# Stat 291 - Recitation 10

Orçun Oltulu

31 / 12 / 2021

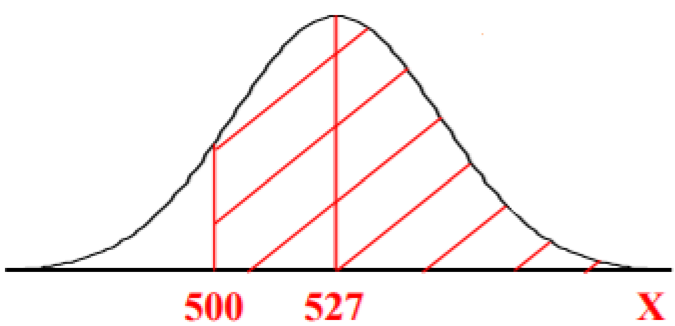
## Statistical Testing and Modeling:

Most graduate schools of business require applicants for admission to take the Graduate Management Admission Council's GMAT examination. Scores on the GMAT are assumed normally distributed with a mean of 527 and a standard deviation of 112.

#### Exercise 1:

#### Part A:

What is the probability of an individual scoring above 500 on the GMAT? $[P(X > 500) = ?]$




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

```
Z <- (500 - 527) / 112
1-pnorm(Z)
```

```
## [1] 0.5952501
```

```
#or
1-pnorm(500, mean = 527, sd = 112)
```

```
## [1] 0.5952501
```

**Exercise 2:**

Write an R function to Construct a Confidence Interval for a Normally Distributed random variable.

**Hint 1:** Your function must take a sample vector and a confidence level then construct C.I.

**Hint 2:**

$$CI = \bar{x} \pm Z_{\frac{\alpha}{2}} * \frac{\sigma}{n}$$

**Hint 3:** qnorm(alpha/2) -> $Z_{\frac{\alpha}{2}}$

```
CI_function <- function(x, conf.level = 0.95){
  alpha <- 1 - conf.level
  lower <- mean(x) - qnorm(1 - alpha/2) * (sd(x) / length(x))
  upper <- mean(x) + qnorm(1 - alpha/2) * (sd(x) / length(x))
  out <- c(lower, upper)
  names(out) <- paste(c("lower ","upper "), "CI ", 1-alpha/2,"%", sep="")
  return(out)
}
```

```
set.seed(291)
mysample <- rnorm(20, mean = 15, sd = 4)

CI_function(mysample)
```

```
## lower CI 0.975% upper CI 0.975%
##        14.77397        15.74292
```

# Hypothesis Testing:

## One sample T-Test:

**Exercise 1:**

Suppose that a researcher collects a sample from METU graduates, records their first year annual salaries and she claims that the mean annual salary for a METU graduate is more than 48k TL.

```
salary <- c(48, 36, 55, 52, 44, 40, 60, 72, 89, 77,
            42, 51, 50, 49, 61, 66, 70, 42, 39, 41)
```

$$H_0: \quad \mu_s \leq 48$$
$$H_A: \quad \mu_s > 48$$

### Part A:

Use t.test() function to test this hypothesis.

```
test_salary <- t.test(salary, mu = 48, alternative = "greater")
test_salary
```

```
##
##  One Sample t-test
##
## data:  salary
## t = 1.9218, df = 19, p-value = 0.03488
## alternative hypothesis: true mean is greater than 48
## 95 percent confidence interval:
##  48.62146      Inf
## sample estimates:
## mean of x
##      54.2
```

### Part B:

Obtain only the p-value for this test and comment on it.

```
test_salary$p.value
```

```
## [1] 0.03488193
```

### Part C:

Now, do the test without using t.test() function, i.e. calculate each step manually, using p_value approach.

$$T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

```
n <- length(salary)
numerator <- mean(salary)-48
denominator <- sd(salary)/sqrt(n)

t_stat <- numerator / denominator

p.val <- 1 - pt(t_stat, df = n-1) # alternative 'greater'

ifelse(p.val < 0.05, "Reject H0", "Cannot Reject H0")
```

```
## [1] "Reject H0"
```

**Exercise 2:**

Suppose that the following data shows the MATH219 midterm grades of 15 students from statistics department.

Test whether mean grades are equal to 50 or not.

```
set.seed(291)
grades <- sample(1:100, replace = T, size = 15)
```

$$H_0: \quad \mu = 50$$

$$H_A: \quad \mu \neq 50$$

```
t.test(grades, mu=50, alternative = "two.sided")
```

```
##
##  One Sample t-test
##
## data:  grades
## t = -2.1678, df = 14, p-value = 0.0479
## alternative hypothesis: true mean is not equal to 50
## 95 percent confidence interval:
##  20.9549 49.8451
## sample estimates:
## mean of x
##      35.4
```

## Two sample T-Test:

**Exercise 3:**

A study was conducted to compare the weights of cats and dogs.

Weights of cats: 31, 20, 21, 35, 13, 21, 10, 17 Weights of dogs: 17, 22, 31, 10, 20, 40

Assume that the population variance to be same for both cats and dogs.

Is there any difference between the weights of cats and dogs?

```
cats <- c(31, 20, 21, 35, 13, 21, 10, 17)
dogs <- c(17, 22, 31, 10, 20, 40)
```

**Part A:** Use t.test() function to test this hypothesis.

```
t.test(cats, dogs, var.equal = T, alternative = "two.sided")
```

```
##
##  Two Sample t-test
##
## data:  cats and dogs
## t = -0.45854, df = 12, p-value = 0.6548
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -13.420369   8.753703
## sample estimates:
## mean of x mean of y
##  21.00000  23.33333
```

**Part B:** Now, do the test without using t.test() function, i.e. calculate each step manually, using critical value approach.

$$T = \frac{(\bar{x_1} - \bar{x_2}) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2) \quad \text{where} \quad S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

```r
sp <- sqrt(
  ((length(cats)-1)*var(cats)+(length(dogs)-1)*var(dogs)) /
    (length(cats)+length(dogs)-2)
  )

t_stat <- (mean(cats)-mean(dogs))/(sp*sqrt( (1/length(cats))+(1/length(dogs))))

t_cv <- qt(0.025,df=length(cats)+length(dogs)-2,lower.tail = TRUE)

ifelse(abs(t_stat) > abs(t_cv), "Rejecet H0", "Cannot Reject H0")
```

```
## [1] "Cannot Reject H0"
```

**Exercise 4:**

Suppose that a researcher claims that the mean sepal widths for 'setosa' species are equal to 'virginica' species. Conduct a t-test and comment on your findings.

$$H_0: \quad \mu_s - \mu_v = 0$$

$$H_A: \quad \mu_s - \mu_v \neq 0$$

```r
library(ISLR)
data(iris)
```

```r
setosa_sepal_width <- iris[iris$Species == "setosa", "Sepal.Width"]
virginica_sepal_width <- iris[iris$Species == "virginica", "Sepal.Width"]

t.test(setosa_sepal_width,virginica_sepal_width, alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  setosa_sepal_width and virginica_sepal_width
## t = 6.4503, df = 95.547, p-value = 4.571e-09
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.3142808 0.5937192
## sample estimates:
## mean of x mean of y
##     3.428     2.974
```

**Exercise 5:**

A study was performed to test whether cars get better mileage on premium gas than on regular gas. Each of 10 cars was first filled with either regular or premium gas, decided by a coin toss, and the mileage for that tank was recorded. The mileage was recorded again for the same cars using the other kind of gasoline. Use a paired t-test to determine whether cars get significantly better mileage with premium gas.

$$H_0: \quad \mu_{prep} \leq \mu_{reg}$$

$$H_A: \quad \mu_{prep} > \mu_{reg}$$

```r
reg = c(16, 20, 21, 22, 23, 22, 27, 25, 27, 28)
prem = c(19, 22, 24, 24, 25, 25, 26, 26, 28, 32)
```

```r
t.test(prem, reg, alternative="greater", paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  prem and reg
## t = 4.4721, df = 9, p-value = 0.0007749
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
```

```
##  1.180207      Inf
## sample estimates:
## mean of the differences
##                       2
```