SM Exercise 7.17

Kristine Villaluna

19/02/2023

SM Problem 7.17:

Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_m$ be independent random samples from continuous distributions F_X and F_Y. We wish to test the hypothesis $H_0 : F_X = F_Y$. Define indicator variables $I_{ij} = 1\{X_i < Y_j\}$ for $i = 1, \ldots, m, j = 1, \ldots, n$, and let

$$U = \sum_{i=1}^{m} \sum_{j=1}^{n} I_{ij}.$$

a) Show that under H_0 $E(U) = mn/2$ and var$(U) = mn(m + n + 1)/12$. For the variance, you will need to consider cov(I_{ij}, I_{ik}), and cov(I_{ij}, I_{kl}), where i, j, k, l are distinct.

For this question, we would like to show that under $H_0 : F_X = F_Y$, $E(U) = mn/2$ and var$(U) = mn(m + n + 1)/12$.

Let us first assume that the null hypothesis is true, and that $F_X = F_Y$. Now, because X and Y are independent and have the same continuous distributions (from our assumption), we can write:

$$P(X < Y) = P(Y < X)$$

and

$$\begin{aligned} P(X < Y) + P(Y < X) &= 1 \quad \text{by axioms of probability} \\ \rightarrow P(X < Y) &= \frac{1}{2} \end{aligned}$$

Thus, we can continue to find the expected value of U as follows:

$$\begin{aligned} E(U) &= \sum_{i=1}^{m} \sum_{j=1}^{n} E(I_{ij}) \\ &= \sum_{i=1}^{m} \sum_{j=1}^{n} P(X_i < Y_i) \cdot 1 \\ &= \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{1}{2} \\ &= \frac{mn}{2} \end{aligned}$$

Next, we are interested in showing the variance is equal to var$(U) = mn(m + n + 1)/12$.

The crux of this proof, as mentioned in the question, is to consider cov(I_{ij}, I_{ik}), and cov(I_{ij}, I_{kl}).

First, let us note that $\text{Var}(I_{ij}) = \frac{1}{2}\frac{1}{2} = \frac{1}{4}$.

Now, let us first find $\text{cov}(I_{ij}, I_{ik})$ for distinct i, j, k, l:

$$\begin{aligned}
\text{Cov}(I_{ij}, I_{ik}) &= E(I_{ij}, I_{ik}) - E(I_{ij})E(I_{ik}) \\
&= E(X_i < Y_j, X_i < Y_k) - E(X_i < Y_j)E(X_i < Y_k) \\
&= P(X_i < Y_j, X_i < Y_k) - P(X_i < Y_j)P(X_i < Y_k)
\end{aligned}$$

Let us compute the first term on the right hand side, $P(X_i < Y_j, X_i < Y_k)$:

$$\begin{aligned}
P(X_i < Y_j, X_i < Y_k) &= P(X_i < Y_j | Y_j < Y_k)P(Y_j < Y_k) + P(X_i < Y_k | Y_k < Y_j)P(Y_k < Y_j) \\
&= \frac{P(X_i < Y_j < Y_k)}{P(Y_j < Y_k)}P(Y_j < Y_k) + \frac{P(X_i < Y_k < Y_j)}{P(Y_k < Y_j)}P(Y_k < Y_j) \\
&= P(X_i < Y_j < Y_k) + P(X_i < Y_k < Y_j) \\
&= \frac{1}{3!}\frac{1}{3!} \\
&= \frac{1}{3}
\end{aligned}$$

Where we know that there are $3! = 6$ unique ways to arrange $X_i, Y_k,$ and Y_j.

Substituting this back in, we get:

$$\begin{aligned}
\text{Cov}(I_{ij}, I_{ik}) &= P(X_i < Y_j, X_i < Y_k) - P(X_i < Y_j)P(X_i < Y_k) \\
&= \frac{1}{3} \cdot \frac{1}{2} \cdot \frac{1}{2} \\
&= \frac{1}{12}
\end{aligned}$$

Now, let us find $\text{cov}(I_{ij}, I_{kl})$ for distinct i, j, k, l:

$$\begin{aligned}
\text{Cov}(I_{ij}, I_{kl}) &= E(I_{ij}, I_{kl}) - E(I_{ij})E(I_{kl}) \\
&= E(X_i < Y_j, X_k < Y_l) - E(X_i < Y_j)E(X_k < Y_l) \\
&= P(X_i < Y_j, X_k < Y_l) - P(X_i < Y_j)P(X_k < Y_l) \\
&= 0 \quad \text{because } X_i < Y_j \text{ and } X_k < Y_l \text{ are independent events}
\end{aligned}$$

Next, we can calculate the variance, $\text{Var}(U)$. First, we must obtain the number of pairs there are of $\text{Cov}(I_{ij}, I_{ik})$. Since k must be distinct from i and j, and $i \in \{1, \ldots, m\}, j \in \{1, \ldots, n\}$, and $k \in \{1, \ldots, m-1\}$ or $k \in \{1, \ldots, n-1\}$, then, there must be a total of $mn(n-1) + mn(m-1)$ pairs.

Thus, the variance can be computed as:

$$\begin{aligned}
\text{Var}(U) &= \sum_{i=1}^{m}\sum_{j=1}^{n} \text{Var}(I_{ij}) + \text{Cov}(I_{ij}, I_{ik}) + \text{Cov}(I_{ij}, I_{kl}) \\
&= \frac{1}{4} \cdot mn + \frac{1}{12} \cdot mn(n-1) + mn(m-1) + 0 \\
&= \frac{mn}{4} + \frac{mn(n-1) + mn(m-1)}{12} \\
&= \frac{3mn}{12} + \frac{mn(n-1) + mn(m-1)}{12} \\
&= \frac{3mn + mn(n-1) + mn(m-1)}{12} \\
&= \frac{3mn + mn^2 - nm + m^2n - mn}{12} \\
&= \frac{mn^2 + m^2n + mn}{12} \\
&= \frac{mn(n+m+1)}{12}
\end{aligned}$$

Thus, we have shown the desired results, $E(U) = mn/2$ and $\text{Var}(U) = mn(m+n+1)/12$.

b) Test H_0 assuming that U is approximately normally distributed using the data of the weight gains in grams of animals fed on high and low protein diets.

```
library(knitr)
```

```
## Warning: package 'knitr' was built under R version 4.0.5
```

```
# first enter the data
u <- 0
high <- c(83, 97, 104, 107, 113, 119, 123, 124, 129, 134, 146, 161)
low <- c(70, 85, 94, 101, 106, 118, 132)
m <- length(high)
n <- length(low)

# for loop to calculate U
for(i in 1:m){
  for (j in 1:n){
    if(high[i]<low[j]){
      u = u+1
    }
  }
}

# expected value
expected_u <- (m*n)/2

# variance
variance_u <- (m*n*(m+n+1))/12

# calculate p-val
z <- (u-expected_u)/sqrt(variance_u)

# two-sided
```

```
pval <- 2 * pnorm(z)
kable(pval, col.names = "P-Value")
```

P-Value
0.075927

Here we see that the calculated p-value is equal to 0.076, which is large, thus we fail to reject the null hypothesis $H_0 : F_X = F_Y$. We do not have sufficient evidence to show that the two distributions are the same.

c) Would the test based on U be more powerful if the experimenter had ensured equal numbers of animals in each group, with the same number of total animals? Explain.

Yes, the test based on U would be more powerful if the experimenter had ensured equal numbers of animals in each group, with the same number of total animals. In general, the statistical power of a hypothesis test that compares groups is highest when we have equal sample sizes. This is because power is based on the smallest sample size.

If you have a specified number of units (in this case, animals) to randomly assign to groups, you will have higher power if you assign them equally. If you add more to the larger group, this won't necessarily help increase the power of the test, but would not hurt it either.

d) The null hypothesis can also be tested using the Cramer-vonMises test statistic

$$T_{m,n} = \frac{mn}{m+n} \int \{\hat{F}_{X,m}(x) - \hat{F}_{Y,n}(x)\}^2 d\hat{F}_{X+Y,m+n}(x),$$

where $\hat{F}_{X,m}(x)$, $\hat{F}_{Y,n}(x)$ are the empirical cumulative distribution functions for the X and Y samples, respectively, and

$$\hat{F}_{X+Y,m+n}(x) = \frac{m}{m+n}\hat{F}_{X,m}(x) + \frac{n}{m+n}\hat{F}_{Y,n}(x)$$

is the empirical cumulative distribution function based on the pooled sample $(X_1, \ldots, X_m, Y_1, \ldots, Y_n)$.

i. Compute $T_{m,n}$ for the data in (b), treating the "high" group as X's and the "low" group as Y'x.

To compute $T_{m,n}$ for the data in (b), I will use the package `twosamples` with the function cvm_stat, which provides a two-sample test based on the Cramer-Von Mises test statistic.

```
library(twosamples)

# compute the test statistic T_{m,n}
cvm_stat(high,low)
```

```
## [1] 1.510629
```

Here we can see that the calculated test statistic is equal to 1.51.

Note: I also tried a different package called CvM2SL2Test: Cramer-von Mises Two Sample Tests. This gave a much smaller value, 0.352.

```
library(CvM2SL2Test)

cvmts.test(high,low)
```

```
## [1] 0.3515038
```

ii. Compute the bootstrap distribution of $T_{m,n}$, where each bootstrap sample involves resampling (with replacement) m X^*'s from the first group and n Y^*'s from the second group. Compare the p-value from the bootstrap to that in (b). The p-value can be computed using any R package that does this test. April 10: A more appropriate bootstrap would be to resample all $m+n$ observations at random with replacement, and then arbitrarily assign m of them to be X^*s and the other n to be Y^*s. This will ensure that we are resampling under the null hypothesis. However, either bootstrap method will be accepted; if you have already completed this question you don't need to re-do it.

The `twosamples` package has a built-in function, `cvm_test`, which does a permutation based two sample Cramer-Von Mises test. The `nboots` parameter lets us specify a number of bootstrap iterations. Let us try with `nboots`=2000.

```
# twosamples has a built in function for bootstrap
cvm_test(high,low, nboots=2000)
```

```
## Test Stat   P-Value
##  1.510629  0.030000
## attr(,"details")
##      n1      n2 n.boots
##      12       7    2000
```

Here we see that the p-value is 0.0305, which is smaller than our calculated p-value from part b (0.076). Since this p-value is smaller than 0.05, we would in this case reject the null hypothesis.

Note: I also tried calculating this using the CvM2SL2Test as above. This time, I computed the bootstrap samples, where each bootstrap sample involves resampling (with replacement) m X^*'s from the first group and n Y^*'s from the second group.

```
set.seed(8)

# bootstrap sample
for(i in 1:2000){
  Xstar <- sample(high,m,replace=TRUE)
  Ystar <- sample(low,n,replace=TRUE)
}

# recompute the test stat and p-value - CvM2SL2Test
cvm <- cvmts.test(Xstar,Ystar)
cvmts.pval(cvm, m,n)
```

```
## [1] 0.2049694
```

This gives us a much larger p-value, 0.2049. This is much higher than the p-value found in part (b). Note that this package has been removed from CRAN, and I am using a past version of it. For the purpose of this assignment question, results from `twosample` package seem more reasonable.