Lecture

INSTITUTE OF ACTUARIAL & QUANTITATIVE STUDIES

Class: FY BSc

Subject : Probability and Statistics -2

Subject Code: PUSASQ 1.2

Chapter: Unit 1 Chp 2

Chapter Name: Sampling Distributions

Index

1 Introduction

- When a sample is taken from a population the sample information can be used to infer certain things about the population. For example, to estimate a population quantity or test the validity of a statement made about the population.

- A population quantity could be its mean or variance, for example. So we might be testing the mean from a normal distribution, say.

- We will look at how to take a sample from a distribution and calculate its mean and variance. If we were to keep taking samples from the same distribution and calculating the mean and variance for each of the samples, we would find that these values also form probability distributions.

1 Introduction

- The statistical method for testing assertions such as ‘smoking reduces life expectancy’, involves selecting a sample of individuals from the population and, on the basis of the attributes of the sample, making statistical inferences about the corresponding attributes of the parent population.

- This is done by assuming that the variation in the attribute in the parent population can be modelled using a statistical distribution. The inference can then be carried out on the basis of the properties of this distribution.

- Theoretically this (technique) deals with samples from infinite populations. Actuaries are concerned with sampling from populations of policyholders, policies, claims, buildings, employees, etc. Such populations may be looked upon as conceptually infinite but even without doing so, they will be very large populations of many thousands and so the methods for infinite populations will be more than adequate.

2 Random Samples

A set of items selected from a parent population is a random sample if:

- the probability that any item in the population is included in the sample is proportional to its frequency in the parent population and
- the inclusion/exclusion of any item in the sample operates independently of the inclusion/exclusion of any other item.

- A random sample is made up of (iid) random variables and so they are denoted by capital X's. We will use the shorthand notation $\underline{X}$ to denote a random sample, that is, $\underline{X} = (X_1, X_2 \ldots, X_n)$. An observed sample will be denoted by $\underline{x} = (x_1, x_2, \ldots, x_n)$.

- The population distribution will be specified by a density (or probability function) denoted by $f(x;\theta)$, where θ denotes the parameter(s) of the distribution.

- Due to the Central Limit Theorem, inference concerning a population mean can be considered without specifying the form of the population, provided the sample size is large enough.

3 Statistic

A statistic is a function of $\underline{X}$ only and does not involve any unknown parameters. Thus $\bar{X} = \frac{\sum X_i}{n}$ **and** $S^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2$ are statistics whereas $\frac{1}{n}\sum(X_i - \mu)^2$ Is not, unless of course μ is known.

- Note here the difference between μ , which is the population mean (ie the mean for all possible observations, which is usually unknown) and $\bar{X}$, which is the sample mean (ie the mean of the sample values which we can calculate for any given sample).
- We might also be interested in statistics such as max X_i , the highest value in the sample.
- A statistic can be generally denoted by g($\underline{X}$) . Since a statistic is a function of random variables, it will be a random variable itself and will have a distribution, its sampling distribution.

4 The sample mean

- Suppose $\boldsymbol{X_i}$ has mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\sigma^2}$. Recall that the sample mean is $\bar{\boldsymbol{X}} = \frac{\sum X_i}{\boldsymbol{n}}$.
- Consider first $\boldsymbol{\sum X_i}$:
- $\boldsymbol{E}[\sum \boldsymbol{X_i}] = \sum \boldsymbol{E}[\boldsymbol{X_i}] = \sum \boldsymbol{\mu} = \boldsymbol{n\mu}$ since they are identically distributed
- $\boldsymbol{var}[\sum \boldsymbol{X_i}] = \sum \boldsymbol{var}[\boldsymbol{X_i}]$ since they are independent
  $= \boldsymbol{n\sigma^2}$ since they are identically distributed

- As $\bar{\boldsymbol{X}} = \frac{\boldsymbol{1}}{\boldsymbol{n}} \sum \boldsymbol{X_i}$, we can now write down that $\boldsymbol{E}[\bar{\boldsymbol{X}}] = \boldsymbol{\mu}$ and $\boldsymbol{var}\,[\bar{\boldsymbol{X}}] = \frac{\boldsymbol{1}}{\boldsymbol{n^2}} * \boldsymbol{n\sigma^2} = \frac{\boldsymbol{\sigma^2}}{\boldsymbol{n}}$

- Note: $\boldsymbol{sd}[\bar{\boldsymbol{X}}] = \frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}}$ is called the standard error of the sample mean.

- So we have established that the sample mean $\bar{\boldsymbol{X}}$ has an expected value of μ (ie the same as the population mean) and a variance of $\boldsymbol{\sigma^2/n}$ (ie the population variance divided by the sample size).

- A consequence of the result for the variance of $\bar{X}$ is that as the sample gets bigger the variance gets smaller. This should be intuitive since a bigger sample produces more accurate results.

5 The sample variance

- Recall that the sample variance is $\boldsymbol{S^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2}$.
- Considering only the mean of $\boldsymbol{S^2}$, it can be proved that $\boldsymbol{E[S^2] = \sigma^2}$ as follows:
- $\boldsymbol{S^2 = \frac{1}{n-1}[\sum X_i^2 - n\bar{X}^2]}$

- Taking expectations and noting that for any random variable Y, $\boldsymbol{E[Y^2] = var[Y] + (E[Y])^2}$ (obtained by rearranging $\boldsymbol{var(Y) = E(Y^2) - E^2(Y)}$ leads to:
- $\boldsymbol{E[S^2] = \frac{1}{n-1}(\sum E[X_i^2] - nE[\bar{X}^2])}$
  - $\boldsymbol{= \frac{1}{n-1}\left\{\sum(\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right)\right\}}$
  - $\boldsymbol{= \frac{1}{n-1}\{n(\sigma^2 + \mu^2) - \sigma^2 - n\mu^2\}}$
  - $\boldsymbol{= \frac{1}{n-1}\{(n-1)\sigma^2\} = \sigma^2}$
- as required.

- To work out $\boldsymbol{E[\bar{X}^2]}$, we've used the general formula just mentioned, which tells us that
- $\boldsymbol{E[\bar{X}^2] = var(\bar{X}) + E^2[\bar{X}]}$ and then we've used the results we just derived for the sample mean.

- So the n - 1 denominator is used to make the mean of S^2 equal to the true value of σ^2 . This is the motivation behind the definition of the sample variance.

Question 1

- The total number of new motor insurance claims reported to a particular branch of an insurance company on successive days during a randomly selected month can be considered to come from a Poisson distribution with $\lambda = 5$. Calculate the mean and variance of a sample mean based on 30 days' figures.

Solution

- The Poisson distribution in the question has mean and variance of 5.
- If the sample size is 30 then $E[\bar{X}] = 5$ and $var[\bar{X}] = \frac{5}{30} = 0.167$

6 Sampling distributions for the normal

The sample mean

- The Central Limit Theorem provides a large-sample approximate sampling distribution for $\bar{X}$ without the need for any distributional assumptions about the population. So for large n:

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1) \quad or \quad \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

- This result is often called the z result.

- It transpires that the above result gives the exact sampling distribution of $\bar{X}$ for random samples from a normal population.

6 Sampling distributions for the normal

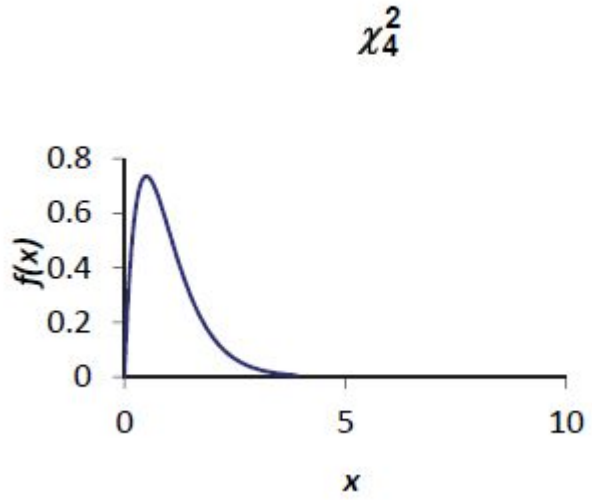
The sample variance

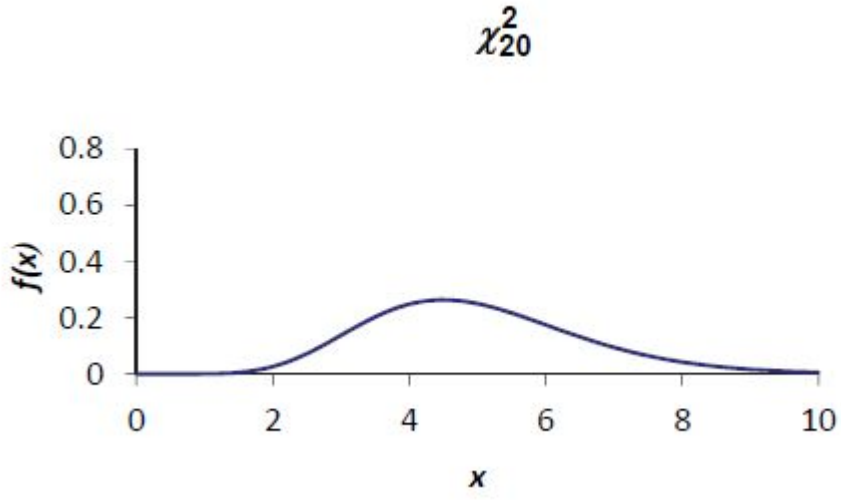
- The sampling distribution of S^2 when sampling from a normal population, with mean μ and variance σ^2 , is:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

6 Sampling distributions for the normal

- Whereas the distribution of $\bar{X}$ is normal and hence symmetrical, the distribution of S^2 is positively skewed especially so for small n but becoming symmetrical for large n.





6 Sampling distributions for the normal

- Using the χ^2 result to investigate the first and second order moments of S^2, when sampling from a normal population, and the fact that the mean and variance of χ_k^2 are k and 2k, respectively:

$$\boldsymbol{E\left[\frac{(n-1)S^2}{\sigma^2}\right] = n-1 \Rightarrow E\left[S^2\right] = \frac{\sigma^2}{n-1} * (n-1) = \sigma^2}$$

- We also have:

$$\boldsymbol{var\left[\frac{(n-1)S^2}{\sigma^2}\right] = 2(n-1) \Rightarrow var\left[S^2\right] = \frac{\sigma^4}{(n-1)^2} * 2(n-1) = \frac{2\sigma^4}{n-1}}$$

- For both $\bar{X}$ and S^2 the variances decrease and tend to zero as the sample size n increases.
- Added to the facts that $E[\bar{X}] = \mu$ and $E[S^2] = \sigma^2$, these imply that X gets closer to μ and S^2 gets closer to σ^2 as the sample size increases. These are desirable properties of estimators of μ and σ^2.

Question 2

- Calculate the probability that, for a random sample of 5 values taken from a $N(100,25^2)$ population

(i) $\bar{X}$ will be between 80 and 120
(ii) S will exceed 41.7.

Solution

(i)

- Since $\bar{X} \sim N(100, 25^2 \div 5) = N(100, 125)$:
- $P(80 < \bar{X} < 120) = P\left(\frac{80-100}{\sqrt{125}} < Z < \frac{120-100}{\sqrt{125}}\right)$
  $= P(-1.789 < Z < 1.789)$
  $= \Phi(1.789) - \Phi(-1.789)$
  =0.96319-(1-0.96319)=0.926

(ii)

- Since $\frac{4S^2}{\sigma^2} \sim \chi_4^2$, we have:

$$P(S > 4.17) = P\left(\frac{4S^2}{\sigma^2} > \frac{4\ X\ 41.7^2}{25^2}\right) = P(\chi_4^2 > 11.13) = 1 - P(\chi_4^2 < 11.13)$$

- Interpolating from the Normal Tables gives:

$$P(S > 41.7) \cong 0.0253$$

7 Independence of the sample mean and variance

- The other important feature when sampling from normal populations is the independence of $\bar{X}$ and S^2 . A full proof of this is not trivial but it is a result that is easily appreciated as follows.
- Suppose that a sample from some normal distribution has been simulated. The value of $\bar{x}$ does not give any information about the value of s^2.
- Remember that changing the mean of a normal distribution shifts the graph to the left or right. Changing the variance squashes the graph up or stretches it out.
- However, if the sample is from some exponential distribution, the value of $\bar{x}$ does give information about the value of $\boldsymbol{s^2}$, as $\boldsymbol{\mu}$ and $\boldsymbol{\sigma^2}$ are related.
- For the exponential distribution these are directly linked since $\boldsymbol{\mu} = \frac{1}{\lambda}$ and $\boldsymbol{\sigma^2} = \frac{1}{\lambda^2}$.
- Other cases such as Poisson, binomial and gamma can be considered in a similar way, but only the normal has the independence property.

8 The *t* result

- The sampling distribution for $\bar{X}$, that is, $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim \boldsymbol{N(0,1)}$ **or** $\bar{\boldsymbol{X}} \sim \boldsymbol{N\left(\mu, \frac{\sigma^2}{n}\right)}$,will be used subsequently for inference concerning $\boldsymbol{\mu}$ when the population variance $\boldsymbol{\sigma^2}$ is known.
- However this is rare in practice, and another result is needed for the realistic situation when $\boldsymbol{\sigma^2}$ is unknown. This is the t result or the t sampling distribution.
- The t result is similar to the z result but with σ replaced by S and N(0,1) replaced by $\boldsymbol{t_{n-1}}$.
- Thus $\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim \boldsymbol{t_{n-1}}$
- is not a sampling distribution for $\bar{X}$ alone as it involves a combination of $\bar{X}$ and S .

8 The *t* result

- The t_k variable is defined by:
- $t_k \equiv \frac{N(0,1)}{\sqrt{\frac{\chi_k^2}{k}}}$ where the ***N(0,1)*** and χ_k^2 random variables are independent.

- Then the ***t*** result above follows from the sampling distributions of the last section, that is, $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ is the ***N(0,1)*** and $\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$ is the χ_k^2, together with their independence, to obtain $\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$ when sampling from a normal population.

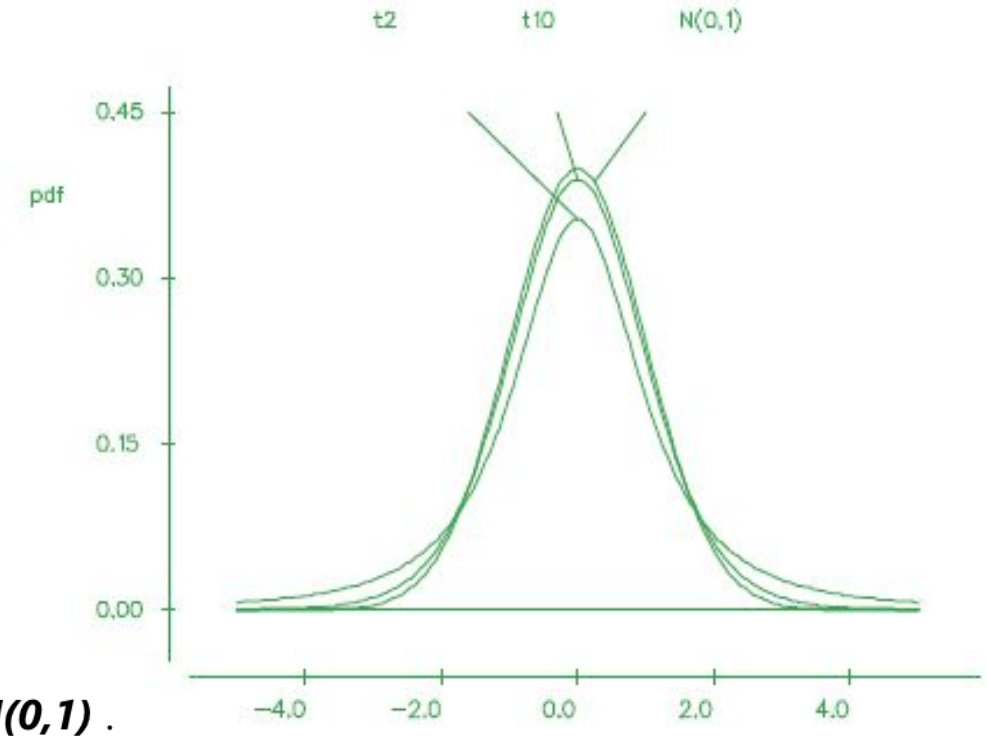
- The ***t*** distribution is symmetrical about zero and its critical points are tabulated.

- There are tables for the ***t*** distribution. It has one parameter, which, like the χ^2 distribution, is called the 'number of degrees of freedom'.

- When you are using the ***t*** distribution, you can work out the number of degrees of freedom by remembering that it is the same as the number you divided by when estimating the variance.

8 The *t* result

- The PDF of the t-distribution looks similar to the standard normal (ie symmetrical) especially for large values of degrees of freedom. The following picture shows a t_2 density, a t_{10} density and a ***N(0,1)*** density for comparison.



- In fact, as $\boldsymbol{k \to \infty, t_k \to}$ ***N(0,1)*** .
- The t_1 distribution is also called the Cauchy distribution and is peculiar in that none of its moments exist, not even its mean. However since samples of size 2 are unrealistic, it should not arise as a sampling distribution.
- For k > 2 , the t_k distribution has mean 0 and variance ***k / (k - 2)*** .

Question 3

- Independent random samples of size n_1 and n_2 are taken from the normal populations $N(\mu_1, \sigma_1^2)$ *and* $N(\mu_2, \sigma_2^2)$ respectively.

i. Write down the sampling distributions of $\bar{X}_1$ and $\bar{X}_2$ and hence determine the sampling distribution of $\bar{X}_1 - \bar{X}_2$, the difference between the sample means.

ii. Now assume that

(a) Express the sampling distribution of $\bar{X}_1 - \bar{X}_2$ in standard normal form.

(b) State the sampling distribution of $\frac{(\boldsymbol{n_1-1})\boldsymbol{S_1^2}+(\boldsymbol{n_2-1})\boldsymbol{S_2^2}}{\boldsymbol{\sigma^2}}$.

(c) Using the N(0,1) distribution from (a) and the χ^2 distribution from (b), apply the definition of the t distribution to find the sampling distribution of $\bar{X}_1 - \bar{X}_2$ when σ^2 is unknown.

Solution

(i)

- $\bar{X}_1$ is $N(\mu_1, \sigma_1^2/n_1)$ and $\bar{X}_2$ is $N(\mu_2, \sigma_2^2/n_2)$
- $\bar{X}_1 - \bar{X}_2$ is the difference between two independent normal variables and so is itself normal, with mean
- $\mu_1 - \mu_2$ and variance $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$

(ii)(a)

- The variance of $\bar{X}_1 - \bar{X}_2$ is now $\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$ and so standardising gives:

$$\frac{((\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$$

(ii)(b)

- As $\frac{(n_1-1)s_1^2}{\sigma^2} \sim \chi^2_{n_1-1}$ and $\frac{(n_2-1)s_2^2}{\sigma^2} \sim \chi^2_{n_2-1}$ are independent, (because the samples are independent), their sum is also χ^2 , with $n_1 + n_2 - 2$ degrees of freedom. This is using the additive property of independent χ^2 distributions (ie $\boldsymbol{\chi^2_m + \chi^2_n \sim \chi^2_{m+n}}$).

Solution

(ii)(c)

- Using the definition of the t distribution:

$$t_k \equiv \frac{N(0,1)}{\sqrt{\chi^2}_k/k}$$

- The distribution in part (ii)(a) was N(0,1) , and the distribution in part (ii)(b) was $\chi^2_{n_1+n_2-2}$

- So:

$$\frac{\dfrac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma^2\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}}{\sqrt{\dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{\sigma^2(n_1+n_2-2)}}} \sim t_{n_1+n_2-2}$$

- The σ^2's cancel to give:

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{(n_1+n_2-2)}\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} \sim t_{n_1+n_2-2}$$

8 The *F* result for variance ratios

- The F distribution is defined by $F = \frac{U/v_1}{V/v_2}$, where U and V are independent χ^2 random variables with v_1 and v_2 degrees of freedom respectively. Thus if independent random samples of size n_1 and n_2 respectively are taken from normal populations with variances

$$\boldsymbol{\sigma_1^2 \text{ and } \sigma_2^2, \text{ then } \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{n_1-1,n_2-1}}$$

- The F distribution gives us the distribution of the variance ratio for two normal populations. v_1 and v_2 can be referred to as the number of degrees of freedom in the numerator and denominator respectively.

8 The *F* result for variance ratios

- It should be noted that it is arbitrary which one is the numerator and which is the denominator and so
- $\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{n_1-1,n_2-1}$.

- Since it is arbitrary which value is the numerator and which is the denominator, and since only the upper critical points are tabulated, it is usually easier to put the larger value of the sample variance into the numerator and the smaller sample variance into the denominator.

- Alternatively, $F \sim F_{n_1-1,n_2-1} \Rightarrow \frac{1}{F} \sim F_{n_2-1,n_1-1}$

- This reciprocal form is needed when using tables of critical points, as only upper tail points are tabulated. See 'Formulae and Tables'.

Question 4

- For random samples of size 10 and 25 from two normal populations with equal variances, use the ***F*** distribution to determine the values of α and β such that $P\left(\frac{s_1^2}{s_2^2} > \alpha\right) = 0.05$ and $P\left(\frac{s_1^2}{s_2^2} < \beta\right) = 0.05$, where subscript 1 represents the sample of size 10 and subscript 2 represents the sample of size 25.

Solution

- Since the population variances are equal, $\frac{S_1^2}{S_2^2} \sim \boldsymbol{F_{9,24}}$ and $\frac{S_2^2}{S_1^2} \sim F_{24,9}$

- From the table of 5% points for the F distribution of the Tables, we find that
- $P(F_{9,24} > 2.300) = 0.05$, and therefore $\alpha = 2.300$.

- Now we know that $\frac{S_1^2}{S_2^2} < \beta$ is equivalent to $\frac{S_2^2}{S_1^2} > 1/\beta$ and $P(F_{24,9} > 2.900) = 0.05$, giving $\beta = \frac{1}{2.900} = 0.345$

Thank You