LAB: TESTS FOR NORMAITY AND EQUAL VARIANCE

Lab 5.2

PART 1. NORMALITY TEST

NORMALITY

- Many statistical procedures such as correlation, regression, t-tests, and ANOVA, namely parametric tests, are based on the normal distribution of data.

Properties of the normal distribution:
Bell-shaped
Symmetrical
Unimodal — it has one "peak"
Mean and median are equal; both are located at the center of the distribution

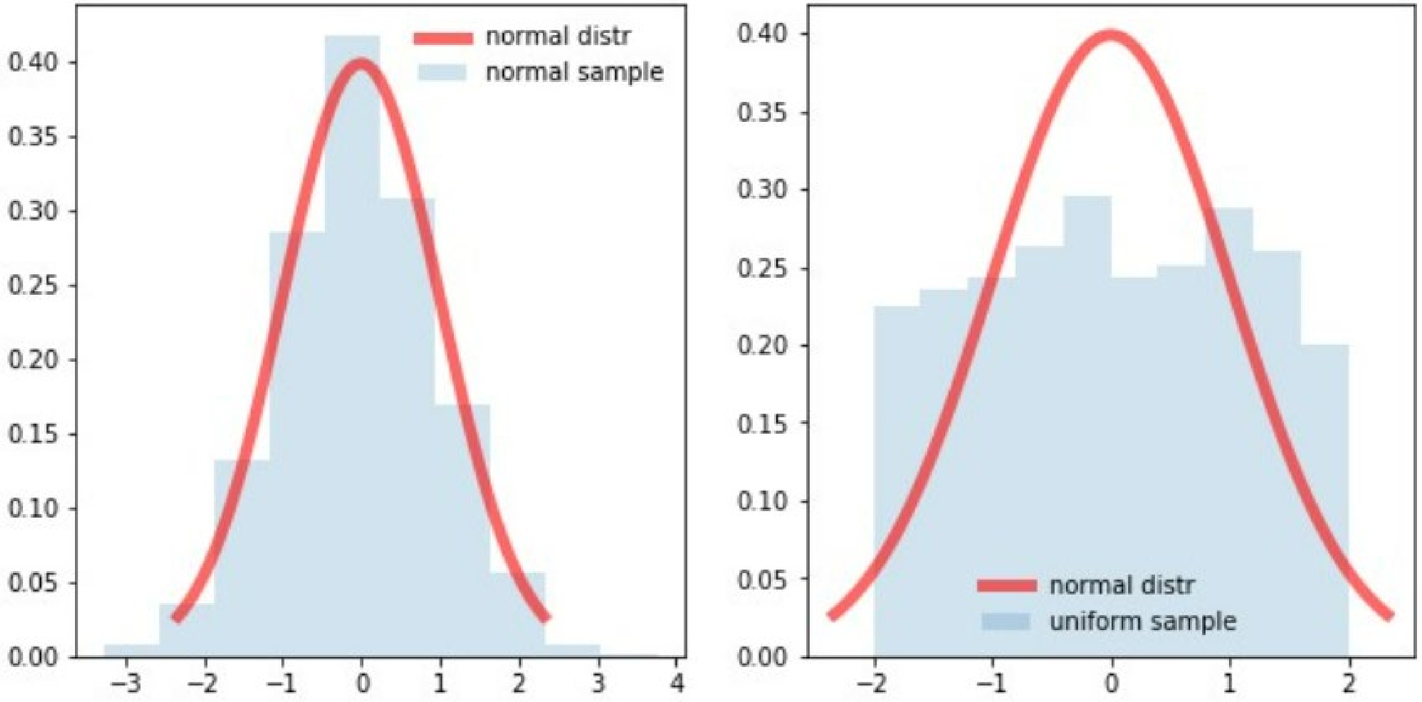
Methods for normality test

- Normality test
  - 1.Graphic
    - Histogram
    - Q-Q
    - P-P
  - 2.Analytical
    - **Shapiro-Wilk normality test**
    - **Kolmogorov-Smirnov Test**

1.1 Histogram

- Y axis: the number of times that the values occurred within the intervals set by the X-axis.

- X axis: intervals that show the scale of values which the measurements fall under.



Bell-shape

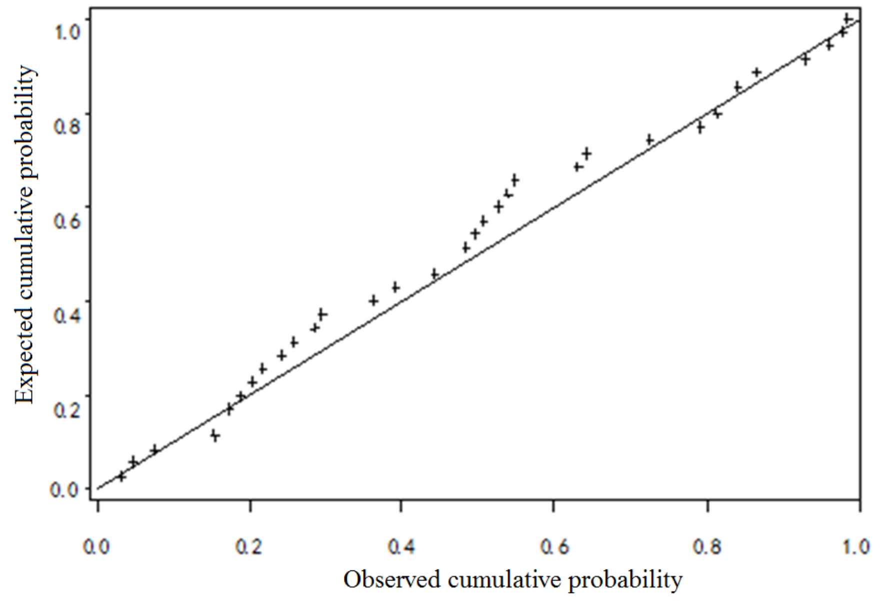
Normal (left) vs. non-normal distribution. The red curve represents an ideal normal (Gaussian) distribution.

1.2 Probability-probability plot (P-P plot)

➢ One axis: the cumulative probability of actual observation values

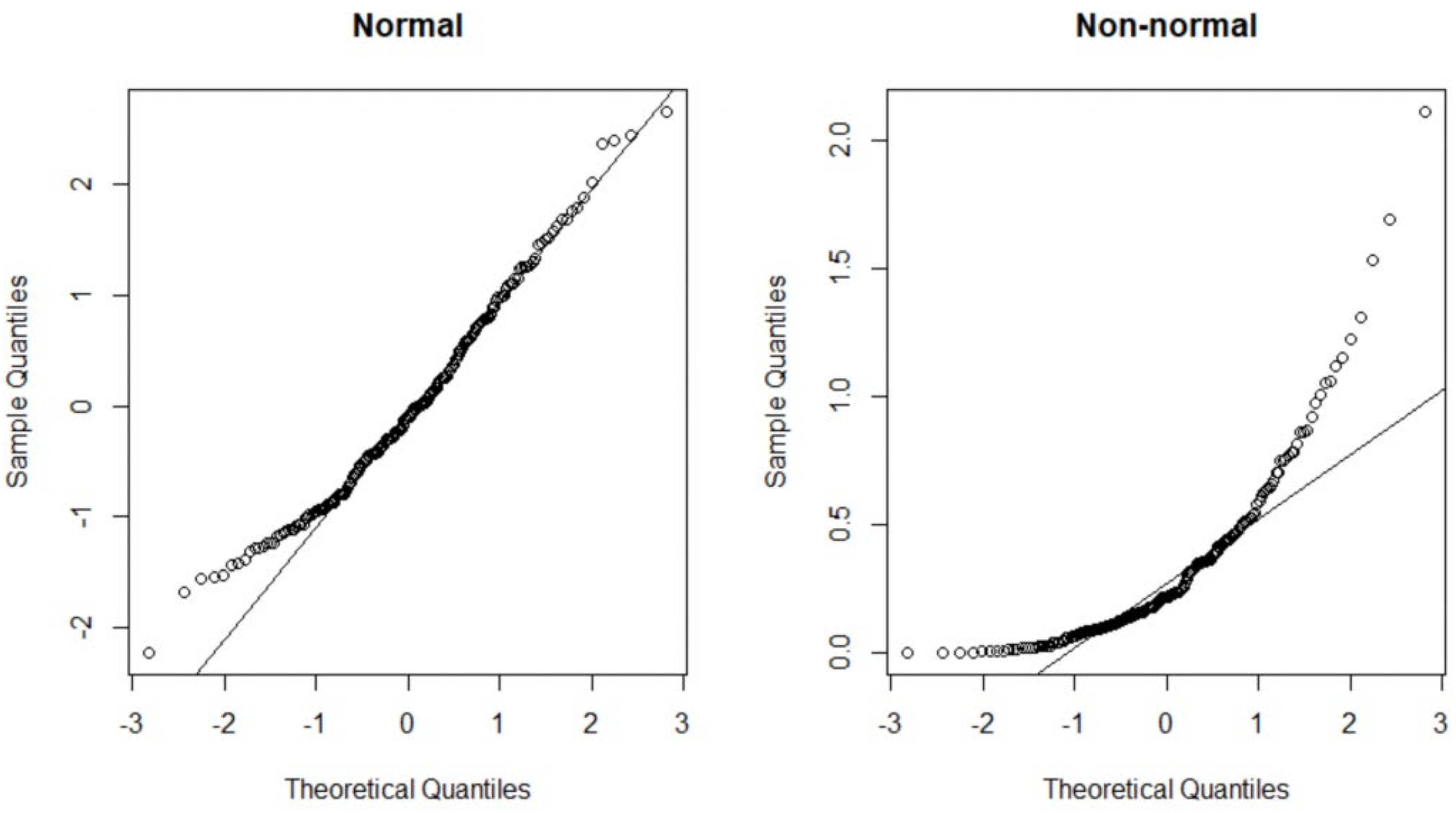
➢ Another axis: the expected/theoretical cumulative probability based on the normal distribution.

➢ A normal distribution means that sample points are distributed around the diagonal of the first quadrant.

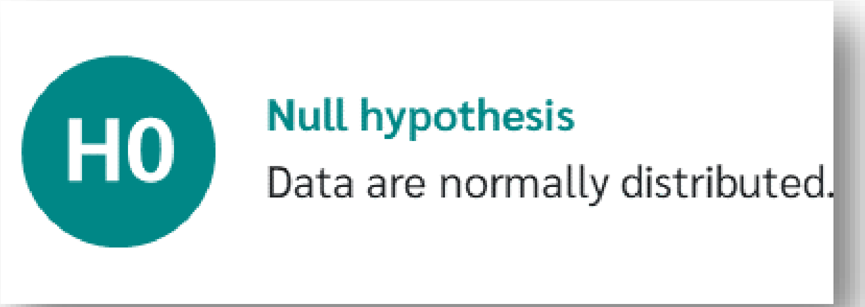


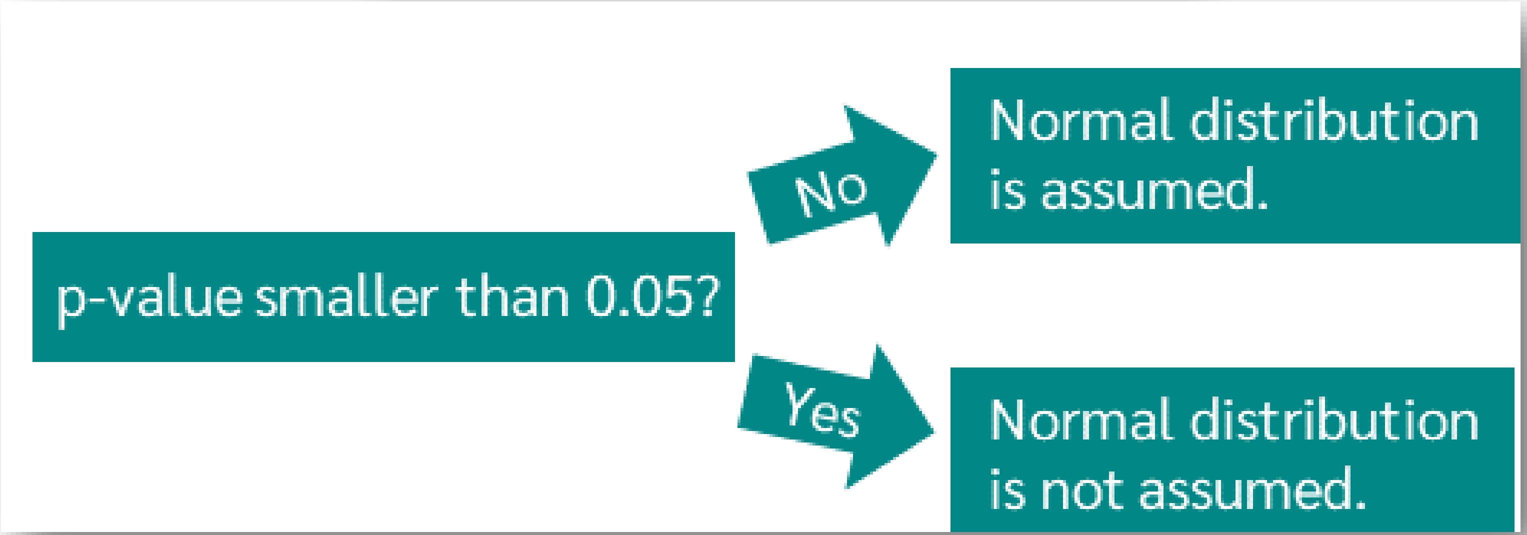
1.3 Quantile-quantile plot (Q-Q plot)

- **One axis: the quartile of sample data.**
- **Another axis: the expected/theoretical quartile based on the normal distribution.**
- **A normal distribution means that sample points are distributed around the diagonal of the first quadrant.**
- **Q-Q plot is more widely used than P-P plot in practice.**


Normal
Sample Quantiles
Theoretical Quantiles
-2
-1
0
1
2
-3
-2
-1
0
1
2
3
Non-normal
Sample Quantiles
Theoretical Quantiles
0.0
0.5
1.0
1.5
2.0
-3
-2
-1
0
1
2
3

2. Hypothesis testing methods





2.1 Shapiro-Wilk test

- **Known as W test and introduced by S.S.Shapiro and M.B.Wilk;**
- **Shapiro-Wilk Original Test is <u>suitable for sample sizes in the range of 3 to 50</u>;**
- **Shapiro-Wilk Expanded Test: a revised approach using the algorithm of J. P. Royston which can handle samples with up to 5,000 (or even more)**

[Shapiro-Wilk Royston Test | Real Statistics Using Excel (real-statistics.com)](real-statistics.com)

Basic approach of Shapiro-Wilk original test

① **Arrange the data in ascending order: x1≤x2 ≤x3….. ≤xn**

② **calculate SS** $SS = \sum_{i=1}^{n}(X_i - \overline{X})$

③ **If n is even, let m = n/2, while if n is odd let m = (n–1)/2**

④ **Calculate b and the <u>test statistic W</u> as follows**

$$b = \sum_{i=1}^{n} a_i(x_{n+1-i} - x_i),\ W = {b^2}/{SS}$$

$$W = \frac{\left[\sum_{i=1}^{n/2} a_i(x_{n+1-i} - x_i)\right]^2}{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}$$

① **taking the ai weights from <u>Shapiro-Wilk Tables</u> (for a given value of n)**

that is closest to W, interpolating if necessary.

2.2 Kolmogorov-Smirnov test

- **Suitable for sample sizes in the range of 50 to 1000;**
- **The formula for the test statistic is:**

$$Y = \frac{\sqrt{n}(D - 0.28209479)}{0.02998598}$$

in which

$$D = \frac{\sum_{i=1}^{n}\left(i - \frac{n+1}{2}\right)x_i}{(\sqrt{n})^3\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

[Kolmogorov-Smirnov Normality | Real Statistics Using Excel (real-statistics.com)](real-statistics.com)

3. Common transformations for non-normal data

- ✓ Square-root for moderate skew:

  sqrt(x) for positively skewed data,

  sqrt(max(x+1) - x) for negatively skewed data

- ✓ Log for greater skew:

  log10(x) for positively skewed data,

  log10(max(x+1) - x) for negatively skewed data

- ✓ Inverse for severe skew:

  1/x for positively skewed data

  1/(max(x+1) - x) for negatively skewed data

PART 2. EQUAL VARIANCE TEST

Equal variances test

The most common statistical tests and procedures that make this assumption of equal variance include:

① ANOVA
② t-test
③ Linear regression

H0 $\sigma_1^2 = \sigma_2^2 = ... = \sigma_n^2$

1. F test

Testing whether two population variances are equal

$$F = \frac{s_1^2\left(bigger\right)}{s_2^2\left(smaller\right)}$$

$$v_1 = n_1 - 1, \quad v_2 = n_2 - 1$$

- ***Sensitive to departures from normality***

2. Levene's test

H$_0$: $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$

H$_a$: $\sigma_i^2 \neq \sigma_j^2$ for at least one pair (i,j).

Critical Region: The Levene test rejects the hypothesis that the variances are equal if

$$W > F_{\alpha,\, k-1,\, N-k}$$

where $F_{\alpha,\, k-1,\, N-k}$ is the upper critical value of the *F* distribution with k-1 and N-k degrees of freedom at a significance level of α.

- **Testing whether two or more population variances are equal**
- Sensitive to departures from normality

$$W = \frac{(N-k)}{(k-1)} \frac{\sum_{i=1}^{k} N_i (\bar{Z}_{i.} - \bar{Z}_{..})^2}{\sum_{i=1}^{k} \sum_{j=1}^{N_i} (Z_{ij} - \bar{Z}_{i.})^2}$$

where Z_{ij} can have one of the following three definitions:

1. $Z_{ij} = |Y_{ij} - \bar{Y}_{i.}|$

   where $\bar{Y}_{i.}$ is the mean of the i-th subgroup.

2. $Z_{ij} = |Y_{ij} - \tilde{Y}_{i.}|$

   where $\tilde{Y}_{i.}$ is the median of the i-th subgroup.

3. $Z_{ij} = |Y_{ij} - \bar{Y}'_{i.}|$

   where $\bar{Y}'_{i.}$ is the 10% trimmed mean of the i-th subgroup.

3. Bartlett's test

The Bartlett test is defined as:

H_0: $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$

H_a: $\sigma_i^2 \neq \sigma_j^2$ for at least one pair (i,j).

Critical Region: The variances are judged to be unequal if,

$$T > \chi^2_{1-\alpha,\, k-1}$$

where $\chi^2_{1-\alpha,\, k-1}$ is the critical value of the chi-square distribution with k - 1 degrees of freedom and a significance level of α.

$$T = \frac{(N-k)\ln s_p^2 - \sum_{i=1}^{k}(N_i-1)\ln s_i^2}{1 + (1/(3(k-1)))((\sum_{i=1}^{k} 1/(N_i-1)) - 1/(N-k))}$$

In the above, s_i^2 is the variance of the ith group, N is the total sample size, N_i is the sample size of the ith group, k is the number of groups, and s_p^2 is the pooled variance. The pooled variance is a weighted average of the group variances and is defined as:

$$s_p^2 = \sum_{i=1}^{k}(N_i-1)s_i^2/(N-k)$$

- Sensitive to departures from normality