Statistics for Computing

MA4413 Lecture 4A and 4B

Kevin O'Brien

Kevin.obrien@ul.ie

Dept. of Mathematics & Statistics,
University *of* Limerick

Autumn Semester 2013

Week 4 (from Lecture 4A)

- The first midterm is to take place Monday of Week 5 at 4pm.
- The first midterm will cover:
  - Basic Probability
  - Descriptive statistics (mean, median variance etc)
  - Discrete probability distributions (binomial and Poisson)
  - The exponential distribution
  - Some of the normal distribution will be included.

Overview of Current Part of Course

Probability Distributions (Question 2 for End Of Year Exam)

- Discrete Probability Distributions
  - Binomial Probability Distribution (Week 3)
  - Geometric Probability Distribution (Week 3)
  - Poisson Probability Distribution (Week 3/4)
- Continuous Probability Distributions
  - Exponential Probability Distribution (Week 4)
  - Uniform Probability Distribution (Week 4)
  - Normal Probability Distribution (Week 4/5)

Current Status (Lecture 4B)

- Mid Term Examination next Monday (Week 5) at 4pm
- Currently covering : Continuous Probability Distributions
- Lecture notes are a bit out of synch with published class notes.
- The Exponential distribution will be examinable in Mid-Term 1
- Next Wednesday, we will start looking at the Normal Distribution.

Binomial Expected Value and Variance

If the random variable X has a binomial distribution with parameters n and p, we write

$$X \sim B(n,p)$$

Expectation and Variance If $X \sim B(n,p)$, then:

- Expected Value of X : E(X) = np
- Variance of X : Var(X) = np(1-p)

Binomial Distribution: Example 1

- Diagrams of the probability mass functions of the two binomial distributions $B(10, 0.5)$ and $B(10, 0.25)$ are shown in the bar-plots (next slide).
- Which is which? Give a reason for your answer.

Binomial Distribution

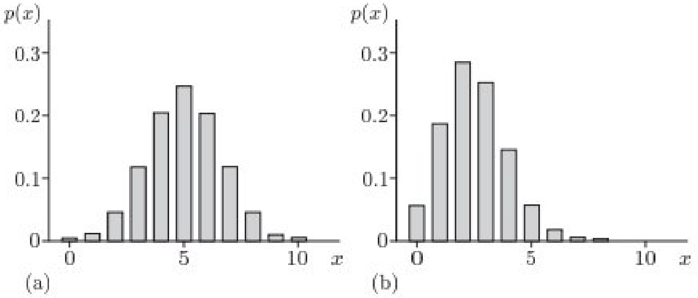


Figure: Bar Charts

Binomial Distribution: Example 1

- Clearly. Figure A is $B(10, 0.5)$ and Figure B is $B(10, 0.25)$.
- The mean of B(10, 0.5) is 5, and the mean of B(10, 0.25) is 2.5. These values correspond to the apex of both distributions on the previous slide.
- Also the variance of a binomial distribution corresponding to $B(10, 0.25)$ is 1.875, while for $B(10, 0.25)$ it is 2.500.
- A visual inspection of the two bar-charts would indicate that Figure A has the higher variance.

Binomial Distribution: Example 2

Example

- Components are placed into containers containing 100 items.
- After an inspection of a large number of containers the average number of defective items was found to be 10 with a standard deviation of three.
- Is the binomial distribution a good useful distribution, given the observed data?

Binomial Distribution: Example 2

- Let the number of containers be the number of independent trials is $n = 100$.
- A success may be defined as a defective component.
- The probability of a success is approximate $p = 0.10$. (The probability of "failure" is $1 - p = 0.9$).
- The expected number of defective components is $np = 10$, which concurs with our observed data.
- The variance is computed as

$$np(1-p) = 100 \times 0.1 \times 0.9 = 9$$

- The observed standard deviation is 3 units, i.e. a variance of 9 square units.
- Yes the binomial distribution is useful in this case.

Poisson Expected Value and Variance

If the random variable X has a Poisson distribution with parameter m, we write

$$X \sim Poisson(m)$$

- Expected Value of X : E(X) = m
- Variance of X : $\text{Var}(X) = m$
- Standard Deviation of X : $SD(X) = \sqrt{m}$

Poisson Distribution : Example

- The number of faults in a fibre optic cable were recorded for each kilometre length of cable.
- The mean number of faults was found to be 4 faults per kilometre.
- The standard deviation of the number of faults was found to be 2 faults per kilometre.
- Is the Poisson Distribution is a useful technique for modelling the number of faults in fibre optic cable?
- (Looking at the last slide, the answer is yes, because the variance and mean are equal).

Poisson Approximation of the Binomial

- The Poisson distribution can sometimes be used to approximate the binomial distribution
- When the number of observations n is large, and the success probability p is small, the $B(n,p)$ distribution approaches the Poisson distribution with the parameter given by $m = np$.
- This is useful since the computations involved in calculating binomial probabilities are greatly reduced.
- As a rule of thumb, n should be greater than 50 with p very small, such that np should be less than 5.
- If the value of p is very high, the definition of what constitutes a "success" or "failure" can be switched.

Poisson Approximation: Example

- Suppose we sample 1000 items from a production line that is producing, on average, 0.1% defective components.
- Using the binomial distribution, the probability of exactly 3 defective items in our sample is

$$P(X=3) = {}^{1000}C_3 \times 0.001^3 \times 0.999^{997}$$

Poisson Approximation: Example

Lets compute each of the component terms individually.

- ${}^{1000}C_3$

$$ {}^{1000}C_3 = \frac{1000 \times 999 \times 998}{3 \times 2 \times 1} = 166,167,000 $$

- 0.001^3

$$ 0.001^3 = 0.000000001 $$

- 0.999^{997}

$$ 0.999^{997} = 0.36880 $$

Multiply these three values to compute the binomial probability $P(X = 3) = 0.06128$

Poisson Approximation: Example

- Lets use the Poisson distribution to approximate a solution.
- First check that $n \geq 50$ and $np < 5$ (Yes to both).
- We choose as our parameter value $m = np = 1000 \times 0.001 = 1$

$$P(X = 3) = \frac{e^{-1} \times 1^3}{3!} = \frac{e^{-1}}{6} = \frac{0.36787}{6} = 0.06131$$

Compare this answer with the Binomial probability P(X = 3) = 0.06128. Very good approximation, with much less computation effort.

Continuous Random variables

- Previously we have been studying discrete random variables, such as the Binomial and the Poisson random variables.
- Now we turn our attention to continuous random variables.
- Recall that a continuous random variable is one which takes an infinite number of possible values, rather than just a countable number of distinct values.
- Continuous random variables are usually measurements.
- Examples include height, weight, the amount of sugar in an orange, the time required to run a mile.

Exact Probabilities

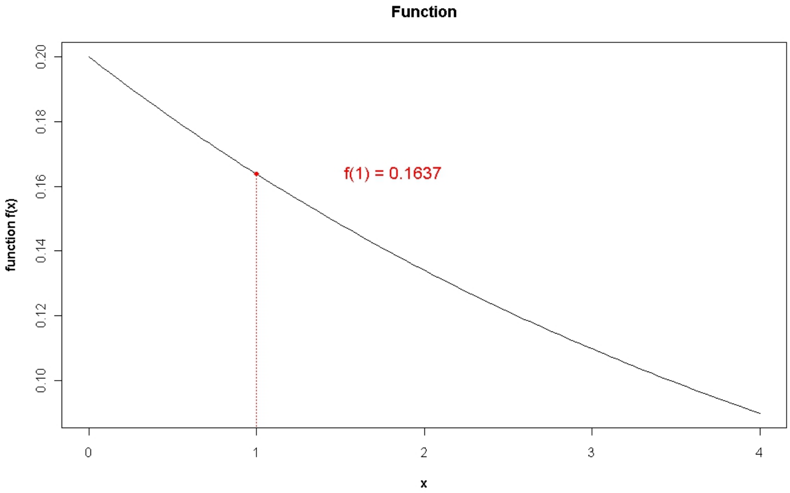
Remarks: This is for continuous distributions only.

- The probability that a continuous random variable will take an exact value is infinitely small. We will usually treat it as if it was zero.
- When we write probabilities for continuous random variables in mathematical notation, we often retain the equality component (i.e. the "...or equal to..").
  For example, we would write expressions $P(X \leq 2)$ or $P(X \geq 5)$.
- Because the probability of an exact value is almost zero, these two expression are equivalent to $P(X < 2)$ or $P(X > 5)$.
- Also, the complement of $P(X \geq k)$ can be written as $P(X \leq k)$.

Functions and Definite integrals

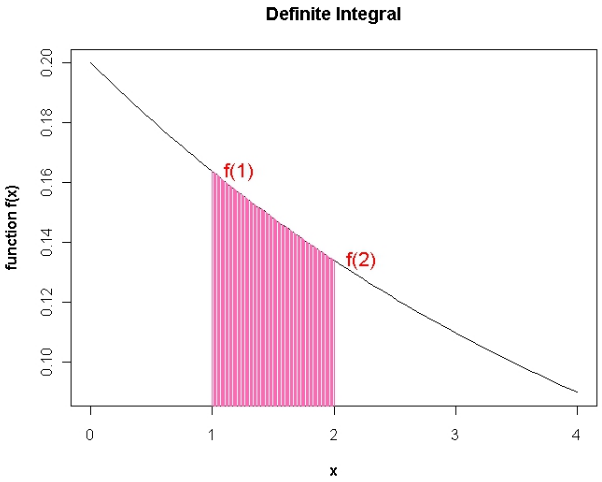
- Integration is not part of the syllabus, and it is assumed that students are not familiar with how to compute definite integrals.
- However, it is useful to know what the purpose of definite integrals are, because we will be using the results derived from definite integrals.
- It is assumed that students are familiar with functions.

Functions



Some function $f(x)$ evaluated at $x = 1$.

Definite Integral



Definite integral of function is area under curve between X=1 and X=2.

Definite Integral

- Definite integrals are used to compute the “area under curves”.
- Definite integrals are defined by a lower and upper limit.
- The area under the curve between X=1 and X=2 is depicted in the previous slide.
- By computing the definite integral, we are able to determine a value for this area.
- Probability can be represented as an area under a curve.

Probability Density Function

- In probability theory, a ***probability density function*** (PDF) (or "density" for short) of a continuous random variable is a function that describes the relative likelihood for this random variable to occur at a given point.
- The PDF for a continuous random variable X is often denoted $f(x)$.
- The probability density function can be integrated to obtain the probability that the random variable takes a value in a given interval.
- The probability for the random variable to fall within a particular interval is given by the integral of this variable's density over the region.
- The probability density function is non-negative everywhere, and its integral over the entire space is equal to one.

Density Curves

- A plot of the PDF is referred to as a ‘***density curve***’.
- A density curve that is always on or above the horizontal axis and has total area underneath equal to one.
- Area under the curve in a range of values indicates the proportion of values in that range.
- Density curves come in a variety of shapes, but the normal distribution’s bell-shaped densities are perhaps the most commonly encountered.
- Remember the density is only an approximation, but it simplifies analysis and is generally accurate enough for practical use.

The Cumulative Distribution Function

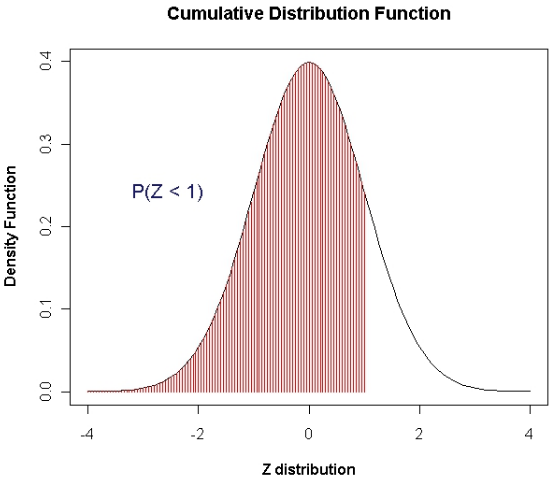
Recall:

- The ***cumulative distribution function*** (CDF), (or just distribution function), describes the probability that a continuous random variable X with a given probability distribution will be found at a value less than or equal to x.

$$F(x) = P(X \leq x)$$

- Intuitively, it is the “area so far” function of the probability distribution.

Cumulative Distribution Function



Cumulative Distribution Function $P(Z \leq 1)$

Here the random variable is called Z (we will see why later)

Continuous Uniform Distribution

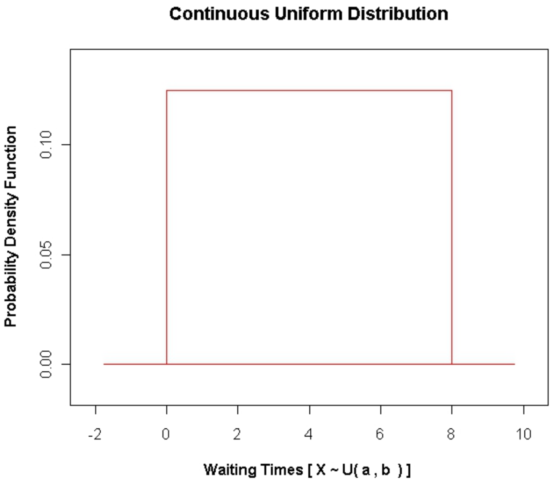
A random variable X is called a continuous uniform random variable over the interval (a,b) if it's probability density function is given by

$$f(x) = \frac{1}{b-a} \qquad \text{when } a \leq x \leq b \text{ (otherwise } f(x) = 0)$$

The corresponding cumulative density function is

$$F(x) = P(X \leq x) = \frac{x-a}{b-a} \qquad \text{when } a \leq x \leq b$$

The Continuous Uniform Distribution



Continuous Uniform Distribution

- The continuous uniform distribution is very simple to understand and implement, and is commonly used in computer applications (e.g. computer simulation).
- It is also known as the 'Rectangle Distribution' for obvious reasons.
- We specify the word "continuous" so as to distinguish it from it's discrete equivalent: the discrete uniform distribution.
- Remark; the dice distribution is a discrete uniform distribution with lower and upper limits 1 and 6 respectively.

Uniform Distribution Parameters

The continuous uniform distribution is characterized by the following parameters

- The lower limit a
- The upper limit b
- We denote a uniform random variable X as $X \sim U(a,b)$

It is not possible to have an outcome that is lower than a or larger than b.

$$P(X \leq a) = P(X \geq b) = 0$$

Interval Probability

- We wish to compute the probability of an outcome being within a range of values.
- We shall call this lower bound of this range L and the upper bound U.
- Necessarily L and U must be possible outcomes.
- The probability of X being between L and U is denoted $P(L \leq X \leq U)$.

$$P(L \leq X \leq U) = \frac{U - L}{b - a}$$

- (This equation is based on a definite integral).

Uniform Distribution: Cumulative Distribution

- For any value "c" between the minimum value a and the maximum value b, we can say
- $P(X \geq c)$

$$P(X \geq c) = \frac{b-c}{b-a}$$

  here b is the upper bound while c is the lower bound
- $P(X \leq c)$

$$P(X \leq c) = \frac{c-a}{b-a}$$

  here c is the upper bound while a is the lower bound.

Uniform Distribution: Mean and Variance

- The mean of the continuous uniform distribution, with parameters a and b is

$$E(X) = \frac{a+b}{2}$$

- The variance is computed as

$$V(X) = \frac{(b-a)^2}{12}$$

Uniform Distribution: Example

- Suppose there is a platform in a subway station in a large large city.
- Subway trains arrive **every three minutes** at this platform.
- What is the shortest possible time a passenger would have to wait for a train?
- What is the longest possible time a passenger will have to wait?

Uniform Distribution: Example

- What is the shortest possible time a passenger would have to wait for a train?
- If the passenger arrives just before the doors close, then the waiting time is zero.

$$a = 0 \text{ minutes } \; = 0 \text{ seconds}$$

Uniform Distribution: Example

- What is the longest possible time a passenger will have to wait?
- If the passenger arrives just after the doors close, and missing the train, then he or she will have to wait the full three minutes for the next one.

$$b = 3 \text{ minutes } = 180 \text{ seconds}$$

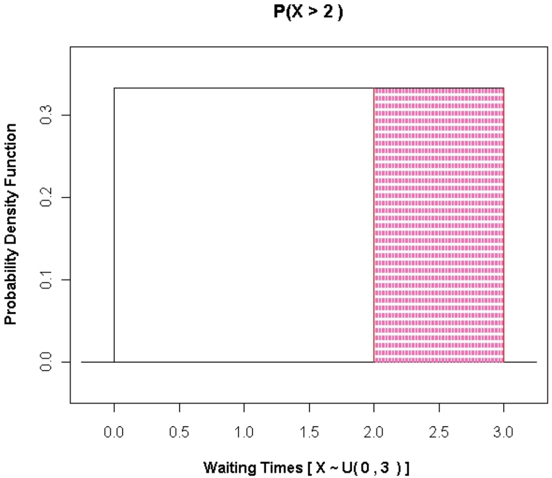
Uniform Distribution: Example

- What is the probability that he will have to wait longer than 2 minutes?

$$P(X \geq 2) = \frac{3-2}{3-0} = 1/3 = 0.33333$$

- See next slide (shaded area is 1/3 of rectangle)

The Continuous Uniform Distribution



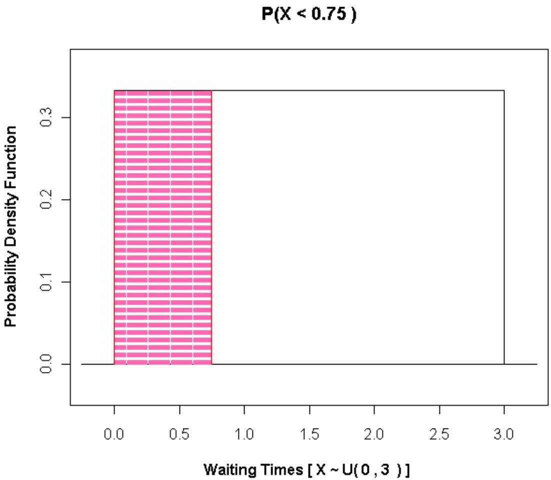
Uniform Distribution: Example

- What is the probability that he will have to wait less than than 45 seconds (i.e. 0.75 minutes)?

$$P(X \leq 0.75) = \frac{0.75 - 0}{3 - 0} = 0.75/3 = 0.250$$

- See next slide (shaded area is 1/4 of rectangle)

The Continuous Uniform Distribution



Uniform Distribution: Expected Value

We are told that, for waiting times, the lower limit a is 0, and the upper limit b is 3 minutes.

The expected waiting time $\mathrm{E}[X]$ is computed as follows

$$\mathrm{E}[X] = \frac{b+a}{2} = \frac{3+0}{2} = 1.5 \text{ minutes}$$

Uniform Distribution: Variance

The variance of the continuous uniform distribution, denoted $\mathrm{V}[X]$, is computed using the following formula

$$\mathrm{V}[X] = \frac{(b-a)^2}{12}$$

For our previous example this is

$$\mathrm{V}[X] = \frac{(3-0)^2}{12} = \frac{3^2}{12} = \frac{9}{12} = 0.75$$

Continuous Distributions: Current Status

- (The Continuous Uniform Distribution, Not examinable)
- The Exponential Distribution (Examinable for midterm)
- (Exponential Distribution is the Cut-off point for Mid-Term 1)
- The Normal Distribution
- The Standard Normal (Z) Distribution.
- Applications of Normal Distribution

Exponential Distribution

The Exponential Distribution may be used to answer the following questions:

- How much time will elapse before an earthquake occurs in a given region?
- How long do we need to wait before a customer enters our shop?
- How long will it take before a call center receives the next phone call?
- How long will a piece of machinery work without breaking down?

Exponential Distribution

- All these questions concern the time we need to wait before a given event occurs. If this waiting time is unknown, it is often appropriate to think of it as a random variable having an exponential distribution.
- Roughly speaking, the time X we need to wait before an event occurs has an exponential distribution if the probability that the event occurs during a certain time interval is proportional to the length of that time interval.

Probability density function

The probability density function (PDF) of an exponential distribution is

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

The parameter λ is called ***rate*** parameter.

Cumulative density function

The cumulative distribution function (CDF) of an exponential distribution is

$$P(X \leq x) = F(x) = \begin{cases} 1 - e^{-\lambda x}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

The complemeent of the CDF (i.e. $P(X \geq x)$ is

$$P(X \geq x) = \begin{cases} e^{-\lambda x}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

Expected Value and Variance

The expected value of an exponential random variable X is:

$$E[X] = \frac{1}{\lambda}$$

The variance of an exponential random variable X is:

$$V[X] = \frac{1}{\lambda^2}$$

Exponential Distribution: Example

Assume that the length of a phone call in minutes is an exponential random variable X with parameter $\lambda = 1/10$. If someone arrives at a phone booth just before you arrive, find the probability that you will have to wait

(a) less than 5 minutes,

(b) between 5 and 10 minutes.

Exponential Distribution: Example

(a) $P(X \leq 5) = 0.39346934$

(b) $P(5 \leq X \leq 10)$
$= P(X \leq 10) - P(X \leq 5)$
$=$ 0.6321- 0.3934
$=$ 0.2386
$=$ 23.86 %

(c) Alternative approach to (b)
$P(5 \leq X \leq 10)$
$= P(X \geq 5) - P(X \geq 10)$
$= e^{-0.5} - e^{-1}$ =0.6065 - 0.3678
$=$ 0.2386 $=$ 23.86 %

Exponential Distribution

- The Exponential Rate
- Related to the Poisson mean (m)
- If we expect 12 occurrences per hour - what is the rate?
- We would expected to wait 5 minutes between occurrences.