Lecture 10 random variables

- All measuring devices have a limit on how accurate they can measure so that all measurements contain *error*.
- *Error* means the difference between the theoretical or predicted outcome and the observed data.
  \[ \epsilon = y - \hat{y} \]
- Sometimes the error is negligible as in tide predictions and can be disregarded but mostly interpretation of data requires addressing the error component.
- If the data arises from different sources, the observed differences may be due to:
  (i) the sources
  (ii) random variation
- Statistics is the science of building mathematical models which attribute observed differences as a combination of systematic and random effects.
- Data \sim \text{Systematic effects} + \text{random effects}
- Correct notation is important in statistics
  
  *A random variable*
  
  - A random experiment has been defined in the introduction to probability.
  - A measurement from a random experiment is a random variable.

  **Formal Definition**
  
  - A variable which cannot be predicted exactly, but whose future outcome can be allocated to specific values with known probabilities.

  *Discrete random variables*
  
  - i.e. integer valued random variables.
  - Examples of discrete random variables are:-
    - binary variables (e.g. T/F),
    - counts,
    - integer scores,
    - the numbers of weeds in a unit area,
    - the number of faulty components in a batch of 100.
  - The collection of probabilities for each value of the random variable is termed a *probability distribution*.

  **Notation**
  
  - Denote the random variable by upper case letters, e.g. \( Y \)
  - particular values that they take are denoted by lower case letters, e.g \( y \).
  - The probability that \( Y \) takes the value \( y \) is \( P(Y = y) \).
  - Note that \( Y \) is a random variable but \( y \) is not.
• Y may represent any one of the 6 outcomes from tossing a die

• y is the observed outcome

• The expression \( Y = y \) denotes the set of outcomes assigned the value y by the random variable Y.

• The pattern of randomness of continuous random variables is the density function.
  
  – Probabilities associated with an interval are determined by areas under the curve.

• Discrete random variables have probability distributions or probability functions
  
  – Probabilities of discrete random variables over an interval are summed.

  Bernoulli distribution

I order 1kg of bacon, and wrap the slices in pairs. What is the distribution of the number of slices left over?

Ans.

There can only be 1 slice or no slices left, so only two discrete outcomes are possible, i.e. \( n(S) = 2 \). In general, unless we know more \( P(X = 0) = P(X = 1) = \frac{1}{2} \), where \( X \) is the number of slices left over.

This is a Bernoulli distribution.

  Binomial distribution

• Place 20 daffodil bulbs in a pot. How many shoot?

• If \( X \) is the number that shoot, we have \( X = 0, 1, 2, 3, \ldots, 20 \).

• What is the distribution of the number that shoot out of 20?

An individual bulb has \( P(X = 0) = 1 - v \), \( P(X = 1) = v \) where \( X \) is the number of viable bulbs, and \( v \) is the viability proportion or probability.

Over 100 days, the number of daily slowdowns due to equipment “failing” are

\[
\begin{array}{cccccccc}
    y & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\
    n(y) & 40 & 25 & 10 & 8 & 7 & 5 & 3 & 2 \\
    P(Y=y) & 0.4 & 0.25 & 0.1 & 0.08 & 0.07 & 0.05 & 0.03 & 0.02 \\
\end{array}
\]

Some results are:-

\[
\begin{align*}
    \bullet P(Y \leq 1) &= P(Y = 0) + P(Y = 1) = 0.65 \\
    \bullet P(1 < Y \leq 3) &= P(Y = 2) + P(Y = 3) = 0.18 \\
    \bullet P(Y > 4) &= 1 - \sum_{n=0}^{4} P(Y = n) = 1 - 0.9 = 0.10
\end{align*}
\]

Probability function
Compound events

(A) \( X \leq x_1 \) \hspace{1cm} x_1 < X \leq x_2 \hspace{1cm} X > x_2

(B) \( X \leq x_1 \) \hspace{1cm} x_1 < X \leq x_2 \hspace{1cm} X > x_2
\[
P(X > x_1) = 1 - P(X \leq x_1) \\
P(X \leq (x_1)) = P[X = 0] + P[X = 1] + \ldots + P[X = x_1] \\
P[X < x_1] = P[X \leq (x_1 - 1)] \\
P[X \geq x_1] = P[X > (x_1 + 1)] \\
P[x_1 \leq X \leq x_2] = P[X = x_1] + P(X = (x_1 + 1)] + P(X = (x_1 + 2)] + \ldots + P[X = x_2] \\
P[x_1 \leq X \leq x_2] = P[(x_1 - 1) < X \leq x_2] = P[X \leq x_2] - P[X \leq (x_1 - 1)]
\]

**Expected value**

- The data from the experiment are interpreted to make inference about the population.
- The *expected value* is the measure of *location* in the population and it is estimated from the sample.
- The operation expected value of a random variable \(X\) is denoted by \(E(X)\).
- Its value is \(\mu\), the Greek letter pronounced “mu”.
- For discrete random variables \(\mu = E(X) = \sum_{i=1}^{n} x_i \times P(X = x_i)\).
- We read \(E(X)\) as “the expected value of \(X\)”.
- Estimates of statistics derived from the sample are denoted by a “hat” over the symbol. This distinction between the statistic and its estimate is very important and must be clear.
- The expected value of the population is estimated by the mean of the sample.

The expected number of slowdowns is

\[
\hat{\mu} = \sum_{i=1}^{7} x_i \times P(X = x_i)
\]

\[
= (0 \times 0.4) + (1 \times 0.25) + (2 \times 0.1) + (3 \times 0.08) + (4 \times 0.07) + \\
(5 \times 0.05) + (6 \times 0.03) + (7 \times 0.02)
\]

\[
= 0 + 0.25 + 0.2 + 0.24 + 0.28 + 0.25 + 0.18 + 0.14 = 1.54
\]

Although the data are discrete, the expected value may be fractional. We interpret this to expect that there will be 1 or 2 slowdowns per day and over 100 days, the total slowdowns is expected to be 154.

**Variance, Standard deviation**

Because the sample is a random variable, the expected value can only be estimated up to the accuracy of an interval.

- \[\text{var}(X) = E \left[ (X - \mu)^2 \right] = \sum_{i=1}^{n} (x_i - \mu)^2 \times P(X = x_i)\]
- \[\text{sd} = \sqrt{E[(X - \mu)^2]}\]
- The notation \( \text{var}(X) \) means “variance of \( X \)” and likewise \( \text{sd}(X) \) is the standard deviation of \( X \).

The variance for slowdowns is

\[
\text{var}(X) = (0 - 1.54)^2 \times 0.4 + (1 - 1.54)^2 \times 0.25 + (2 - 1.54)^2 \times 0.1 \\
+ (3 - 1.54)^2 \times 0.08 + (4 - 1.54)^2 \times 0.07 + (5 - 1.54)^2 \times 0.05 + \\
(6 - 1.54)^2 \times 0.03 + (7 - 1.54)^2 \times 0.02 \\
= 0.95 + 0.07 + 0.02 + 0.17 + 0.42 + 0.60 + 0.60 + 0.60 \\
= 3.46
\]

The standard deviation is 1.86.

\textbf{R calculations}

We can envisage the sample before summary as:-

\[
x <- c(0, 0, 0, \ldots, 0, 1, 1, 1, \ldots, 1, 2, 2, 2, \ldots, 2, 3, \ldots)
\]

That is 40 zeroes, 25 ones, 10 twos and so on.

The mean and variance can be calculated by \texttt{mean(x)} and \texttt{sd(x)}.

\[
> \text{rep(3,6)}
\]

\[
[1] 3 3 3 3 3 3
\]

\textbf{Lists in parallel}

The replication would be done in pairs,

\[
> \text{rep(c(3,4),c(6,2))}
\]

\[
[1] 3 3 3 3 3 3 4 4
\]

That is the \texttt{rep(3,6)} was followed by \texttt{rep(4,2)}.

\textbf{R calculations of the mean and standard deviation}

```{r}
options(digits=3)

nslowdowns <- 0:7
counts <- c(40,25,10,8,7,5,3,2)
Xsample <- rep(nslowdowns,counts)

mnx <- weighted.mean(x=nslowdowns,w=counts)
vx <- var(Xsample)
sdx <- sd(Xsample)

print(table(Xsample))
cat("_____________
")
print(c(mnx,vx,sdx))

> source("mnvar.r")

Xsample

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td>25</td>
<td>10</td>
<td>8</td>
<td>7</td>
<td>5</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

[1] 1.54 3.46 1.86
```

>
Lecture 11 - Binomial Random Variables

- \[ P(E) = \frac{n(E)}{n(S)} \]
  \[ 0 \leq P(E) \leq 1 \]
  \[ P(\bar{E}) = 1 - P(E) \]

- The definition of a random experiment included that the set of possible outcomes was known and that after many samplings, a pattern may appear.
- Rather than always returning to the first principles of axioms, we can establish direct results on probability when the pattern of outcomes can be represented by a mathematical form.

**Binomial**

A variable that has only 2 possible outcomes is a *binary* variable.

Examples of binary outcomes from an experiment are

- True/False,
- success/fail.
- The outcome from coin tossing is often used as an analogy for binary variables.

We can simulate the outcomes of a random experiment of tossing a biassed coin 20 times. The coin is weighted so that the probability of a head is 0.6 (a fair coin would have \( p = 0.5 \)).

```
#________ coin.r ________________
samplesz <- 20
side.up <- sample(c("H","T"),size=samplesz,
replace=T,prob=c(0.6,0.4))
HT.tab <- table(side.up)
print(HT.tab)
> source("coin.r")
side.up
H T
13 7
> source("coin.r")
side.up
H T
15 5
> source("coin.r")
side.up
H T
11 9
```

Over a long run, we would expect the average number of heads to be 12 out of 20 but each sample is not exactly 12.

**Binomial experiment**

Displays the following characteristics:

(i) There are \( n \) identical, independent trials. That is, the conditions of the experiment remain the same for each trial and the outcome of one trial does not influence any other trial.

(ii) There are only two possible outcomes for each trial: “Success” or “Failure”. (Success is defined by the experimenter).

(iii) The probability of “Success” is denoted \( p \) and is the same for each trial. (Probability of “Failure” is \( 1 - p = q \)).

For an underlying population probability of success \( p \), the probability of that the sample from \( n \) independent binomial trails will have \( r \) successes is :-
\[ P(X = r) = \binom{n}{r} p^r (1 - p)^{n-r} \]

\[
\binom{n}{r} = \frac{n!}{r!(n-r)!}.
\]

The notation for describing a binomial distribution is \( B(n, p) \).

**Survival probabilities**

The survival rate during a risky operation for patients with no other hope of survival is 80%.

(a) What is the probability that exactly four of the next five patients survive this operation?

(b) What is the probability that no more than two will die?

Write down

- Sample size \( n = 5 \)
- Survive \( p = 0.2 \), Die \( p = 0.8 \)
- \( P(X = 4) \) when \( p = 0.8 \) or \( P(X = 1) \) when \( p = 0.2 \)
- \( P(X \leq 2) \) when \( p = 0.2 \) or \( P(X > 3) \) when \( p = 0.8 \)

**R functions for calculating binomial probabilities**

Using Rcmdr

**Distributions → Discrete distributions → Binomial**

(a) “point” probabilities are given by `dbinom`. This is the function used when you use the following menu in Rcmdr

- the probability that there will be 4 survivors out of 5 operations, where probability of success is 0.8,
- choose the menu for binomial **probabilities**
- Binomial trials is 5 and probability of success is 0.8

\[
> \text{dbinom}(4, size=5, prob=0.8)
\]

```
[1] 0.000322
```

Probability of 4 successes is 0.0032

(b) Cumulative probabilities are determined with `pbinom`. Observe the use of this function in Rcmdr

- probability that there will be up to 2 deaths \( (p = 0.2) \),
- choose the menu for binomial **tail probabilities**. This will be the cumulative probability of the lower.tail by default or the upper.tail if we specify that.
- binomial trials is 5, probability is 0.2 (for failures), Variable values is 2. this is a quantile.

\[
> \text{pbinom}(c(2), size=5, prob=0.2, lower.tail=TRUE)
\]

```
[1] 0.94208
```
dx <- dbinom(x=0:5,prob=0.8,size=5)

px <- pbinom(q=0:5,prob=0.8,size=5)

\[
P(X = x)
\]

\[
P(X = 3) \text{ dbinom}(x=3,p=0.8,\text{size}=5)
\]
\[
P(X > q) = 1 - P(X \leq q)
\]

- \(P(X \leq 3) \text{ pbinom}(q=3,p=0.8,\text{size}=5)\)
- \(P(X > 3) = 1 - \text{ pbinom}(q=3,p=0.8,\text{size}=5)\) or \(\text{ pbinom}(q=3,p=0.8,\text{size}=5,\text{lower.tail}=F)\) or \(\text{ pbinom}(q=3,p=0.2,\text{size}=5)\)

Note that in this *discrete* case,

- \(P(X < 3) = P(X \leq 2) \text{ pbinom}(q=2,p=0.8,\text{size}=5)\)
- \(P(X \geq 3) = P(X > 2) = 1 - \text{ pbinom}(q=2,p=0.8,\text{size}=5)\) or \(\text{ pbinom}(q=2,p=0.8,\text{size}=5,\text{lower.tail}=F)\) or \(\text{ pbinom}(q=2,p=0.2,\text{size}=5)\)
# example1.r

```r
options(digits=2)
print(dbinom(x=4,size=5,prob=0.8))
print(pbinom(q=2,size=5,prob=0.2))
```

```r
c
```

Hence

\[
P(\text{number of survivors} = 4) = 0.41
\]

\[
P(\text{number of deaths} \leq 2) = 0.94
\]

**Expected value and variance for binomial distribution**

The mathematical form of the binomial distribution allows ready calculation of the expected value and the variance, standard deviation of an observation from a binomial distribution. For sample size \( n \),

- \( E(X) = np \)
- \( \text{var}(X) = np(1 - p) \)

The expected value and variance of a proportion

- \( \hat{p} = \frac{X}{n} \)
- \( E(\hat{p}) = p \)
- \( \text{var}(\hat{p}) = \frac{p(1-p)}{n} \)

For the “operations” example,

- The expected number of survivors from the next 5 operations is
  \[
  E(X) = \mu_X = 5 \times 0.8 = 4
  \]

- The variance is
  \[
  \text{var}(X) = \sigma_X^2 = 5 \times 0.8 \times 0.2 = 0.8
  \]
Lecture 12 - Poisson random variables

The Poisson distribution is commonly used for modelling the randomness of count data such as weeds in a field, colonies on a microscope slide, accidents.

Define the Poisson probability function by

$$ Pr(X = x) = \frac{e^{-\lambda} \times \lambda^x}{x!} $$

- $e$ is the natural base of logarithms $\approx 2.72$
- $X$ is the variable
- $x$ is a particular value it might take.
- The parameter $\lambda$ is the mean rate at which incidents occur.$^2$

$$ Pr(X = 2) = \frac{e^{-\lambda} \times \lambda^2}{2!} $$

and $\lambda$ would be estimated by the average value of the counts over the sample. Even though the data are discrete, $\lambda$ is not an integer, e.g. the mean of 0, 2, 3, 5, 1 is 2.2.

**Poisson example 1**

Suppose that a machine which produces aluminium pins for aeroplanes produces defective pins at the long term rate of 2 pins per hour, ie $\lambda = 2$. Of course, some hour runs do better and some do worse.

What is the probability that a 1 hour run will produce 0, 1 and 2 defective pins?

$$ Pr(\text{defects} = 0) = \frac{e^{-2 \times 2^0}}{0!} = 0.14 $$

$$ Pr(\text{defects} = 1) = \frac{e^{-2 \times 2^1}}{1!} = 0.27 $$

$$ Pr(\text{defects} = 2) = \frac{e^{-2 \times 2^2}}{2!} = 0.27 $$

Suppose we wished to know the probability of finding no more than 2 defects. That is $Pr(\text{defects} = 0) + Pr(\text{defects} = 1) + Pr(\text{defects} = 2)$ or

$$ \sum_{i=0}^{2} Pr(\text{defects} = i) = \sum_{i=0}^{2} \frac{e^{-2} \times 2^i}{i!} . $$

The probability that there are more than 2 defects is

$$ Pr(\text{defects} > 2) = \sum_{i=3}^{\infty} \frac{e^{-2} \times 2^i}{i!} $$

and as $i$ gets large, the contributions to the sum become negligible. However, the more convenient way of evaluating this is

$$ 1 - Pr(\text{defects} \leq 2) = 1 - \sum_{i=0}^{2} \frac{e^{-2} \times 2^i}{i!} , $$

since $\sum_{i=0}^{\infty} Pr(X = i) = 1$.

R program for Poisson probabilities

```r
#_____ poisprobs.r _____
options(digits=2)
v <- 0:6
dv <- dpois(lam=2,x=v)
pv <- ppois(lam=2,q=v)
print(rbind(v,dv,pv))
```

> source("poisprobs.r")

<table>
<thead>
<tr>
<th>v</th>
<th>dv</th>
<th>pv</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.14</td>
<td>0.14</td>
</tr>
<tr>
<td>1</td>
<td>0.27</td>
<td>0.41</td>
</tr>
<tr>
<td>2</td>
<td>0.27</td>
<td>0.68</td>
</tr>
<tr>
<td>3</td>
<td>0.18</td>
<td>0.86</td>
</tr>
<tr>
<td>4</td>
<td>0.09</td>
<td>0.95</td>
</tr>
<tr>
<td>5</td>
<td>0.036</td>
<td>0.983</td>
</tr>
<tr>
<td>6</td>
<td>0.012</td>
<td>0.995</td>
</tr>
</tbody>
</table>

$^1$Named after a French mathematician, early 1800's
$^2$Greek letter “lambda”
\[ P(X \leq 2) = 0.68, \ P(X > 2) = 1 - 0.68 = 0.32. \]
This can be calculated directly in R by

```r
> ppois(lam=2,q=2,lower.tail=F)
[1] 0.32
```
• The lower tail of the distribution is \( X \leq 2 \)
• The upper tail is \( X > 2 \).
• The default for the probability and distribution functions is to return the lower tail probability
• The upper tail probability is obtained by so by specifying `lower.tail=F` or calculating \( 1 - p \). The calculation here is \( P(X > 2) \), with `lower.tail=F`.

In Rcmdr,
• Distributions → Discrete distributions → Poisson → Poisson probabilities
• Enter mean 2

```r
> .Table
   Pr
0 0.1353
1 0.2707
2 0.2707
3 0.1804
4 0.0902
5 0.0361
6 0.0120
7 0.0034
8 0.0009
9 0.0002
```
• Distributions → Discrete distributions → Poisson → Poisson tail probabilities
• Enter mean 2, Variable values 2. Check Upper tail since we want \( P(X > 2) \).

```r
> ppois(c(3), lambda=0.5, lower.tail=FALSE)
[1] 0.001751623
```

**Poisson example 2**
If more than three cars per minute enter a tunnel it is known that a hazardous situation will develop. It is known that on average 1 car enters the tunnel per 2 minute interval.

\[
P(X = x) = \frac{e^{-0.5} \times 0.5^x}{x!}, \quad x = 0, 1, 2, \ldots
\]

R code for Poisson example 2

```r
Distributions → Discrete distributions → Poisson → Poisson tail probabilities
Enter mean 0.5, variable value 3, check upper tail
```

```r
> ppois(c(3), lambda=0.5, lower.tail=FALSE)
[1] 0.001751623
```
\[
P(X > 3) = \sum_{x=4}^{\infty} \frac{e^{-0.5}(0.5)^x}{x!}
\]
\[
= 1 - \sum_{x=0}^{3} \frac{e^{-0.5}(0.5)^x}{x!}
\]
\[
= 1 - 0.998 = 0.002
\]

**Expected value, variance**

The mathematical form of the Poisson distribution leads to explicit results for expected value (or mean) and variance without resorting to the first principles.

\[
E(X) = \lambda
\]
\[
\text{var}(X) = \lambda
\]

**An example of mean-variance relationship**

Warpbreaks tabulates the number of warp breaks per loom in yarn during weaving. There are 2 explanatory factors, wool with levels A and B and tension with levels L, M, H.