Lecture 1

IAQS | INSTITUTE OF ACTUARIAL & QUANTITATIVE STUDIES

Class: FY BSc

Subject : Probability and Statistics -2

Subject Code: PUSASQ 1.2

Chapter: Unit 1 Chapter 1

Chapter Name: Central Limit Theorem

Index

1 Introduction

- The Central Limit Theorem is perhaps the most important result in statistics. It provides the basis for large-sample inference about a population mean when the population distribution is unknown and more importantly does not need to be known.

- It also provides the basis for large-sample inference about a population proportion, for example, in initial mortality rates at given age x, or in opinion polls and surveys. It is one of the reasons for the importance of the normal distribution in statistics.

2 The Central Limit Theorem

If $X_1, X_2, \ldots X_n$ is a sequence of independent, identically distributed (iid) random variables with finite mean μ and finite (non-zero) variance σ^2 then the distribution of $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ approaches the standard normal distribution, ***N(0,1)*** , as $n \to \infty$.

- Note: It is not necessary to be able to prove this result. Remember that $\bar{X}$ is the sample mean, calculated as

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

3 Practical uses

- The way the Central Limit Theorem is used in practice is to provide useful normal approximations to the distributions of particular functions of a set of *iid* random variables.
- Therefore both $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ and $\frac{\sum X_i - n\mu}{\sqrt{n\sigma^2}}$ are approximately distributed as $N(0,1)$ for large n.
- Alternatively the unstandardized forms can be used. Thus $\bar{X}$ is approximately $N(\mu, \sigma^2/n)$ and $\sum X_i$ is approximately $N(n\mu, n\sigma^2)$.
- As a notation the symbol $\sim$ is used to mean 'is approximately distributed', so we can write the statements in the preceding paragraph as $\bar{X} \sim N(\mu, \sigma^2/n)$ and $\sum X_i \sim N(n\mu, n\sigma^2)$.

3 Practical uses

- An obvious question is: what is large n?
- A common answer is simply n≥30 but this is too simple an answer. A fuller answer is that it depends on the shape of the population, that is, the distribution of X_i, and in particular how skewed it is.
- If this population distribution is fairly symmetric even though non-normal, then n = 10 may be large enough; whereas if the distribution is very skewed, n = 50 or more may be necessary.

Examples

Question 1

It is assumed that the number of claims arriving at an insurance company per working day has a mean of 40 and a standard deviation of 12. A survey was conducted over 50 working days. Calculate the probability that the sample mean number of claims arriving per working day was less than 35.

Examples

Solution

Using the standard notation $\mu = 40, \sigma = 12, n = 50$.

The Central Limit Theorem states that $\bar{X} \sim N(40, 12^2/50)$.

We want $P(\bar{X} < 35)$. Standardising in the usual way:

$$P(\bar{X} < 35) \approx P\left(Z < \frac{35 - 40}{\sqrt{\frac{12^2}{50}}}\right)$$

$$= P(Z < -2.946) = 1 - P(Z, 2.946) = 1 - 0.99839 = 0.00161$$

4 Normal approximations

- We can use Central Limit Theorem to obtain approximations to the binomial, Poisson and gamma distributions. This is useful for calculating probabilities and obtaining confidence intervals and carrying out hypothesis tests on a piece of paper.

- However, it is easy for a computer to calculate exact probabilities, confidence intervals and hypothesis tests. Hence, these approximations are not as important as they used to be.

5 Binomial distribution *Bin(n,p)*

- Let X_i be iid Bernoulli random variables, that is, **Bin(1,p)** , so that

$$P(X_i = 1) = p$$
$$P(X_i = 0) = 1 - p$$

- In other words X_i is the number of successes in a single Bernoulli trial.

- Consider $X_1, X_2 \ldots, X_n$, a sequence of such variables. This is precisely the binomial situation and $X = \sum X_i$ is the number of successes in the **n** trials.

- So $X = \sum X_i$ ~ **Bin(n,p)** . Also note that $\frac{X}{n} = \overline{X}$. As a result of the Central Limit Theorem it can be said that, for large **n**:

$$\overline{X} \doteq N(\mu, \sigma^2/n) \quad \text{or} \quad \sum X_i \doteq N(n\mu, n\sigma^2)$$

- For the Bernoulli distribution:

$$\mu = E[X_i] = p \quad \text{and} \quad \sigma^2 = var[X_i] = p(1-p)$$

- Therefore $\sum X_i \doteq$ **N(np, np(1-p))** for large **n**, which is of course the normal approximation to the binomial.

- Basically, we approximate using a normal distribution, which has the same mean and variance as the binomial distribution.

5 Binomial distribution *Bin(n,p)*

Question

Given that ***X ~ Bin(n,p)*** , derive the mean and variance of $\bar{X}$, and hence write down the distribution of $\bar{X}$.

Solution

Since $\bar{X} = \frac{\sum X_i}{n}$, then:

$$E(\bar{X}) = \mathrm{E}\left(\frac{\sum X_i}{n}\right) = \frac{1}{n}E(\sum X_i) = \frac{1}{n}np = p$$

$$var(\bar{X}) = var\left(\frac{\sum X_i}{n}\right) = \frac{1}{n^2}var(\sum X_i) = \frac{1}{n^2}np(1-p) = \frac{p(1-p)}{n}$$

Therefore $\bar{X} \sim N\left(p, \frac{p(1-p)}{n}\right)$

5 Binomial distribution *Bin(n,p)*

- What is ‘large n’? A commonly quoted rule of thumb is that the approximation can be used only when both np and n(1- p) are greater than 5. The ‘only when’ is a bit severe. It is more a case of the approximation is less good if either is less than 5. However, this rule of thumb agrees with the answer that it depends on the symmetry/skewness of the population.

- Note that when p = 0.5 the Bernoulli distribution is symmetrical. In this case both np and n(1- p) equal 5 when n = 10 , and so the rule of thumb suggests that n = 10 is large enough.

- As p moves away from 0.5 towards either 0 or 1 the Bernoulli distribution becomes more severely skewed. For example, when p = 0.2 or 0.8 the rule of thumb gives n = 25 as large enough, but, when p = 0.05 or 0.95 the rule of thumb gives n = 100 as large enough.

6 Poisson distribution

- Let X_i , $i = 1,2,\ldots,n$ be iid $Poi(\lambda)$ random variables.
- So $\mu = E[X_i] = \lambda$ and $\sigma^2 = var[X_i] = \lambda$.
- The Central Limit Theorem implies that
$$\sum X_i \dot{\sim} N(n\lambda, n\lambda) \text{ for } \textbf{large } n.$$
- But $\sum X_i \sim Poi(n\lambda)$ and so, for **large** n, $Poi(n\lambda) \dot{\sim} N(n\lambda, n\lambda)$, or, equivalently, $Poi(\lambda) \dot{\sim} N(\lambda, \lambda)$ for **large** λ.
- Again, we are approximating using a normal distribution, which has the same mean and variance as the Poisson distribution.

6 Poisson distribution

Question

Show that $\sum X_i \sim Poi(n\lambda)$, where X_i is $Poi(\lambda)$ for all i.

Solution

Recall that the Poisson distribution is additive, ie:

$$X \sim Poi(\lambda) \text{ and } Y \sim Poi(\mu) => X + Y \sim Poi(\lambda + \mu)$$

Therefore $\sum X_i \sim Poi(n\lambda)$.

6 Poisson distribution

- A rule of thumb for this one is that the approximation is good if $\lambda > 5$. However since extensive tables for a range of values of λ are available, it is only needed in practice for much larger values of λ.

- Remember that the Poisson distribution is the limiting case of the binomial with $\lambda = np$ as $n \to \infty$ and $p \to 0$. So this is consistent with the rule for the binomial.

- The normal approximations to the binomial and Poisson distributions (both discrete) are the most commonly used in practice, and they are needed as the direct calculation of probabilities is computationally awkward without them.

7 Gamma distribution

- Let X_i , i = 1,2,...,n be a sequence of iid exponential (λ) variables and let Y be their sum.
- The exponential distribution has mean $\mu = 1/\lambda$ and variance $\sigma^2 = 1/\lambda^2$.
- Therefore for large n, $Y = \sum X_i \sim N(n/\lambda, n/\lambda^2)$.
- Therefore ***Y***, which is ***Gamma(n,λ)*** , will have a normal approximation for large values of ***n***.
- Recall that if $\boldsymbol{X_i \sim Exp(\lambda)}$ then $\boldsymbol{\sum X_i \sim Gamma(n, \lambda)}$.
- Since $\boldsymbol{\chi_k^2 \equiv Gamma(k/2, 1/2)}$, $\boldsymbol{\chi_k^2}$ will have a normal approximation ***N(k,2k)*** for large values of its degrees of freedom ***k***.
- These approximations are poorer than those used for the binomial and Poisson distributions due to the skewness of the Gamma distribution. It is therefore preferable to make use of the exact result.

8 The continuity correction

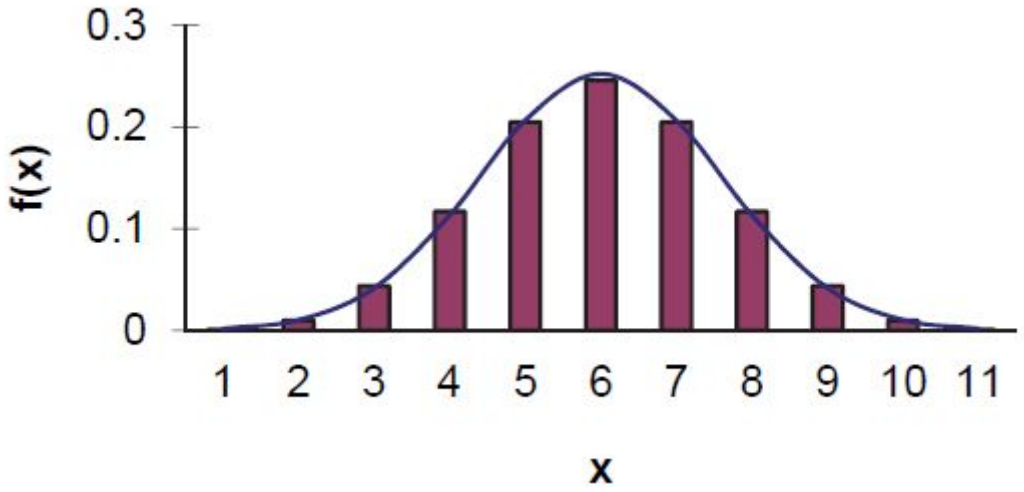
- When dealing with the normal approximations to the binomial and Poisson distributions, which are both discrete, a discrete distribution is being approximated by a continuous one.
- When using such an approximation the change from discrete to continuous must be allowed for.
- For an integer-valued discrete distribution, such as the binomial or Poisson, it is perfectly reasonable to consider individual probabilities such as P(X = 4) . However if X is continuous, such as the normal, P(X = 4) is not meaningful and is taken to be zero.
- For a continuous variable it is sensible to consider only the probability that X lies in some interval.
- For a continuous distribution it is not useful to think about the probability of a random variable being exactly equal to a value: for example, for a continuous distribution:

$$\boldsymbol{P(X = 4) = P(4 \leq X \leq 4) = \int_4^4 f(x)dx = 0}$$

- To allow for this a continuity correction must be used. Essentially it corresponds to treating the integer values as being rounded to the nearest integer.

8 The continuity correction

- The diagram below illustrates the problem. The bars correspond to the probabilities for a Bin(10,0.5) distribution, whereas the graph corresponds to the probability density function for the normal approximation.

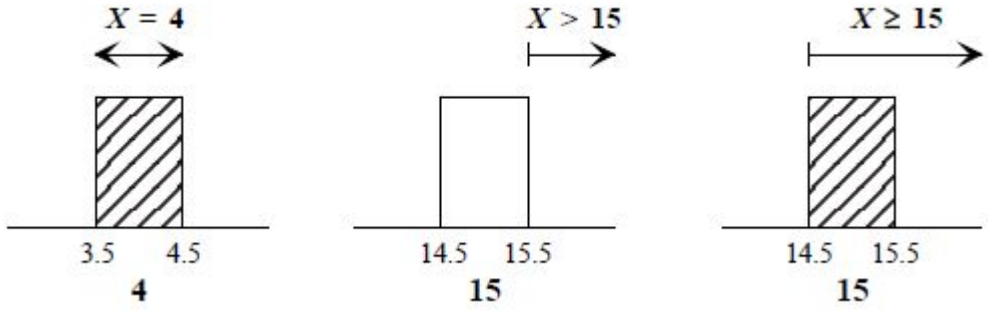


- Since the binomial is a discrete distribution there are no probabilities for non-integer values, whereas the normal approximation can take any value. To compensate for the 'gaps' between the bars, we suppose that they are actually rounded to the nearest integer. For example, the x = 6 bar is assumed to represent values between x = 5.5 and x = 6.5 .

8 The continuity correction

- So to use the continuity correction in practice, for example,
  $X = 4$ is equivalent to '$3.5 < X < 4.5$'
  $X > 15$ is equivalent to '$X > 15.5$'
  $X \geq 15$ is equivalent to '$X > 14.5$'

- Alternatively, considering the bars on the graph:



- X= 4 must, obviously, include all of the X = 4 bar which goes from 3.5 to 4.5.
- X >15 must not include the X =15 bar (as it is a strict inequality), therefore it should start from 15.5 (the upper end of the 15 bar).
- X ≥15 includes the X = 15 bar and higher, therefore it should start from 14.5 (the lower end of the 15 bar).

Examples

Question 2

Let X be a Poisson variable with parameter 20. Use the normal approximation to obtain a value for $P(X \leq 15)$ and use tables to compare with the exact value.

Examples

Solution

We have:

X~Poi(20)⇒ $X \quad N(20,20) \Rightarrow \frac{X-20}{\sqrt{20}} \quad N(0,1)$

$P(X \leq 15) \equiv P(X < 15.5)$:using continuity correction

$$\approx P\left(Z < \frac{15.5 - 20}{\sqrt{20}}\right) = P(Z < -1.006)$$

$= 1 - 0.84279$, interpolating in tables to be as accurate as possible

=0.15721

From Poisson tables, $P(X \leq 15) = 0.15651$

Error=0.0007 or a 0.45% relative error.

Examples

Question 3

Use a normal approximation to calculate an approximate value for the probability that an observation from a Gamma(25,50) random variable falls between 0.4 and 0.8.

Examples

Solution

The mean and variance of a general gamma distribution are $\frac{\alpha}{\lambda}$ and $\frac{\alpha}{\lambda^2}$ so here the mean and variance are 0.5 and 0.01 respectively. If X is the gamma random variable, then we will use X ~ N(0.5,0.01) :

$$
\begin{aligned}
P(0.4 < X < 0.8) &\approx P(-1 < Z < 3) \\
&= \Phi(3) - \Phi(-1) \\
&= \Phi(3) - [1 - \Phi(1)] \\
&= 0.99865 - 0.15866 = 0.840
\end{aligned}
$$

No continuity correction is required, as we started with a continuous distribution.

The exact answer is 0.8387.

Summary

- According to Central Limit Theorem (CLT) ,if $X_1 \dots X_n$ are independent and identically distributed random variables with mean μ and variance σ^2 , then:
- $\sum X \doteq N(n\mu, n\sigma^2) \Rightarrow \frac{\sum X_i - n\mu}{\sqrt{n\sigma^2}} \quad N(0,1) \quad as\ n \to \infty$
- $\bar{X} \doteq N\left(\mu, \frac{\sigma^2}{n}\right) \Rightarrow \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \quad N(0,1) \quad as\ n \to \infty$

- Normal approximations:
- Bin(n,p) $\doteq N(np, npq)$ np>5,nq>5 with continuity correction
- $Poi(\lambda) \doteq N(\lambda, \lambda)$ λ large with continuity correction
- Gamma$(\alpha, \lambda) \doteq N\left(\frac{\alpha}{\lambda}, \frac{\alpha}{\lambda^2}\right)$ α large
- $\chi_k^2 \doteq N(k, 2k)$ k large

Thank You