Lecture

IAQS | INSTITUTE OF ACTUARIAL & QUANTITATIVE STUDIES

Class: FY BSc

Subject : Probability and Statistics -2

Subject Code: PUSASQ 1.2

Chapter: Unit 2 Chp 2

Chapter Name: Hypothesis Testing - 1

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 hupothesis ‘foundation’, from hupo ‘under’ + thesis ‘placing’.

A statistical hypothesis is a hypothesis that is testable on the basis of observed data modelled as the realised 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.

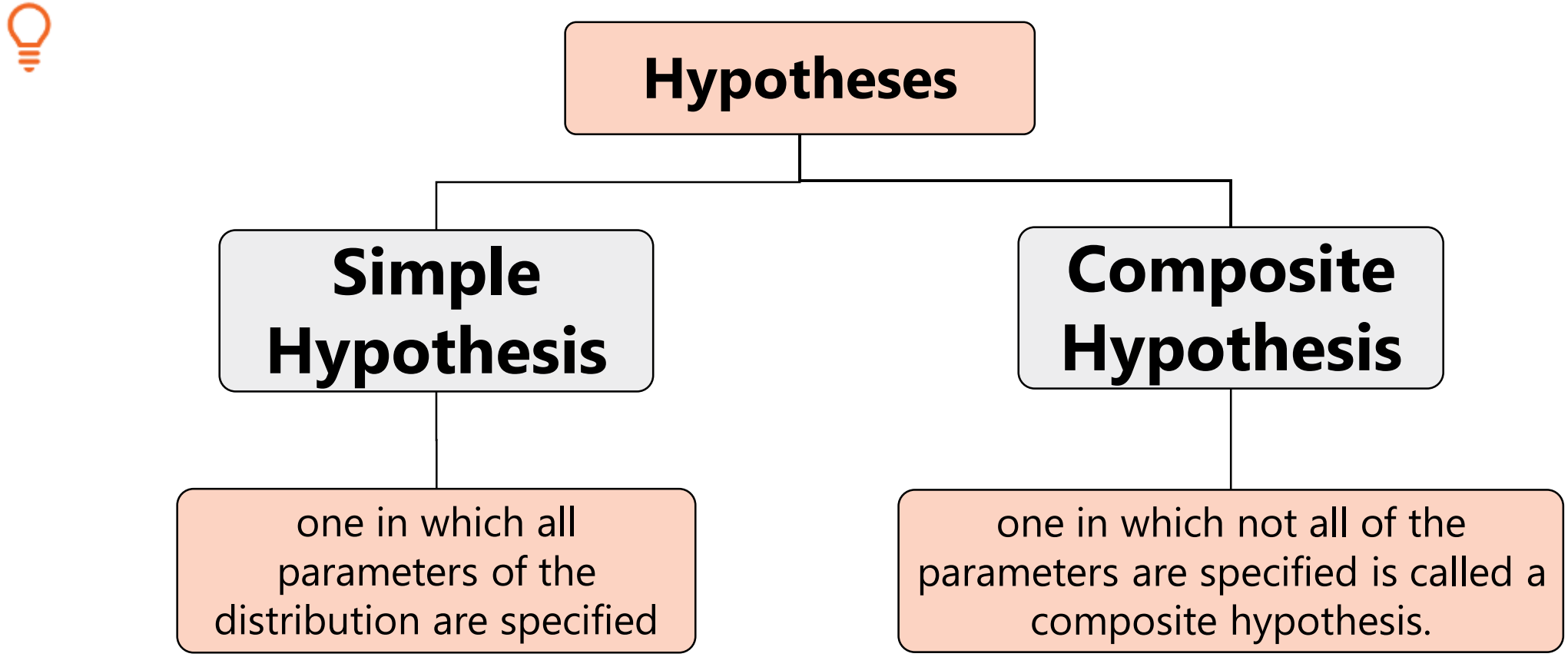
Alternate Hypothesis H_1

- 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 ?'.

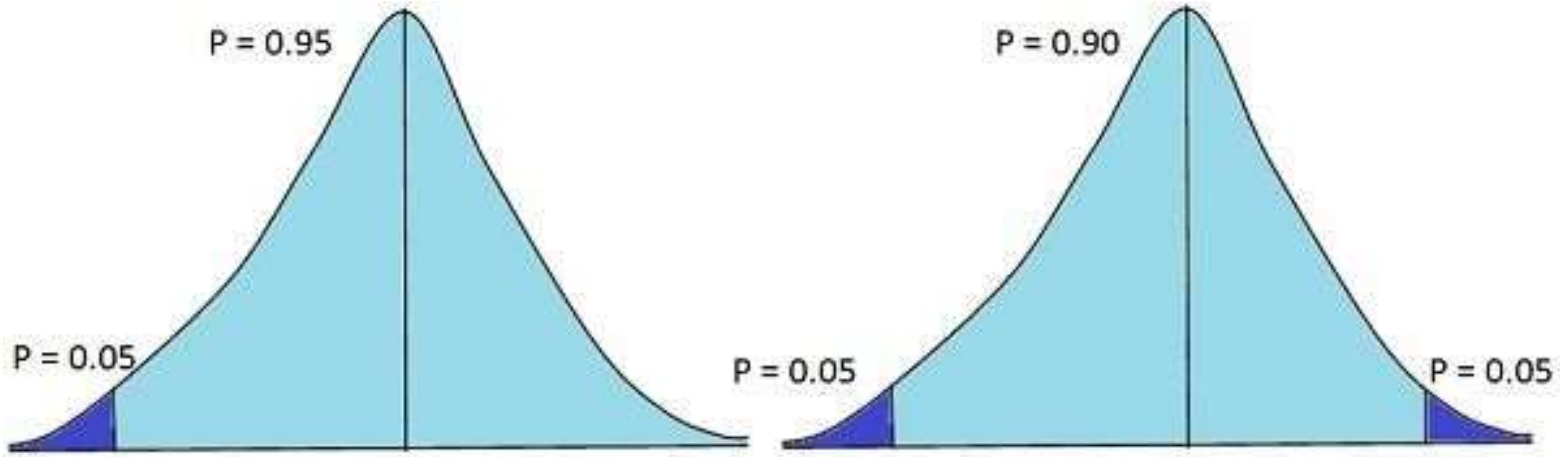
6 One-sided and two-sided tests

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.

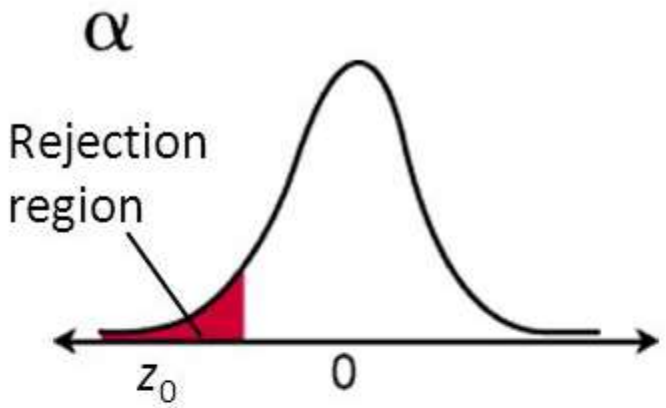
- The level of significance is stated to be the probability of type I error and is preset by the researcher with the outcomes of error.

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.

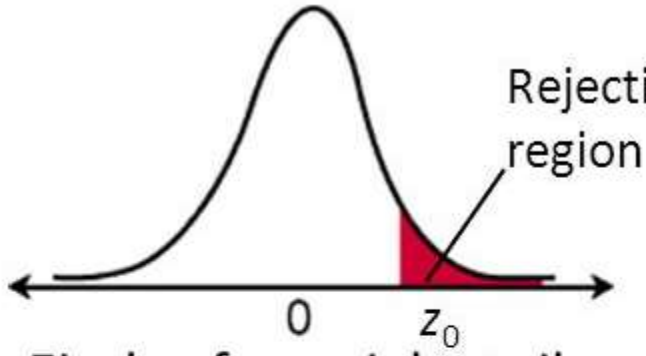
- If the observed test statistic is in the critical region then we reject the null hypothesis and accept the alternative hypothesis.

9 Critical Region



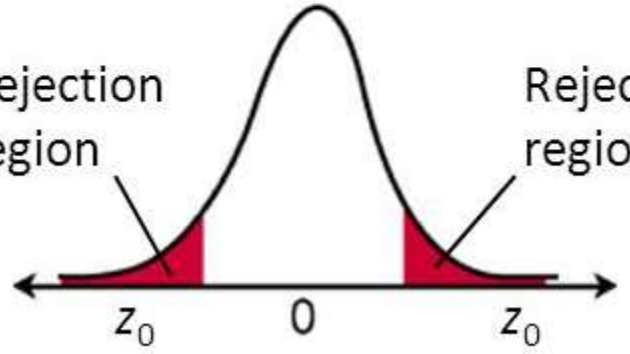
Find z_0 for a left-tail test with $\alpha = .01$.

$z_0 = -2.33$



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 = .

Questions

The average IQ of a sample of 50 university students was found to be 105. Carry out a statistical test to determine whether the average IQ of university students is greater than 100, assuming that IQs are normally distributed. It is known from previous studies that the standard deviation of IQs among students is approximately 20.

Questions

The annual rainfall in centimetres at a certain weather station over the last ten years has been as follows:

17.2 28.1 25.3 26.2 30.7 19.2 23.4 27.5 29.5 31.6

Assuming this data is taken from a normal distribution test at the 5% level whether the standard deviation of the annual rainfall at the weather station is equal to 4 cm.

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.

10 Errors & Power

Question

A random variable X is believed to follow an $Exp(\lambda)$ distribution. In order to test the null hypothesis $\mu = 20$ against the alternative hypothesis $\mu = 30$, where $\mu = 1/\lambda$, a single value is observed from the distribution. If this value is less than 28, H_0 is accepted, otherwise H_0 is rejected. Calculate the probabilities of:

(i) a Type I error

(ii) a Type II error.

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:

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

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:** $H_0: \mu = \mu_0$

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

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

- 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 $\frac{(\boldsymbol{n-1})\boldsymbol{S^2}}{\boldsymbol{\sigma_0^2}} \sim \boldsymbol{\chi^2_{n-1}}$ $under\ H_0$
- 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 Goodness of fit test

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.

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

where f_i and e_i are the observed and expected frequencies respectively in the i^{th} 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.

19 Goodness of fit test

In testing whether a die is fair, a suitable model is:

$$P(X = i) = \frac{1}{6}, \quad i = 1,2,3,4,5,6 \text{ where } X \text{ is the number thrown}$$

and the hypotheses may be:

H_0 : Number thrown has the distribution specified in the model

H_1 : Number thrown does not have the distribution specified in the model

If the die is thrown 300 times, with the following results,

x:	1	2	3	4	5	6
f_i :	43	56	54	47	41	59

Carry out a χ^2 test to determine whether the data comes from a fair die.

19 Goodness of fit test

The numbers of claims made last year by individual motor insurance policyholders were:

Number of claims	0	1	2	3	4+
Number of policyholders	2,962	382	47	25	4

Carry out a chi-square test to determine whether these frequencies can be considered to conform to a Poisson distribution.

20 Contingency tables

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:

$$\frac{\text{row total} \times \text{column total}}{\text{table total}}$$

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.

20 Contingency tables

For each of three insurance companies, A, B, and C, a random sample of non-life policies of a particular kind is examined. It turns out that a claim (or claims) have arisen in the past year in 23% of the sampled policies for A, in 28% of those for B, and in 20% of those for C.

Test for differences in the underlying proportions of policies of this kind which have given rise to claims in the past year among the three companies in the two situations:

(a) the sample sizes were 100, 100, and 200 respectively

(b) the sample sizes were 300, 300, and 600 respectively.

Comment briefly on your results.