STAT011 Statistical Methods I

Lecture 15 One-Sample *t* Procedures

Lu Chen
Swarthmore College
3/21/2019

Review - Statistical inference

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

- Level C **confidence interval** for population mean μ: $\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$
- Level α z *test* for a population mean μ:
  - $H_0 : \mu = \mu_0$; $H_a : \mu > \mu_0$ or $\mu < \mu_0$ or $\mu \neq \mu_0$
  - $z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} \stackrel{approx.}{\sim} N(0, 1)$
  - P-value is computed based on H_a
  - $P \leq \alpha$, reject H_0; $P > \alpha$, fail to reject H_0.
- For both, we assume unknown population mean μ and known population standard deviation σ.
- What if σ is unknown?

Outline

- Sample standard deviation (SD)
- Degree of freedom
- Standard error (SE)
- t distribution
- One-sample t procedures: statistical inference for a population mean based on t distribution
  - One-sample t confidence interval
  - One-sample t test
- Examples

Sample standard deviation (SD)

- When population standard devation σ is unknown, we use the sample standard deviation s to estimate σ.

$$s = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1}}$$

- σ is a population parameter; s is a sample statistic.
- Why $n - 1$?

Sample standard deviation (SD)

- Ultimately, we want s to be an **unbiased estimator** of σ.
- Let's use simulation to compare three possible ways of calculating sample standard deviation:

1. $$s_1 = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1}}$$

2. $$s_2 = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n}}$$

3. $$s_3 = \sqrt{\frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_n - \mu)^2}{n}}$$

- **Note**: s_3 is the formula for calculating SD when population mean μ is known.

Sample standard deviation (SD)

```r
set.seed(10)
n <- 25; s1 <- s2 <- s3 <- NULL
for(i in 1:1000){
  x <- rnorm(n) # mu = 0, sigma = 1
  s1[i] <- sd(x) # sqrt(sum((x-mean(x))^2)/(n-1))
  s2[i] <- sqrt(sum((x-mean(x))^2)/n)
  s3[i] <- sqrt(sum((x-0)^2)/n) # mu=0
}
mean(s1)
```

```
## [1] 0.9957356
```

```r
mean(s2)
```

```
## [1] 0.9756177
```

```r
mean(s3)
```

```
## [1] 0.9965334
```

Sample standard deviation (SD)

1. $$s_1 = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1}}$$

2. $$s_2 = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n}}$$

3. $$s_3 = \sqrt{\frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_n - \mu)^2}{n}}$$

- By simulation, mean of s_3 is the closest to the true population SD $\sigma = 1$.
- In reality, since we do not know population mean μ, we cannot apply the formula s_3. We use sample mean $\bar{x}$, which is an unbiased estimator of μ, to compute the sample SD.
- Using $\bar{x}$ brings more uncertainness (μ is fixed and $\bar{x}$ changes from sample to sample) into the estimation. s_2 turns out to be a biased estimator of σ.

Sample standard deviation (SD)

1. $$s_1 = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1}}$$

2. $$s_2 = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n}}$$

3. $$s_3 = \sqrt{\frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_n - \mu)^2}{n}}$$

- SD measures the variability of n **random values** $x_1, x_2, \cdots, x_n$. However, once $\bar{x}$ is used (s_2), knowing $x_1, x_2, \cdots, x_{n-1}$ and $\bar{x}$, we will know x_n for sure. It measures the variability of only $n-1$ **random values** that are free to vary.
- Therefore, in the formula of sample standard deviation (s_1), the denominator is $n-1$, which results in an unbiased estimator of σ.

Degree of freedom

> **Degree of freedom** is the number of values in the final calculation of a statistic that are **free to vary**.

- It is calculated as the difference between
  - Number of independent values that go into the estimate: n
  - Number of statistics used as intermediate steps: 1
- For

$$s = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1}}$$

The degree of freedom for sample SD s is $n - 1$.

Standard error (SE)

By CLT,

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

- When population SD σ is unknown, we use sample SD s to replace it.
- The SD of $\bar{x}$ becomes

$$\frac{s}{\sqrt{n}}$$

- This is called the **standard error (SE)** of $\bar{x}$.

Standard error (SE)

> When the standard deviation of a **statistic** is **estimated from the data**, the result is called the **standard error (SE)** of the statistic.

- Population SD of a variable: σ
- Sample SD of a variable: s
- SD of $\bar{x}$ (when σ is known):

$$\frac{\sigma}{\sqrt{n}}$$

- SE of $\bar{x}$ (when σ is unknown):

$$\frac{s}{\sqrt{n}}$$

Distribution of sample mean

When σ is known,

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

And by standardization of Normal distribution,

$$z = \frac{\bar{x} - \mu}{\sigma \big/ \sqrt{n}} \overset{approx}{\sim} N(0, 1)$$

When σ is unknown and estimated by s,

$$t = \frac{\bar{x} - \mu}{s \big/ \sqrt{n}} \overset{approx}{\sim} t(n-1)$$

The t distribution

Suppose that an SRS of size n is drawn from an $N(\mu, \sigma)$ population. Then the one-sample t statistic

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \sim t(n-1)$$

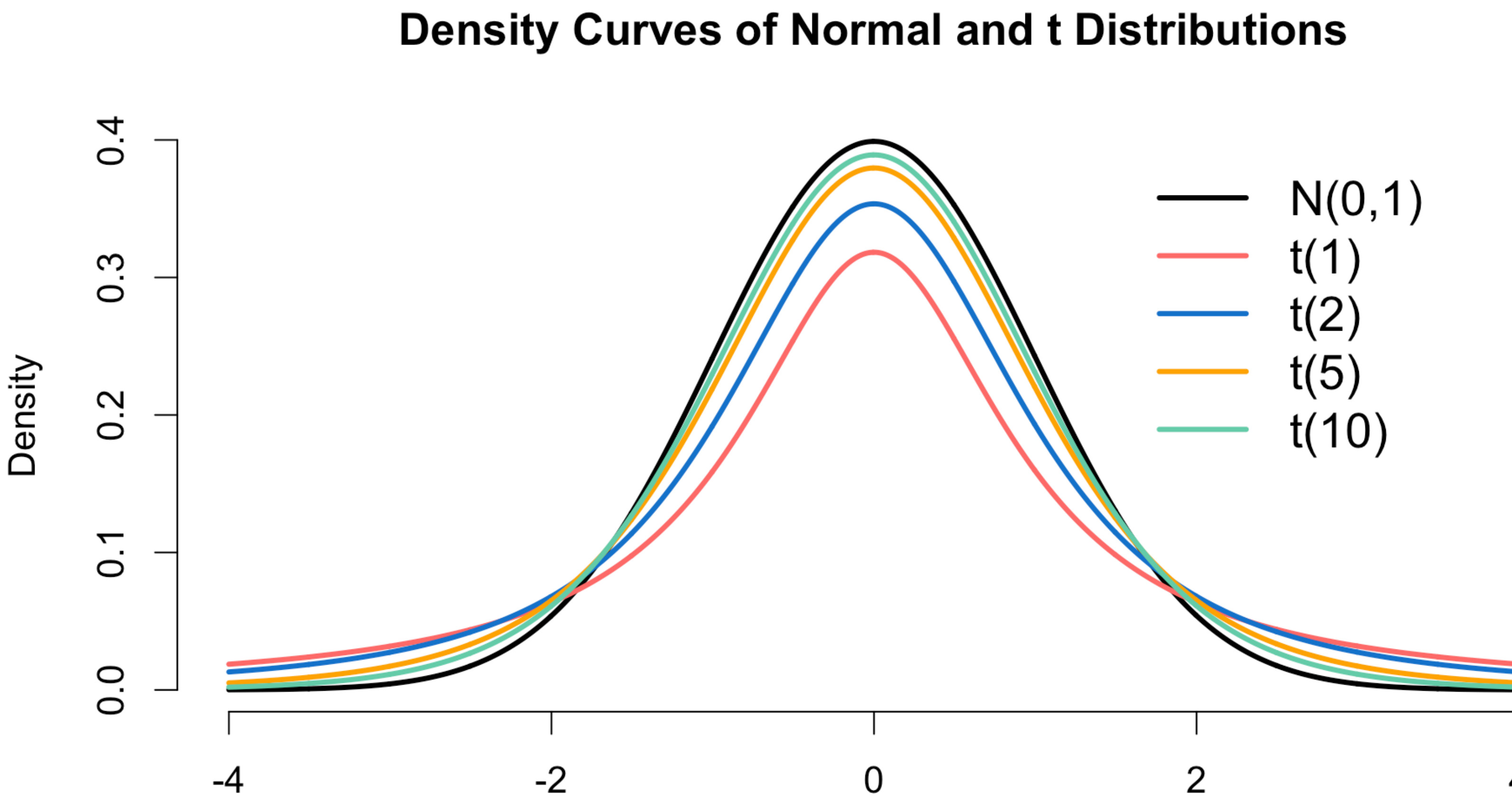
has the t **distribution** with $n-1$ **degrees of freedom**.

When the population distribution is not Normal,

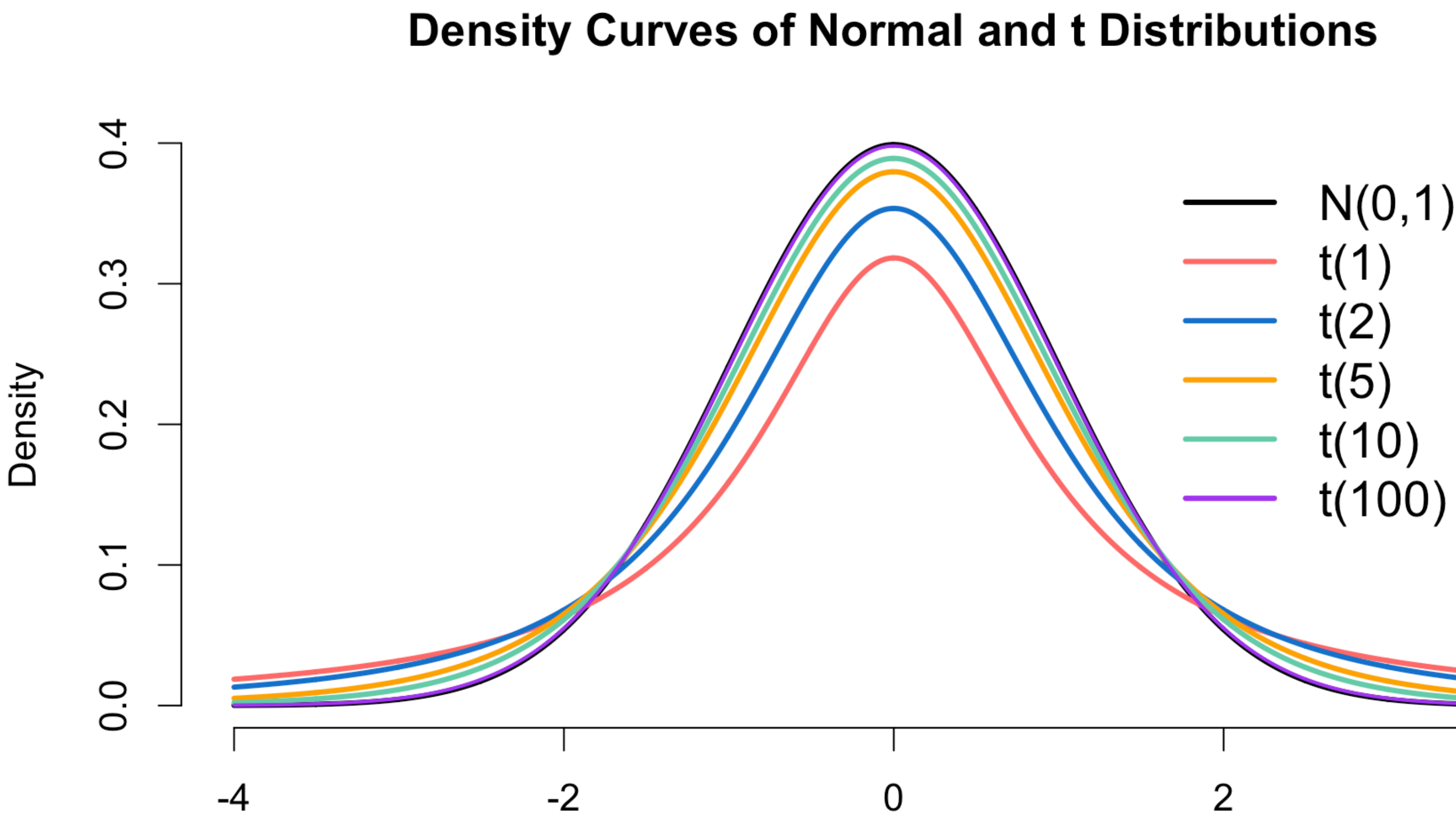
$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \overset{approx}{\sim} t(n-1)$$

has an approximate t **distribution** with $n-1$ **degrees of freedom**.

The t distribution



The *t* distribution



The t distribution

- Symmetric, unimodal, bell-shaped.
- Approximates the Normal distribution when n is large.
- Has heavier tails than the Normal distribution
  - Using s instead of σ introduces more variability to $\frac{\bar{x}-\mu}{s/\sqrt{n}}$
  - Using t distribution results in wider C.I. and larger P-value than Normal dsitribution.
  - We are less sure about the inference of population mean when population SD is unknown.

t distribution in R

```r
# dnorm() and dt( , df = n-1)
dnorm(0); dt(0, df=5); dt(0, df=100)
```

```
## [1] 0.3989423

## [1] 0.3796067

## [1] 0.3979462
```

```r
# pnorm() and pt( , df = n-1)
pnorm(0); pt(0, df=5); pt(0, df=100)
```

```
## [1] 0.5

## [1] 0.5

## [1] 0.5
```

t distribution in R

```r
# pnorm() and pt( , df = n-1)
pnorm(-1.96); pt(-1.96, df=5); pt(-1.96, df=100)
```

```
## [1] 0.0249979

## [1] 0.05364398

## [1] 0.02638945
```

```r
# qnorm() and qt( , df = n-1)
qnorm(0.975); qt(0.975, df=5); qt(0.975, df=100)
```

```
## [1] 1.959964

## [1] 2.570582

## [1] 1.983972
```

One-sample t confidence interval

Suppose that an SRS of size n is drawn from a population having unknown mean μ. A **level C confidence interval** for μ is

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$$

where t^* is the value for the $t(n-1)$ density curve with area C between $-t^*$ and t^*. The quantity

$$t^* \frac{s}{\sqrt{n}}$$

is the **margin of error**. The confidence level is exactly C when the population distribution is Normal and is approximately correct for large n in other cases.

One-sample t test

Suppose that an SRS of size n is drawn from a population having unknown mean μ. To test the hypothesis $H_0 : \mu = \mu_0$, compute the **one-sample t statistic**

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

In terms of a random variable T having the $t(n-1)$ distribution, the P-value for a test of H_0 against

$$\begin{aligned} H_a : \mu > \mu_0 &\quad \text{is } P(T \geq t) \\ H_a : \mu < \mu_0 &\quad \text{is } P(T \leq t) \\ H_a : \mu \neq \mu_0 &\quad \text{is } 2P(T \geq |t|) \end{aligned}$$

These P-values are exact if the population distribution is Normal and are approximately correct for large n in other cases.

Guidelines for one-sample t procedures

For sample size n,

- $n < 15$: Use t procedures if the data are close to Normal. If the data are clearly non-Normal or if outliers are present, do not use t.
- $15 \leq n < 40$: The t procedures can be used except in the presence of outliers or strong skewness.
- $n \geq 40$: The t procedures can be used even for clearly skewed distributions when the sample is large.

The t procedures are quite **robust**.

- A statistical inference procedure is called **robust** if it is insensitive to violations of the assumptions made.

Comparing z and t procedures

	z **procedures**	t **procedures**
Population SD σ	Known	Unknown, use sample SD s
Level C **C.I.**	$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$ $z^* =$ `qnorm(1-(1-C)/2)`	$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$ $t^* =$ `qt(1-(1-C)/2, df=n-1)`
Level α **significance test**	$H_0 : \mu = \mu_0$ $H_a : \mu > \mu_0, \mu < \mu_0, \mu \neq \mu_0$ $z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} \overset{approx.}{\sim} N(0, 1)$ $P(Z \leq z)$, `pnorm(z)` $P(Z \geq z)$, `1-pnorm(z)` $2P(Z \geq \|z\|)$,`2*(1-pnorm(abs(z)))`	$H_0 : \mu = \mu_0$ $H_a : \mu > \mu_0, \mu < \mu_0, \mu \neq \mu_0$ $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} \overset{approx.}{\sim} t(n-1)$ $P(T \leq t)$, `pt(t,df=n-1)` $P(T \geq t)$, `1-pt(t,df=n-1)` $2P(T \geq \|t\|)$, `2*(1-pt(abs(t),df=n-1))`

Example 1

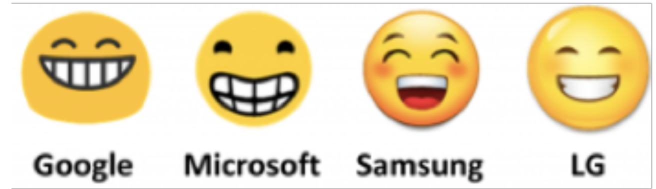
Within-platform score of mis-communication (25 emoji for each platform)

	Apple	Google	Microsoft	Samsung	LG
Top 3	3.64	3.26	4.40	3.69	2.59
	3.50	2.66	2.94	2.36	2.53
	2.72	2.61	2.35	2.29	2.51
•••	•••				
Bottom 3	1.25	1.13	1.12	1.23	1.30
	0.65	1.06	1.08	1.09	1.26
	0.45	0.62	0.66	1.08	0.63

Google, MS, Samsung and LG together

- Average score of mis-communication: 1.84
- Number of emoji's: 100
- Population standard deviation: 0.50
  - In fact, this is sample SD.

Example 1



Assume population SD is known.

$\bar{x} = 1.84$, $\sigma = 0.5$, $n = 100$, $C = 0.95$

95% confidence interval

- $C = 0.95$, $z^* = 1.96$ `qnorm(0.975)`
- Margin of error

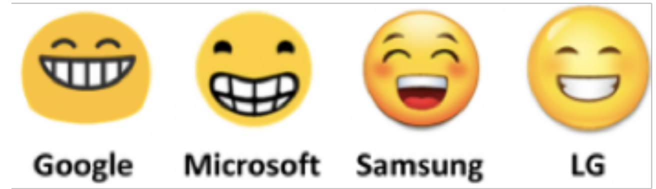
$$m = z^* \frac{\sigma}{\sqrt{n}} = 1.96 \frac{0.5}{\sqrt{100}} = 0.098$$

- 95% confidence interval $\bar{x} \pm m = 1.84 \pm 0.098$
- We are 95% confident (about the method) that the population mean score of mis-communication for the four platforms will be within [1.742, 1.938]

Example 1

Population SD is in fact unknown.

$\bar{x} = 1.84$, $s = 0.5$, $n = 100$, $C = 0.95$



95% confidence interval

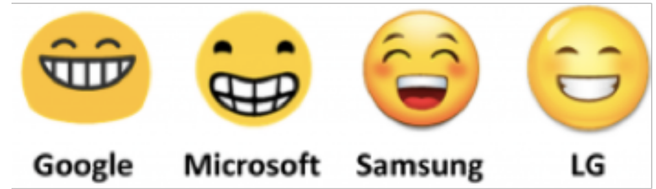
- $C = 0.95$, $t^* = 1.98$ `qt(0.975, df = 99)`
- Margin of error

$$m = t^* \frac{s}{\sqrt{n}} = 1.98 \frac{0.5}{\sqrt{100}} = 0.099$$

- 95% confidence interval $\bar{x} \pm m = 1.84 \pm 0.099$
- We are 95% confident (about the method) that the population mean score of mis-communication for the four platforms will be within [1.741, 1.939]

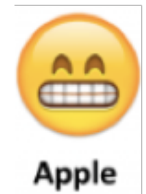
Example 1

Is the average score of mis-communication of the four from 2, which is the mean score of Apple emoji?



Assume population SD is known.

VS.

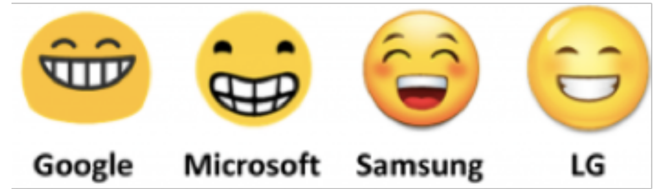


$\bar{x} = 1.84, \sigma = 0.5, n = 100, \alpha = 0.05$

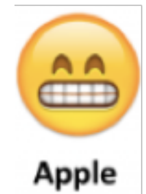
- $H_0 : \mu = 2; H_a : \mu \neq 2$
- $$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{1.84 - 2}{0.5/\sqrt{100}} = \frac{-0.12}{0.05} = -3.2$$
- $2P(Z \geq |z|) = 2P(Z \geq 3.2) = 0.0014 < 0.05$ `2*(1-pnorm(3.2))`
- The test is significant at level 0.05 and we reject H_0. The mean score of mis-communication of the four platforms is significantly different from 2.

Example 1

Is the average score of mis-communication of the four from 2, which is the mean score of Apple emoji?



VS.



Population SD is in fact unknown.

$\bar{x} = 1.84$, $s = 0.5$, $n = 100$, $\alpha = 0.05$

- $H_0 : \mu = 2; H_a : \mu \neq 2$
- $$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{1.84 - 2}{0.5/\sqrt{100}} = \frac{-0.12}{0.05} = -3.2$$
- $2P(T \geq |t|) = 2P(T \geq 3.2) = 0.0018 < 0.05$ `2*(1-pt(3.2, df=99))`
- The test is significant at level 0.05 and we reject H_0. The mean score of mis-communication of the four platforms is significantly different from 2.

Example 2

The mean percentage of dialogue spoken by men for the 62 screenplays in 2015 is 0.668. The SD of the 62 screenplays is 0.241.

95% confidence interval

- $\bar{x} \pm t^* \frac{s}{\sqrt{n}} = 0.668 \pm 2.00 \times \frac{0.241}{\sqrt{62}} = 0.668 \pm 0.061$
  `t* = qt(0.975, df = 61) = 1.999624`
  We are 95% confident (about the method) that the population mean percentage of dialogue spoken by men is within [0.607, 0.729]

Level 0.05 significance test whether population mean greater than 0.5

- $H_0 : \mu = 0.5, H_a : \mu > 0.5.\ t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} = \frac{0.668-0.5}{0.241/\sqrt{62}} = 5.49.$
  $t > t^* = 1.7$ `qt(0.95,df=61)` or $P = 4 \times 10^{-7} < 0.05$ `1-pt(5.49,df=61)`
  The test is highly significant at level 0.05. We reject H_0 and conclude that the population mean percentage of dialogue spoken by men is significantly greater than 0.5.

One-sample *t* procedures in R

```r
percent_men # 62 percentage values of dialogue spoken by men for 2015 movies
```

```
##  [1] 0.98457660 0.97053407 0.32418830 0.27857449 0.77425697 0.84956568
##  [7] 0.53103976 0.10586256 0.50441158 0.25436772 0.35793946 0.73917869
## [13] 0.83809736 0.15171331 0.89037260 0.38880671 1.00000000 0.76046885
## [19] 0.39227316 0.28032892 0.91113709 0.54406303 0.84768212 1.00000000
## [25] 0.78186381 0.62622438 0.56980907 0.78833910 0.72210815 0.79612088
## [31] 0.82716454 0.64336662 0.89454643 0.80753437 0.50219759 0.63798364
## [37] 0.25218825 0.74342258 0.79141282 0.93589744 0.80880134 0.68960030
## [43] 0.81827042 0.89763325 0.46960452 0.59691068 0.80664427 0.49614112
## [49] 0.53015726 0.83555121 0.93586918 0.76235198 0.72157216 0.45707300
## [55] 0.65388303 0.80549821 1.00000000 0.90663453 0.74396939 0.71404924
## [61] 0.71968288 0.03445006
```

One-sample *t* procedures in R

```r
t.test(percent_men, conf.level = 0.95) # 95% confidence interval
```

```
##
##  One Sample t-test
##
## data:  percent_men
## t = 21.841, df = 61, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.6066667 0.7289451
## sample estimates:
## mean of x
## 0.6678059
```

- The 95% confidence interval for the true mean percentage of dialogue spoken by men in 2015 movies is [0.607, 0.729].
- Here the `t.test()` function automatically runs a two-sided test, but we are interested in a one-side test.

One-sample *t* procedures in R

```r
t.test(percent_men, alternative = "greater", mu = 0.5)
```

```
##                       ## alternative = "greater", "less" or "two.sided"
##  One Sample t-test
##
## data:  percent_men
## t = 5.4883, df = 61, p-value = 4.15e-07
## alternative hypothesis: true mean is greater than 0.5
## 95 percent confidence interval:
##  0.6167384       Inf
## sample estimates:
## mean of x
## 0.6678059
```

- $H_0 : \mu = 0.5$; $H_a : \mu > 0.5$; $t = 5.49$; $P = 4.15 \times 10^{-7} < 0.05$. We reject H_0 at level 0.05. Men speak significantly more dialogue than women in 2015 movies.
- When `t.test()` function is run for a one-sided test, it generates a "one-sided" confidence interval at the same time, which is NOT the correct confidence interval - so ignore it.

One-sample *t* procedures in R

```r
t.test(percent_men, alternative = "two.sided", mu = 0.5)
```

```
##                         ## alternative = "greater", "less" or "two.sided"
##  One Sample t-test
##
## data:  percent_men
## t = 5.4883, df = 61, p-value = 8.299e-07
## alternative hypothesis: true mean is not equal to 0.5
## 95 percent confidence interval:
##  0.6066667 0.7289451
## sample estimates:
## mean of x
## 0.6678059
```

- $H_0 : \mu = 0.5$; $H_a : \mu > 0.5$; $t = 5.49$; $P = 4.15 \times 10^{-7} < 0.05$. We reject H_0 at level 0.05. Men speak significantly more dialogue than women in 2015 movies. The 95% confidence interval for μ is $[0.607, 0.729]$.
- When `t.test()` function is run for a two-sided test, it gives the results for the test as well as the confidence interval.

About homework

- Some questions may ask you to calculate the confidence interval and conduct a t test "**using R**". You should use the `t.test()` function to do the analysis and write down the four steps of the test and report the CI as in Slide 30~32.
- Some other questions may ask you to do the analysis "**by hand**". Then you should apply the formulas in the definitions of the confidence interval and the test and write everything down in math mode. You may still use R as a calculator and to compute t^* values and P-values.
- This guidance applies to all the subsequent problem sets (Homework 6 to 10).

Summary

- Sample standard deviation (SD)
- Degree of freedom
- Standard error (SE)
  - *SD of a statistic estimated from sample data*
- t distribution `dt( , df = )`, `pt( , df = )`, `qt( , df = )`
- Statistical inference for a population mean based on t distribution
  - One-sample t confidence interval $\bar{x} \pm t^* \frac{s}{\sqrt{n}}$
  - One-sample t test $H_0 : \mu = \mu_0, t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} \overset{approx.}{\sim} t(n-1)$
- Examples

  `t.test( , conf.level = )`, `t.test( , alternative = , mu = )`