# Lab: Tests for Normaity and Equal Variance

Lab 6

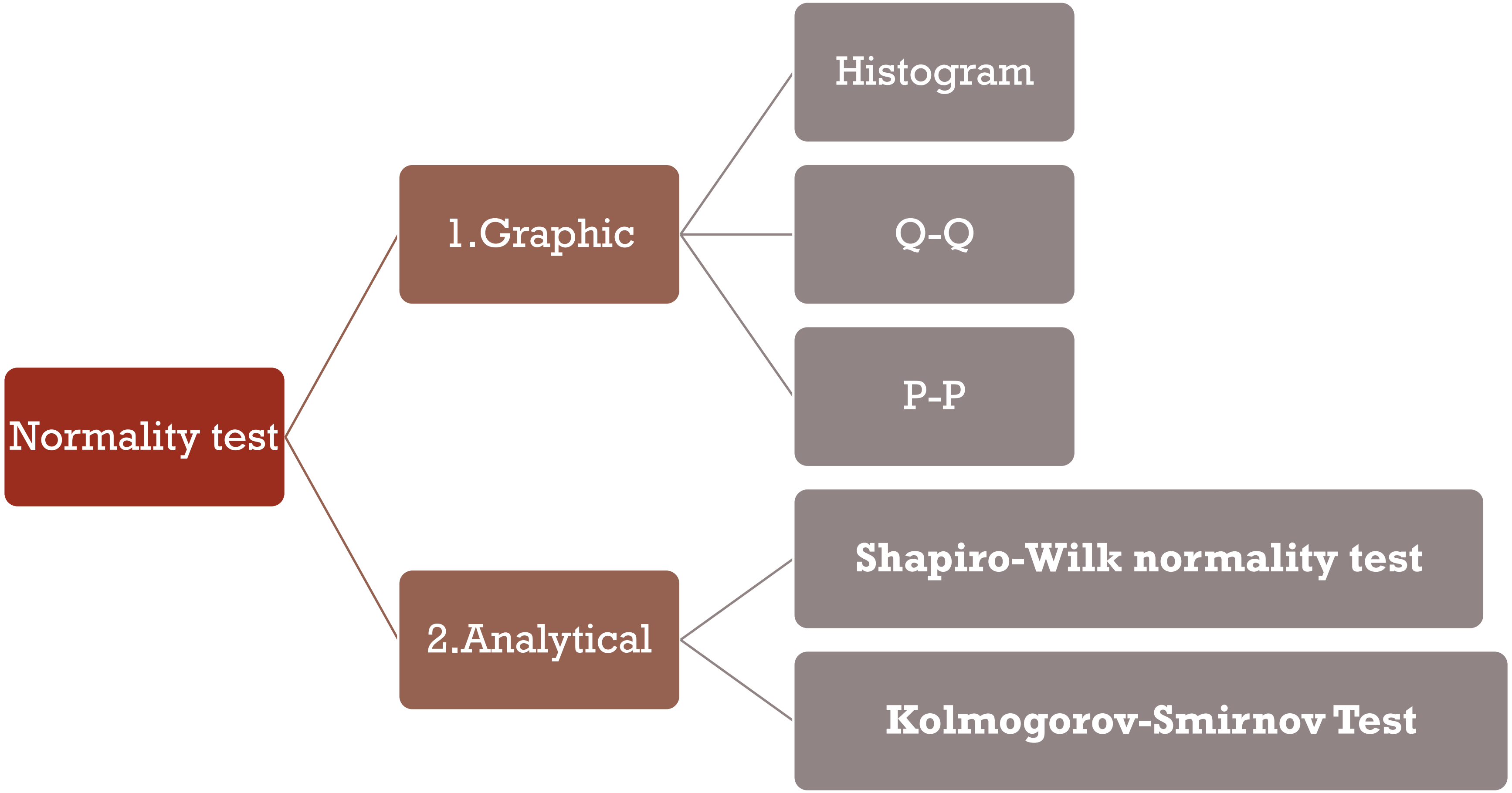
# PART 1. NORMALITY TEST

# NORMALITY

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

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***

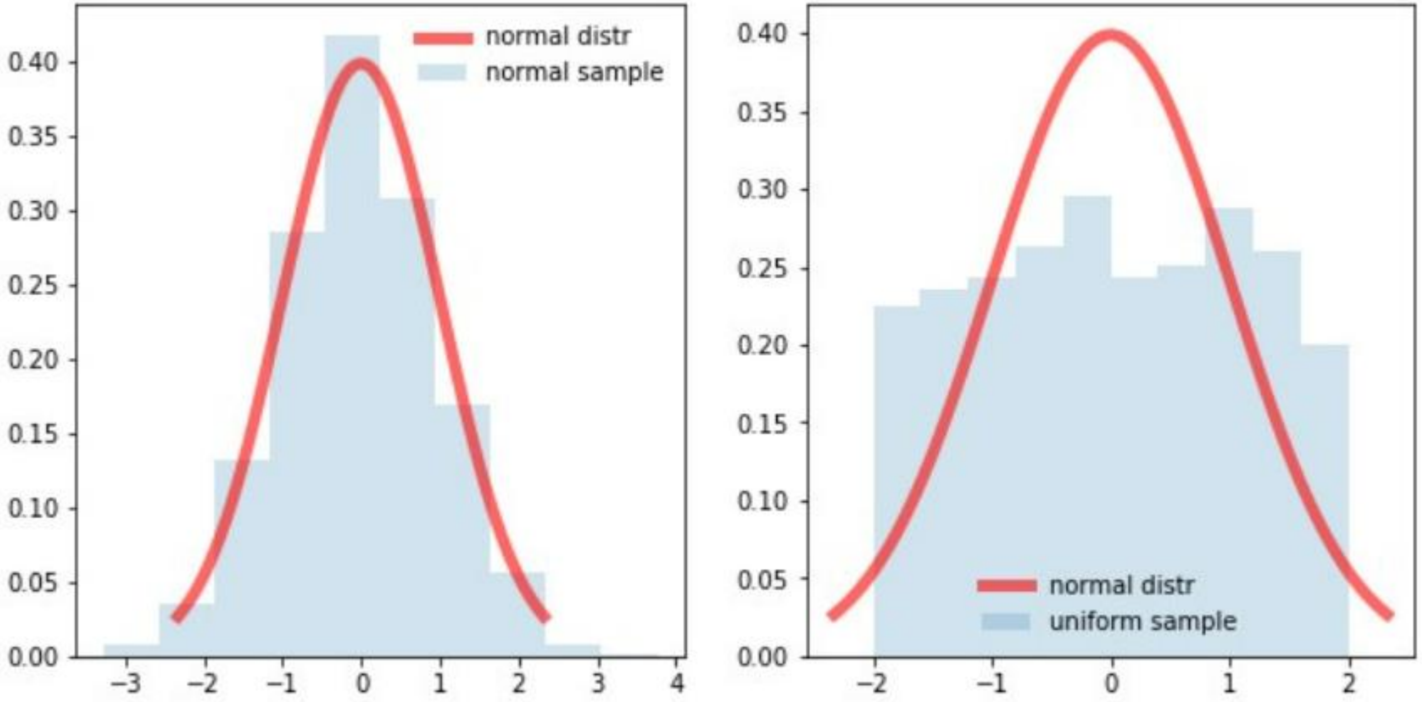


# 1.1 Histogram

- 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.

not very useful when the sample size is small




Bell-shape

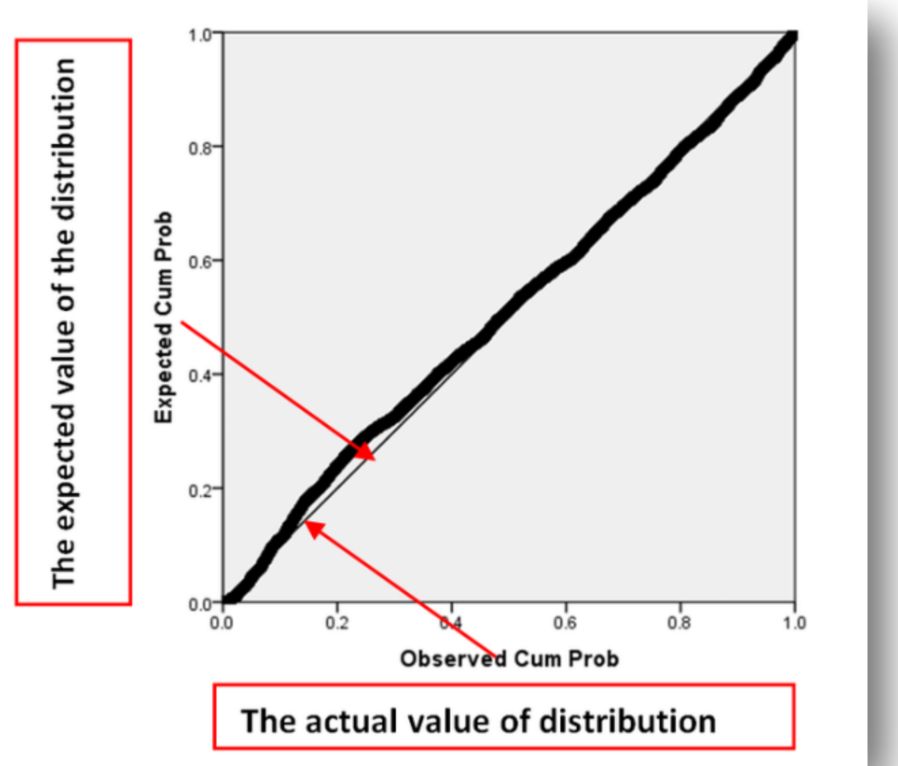
Normal (left) vs. non-normal distribution. The red curve represents an ideal normal (Gaussian) distribution.

# 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.



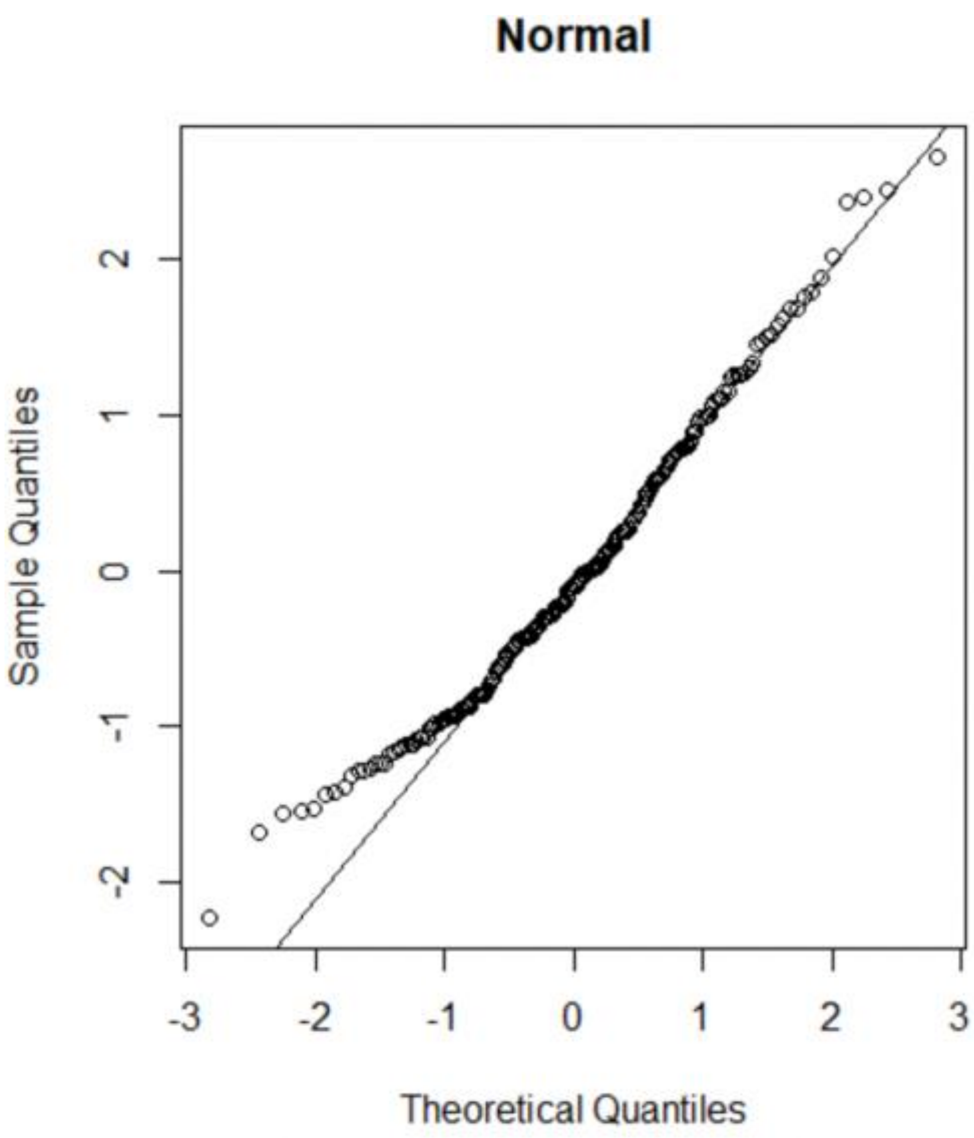
# *1.3 Quantile-quantile plot (Q-Q plot)*

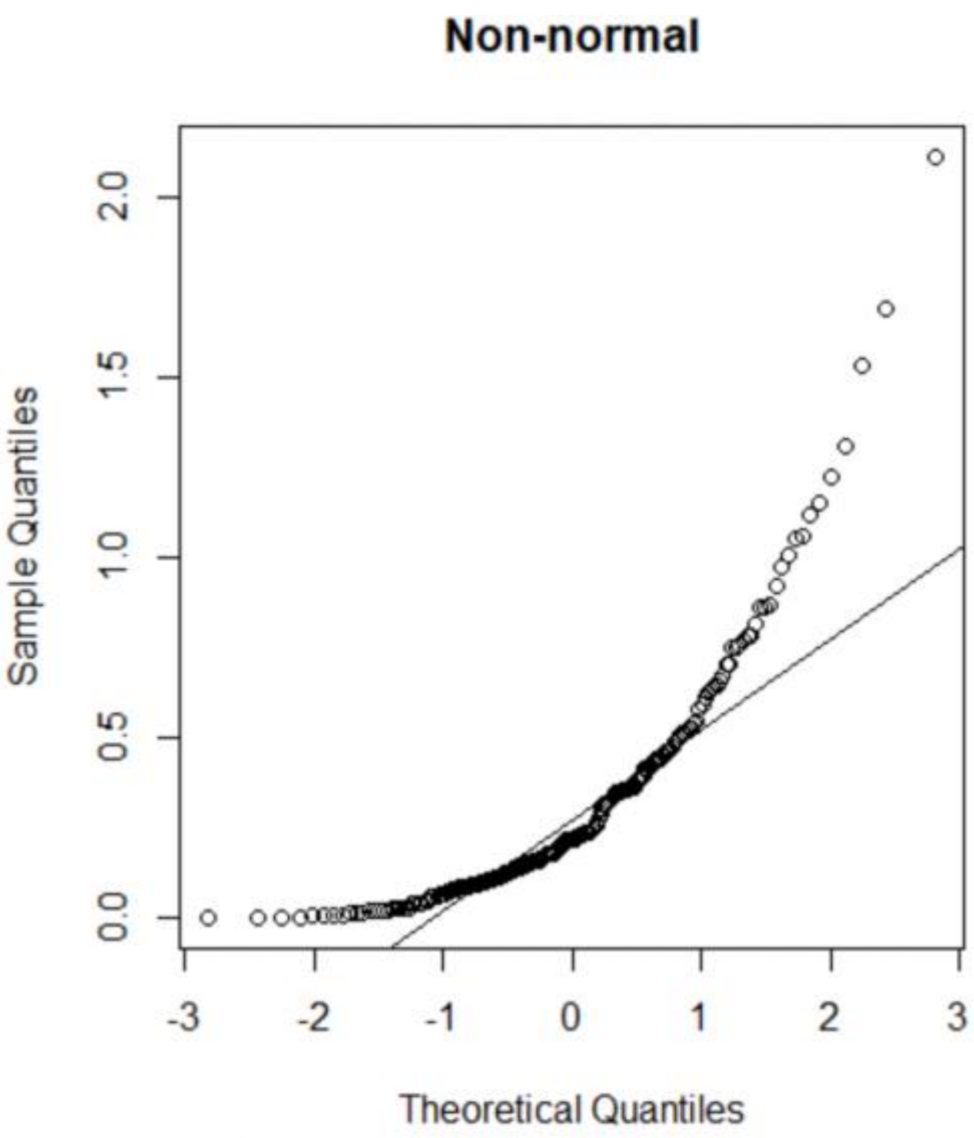
- **One axis: the quartile of sample data.**

- **Another axis: the expected/theoretical quartile based on the normal distribution.**

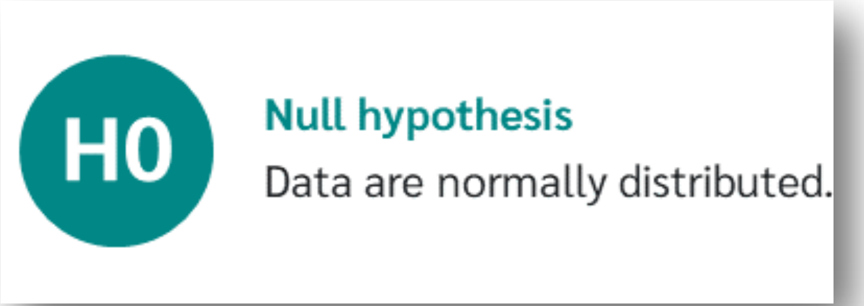
- **A normal distribution means that sample points are distributed around the diagonal of the first quadrant.**

- **Q-Q plot is more widely used than P-P plot in practice.**


Normal
Sample Quantiles
Theoretical Quantiles
-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*





Some statistician uses a=0.1

# *2.1 Shapiro-Wilk test*

- **Known as W test and introduced by S.S.Shapiro and M.B.Wilk;**
- **Shapiro-Wilk Original Test is suitable for sample sizes in the range of 3 to 50;**
- **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)]

# *Basic approach of Shapiro-Wilk original test*

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

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

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

④ **Calculate b and the test statistic W as follows**

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

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

① **taking the ai weights from Shapiro-Wilk Tables (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**

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

[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/Reciprocal for severe skew (for non-zero values):

  1/x for positively skewed data

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

Box-cox transformation

Yeo-Johnson Transformation

# 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**

**If two population variances are equal**

$$F = \frac{s_1^2}{s_2^2} \qquad usually, F = \frac{s_1^2 (bigger)}{s_2^2 (smaller)}$$

$$v_1 = \mathrm{n}_1\text{-}1, \quad v_2 = \mathrm{n}_2\text{-}1$$

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

- Suppose we have 2 independent populations

PROOF

$$X_1 \sim N(\mu_1, \sigma_1^2), \quad X_2 \sim N(\mu_2, \sigma_2^2)$$

$$\bar{X}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} X_{i1}, \quad s_1^2 = \frac{1}{n_1-1} \sum_{i=1}^{n_1} (X_{i1} - \bar{X}_1)^2$$
$$\bar{X}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} X_{i2}, \quad s_2^2 = \frac{1}{n_2-1} \sum_{i=1}^{n_2} (X_{i2} - \bar{X}_2)^2$$

$$then \ \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} = \frac{\left. \dfrac{\sum_{i=1}^{n1} (X_{1i} - \bar{X}_1)^2}{n_1 - 1} \middle/ \sigma_1^2 \right.}{\left. \dfrac{\sum_{i=1}^{n2} (X_{2i} - \bar{X}_2)^2}{n_2 - 1} \middle/ \sigma_2^2 \right.} = \frac{\left. \dfrac{\sum_{i=1}^{n1} (X_{1i} - \bar{X}_1)^2}{\sigma_1^2} \middle/ n_1 - 1 \right.}{\left. \dfrac{\sum_{i=1}^{n2} (X_{2i} - \bar{X}_2)^2}{\sigma_2^2} \middle/ n_2 - 1 \right.} \sim F(n_1 - 1, n_2 - 1)$$

If σ1=σ2, then $\dfrac{s_1^2}{s_2^2} \sim F(n_1 - 1, n_2 - 1)$

Specifically, note that under $H_0$,F follows an $F_{d2,d1}$ distribution. Therefore,

$$Pr\left(S_2^2 / S_1^2 \geq F_{d_2,d_1,1-p}\right) = p$$

By taking the inverse of each side and reversing the direction of the inequality, we get

$$Pr\left(\frac{S_1^2}{S_2^2} \leq \frac{1}{F_{d_2,d_1,1-p}}\right) = p \qquad Pr\left(\frac{S_1^2}{S_2^2} \leq F_{d_1,d_2,p}\right) = p$$

The **lower *p*th percentile** of an *F* distribution with $d_1$ and $d_2$ *df is* the reciprocal of the **upper *p*th percentile** of an *F* distribution with $d_2$ and $d_1$ *df*. In symbols,

$$F_{d_1,d_2,p} = 1/F_{d_2,d_1,1-p}$$

(1) If F~F(n, m), then 1/F~F(m, n)

Proof $X \sim x_n^2, Y \sim x_m^2$

$$F_{n,m} = \frac{X/n}{Y/m}, then$$

$$1/F = \frac{Y/m}{X/n} \sim F_{m,n}$$

## *F* Test for the Equality of Two Variances

Suppose we want to conduct a test of the hypothesis $H_0$: $\sigma_1^2 = \sigma_2^2$ vs. $H_1$: $\sigma_1^2 \neq \sigma_2^2$ with significance level $\alpha$.

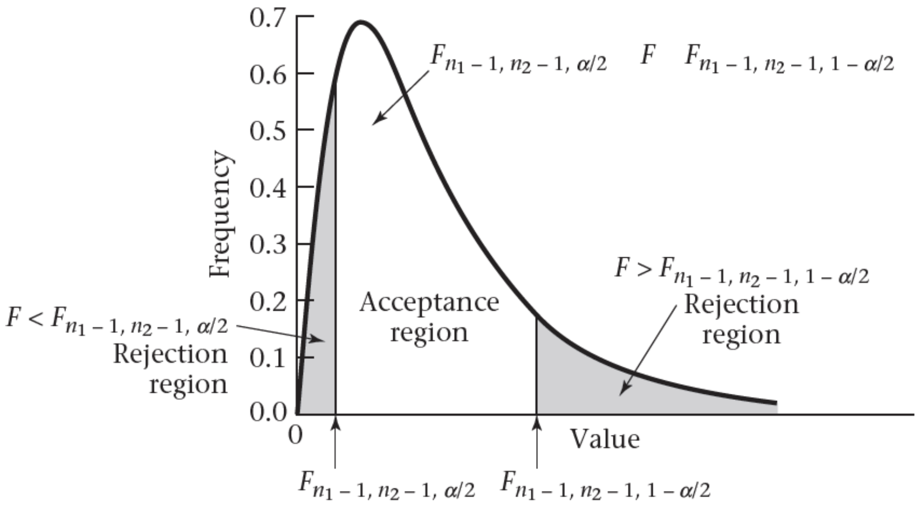
Compute the test statistic $F = s_1^2 / s_2^2$.

If $\quad F > F_{n_1-1,n_2-1,1-\alpha/2} \quad$ or $\quad F < F_{n_1-1,n_2-1,\alpha/2}$

then $H_0$ is rejected.

If $\quad F_{n_1-1,n_2-1,\alpha/2} \le F \le F_{n_1-1,n_2-1,1-\alpha/2}$

then $H_0$ is accepted. The acceptance and rejection regions for this test are shown




$F_{n_1-1,\,n_2-1}$ distribution = distribution of $F = S_1^2/S_2^2$ under $H_0$

If $F \ge 1$, then $\quad p = 2 \times Pr\left(F_{n_1-1,n_2-1} > F\right)$

If $F < 1$, then $\quad p = 2 \times Pr\left(F_{n_1-1,n_2-1} < F\right)$

# 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 $\alpha$.

- **Testing whether two or more population variances are equal**
- Less Sensitive to departures from normality than Bartlett's test

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

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 $\alpha$.

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

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 $i$th 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

# EXERCISE

- 1. In “example data.xls”, it has several numeric variables, such as bun, creatinine, tc, SBP and DBP. Using graphic and analytical methods to test whether these variables follow normal distribution or not?

- 2. Create a new variable BMI=weight (kg)/ height ($m^2$). Suppose BMI is normally distributed. The equality of variance of BMI between different genders before we conduct a two independent samples’ t-test.

# PART 3. R FUNCTIONS

## hist: Histograms

- The generic function hist computes a histogram of the given data values.

```
hist(x, …)

# S3 method for default
hist(x, breaks = "Sturges",
     freq = NULL, probability = !freq,
     include.lowest = TRUE, right = TRUE,
     density = NULL, angle = 45, col = NULL, border = NULL,
     main = paste("Histogram of" , xname),
     xlim = range(breaks), ylim = NULL,
     xlab = xname, ylab,
     axes = TRUE, plot = TRUE, labels = FALSE,
     nclass = NULL, warn.unused = TRUE, …)
```

- ‘**qqnorm**’ is a generic function the default method of which produces a normal QQ plot of the values in y.
- ‘**qqline**’ adds a line to a ‘‘theoretical’’, by default normal, quantile-quantile plot which passes through the probs quantiles, by default the first and third quartiles.
- ‘**qqplot**’ produces a QQ plot of two datasets.

```
qqnorm(y, …)
# S3 method for default
qqnorm(y, ylim, main = "Normal Q-Q Plot",
       xlab = "Theoretical Quantiles", ylab = "Sample Quantiles",
       plot.it = TRUE, datax = FALSE, …)

qqline(y, datax = FALSE, distribution = qnorm,
       probs = c(0.25, 0.75), qtype = 7, …)

qqplot(x, y, plot.it = TRUE, xlab = deparse(substitute(x)),
       ylab = deparse(substitute(y)), …)
```

- **shapiro.test**: Shapiro-Wilk Normality Test
- Performs the Shapiro-Wilk test of normality.

```
shapiro.test(x)
```

- **ks.test**: Kolmogorov-Smirnov Tests
- Performs one or two sample Kolmogorov-Smirnov tests.

```
ks.test(x, y, …,
        alternative = c("two.sided", "less", "greater"),
        exact = NULL, tol=1e-8, simulate.p.value=FALSE, B=2000)
```

- **lillie.test**: Lilliefors (Kolmogorov-Smirnov) test for normality
- Performs the Lilliefors (Kolmogorov-Smirnov) test for the composite hypothesis of normality

```
lillie.test(x)
```

- **Bartlett.test**: Bartlett Test of Homogeneity of Variances
- Performs Bartlett's test of the null that the variances in each of the groups (samples) are the same.

```
bartlett.test(x, …)

# S3 method for default
bartlett.test(x, g, …)

# S3 method for formula
bartlett.test(formula, data, subset, na.action, …)
```

X=data value
G=group

## leveneTest: Levene's Test

- Computes Levene's test for homogeneity of variance across groups.

```
leveneTest(y, ...)
# S3 method for formula
leveneTest(y, data, ...)
# S3 method for lm
leveneTest(y, ...)
# S3 method for default
leveneTest(y, group, center=median, ...)
```