

1. When conducting the work of Lab 11, we conducted the test that uses the Central Limit Theorem even though the sample size was “small” (i.e., $n < 30$). It turns out, that how “far off” the t -test is can be computed using a first-order Edgeworth approximation for the error. Below, we will do this for the the further observations.

(a) Boos and Hughes-Oliver (2000) note that

$$P(T \leq t) \approx F_Z(t) + \underbrace{\frac{\text{skew}}{\sqrt{n}} \frac{(2t^2 + 1)}{6}}_{\text{error}} f_Z(t),$$

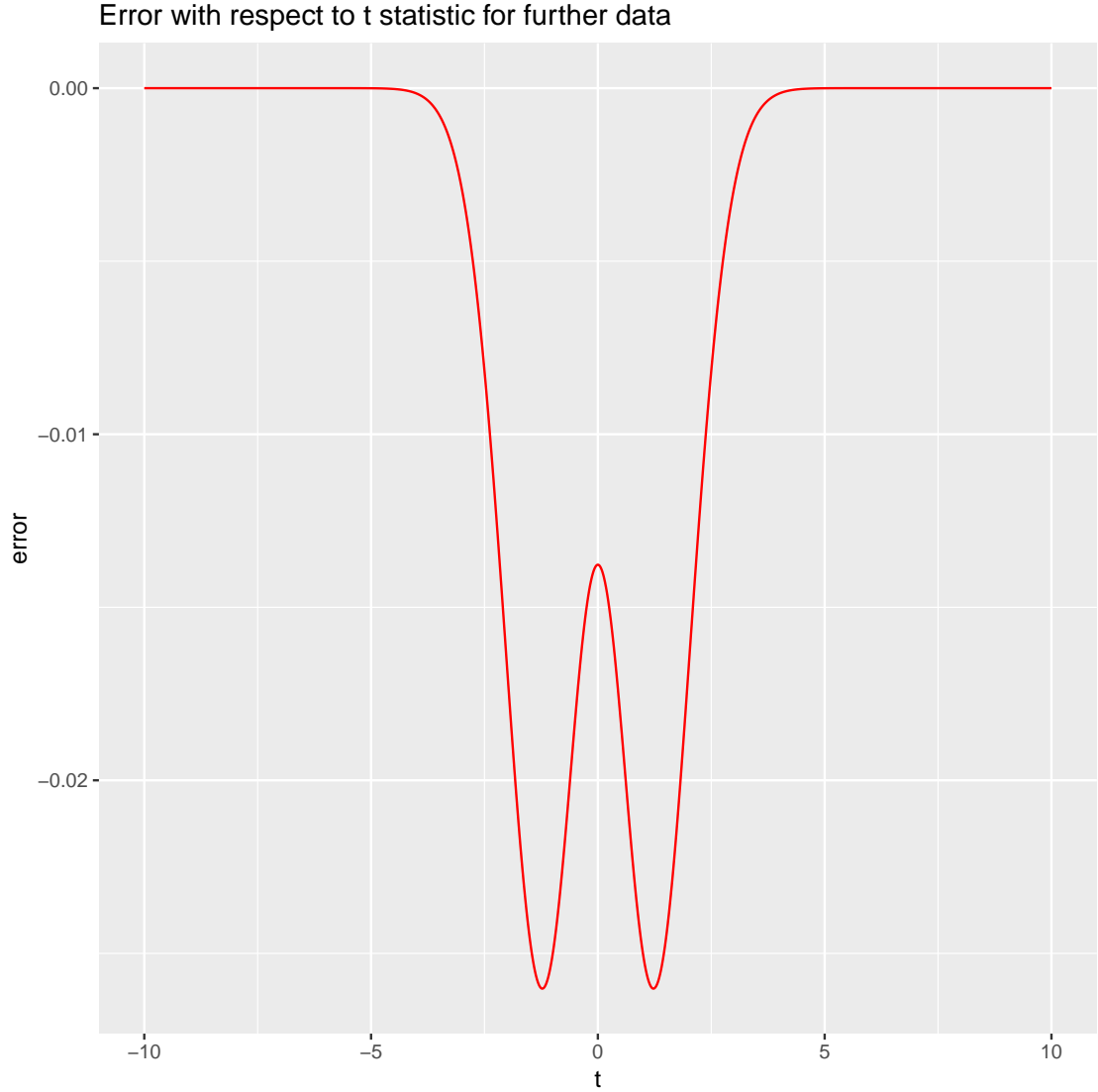
where $f_Z(\cdot)$ and $F_Z(\cdot)$ are the Gaussian PDF and CDF and skew is the skewness of the data. What is the potential error in the computation of the p -value when testing $H_0 : \mu_X = 0; H_a : \mu_X < 0$ using the zebra finch further data?

```
library(tidyverse)
library(e1071)
data = read_csv("zebrafinches.csv")
far.vec = data$farther #farther data
mu0 = 0
n = 25
far.t.stat = (mean(far.vec) - mu0)/(sd(far.vec)/sqrt(n)) #far t statistic
far.pdf = dnorm(far.t.stat) #pdf using the far t statistic
#potential error for further data
(far.error = (skewness(far.vec)/sqrt(n))*((2*far.t.stat^2 + 1)/6)*far.pdf)
## [1] -1.226006e-13
```

As you can see, the potential error is essentially zero.

- (b) Compute the error for t statistics from -10 to 10 and plot a line that shows the error across t . Continue to use the skewness and the sample size for the zebra finch further data.

```
t.stat = seq(-10,10, by = 0.01)
error.vec = (skewness(far.vec)/sqrt(n))*((2*t.stat^2 + 1)/6)*dnorm(t.stat)
error.dat = data.frame(t = t.stat, error = error.vec)
error.plot = ggplot(data = error.dat, aes(x = t, y = error))+
  geom_line(color = "red") +
  labs(title = "Error with respect to t statistic for further data")
error.plot
```



As you can see, the error is essentially zero when the magnitude of the t statistic is between 5 and 10. This is because small shifts don't matter very much when the magnitude of the t statistic is so large. As the magnitude of t gets smaller, the magnitude of the error begins to increase towards -0.026. However, between t values of roughly -1 and 1, the error gets a little smaller (approaches -0.013). This is because there is an inflection point somewhere near 1 and -1.

- (c) Suppose we wanted to have a tail probability within 10% of the desired $\alpha = 0.05$. Recall we did a left-tailed test using the further data. How large of a sample size would we need? That is, we need to solve the error formula equal to 10% of the desired left-tail probability:

$$0.10\alpha \stackrel{\text{set}}{=} \underbrace{\frac{\text{skew}}{\sqrt{n}} \frac{(2t^2 + 1)}{6} f_Z(t)}_{\text{error}},$$

which yields

$$n = \left(\frac{\text{skew}}{6(0.10\alpha)} (2t^2 + 1) f_Z(t) \right)^2.$$

```
alpha = 0.05
t = qnorm(alpha) #t statistic at alpha = 0.05
#required sample size to get a tail probability within 10% of alpha
(n.required = ((skewness(far.vec)/(6*.1*alpha))*((2*t^2 + 1))*dnorm(t))^2)
## [1] 520.8876
```

It would require a sample size of 521 in order to get a tail probability within 10% of alpha (0.05).

2. Complete the following steps to revisit the analyses from lab 11 using the bootstrap procedure.

- (a) Now, consider the zebra finch data. We do not know the generating distributions for the closer, further, and difference data, so perform resampling to approximate the sampling distribution of the T statistic:

$$T = \frac{\bar{x}_r - 0}{s/\sqrt{n}},$$

where \bar{x}_r is the mean computed on the r^{th} resample and s is the sample standard deviation from the original samples. At the end, create an object called `resamples.null.closer`, for example, and store the resamples shifted to ensure they are consistent with the null hypotheses at the average (i.e., here ensure the shifted resamples are 0 on average, corresponding to $t = 0$, for each case).

```
far.vec = data$farther
close.vec = data$closer
diff.vec = data$diff
R = 10000 #simulations
n = 25
far.resamples = tibble(t.stats = rep(NA, R))
close.resamples = tibble(t.stats = rep(NA, R))
diff.resamples = tibble(t.stats = rep(NA, R))
#resamples for all 3 types of data
for(i in 1:R){
  curr.far.resample = sample(far.vec,
                             size = n,
                             replace = T)
  curr.close.resample = sample(close.vec,
                                size = n,
                                replace = T)
  curr.diff.resample = sample(diff.vec,
                              size = n,
                              replace = T)
  #resampled t statistics for close, far, and diff data
  far.resamples$t.stats[i] = mean(curr.far.resample)/(sd(far.vec)/sqrt(n))
  close.resamples$t.stats[i] = mean(curr.close.resample)/(sd(close.vec)/sqrt(n))
  diff.resamples$t.stats[i] = mean(curr.diff.resample)/(sd(diff.vec)/sqrt(n))
}
#shifted resamples
resamples.null.far = far.resamples$t.stats - mean(far.resamples$t.stats)
resamples.null.close = close.resamples$t.stats - mean(close.resamples$t.stats)
resamples.null.diff = diff.resamples$t.stats - mean(diff.resamples$t.stats)
#mean of shifted resamples
mean(resamples.null.far)
## [1] -1.688871e-16
```

```
mean(resamples.null.close)
## [1] -3.936351e-16
mean(resamples.null.diff)
## [1] 3.467046e-16
```

- (b) Compute the bootstrap p -value for each test using the shifted resamples. How do these compare to the t -test p -values?
 - (c) What is the 5th percentile of the shifted resamples under the null hypothesis? Note this value approximates $t_{0.05, n-1}$. Compare these values in each case.
 - (d) Compute the bootstrap confidence intervals using the resamples. How do these compare to the t -test confidence intervals?
3. Complete the following steps to revisit the analyses from lab 11 using the randomization procedure.
- (a) Now, consider the zebra finch data. We do not know the generating distributions for the closer, further, and difference data, so perform the randomization procedure
 - (b) Compute the randomization test p -value for each test.
 - (c) Compute the randomization confidence interval by iterating over values of μ_0 .
Hint: You can “search” for the lower bound from Q_1 and subtracting by 0.0001, and the upper bound using Q_3 and increasing by 0.0001. You will continue until you find the first value for which the two-sided p -value is greater than or equal to 0.05.
4. **Optional Challenge:** In this lab, you performed resampling to approximate the sampling distribution of the T statistic using

$$T = \frac{\bar{x}_r - 0}{s/\sqrt{n}}.$$

I’m curious whether it is better/worse/similar if we computed the statistics using the sample standard deviation of the resamples (s_r), instead of the original sample (s)

$$T = \frac{\bar{x}_r - 0}{s_r/\sqrt{n}}.$$

- (a) Perform a simulation study to evaluate the Type I error for conducting this hypothesis test both ways.
- (b) Using the same test case(s) as part (a), compute bootstrap confidence intervals and assess their coverage – how often do we ‘capture’ the parameter of interest?

References

Boos, D. D. and Hughes-Oliver, J. M. (2000). How large does n have to be for z and t intervals? *The American Statistician*, 54(2):121–128.