STAT 308 – Chapter 3

Background Information

Important Definitions

Statistics: the science and art of collecting, analyzing, and drawing conclusions from data.

Population of Interest: Group of individuals we wish to know more information about

Sample: Subset of the population of interest from which we can obtain information

Individuals: the subjects/objects of the population of interest; can be people, but also business firms, common stocks, or any other object we want to study.

Variable: any characteristic of an individual that we can measure and observe.

Uploading a dataset to R

```
# This is a very common R package that easily creates tidy data easier for
# analysis and make beautiful graphs!
# library(tidyverse)

# This tells R to find all files in the below folder
setwd("C:/Users/mstuart1/OneDrive - Loyola University Chicago/Classes/Fall 2022/STAT 308/Chapter 3")
# Slashes must be either \\ or / for R to read it

# This tells R to upload a text file as a data frame
airfares <- read.csv("../Data/airfares.csv")
head(airfares)
```

```
##   City Fare Distance
## 1    1  360     1463
## 2    2  360     1448
## 3    3  207      681
## 4    4  111      270
## 5    5   93      190
## 6    6  141      393
```

```r
# This tells R to define an object as a variable from the data frame
# R code is data.frame$variable
dist <- airfares$Distance
```

Parameters and Statistics

Population Parameter: A numeric value that describes the characteristics of an entire population

Sample Statistic: A numeric value that describes the characteristics of the observed data from a sample

Recall, we use **sample statistics** to make inference about **population parameters**.

Some important sample statistics:

Sample Mean: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
Sample Variance: $s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
Sample Standard Deviation: $s_x = \sqrt{s_x^2}$

Summary Statistics in R

```r
mean(dist)
```

```
## [1] 816.5294
```

```r
median(dist)
```

```
## [1] 681
```

```r
var(dist)
```

```
## [1] 346679
```

```r
sd(dist)
```

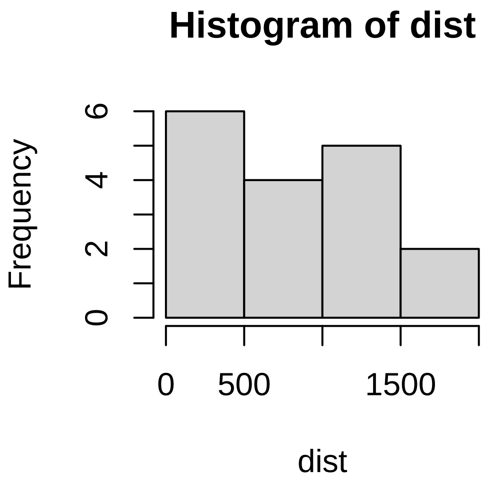
```
## [1] 588.7945
```

```r
summary(dist)
```

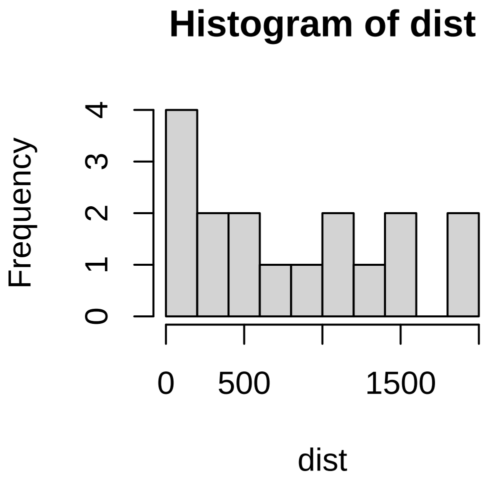
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    90.0   270.0   681.0   816.5  1204.0  1828.0
```

Summary Graphs in R

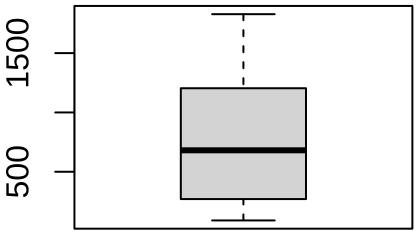
```
# Create a histogram in R
hist(dist)
```



```
# Change the breaks (number of bars) in a histogram in R
hist(dist,breaks=7)
```



```
# Create a boxplot in R
boxplot(dist)
```


1500
500

Random Variables and Distributions

Random Variable: denote a variable whose observed values may be considered outcomes of a stochastic or random experiment. Random variables are typically denoted by a capital letter X, Y, etc., while observations are typically denoted by lowercase letters x, y, etc.

Recall, a data frame contains **observations** from multiple **random variables** from a particular **sample** from the **population of interest**.

Normal Distribution

If a random variable X is normally distributed, this is denoted as

$$X \sim \mathcal{N}(\mu_x, \sigma_x^2)$$

where μ_x is the mean of X and σ_x is the standard deviation of X.

Example

Suppose $X \sim \mathcal{N}(2, 4)$.

a

What is $Pr(X > 3.5)$?

Recall that $Pr(X > 3.5) = 1 - Pr(X \leq 3.5)$

```r
# pnorm calculates a probability less than or equal x with a particular mean
# and standard deviation
1 - pnorm(3.5,mean=2,sd=2)
```

```
## [1] 0.2266274
```

```r
# If you set lower.tail = FALSE, this calculates probability greater than
pnorm(3.5,mean=2,sd=2,lower.tail=FALSE)
```

```
## [1] 0.2266274
```

b

What is the 0.35 quantile/35^{th} percentile of X?

```r
# qnorm a quantile for a given probability with a particular mean
# and standard deviation

qnorm(0.35,mean=2,sd=2)
```

```
## [1] 1.229359
```

Central Limit Theorem

Define $\bar{X}$ as the random variable associated with the mean of a sample $\bar{x}$.

If a random variable X is normally distributed with mean μ_x and standard deviation σ_x OR the sample size n_x is sufficiently large ($n_x > 30$), then the **sampling distribution of the sample mean**,

$$\bar{X} \approx \mathcal{N}\left(\mu_x, \frac{\sigma_x^2}{n_x}\right)$$

or, in other words,

$$\frac{\bar{X} - \mu_x}{\frac{\sigma_x}{\sqrt{n_x}}} \approx \mathcal{N}(0,1).$$

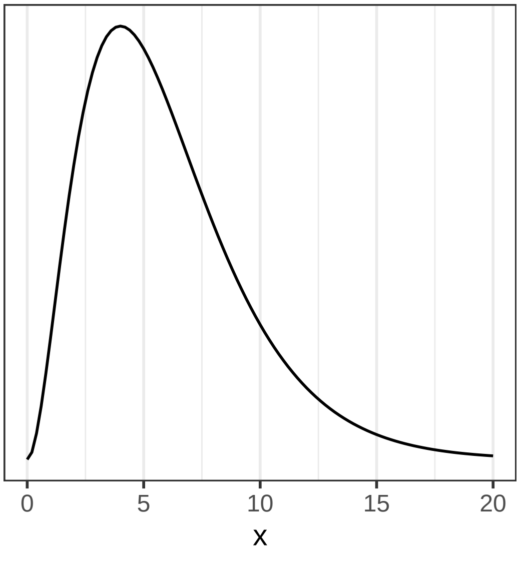
Recall $\frac{\sigma_x}{\sqrt{n_x}}$ is referred to as the standard error. This is an important theorem used in estimation and inference, and will be used throughout the semester.

Chi-squared χ^2 Distribution

Let S_x^2 be the random variable associated with the sample variance s_x^2. The chi-squared distribution can be used to describe the distribution of S_x^2, among other types of random variables. More specifically,

$$\frac{(n_x - 1)S_x^2}{\sigma_x^2} \sim \chi^2_{df=n_x-1}.$$

The chi-squared distribution applies only to positive random variables and is significantly skewed to the right.



Example

Suppose $X \sim {\chi^2}_{df=10}$.

a

What is $Pr(X < 5)$?

```
pchisq(5,df=10)
```

```
## [1] 0.108822
```

b

Find x such that $Pr(X > x) = 0.6$.

```
qchisq(0.6,df=10,lower.tail=FALSE)
```

```
## [1] 8.295472
```

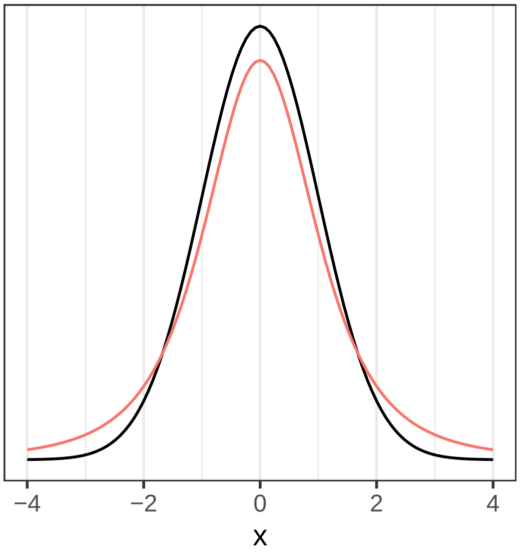
t Distribution

Often times, the population standard deviation σ_x is unknown in the purposes of the sampling distribution. If this is the case, then we can substitute the sample standard deviation s_x for the population standard deviation, σ_x. And, in that case,

$$\frac{\bar{X} - \mu_x}{\frac{s_x}{\sqrt{n_x}}} \sim t_{df=n_x-1}$$

.

Like the normal distribution, the t distribution is also symmetric and unimodal, but has fatter tails to account for the fact that we are using an estimate s_x instead of σ_x.



Example

Suppose $X \sim t_{df=10}$.

a

What is $Pr(X < 2.2)$?

```
pt(2.2,df=10)
```

```
## [1] 0.9737795
```

b

What is the 0.8 quantile/80^{th} percentile of X?

```
qt(0.8,df=10)
```

```
## [1] 0.8790578
```

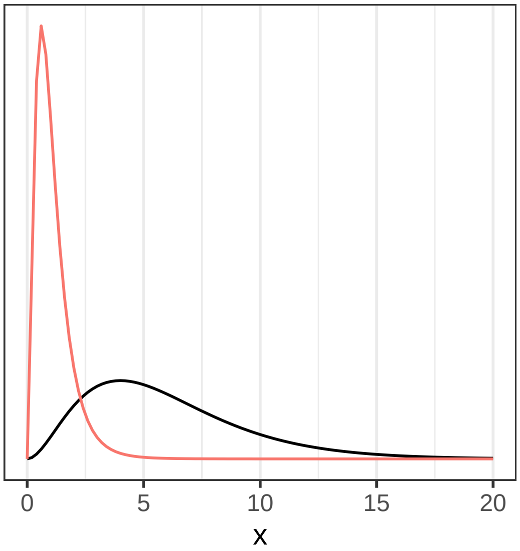
F Distribution

Suppose now we have a new set of data from a random variable Y with population mean μ_y and population variance σ_y^2. Suppose the observed data has a sample mean $\bar{y}$ and sample variance s_y^2. The F distribution is an appropriate distribution for the ratio of the variances of the two random variables. More specifically,

$$\frac{S_x^2/\sigma_x^2}{S_y^2/\sigma_y^2} \sim F_{df1=n_x-1,df2=n_y-1}$$

where $df1$ is denoted as the **numerator degrees of freedom** and $df2$ is denoted as the **denominator degrees of freedom**.

Like the χ^2 distribution, the F distribution is skewed to the right.



Example

Suppose $X \sim F_{df1=6,df2=21}$.

a

What is $Pr(X < 1.3)$?

```
pf(1.3,df1=6,df2=21)
```

```
## [1] 0.6998712
```

b

Find x such that $Pr(X > x) = 0.4$.

```
qf(0.4,df1=6,df2=21,lower.tail=FALSE)
```

```
## [1] 1.090613
```

The F distribution is related to the t distribution because if a random variable $T \sim t_{df=\nu}$, then $T^2 \sim F_{df1=1,df2=\nu}$

Notes for distribution calculations in R

p – calculates probabilty for given quantile q – calculates quantile for a given probabilty lower.tail == FALSE – if using greater than probability

norm – Normal distribution chisq – Chi-squared distribution t – t distribution F – F distribution

Statistical Inference

Estimation

Estimation: The category of statistical inference concerned with quantifying the specific value of a population parameter.

For example, if we have a random sample of data $x_1, x_2, \ldots, x_n$ from a population, we can obtain an estimate of the population mean, μ, by the sample mean $\bar{x}$.

Can we say that $\bar{x}$ equivalent to μ?

NO, different samples produce different sample means.

We need to find a way to quantify the uncertainty of our estimate of the population mean (or other population parameter).

Confidence Interval: A pair of values that provides a range of *plausible* values for the population parameter for a given level of confidence $C = 100 \times (1 - \alpha)$.

Assumptions needed to calculate a confidence interval

- Data comes from a random sample from the population of interest

Confidence intervals take the following general form:

(Parameter Estimate) ± (Critical Value from t-distribution) × (Estimate of Std. Error of Estimate).

A $C\%$ confidence interval for a population mean, μ is written as

$$\bar{x} \pm t_{n-1,1-\frac{\alpha}{2}} \times s_x,$$

where $t_{n-1,1-\frac{\alpha}{2}}$ is the $1 - \frac{\alpha}{2}$ quantile of the t-distribution with $n - 1$ degrees of freedom.

Example

Recall the airfares dataset we previously uploaded into our `R` session. Assume that the data comes from the population of interest, which in this case is all domestic flights out of O'Hare International Airport. Calculate a 95% confidence interval for the mean flight distance.

Interpreting a confidence interval

We are ***C%*** confident that the **true population parameter in the context of the given problem** is between **lower bound with units** and **upper bound with units**.

Go back to the previous example. Interpret the 95% CI in the context of the problem.

Hypothesis testing

Hypothesis Testing: The category of statistical inference concerned with testing whether our estimated value for the population parameter is different enough from the hypothesized value

Procedure for performing a hypothesis test

1. Check that the assumptions needed to perform a hypothesis test are met.
    - Data comes from a random sample from the population of interest.
2. Specifically state the null hypothesis, H_0, and the alternative hypothesis, H_a.
3. Specify the level of significance, α.
4. Calculate the test statistic.
5. Calculate the appropriate p-value for the hypothesis test.
6. Form a decision to either reject H_0 or fail to reject H_0.
7. State your conclusion.

Example

In the airfares dataset, suppose it is believed that the average distance for domestic flights from O'Hare is 1000 miles. Perform a hypothesis test for this belief with $\alpha = 0.05$.

Connection between confidence intervals and hypothesis testing.

CI and HT connection: If a confidence interval and hypothesis test are calculated on the **same observed dataset** where $H_a : \mu \neq \mu_0$ and the same α is used in both calculations, then

$$\mu_0 \text{ is not inside the C\% CI} \Leftrightarrow H_0 \text{ is rejected}$$

and

$$\mu_0 \text{ is inside the C\% CI} \Leftrightarrow H_0 \text{ is not rejected.}$$

Example

Return to the confidence interval and hypothesis test we just conducted. Are these answers compatible?