REVISION SHEET – STATISTICS 2 (OCR)

CONTINUOUS RANDOM VARIABLES

The main ideas are:

- Properties of Continuous Random Variables
- Mean, Median and Mode
- Normal approximations to other distributions

Before the exam you should know:

- The properties of continuous random variables, including the p.d.f. function.
- How to calculate the mean, variance, median and mode.
- And be able to use the cumulative distribution function.
- How to approximate to the normal distribution from other distributions.

Continuous Random Variables

A continuous random variable is a random variable that can take any value within a range, i.e. height or weight. It is described by a *probability density function* (p.d.f.). A probability density function may be found from the results of an experiment, or it may be given as an algebraic expression. For a continuous random variable, the total area under the curve of the probability density function must be 1.

The expectation $\mathrm{E}(X) = \mu = \int x\mathrm{f}(x)dx$ and

$$\mathrm{Var}(X) = \int (x - \mu)^2 \mathrm{f}(x)\, dx \qquad \textit{where } \mu = E(X)$$

$$= \int (x^2\mathrm{f}(x) - 2\mu x\mathrm{f}(x) + \mu^2\mathrm{f}(x))\, dx$$

$$= \int x^2\mathrm{f}(x)\, dx - 2\mu \int x\mathrm{f}(x)\, dx + \mu^2 \int \mathrm{f}(x)\, dx$$

$$= \int x^2\mathrm{f}(x)\, dx - 2\mu^2 + \mu^2 \qquad \textit{since } \int x\mathrm{f}(x)dx = E(X) = \mu$$

$$\textit{and } \int \mathrm{f}(x)dx = 1$$

$$= \int x^2\mathrm{f}(x)\, dx - \mu^2$$

$$\mathrm{Var}(X) = \int x^2\mathrm{f}(x)\, dx - [\mathrm{E}(X)]^2$$

Example

A continuous random variable X has p.d.f f(x), where:

$$\mathrm{f}(x) = \begin{cases} \frac{1}{3}(x-1) & \text{for } 1 \le x \le 3 \\ 0 & \text{otherwise} \end{cases}$$

Find the expectation and variance of X.

Solution

$$\mathrm{E}(X) = \int_1^3 \tfrac{1}{3}x(x-1)dx = \tfrac{1}{3}\int_1^3 (x^2 - x)dx = \tfrac{1}{3}\left[\tfrac{1}{3}x^3 - \tfrac{1}{2}x^2\right]_1^3 = \frac{14}{9}$$

$$\mathrm{Var}(X) = \int_1^3 \tfrac{1}{3}x^2(x-1)dx - [\mathrm{E}(X)]^2 \quad = \tfrac{1}{3}\int_1^3 (x^3 - x^2)dx - \left[\tfrac{14}{9}\right]^2$$

$$= \tfrac{1}{3}\left[\tfrac{1}{4}x^4 - \tfrac{1}{3}x^3\right]_1^3 - \left[\tfrac{14}{9}\right]^2$$

$$= \frac{100}{81}$$

Disclaimer: Every effort has gone into ensuring the accuracy of this document. However, the FM Network can accept no responsibility for its content matching each specification exactly.

Median

The value m for which P($X < m$) = 0.5.
The median can be found by using

$$\int_a^m f(x)dx = 0.5 \qquad \text{where } a \text{ is the lower limit of f}(x)$$

or $$\int_m^b f(x)dx = 0.5 \qquad \text{where } b \text{ is the upper limit of f}(x)$$

or F(x) – F(a) = 0.5 where F(x) is the cumulative distribution function and a is the lower limit of f(x).

Mode

The mode of a continuous random variable is the value of x for which f(x) has its highest value. If the mode is at a stationary point, it can be found by differentiation; otherwise it can usually be found by inspection of the graph of f(x).

Rectangular distribution

A distribution with for which f(x) is a constant within a particular range and zero elsewhere.

Its p.d.f. is given by:

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{for } a \le x \le b \\ 0 & \text{otherwise.} \end{cases}$$

The expectation is $\dfrac{a+b}{2}$ and the variance is $\dfrac{1}{12}(b-a)^2$.

Cumulative distribution function (c.d.f.)

The cumulative distribution function F(x) = P($X \le x$).
It can be found from the p.d.f. f(x) as follows:

$$\begin{aligned} F(x) &= 0 && x < a \\ &= \int_a^x f(u)du && a \le x \le b \\ &= 1 && x > b \end{aligned}$$

where a is the lower limit of f(x), and b is the upper limit of f(x).

Normal Distribution as an approximation to the Binomial Distribution

Suppose $X \sim \text{Binomial}(n, p)$.
If n is large and p is not too close to 0 or 1 (i.e. the distribution is reasonably symmetrical), then using the mean (np) and variance (npq) of a binomial distribution we can approximate using the normal distribution.

$$X \sim N(np, npq)$$

Normal Distribution as an approximation to the Poisson Distribution

Suppose $X \sim Poisson(\lambda)$
If λ is large, then the Poisson distribution is reasonably symmetrical.
Then using the mean (λ) and variance (λ) of a Poisson distribution we can approximate using the normal distribution.

$$X \sim N(\lambda, \lambda)$$

Important: In both cases above we are using a continuous distribution to approximate a discrete one and as such we must use continuity correcting when calculating a probability. Make sure you understand how to do this.

REVISION SHEET – STATISTICS 2 (OCR)

HYPOTHESIS TESTING USING THE BINOMIAL DISTRIBUTION

The main ideas are:

- Establishing the null and alternative hypotheses
- Conducting the test, doing the necessary calculations
- Interpreting the results

Before the exam you should know:

- The vocabulary associated with hypothesis testing.
- How to write the null and alternative hypotheses.
- How to decide whether the hypothesis test is one or two tailed.
- How to compare a value to the significance level.
- How to find critical values/regions.
- How to decided whether to reject H_0 or not and how to write a conclusion based on the situation.
- How to carry out a 2-tail test.

Vocabulary

You should be familiar with the following terms/notation for binomial hypothesis tests

Probability of success: p **Number of trials:** n **Number of successes:** X

Null Hypothesis (H_0): The statement that the probability of success is equal to a certain value.

Alternative Hypothesis (H_1): The statement that the probability of success is actually <, > or ≠ to the value in given in H_0.

Significance level: The probability at which you make the decision that an observed outcome hasn't happened by chance, given the probability of success in H_0.

1-tail test: A test based on the probability in H_0 being either too high or too low (but not both).

2-tail test: A test based on the probability in H_0 being incorrect (too high or too low).

Critical value: The maximum (for <) or minimum (for >) value, X, for the number of successes that would result in rejecting H_0.

Critical region: The set of values of X for the number of successes that would result in rejecting H_0.

Acceptance region: The set of values of X for the number of successes that would result in accepting H_0.

Errors P(type I error) = P(reject H_0 | H_0 true), P(type II error) = P(accept H_0 | H_0 false).

Hypothesis Tests

Hypothesis testing is based on assuming that the probability of success, p, takes a certain value, then conducting an experiment to test it. Given this assumption, if the result of the experiment is sufficiently rare (i.e. unlikely to have happened by chance) you can conclude that the probability, p, is likely to be incorrect.

Setting up

The statement of the value of this assumed probability, p, is known as the Null Hypothesis (H_0) (this is what you are testing). You must then decide if the situation leads you to think this value is too high, too low or, in the case of a 2-tailed test, incorrect.

Conducting

The probability of obtaining the value recorded in the experiment, x, or something more extreme is compared to the significance level to see if it is sufficiently rare to reject the null hypothesis. You must use $P(X \le x)$ or $P(X \ge x)$ as opposed to $P(X = x)$.

Drawing conclusions

If the probability is smaller than the significance level then reject H_0 in favour of H_1, otherwise you accept H_0 at the stated significance level.

Disclaimer: Every effort has gone into ensuring the accuracy of this document. However, the FM Network can accept no responsibility for its content matching each specification exactly.

Example

The makers of the drink Fizzicola claim that three-quarters of people prefer their drink to any other brand of cola. A rival company suspects that the claim by Fizzicola is exaggerated. They wish to carry out a hypothesis test to test this claim.

(i) State suitable Null and Alternative Hypotheses.

The rival company take a sample of 15 cola drinkers of whom 9 say they prefer Fizzicola to any other brand.

(ii) Using these data, carry out a hypothesis test at the 5% level stating your conclusion carefully.

Solution

(i) H_0: $p = 0.75$; The probability of a person chosen at random preferring Fizzicola is 0.75.

H_1: $p < 0.75$; The claim is exaggerated, the probability of a person chosen at random preferring Fizzicola is less than 0.75.

The alternative hypothesis is based on the rival branding thinking the claim is exaggerated, i.e. the proportion stated is too high.

(ii) From the tables: $P(X \leq 9) = 0.1484$. This value is not significant at the 5% level, therefore we accept H_0. There is not sufficient evidence to suggest Fizzicola are overestimating the proportion.

The probability of 9 or fewer is used, as opposed to exactly 9, as if you would accept 9 as evidence of overestimating then you would have also accepted 8, 7, 6, …

The significance level tells you the value at which a probability is considered so rare that is unlikely to have happened by chance. In this example case 5% is used, so an event with probability smaller than 0.05 is considered rare: 0.1484 is not smaller than 0.05 so the event is not rare.

As the event is not rare, it is likely that it occurred by chance, so there is no evidence to suggest that the makers of Fizzicola were overestimating. Note that you are not saying that they are correct, just that you don't have strong enough evidence to contradict them.

Alternative solution using critical value/critical region

(ii) From the tables: $P(X \leq 7) = 0.0173$, $P(X \leq 8) = 0.0566$. The critical value is 7, (the critical region is 0-7). 9 is not in the critical region (it is in the acceptance region), therefore we do not reject H_0. There is not sufficient evidence to suggest Fizzicola are overestimating the proportion.

The critical value is the largest (because H_1 is <) value of x such that P*(X < x) is smaller than the significance level.*

This example used an alternative hypothesis of the form H_1: $p < 0.75$ (because the rival firm thought the company was overestimating). This made it easy to read the values for $P(X \leq 7)$, $P(X \leq 8)$ and $P(X \leq 9)$ from the tables. If the alternative hypothesis had been of the form H_1: $p > 0.75$ (e.g. if the firm thought 0.75 was an underestimate), you would need to work with ≥ probabilities, using the complement of the values in the table.

e.g. If the alternative hypothesis had been H_1: $p > 0.75$ you would have calculated $P(X \geq 9)$.
$P(X \geq 9) = 1 - P(X \leq 8) = 1 - 0.0566 = 0.9434$: this is not smaller than 0.05 so you do not reject H_0.

1-tail vs 2-tail tests.

If there is no indication in the situation as to whether the probability used in H_0 is too high or too low you use a 2-tailed test, splitting the significance level in half and using half at each end.

Example

A teacher is forming a 12-person committee of students. She does not want the selection system to unfairly favour either boys or girls. Construct a hypothesis test at the 5% level to test this.

Solution

H_0: $p = 0.5$, There is an equal chance of a boy or girls being chosen.
H_1: $p \neq 0.5$, The selection system favours one gender.
You then split the significance level in half forming two critical regions of 2.5% at the top and bottom, totalling 5%.
Critical regions: 0 – 2 and 10 – 12.

Disclaimer: Every effort has gone into ensuring the accuracy of this document. However, the FM Network can accept no responsibility for its content matching each specification exactly.

REVISION SHEET – STATISTICS 2 (OCR)

NORMAL DISTRIBUTION

The main ideas are:

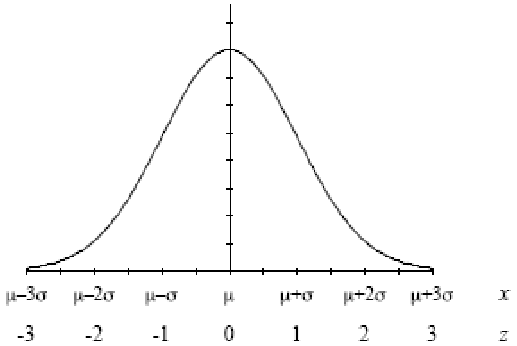
- Properties of the Normal Distribution
- Mean, SD and Var
- Approximating from other distributions

Before the exam you should know:

- All of the properties of the Normal Distribution.
- How to use the relevant tables.
- How to calculate mean, standard deviation and variance.
- How to approximate to the normal distribution from others.

Definition

A continuous random variable X which is bellshaped and has mean (expectation) μ and standard deviation σ is said to follow a **Normal Distribution** with **parameters** μ and σ.
In shorthand, $\boldsymbol{X \sim \mathrm{N}(\mu, \sigma^2)}$



This may be given in *standardised* form by using the transformation

$$z = \frac{x-\mu}{\sigma} \Rightarrow x = \sigma z + \mu \text{, where } Z \sim \mathbf{N}(0, 1)$$

Calculating Probabilities

The area to the left of the value z, representing $P(Z \le z)$, is denoted by $\Phi(z)$ and is read from tables for $z \ge 0$.

Useful techniques for $z \ge 0$:

- $P(Z > z) = 1 - P(Z \le z)$
- $P(Z > -z) = P(Z \le z)$
- $P(Z < -z) = 1 - P(Z \le z)$

The *inverse normal tables* may be used to find $z = \Phi^{-1}(p)$ for $p \ge 0.5$. For $p < 0.5$, use symmetry properties of the Normal distribution.

99.73% of values lie within 3 s.d. of the mean

Estimating μ and/or σ

Use (simultaneous) equations of the form: $x = \sigma z + \mu$ for matching (x, z) pairs – where z is given or may be deduced from $\Phi^{-1}(p)$ for given value(s) of x.

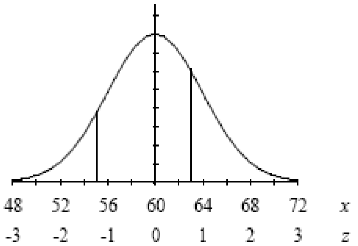
Example 1

$X \sim \mathrm{N}(60, 16) \Rightarrow z = \dfrac{x-60}{4}$;

find (a) $P(X < 66)$, (b) $P(X \ge 66)$, (c) $P(55 \le X \le 63)$,
(d) x_0 s.t. $P(X > x_0) = 99\%$

(a) $P(X < 66) = P(Z < 1.5) = \mathbf{0.9332}$

(b) $P(X \ge 66) = 1 - P(X < 66) = 1 - 0.9332 = \mathbf{0.0668}$



(c) $P(55 \le X \le 63) = P(-1.25 \le Z \le 0.75)$
$= P(Z \le 0.75) - P(Z < -1.25)$
$= P(Z \le 0.75) - P(Z > 1.25)$
$= P(Z \le 0.75) - [1 - P(Z \le 1.25)]$
$= 0.7734 - [1 - 0.8944] = \mathbf{0.6678}$

(d) $P(Z > -2.326) = 0.99$ from tables

Since $z = \dfrac{x-60}{4}$, $x = 4z + 60$

$\Rightarrow x_0 = 60 + 4\times(-2.326) = \mathbf{50.7}$ (to 3 s.f.)

Example 2

For a certain type of apple, 20% have a mass greater than 130g and 30% have a mass less than 110g.

(a) Estimate μ and σ.

(b) When 5 apples are chosen at random, find the probability that all five have a mass exceeding 115g

(a) $P(Z > 0.8416) = 0.2$ $(X = 130)$
$P(Z < -0.5244) = 0.3$ $(X = 110)$
$\Rightarrow 130 = 0.8416\,\sigma + \mu$
$110 = -0.5244\,\sigma + \mu$
Solving equations simultaneously gives: $\boldsymbol{\mu} = \mathbf{117.68}$, $\boldsymbol{\sigma} = \mathbf{14.64}$

(b) $X \sim \mathrm{N}(117.68, 14.64^2) \Rightarrow z = \dfrac{x-117.68}{14.64}$;
$P(X > 115)^5 = P(Z > -0.183)^5 = 0.5726^5 = \mathbf{0.0616}$ (to 3 s.f.)

Acknowledgement: Material on this page was originally created by Bob Francis and we acknowledge his permission to reproduce it here.

Disclaimer: Every effort has gone into ensuring the accuracy of this document. However, the FM Network can accept no responsibility for its content matching each specification exactly.

Further examples

Example 1

Suppose that $X \sim N(12,4)$. Calculate $P(X<13)$.

Solution

$$P(X<13)=\Phi\left(\frac{13-12}{2}\right)=\Phi(0.5)=$$

This is the number of standard deviations between 13 and the mean, 12.

Example 2

Suppose that the time taken for a journey to work is normally distributed with a mean of 20 and a standard deviation of 3. Calculate the probability that a journey to work takes between 17 and 21 minutes.

Solution

Let X = time taken for the journey to work. Then $X \sim N(20,9)$.

$$P(17<X<21)=\text{Area A} - \text{Area B}$$

This is the number of standard deviations between 21 and the mean, 20.

$$=\Phi\left(\frac{21-20}{3}\right)-\left(1-\Phi\left(\frac{20-17}{3}\right)\right)$$

This is the number of standard deviations between 17 and the mean, 20.

$$=\Phi\left(\frac{1}{3}\right)-\left(1-\Phi(1)\right)$$

$$=0.6301-(1-0.8413)$$

$$=0.4714$$

Example 3

Suppose that $X \sim N(\mu,16)$. If $P(X>10)=0.3$, find μ.

Solution

Since $P(X>10)=0.3$ μ must be less than 10.

So, $0.3=P(X>10)=1-\Phi\left(\frac{10-\mu}{4}\right)$ and so $\Phi\left(\frac{10-\mu}{4}\right)=0.7$.

This gives that:

$$\frac{10-\mu}{4}=\Phi^{-1}(0.7)=0.5244$$

$$\Rightarrow \mu=10-(4\times 0.5244)=7.9024$$

Approximating the Binomial Distribution

Suppose $X \sim \text{Binomial}(n,p)$.

If n is large and p is not too close to 0 or 1 (i.e. the distribution is reasonably symmetrical), then using the mean (*np*) and variance (*npq*) of a binomial distribution we can approximate using the normal distribution.

$$X \sim N(np,npq)$$

Approximating the Poisson Distribution

Suppose $X \sim Poisson(\lambda)$

If λ is large, then the Poisson distribution is reasonably symmetrical.
Then using the mean (λ) and variance (λ) of a Poisson distribution we can approximate using the normal distribution.

$$X \sim N(\lambda,\lambda)$$

Important: In both cases above we are using a continuous distribution to approximate a discrete one and as such we must use continuity correcting when calculating a probability. Make sure you understand how to do this.

REVISION SHEET – STATISTICS 2 (OCR)

POISSON DISTRIBUTION

The main ideas are:

- Calculations using the Poisson Distribution
- Modelling the Binomial distribution with the Poisson distribution

Before the exam you should know:

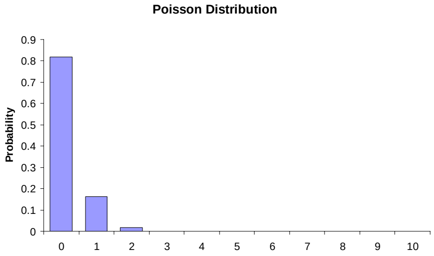
- When the Poisson distribution is an appropriate model for a given situation.
- The relationship $e^y = x \Leftrightarrow y = \ln x$, this is sometimes useful in questions.
- How to use the formula $P(X = r) = e^{-\lambda}\dfrac{\lambda^r}{r!}$ (without getting confused between λ and *r*).
- How to look up $P(X \le r)$ in the tables given.

Poisson Distribution

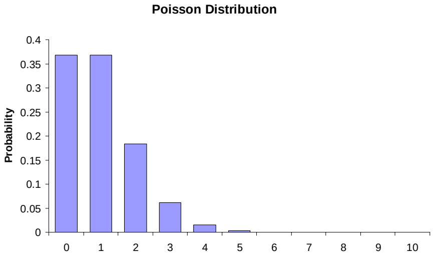
This models events which are random, independent, which occur singly and with a uniform likelihood.

If $X \sim$ Poisson(λ) then: $P(X = r) = \dfrac{e^{-\lambda}\lambda^r}{r!}$**,** where E(*X*) = $\mu = \lambda$ and Var(*X*) = $\sigma^2 = \lambda$.

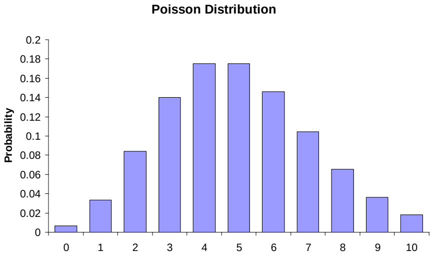
The Poisson Distribution for various values of λ is shown below.



λ = 0.2



λ = 1



λ = 5

Calculations using the Poisson Distribution

You should be able to use the formula $P(X = r) = e^{-\lambda}\dfrac{\lambda^r}{r!}$ and the cumulative Poisson tables (which give $P(X \le r)$ for various values of λ) to find simple probabilities.

Example 1
The number of goals, *X*, scored by a team playing at home in the Premier League is modelled by a Poisson distribution with a mean of 1.6. What is the probability that the team scores

a) 3 goals in a game
b) More than 4 goals in a game

Solution

a) The probability of the team scoring 3 goals in a game is: $P(X = 3) = e^{-1.6}\dfrac{1.6^3}{3!} = 0.138$ (to 3 d.p.)

b) The probability of the team scoring more than 4 goals in a game is:

$$P(X > 4) = 1 - P(X \le 4) = 1 - 0.9763 = 0.0237$$

Disclaimer: Every effort has gone into ensuring the accuracy of this document. However, the FM Network can accept no responsibility for its content matching each specification exactly.

More Complicated Questions

In other questions you will need to use the following properties of the Poisson Distribution:

If $X \sim$ Poisson(λ_1) and $Y \sim$ Poisson(λ_2) then: $nX \sim$ Poisson($n\lambda_1$) and $X+Y \sim$ Poisson($\lambda_1+\lambda_2$)

Example 2
The mean number of burgers sold per minute at a snack bar is 0.7. The mean number of hotdogs sold per minute is 0.5. Using a Poisson distribution model calculate the probability that the burger bar sells:

a) *5 burgers in a 5 minute period.*
b) *No hot dogs or burgers in a 1 minute period.*

Solution

a) The mean number of burgers sold in one minute is 0.7. Therefore the mean number of burgers sold in five minutes is $0.7 \times 5 = 3.5$. So, $5X$ is the number of burgers sold in 5 minutes and we have that $5X \sim$ Poisson(3.5). Therefore,

$$P(5X = 5) = e^{-3.5}\frac{3.5^5}{5!} = 0.132 \text{ (to 3 d.p.)}$$

b) If X is the number of burgers sold in a minute then $X \sim$ Poisson(0.7). If Y is the number of burgers sold in a minute then $Y \sim$ Poisson(0.5). So the total number of hotdogs and burgers sold in a minute is $X + Y$ and $X+Y \sim$ Poisson(0.7 + 0.5 = 1.2). Therefore,

$$P(X+Y=0) = e^{-1.2}\frac{1.2^0}{0!} = 0.301$$

Approximating the Binomial Distribution with the Poisson Distribution

If $X \sim \text{Binomial}(n, p)$ a Poisson approximation of $X \sim$ Poisson(np) can be used when

- n is large
- p is small (i.e. it is a rare event)

but it is only useful if np is not too large.

For example if n = 1000, p = 0.002, then np = 2. Under the binomial distribution $X \sim \text{Binomial}(1000, 0.02)$

$$P(X=10) = {}^{1000}C_{10} \times 0.002^{10} \times 0.998^{990} = 0.000037 \text{ to (6 d.p)}$$

With the Poisson Distribution $X \sim$ Poisson(2)

$$P(X=10) = e^{-2}\frac{2^{10}}{10!} = 0.000038 \text{ (to 6 d.p)}$$

The difference between these two values is only 0.000001

Disclaimer: Every effort has gone into ensuring the accuracy of this document. However, the FM Network can accept no responsibility for its content matching each specification exactly.