PSY 3307

One Sample t-test

Jonathan A. Pedroza, MS, MA

Cal Poly Pomona

2021-09-28

Agenda

- Understanding the One-Sample t-test
- Performing a One-Sample t-test
- Interpreting a One-Sample t-test
- Estimating Mu by Computing Confidence Intervals
- Reporting a One-Sample t-test

What is a One-Sample t-test

- It's pretty similar to a z-test
  - t-test used more often in behavioral research
- z-test requires we know population standard deviation
  - often not possible in behavioral research
- uses unbiased estimators (N - 1 formulas)
- computes something like the z-score for our sample mean
  - t-score

One-Sample t-test

- parametric test for when the population standard deviation is unknown
- still compares the sample mean to the population mean

Steps to One-Sample t-test

1. Statistical Hypotheses
    - what is the population mean and is your sample mean different from that population mean
    - H0: sample mean equals the population mean
    - H1: sample mean is different from the population mean
2. Select an alpha
3. Check assumptions
    - Outcome needs to be continuous (interval or ratio scale)
    - Population score forms a normal distribution
    - variability of raw score population is estimated from the sample

Steps to One-Sample t-test

- All we need to know is the t critical value and if the t obtained value is within the regions of rejection

Steps to a z-test/One-Sample t-test

- get population variance/standard deviation (z-test)
- get estimated variance/standard deviation (t-test)
- get the standard error (SE) of the mean (z-test)
- get the **estimated** SE (t-test)
- calculate the score by subtracting the population mean from the sample mean and dividing by the SE
  - either obtained z or t value

Changes between the z-test and t-test

$$s_x^2 = \frac{\Sigma X^2 - \frac{(\Sigma X)^2}{N}}{N - 1}$$

z-test	t-test
$\sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{N}}$	$s_{\overline{X}} = \frac{s_X}{\sqrt{N}}$
`$$`	
$z_{obt} = \frac{\overline{X} - \mu}{\sigma_{\overline{X}}}$	$t_{obt} = \frac{\overline{X} - \mu}{s_{\overline{X}}}$

Small Change in Formulas

- SE calculation will start to look slightly different as it will use the variance squared

- Due to future formulas using slightly different notation, we will adopt that for our SE

$$s_{\overline{X}} = \frac{\sqrt{s_X^2}}{\sqrt{N}}$$

$$s_{\overline{X}} = \sqrt{\frac{s_X^2}{N}}$$

Example

```r
set.seed(092221)

numbers = rnorm(10, mean = 5, sd = 1.2)

numbers
```

```
##  [1] 4.770670 5.271329 6.899812 5.598090 4.219022 6.828034 6.046497 2.92124
##  [9] 4.993870 5.948126
```

1. Calculate the Variance
2. Calculate the SE
3. Compute t

```r
# population mean is 10

4.770670 + 5.271329 + 6.899812 + 5.598090 + 4.219022 + 6.828034 + 6.0464
```

```
## [1] 53.4967
```

```r
# 53.50

4.770670^2 + 5.271329^2 + 6.899812^2 + 5.598090^2 + 4.219022^2 + 6.82803
```

```
## [1] 299.3272
```

```r
# 299.33

53.50^2
```

```
## [1] 2862.25
```

```r
# 2862.25
```

```r
2862.25/10
```

```
## [1] 286.225
```

```r
# 286.23

299.33 - 286.23
```

```
## [1] 13.1
```

```r
# 13.1

13.1/9
```

```
## [1] 1.455556
```

```r
# 1.46 variance
```

```r
# se is...
1.46/10
```

```
## [1] 0.146
```

```r
sqrt(1.46/10)
```

```
## [1] 0.3820995
```

```r
sqrt(.146)
```

```
## [1] 0.3820995
```

```r
# compute t
(5.35 - 10)/.38
```

```
## [1] -12.23684
```

```r
# t value of -12.24
```

```r
53.50/10
```

```
## [1] 5.35
```

```r
# 5.35 going to need the mean later
```

t-distribution & Degrees of Freedom (df)

- we will now be working with the t-distribution
  - this also means we'll be working with a t-table
- **t-distribution** is the sampling distribution of all values of t when samples of a particular size (differing N size) are selected from the raw score population in the null hypothesis

t-distribution & Degrees of Freedom (df)

- higher values on the t-distribution are to the right of the population mean, lower values to the left of the population mean
- t-tests also have regions of rejection
- doesn't always represent a perfectly normal distribution
  - dependent on N value
  - larger the sample the more normal the distribution looks
- the different shapes are important because our regions of rejection will look different dependent on the sample size

t-distribution & Degrees of Freedom (df)

- the distribution changes based on the sample size, which then means that the 5% of the regions of rejection and critical value change
- remember to be conservative about estimating variance and SD, we have been using `N - 1`
- the name of that is the **degrees of freedom** or df
  - number of scores in a sample that reflect the variability in the population
  - determines shape of sampling distribution when estimating standard deviation for the population

t-distribution & Degrees of Freedom (df)

- since the df is the sample size - 1, the larger the df, the closer to resembling a normal distribution our data becomes
  - df of 120+ is the same as a z-distribution

Using the t-table

- the t-table is different from the z-table
- has df, \alpha = .05 and \alpha = .01
    - this is dependent on our sample size - 1, and what our alpha is `a priori`

t-table

- we need to figure out our `t critical value`
- we need our sample size, and a decision on what alpha we want to use (.05 or .01)
- since not all df are listed, if your df is between two values, a statistically significant finding is a t-value larger than the larger df and smaller than the smaller df

Examples

- sample size = 200
  - alpha = .05
- sample size = 90
  - alpha = .05
- sample size = 37
  - alpha = .01

t-test Interpretation

- If a statistically significant finding is found
  - your sample is significantly different from the population in whatever the outcome was

One-tailed test

- if you know if your sample will do better or worse than the population, you'd use a one-tailed test
- Example: you know that your sample will get higher grades than the population

Confidence Intervals

- **point estimation** is a way to estimate a point where you think the population's outcome value will be
    - this is why we can't say we're certain mu is a specific number and have to say *around* that number
- **interval estimation** is when we state that mu will fall within a range of values
    - margin of error, such as getting an exam and stating that the average test score was 84 plus or minus 3 points
    - due to sampling error
- **confidence intervals** are a range of values which we are certain our value falls within
    - when we say *around* a value, we are saying that we got one value but we are certain it is within a range of values
    - around 84 points on an exam, but we are certain the correct value is between 80 and 87

Confidence Intervals

- We're choosing a range of values that are not significantly different from our sample mean
- we compute confidence intervals after we have a statistically significant finding
- It is often stated as:
    - We got a statistically significant finding where our sample scored ***points compared to the population's score*** ; t(df) = t-value, p-value
    - Example: $t(31) = 4.7$, $p = .037$

```r
coffee <- read_csv('https://raw.githubusercontent.com/rfordatascience/t
  mutate(species = as.factor(species),
         process = recode(processing_method, "Washed / Wet" = "washed",
                          "Semi-washed / Semi-pulped" = "not_washed",
                          "Pulped natural / honey" = "not_washed",
                          "Other" = "not_washed",
                          "Natural / Dry" = "not_washed",
                          "NA" = NA_character_),
         process = as.factor(process),
         species = as.factor(species),
         country_of_origin = as.factor(country_of_origin),
         variety = as.factor(variety)) %>%
  drop_na(process, color)
```

```
##
## -- Column specification --------------------------------------------
## cols(
##   .default = col_character(),
##   total_cup_points = col_double(),
##   number_of_bags = col_double(),
##   aroma = col_double(),
##   flavor = col_double(),
##   aftertaste = col_double(),
##   acidity = col_double(),
##   body = col_double(),
##   balance = col_double(),
##   uniformity = col_double(),
```

```r
psych::describe(coffee$total_cup_points, na.rm = TRUE)
```

```
##    vars    n  mean   sd median trimmed  mad   min   max range  skew kurtosi
## X1    1 1071 82.03 2.67  82.42    82.3 1.85 59.83 90.58 30.75 -2.11    10.5
##      se
## X1 0.08
```

```r
# mean is 82.03
# SE is .08
# sample size is 1071
```

```r
t.test(coffee$total_cup_points, mu = 85) #conf int only works for two ta
```

```
##
##      One Sample t-test
##
## data:  coffee$total_cup_points
## t = -36.39, df = 1070, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 85
## 95 percent confidence interval:
##  81.87385 82.19373
## sample estimates:
## mean of x
##  82.03379
```

```r
t.test(coffee$total_cup_points, mu = 85, alternative = "less")
```

```
##
##  One Sample t-test
##
## data:  coffee$total_cup_points
## t = -36.39, df = 1070, p-value < 2.2e-16
## alternative hypothesis: true mean is less than 85
## 95 percent confidence interval:
##      -Inf 82.16798
## sample estimates:
## mean of x
##  82.03379
```

```r
t.test(coffee$total_cup_points, mu = 85, alternative = "greater")
```

```
## 
##  One Sample t-test
## 
## data:  coffee$total_cup_points
## t = -36.39, df = 1070, p-value = 1
## alternative hypothesis: true mean is greater than 85
## 95 percent confidence interval:
##  81.8996     Inf
## sample estimates:
## mean of x 
##  82.03379
```

Confidence Interval Calculations

$$ (s{\overline{x}})(-t{crit}) + \overline{X} \; \leq \; \mu \; \leq \; (s{\overline{x}})(t{crit}) + \overline{X} $$

```r
# t critical value is 1.96 since we have such a large sample and df

# mu = 85
# sample mean = 82.03
# SE = .08
# df = 1070

# lower
.08*-1.96 + 82.03
```

```
## [1] 81.8732
```

```r
# 81.8732

# higher
.08*1.96 + 82.03
```

```
## [1] 82.1868
```

```r
t.test(coffee$total_cup_points, mu = 85)
```

```
##
##     One Sample t-test
##
## data:  coffee$total_cup_points
## t = -36.39, df = 1070, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 85
## 95 percent confidence interval:
##  81.87385 82.19373
## sample estimates:
## mean of x
##  82.03379
```

$t(1070) = -36.39$, $p < .05$, 95% CI [81.87, 82.19]

Our one-sample t-test comparing a sample of coffee ratings ($M = 82.03$, $SD = 2.67$) to the population of coffee ratings ($M = 85$) showed evidence of a statistically significant difference. Specifically, the sample's average coffee rating was significantly lower than the population's average coffee rating; $t(1070) = -36.39$, $p < .05$, 95% CI [81.87, 82.19]. We are 95% certain that the actual sample mean is between 81.87 and 82.19.