**A7Wa Common assumptions about data**

Statistical inference uses data from a sample of individuals to reach conclusions about the whole population. When attempting to make inferences from sample data, you must check your assumptions. Violating any of these assumptions can result in false positives or false negatives, thus invalidating your results.

**The common data assumptions are:**

* **Random samples.**
* **Independence.**
* **Normality.**
* **Equal variance.**
* **Stability.**
* **Measurement system is accurate and precise.**

**What is the assumption of random samples?**

A sample is random when each data point in your population has an equal chance of being included in the sample; therefore, selection of any individual happens by chance, rather than by choice. This reduces the chance that differences in conditions strongly bias your results.

Random samples are more likely to be representative of the population that you are measuring, and you can be more confident with the analysis you undertake on your collected sample data. It is important to realise that the statistical tests we apply to solve problems involving hypothesis tests (including point and interval estimates) rely on random sampling. Following good sampling techniques will help to ensure your samples are random.

**What is the assumption of statistical independence?**

Statistical independence is a critical assumption for many statistical tests, such as the one and two sample z and t tests discussed in textbook Chapter 7. Independence means the value of one observation does not influence or affect the value of other observations. Non-independent observations introduce bias and can make your statistical test give too many false positives. Following good sampling techniques will help to ensure your samples are independent.

You can test for independence using the chi-square test for association where the claim is that the row and column variables are independent of each other. This is the null hypothesis. The multiplication rule said that if two events were independent, then the probability of both occurring was the product of the probabilities of each occurring.

This is key to working the test for independence. If you end up rejecting the null hypothesis, then the assumption must have been wrong, and the row and column variable are dependent. Remember, all hypothesis testing is done under the assumption the null hypothesis is true.

**What is the assumption of normality?**

Before you perform a statistical test, you should find out the distribution of your data. If you don’t, you risk selecting an inappropriate statistical test. Many statistical methods start with the assumption your data follow the normal distribution, including the one and two sample z and t tests.

 If you don’t have normally distributed data, you might use an equivalent non-parametric test based on the median instead of the mean or try data transformation to transform your non-normal data into a normal distribution.



Figure 1 Distribution shapes: left skewed



Figure 2 Distribution shapes: normal



Figure 3 Distribution shapes: right skewed

In fact, nonparametric statistics don't assume your data follow any distribution at all.

**Parametric vs non-parametric tests**

Both parametric and nonparametric tests draw inferences about populations based on samples, but parametric tests focus on sample parameters like the mean and the standard deviation and make various assumptions about your data—for example, that it follows a normal distribution.

But keep in mind that many statistical tools based on the assumption of normality do not actually require normally distributed data if the sample sizes are at least 30. But if sample sizes are less than 30 and the data are not normally distributed, the p-value may be inaccurate, and you should interpret the results with caution.

In contrast, non-parametric tests are unaffected by the distribution of your data. Non-parametric tests also accommodate many conditions that parametric tests do not handle, including small sample sizes, ordered outcomes, and outliers. Consequently, they can be used in a wider range of situations and with more types of data than traditional parametric tests.

Many people also feel that non-parametric analyses are more intuitive. But non-parametric tests are not completely free from assumptions—they do require data to be an independent random sample, for example. Furthermore, they typically have less statistical power than parametric equivalents. Power is the probability that you will correctly reject the null hypothesis when it is false. That means you have an increased chance making a Type II error with these tests.

Table 1 lists common parametric tests, their equivalent nonparametric tests, and the main characteristics of each.

In practical terms, that means non-parametric tests are less likely to detect an effect or association when one really exists. So, if you want to draw conclusions with the same confidence level you'd get using an equivalent parametric test, you will need larger sample sizes.



Table 1 Parametric vs non-parametric tests

**Testing for normality**

There are several methods to determine normality including graphical and statistical methods:

1. A graphical tool for assessing normality is the normal probability plot, a quantile-quantile plot (QQ plot) of the standardized data against the standard normal distribution. Here the correlation between the sample data and normal quantiles (a measure of the goodness of fit) measures how well the data are modelled by a normal distribution. For normal data the points plotted in the QQ plot should fall approximately on a straight line, indicating high positive correlation. These plots are easy to interpret and also have the benefit that outliers are easily identified.
2. Statistical tests are available to test for normality, including the Shapiro–Wilk’s test.

**Use of SPSS to assess normality**

SPSS’s normality test will generate a probability plot and perform a one-sample hypothesis test to determine whether the population from which you draw your sample is non-normal.

**Example 1**

Consider the data collected from the local badminton club which measures the age of members. Is this sample data normally distributed?

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 65 | 61 | 63 | 86 | 70 | 55 | 74 | 35 |
| 72 | 68 | 45 | 58 | 58 | 45 | 68 | 72 |
| 35 | 74 | 55 | 70 | 86 | 63 | 61 | 65 |

Table 2

1. Null hypothesis states that the population is normal.
2. The alternative hypothesis states that the population is non-normal.

SPSS test

Enter data into SPSS and run the appropriate test

Click Analyze > Descriptive Statistics > Explore...

Transfer the variable that needs to be tested for normality into the Dependent List: box

Click the Statistics button and choose Descriptives

Click the Plots button. Change the options so that Normality plots with tests is chosen



Figure 4 SPSS solution

Click OK

When testing for normality, we are mainly interested in the Tests of Normality table and the Normal Q-Q Plots, our numerical and graphical methods to test for the normality of data, respectively.

SPSS output



Figure 5

The above presents the results from two well-known tests of normality, namely the Kolmogorov-Smirnov Test and the Shapiro-Wilk Test. The Shapiro-Wilk Test is more appropriate for small sample sizes (< 50 samples) but can also handle sample sizes as large as 2000. If we have sample sizes above 2000, we use Kolmogorov‐Smirnov test. In the example above we have a sample size of 24 so we will use the Shapiro-Wilk test as our numerical means of assessing normality.

Shapiro-Wilk test statistic p-value = 0.25 > 0.05, accept null hypothesis (statistical evidence suggests data is normally distributed) and reject alternative hypothesis (statistical evidence suggests data is not normally distributed). We can see from the table that "Age" is normally distributed. How do we know this? If the Sig. value of the Shapiro-Wilk Test is greater than 0.05, the data is normal. If it is below 0.05, the data significantly deviate from a normal distribution.

Note: you could use skewness and kurtosis values to determine normality, rather the Shapiro-Wilk test.

Normal Q-Q Plot

In order to determine normality graphically, we can use the output of a normal Q-Q Plot. If the data are normally distributed, the data points will be close to the diagonal line. If the data points stray from the line in an obvious non-linear fashion, the data are not normally distributed.



Figure 6

As we can see from the normal Q-Q plot below, the data is normally distributed. If you are at all unsure of being able to correctly interpret the graph, rely on the numerical methods instead because it can take a fair bit of experience to correctly judge the normality of data based on plots.

**What is the assumption of equal variance?**

In simple terms, variance refers to the data spread or scatter. Statistical tests, such as analysis of variance (ANOVA), assume that although different samples can come from populations with different means, they have the same variance.

1. Equal variances (homoscedasticity) is when the variances are approximately the same across the samples.
2. Unequal variances (heteroscedasticity) can affect the Type I error rate and lead to false positives.

If you are comparing two or more sample means, as in the 2-Sample t-test and ANOVA test which compare more than 2 samples, a significantly different variance could overshadow the differences between means and lead to incorrect conclusions. To test for homogeneity of variance, there are several statistical tests that can be used:

1. F test for equality of variance.
2. Levene’s test for equality of variance.

The F-test is sensitive to the normality assumption and if this is an issue then the non-parametric Levene’s test should be used to assess equality of variance. For further information see W7.7.6. If this assumption is violated then you can use a non-parametric test to undertake the hypothesis testing.

**What Is the Assumption of Stability?**

A stable process is one in which the inputs and conditions are consistent over time. When a process is stable, it is said to be “in control.” This means the sources of variation are consistent over time, and the process does not exhibit unpredictable variation. In contrast, if a process is unstable and changing over time, the sources of variation are inconsistent and unpredictable. As a result of the instability, you cannot be confident in your statistical test results.

**What Is the Assumption for Measurement Systems?**

All the assumptions I’ve mentioned so far assume the data reflects reality. But does it? The measurement system is one potential source of variability when measuring a process. When a measurement system is poor, you lose the ability to measure what you are trying to measure. A poor measurement system leads to incorrect conclusions and flawed implementation.