Mathematical Statistics 2B

Practical 3: Checking Parametric Assumptions

Department of Statistics - University of Johannesburg

1. Introduction

In statistics, parametric tests are tests that make assumptions about the underlying distribution of data. The most common parametric assumption is that data are approximately normally distributed. If the parametric assumptions are violated, it may change the conclusion of the research and the interpretation of the results. The four main assumptions for a parametric test to yield valid results are normality, equality of variance, independence and no outliers.

2. Normality

Most of the parametric tests require that the assumption of normality be met. Normality means that the distribution of the test is normally distributed (or bell-shaped). The assumption of normality can be assessed visually or numerically.

2.1 Visual methods

Histogram

The simplest graph to visually assess if the data are normally distributed is the histogram. This is a very informal method and should be used together with formal tests.

Cumulative distribution

A comparison of the empirical cumulative distribution function (ECDF) with the theoretical cumulative distribution function (CDF) is a useful visual representation. If the ECDF is S-shaped, it indicates normality. However, it should also be supported with formal tests. This plot was dealt with in Practical 2.

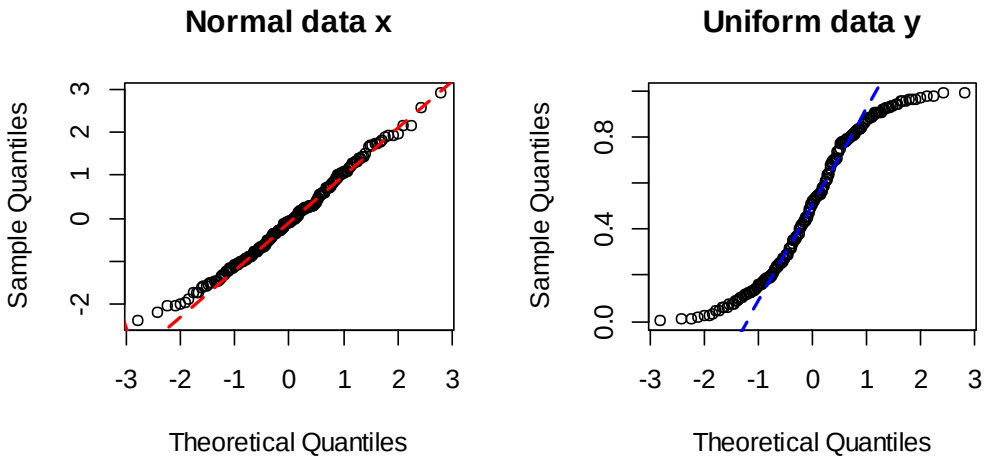
Probability plots

In probability plots, the data density distribution is transformed into a linear plot. The Q-Q plot and the P-P plot are both probability plots. A P-P plot compares the ECDF of a dataset with a specified theoretical CDF, whereas 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 the most part, the normal P-P plot is better at finding deviations from normality in the centre of the distribution, and the normal Q-Q plot is better at finding deviations in the tails. Q-Q plots tend to be preferred in research situations as data analysts are typically more concerned about the tails of a distribution, which will have more effect on inference. Both Q-Q and P-P plots can be used for distributions other than normal. The *qqnorm()* function for Q-Q plots is available in R base. The *qqline()* function adds a line to a theoretical (default normal) Q-Q plot which passes through the first and third quartiles. The P-P plot function is not available in R base.

Q-Q plot example

```
#let's simulate some data from two different distributions to illustrate the application
x = rnorm(200)
y = runif(200)

#Q-Q plots with the theoretical line from the normal distribution for both x and y in a 1 × 2 plot window
par(mfrow=c(1,2))
qqnorm(x,main="Normal data x")
qqline(x,col="red",lwd=2,lty=2)
qqnorm(y,main="Uniform data y")
qqline(y,col="blue",lwd=2,lty=2)
```



2.2 Numerical methods

Skewness and kurtosis

To test the assumption of normality, skewness should be within the range ±2 and kurtosis values should be within the range of ±7. In R, the *skewness()* and *kurtosis()* functions are available in the *moments* library, not in base R.

Chi-squared goodness-of-fit test

The χ^2 goodness-of-fit test is based on how many data points fall into different intervals based on what we would expect from a sample from a normal distribution. We already dealt with this test in practical 2.

Kolmogorov-Smirnov test

The Kolmogorov-Smirnov (K-S) test is used to compare two distributions to determine if they come from the same underlying distribution (null hypothesis). The K-S test statistic (D) is based on the maximum distance between the ECDF of a sample and the CDF of a specific reference distribution, or between the ECDF of one sample and the ECDF of another sample.

The first two arguments of the *ks.test()* function are "*x*" and "*y*", where "*x*" is a numeric vector and "*y*" is either a numeric vector or a character string identifying the specific CDF, such as "*pnorm*". If both *x* and *y* are numeric vectors, a two-sample K-S test is performed. If *y* denotes a specific distribution, a one-sample K-S test is performed. In this case we must also specify the parameters of the distribution. The K-S test is not the best test when estimating the parameters from the data as this may lead to ties in the data, which may result in a conservative *p*-value. Although the K-S test can still be used in such cases, an alternative is the Anderson-Darling test for normality which does not require specifying the mean and standard deviation. This test is not available in base R.

Shapiro-Wilk's W test

Shapiro-Wilk's method is widely recommended for testing normality, and it provides better power than K-S. It is based on the correlation between the data and the corresponding normal scores. It is an easy-to-use statistical tool that can help us find an answer to the normality check we need, but it has one flaw in that it does not work well with large data sets. In such cases alternative tests such as K-S can be used.

The *shapiro.test()* function is used to perform the Shapiro-Wilk test of normality for one variable (univariate). It has only one argument, namely the data vector.

K-S Example 1

#let's simulate some data to illustrate the application

x = rnorm(50)

y = runif(30)

#test the following hypothesis

H_0: *x* and *y* come from the same distribution vs. H_1: *x* and *y* do not come from the same distribution

ks.test(x, y)

#output

Exact two-sample Kolmogorov-Smirnov test

data: x and y

D = 0.62, p-value = 2.964e-07

alternative hypothesis: two-sided

#interpretation

p-value < 0.01 → reject H_0 at the 1% level of significance → very strong evidence that *x* and *y* do not come from the same distribution

K-S Example 2

#let's simulate some data to illustrate the application

x = rnorm(50,mean=100,sd=12)

#test the following hypothesis

H_0: *x* follows a normal distribution with a mean of 100 and standard deviation of 12

H_1: *x* does not follow a normal distribution with a mean of 100 and standard deviation of 12

ks.test(x, "pnorm",100,12)

#output

Exact one-sample Kolmogorov-Smirnov test

data: x

D = 0.10834, p-value = 0.5632

alternative hypothesis: two-sided

#interpretation

p-value > 0.1 → fail to reject H_0 at any level of significance → weak/no evidence for H_1 → it appears that *x* follows a normal distribution with a mean of 100 and standard deviation of 12

S-W Example 1

#let's simulate normal data to illustrate the application

x=rnorm(100, mean = 5, sd = 3)

#test the following hypothesis

H_0: *x* follows a normal distribution vs. H_1: *x* does not follow a normal distribution

shapiro.test(x)

#output

Shapiro-Wilk normality test

data: rnorm(100, mean = 5, sd = 3)

W = 0.99156, p-value = 0.7884

#interpretation

p-value > 0.1 → fail to reject H_0 at any level of significance → weak/no evidence for H_1 → it appears that *x* follows a normal distribution

S-W Example 2

#let's simulate uniform data to illustrate the application

x=runif(100, min = 2, max = 4)

#test the following hypothesis

H_0: *x* follows a normal distribution vs. H_1: *x* does not follow a normal distribution

shapiro.test(x)

#output

Shapiro-Wilk normality test

data: runif(100, min = 2, max = 4)

W = 0.95902, p-value = 0.003424

#interpretation

p-value < 0.01 → reject H0 at the 1% level of significance → very strong evidence that *x* does not follow a normal distribution

3. Equal Variance

Equality of variance, also referred to as homogeneity of variance or homoscedasticity, is an assumption underlying the independent samples *t*-test and ANOVA in which the population variances of two or more independent samples are considered equal. There are a number of different methods to assess the equality of variances. All methods test the following hypothesis:

H_0: the variances are equal vs. H_1: the variances are not equal

F-test

The F-test is used to compare variances of two samples selected from normal populations. The test statistic is calculated as the ratio of the two variances. If the ratio deviates more from 1, there is stronger evidence of unequal variances. A major assumption of the F-test is that the data should be normally distributed. The *var.test()* function in R is used for the F-test.

Bartlett's test

Bartlett's test is used to compare variances of two or more groups. It is used for normally distributed data. The *bartlett.test()* function in R is used for the Bartlett's test.

Levene's test

Levene's test for equality of variance is an alternative to Bartlett's test, especially if the data are not normally distributed. The *leveneTest()* function is not available in R base.

Fligner-Killeen's test

The Fligner-Killeen test is the most robust test for equality of variance if the assumption of normality is not met, or if there are outliers in the data, as it is a non-parametric test based on ranks. The *fligner.test()* function in R is used for the Fligner-Killeen test.

Tests for equality of variance example

```
#let's simulate data for two independent groups with a) equal variance and b) unequal variance
data1=as.data.frame(rbind(cbind(1,rnorm(100,mean = 100,sd = 10)),
                          cbind(2,rnorm(100,mean = 60,sd = 10))))
colnames(data1)=c("group","Same_SD")
data2=as.data.frame(rbind(cbind(1,rnorm(100,mean = 100,sd = 10)),
                          cbind(2,rnorm(100,mean = 100,sd = 20))))
colnames(data2)=c("group","Diff_SD")
```

```
#apply the F-test, Bartlett's test and the Fligner-Killeen test to both datasets
#I am only showing you the resulting p-values from each test here, there is more output in R
#data with equal variances
> var.test(x=data1$Same_SD[data1$group==1], y=data1$Same_SD[data1$group==2])
p-value = 0.3135

> bartlett.test(data1$Same_SD~data1$group)
p-value = 0.3135

> fligner.test(data1$Same_SD~data1$group)
p-value = 0.4197

#data with unequal variances
> var.test(x=data2$Diff_SD[data2$group==1], y=data2$Diff_SD[data2$group==2])
p-value = 5.322e-12

> bartlett.test(data2$Diff_SD~data2$group)
p-value = 5.346e-12

> fligner.test(data2$Diff_SD~data2$group)
p-value = 4.952e-08
```

4. Independence

Parametric tests assume that the observations are independent of one another. The easiest way to check this assumption is to verify that the data were collected using a probability sampling method, where every member in a population has an equal probability of being selected to be in the sample, and selection is done at random. This was dealt with in Practical 1.

5. Outliers

Parametric tests assume that there are no extreme outliers in the data that could adversely affect the results of the test. We can check this through frequency tables, or visually using boxplots or histograms. There are different tests that can be used to identify outliers in the data, for univariate, bivariate and multivariate statistical analysis. In this module we will not formally test for the presence of outliers and will only assess this visually.