Additional Hypothesis Tests

Sean Hellingman ©

Introduction to Statistical Data Analysis (ADSC1000)

shellingman@tru.ca

Fall 2024

Topics

Introduction

- We are continuing to make statistical inferences about target populations.
- We are going to cover a two more statistical tests:
  - Two-Sample Kolmogorov-Smirnov Test
  - Mann-Whitney U test
- Note: *The formulas are more complicated so we will focus on the R applications.*

Two-Sample Kolmogorov-Smirnov Test

Two-Sample K-S Test

- **K-S test** is a non-parametric test for the equality of continuous distributions.
- One-sample tests can be used to see if samples are drawn from known distributions.
  - `ks.test(sample.data, "pnorm")`
- We can use two-sample tests to see if samples are drawn from the same distribution.
- The test is based on the maximum difference between the empirical distribution functions (similar to CDF) of the samples.

Example 1

- Run the example code to generate three random samples.
- Generate the ECDF plot.
- What do you notice?

Two-Sample K-S Test in R

- To perform the test in R:

```
ks.test(Sample1, Sample2)
```

- Null Hypothesis: *The samples are drawn from the same distribution.*
- Alternative Hypothesis: *The samples are not drawn from the same distribution*

Example 2

1. Use the K-S test to determine if the samples from Example 1 are drawn from the same distribution.

2. Import the *Wages.csv* dataset into R to determine if `Salary` and `Salary.5` are drawn from the same distribution.

Mann-Whitney U test

Mann-Whitney U test

- The **Mann-Whitney U test** is a non-parametric equivalent to the two-sample t-test.
- Used to compare the distributions of two independent samples (numeric response variable).
- This test is based on the sum of ranks of the elements.

Mann-Whitney U test in R

- To perform the test in R:

```
wilcox.test(Response ~ Group)
```

- Null Hypothesis: *The distributions of both populations are identical.*
- Alternative Hypothesis: *The distributions of both populations are **not** identical.*

Example 3

- Using the *Wages.csv* dataset do the following:
  1. Is the distribution of `Salary` the same between *Metro* and *Rural* areas?
  2. Is the distribution of `Salary` the same between *Data Scientists* and *Biologists*?
  3. Can you think of any situations where this test may be useful?

Exercise 1

- Load the *Cars93* dataset from the *MASS* R package to do the following:

  1. Are the *Min.Price* and the *Max.Price* drawn from the same distribution?

  2. Are the *MPG.city* and the *MPG.highway* drawn from the same distribution?

     - *Feel free to use visualizations to support your conclusions.*

Exercise 2

- Load the *Cars93* dataset from the *MASS* R package to do the following:

  1. Is the distribution of the *Price* the same for cars with *4* Cylinders and cars with *6* Cylinders?

  2. Is the distribution of the *Price* the same for cars from the USA or non-USA (*Origin*)?

     - *Feel free to use visualizations (histograms) to support your conclusions.*

Exercise 3

- Load the *Cars93* dataset from the *MASS* R package and use the Chi-squared tests to test the following:

  1. If the horsepower *(Horsepower)* is independent of the vehicle type (*Type*).
  2. If the highway miles per gallon (*MPG.highway*) is independent of the number of the number of airbags *(AirBags)*
  3. If the price (*Price*) is independent of the number of airbags *(AirBags)*.

References & Resources

- ks.test()
- Mann Whitney U