Lecture 7.
Sampling Distributions

Data are observations of outcomes of a random process. The simplest, and quite common, situation occurs when all the data measure the same thing – i.e. can be regarded as multiple realisation from the same probability model. But sometimes more general models are required.
A *sample* is a set of observations.

Types of data:
1. Continuous measurements of the same quantity.
VOL data , weighing the 10g Standard Data , height of people, weight gains of pigs, plank strengths etc. are examples.

2. Discrete measurements of the same *ordinal* quantity.
Car Sales, number of goals (by one or both teams (sum)) , numbers of students passing /failing exam, numbers of marks obtained by students on a question (out of 20 say). Marks obtained in an exam ? – these are often treated as continuous!
The issue here is whether the numbers are confined to a small set of values – need a discrete model or many values in which case a simpler cts model is adequate.

3.Discrete measurements of a *nominal* response.
Binary responses such as yes/no, male/female etc.
Reponses confined to a few values : Windows XP, Windows 2000, Linux . Bad, Average, Good etc.
In such situations (even when numerically coded) the average value may not make sense.

4.Data from different models

Pairs (x,Y), xs are know constants and some aspect of Y is a function of x.
Eg. x=size of file in KB, Y=download time – the mean of Y depends on x.

$$E(Y) = a + bx$$

or

x=number of records in data base, Y=retrieval time.

The data are still univariate – the x are assumed known.

Time series – RAM prices by month time plays the role of x here.

5.Data are vectors

Eg . (Height,Weight) of a person, (Processor speed, RAM, Storage) characteristics of a PC.
Here the interest is on relationships between the coordinates.

6. More complex structures.

Curves, Images etc.

Extracting information from a sample

Ex.

Data has been collected on the number of Spam messages that arrive at email account during a 24 hour period.

Sample of 10 24h periods

28	20	24	34	30	36	18	26	24	27

So what?

Why was the information collected?

Say:
1. Average number of Spam messages?
2. Probability more than 35, 40 etc.
3. Probability less than 20?

We can answer these questions directly from the sample:

$$Average = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = 26.7$$

$$P(X > 35) = \frac{1}{10} \qquad P(X > 40) = 0 \qquad P(X < 20) = \frac{1}{10}$$

These are estimates of "interesting" based on data only – empirical estimates.

A new sample may give different answers.

Using a model to augment the sample.

Suppose we consider the arrival of messages as a random process. Specifically as a Poisson process with rate λ per 24h.

Not unreasonable

Then there is only one thing to estimate the rate λ.

For a Poisson random variable E(X)=λ

We estimate E(X) by $\bar{X}$ the sample equivalent of E(X).

So $\hat{\lambda} = 26.7$

We now can calculate

P(X>35) = 1-Poisson(35,26.7,1)=0.049
P(X<20)=Poisson(19,26.7,1)= 0.076

There are two things that we need to say about this procedure:
The quality of the estimates depends on:

1.The Poisson process assumption
2.The accuracy of the estimate of λ.

With so few observations it is not possible to check 1 with any certainty (with more than a 100 we probably could). But if the real process is *similar* to a P.P then the estimates will be reasonable.

One check that we can make is:

For Poisson Var(X) = λ.

The sample equivalent (estimate of) variance of X is:

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

For this sample it is 32.01 not sufficiently different from 26.7 to be worrying.
(We could verify this by simulating samples of size 10 from Poisson(26.7) and conclude that values ≥ 32.01 are quite likely – or use algebra)
This doesn't prove the P.P. but values of variance say 3 or 150 would disprove it. So it is a check we should make.

2. This aspect is more serious $\hat{\lambda} = 26.7$ is only an estimate- it could be 25, 30, or maybe even 10 or 40. Later we shall see how to establish a plausible range for λ based on the sample.
The procedure relies on the fact that we would be unlikely to observe a sample average of 26.7 if λ=10.

Based on this sample we can conclude that λ is likely to lie in the range:

(23.1, 30.3)

Values in this range lead to a reasonable probability of observing an average of 26.7

This leads to a likely range for the probability P(X<20)

Poisson(19,30.3,1)< P(X<20) < Poisson(19,23.1,1)

0.02 < P(X<20) < 0.23

Which is not very impressive! A very wide range !

The issue here is that we discover that we don't have enough data to estimate this probability.

The extraction of information from data is a complicated process.

We can calculate basic summaries but this is not the whole story. We have to acknowledge that we are dealing with samples so we must allow for sampling variation.

Of course if we have lots of data some of the problems disappear. With a 1000 24h periods the simple, direct estimates will be good.

But we would probably discover that they don't actually make sense !

The number of Spam messages probably depends on the day of the week.
1000 * 24h is nearly 3 years –the numbers will change over the course of the period.

The questions we would ask would be more complex:

> How has the rate changed?
> Is there periodicity (day of the week effect).
> We would probably investigate different types of messages.

We still need statistical procedures to address these.

Mean and Variance of $\bar{X}$

The main issue in the problem above was the uncertainty associated with our estimate of λ.

$$\hat{\lambda} = \bar{X}$$

It turns out that for almost all estimation problems revolve round $\bar{X}$.

Binomial
Number of defectives in sample of size n.

Need to estimate p=$\dfrac{np}{n} = \dfrac{E(X)}{n}$.

Here we have 1 observation X so $\bar{X}$=X

$$\hat{p} = \frac{\bar{X}}{n}$$

Geometric

$$E(X) = \frac{1}{p} \quad \text{->} \quad \hat{p} = \frac{1}{\bar{X}}$$

Exponential

$$E(X) = \frac{1}{\lambda} \Rightarrow \hat{\lambda} = \frac{1}{\bar{X}}$$

Gamma
has two parameters a and λ.
A bit more tricky here but if a is known (assumed).

$$E(X) = \frac{a}{\lambda} \Rightarrow \hat{\lambda} = \frac{\bar{X}}{a},$$

λ known $\quad \hat{a} = \lambda \bar{X}$

If both a and λ have to be estimated the best estimates involve solving equations involving gamma functions.

Weibull a mess.

So not always, but in most cases $\bar{X}$ is the key. If its not life gets complicated.

For this reason we want to take a closer look at the properties of $\bar{X}$.

Mean

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

Where $\mu = E(X)$.

So the expected value of the sample average is the same as the expected value of the individual observations.

Variance
$E(X) = \mu, \quad V(X) = \sigma^2$

$$V(\bar{X}) = V\left(\frac{1}{n}\sum X_i\right) = \frac{1}{n^2}V(\sum X_i)$$

$= \frac{1}{n^2}\sum V(X_i)$ note this step requires independence of X_i.

$$= \frac{1}{n^2}\sum \sigma^2 = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}$$

So $\bar{X}$ is less variable than the Xs , the variance is n times smaller.

As $n \to \infty$, the variance tends to 0. $\bar{X}$ tends to a constant (no variation) . This constant is μ.

So as n increases $\bar{X}$ tends to the true value of E(X). That is why we use $\bar{X}$ for all these estimates. The rate of convergence is $\frac{1}{\sqrt{n}}$ (in the same units as E(X))

This is what guarantees that $\bar{X}$ is close to E(X). How close? This will depend on n, σ^2 and the distribution of $\bar{X}$.

The exact distribution of $\bar{X}$ will depend on the underlying distribution of X, but as we shall see in the next lecture not a lot !