Module 2: Review of frequentist hypothesis testing for two independent samples

## Disclaimer

Module 2 is a review of frequentist statistical methods. If the material covered in Module 2 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 3 and 4.

## The Example

Imagine a nursing home has hired you to conduct some statistical analyses. In Module 1, you measured the cognitive state of a sample of the nursing home’s patients using the Mini-Mental State Examination (MMSE; Folstein et al. 1975) and estimated the mean and standard deviation. Now you will conduct another statistical analysis for the nursing home.

Imagine that in your literature review you came across a few articles on the increased risk of cognitive impairment for individuals with low education (see, Stern et al. 1994; Letenneur et al. 1999; Le Carret et al. 2003). **The nursing home would now like to know whether the mean cognitive ability of their patients with low levels of education is significantly lower than the mean cognitive ability of their patients with high levels of education.**

## The Data

Imagine you administer the MMSE to a new sample of 60 nursing home patients to obtain a measure of their cognitive state. **The MMSE scores will serve as the outcome variable in the analysis.** The nursing home will give you access to demographic information on their patients, including which of them completed no more than a high school education and which of them obtained some form of education beyond high school. **This binary education level variable will serve as the predictor in the analysis.**

Because this is a hypothetical example, we will simulate the 60 MMSE-like scores along with a binary variable representing education level. The data will be simulated such that the mean of the simulated MMSE scores is lower for the low education group than for the high education group. “Low education” is defined as completing no more than a high school education, and “high education” is defined as completing some form of education beyond high school. In the simulation below, the low education group is given the label “0” and the high education group is given the label “1”. To run the simulation, open a new R script in RStudio, copy and paste the following code, and run it.

set.seed(1)  
nrep <- 30 # Tell the simulation that there will be 30 observations per group  
ngroup <- 2 # Tell the simulation that there will be two groups (low and high education)  
b0 <- 22 # Tell the simulation that the low education group's MMSE scores will be simulated from a distribution with a mean of 22  
b1 <- 0.75 # Tell the simulation that the high education group's MMSE scores will be simulated from a distribution with a mean of 22.5 (b0 + b1)  
educ <- rep(c("0", "1"), each = nrep) # Simulate the predictor variable  
err <- rnorm(n = ngroup\*nrep, mean = 0, sd = 1.25) # Simulate 60 residuals from a distribution with a mean of 0 and standard deviation of 1.25  
MMSE <- b0 + b1\*(educ == "1") + err # Simulate the outcome variable  
simdata <- data.frame(MMSE, educ) # Create a dataframe containing the simulated variables MMSE and educ

Let’s view the first few lines of the simulated dataset, which should consist of MMSE scores for 60 nursing home patients and a binary variable indicating whether each individual falls into group “0” or group “1”. Copy and paste the following code into your R script and run it.

head(simdata)

Let’s also output the mean of the MMSE scores for each group. Copy and paste the following code into your R script and run it.

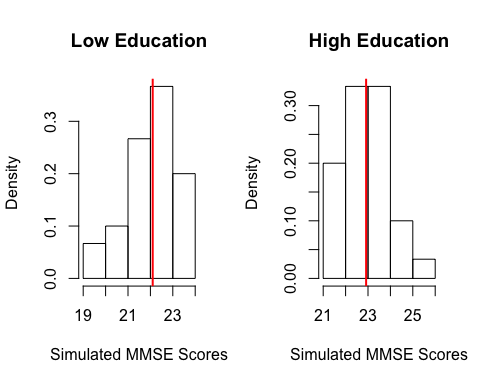
mean(simdata$MMSE[educ=="0"]) # Calculate the mean of the simulated MMSE scores for the low education group (group "0")  
mean(simdata$MMSE[educ=="1"]) # Calculate the mean of the simulated MMSE scores for the high education group (group "1")

Note that the MMSE scores were simulated such that the mean of MMSE scores for group “0” is less than the mean of MMSE scores for group “1”.

## Visualizing the Simulated Scores

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

par(mfrow = c(1, 2)) # Create a plotting window that will display the histograms side-by-side  
hist(simdata$MMSE[educ=="0"], main = 'Low Education', breaks = 5, prob = T, xlab = "Simulated MMSE Scores")   
abline(v = mean(simdata$MMSE[educ=="0"]), col = "red", lwd=2) # Overlay a red line at the sample mean for group "0"  
hist(simdata$MMSE[educ=="1"], main = 'High Education', breaks = 5, prob = T, xlab = "Simulated MMSE Scores")   
abline(v = mean(simdata$MMSE[educ=="1"]), col = "red", lwd=2) # Overlay a red line at the sample mean for group "1"



## 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$MMSE[educ=="0"]) # Shapiro-Wilk test of normality for group "0"  
shapiro.test(simdata$MMSE[educ=="1"]) # Shapiro-Wilk test of normality for group "1"

The results of the Shapiro-Wilk tests 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

In the simulation above, we verified that the data were simulated such that the mean of the simulated MMSE scores for the low education group (group “0”) is less than the mean of the simulated MMSE scores for the high education group (group “1”). Now we need to conduct an analysis to determine whether this observed difference in means is statistically significant. That is, the statistical analysis determines whether the observed difference in means is likely to be due to some real difference in the population as opposed to sampling variability.

You probably remember from your undergraduate psychology statistics course that **two independent samples *t* tests are used to compare the means of two independent groups**. We are going to conduct a two-tailed *t* test, so the null and alternative hypotheses are:

and

, where is the population mean for the low education group, is the population mean for the high education group, and is the estimated difference between and .

In Module 1, we fit a linear model to the MMSE scores, producing estimates of the population mean , the standard error of the estimate of , and the residual standard error . In this module, we will also fit a linear model to the MMSE scores but we will include education level as a binary predictor. Although you were probably taught in your undergraduate psychology statistics course to think of the *t* test as a stand-alone statistical tool, **the *t* test is actually just a special case of the linear model**. Let us explain.

Output from the linear model specified in the upcoming code will include:

1. an estimate of the population mean for the low education group ,
2. the standard error of the estimate of ,
3. an estimate of (denoted as ),
4. the standard error of the estimate of , and
5. the population standard deviation (referred to as the residual standard error).

**The null and alternative hypotheses for this linear model look different from those of the *t* test you’re familiar with, but you’ll soon see that retaining/rejecting the null hypothesis results in the same conclusion.** The null and alternative hypotheses are:

and

, where represents .

**If is statistically significant (*p* < .05), we reject the null hypothesis and conclude that is significantly different from zero.** Copy and paste the following code into your R script and run it.

mod2 <- lm(MMSE ~ educ, data = simdata) # Fit a linear model   
summary(mod2) # Examine the model output

## Interpreting the Model Results

The line of code

summary(mod2)$coefficient[1]

outputs **the sample mean for group “0”** (the low education group). is a point estimate of (the population mean of , where represents the simulated MMSE scores for group “0”). The line of code

summary(mod2)$coefficient[2]

outputs **the estimate of** , the estimated difference between and . The line of code

summary(mod2)$coefficient[1] + summary(mod2)$coefficient[2]

outputs **the estimate of** . Since is the estimated difference between and , it follows that . In other words, the above line of code outputs the sample mean for group “1” (the high education group), and is a point estimate of (where represents the simulated MMSE scores for group “1”). The line of code

summary(mod2)$coefficient[ , 2]

outputs **the standard errors of and** , repsectively. Remember from Module 1, the standard error is a measure of sampling variability, and sampling variability tells us to what degree we should be confident that our sample statistic is an accurate estimate of the population parameter. That is, sampling variability tells us how far the typical sample statistic is from the “true” value of the parameter in the population. If you need a refresher on the concept of sampling variability, return to the demonstration of the sampling distribution in the section titled Interpreting the Model Results from Module 1.

Finally, the line of code

summary(mod2)$sigma

outputs **the sample standard deviation** , a point estimate of (the population standard deviation). represents the spread of about its mean, where denotes the simulated MMSE scores. You may have noticed that the model output only contains a single estimate of (as opposed to containing estimates of both and ). This is because we’ve assumed the population standard deviation is the same for both groups. In practice, you would test this assumption. (As a note, we didn’t test this assumption because we happen to know that the distributions from which the MMSE scores were simulated had the same standard deviation (). If in practice, you ran a test of homoscedasticity of variance on your sample and found that the assumption was likely violated, you would need to specify a model that estimates separate standard deviations for each group. In your undergraduate psychology statistics course, you may have learned about the Welch’s *t* test, which is a *t* test that addresses violated assumptions of homescedasticity by estimating a separate standard deviation for each group.)

**The main piece of model output that must be evaluated to address our research question is whether or not the estimate of (the estimated difference between and ) is statistically significant (*p* < .05)**. If *p* < .05, we reject the null hypothesis and conclude that is significantly different from zero. Note here that the null hypothesis is never accepted; we either reject the null hypothesis or fail to obtain enough evidence to reject the null hypothesis.

## Interpreting the *p*-value

The “*p*” we’re referring to when we evaluate whether or not “*p* < .05” is the *p*-value, whose interpretation relies on the concept of long-run frequencies as discussed in the demonstration of the sampling distribution in Module 1.

To better understand the long-run interpretation of the *p*-value in the context of our current analysis, imagine drawing many samples of size *n* from a population, conducting a two-tailed, two independent samples *t* test on each sample, and saving the test statistic resulting from each *t* test conducted. A histogram of these test statistics would represent the sampling distribution of the test statistic.

Now, think about our current analysis. The line of code

summary(mod2)$coefficient[2, 3]

outputs **the test statistic we obtained from our two-tailed, two independent samples *t* test**. The line of code

summary(mod2)$coefficient[2, 4]

outputs **the corresponding *p*-value, which is an estimate of the proportion of test statistics in the sampling distribution of the test statistic that are equal to or more extreme than the test statistic we obtained, assuming that the null hypothesis is true.** In other words, we estimate that

summary(mod2)$coefficient[2, 4]\*100

percent of the test statistics in the sampling distribution of the test statistic would be equal to or more extreme than the test statistic we obtained if the null hypothesis were true. Put simply, **it would be extremely rare to observe the test statistic we obtained if the null hypothesis were true, so we conclude the evidence suggests the null hypothesis is false.**

**It is important to note that the *p-*value cannot be interpreted as the probability of the null hypothesis being true (Gigerenzer et al. 2004; Kruschke and Liddell 2018).** To reiterate, the *p*-value represents the proportion of test statistics in the sampling distribution of the test statistic that are equal to or more extreme than the test statistic we obtained, assuming the null hypothesis is true. There are, however, statistical tools available in the Bayesian perspective for obtaining the probability of a null hypothesis being true. To learn more, continue on to Modules 3 and 4.

## 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.

Gigerenzer, G., Krauss, S., and Vitouch, O. (2004), “The null ritual”, *The Sage handbook of quantitative methodology for the social sciences*, 391-408.

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.

Le Carret, N., Lafont, S., Letenneur, L., Dartigues, J. F., Mayo, W., and Fabrigoule, C. (2003), “The effect of education on cognitive performances and its implication for the constitution of the cognitive reserve”, *Developmental neuropsychology*, 23(3), 317-337.

Letenneur, L., Gilleron, V., Commenges, D., Helmer, C., Orgogozo, J. M., and Dartigues, J. F. (1999), “Are sex and educational level independent predictors of dementia and Alzheimer’s disease? Incidence data from the PAQUID project”, *Journal of Neurology, Neurosurgery and Psychiatry*, 66(2), 177–183.

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.

Stern, Y., Gurland, B., Tatemichi, T. K., Tang, M. X., Wilder, D., and Mayeux, R. (1994), “Influence of education and occupation on the incidence of Alzheimer’s disease”, *Journal of the American Medical Association*, 271, 1004–1010.

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