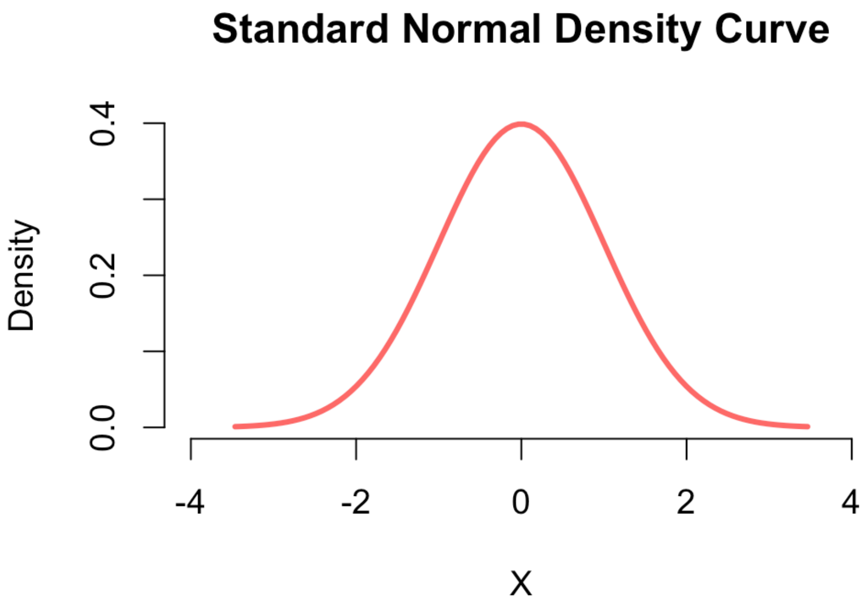
Normal distribution example

NY Times report; Based on Hansen et al., 2012



Normal distribution

> The **standard Normal distribution** is the Normal distribution with mean 0 and standard deviation 1.


Standard Normal Density Curve

- $Z \sim N(0, 1)$
- $\mu = 0, \sigma = 1$

Any Normal distribution $N(\mu, \sigma)$ (e.g., $N(5, 2)$) can be tranformed to the standard Normal distribution. This process is called **standardization**.

- If $X \sim N(\mu, \sigma)$, then

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

- For example, $X \sim N(5, 2)$, then $Z = \frac{X-5}{2} \sim N(0, 1)$

Normal distribution

Important properties of Normal distribution

1. **Standardization**. For $X \sim N(\mu, \sigma)$,

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

2. **Linear tranformation**. For $X \sim N(\mu, \sigma)$,

$$Y = a + bX \sim N(a + b\mu, b\sigma)$$

3. **Linear combination**. For $X_1 \sim N(\mu_1, \sigma_1)$ independent of $X_2 \sim N(\mu_2, \sigma_2)$,

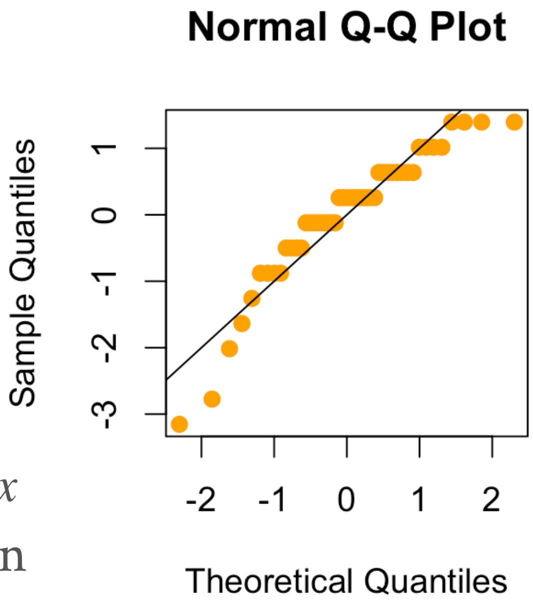
$$X_1 + X_2 \sim N\left(\mu_1 + \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2}\right)$$

Normal distribution

```
ht <- horse$Height[!is.na(horse$Height)] # remove NAs in the Height variable
s.ht <- (ht - mean(ht))/sd(ht) # standardization
qqnorm(s.ht, pch=19, col="orange") # Q-Q plot
abline(a=0, b=1) # Add the y=x line
```

How to know a variable is Normally distributed?

- Look at the histogram
  - Unimodal? Symmetric? Bell-shaped?
- Check Normal Q-Q (quantile-quantile) plot
  - Compare the the quantiles of a variable of interest to the standard Normal variable and see how close they are.
  - For the *Height* variable, most points lie closely to the $y = x$ line, only the points at the tails are a little off. Distribution of *Height* is quite close to Normal.



Population, sample, parameter, statistic

Population: The entire group of individuals that we want information about.

- eg. horses for sale on the internet

Sample: A part of the population that we actually examine in order to gather information.

- eg. 50 horses we randomly picked from the internet

A **parameter** is a number that describes the **population**.

- Fixed and unknown
- eg. mean price of horses on the internet μ (or SD σ)

A **statistic** is a number that describes a **sample**.

- Changes from sample to sample; estimates the parameter
- eg. mean price of the 50 horses we picked $\bar{x}$ (or SD s)

Random sampling

A sample statistic (mean of the 50 horses) can be used to estimate an unknown population parameter (mean of all horses for sale on the internet) only if the sampling process is **random**. For example, sample mean

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

is an estimate of the population mean μ only if $x_1, x_2, \cdots, x_n$ are **randomly generated** from the population distribution. If the sampling process is NOT random, the sample is biased. Any statistic calculated from the sample is also biased.

> A **simple random sample (SRS)** of size n is generated in such a way that each of the n individuals in the sample has an equal chance to be chosen from the population.

Random sampling

```r
# Randomly generate 5 observations from population distribution N(10, 2)
x1 <- rnorm(5, mean = 10, sd = 2); x1 # Sample 1
```

```
## [1] 11.586026 11.044503 13.492444  7.457328 14.394779
```

```r
x2 <- rnorm(5, mean = 10, sd = 2); x2 # Sample 2
```

```
## [1] 10.866262  6.859601  8.130189 10.126987  9.995213
```

```r
x3 <- rnorm(5, mean = 10, sd = 2); x3 # Sample 3
```

```
## [1]  5.446438 11.514824  8.903189 10.345099 11.125706
```

```r
c(mean(x1), mean(x2), mean(x3)) # mean of x1, x2 and x3
```

```
## [1] 11.595016  9.195650  9.467051
```

```r
c(sd(x1), sd(x2), sd(x3)) # SD of x1, x2 and x3
```

```
## [1] 2.686192 1.649997 2.459611
```

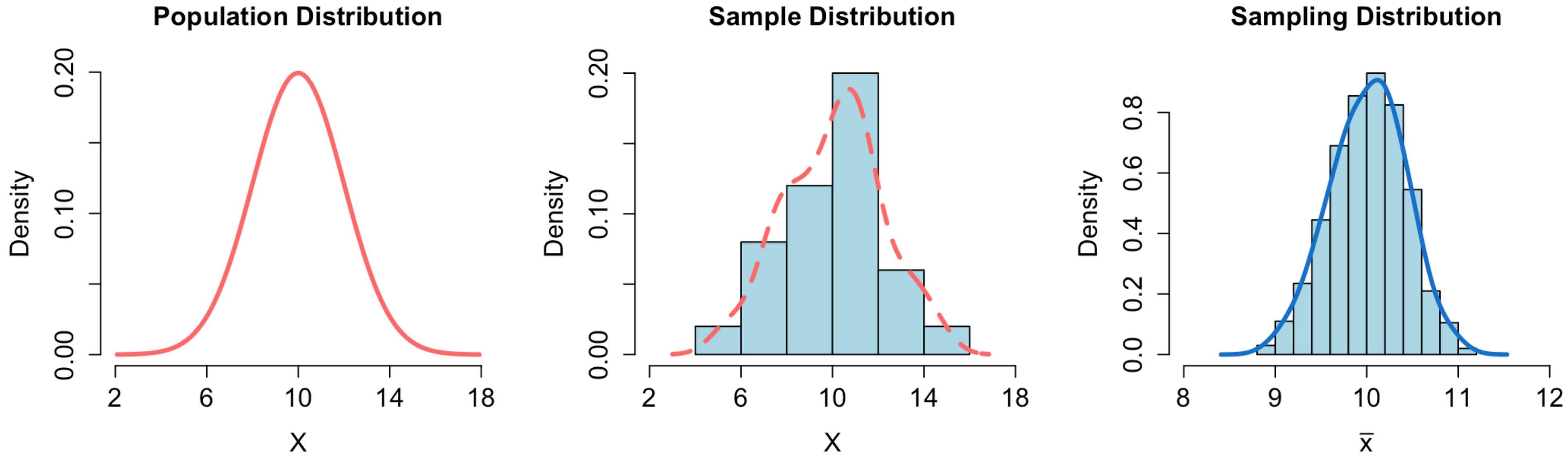
Sampling distribution

- The 3 samples are different.
- The statistics computed from the 3 samples are different - **sampling variability**.
  - $\bar{x}_1 = 11.6, \bar{x}_2 = 9.2, \bar{x}_3 = 9.5$
  - $s_1 = 2.7, s_2 = 1.6, s_3 = 2.5$

> The **sampling distribution** of **a statistic** is the distribution of values taken by the statistic in *all possible samples* of **the same size** from **the same population**.

- **Population distribution**: distribution of a population
- **Sample distribution**: distribution of a sample
- **Sampling distribution**: distribution of a statistic in repeated sampling

Sampling distribution



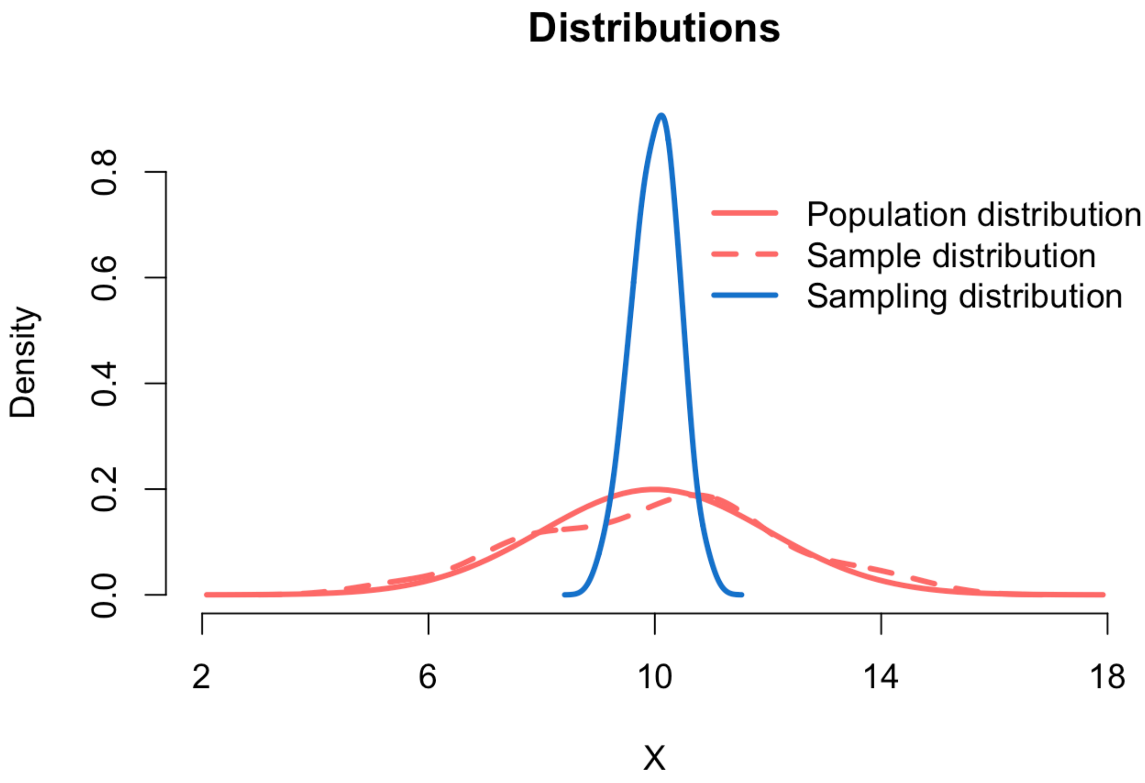
Population Distribution

Sample Distribution

Sampling Distribution

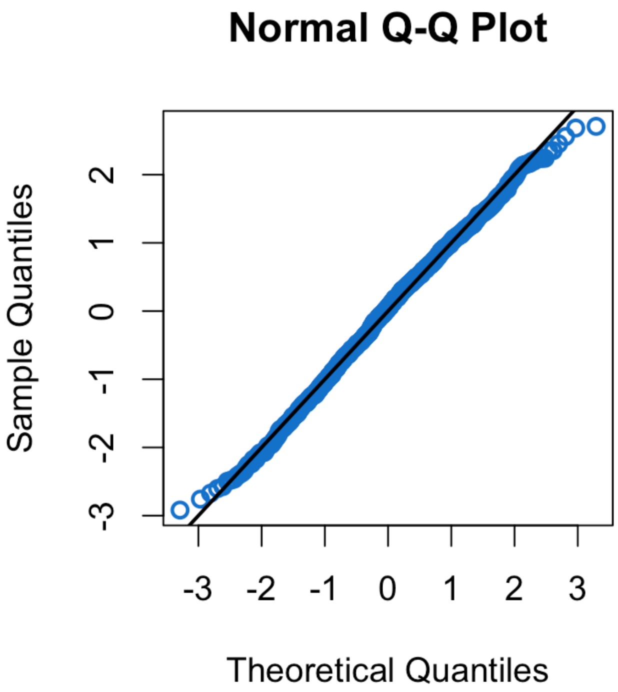
- Population distribution: $X \sim N(10, 2)$.
- Sample distribution: distribution a single sample of size 25; if random, it should be close to the population distribution.
- Sampling distribution of $\bar{x}$: distribution of mean of X from 1000 samples of size 25.

Sampling distribution



- Distribution of the sample is close to the population distribution if the sample is a simple random sample (SRS).
- Sampling distribution of mean $\bar{x}$ is much narrower than the population or sample distribution.
- What kind of distribution do you think it is?
- Unimodal, symmetric, bell-shaped - Normal?
- Let's check.

Sampling distribution



- The distribution of $\bar{x}$ is Normal. In fact, for $X \sim N(10, 2)$, the mean $\bar{x}$ from samples of size 25 has distribution

$$\bar{x} \sim N(10, 0.4)$$

> Let $\bar{x}$ be the mean of an SRS of size n from a population having Normal distribution with mean μ and standard deviation σ. The mean and standard deviation of $\bar{x}$ are $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. And
>
> $$\bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Central Limit Theorem

Draw an SRS of size n from **any population** with mean μ and finite standard deviation σ. When **n is large**, the sampling distribution of the sample mean $\bar{x}$ is approximately Normal:

$$\bar{x} \text{ is approximately } N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

$$\bar{x} \overset{approx.}{\sim} N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

- Central limit theorem holds for ANY population distribution.
- The population distribution is NOT necessarily Normal. As long as it has mean μ and SD σ, the mean $\bar{x}$ from samples of size n has approximately Normal distribution with mean μ and $\sigma/\sqrt{n}$ when n is large.

Central Limit Theorem

Link

Review

- Data structure, variables and relationships
  - *Rows (observations) and columns (variables)*
  - *Response variable and explanatory variable*
  - *Association and causation*
- Distribution
- Normal distribution $X \sim N(\mu, \sigma)$
  - *68-95-99.7 rule* `dnorm()`, `pnorm()`, `qnorm()`
  - *Standard Normal distribution* $Z \sim N(0, 1)$
  - *Normal Q-Q plot* `qqnorm()`
- Population, sample, parameter, statistic
- Random sampling
- Sampling distribution
- Central Limit Theorem (CLT)