Masco: follow-along with R

Academic year 2025-2026

2025-11-14

Testing the equality of two means of normal populations

In this exercise, we explore methods for comparing the means of two populations under the assumption of normality. We will use the iris dataset, which contains measurements of sepal and petal dimensions for three species of iris flowers. The goal is to visualize and statistically compare these groups to assess whether their mean values differ significantly.

```r
library(datasets)
data(iris)

X = iris$Sepal.Length[iris$Species == "versicolor"]
Y = iris$Sepal.Length[iris$Species == "virginica"]
Z = iris$Sepal.Length[iris$Species == "setosa"]
```

1. Plot the distributions of all four iris measurements (Sepal.Length, Sepal.Width, Petal.Length, and Petal.Width) by species using boxplots.
2. Check the normality of the populations X, Y, and Z.
3. Test whether X and Y have the same variance, and do the same for X and Z.
4. Perform a t-test to compare the populations X and Y, then X and Z, and comment on your findings.
5. Explain how the 95% confidence interval provided by the function `t.test` is computed.

Bayesian two-sample test

We now apply a Bayesian approach to compare the populations `X` and `Y` from the previous exercise.

Sampling model.

$$\begin{aligned} x_{1,i} \mid \mu_1, \sigma_1^2 &\sim \mathcal{N}(\mu_1, \sigma_1^2), \\ x_{2,i} \mid \mu_2, \sigma_2^2 &\sim \mathcal{N}(\mu_2, \sigma_2^2). \end{aligned}$$

Prior specification.

$$\begin{aligned} \mu_1 &\sim \mathcal{N}(0, 1000), \\ \mu_2 &\sim \mathcal{N}(0, 1000), \\ \sigma_1^{-2} &\sim \mathcal{G}amma(0.001, 0.001), \\ \sigma_2^{-2} &\sim \mathcal{G}amma(0.001, 0.001). \end{aligned}$$

1. Discuss the modeling choices described above.
2. Using `Rjags`, compute the posterior distribution of the difference between the means.
3. Provide a 90% credible interval for this difference and comment on the result.

Nonparametric Wilcoxon test

Let F_X, F_Y be continuous cumulative distribution functions. We want to test whether $H_0 : F_X = F_Y$ against $H_1 : F_X \neq F_Y$, nevertheless, no test can be very powerful for such a very general alternative.

The *Wilcoxon* test for location alternative hypothesis $H1 : F_Y(x) = F_X(x - \theta), \theta \in \mathbb{R}^*$. In other words, the alternative is that the median of the two populations differs. Note that under the null $\text{median}(F_X) = \text{median}(F_Y)$.

Consider an i.i.d sample $X_1, \cdots, X_m$ (resp. $Y_1, \cdots, Y_n$) with CDF F_X (resp. F_Y), then take the global sample

$$X_1, X_2, \cdots, X_m, Y_1, Y_2, \cdots, Y_n.$$

Let $r(X_i)$ be the **rank** of X_i in the ordered global sample. For example if $X_2 < Y_1 < X_n < \cdots$ then $r(X_2) = 1, r(Y_1) = 2$ and $r(X_n) = 3, \cdots$. Note that $r(X_i)$ is a random variable.

Testing procedure

- Sort the two samples in increasing order while keeping track of the original sample.
- Give a rank to each observation, beware for ties.
- Compute the test statistic $T_W = \min(T_X, T_Y)$, where $T_X = \sum_{k=1}^{m} r(X_k)$ and $T_Y = \sum_{k=1}^{n} r(Y_k)$.

Some details

$$P(r(X_i) = k) = \frac{1}{m+n}, \; k = 1, \cdots, m+n.$$

$$\mathbb{E}[r(X_i)] = \sum_{k=1}^{m+n} k\frac{1}{m+n} = \frac{m+n+1}{2}.$$

$$\text{Var}(r(X_i)) = \frac{(m+n)^2}{12}.$$

Recall that $T_X = \sum_{k=1}^{m} r(X_i)$, then

$$\mathbb{E}[T_X] = m\frac{m+n+1}{2},$$

$$\text{Var}(T_X) = \frac{mn(m+n+1)}{12},$$

$$T_X + T_Y = \frac{(m+n)(m+n+1)}{2}.$$

We want to use this test to compare between the effectiveness of various feed supplements on the growth rate of chickens.

```r
library(datasets)
data("chickwts")
#help("chickwts")

X = chickwts$weight[chickwts$feed == "soybean"]
Y = chickwts$weight[chickwts$feed == "horsebean"]
```

Questions

1. Using the Central Limit Theorem (CLT), provide an approximation of the distribution of T_X.
2. Under the null hypothesis, does the distribution of T_X depend on F_X and F_Y?
3. Implement the nonparametric Wilcoxon test.
4. Using your implemented function, compare the weights of chicks fed with `soybean` and `horsebean`.
5. Compare your results with those obtained using the built-in `wilcox.test` function.
6. Compare the weights of chicks fed with `soybean` and `linseed`. What conclusions can you draw?
7. In the same figure, plot the boxplots for all feed types. How can you compare all feed types at once?