Hypothesis Testing III

Sean Hellingman ©

Introduction to Statistical Data Analysis (ADSC1000)

shellingman@tru.ca

Fall 2024

THOMPSON RIVERS UNIVERSITY

Topics

2. Introduction
3. Tests for Normality
4. Kolmogorov–Smirnov test
5. Shapiro–Wilk test
6. ANOVA
7. Independence
8. Exercises and References

Introduction

- We are continuing to make statistical inferences about target populations.
- We are going to cover a few more statistical tests:
  - Tests for normality of data.
  - Differences in several means (ANOVA).
  - Chi-squared test for independence.
- Note: *The formulas are more complicated so we will focus on the R applications.*

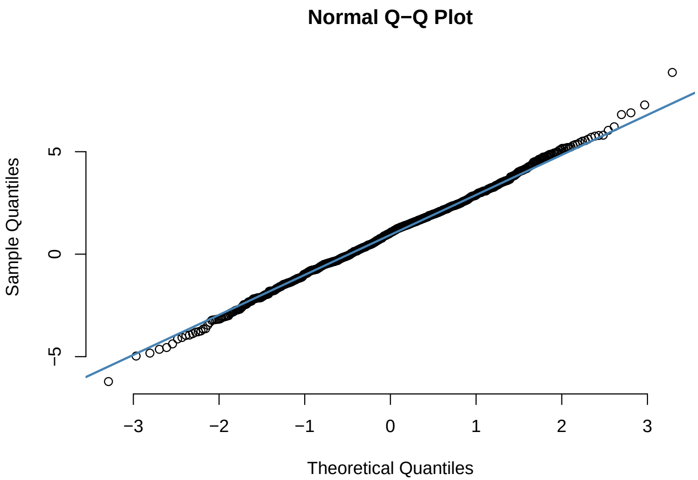
Normality Assumptions

- Some statistical tests are only valid when assumptions about the distribution hold.
- Data being drawn from a Normal distribution is common assumption for many tests.
- Some methods to determine if our data is normally distributed or not:
    - Quantile-Quantile plots (not a formal test).
    - Kolmogorov–Smirnov test.
    - Shapiro-Wilk test of normality.

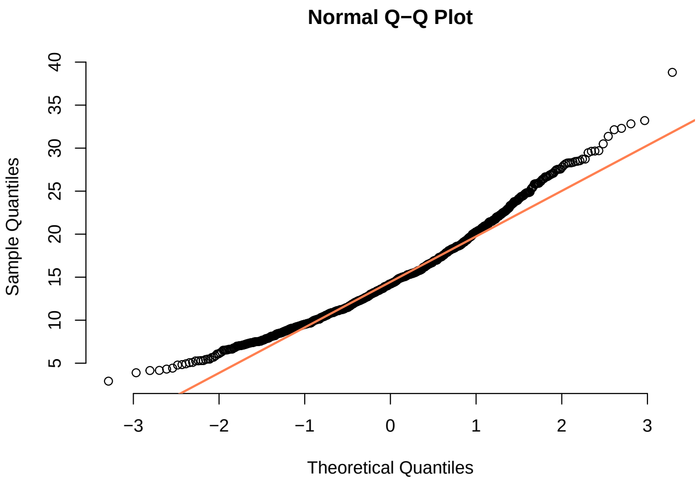
Quantile-Quantile (Q-Q) plots

- **(Q-Q) plot** is a visualisation method to determine if two distributions are the same.
- Use the theoretical quantiles from a normal distribution and plot them against the quantiles from the sample.
- If the two distributions are identical the Q-Q plot will follow the 45^o line.
- Otherwise we can conclude that the distributions are different (not normally distributed).

Q-Q Normal


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
−5
0
5
−3
−2
−1
0
1
2
3

Q-Q Not Normal



Quantile-Quantile (Q-Q) plots in R

- We can use base R or *ggplot2* to generate Q-Q plots.

- Base R:
  - `qqnorm(sample.data)` To create scatterplot
  - `qqline(sample.data, col = "color")` To add quantile line

- *ggplot2*:
  - ```
    ggplot(data = data.frame, aes(sample = sample.data)) +
    stat_qq() +
    stat_qq_line()
    ```
  - *You can add additional layers to make your plot more informative*
    ```

Example 1

- Load the *Cars93* dataset from the *MASS* R package and use (Q-Q) plots to determine if any of the following variables are normally distributed:

  1. *Price*
  2. *Horsepower*
  3. *Width*
  4. *Weight*

Kolmogorov–Smirnov Test

- **K-S test** is a non-parametric test for the equality of continuous distributions.
- We will use the one-sample test to compare our sample with a reference probability distribution (normal distribution).
- The test is based on the maximum difference between the empirical distribution function (similar to CDF) of the sample and the cumulative distribution function (CDF) of the reference distribution.
- If there is a large enough difference we can assume that the sample does not come from the reference distribution.

One Sample K-S Test

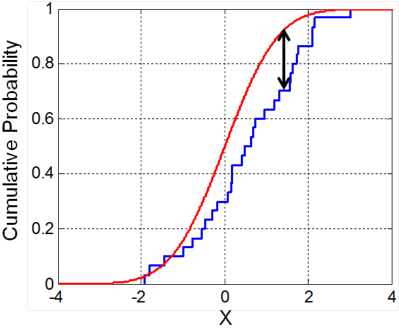


Figure: *Source: (3)*

One Sample K-S Test in R

- The null hypothesis for this test is that the data come from a normal distribution.
- R code: `ks.test(sample.data, "pnorm")`
  - A small p-value indicates that our data is not from a normal distribution.
- *Sensitive to sample sizes.*

Example 2

- Load the *Cars93* dataset from the *MASS* R package and use One Sample K-S tests to determine if any of the following variables are normally distributed:

    1. *Price*
    2. *Horsepower*
    3. *Width*
    4. *Weight*

Shapiro–Wilk Test

- **Shapiro–Wilk test** is based on the ordered sample.
- The test statistic is complicated to calculate so we will use R.
- The null hypothesis is that the population is normally distributed.
- If we obtain a small p-value we can reject the null hypothesis of normality.

Shapiro–Wilk Test in R

- **The null hypothesis for this test is that the data come from a normal distribution.**

- R code: `shapiro.test(sample.data)`
  - A small p-value indicates that our data is not from a normal distribution.

- *Sensitive to sample sizes.*
- **Generally the preferred method for determining normality**

Example 3

- Load the *Cars93* dataset from the *MASS* R package and use Shapiro–Wilk Tests to determine if any of the following variables are normally distributed:

  1. *Price*
  2. *Horsepower*
  3. *Width*
  4. *Weight*

Testing Differences in Several Means

- Sometimes we would like to compare the means of multiple groups for equality.
    - Useful when examining different factors in results from designed experiments.

- Instead of doing multiple pairwise tests we use **analysis of variance (ANOVA)** tools.

- General hypothesis for m groups:

  H_0: $\mu_1 = \mu_2 = \ldots\ \mu_m$
  H_1: At least one mean is different from the others

ANOVA

- ANOVA derives its name from the fact that we are analyzing variances in the data.
- Basically, ANOVA computes a measure of the variance between the means of each group, and a measure of the variance within the groups.
- If the null hypothesis (same means) is true, the between-group variance should be small.
- Under the null $F = \textit{Mean Square between groups}/\textit{Mean Square within groups}$ follows an F-Distribution (similar to the test for equality in variances).
  - Large ratio implies that at least one mean is different (H_1)

ANOVA Assumptions

1. Observations are randomly and independently obtained.
2. Observations are normally distributed.
   - We have covered tests needed to validate this assumption.
3. Observations have equal variances.
   - We can use a **Bartlett's test** (among others) to validate this assumption.

Bartlett's Test in R

- Observations **must** be normally distributed.
- The null hypothesis is that the variances are the same
  - A small p-value indicates that at least one variance is different.
- In R: `bartlett.test(value ~ group, data = data.frame)`
- *We can also use a Levene's test for non-normal data.*

ANOVA Test in R

- After checking the assumptions we can conduct an ANOVA test:
- In R:
  - `oneway.test(value ~ group, data = data.frame, var.equal = TRUE)`
- **If the equal variances assumpton is violated we can use:**
  - `oneway.test(value ~ group, data = data.frame, var.equal = FALSE)`
  - (Welch ANOVA)

Example 4

- Load the *iris* dataset from base R and follow the steps to perform an ANOVA test to determine if there are differences in the *Petal.Length* of the three *Species*:

  1. Validate the normality assumption of the three groups.
  2. Validate the equal variance assumption of the three groups.
  3. Perform an appropriate ANOVA test and comment on your results.

ANOVA Comments

- The ANOVA test requires a strong set of assumptions.
- Usually we can visually see which mean(s) are different but there are more formal tests for determining which mean(s) are different:
  - Tukey-Kramer test
  - Kruskal-Wallis test:

```
kruskal.test(value ~ group, data = data.frame)
```

Independence of Categorical Variables

- Sometimes we want to assess the independence of categorical variables.
- Example: If we are examining the laptop preferences of students at TRU and UBC-O.
  - Are the preferences independent of which university the student attends?
- Sampling error can make it difficult to properly assess independence of categorical variables.
- We can do this using the **chi-squared test** for independence.

Chi-squared Test

- The **chi-squared test** test is used to examine whether two categorical variables are independent.
  - The test is generally applied to a two-dimensional frequency table.
- **The null hypothesis is that the two categorical variables are independent.**
- Under the null hypothesis we expect that the same proportions for each group exist across the other group.
  - *We would expect the proportions of laptop preferences to be roughly the same across the two universities.*
- The procedure uses the observed and expected frequencies to compute a test statistic.

Chi-squared Test Formulation

- The test statistic for the Chi-squared test:

$$\chi^2 = \sum \frac{(O - E)^2}{E} \ . \tag{1}$$

- Where O is the observed frequency and E is the expected frequency.
- *To compute the expected frequency for a particular cell in the table, simply multiply the row total by the column total and divide by the grand total:*
    - $E = \frac{\text{row sum} \cdot \text{column sum}}{\text{grand total}}$

Chi-squared Test in R

- Chi-squared tests with Yates' continuity correction:

1.
```
chisq.test(data.frame$numeric.variable,
    data.frame$factor.variable)
```

2.
```
chisq.test(frequency.table.matrix)
```

Example 5

- 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 availability of a manual transmission (*Man.trans.avail*).
    2. If the highway miles per gallon (*MPG.highway*) is independent of the number of the number of cylinders *(Cylinders)*
    3. If the price (*Price*) is independent of the availability of a manual transmission (*Man.trans.avail*).

Exercise 1

- Load the *Cars93* dataset from the *MASS* R package and follow the steps to perform an ANOVA test to determine if there are differences in the *Horsepower* of the three *Cylinders*:

    1. Validate the normality assumption of the three groups.
    2. Validate the equal variance assumption of the three groups.
    3. Perform an appropriate ANOVA test and comment on your results.

Exercise 2

- Load the *Cars93* dataset from the *MASS* R package and follow the steps to perform an ANOVA test to determine if there are differences in the *Price* of the three *Type*:

  1. Validate the normality assumption of the three groups.
  2. Validate the equal variance assumption of the three groups.
  3. Perform an appropriate ANOVA test and comment on your results.

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

1. Evans, J. R., Olson, D. L., & Olson, D. L. (2007). *Statistics, data analysis, and decision modeling.* Upper Saddle River, NJ: Pearson/Prentice Hall.

2. Devore, J. L., Berk, K. N., & Carlton, M. A. (2012). *Modern mathematical statistics with applications (Second Edition).* New York: Springer.

3. By Bscan - Own work, CC0, https://commons.wikimedia.org/w/index.php?curid=25222928

- https://ggplot2.tidyverse.org/reference/geom_qq.html
- https://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test
- http://www.sthda.com/english/wiki/compare-multiple-sample-variances-in-r
- http://www.sthda.com/english/wiki/chi-square-test-of-independence-in-r