Chapter 3: Probability distributions

- We extend the probability analysis by considering random variables (usually the outcome of a probability experiment)
- These (usually) have an associated probability distribution
- Once we work out the relevant distribution, solving the problem is usually straightforward

Random variables

- Most statistics (e.g. the sample mean) are random variables
- Many random variables have well-known probability distributions associated with them
- To understand random variables, we need to know about probability distributions

Some standard probability distributions

- Binomial distribution
- Normal distribution
- Poisson distribution

When do they arise?

- Binomial - when the underlying probability experiment has only two possible outcomes (e.g. tossing a coin)

- Normal - when many small independent factors influence a variable (e.g. IQ, influenced by genes, diet, etc.)

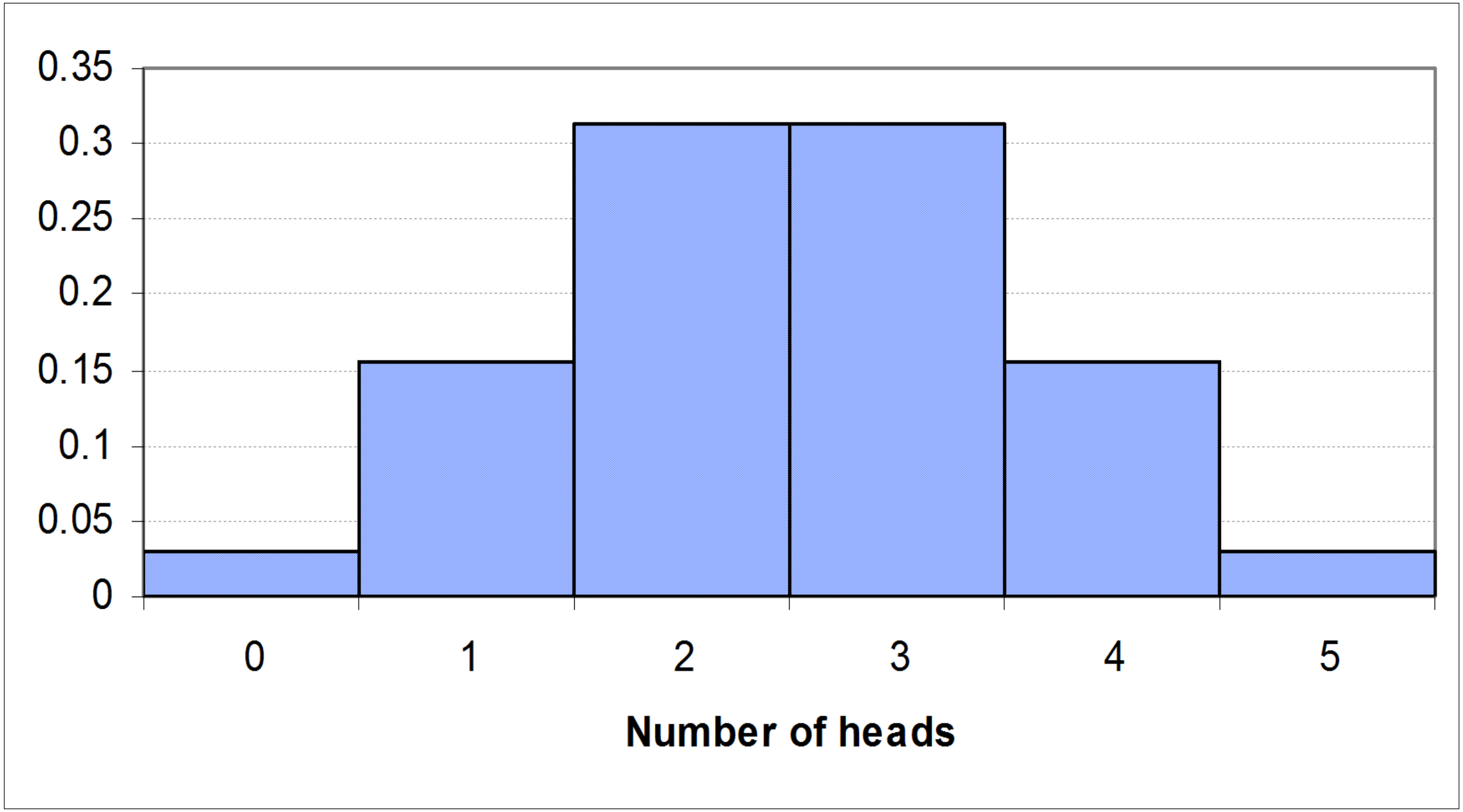
- Poisson - for rare events, when the probability of occurrence is low

The Binomial distribution

- Pr(*r* Heads in five tosses of a coin)
  - $\Pr(r = 0) = (\tfrac{1}{2})^0 \times (\tfrac{1}{2})^5 \times 5C0 = {}^1/_{32} \times 1 = {}^1/_{32}$
  - $\Pr(r = 1) = (\tfrac{1}{2})^1 \times (\tfrac{1}{2})^4 \times 5C1 = {}^1/_{32} \times 5 = {}^5/_{32}$
  - $\Pr(r = 2) = (\tfrac{1}{2})^2 \times (\tfrac{1}{2})^3 \times 5C2 = {}^1/_{32} \times 10 = {}^{10}/_{32}$
  - $\Pr(r = 3) = (\tfrac{1}{2})^3 \times (\tfrac{1}{2})^2 \times 5C3 = {}^1/_{32} \times 10 = {}^{10}/_{32}$
  - $\Pr(r = 4) = (\tfrac{1}{2})^4 \times (\tfrac{1}{2})^1 \times 5C4 = {}^1/_{32} \times 5 = {}^5/_{32}$
  - $\Pr(r = 5) = (\tfrac{1}{2})^5 \times (\tfrac{1}{2})^0 \times 5C5 = {}^1/_{32} \times 1 = {}^1/_{32}$

Slide 3.6

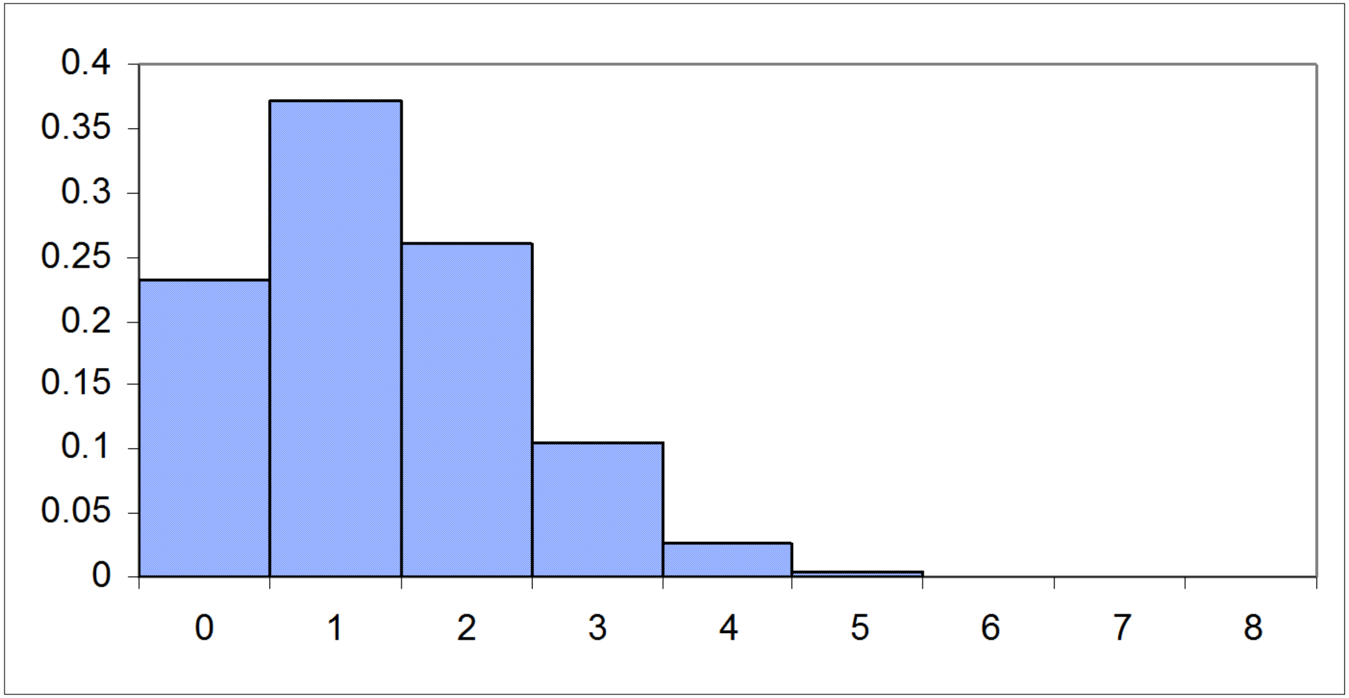
The probability distribution of five tosses of a coin



Slide 3.7

The Binomial distribution with different parameters

- Eight tosses of an unfair coin ($P = {}^{1}/_{6}$)



The Binomial 'family'

- Like other distributions, the Binomial is a family of distributions, members being distinguished by their different parameters.

- The parameters of the Binomial are:
  - P - the probability of 'success'
  - n - the number of trials

- Notation: $r \sim B(n, P)$

The Binomial ‘family’ (continued)

- $r \sim \mathrm{B}(n, P)$ means

  $\Pr(r) = P^r \times (1-P)^{(n-r)} \times nCr$

- E.g. $r \sim \mathrm{B}(5, ½)$ means

  $\Pr(r) = (½)^r \times (1- ½)^{(5-r)} \times 5Cr$

- and from this we can work out $\Pr(r)$ for any value of r.

Mean and variance of the Binomial

- From the diagram of the Binomial it is evident that we should be able to calculate its mean and variance
  - Mean = $n \times P$
  - Variance = $n \times P \times (1-P)$

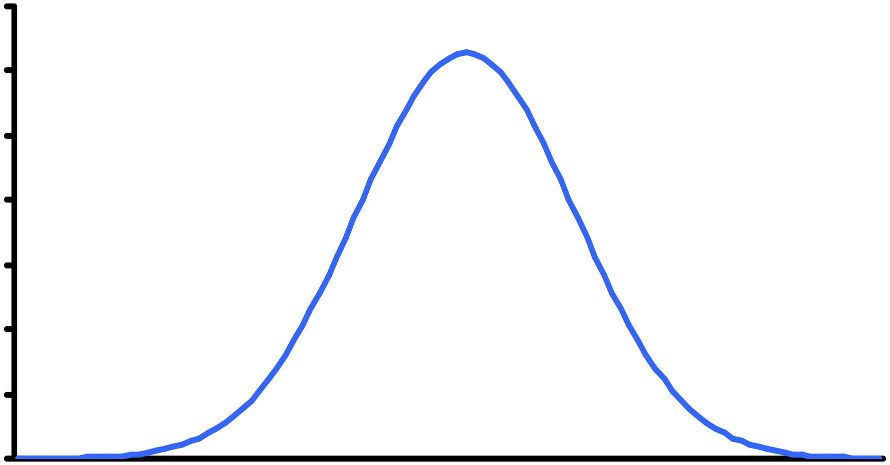
- On average, you would expect 10 Heads from (n =) 20 tosses of a fair (P = ½) coin (10 = 20 × ½)

The Normal distribution

- Examples of Normally distributed variables:
  - IQ
  - Men’s heights
  - Women’s heights
  - The sample mean

The Normal distribution (continued)

- The Normal distribution is
  - bell shaped
  - symmetric
  - unimodal
  - and extends from $x = -\infty$ to $+\infty$ (in theory)

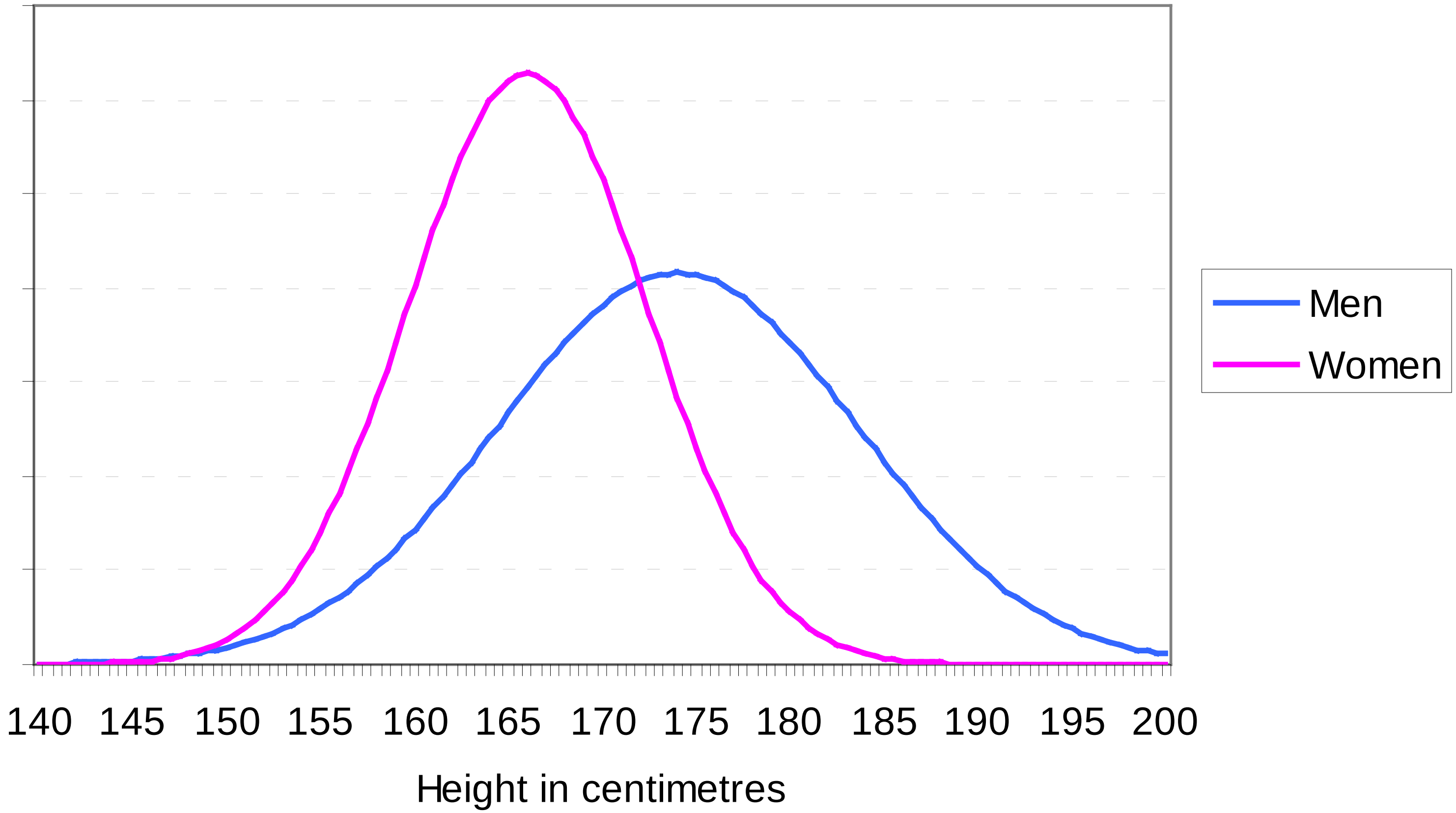


Parameters of the distribution

- The two parameters of the Normal distribution are the mean μ and the variance σ^2
  - $x \sim N(\mu, \sigma^2)$
- Men's heights are Normally distributed with mean 174 cm and variance 92.16
  - $x_M \sim N(174, 92.16)$
- Women's heights are Normally distributed with a mean of 166 cm and variance 40.32
  - $x_W \sim N(166, 40.32)$

Slide 3.14

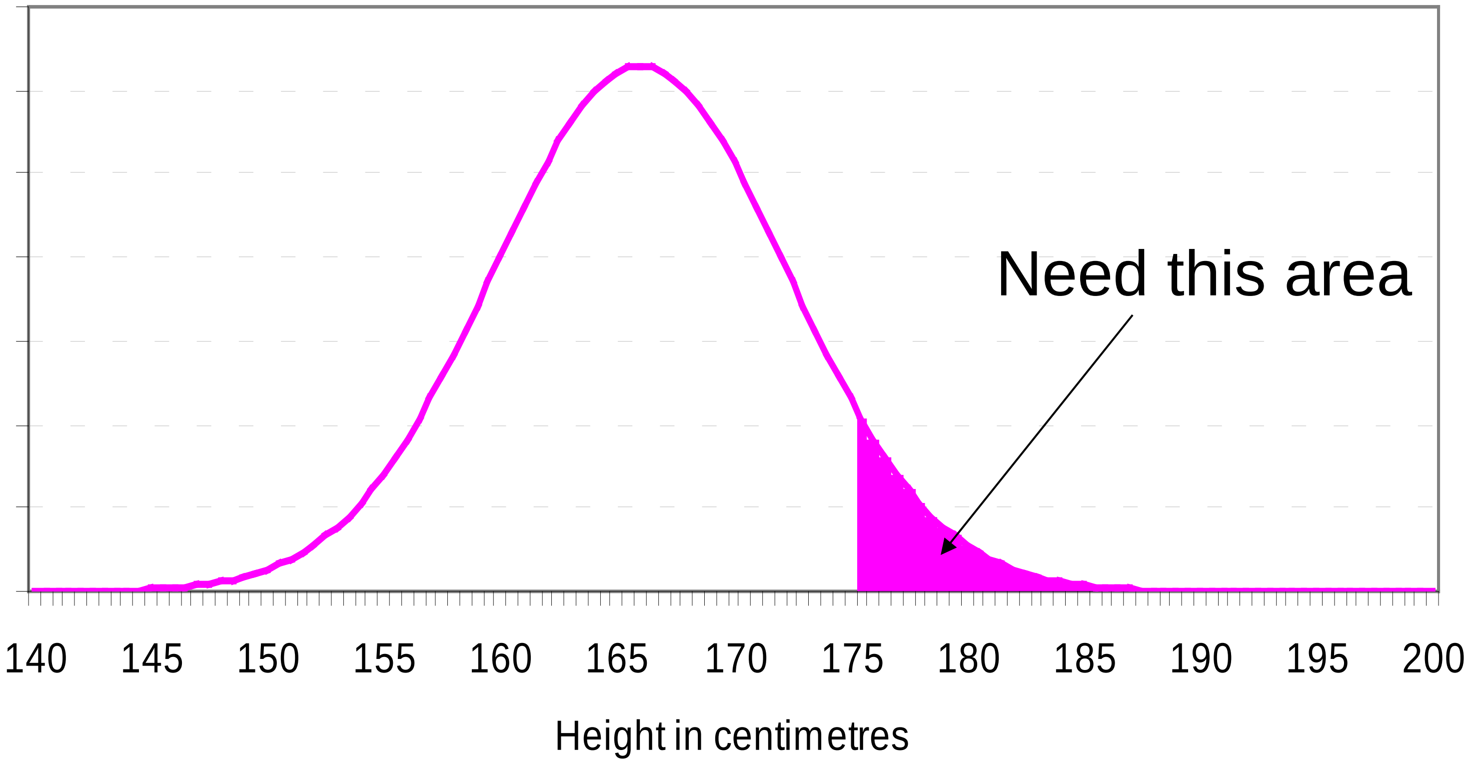
Graph of men’s and women’s heights



Slide 3.15

Areas under the distribution

- What proportion of women are taller than 175 cm?



Areas under the distribution (continued)

- How many standard deviations is 175 above 166?
- The standard deviation is $\sqrt{40.32} = 6.35$, hence

$$z = \frac{175 - 166}{6.35} = 1.42$$

- so 175 lies 1.42 s.d's above the mean
- How much of the Normal distribution lies beyond 1.42 s.d's above the mean? Use tables...

Slide 3.17

Table A2 The standard Normal distribution

z	0.00	0.01	0.02	0.03	0.04	0.05	…
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	
1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	

Answer

- 7.78% of women are taller than 175 cm.
- Summary: to find the area in the tail of the distribution, calculate the *z*-score, giving the number of standard deviations between the mean and the desired height. Then look the *z*-score up in tables.

The distribution of the sample mean

- If samples of size *n* are randomly drawn from a Normally distributed population of mean μ and variance σ^2, the sample mean is distributed as

$$\bar{x} \sim N\left(\mu, \sigma^2/n\right)$$

- E.g. if samples of 50 women are chosen, the sample mean is distributed

$$\bar{x} \sim N\left(166, 40.32/50\right)$$

Example

- What is the probability of drawing a sample of 50 women whose *average* height is > 168 cm?

$$z = \frac{168 - 166}{\sqrt{40.32/50}} = 2.23$$

- z = 2.23 cuts off 1.29% in the upper tail of the standard Normal distribution

The distributions of *x* and of $\bar{x}$

- Note the distinction between

$$x \sim N\left(\mu, \sigma^2\right)$$

  - and

$$\bar{x} \sim N\left(\mu, \sigma^2 / n\right)$$

- The former refers to the population (or equivalently, a typical member of the population), the latter to the sample mean

The Central Limit Theorem

- If the sample size is large (n > 25) the population does not have to be Normally distributed, the sample mean is (approximately) Normal whatever the shape of the population distribution.

- The approximation gets better, the larger the sample size. 25 is a safe minimum to use.

The Poisson distribution

- The ‘rare event’ distribution
- Use in place of the Binomial where *nP* < 5

$$\Pr(x) = \frac{\mu^x e^{-\mu}}{x!}$$

- where μ is the mean of the distribution

Example

- A manufacturer claims a failure rate of 0.2% for its hard disk drives. In an assignment of 500 drives, what is the probability, none are faulty, one is faulty, etc?

- On average, 1 drive (0.2% of 500) should be faulty, so $\mu = 1$.

Example (continued)

- The probability of no faulty drives is

$$\Pr(x=0)=\frac{1^0 e^{-1}}{0!}=0.368$$

- The probability of one faulty drive is

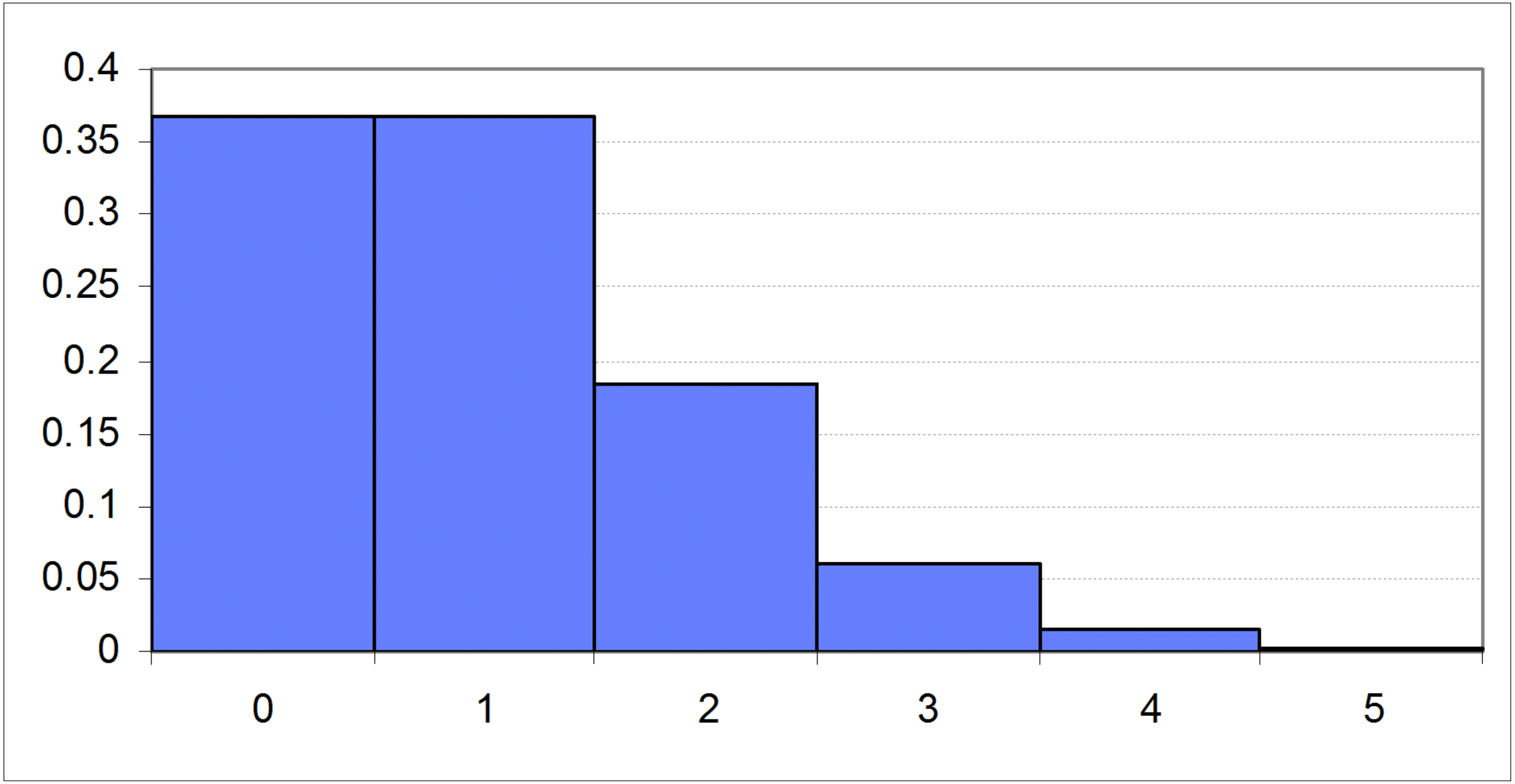
$$\Pr(x=1)=\frac{1^1 e^{-1}}{1!}=0.368$$

- and

$$\Pr(x=2)=\frac{1^2 e^{-1}}{2!}=0.184$$

Slide 3.26

Graph of the Poisson distribution, $\mu = 1$



Example 2

- An ambulance station receives, on average, ten emergency calls over an eight hour period. What is the probability of no emergency calls in a 15 minute period?

- $\mu = 10 \times 15/480 = 0.3125$

$$\Pr(x=0) = \frac{0.3125^0 e^{-0.3125}}{0!} = 0.732$$

Summary

- Most statistical problems concern random variables which have an associated probability distribution
- Common distributions are the Binomial, Normal and Poisson (there many others)
- Once the appropriate distribution for the problem is recognised, the solution is relatively straightforward