University of New Mexico

Hypothesis Testing-3 (Fall 2018)

BIOM 505: Biostatistical Methods I (by Fares Qeadan)

Hypothesis Testing for Normality:

The null and alternative hypotheses are:

$$H_0 : \text{The data follow the normal distribution}$$
$$H_1 : \text{The data do not follow the normal distribution}$$

We can test normality, in SAS, by either graphical or numerical methods. The graphical methods include drawing

- stem-and-leaf plot
- box-plot
- histogram
- probability-probability (P-P) plot
- quantile-quantile (Q-Q) plot

The numerical methods involve computing the [1]

- Shapiro-Wilk test
- Kolmogorov-Smirnov test
- Cramer-von Mises test
- Anderson-Darling
- Chi-Square

Notes:

1. A P-P plot compares the empirical cumulative distribution function of a data set with a specified theoretical cumulative distribution function F(). A Q-Q plot compares the quantiles of a data distribution with the quantiles of a standardized theoretical distribution from a specified family of distributions. For normally distributed data this plot should lie on a 45° line between (0, 0) and (1, 1)

2. Skewness is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the center point. Kurtosis is a measure of whether the data are peaked or flat relative to a normal distribution.

- **Skewness > 0:** Right skewed distribution - most values are concentrated on left of the mean, with extreme values to the right.

- **Skewness < 0:** Left skewed distribution - most values are concentrated on the right of the mean, with extreme values to the left.

- **Skewness $= 0$:** mean = median, the distribution is symmetrical around the mean.

- **Kurtosis > 3:** Leptokurtic distribution, sharper than a normal distribution, with values concentrated around the mean and thicker tails. This means high probability for extreme values.

- **Kurtosis < 3:** Platykurtic distribution, flatter than a normal distribution with a wider peak. The probability for extreme values is less than for a normal distribution, and the values are wider spread around the mean.

- **Kurtosis $= 3$:** Mesokurtic distribution - normal distribution for example.

[1] The three different tests might give different and conflicting results especially at the marginally significant cases

(1) Consider the following hypertension dataset (http://www.mathalpha.com/BIOM-505/hypertensionfall17.sas7bdat) [Data set 2 on the course webpage]. This dataset is courtesy of Dr Waldon Garris, University of Virginia School of Medicine. Dr Garriss collected the data in a pilot study during his work in the Dominican Republic in 1997. The subjects are persons who came to medical clinics in several villages, for a variety of complaints. Data on gender, age, systolic and diastolic blood pressure were collected. Test whether the diastolic blood pressures follow the normal distribution at the significance level of $\alpha = 0.05$?

(a) The significance level α is:

(b) The null and alternative hypotheses are:

(c) The decision rule (about H_0) is:

(d) Conduct the test in SAS:

```
proc capability data=biom505.hypertensionfall17 graphics normaltest;
var dbp;
histogram dbp/normal;
run;

data dbp;
set biom505.hypertensionfall17;
keep dbp;
run;

proc univariate data=dbp normal plot;
histogram dbp/normal;
qqplot dbp;
ppplot dbp;
run;
```

(e) Get the p-value from SAS output:

(f) Decision:

(g) Conclusion:

(2) Consider the following hypertension dataset (http://www.mathalpha.com/BIOM-505/hypertensionfall17.sas7bdat) [Data set 2 on the course webpage]. This dataset is courtesy of Dr Waldon Garris, University of Virginia School of Medicine. Dr Garriss collected the data in a pilot study during his work in the Dominican Republic in 1997. The subjects are persons who came to medical clinics in several villages, for a variety of complaints. Data on gender, age, systolic and diastolic blood pressure were collected. Test whether the diastolic blood pressures, among subjects of the ages between 48 and 60 years old, follow the normal distribution at the significance level of $\alpha = 0.05$?

(a) The significance level α is:

(b) The null and alternative hypotheses are:

(c) The decision rule (about H_0) is:

(d) Conduct the test in STATA:

```
proc capability data=biom505.hypertensionfall17 graphics normaltest;
var dbp;
histogram dbp/normal;
where age gt 48 and age lt 60;
run;

data dbp2;
set biom505.hypertensionfall17;
keep dbp;
where age gt 48 and age lt 60;
run;

proc univariate data=dbp2 normal plot;
histogram dbp/normal;
qqplot dbp;
ppplot dbp;
run;
```

(e) Get the p-value from SAS output:

(f) Decision:

(g) Conclusion: