Test for Normality
Tests for Independence
Measures of association

R-Test for Normality

- It's possible to use a **significance test** comparing the sample distribution to a normal one in order to ascertain whether data show or not a serious deviation from normality.
- There are several methods for **normality test** such as **Kolmogorov-Smirnov (K-S) normality test** and **Shapiro-Wilk's test**.
- The null hypothesis of these tests is that "sample distribution is normal". If the test is significant, the distribution is non-normal.

R-Test for Normality

- Shapiro-Wilk's method is widely recommended for normality test and it provides better power than K-S.
- It is based on the correlation between the data and the corresponding normal scores.
- Note that, normality test is sensitive to sample size. Small samples most often pass normality tests.
- Therefore, it's important to combine visual inspection and significance test in order to take the right decision.

R-Test for Normality

- The R function shapiro.test() can be used to perform the Shapiro-Wilk test of normality for one variable (univariate):
- shapiro.test(my_data$len)

```
Shapiro-Wilk normality test
data: my_data$len
W = 0.96743, p-value = 0.1091
```

- From the output, the p-value > 0.05 implying that the distribution of the data are not significantly different from normal distribution. In other words, we can assume the normality.

R-Tests of independence

- R provides several methods of testing the independence of categorical variables.
- The three tests are
  - The chi-square test of independence,
  - The Fisher exact test, and
  - The Cochran–Mantel–Haenszel test.
- Example:

CHI-SQUARE TEST OF INDEPENDENCE

- We can apply the function chisq.test() to a two-way table in order to produce a chisquare test of independence of the row and column variables.

CHI-SQUARE TEST OF INDEPENDENCE

```
> library(vcd)
> mytable <- xtabs(~ Treatment+Improved, data=Arthritis)
> mytable
           Improved
Treatment  None  Some  Marked
  Placebo   29     7       7
  Treated   13     7      21
> chisq.test(mytable)
        Pearson's Chi-squared test
data:  mytable
 X-squared = 13.1, df = 2, p-value = 0.001463
```

1 Treatment and Improved aren't independent.

- There appears to be a relationship between treatment received and level of improvement (p < .01).
- The p-values are the probability of obtaining the sampled results, assuming independence of the row and column variables in the population.
- Because the probability is small for (1), you reject the hypothesis that treatment type and outcome are independent.

CHI-SQUARE TEST OF INDEPENDENCE

```
> mytable <- xtabs(~Improved+Sex, data=Arthritis)
> mytable
         Sex
Improved Female Male
  None       25   17
  Some       12    2
  Marked     22    6
> chisq.test(mytable)
```

```
> chisq.test(mytable)
        Pearson's Chi-squared test
data:  mytable
 X-squared = 4.84, df = 2, p-value = 0.0889

Warning message:
In chisq.test(mytable) : Chi-squared approximation may be incorrect
```

2 Gender and Improved are independent.

CHI-SQUARE TEST OF INDEPENDENCE

- But there doesn't appear to be a relationship (2) between patient sex and improvement (p > .05).
- The p-values are the probability of obtaining the sampled results, assuming independence of the row and column variables in the population.
- Because the probability for (2) isn't small, it's not unreasonable to assume that outcome and gender are independent.
- The warning message is produced because one of the six cells in the table (male-some improvement) has an expected value less than five, which may invalidate the chi-square approximation.

FISHER'S EXACT TEST

- Fisher's exact test evaluates the null hypothesis of independence of rows and columns in a contingency table with fixed marginals.
- The format is fisher.test(mytable)
  - Where mytable is a two-way table.

```
> mytable <- xtabs(~ Treatment+Improved, data=Arthritis)
> mytable
          Improved
Treatment  None  Some  Marked
  Placebo   29     7      7
  Treated   13     7     21

> fisher.test(mytable)
        Fisher's Exact Test for Count Data
data:  mytable
p-value = 0.001393
alternative hypothesis: two.sided
```

- The fisher.test() function can be applied to any two-way table with two or more rows and columns, not a 2 × 2 table.

COCHRAN–MANTEL–HAENSZEL TEST

- The mantelhaen.test() function provides a Cochran–Mantel–Haenszel chi-square test of the null hypothesis that two nominal variables are conditionally independent in each stratum of a third variable.
- The following code tests the hypothesis that the Treatment and Improved variables are independent within each level for Gender.
- The test assumes that there's no three-way (Treatment × Improved × Gender) interaction:

```
> mytable <- xtabs(~Treatment+Improved+Sex, data=Arthritis)
> mytable
, , Sex = Female

          Improved
Treatment None Some Marked
  Placebo   19    7      6
  Treated    6    5     16

, , Sex = Male

          Improved
Treatment None Some Marked
  Placebo   10    0      1
  Treated    7    2      5

>
>
```

COCHRAN–MANTEL–HAENSZEL TEST

```
> mantelhaen.test(mytable)
        Cochran-Mantel-Haenszel test
data:  mytable
Cochran-Mantel-Haenszel M^2 = 14.6, df = 2, p-value = 0.0006647
```

- The results suggest that the treatment received and the improvement reported aren't independent within each level of Gender.

Measures of association

- The significance tests in the previous section evaluate whether sufficient evidence exists to reject a null hypothesis of independence between variables.
- If you can reject the null hypothesis(variables are not independent), your interest turns naturally to measures of association in order to gauge the strength of the relationships present.
- The assocstats() function in the vcd package can be used to calculate the phi coefficient, contingency coefficient, and Cramer's V for a two-way table.

```
> library(vcd)
> mytable <- xtabs(~Treatment+Improved, data=Arthritis)
> assocstats(mytable)

                    X^2 df  P(> X^2)
Likelihood Ratio 13.530  2 0.0011536
Pearson          13.055  2 0.0014626

Phi-Coefficient   : 0.394
Contingency Coeff.: 0.367
Cramer's V        : 0.394
```

- In general, larger magnitudes indicate stronger associations.
- The vcd package also provides a kappa() function that can calculate Cohen's kappa and weighted kappa for a confusion matrix.

Visualizing results

- R has mechanisms for visually exploring the relationships among categorical variables that go well beyond those found in most other statistical platforms.
- We typically use bar charts to visualize frequencies in one dimension. The vcd package has excellent functions for visualizing relationships among categorical variables in multidimensional datasets using mosaic and association plots.
- Finally, correspondence-analysis functions in the ca package allow you to visually explore relationships between rows and columns in contingency tables using various geometric representations.