INSTITUTE OF ACTUARIAL & QUANTITATIVE STUDIES

Class: MSc

Subject : Probability and Statistics -2

Chapter: Unit 3 Chapter 2

Chapter Name: Hypothesis Testing - 1

Index

Index

Index

1 Introduction

- In many research areas, such as medicine, education, advertising and insurance, it is necessary to carry out statistical tests. These tests enable researchers to use the results of their experiments to answer questions such as:
  - Is drug ***A*** a more effective treatment for AIDS than drug ***B***?
  - Does training program T lead to improved staff efficiency?
  - Are the severities of large individual private motor insurance claims consistent with a lognormal distribution?

- A hypothesis is where we make a statement about something; for example the mean lifetime of smokers is less than that of non-smokers. A hypothesis test is where we collect a representative sample and examine it to see if our hypothesis holds true.

2 Hypothesis

Hypothesis: late 16th century: via late Latin from Greek hypothesis ‘foundation’, from hypo ‘under’ + thesis ‘placing’.

A statistical hypothesis is a hypothesis that is testable on the basis of observed data modelled as the realized values taken by a collection of random variables.

3 Testing of Hypothesis

The standard approach to carrying out a statistical test involves the following steps:

- specify the hypothesis to be tested
- select a suitable statistical model
- design and carry out an experiment/study
- calculate a test statistic
- calculate the probability value
- determine the conclusion of the test

3 Testing of Hypothesis

Null Hypothesis H_0

- The null hypothesis states that a population parameter (such as the mean, the standard deviation, and so on) is equal to a hypothesized value. The null hypothesis is often an initial claim that is based on previous analyses or specialized knowledge.

- The basic hypothesis being tested is the null hypothesis, denoted H_0 – it can sometimes be regarded as representing the current state of knowledge or belief about the value of the parameter being tested (the 'status quo' hypothesis). In many situations a difference between two populations is being tested and the null hypothesis is that there is no difference.

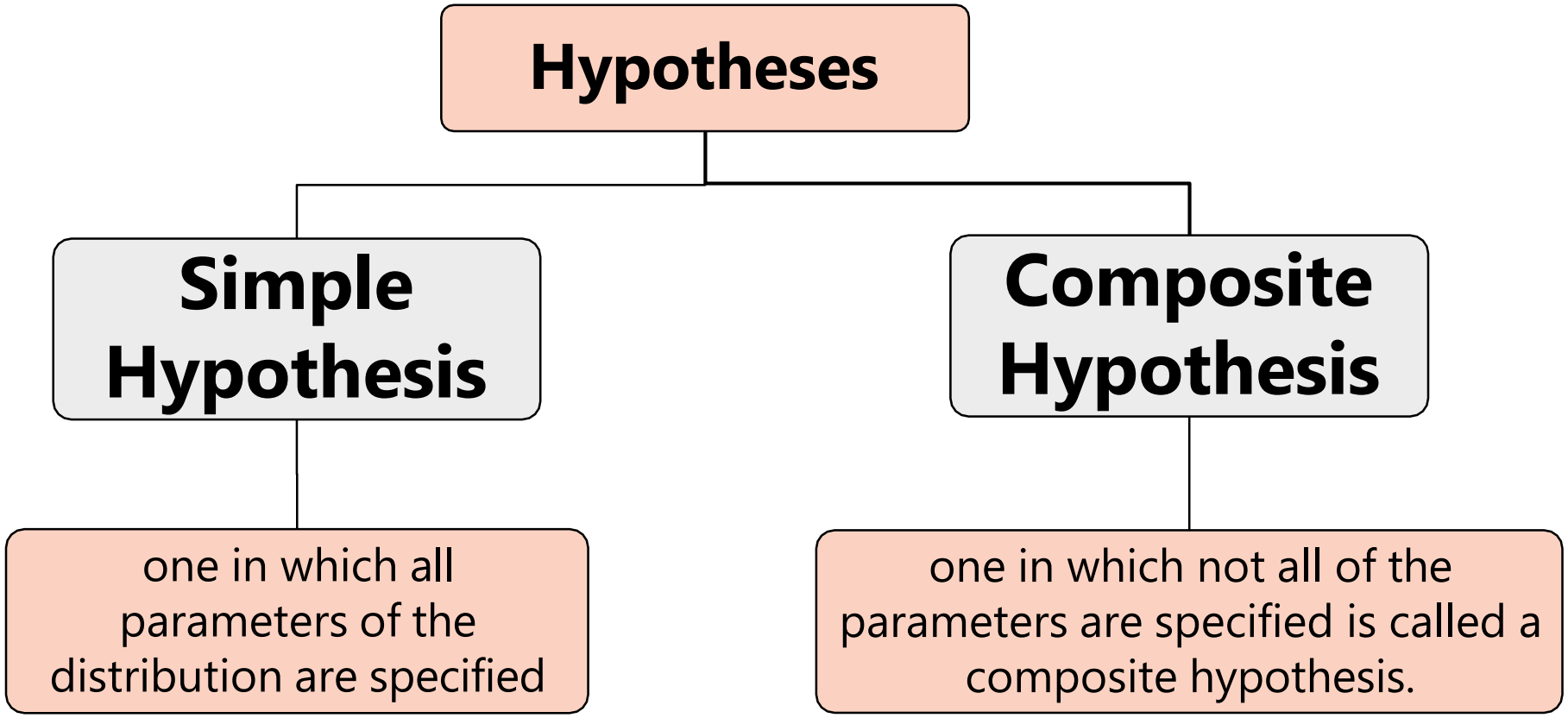
Alternate Hypothesis H_1

- The alternative hypothesis states that a population parameter is smaller, greater, or different than the hypothesized value in the null hypothesis. The alternative hypothesis is what you might believe to be true or hope to prove true.

- In a test, the null hypothesis is contrasted with the alternative hypothesis, denoted H_1.

- The null and alternative hypotheses are two mutually exclusive statements about a population. A hypothesis test uses sample data to determine whether to reject the null hypothesis.

4 Types of Hypotheses



4 Types of Hypotheses

Case I
- Normal Distribution: H_0: $\mu = 175, \sigma^2 < 4$

Case II
- Normal Distribution: H_0: $\mu = 175, \sigma^2 = 9$

Case III
- Binomial Distribution : n=12, p=0.5

Case IV
- Binomial Distribution : n =12, p$\leq$ 0.5

5 Test

- A test is a rule which divides the sample space (the set of possible values of the data) into two subsets, a region in which the data are consistent with H_0 , and its complement, in which the data are inconsistent with H_0 . The tests discussed here are designed to answer the question 'Do the data provide sufficient evidence to justify our rejecting H_0 ?'.

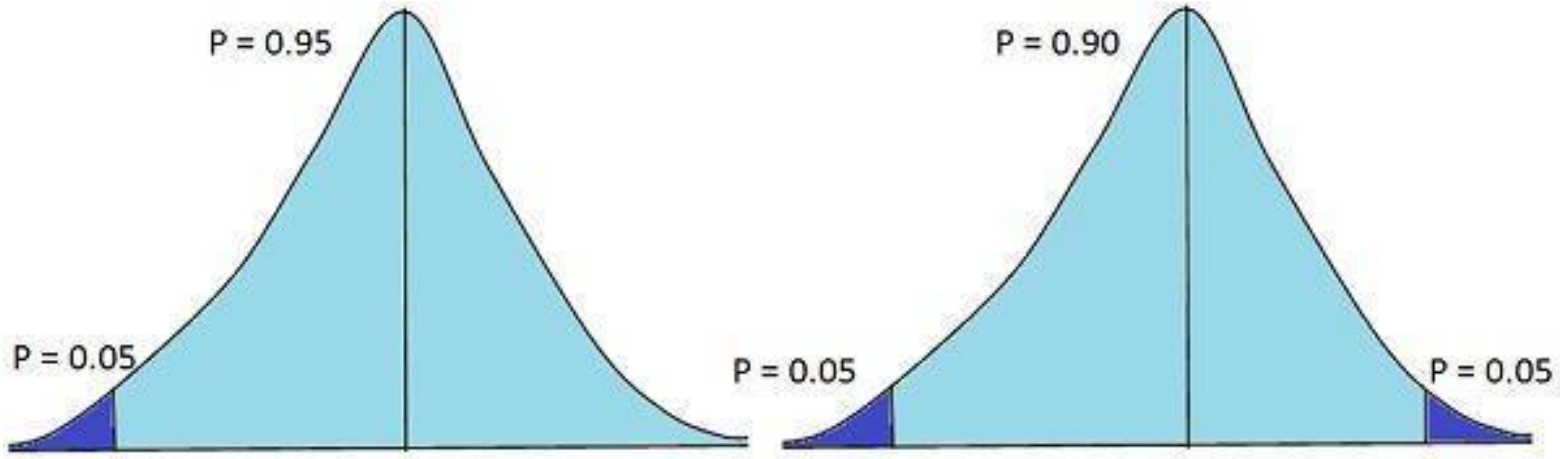
6 One-sided and two-sided tests

One-Tailed Test

- A one-tailed test is a statistical test in which the critical area of a distribution is one-sided so that it is either greater than or less than a certain value, but not both.

Two-Tailed Test

- A two-tailed test, in statistics, is a method in which the critical area of a distribution is two-sided and tests whether a sample is greater than or less than a certain range of values..



One-tailed Test Vs Two-tailed Test

6 One-sided and two-sided tests

- In a test of whether smoking reduces life expectancies, the hypotheses would be:
  - H_0 : smoking makes no difference to life expectancy
  - H_1 : smoking reduces life expectancy

- This is an example of a one-sided test, since we are only considering the possibility of a reduction in life expectancy ie a change in one direction. However we could have specified the hypotheses:

  - H_0 : smoking makes no difference to life expectancy
  - H_1 : smoking affects life expectancy

- This is a two-sided test, since the alternative hypothesis considers the possibility of a change in either direction, ie an increase or a decrease.

Example

Which Test would you use?

- Testing a new drug against an existing treatment.
- A certain course claiming 50% higher chances of employment after completion.
- There are two movies that caught your eye, but you're not really sure which one is better.

7 Test Statistics

A test statistic is a statistic (a quantity derived from the sample) used in statistical hypothesis testing.

- A hypothesis test is typically specified in terms of a test statistic, considered as a numerical summary of a data-set that reduces the data to one value that can be used to perform the hypothesis test.
- In general, a test statistic is selected or defined in such a way as to quantify, within observed data, behavior that would distinguish the null from the alternative hypothesis, where such an alternative is prescribed, or that would characterize the null hypothesis if there is no explicitly stated alternative hypothesis.
- The actual decision is based on the value of a suitable function of the data, the test statistic. The set of possible values of the test statistic itself divides into two subsets, a region in which the value of the test statistic is consistent with H_0 , and its complement, the critical region (or rejection region), in which the value of the test statistic is inconsistent with H_0.
- If the test statistic has a value in the critical region, H_0 is rejected. The test statistic (like any statistic) must be such that its distribution is completely specified when the value of the parameter itself is specified (and in particular 'under H_0 ' ie when H_0 is true).

8 Level of Significance (α)

The level of significance is defined as the fixed probability of wrong elimination of null hypothesis when in fact, it is true.

- The level of significance is the measurement of the statistical significance. It defines whether the null hypothesis is assumed to be accepted or rejected.
- It is expected to identify if the result is statistically significant for the null hypothesis to be false or rejected.

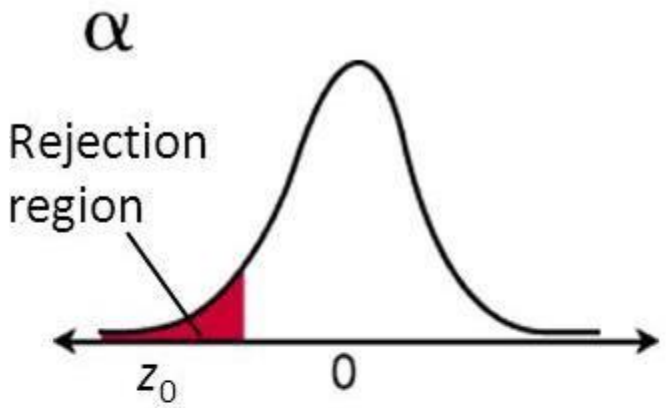
- The level of significance is stated to be the probability of type I error and is preset by the researcher with the outcomes of error.

9 Critical Region

A critical region, also known as the rejection region, is a set of values for the test statistic for which the null hypothesis is rejected.

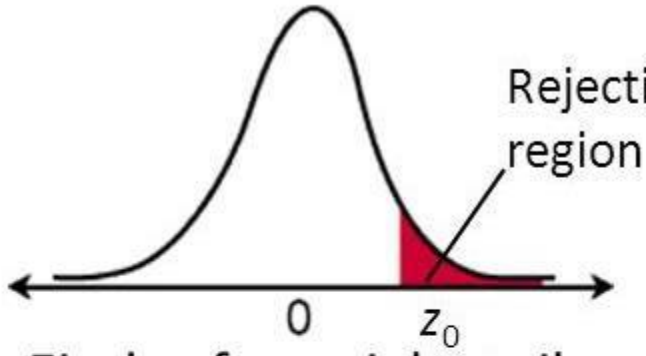
- If the observed test statistic is in the critical region then we reject the null hypothesis and accept the alternative hypothesis.

9 Critical Region



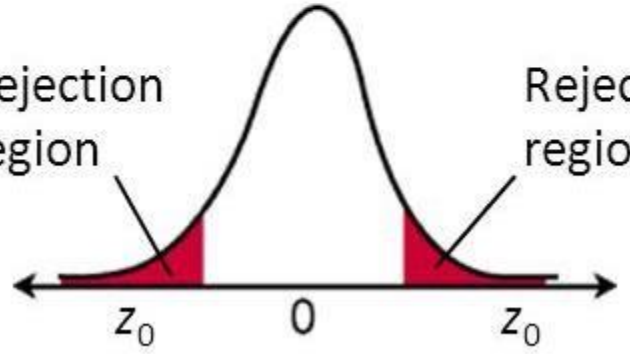
Find z_0 for a left-tail test with $\alpha = .01$.

$z_0 = -2.33$



Find z_0 for a right-tail test with $\alpha = .05$.

$z_0 = 1.645$



$-z_0 = -2.575$
and $z_0 = 2.575$

Find $-z_0$ and z_0 for a two-tail test with α.

10 Errors & Power

- The level of significance of the test, denoted α , is the probability of committing a Type I error, ie it is the probability of rejecting H_0 when it is in fact true.
- The probability of committing a Type II error, denoted β , is the probability of accepting H_0 when it is false.
- An ideal test would be one which simultaneously minimises α and β – this ideal however is not attainable in practice.
- The power of a test is the probability of rejecting H_0 when it is false, so that the power equals $1 - \beta$.
- In general, this will be a function of the unknown parameter value. For simple hypotheses the power is a single value, but for composite hypotheses it is a function being defined at all points in the alternative hypothesis.

11 Best Tests

- The classical approach to finding a 'good' test (called the Neyman-Pearson theory) fixes the value of α , ie the level of significance required and then tries to find such a test for which the other error probability, β , is as small as possible for every value of the parameter specified by the alternative hypothesis. This can also be described as finding the 'most powerful' test.

- The key result in the search for such a test is the Neyman-Pearson lemma, which provides the 'best' test (smallest β) in the case of two simple hypotheses. For a given level, the critical region (and in fact the test statistic) for the best test is determined by setting an upper bound on the likelihood ratio L_0/L_1 , where L_0 and L_1 are the likelihood functions of the data under H_0 and H_1 respectively.

12 The Neyman-Pearson lemma

- Formally, if C is a critical region of size α and there exists a constant k such that $\frac{L_0}{L_1} \leq k$ inside C and $\frac{L_0}{L_1} \geq k$ outside C, then C is a most powerful critical region of size α for testing the simple hypothesis $\theta = \theta_0$ against the simple alternative hypothesis $\theta = \theta_1$.

- So a Neyman-Pearson test rejects H_0 if:

$$\frac{\boldsymbol{Likelihood\ under\ H_0}}{\boldsymbol{Likelihood\ under\ H_1}} < \boldsymbol{critical\ value}$$

- Common tests are often such that the null hypothesis is simple, eg $H_0: \theta = \theta_0$, against a composite alternative, eg $H_1: \theta \neq \theta_0$, which is two-sided, and $H_1: \theta > \theta_0$ or $H_1: \theta < \theta_0$, which are one-sided.

- Here it is only in certain special cases (usually one-sided cases) that a single test is available which is best (ie uniformly most powerful) for all parameter values. In cases where a single best test in the sense of the Neyman-Pearson Lemma is unavailable, another approach is used to derive sensible tests. This approach, which is a generalisation of the Lemma, produces tests which are referred to as likelihood ratio tests.

13 Likelihood ratio tests

- The critical region (and test statistic) for the test are determined by setting an upper bound on the ratio (max L_0 /max L), where max L_0 is the maximum value of the likelihood L under the restrictions imposed by the null hypothesis, and max L is the overall maximum value of L for all allowable values of all parameters involved.

- In the most common case when H_0 and H_1 together cover all possible values for the parameters, this generalised test rejects H_0 if:

$$\frac{\max(\boldsymbol{Likelihood\ under\ H_0})}{\max(\boldsymbol{Likelihood\ under\ H_0 + H_1})} < \boldsymbol{critical\ value}$$

13 Likelihood ratio tests

- Important results include the case of sampling from a N(μ, σ^2) distribution. The method leads to the test statistic:

$$\frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1} \quad under\ H_0: \mu = \mu_0$$

- for tests on the value of the mean μ .

- We're assuming here that σ^2 is unknown. If it is known, then the z-test is the 'best' test.
- The method also leads to the test statistic::

$$\frac{\left((n-1)S^2\right)}{\sigma_0^2} \sim \chi^2_{n-1} \quad under\ H_0: \sigma^2 = \sigma_0^2$$

- for tests on the value of the variance σ^2.

14 P-values

- Under the 'classical' Neyman-Pearson approach, with a fixed predetermined value of α , a test will produce a decision as to whether to reject H_0 . But merely comparing the observed test statistic with some critical value and concluding eg 'using a 5% test, reject H_0 ' or 'reject H_0 with significance level 5%' or 'result significant at 5%' (all equivalent statements) does not provide the recipient of the results with clear detailed information on the strength of the evidence against H_0 .

- A more informative approach is to calculate and quote the probability value (p-value) of the observed test statistic. This is the observed significance level of the test statistic – the probability, assuming H_0 is true, of observing a test statistic at least as 'extreme' (inconsistent with H_0) as the value observed.

14 P-values

- The p-value is the lowest level at which H_0 can be rejected.
- The smaller the p-value, the stronger is the evidence against the null hypothesis.
- For example, when testing $H_0: \theta = 0.5$ vs $H_1: \theta = 0.4$, where θ is the probability of a coin coming up heads, and 82 heads have been observed in 200 tosses, the p-value of the result is:

$$\boldsymbol{P(X \leq 82) \ \ where \ \ X \sim Bin(200, 0.5)}$$

$$\boldsymbol{P\left(Z < \frac{82.5 - 100}{\sqrt{50}}\right) = P(Z < -2.475) = 0.0067}$$

- H_0 is therefore extremely unlikely – probability < 0.01– and there is very strong evidence against H_0 and in favour of H_1. A good way of expressing the result is: 'we have very strong evidence against the hypothesis that the coin is fair (p-value 0.007) and conclude that it is biased against heads'.

14 P-values

- Testing does not prove that any hypothesis is true or untrue. Failure to detect a departure from H_0 means that there is not enough evidence to justify rejecting H_0 , so H_0 is accepted in this sense only, whilst realising that it may not be true. This attitude to the acceptance of H_0 is a feature of the fact that H_0 is usually a precise statement, which is almost certainly not exactly true.

15 Testing the value of a population mean

- **Situation**: random sample, size n, from N(μ, σ^2) – sample mean $\bar{X}$
- **Testing:** $H_0: \mu = \mu_0$

$$(a)\ \sigma\ known\ :test\ statistic\ is\ \bar{X},\quad and\ \frac{(\bar{X}-\mu_0)}{\sigma/\sqrt{n}} \sim N(0,1)\ under\ H_0$$

$$(b)\ \sigma\ unknown\ :test\ statistic\ is\ \frac{(\bar{X}-\mu_0)}{S/\sqrt{n}} \sim t_{n-1}\ under\ H_0$$

- For large samples, N(0,1) can be used in place of t_{n-1} . Further, the Central Limit Theorem justifies the use of a normal approximation for the distribution of $\bar{X}$ in sampling from any reasonable population, and s^2 is a good estimate of σ^2 , so the requirement that we are sampling from a normal distribution is not necessary in either case (a) or (b) when we have a large sample.

16 Testing the value of a population variance

- **Situation**: random sample, size n, from N(μ, σ^2) – sample variance S^2.
- **Testing**: $H_0: \sigma^2 = \sigma_0^2$
- **Test statistic** is $\frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1}$ $under\ H_0$
- For large samples, the test works well even if the population is not normally distributed.

17 Testing the value of a population proportion

- **Situation**: n binomial trials with P(success) = p ; we observe x successes.
- **Testing**: $H_0 : p = p_0$.
- **Test statistic** is X ~ Bin(n,p_0)under H_0 .
- For large n, use the normal approximation to the binomial (with continuity correction), ie use:

$$\frac{\left(\frac{\left(X \pm \frac{1}{2}\right)}{n} - p\right)}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$$

- or:

$$\frac{X \pm \frac{1}{2} - np}{\sqrt{np(1-p)}} \sim N(0,1)$$

18 Testing the value of the mean of a Poisson distribution

- **Situation**: random sample, size n, from Poi (λ) distribution.
- **Testing:** $H_0: \lambda = \lambda_0$
- **Test statistic** is sample sum $\sum X_i \sim Poi(n\lambda_0)$ under H_0 . In the case where n is small and $n\lambda_0$ is of moderate size, probabilities can be evaluated directly (or found from tables, if available).
- For large samples (or indeed whenever the Poisson mean is large) a normal approximation can be used for the distribution of the sample sum or sample mean. Recall that
- $\sum X_i \sim Poi(n\lambda) \rightarrow N(n\lambda, n\lambda)$.
- Test statistic is $\bar{X}$,and $\frac{\bar{X}-\lambda_0}{\sqrt{\lambda_0/n}} \sim \boldsymbol{N(0,1)}$ under H_0.
- or we can use $\sum X_i$, and $\frac{\sum X_i - n\lambda_0}{\sqrt{n\lambda_0}} \sim \boldsymbol{N(0,1)}$ $under$ H_0.
- Using the second version it is easier to incorporate a continuity correction. The first version has continuity correction 0.5/n , whereas the second version has continuity correction 0.5.

19 Testing the value of the difference between two population means

- **Situation**: independent random samples, sizes n_1 and n_2 from $N(\mu_1, \sigma_1^2), N(\mu_2, \sigma_2^2)$ respectively.
- **Testing:** $H_0: \mu_1 - \mu_2 = \delta$
- (a) σ_1^2, σ_2^2 known

$$\textbf{Test statistic: } z = \frac{\bar{x}_1 - \bar{x}_2 - \delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

- (b) σ_1^2, σ_2^2 unknown-much the more usual situation

19 Testing the value of the difference between two population means

- Large samples: use S_i^2 to estimate σ_i^2. We will now use a t distribution.
- Further, the Central Limit Theorem justifies the use of a normal approximation for the distribution of the test statistic in sampling from any reasonable populations, so the requirement that we are sampling from normal distributions is not necessary when we have large samples.
- Small samples: under the assumption $\sigma_1^2 = \sigma_2^2 (= \sigma^2$ say) , this common variance is estimated by Sp^2, and the **test statistic** is $\boldsymbol{t = \frac{\bar{x}_1 - \bar{x}_2 - \delta}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}}$(which is distributed as t with $n_1 + n_2 - 2$ degrees of freedom under H_0 .
- Remember that $\boldsymbol{s_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$

20 Testing the value of the ratio of two population means

- **Situation**: independent random samples, sizes n_1 and n_2 from $N(\mu_1, \sigma_1^2), N(\mu_2, \sigma_2^2)$ respectively. Sample variances S_1^2 and S_2^2.
- **Testing**: $H_0: \sigma_1^2 = \sigma_2^2 \quad vs \quad H_1: \sigma_1^2 \neq \sigma_2^2$
- This test is a formal prerequisite for the two-sample t test, for which the assumption $\sigma_1^2 = \sigma_2^2$ is required. In practice, however, a simple plot of the data is often sufficient to justify the assumption – only if the population variances are very different in size is there any problem with the t test.
- **Test statistic**: $S_1^2/S_2^2 \sim F_{n_1-1, n_2-1}$ under H_0**.**

21 Testing the value of the difference between two population proportions

- Both one-sided and two-sided tests can easily be performed on the difference between two binomial probabilities – at least for large samples.

- **Situation**:

  n_1 (large) trials with P(success) = p_1; observe x_1 successes.
  n_2 (large) trials with P(success) = p_2 ; observe x_2 successes.

- **Testing**: H_0: $p_1 = p_2$

- **Test statistic** $\frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}} \sim N(0,1)$ under H_0.

- Where $\hat{p}_1, \hat{p}_2$ are the maximum likelihood estimates (MLEs) of p_1 and p_2 respectively, (the sample proportions $\left(\frac{X_1}{n_1}, \frac{X_2}{n_2}\right)$ and p is the MLE of the common p under the null hypothesis, which is the overall sample proportion, namely $\frac{X_1+X_2}{n_1+n_2}$.

22 Testing the value of the difference between two Poisson means

- **Situation**: independent random samples, sizes n_1 and n_2 , from Poi(λ_1) and Poi(λ_2) distributions. Considering the case in which normal approximations can be used – which is so whenever the sample sizes are large and/or the parameter values are large:

- **Testing**: H_0: $\lambda_1 = \lambda_2$

- **Test statistic**: $\frac{(\hat{\lambda}_1 - \hat{\lambda}_2)}{\sqrt{\frac{\hat{\lambda}}{n_1} + \frac{\hat{\lambda}}{n_2}}} \sim \boldsymbol{N(0,1)}$ under H_0.

- where are the maximum likelihood estimates (MLEs) (the sample means $\bar{X}_1 \bar{X}_2$respectively) and $\hat{\lambda}$ is the MLE of the common λ under the null hypothesis, which is the overall sample mean.

23 Basic test – paired data

- In testing for a difference between two population means, the use of independent samples can have a major drawback. Even if a real difference does exist, the variability among the responses within each sample can be large enough to mask it. The random variation within the samples will mask the real difference between the populations from which they come.

- One way to control this variability external to the issue in question is to use a pair of responses from each subject, and then work with the differences within the pairs. The aim is to remove as far as possible the subject-to-subject variation from the analysis, and thus to 'home in' on any real difference between the populations.

- **Assumption**: differences constitute a random sample from a normal distribution.

- **Testing**: H_0: $\mu_D\,(=\mu_1-\mu_2\,)=\delta$

- **Test statistic**: $\frac{\bar{D}-\delta}{S_D/\sqrt{n}}\sim\boldsymbol{t_{n-1}}$ under H_0.

- We can use *N(0,1)* for t, and do not require the 'normal' assumption, if *n* is large.

24 Tests and confidence intervals

- There are very close parallels between the inferential methods for tests and confidence intervals. In many situations there is a direct link between a confidence interval for a parameter and tests of hypothesised values for it.

- A confidence interval for θ can be regarded as a set of acceptable hypothetical values for θ , so a value θ_0 contained in the confidence interval should be such that the hypothesis $H_0: \theta = \theta_0$ will be accepted in a corresponding test. This generally proves to be the case.

- In some situations there is a difference between the manner of construction of the confidence interval and that of the construction of the test statistic which is actually used. For example the confidence interval for the difference between two proportions (based on normal approximations) is constructed in a different way from that used for the test statistic in the corresponding test, where an estimate of a common proportion (under H_0) is used.

- As a result, in this and similar cases there is only an approximate match (albeit a good one) between the confidence interval and the corresponding test.

25 Non-parametric tests

- The tests we have been considering so far all make assumptions about the distribution of the variables of interest within the population. If these assumptions are not correct, then the level of statistical significance can be affected.

- It is possible to devise tests which make no distributional assumptions. Such tests are termed non-parametric. They have the advantages of being applicable under conditions in which the tests in the previous sections should not be used.

26 Chi-square tests

- These tests are relevant to category or count data. Each sample value falls into one or other of several categories or cells. The test is then based on comparing the frequencies actually observed in the categories/cells with the frequencies expected under some hypothesis, using the test statistic

$$\sum \frac{(\boldsymbol{f_i} - \boldsymbol{e_i})^2}{\boldsymbol{e_i}}$$

- where f_i and e_i are the observed and expected frequencies respectively in the ith category/cell, and the summation is taken over all categories/cells involved. This statistic has, approximately, a chi-square (χ^2) distribution under the hypothesis on the basis of which the expected frequencies were calculated.

26 Chi-square tests

Goodness of fit

- This is investigating whether it is reasonable to regard a random sample as coming from a specified distribution, ie whether a particular model provides a 'good fit' to the data.

Degrees of freedom

- Suppose there are k cells, so k terms in the summation which produces the statistic, and that the sample size is $n = \sum f_i$. The expected frequencies also sum to n, so knowing any k-1 of them automatically gives you the last one. There is a dependence built in to the k terms which are added up to produce the statistic – and this is the reason why the degrees of freedom of the basic statistic is k-1 and not k.

- Further, for each parameter of the distribution specified by the null hypothesis which must be estimated from the observed data, another degree of dependence is introduced in the expected frequencies – for each parameter estimated another degree of freedom is lost. The theory behind this assumes that the maximum likelihood estimators are used. So the number of degrees of freedom is reduced by the number of parameters estimated from the observed data.

26 Chi-square tests

The 'accuracy' of the chi-square approximation

- The test statistic is only approximately, not exactly, distributed as χ^2 . The presence of the expected frequencies e_i in the denominators of the terms to be added up is important — dividing by very small e_i values causes the resulting terms to be somewhat large and 'erratic', and the tail of the distribution of the statistic may not match that of the χ^2 distribution very well. So, in practice, it is best not to have too many small e_i values, which can be done by combining cells and suffering the consequent loss of information/degrees of freedom. The most common recommendation is not to use any e_i which is less than 5.

- (However, the statistic is more robust than that and in practice a less conservative approach, such as ensuring that all e_i are greater than 1 and that not more than 20% of them are less than 5, may be taken.)

27 Contingency tables

- A contingency table is a two-way table of counts obtained when sample items (people, companies, policies, claims etc) are classified according to two category variables. The question of interest is whether the two classification criteria are independent.

- H_0: the two classification criteria are independent.

- The simple rule for calculating the expected frequency for any cell is then:

$$\frac{\textbf{row total x column total}}{\textbf{table total}}$$

- (ie the proportion of data in row i is $f_i./f$ so if the criteria are independent, the number expected in cell (i, j) is (f_i./f) x f_j)

- The degrees of freedom associated with a table with r rows and c columns is:

$$(rc - 1) - (r - 1) - (c - 1) = (r - 1)(c - 1)$$

- since the column totals and row totals reduce the number of degrees of freedom.

- An important use of this method is with a table of dimension 2 x c (or r x 2) which gives a test for differences among 2 or more population proportions.