Module 1: Review of frequentist estimation of a population mean () and standard deviation ()

## Disclaimer

Module 1 is a review of frequentist statistical methods. If the material covered in Module 1 is new for you, we recommend consulting a statistics textbook, such as Peck and Devore (2011), Witte and Witte (2017), or McClave and Sincih (2017), before continuing on to Modules 2, 3 and 4.

## The Example

Imagine a nursing home has hired you to conduct some statistical analyses for them. One of the things the nursing home would like to know is the mean and standard deviation of cognitive ability for their patients.

## The Data

To measure the patients’ cognitive ability, you decide to administer the Mini-Mental State Examination (MMSE; Folstein et al. 1975) to only a sample of the nursing home’s patients (due to logistic constraints). The MMSE is an examination developed to measure cognitive ability and is often used as a diagnostic test for dementia (Siu 1991). The MMSE awards points for correct question responses up to a maximum score of 30. That is, scores range from 0-30, with higher scores indicating higher cognitive ability.

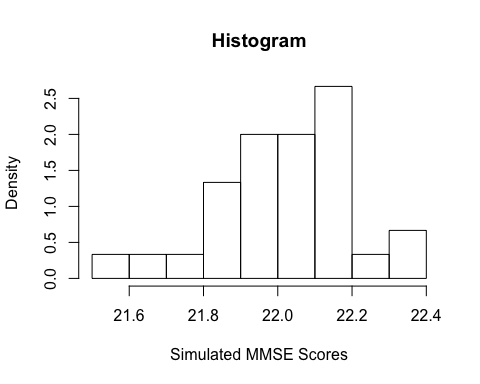
Because this is a hypothetical example, we will simulate the MMSE scores. Our simulation will produce MMSE-like scores for a sample of 30 nursing home patients who are, on average, cognitively impaired. (According to Kochhan (2010), MMSE scores below 23 may indicate cognitive impairment.) Scores are simulated from a normal distribution with and . To simulate the MMSE scores, open a new R script in RStudio, copy and paste the following code, and run it.

set.seed(1) # Setting the seed ensures that your random numbers match my random numbers   
 # As long as our seed values match, the chosen seed value is arbitrary  
MMSE <- rnorm(n = 30, mean = 22, sd = 0.2) # Simulate 30 observations from a normal distribution with a mean of 22 and standard deviation of 0.2  
simdata <- as.data.frame(MMSE)

## Visualizing the Simulated Scores

Copy and paste the following code into your R script and run it to produce a histogram of the simulated scores.

hist(simdata$MMSE, main = "Histogram", breaks = 10, prob = T, xlab = "Simulated MMSE Scores")



## Testing the Assumption of Normality

In the following analysis, we will fit a statistical model that assumes the outcome variable is normally distributed in the population. Although the histograms above suggest that the data could have been sampled from a normal population distribution, let’s use the Shapiro-Wilk test of normality to evaluate this. Copy and paste the following code into your R script and run it.

shapiro.test(simdata$MMSE) # Shapiro-Wilk test of normality

The results of the Shapiro-Wilk test are insignificant (p > .05), so we fail to reject the null hypothesis. In other words, there is not enough evidence to suggest that the simulated MMSE scores are non-normally distributed.

## The Analysis

To **estimate and**  using the simulated MMSE scores, we fit a linear model to the scores using the lm() function. The linear model specified below will produce estimates of the population mean , the standard error of the estimate of , and the residual standard error . The residual standard error represents the spread of the simulated MMSE scores about their mean. To obtain estimates of , the standard error of the estimate of , and , copy and paste the following code into your R script and run it.

mod1 <- lm(MMSE ~ 1, data = simdata) # Fit a linear model   
summary(mod1) # Examine the model output  
summary(mod1)$coefficient[1] # This is an estimate of the population mean of MMSE scores  
summary(mod1)$coefficient[2] # This is the standard error of the estimate of the mean of MMSE scores  
summary(mod1)$sigma # This is an estimate of the population standard deviation of MMSE scores

## Interpreting the Model Results

The line of code

summary(mod1)$coefficient[1]

outputs the sample mean , a point estimate of (the population mean of , where represents the simulated MMSE scores). represents the central tendency of . The line of code

summary(mod1)$sigma

outputs the sample standard deviation , a point estimate of (the population standard deviation of ). represents the spread of about its mean. The line of code

summary(mod1)$coefficient[2]

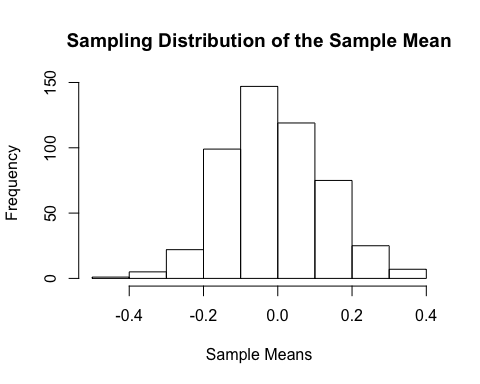
outputs the standard error of . Because **it’s so important to understand the concept of the sampling distribution when interpreting the standard error**, we present a demonstration of the concept below to refresh your memory.

Imagine if we were to draw many samples of size *n* from a population, estimate for each sample, and plot a histogram of every . This histogram would represent the sampling distribution of , and its standard deviation would represent the sampling variability of . The standard deviation of the sampling distribution gives us information about how far the typical sample mean is from the “true” value of the mean in the population.

The sampling distribution is a theroetical concept. In practice, we don’t draw many samples of size *n* from the population because we only have the resources (time, money, etc.) to draw a single sample. For this reason, we don’t know the standard deviation of the sampling distribution. Rather, we estimate the standard deviation of the sampling distribution of the sample mean (the standard error of the mean) by dividing the sample standard deviation (or the population standard deviation, if you know it) of the variable by the square root of *n*.

The following code (1) simulates a population of scores, (2) draws many (in this case, 500) samples of size *n* = 50 from this population, (3) calculates the mean of each sample, and (4) plots a histogram of the sample means. The histogram is a visualization of the theoretical sampling distribution. Copy and paste the following code into your R script and run it.

# (1) Simulate a population of scores  
set.seed(1)  
population <- rnorm(n = 400000, mean = 0, sd = 1)   
  
# (2) Draw many (in this case 500) samples of size N = 50 from the population and (3) Estimate the mean of each sample  
means <- unlist(lapply(50, function(x){  
 lapply(1:500, function(y){  
 mean(sample(population,x,replace=FALSE))  
 })  
}))  
  
# (4) Plot a histogram of the means (a.k.a. the sampling distribution of the mean)  
hist(means, main = 'Sampling Distribution of the Sample Mean', xlab = "Sample Means")



According to Peck and Devore (2011), the first two general properties of the sampling distribution are that:

1. (the mean of the sampling distribution)
2. (the standard deviation of the sampling distribution)

Using the demonstration of the sampling distribution in the block of code above, let’s verify the first two general properties of the sampling distribution. Copy and paste the following code into your R script and run it.

mean(means) # Calculate the mean of the 500 sample means   
sd(means) # Calculate the standard deviation of the 500 sample means  
1/sqrt(50) # Calculate the population standard deviation divided by the square root of the size of the 500 samples

The line of code

mean(means)

outputs (the mean of the sampling distribution of ). From this output, we see that . The line of code

sd(means)

outputs (the standard deviation of the sampling distribution of ), and the line of code

1/sqrt(50)

outputs . From the output of these two lines of code, we see that .

## Interpreting a Confidence Interval for

Frequentist confidence intervals provide another measure of the uncertainty of the value of a parameter estimate (Kruschke and Liddell 2018). The higher the specified level of confidence (typically 95% in psychology), the wider the interval (and vice versa).

The interpretation of a frequentist confidence interval relies on the concept of long-run frequencies as discussed in the previous illustration of the sampling distribution. For example, one interpretation of a 95% confidence interval is that 100 samples drawn from a population yield confidence intervals, of which 95 are expected to contain the “true” population parameter. This interpretation applies the statement of probability to the sampling distribution.

An **incorrect interpretation** of a frequentist confidence interval says that the confidence interval contains the 95% most probable values of a parameter. This interpretation applies the statement of probability to the parameter as if it has a probability distribution. In the frequentist perspective, parameters do not have statements of probability attached to them because they are not random variables with probability distributions. That is, they are conceptualized as fixed at some “true” value in the population. Therefore, the interval either contains the parameter or it does not.

In summary, statements of probability in the frequentist perspective apply to the sampling distribution, not the parameter (Gelman et al. 2013). This distinction will become clearer as we introduce concepts from the Bayesian perspective in Modules 3 and 4.

The following code calculates a 95% confidence interval for . Copy and paste the following code into your R script and run it.

confint(mod1)

## References

Folstein, M. F., Folstein, S. E., and McHugh, P. R. (1975), “Mini-mental state: A practical method for grading the cognitive state of patients for the clinician”, *Journal of Psychiatric Research*, *12*(3), 189-198.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013), *Bayesian data analysis*, 3rd Edition, CRC press.

Kochhann, R., Varela, J. S., de Macedo Lisboa, C. S., and Chaves, M. L. F. (2010), “The Mini Mental State Examination: review of cutoff points adjusted for schooling in a large Southern Brazilian sample”, *Dementia and Neuropsychologia*, 4(1), 35.

Kruschke, J. K., and Liddell, T. M. (2018), “The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective”, *Psychonomic Bulletin and Review*, *25*(1), 178-206.

McClave, J. T., and Sincich, T. (2016), *Statistics*, 13th Edition, Pearson Education.

Peck, R., and Devore, J. L. (2012), *Statistics: The exploration and analysis of data*, 7th Edition, Brooks/Cole.

Siu, A. L. (1991), “Screening for dementia and investigating its causes”, *Annals of Internal Medicine*, 115(2), 122-132.

Witte, R. S., and Witte, J. S. (2017), *Statistics*, 11th Edition, Wiley.