Stat. Inf. II: Assignment 1

Naimul Chowdhury

Reading:

a. Read the remaining Sections in Chapter 9: 9.3, 9.5 and 9.6. And, read the final section about errors and power on the R handout for inference about a proportion.

Problem #1

Prelude.

We know that, for $x_1, x_2, \ldots, x_n \sim N(\mu, \sigma^2)$, the following is true:

$$\bar{X} \sim N(\mu, \sigma/\sqrt{n}),$$

which leads to

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

Typically, we don't have access to population standard deviation σ, hence we substitute it by sample standard deviation $s = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n-1}}$. Then, the test statistic becomes

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1},$$

where t-distribution is

- like $N(0,1)$, symmetric, bell-shaped and centered at 0,
- but has heavier tails (see https://istats.shinyapps.io/tdist/ for demo).

By **mathematical definition**, random variable T has a t-distribution with df degrees of freedom if

$$T = \frac{Z}{\sqrt{X^2_{df}/df}} \sim t_{df},$$

where $Z \sim N(0,1)$, $X^2_{df} \sim \chi^2_{df}$ (for definition of χ^2_{df}, please see the https://en.wikipedia.org/wiki/Chi-squared_distribution or the last set of slides from previous semester).

By **Cochran's theorem**, the following is true:

$$\mathbf{z}' A \mathbf{z} \sim \chi^2_{rank(A)},$$

where

- $\mathbf{z} = (z_1, z_2, \ldots, z_n)'$, with $z_1, z_2, \ldots, z_n$ being independent standard normal random variables (in short, $z_1, z_2, \ldots, z_n \sim_{ind} N(0,1)$).
- A is a symmetric, idempotent matrix,

Actual problem. Piece-by-piece, we will proceed to show that, for $x_1, \ldots, x_n \sim N(\mu, \sigma^2)$, the test statistic $T = \frac{\bar{X}-\mu}{s/\sqrt{n}}$ follows the **mathematical definition** of t_{n-1} distribution, as in

Main statement:

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{Z}{\sqrt{X_{n-1}^2/(n-1)}} \sim t_{n-1},$$

where $Z \sim N(0,1)$, $X_{n-1}^2 \sim \chi_{n-1}^2$.

1. Show that

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

Solution 1.

The χ^2 distribution is the square of independent standard normal distribution,

$$\chi_{df}^2 = \sum_i^{df} Z_i^2.$$

Also note that the sample mean $\bar{X}$ has follows a normal distribution with mean μ and standard deviation σ,

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}).$$

Finally, note that Z is given by

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}.$$

Thus by simple algebra we arrive at the desired conclusion.

$$\bar{X} - \mu = \frac{\sigma}{\sqrt{n}} Z.$$

2. Show that

$$\sum_i (x_i - \bar{x})^2 = \sigma^2 \sum_i (z_i - \bar{z})^2,$$

where $z_1, z_2, \ldots, z_n \sim_{ind} N(0,1)$$.

Solution 2.

We first recall that

$$X_i \sim N(\mu, \sigma^2).$$

and thus,

$$z_i = \frac{x_i - \mu}{\sigma} \sim N(0,1).$$

By simple algebra we can deduce that

$$x_i = z_i\sigma + \mu.$$

Consider $\sum_i (x_i - \bar{x})^2$. By part 1, we know that $\bar{x} = \frac{\sigma z}{\sqrt{n}} + \mu$.

Thus,

$$\sum_i (x_i - \bar{x})^2 = \sum_i [(z_i\sigma + \mu) - (\frac{\sigma z}{\sqrt{n}} + \mu)],$$

as required.

3. Applying the basic "row-by-column" matrix multiplication, proceed to calculate $\mathbf{z}' A_{n\times n} = (z_1, z_2, \ldots, z_n) \times A_{n\times n}$, where

$$A = \mathbf{I}_{n\times n} - \frac{1}{n}\mathbf{1}_{n\times n}$$

$$\mathbf{I}_{n\times n} = diag(\underbrace{1, 1, \ldots, 1}_{n}) = \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 & 0 \\ 0 & 1 & 0 & \ldots & 0 & 0 \\ \ldots & & & & & \\ 0 & 0 & 0 & \ldots & 0 & 1 \end{pmatrix}, \quad \mathbf{1}_{n\times n} = \begin{pmatrix} 1 & 1 & 1 & \ldots & 1 & 1 \\ 1 & 1 & 1 & \ldots & 1 & 1 \\ \ldots & & & & & \\ 1 & 1 & 1 & \ldots & 1 & 1 \end{pmatrix}$$

Solution 3.

It follows by the linear algebra operations that,

$$\mathbf{z}'A = \mathbf{z}'(\mathbf{I} - \frac{1}{n}\mathbf{1}).$$

$$= \mathbf{z}'\mathbf{I} - \frac{1}{n}\mathbf{z}'\mathbf{1}$$

$$= \mathbf{z}' - (\frac{1}{n}\sum_i^n z_i, \frac{1}{n}\sum_i^n z_i, ..., \frac{1}{n}\sum_i^n z_i)_{1\times n}$$

$$= \mathbf{z}' - (\bar{z}, \bar{z}, ..., \bar{z})_{1\times n}$$

$$= \mathbf{z}' - \ \bar{\mathbf{z}}$$

Which is the $1 \times n$ vector,

$$= (z_i - \bar{z})_i^n.$$

4. Using your result from part 3, show that $(\mathbf{z}'A_{n\times n})(\mathbf{z}'A_{n\times n})' = \sum_i (z_i - \bar{z})^2$.

Solution 4.

We have already shown that $\mathbf{z}'A_{n\times n} = (z_1 - \bar{z}, z_2 - \bar{z}, ..., z_n - \bar{z})_{1\times n}$.

It follows that

$$(\mathbf{z}'A_{n\times n})(\mathbf{z}'A_{n\times n})' = (z_1 - \bar{z}, z_2 - \bar{z}, ..., z_n - \bar{z})_{1\times n} \times (z_1 - \bar{z}, z_2 - \bar{z}, ..., z_n - \bar{z})_{n\times 1}$$

$$= (z_1 - \bar{z})^2 + (z_2 - \bar{z})^2 + ... + (z_n - \bar{z})^2$$

$$= \sum_i^n (z_i - \bar{z})^2,$$

as we wished to show.

5. For matrix $A_{n\times n}$ as defined in part 3, show that it is **idempotent**, as in

$$A_{n\times n} \times A_{n\times n} = A_{n\times n},$$

hence, combined with part 4, leading to the fact that

$$\sum_i (z_i - \bar{z})^2 = (\mathbf{z}'A_{n\times n})(\mathbf{z}'A_{n\times n})' = \mathbf{z}'A_{n\times n}A'_{n\times n}\mathbf{z} = \mathbf{z}'A_{n\times n}\mathbf{z}$$

Solution 5.

Recall that $A_{n\times n} = \mathbf{I}_{n\times n} - \frac{1}{n}\mathbf{1}_{n\times n}$.

Thus

$$A_{n\times n} \times A_{n\times n} = (\mathbf{I} - \frac{1}{n}\mathbf{1})(\mathbf{I} - \frac{1}{n}\mathbf{1})$$

$$= (\mathbf{I}^2 - \frac{2}{n}\mathbf{1} + \frac{1}{n^2}\mathbf{1}^2).$$

Here, we note that $\mathbf{1}^2 = \mathbf{n}$, where $\mathbf{n} \in \mathbb{R}^{n\times n}$ such that every element $\ltimes_{(i,j)} = n$, for all (i, j).

Thus

$$(\mathbf{I}^2 - \frac{2}{n}\mathbf{1} + \frac{1}{n^2}\mathbf{1}^2)$$

$$= (\mathbf{I}^2 - \frac{2}{n}\mathbf{1} + \frac{1}{n^2}\mathbf{n})$$

$$= (\mathbf{I}^2 - \frac{2}{n}\mathbf{1} + \frac{1}{n}\mathbf{1})$$

$$= (\mathbf{I}^2 - \frac{1}{n}\mathbf{1}),$$

as desired.

6. Matrix $A_{n\times n}$ as defined in part 3, has rank of $n - 1$. Why not n?

Solution 6.

We first consider $A_{n\times n}$.

$$A_{n\times n} = \begin{pmatrix} (1-\frac{1}{n}) & -\frac{1}{n} & -\frac{1}{n} & \dots & -\frac{1}{n} & -\frac{1}{n} \\ -\frac{1}{n} & (1-\frac{1}{n}) & -\frac{1}{n} & \dots & -\frac{1}{n} & -\frac{1}{n} \\ \dots \\ -\frac{1}{n} & -\frac{1}{n} & -\frac{1}{n} & \dots & -\frac{1}{n} & (1-\frac{1}{n}) \end{pmatrix}$$

$$= \frac{1}{n}\begin{pmatrix} (n-1) & 1 & 1 & \dots & 1 & 1 \\ 1 & (n-1) & 1 & \dots & 1 & 1 \\ \dots \\ 1 & 1 & 1 & \dots & 1 & (n-1) \end{pmatrix}.$$

Choose any row, such as the first row, and subtract all other rows from it. The resulting row is a vector $(0, 0, ..., 0))1 \times n$. Since this row can be expressed as a linear combination of the other rows, the matrix is not linearly independent. Thus it cannot have a rank n.

7. Combine the parts $2, 5$ and 6 with **Cochran's theorem** (see page 1) to show that

$$s/\sqrt{n} = \sqrt{\sum_i (x_i - \bar{x})^2/(n-1)} \times \frac{1}{\sqrt{n}} = \sqrt{\sigma^2 X^2_{n-1}/(n-1)} \times \frac{1}{\sqrt{n}}, \quad X^2_{n-1} \sim \chi^2_{n-1}$$

8. Combine parts 2 and 7 to prove the **main statement**.

Note: Some extra details could be found here (see p. 1-2 of the main post): https://stats.stackexchange.co m/questions/306937/quadratic-form-and-chi-squared-distribution.

Problem 2:

1. Write a *prop.sample.size()* function that will output the sample size needed for a one-sample proportion test to achieve

- a desired margin of error (argument #1)
- for a given confidence level (argument #2)

in the "worst-case scenario" (as was explained in class). What was meant by the "worst-case scenario"?

Solution 1.

We recall that

$$m = z_{1-\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

We wish to compute the desired sample size given margin of error m and confidence level $1 - \frac{\alpha}{2}$.

In particular, we wish to account for the "worst-case scenario", where the standard error is maximized by our choice of $\hat{p}$.

```r
prop.sample.size <- function(m, conf) {
  n <- (qnorm(conf)^2 * 0.5^2)/(m^2)
  response <- c("The sample size for a one-sample proportion test with a desired margin of error", m, "
  return(response)
}
```

For example, suppose $m = 0.4$, and the given confidence level is 0.995. Then,

```r
prop.sample.size(0.4, .995)
```

```
## [1] "The sample size for a one-sample proportion test with a desired margin of error"
## [2] "0.4"
## [3] " and given confidence level"
## [4] "0.995"
## [5] " is "
## [6] "10.3670259390956"
```

2. use your *prop.sample.size()* from part 1 to do exercise 8.50 from the Agresti book.

How many businesses fail? A study is planned to estimate the proportion of businesses started in the year 2006 that failed within five years of their start-up. How large a sample size is needed to guarantee estimating this proportion correct to within

a. 0.10 with probability 0.95?

```r
prop.sample.size(0.10, 0.95)
```

```
## [1] "The sample size for a one-sample proportion test with a desired margin of error"
## [2] "0.1"
## [3] " and given confidence level"
## [4] "0.95"
## [5] " is "
## [6] "67.6385863523852"
```

b. 0.05 with probability 0.95?

```r
prop.sample.size(0.05, 0.95)
```

```
## [1] "The sample size for a one-sample proportion test with a desired margin of error"
## [2] "0.05"
## [3] " and given confidence level"
## [4] "0.95"
## [5] " is "
## [6] "270.554345409541"
```

c. 0.05 with probability 0.99?

```r
prop.sample.size(0.05, 0.99)
```

```
## [1] "The sample size for a one-sample proportion test with a desired margin of error"
## [2] "0.05"
## [3] " and given confidence level"
## [4] "0.99"
## [5] " is "
## [6] "541.189443105434"
```

d. Compare sample sizes for parts a and b, and b and c, and summarize the effects of decreasing the margin of error and increasing the confidence level.

3. Write a *mean.sample.size()* function that will output the sample size needed for a one-sample mean test to achieve

- a desired margin of error (argument #1)
- for a given confidence level (argument #2)
- for a given standard deviation (argument #3)

Proceed to use that function in order to do exercise 8.53 from the Agresti book.

Solution 3

Recall that

$$t_{(n-1),\ 1-\alpha/2} \cdot \frac{s}{\sqrt{n}} = m,$$

implying,

$$n = \frac{\sigma^2 z^2}{m^2}.$$

This is working under the assumption that we have a population size of at least $n = 30$, since this would make our distrubution approach the normal distribution.

Thus we can construct our function.

```
mean.sample.size <- function(m, conf, s) {
  n <- ((s * qnorm(conf))/m)^2
  response <- c("The sample size for a one-sample mean test with a desired margin of error,", m, " giver
  return(response)
}
```

(8.53) *Income of the Native Americans* How large a sample size do we need to estimate the mean annual income of Native Americans in onondaga County, New York, correct to within \$1000 with probability 0.99? No information is available to us about the standar deviation of their annual income. We guess that nearly all of the incomes fall between \$0 and \$120,000 and that this distribution is approximately bell shaped.

Solution

We are given that $m = 1000$, and the confidence level is 0.99. Although we are not given the standard deviation σ, we recall that we may approximate σ by

$$\sigma \approx \frac{range}{6},$$

since 99% of the data lies within 3 standar deviations from the average, or 1/6 of the range.

Thus

$$\sigma \approx \frac{120,000}{6} = 20,000.$$

Using the function written above,

```
mean.sample.size(1000, 0.99, 20000)
```

```
## [1] "The sample size for a one-sample mean test with a desired margin of error,"
## [2] "1000"
## [3] " given confidence level"
## [4] "0.99"
## [5] "and standard deviation"
## [6] "20000"
## [7] " is "
## [8] "2164.75777242174"
```

We find that 2165 participants are needed to reach the desired criteria.