AGH Practicals 10 April 2021

Background:

Suppose one has two independent samples, $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ and wishes to test the hypothesis twidehat the mean of the x population is equal to the mean of the y population:

$$H_0 : \mu_x = \mu_y$$

Let $\overline{X}$ and $\overline{Y}$ denote the sample means of the x's and y's and let s_x and s_y denote the respective standard deviations. The standard test of this hypothesis H_0 is based on the t statistic

$$T = \frac{\overline{X} - \overline{Y}}{s_p\sqrt{1/m + 1/n}}$$

where $s_p = \sqrt{\frac{(m-1)s_x^2+(n-1)s_y^2}{m+n-2}}$ is the pooled standard deviation.

Under the null hypothesis, the test statistic T has a t distibution with $m + n - 2$ degrees of freedom when

- both the x's and y's are independent random samples from normal distributions
- the standard deviations of the x and y populations, σ_x and σ_y are equal.

Suppose the level of significance of the test is set at α. Then one will reject the null hypothesis when

$$|T| \geq t_{n+m-2,\alpha/2}$$

where $t_{\nu,\alpha}$ is the $(1 - \alpha)$ quantile of a t random variable with ν degrees of freedom.

Writing a function to compute the t Statistic

```
tstatistic <- function(x, y) {
  m <- length(x)
  n <- length(y)
  sp <- sqrt(((m - 1)*sd(x)^2 + (n - 1)*sd(y)^2)/(m + n - 2))
  t.stat <- (mean(x) - mean(y))/(sp*sqrt(1/m + 1/n))
}

data.x <- c(1,4,3,6,5)
data.y <- c(5,4,7,6,10)
print(tstatistic(data.x, data.y))
```

```
## [1] -1.937926
```

```
x <- rnorm(10, mean = 50, sd = 10)
y <- rnorm(10, mean = 50, sd = 10)
print(tstatistic(x, y))
```

```
## [1] 0.4490436
```

Simulation Algorithm

Suppose we are interested in learning about the true significance level for the t statistic when the populations do not follow the standard assumption of normality and equality of variances. In general, the true significance level will depend on

- the stated level of significance α
- the shape of the populations (normal, skewed, heavy-tailed, etc)
- the spreads of the two populations as measured by the two standard deviations
- the sample sizes m and n

Given a particular choice of α, shape, spreads, and sample sizes, we wish to estimate the true significance level given by

$$\alpha = P(|T| \geq t_{n+m-2,\alpha/2}) \tag{1}$$

Outline of the simulation algorithm to compute α:

1. Simulate a random sample $x_1, \ldots, x_m$ from the first population and $y_1, \ldots, y_n$ from the second population.
2. Compute the T statistic from the two samples.
3. Decide if $|T|$ exceeds the critical point and H_0 is rejected.
4. Repeat (1) to (3) N times. Estimate the true significance level by

$$\widehat{\alpha} = \text{number of rejections of } \ H_0 \ /N \tag{2}$$

```
mc_tstat <- function(alpha, N, pars, seed) {
  set.seed(seed)
  n.reject <- 0

  for (i in 1:N) {
    x <- rnorm(pars$m, pars$mu1, pars$sigma1)
    y <- rnorm(pars$n, pars$mu2, pars$sigma2)
    t.stat <- tstatistic(x, y)
    if (abs(t.stat) > qt(1 - alpha/2, pars$n + pars$m - 2))
      {n.reject <-  n.reject + 1}
    est.sig.level <- n.reject / N
  }

  print(est.sig.level)
}
```

Case 1: Normal populations with zero means and equal spreads

```
pars <- list(m = 10, n = 10, mu1 = 0, mu2 = 0, sigma1 = 1, sigma2 = 1)
mc_tstat(alpha = 0.1, N = 10000, pars, seed = 1234)
```

```
## [1] 0.0985
```

Case 2: Normal populations with zero means and very different spreads

```
pars <- list(m = 10, n = 10, mu1 = 0, mu2 = 0, sigma1 = 1, sigma2 = 10)
mc_tstat(alpha = 0.1, N = 10000, pars, seed = 1234)
```

```
## [1] 0.1114
```

Case 3: t populations, 4 df and equal spreads

```
pars <- list(m = 10, n = 10, df = 4)
mc_tstat(alpha = 0.1, N = 10000, pars, seed = 1234)
```

```
## [1] 0.0922
```

Case 4: Exponential populations, equal rates

```
pars <- list(m = 10, n = 10, rate = 1)
mc_tstat(alpha = 0.1, N = 10000, pars, seed = 1234)
```

```
## [1] 0.0972
```

Case 5: One normal population, one exponential population

```
pars <- list(m = 10, n = 10, mu = 10, sigma = 2, rate = 1/10)
mc5 <- mc_tstat(alpha = 0.1, N = 10000, pars, seed = 1234)
```

```
## [1] 0.1527
```

Summary:

Populations	True Significance Level
Normal populations with equal spreads	0.0985
Normal populations with unequal spreads	0.1114
t(4) distributions with equal spreads	0.0922
Exponential populations with equal spreads	0.0972
Normal and exponential populations with unequal spreads	0.1527

Exact sampling distribution for Case 5

```
m <- 10
n <- 10
t.simulation <- function() {
  tstatistic(rnorm(m, mean = 10, sd = 2), rexp(n, rate = 1/10))
}

tstat.vector <- replicate(10000, t.simulation())
plot(density(tstat.vector), xlim = c(-5, 8), ylim = c(0, 0.4), lwd = 3, main = "")
```

```r
curve(dt(x, df = 18), add = TRUE)
legend(4, 0.3, c("exact", "t(18)"), lwd = c(3, 1))
```