Theme 5. Hypothesis testing

A statistical hypothesis is some statement about the probability distribution, characterising a population on the basis of the information available from the sample observations. In the formulation and testing of hypothesis, statistical methods are extremely useful. Whether crop yield has increased because of the use of new fertilizer or whether the new medicine is effective in eliminating a particular disease are some examples of statements of hypothesis and these are tested by proper statistical tools.

We define five stages when carrying out a hypothesis test:

1. Define the null and alternative hypotheses under study.
2. Collect relevant data from a sample of individuals
3. Calculate the value of the test statistic specific to the null hypothesis.
4. Compare the value of the test statistic to values from a known probability distribution.
5. Interpret the P-value and results.

We usually test *the null hypothesis* ($\boldsymbol{H_0}$) which assumes no effect (e.g. the difference in means equals zero) in the population. For example, if we are interested in comparing smoking rates in men and women in the population, the null hypothesis would be:

H_0: smoking rates are the same in men and women in the population.

We then define *the alternative hypothesis* ($\boldsymbol{H_1}$) which holds if the null hypothesis is not true. The alternative hypothesis relates more directly to the theory we wish to investigate. So, in the example, we might have:

$\boldsymbol{H_1}$: the smoking rates are different in men and women in the population.

All test statistics follow known theoretical probability distributions. We relate the value of the test statistic obtained from the sample to the known distribution to obtain the P-value, the area in both (or occasionally one) tails of the probability distribution. ***The P-value is the probability of obtaining our results, or something more extreme, if the null hypothesis is true.***

The smaller the P-value, the greater the evidence against the null hypothesis.

• Conventionally, we consider that if the P-value is less than 0.05, there is sufficient evidence to reject the null hypothesis, as there is only a small chance of the results occurring if the null hypothesis were true. We then reject the null hypothesis and say that the results are significant at the 5% level (Fig. 6.1).

• In contrast, if the P-value is equal to or greater than 0.05, we usually conclude that there is insufficient evidence to reject the null hypothesis. We do not reject the null hypothesis, and we say that the results are not significant at the 5% level (Fig. 9). This does not mean that the null hypothesis is true; simply that we do not have enough evidence to reject it. The P-value (e.g. 0.05 or 0.01) is called the ***significance level*** of the test; it must be chosen before the data are collected.

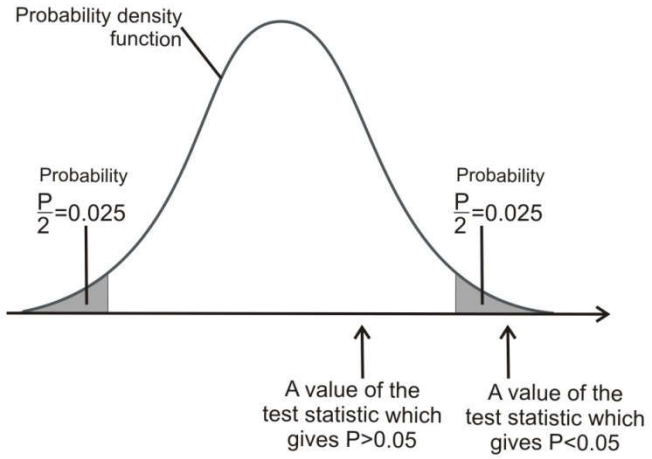

Fig.6.1. Probability distribution of the test statistic showing a two-tailed probability, P= 0.05.

<u>Parametric and non-parametric tests</u>

Hypothesis tests which are based on knowledge of the probability distributions that the data follow are known as ***parametric tests***. Often data do not conform to the assumptions that underly these methods. In these instances we can use ***non-parametric tests*** (sometimes referred to as ***distribution-free tests***, or ***rank methods***).

Non-parametric tests are particularly useful when the sample size is small (so that it is impossible to assess the distribution of the data), and/or when the data is measured on a categorical scale.

Errors in hypothesis testing

Although we hope we will draw the correct conclusion about the null hypothesis, we have to recognize that, because we only have a sample of information, we may make the wrong decision when we reject/do not reject the null hypothesis. The possible mistakes we can make are shown in Table 6.1.

• ***Type I error***: we reject the null hypothesis when it is true, and conclude that there is an effect when, in reality, there is none. The maximum chance (probability) of making a *Type I error* is denoted by ***α (alpha)***. This is the significance level of the test; we reject the null hypothesis if our P-value is less than the significance level, i.e. if $P < \alpha$.

• ***Type II error***: we do not reject the null hypothesis when it is false, and conclude that there is no effect when one really exists. The chance of making a Type II error is denoted by ***β (beta)***; its compliment, ***(1 - β)***, is the ***power of the test***. The power, therefore, is the probability of rejecting the null hypothesis when it is false; i.e. it is the chance (usually expressed as a percentage) of detecting, as statistically significant, a real treatment effect of a given size.

Table 6.1.

The consequences of hypothesis testing

	Reject H_0	***Do not reject H_0***
H_0 true	Type I error	No error
H_0 false	No error	Type II error

Power and related factors

It is essential that we know the power of a proposed test at the planning stage of our investigation. Clearly, we should only embark on a study if we believe that it has a 'good' chance of detecting a clinically relevant effect, if one exists. By 'good' we mean that the power should be at least 80%. It is ethically irresponsible, and wasteful of time and resources, to undertake a clinical trial that has, say, only a 40% chance of detecting a real treatment effect.

A number of factors have a direct bearing on power for a given test.

• ***The sample size***: power increases with increasing sample size. This means that a large sample has a greater ability than a small sample to detect a clinically important effect if it exists. When the sample size is very small, the test may have an inadequate power to detect a particular effect.

• ***The variability of the observations***: power increases as the variability of the observations decreases.

• ***The effect of interest***: the power of the test is greater for larger effects. A hypothesis test thus has a greater chance of detecting a large real effect than a small one.

• ***The significance level***: the power is greater if the significance level is larger. This is equivalent to the probability of the *Type I error* (α) increasing as the probability of the *Type II error* (β) decreases. So, we are more likely to detect a real effect if we decide at the planning stage that we will regard our P-value as significant if it is less than 0.05 rather than less than 0.01.

Numerical data: a single group

We have a sample from a single group of individuals and one numerical or ordinal variable of interest.

The one-sample t-test

In the population, the variable is normally distributed with a given (usually unknown) variance. In addition, we have taken a reasonable sample size so that we can check the assumption of normality.

We are interested in whether the mean μ of the variable in the population of interest differs from some hypothesized value, μ_1. We use a test statistic that is based on the difference between the sample mean $\bar{x}$ and μ_1. Assuming that we do not know the population variance, then this test statistic, often referred to as *t*, follows the *t*-distribution.

Notation

Our sample is of size *n* and the estimated standard deviation is *s*.

1. Define the null and alternative hypotheses under study

H_0: the mean in the population μ equals μ_1.

H_1: the mean in the population does not equal μ_1.

2. Collect relevant data from a sample of individuals

3. Calculate the value of the test statistic specific to H_0

$t = \frac{(\bar{x} - \mu_1)}{s/\sqrt{n}}$,which follows the t-distribution with (n - 1) degrees of freedom.

4. Compare the value of the test statistic to values from a known probability distribution

5 Interpret the P-value and results

The 95% confidence interval is given by:

$\bar{x} \pm t_{0.05} \times (s/\sqrt{n})$,

where $t_{0.05}$ is the percentage point of the t-distribution with (n - 1) degrees of freedom which gives a two-tailed probability of 0.05.

The 95% confidence interval provides a range of values in which we are 95% certain that the true population mean lies. If the 95% confidence interval does not include the hypothesized value for the mean μ_1 , we reject the null hypothesis at the 5% level. If, however, the confidence interval includes μ_1 , then we fail to reject the null hypothesis at that level.

The sign test

The sign test is a simple test based on the median of the distribution. We have some hypothesized value λ for the median in the population. If our sample comes from this population, then approximately half of the values in our sample should be greater than λ and half should be less than λ (after excluding any values which equal λ). The sign test considers the number of values in our sample that are greater (or less) than λ.

1. Define the null and alternative hypotheses under study

H_0: the median in the population equals λ.

H_1: the median in the population does not equal λ.

2. Collect relevant data from a sample of individuals

3. Calculate the value of the test statistic specific to H_0

4. Compare the value of the test statistic to values from a known probability distribution

5. Interpret the P-value and results

Interpret the *P*-value and calculate a confidence interval for the median; if not, we can rank the values in order of size to identify the ranks of the values that are to be used to define the limits of the confidence interval. In general, confidence intervals for the median will be wider than those for the mean.

Numerical data: two related groups

We have two samples that are related to each other and one numerical or ordinal variable of interest.

• The variable may be measured on each individual in two circumstances. For example, in a cross-over trial, each patient has two measurements on the variable, one while taking active treatment and one while taking placebo.

• The individuals in each sample may be different, but are linked to each other in some way. For example, patients in one group may be individually matched to patients in the other group in a case–control study.

The paired t-test

In the population of interest, the individual differences are normally distributed with a given (usually unknown) variance. We have a reasonable sample size so that we can check the assumption of normality.

If the two sets of measurements were the same, then we would expect the mean of the differences between each pair of measurements to be zero in the population of interest. Therefore, our test statistic simplifies to a one-sample *t*-test on the differences, where the hypothesized value for the

mean difference in the population is zero.

Notation

Because of the paired nature of the data, our two samples must be of the same size, n. We have n differences, with sample mean $\bar{x}$ and estimated standard deviation s_d.

1. Define the null and alternative hypotheses under study

H_0: the mean difference in the population equals zero

H_1: the mean difference in the population does not equal zero.

2. Collect relevant data from two related samples

3. Calculate the value of the test statistic specific to H_0

4. Compare the value of the test statistic to values from a known probability distribution

5 Interpret the P-value and results

The Wilcoxon signed ranks test

The Wilcoxon signed ranks test takes account not only of the signs of the differences but also their magnitude, and therefore is a more powerful test. The individual difference is calculated for each pair of results. Ignoring zero differences, these are then classed as being either positive or negative. In addition, the differences are placed in order of size, ignoring their signs, and are ranked accordingly.

1. Define the null and alternative hypotheses under study

H_0: the median difference in the population equals zero

H_1: the median difference in the population does not equal zero.

2. Collect relevant data from two related samples

3. Calculate the value of the test statistic specific to H_0

4. Compare the value of the test statistic to values from a known probability distribution

5. Interpret the P-value and results

Interpret the *P*-value and calculate a confidence interval for the median difference in the entire sample.

Numerical data: two unrelated groups

We have samples from two independent (unrelated) groups of individuals and one numerical or ordinal variable of interest. We are interested in whether the mean or distribution of the variable is the same in the two groups. For example, we may wish to compare the weights in two groups of children, each child being randomly allocated to receive either a dietary supplement or placebo.

The unpaired (two-sample) t-test

In the population, the variable is normally distributed in each group and the variances of the two groups are the same. We consider the difference in the means of the two groups. Under the null hypothesis that the population means in the two groups are the same, this difference will equal zero. Therefore, we use a test statistic that is based on the difference in the two sample means, and on the value of the difference in population means under the null hypothesis (i.e. zero). This test statistic, often referred to as t, follows the t-distribution.

1. Define the null and alternative hypotheses under study

H_0: the population means in the two groups are equal

H_1: the population means in the two groups are not equal.

2. Collect relevant data from two samples of individuals

3. Calculate the value of the test statistic specific to H_0

4. Compare the value of the test statistic to values from a known probability distribution

Refer t to t-distribution table. When the sample sizes in the two groups are large, the t-distribution approximates a Normal distribution, and then we reject the null hypothesis at the 5% level if the absolute value (i.e. ignoring the sign) of t is greater than 1.96.

5. Interpret the P-value and results

Interpret the *P*-value and calculate a confidence interval for the difference in the two means. The 95% confidence interval, assuming equal variances, is given by:

Numerical data: more than two groups

We have samples from a number of independent groups. We have a single numerical or ordinal variable and are interested in whether the average value of the variable varies in the different groups. Although we could perform tests to compare the averages in each pair of groups, the high *Type I error rate*, resulting from the large number of comparisons, means that we may draw incorrect conclusions. Therefore, we carry out a single global test to determine whether the averages differ in any groups.

One-way analysis of variance

The groups are defined by the levels of a single factor. In the population of interest, the variable is normally distributed in each group and the variance in each group is the same. We have a reasonable sample size so that we can check these assumptions.

The one-way analysis of variance separates the total variability in the data into that which can be attributed to differences between the individuals from the different groups (the ***between-group variation***), and to the random variation between the individuals within each group (the ***within-group variation***, sometimes called ***unexplained*** or ***residual variation***). These components of variation are measured using variances, hence the name ***analysis of variance*** (***ANOVA***). Under the null hypothesis that the group *means* are the same, the between-group variance will be similar to the within-group variance. If, however, there are differences between the groups, then the between-group variance will be larger than the within-group variance. The test is based on the ratio of these two variances.

The Kruskal–Wallis test

This non-parametric test is an extension of the Wilcoxon rank sum test. Under the null hypothesis of no differences in the distributions between the groups, the sums of the ranks in each of the k groups should be comparable after allowing for any differences in sample size.

1. Define the null and alternative hypotheses under study

H_0: each group has the same distribution of values in the population

H_1: each group does not have the same distribution of valuesin the population.

2. Collect relevant data from samples of individuals

3. Calculate the value of the test statistic specific to H_0

4. Compare the value of the test statistic to values from a known probability distribution

5. Interpret the P-value and results

Categorical data: a single proportion

We have a single sample of n individuals; each individual either 'possesses'a characteristic of interest (e.g. is male, is pregnant, has died) or does not possess that characteristic (e.g. is female, is not pregnant, is still alive).

The test of a single proportion

Our sample of individuals is selected from the population of interest. Each individual either has or does not have the particular characteristic.

1. Define the null and alternative hypotheses under study

H_0: the population proportion p is equal to a particular value p_1

H_1: the population proportion p is not equal to p_1.

2. Collect relevant data from a sample of individuals

3. Calculate the value of the test statistic specific to H_0

4. Compare the value of the test statistic to values from a known probability distribution

5 Interpret the P-value and results

We can use this confidence interval to assess the clinical or biological importance of the results. A wide confidence interval is an indication that our estimate has poor precision.

Independent groups: the Chi-squared test

The data are obtained, initially, as ***frequencies***, i.e. the numbers with and without the characteristic in each sample. A table in which the entries are frequencies is called a ***contingency table***; when this table has *two rows* and *two columns* it is called a ***2 × 2 table***. Table 4 shows the ***observed*** frequencies in the four cells corresponding to each row/column combination, the four ***marginal totals*** (the frequency in a specific row or column, e.g. a + b), and the ***overall total***, *n*. We can calculate the frequency that we would expect in each of the four cells of the table if H_0 were true (the ***expected frequencies***).

1. Define the null and alternative hypotheses under study

H_0: the proportions of individuals with the characteristic are equal in the two groups in the population

H_1: these population proportions are not equal.

2. Collect relevant data from samples of individuals

3. Calculate the value of the test statistic specific to H_0

4. Compare the value of the test statistic to values from a known probability distribution

5. Interpret the P-value and results

If the assumptions are not satisfied

If E < 5 in any one cell, we use ***Fisher's exact test*** to obtain a P-value that does not rely on the approximation to the Chi-squared distribution.

Related groups: McNemar's test

The two groups are related or dependent, e.g. each individual may be measured in two different circumstances. Every individual is classified according to whether the characteristic is present in both circumstances, one circumstance only, or in neither

1. Define the null and alternative hypotheses under study

H_0: the proportions with the characteristic are equal in the two groups in the population

H_1: these population proportions are not equal.

2. Collect relevant data from two samples

3. Calculate the value of the test statistic specific to H_0.

4. Compare the value of the test statistic with values from a known probability distribution

5. Interpret the P-value and results

Categorical data: more than two categories

Chi-squared test: large contingency tables

Individuals can be classified by two factors. For example, one factor may represent disease severity (mild, moderate or severe) and the other factor may represent blood group (A, B, O, AB). We are interested in whether the two factors are associated. Are individuals of a particular blood group likely to be more severely ill?

Assumptions

The data may be presented in an $r \times c$ contingency table with r rows and c columns. The entries in the table are ***frequencies***; each cell contains the number of individuals in a particular row and a particular column. Every individual is represented once, and can only belong in one row and in one column, i.e. the categories of each factor are mutually exclusive. At least 80% of the expected frequencies are greater than or equal to 5.

1. Define the null and alternative hypotheses under study

H_0: there is no association between the categories of one factor and the categories of the other factor in the population

H_1: the two factors are associated in the population.

2. Collect relevant data from a sample of individuals

3. Calculate the value of the test statistic specific to H_0

4. Compare the value of the test statistic to values from a known probability distribution

5. Interpret the P-value and results

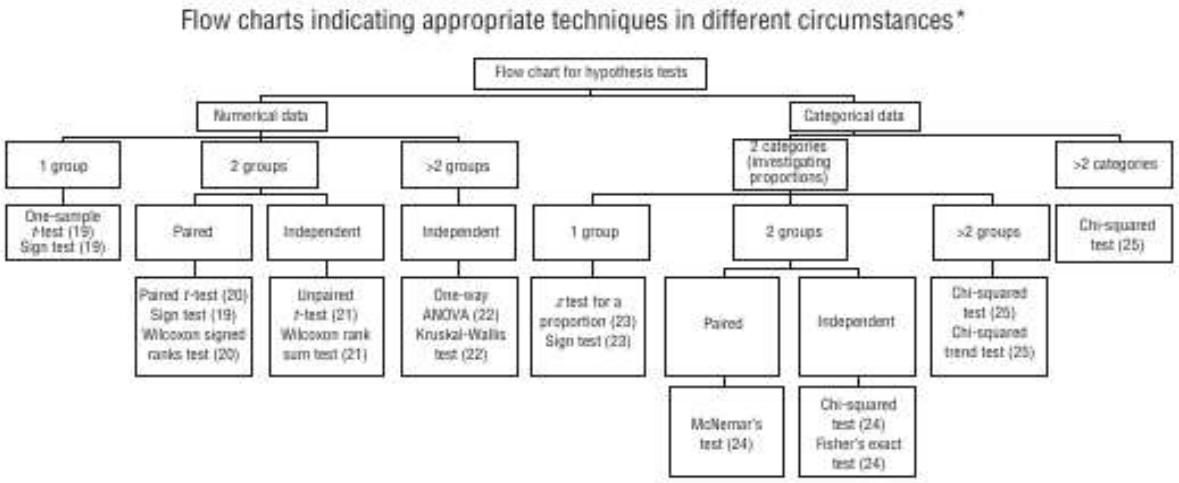

Fig. 6.10. Flow chart indicating appropriate techniques in different circumstances.

Control questions

1. What is hypothesis testing?
2. List stages when carrying out a hypothesis test.
3. What are the null hypothesis and the alternative hypothesis?
4. What are parametric and non-parametric tests?
5. Which parametrical tests for numerical data do you know?
6. Which non-parametrical tests for numerical data do you know?
7. Which tests for categorical data do you know?

References

1. Medical statistics at a glance / Aviva Petrie, Caroline Sabin.—2nd ed., 2005. – 157 p.

2. Using and understanding medical statis tics / David E. Matthews Vernon T. Farewell. – 4th, completely rev. and enl. ed., 2007. – 322 p.

3. Handbook of statistics. Epidemiology and Medical Statistics / C.R. Rao, J.P. Miller, D.C. Rao, 2008. – 852 p.