STAT2402: Analysis of Observations

R. Nazim Khan
Department of Mathematics and Statistics
nazim.khan@uwa.edu.au

The University of Western Australia

Week 3—Part II

3. Discrete Random Variables

What is a random variable?

1. A random variable is a map from a sample space to the real numbers.
2. As a simple example, toss a coin once and let $X =$ ~~0~~ 1 if a H is tossed, otherwise $X = 0$. Then X is a random variable—its value depends on chance.
3. Probabilities ofr random variables are obtained from the corresponding sample space. So for the above example $\mathrm{P}(X = 1) = 0.5, \mathrm{P}(X = 0) = 0.5$ if the coin is fair.
4. Two types of random variables. **Discrete** random variables take only certain fixed values. Usually arise from some counting process. **Continuous** random variables take all values in some fixed interval. Usually arise from some measurement process.

1.2 – 2.5

3.1 Discrete Random Variables

A *discrete* random variable is one that takes only certain fixed values. For example, toss a fair coin twice, and let the random variable X denote the number of heads tossed. Then X takes values 0 and 1, with probability 0.5 each. We can tabulate the values of th random variable.

x	0	1	2
$p_X(x)$	0.25	0.5	0.25

Note that

$$\boxed{p_X(x)} = \underline{\mathrm{P}(X = x)},$$

where X denotes the random variable and x denotes a realised value of the random variable.

3.2 Mean or Expected Value

The *mean* or *expected value* of a discrete random variable X, denoted $\mathbb{E}(X)$ is given by

$$\mathbb{E}(X) = \sum_x x \; P_X(x)$$

In the coin toss example, the mean number of heads is

$$\mathbb{E}(X) = \underline{0} \times 0.25 + 1 \times 0.5 + 2 \times 0.25 = 1.$$

Mean is a measure of central tendency.
For any function $f(X)$ of a random variable the mean is given by

$$\mathbb{E}\,(f(X)) = \sum_x f(x) \; p_X(x).$$

For the coin toss example,

$$\mathbb{E}\,(X^2) = 0^2 \times 0.25 + 1^2 \times 0.5 + 2^2 \times 0.25 = \underline{\underline{1.5}}.$$

3.3 Variance

The *variance* of a discrete random variable X, denoted $\mathrm{Var(X)}$ is given by

$$\mathrm{Var(X)} = \mathbb{E}\left(\mathrm{X}^2\right) - [\mathbb{E}(\mathrm{X})]^2$$

The *standard deviation* denoted σ_X is is given by

$$\sigma_X = \sqrt{\mathrm{Var(X)}}.$$

For the coin toss example,

$$\mathrm{Var(X)} = \mathbb{E}\left(\boxed{\mathrm{X}^2}\right) - [\mathbb{E}(\mathrm{X})]^2 = \underline{\underline{1.5}} - \underline{1}^2 = \underline{\underline{0.5}},$$

giving the $\underline{\text{standard deviation as } \sigma_X}$ $= \sqrt{0.5}$. Variance is a measure of spread. Standard deviation is also a measure of spread, and is better because it has the same units as the random variable.

Example

Compute the mean and standard deviation of the distribution below.

x	-1	0	1	2
$p_X(x)$	0.1	0.2	0.3	0.4

Mean $E(X) = (-1)(0.1)+(0)(0.2)+(1)(0.3)$
$+ (2)(0.4) = 1$

$E(X^2) = (-1)^2(0.1)+(0)^2(0.2)+(1)^2(0.3)$
$+ (2)^2(0.4) = 2$

$\text{Var}(X) = 2 - (1)^2 = 1$

$\sigma = \sqrt{1} = 1$

Example

Grizzly bear litter sizes are between 1 and 5 cubs, although litters with more than 3 cubs are rare. From data collected, Shideler and Hechtel (Grizzly bear, in *The natural history of an Arctic oil field*, Chapter 6, 2000. Editors: Truett and Johnson, Elsevier Inc.) estimated the probabilities of litter sizes as follows: 11% single-cub litters, 50% two-cub litters, 39% three-litter cubs, 2% four-cub litters and 1% five-cub litters.

1. Calculate the mean and variance of gizzly bear litter size.
2. Ecologists estimate that for a viable population of grizzly bears the number of cubs should not be lower than 2. Is this population of grizzly bears viable? Justify your decision.

Solution

- *Define an appropriate random variable.* Let the random variable X denote the number of cubs in a litter.
- *Determine its distribution.*

Example (ctd)

x	1	2	3	4	5
$p_X(x)$	0.11	0.47	0.39	0.02	0.01

1. The mean number of cubs per litter is

$$\mathbb{E}(X) = 1(0.11) + 2(0.47) + 3(0.39) + 4(0.02) + 5(0.01) = 2.35$$

$$\mathbb{E}\left(X^2\right) = 1^2(0.11) + 2^2(0.47) + 3^2(0.39) + 4^2(0.02) + 5^2(0.01) = 6$$

$$\mathrm{Var(X)} = \sigma^2 = \mathbb{E}\left(X^2\right) - [\mathbb{E}(X)]^2 = 6.07 - 2.35^2 = 0.5475$$

$$\sigma = \sqrt{0.5475} = 0.7399$$

2. The mean number of cubs per litter is 2.35. Further, $P(X < 2) = 0.11$, and the standard deviation of the number of cubs per litter is 0.7399. So it is unlikely that the number of cubs is less than 2. Thus this population of grizzlies is viable.

3.4 The Bernoulli Distribution

Consider a trial with only two possible outcomes, Success (S) or failure, with $\mathrm{P}(S) = p$. Let the random variable $X = 1$ if the trial is a success, and 0 otherwise. Then X has a Bernoulli distribution with parameter p, written as $X \sim \mathrm{Bern}(p)$. It takes only two values, 0 and 1, with $\mathrm{P}\,(X = 1) = p$. The distribution can be tabulated as below.

x	0	1
$P_X(x)$	$1 - p$	p

Then

$$
\begin{aligned}
\mathbb{E}(X) &= (0)(1-p) + (1)(p) = \boxed{p} \\
\mathbb{E}(X^2) &= (0)^2(1-p) + (1)^2 p = p \\
\mathrm{Var(X)} &= p - p^2 = \boxed{p(1-p)} \\
\sigma_X &= \underline{\sqrt{p(1-p)}.}
\end{aligned}
$$

A genetics example.

Both parents are carriers for cystic fibrosis, a particularly unpleasant genetically inherited disease. Let ”A” be the dominant normal allele and ”a” as the recessive abnormal one that is responsible for cystic fibrosis. As carriers, both parents are heterozygous (Aa). This disease only afflicts those who are homozygous recessive (aa). The Punnett square below makes it clear that at each birth, there will be a 25% chance of you having a normal homozygous (AA) child, a 50% chance of a healthy heterozygous (Aa) carrier child like you and your mate, and a 25% chance of a homozygous recessive (aa) child who probably will eventually die from this condition. Let the random variable X denote the event that the child is homozygous recessive, that is, $X = 1$ if the child is aa, and 0 otherwise. Then $X \sim \mathrm{Bern}(0.25)$. The mean and variance of X are

	A	a
A	AA	Aa
a	Aa	aa

$$\mathbb{E}(X) = p = 0.25, \mathrm{Var}(\mathrm{X}) = \mathrm{p}(1-\mathrm{p}) = 0.25 \times 0.75 = 0.1875.$$

3.5 Binomial distribution

Toss a fair coin ten times. What is the probability of obtaining five heads? This example encapsulates some key features that are common in many situations. We formalise this below.

Binomial distribution

1. A fixed number n of independent and identical trials.
2. Each trial has exactly two possible outcomes, denoted **success** (S) and **failure** (F).
3. The probability of success is p which is fixed throughout the trials.
4. Let the rv X denote the number of successes in these trials.

Then X has a **binomial distribution** with parameters n and p. We write $X \sim \mathrm{Bin}(n, p)$.

Note that in this notation, $\mathrm{Bern}(p) \equiv \mathrm{Binom}(1, p)$.

Probabilities for Bin(n, p)

The probability that stock market falls on any given day is 0.3. Assume that the market falls or rises independently of the previous day's performance. What is the probability that in a week of six trading days, the market fall in two of them?

We can list the sample space here. Let F denote that the market falls and N that it does not. Then the sample space is NNNNNN, FNNNNN, NFNNNN, NNFNNN, NNNFNN, NNNNFN, NNNNNF, FFNNNN, FNFNNN, FNNFNN, ...

Now by independence,

$$P(FNNNNN) = 0.3(0.7)^5$$

Observations

The following observations can be made.

1. The two Falls can occur on any two of the six days.
2. The probability of any of the sequences containing two Falls is the same, that is, $0.3^2\ 0.7^4$.
3. If we can COUNT the number of ways of getting two falls out of six, then we can find the probability of two Falls in six days by multiplying this number with the probability $0.3^2\ 0.7^4$.

COUNTING TECHNIQUES — Combinatorics

In effect in the above example we need to select the two places that falls can occur out of six. In general, the number of ways of selecting r things out of n is given by

$$C_r^n = \binom{n}{r} = \frac{n!}{r!\,(n-r)!},$$

where

$$n! = \text{n factorial } = n(n-1)(n-2)\dots(2)(1).$$

For this notation to make sense, we define $0! = 1$. For example,

$$\binom{n}{n} = \frac{n!}{n!\,\underbrace{(n-n)!}_{=0!=1}} = 1, \binom{n}{1} = \frac{n!}{1!(n-1)!} = \frac{n(n-1)!}{1!(n-1)!} = n,$$

$$\binom{6}{2} = \frac{6!}{\underbrace{2!4!}_{Add\ \ to\ 6}} = \frac{6.5.4!}{2!4!} = 15, \binom{10}{2} = \frac{10!}{2!8!} = \frac{10.9.8!}{\underbrace{2!8!}_{Add\ \ to\ 10}} = 45.$$

Example

(a) In how many ways can two stocks be selected for purchase out of 10?

$$\binom{10}{2} = \frac{10!}{2!8!} = \frac{10.9.8!}{2!8!} = 45.$$

(b) A part-time worker works only two day a week. In how many possible ways can he select his weekly roster? Assume a five day week.

$$\binom{5}{2} = \frac{5!}{2!3!} = \frac{5.4.3!}{2!3!} = 10.$$

Example (ctd)

(c) In how many different ways can two lab mice be selected for an experiment?

$$\binom{6}{2} = \frac{6!}{2!4!} = \frac{6.5.4!}{2!4!} = 15.$$

Note On most calculators you can quickly compute combinations using the button

$$\binom{n}{r} \qquad \text{or} \qquad {}^{n}C_{r}$$

Probabilities for $X \sim \mathrm{Binom}(n, p)$

First note that $P(X = k) = 0$ for $k < 0$ and $k > n$. Next, $P(X = k)$ for $k = 0, 1, \ldots, n$ indicates that there are k successes and the *remaining* $n - k$ *trials are failures*. Thus the probability mass function for $X \sim \mathrm{Bin}(n, p)$ is

$$p_X(k) = \boxed{P(X = k) = \binom{n}{k}\, p^k\, (1 - p)^{n-k},} \quad k = 0, 1, 2, \ldots, n$$

where

- $$\binom{n}{r} = \frac{n!}{r!\,(n - r)!}$$ is the number of ways of obtaining k successes from n trials;
- p^k is the probability of the k successes;
- $(1 - p)^{n-k}$ is the probability of the $n - k$ failures.

The cumulative distribution function

The cdf for $X \sim \mathrm{Bin}(n, p)$ is

$$\begin{aligned} F_X(k) &= P(X \leq k) \\ &= P(X = 0) + P(X = 1) + \ldots + P(X = k) \\ &= \sum_{x=0}^{k} \binom{n}{x} \; p^x \; (1 - p)^{n-x}, \quad k = 0, 1, 2, \ldots, n. \end{aligned}$$

Note

Probabilities for the binomial distribution can be obtained:

- using the formula — just for fun!;
- using software such as `R` — this is what we will use.

Example

The probability a child from heterogygote parents is homogygote recessive aa for cystic fibrosis is $\boxed{0.25.}$ Assume that children from the same parents inherit this condition independently.

(a) What is the probability that in a family of of six children, two of them are aa?

Let the rv X denote the number of children who are aa out of six. Then $X \sim \mathrm{Binom}(\underline{6, 0.25})$, and

$$\underline{P(X=2)} = \binom{6}{2}\, 0.25^2\, 0.75^4 = 0.2966.$$

(b) What is the probability that at least one child is aa?

Now we need

$$P(X \geq 1) = 1 - P(X=0) = 1 - \binom{6}{0}\, 0.25^0\, 0.75^6$$

$$= 1 - 0.1780 = 0.8220.$$

Example

The probability that an pumpkin blossom produces fruit is 0.2. Assume that blossoms produce fruit independently of each other. The pumpkin vine has 20 blossoms. What is the probability that these blossoms produce

(a) no pumpkins,

(b) exactly 5 pumpkins.

Solution

Let the rv X denote the number of pumpkins produced by the 20 blossoms. Then $X \sim \mathrm{Binom}(20, 0.2)$.

(a) $P(X = 0) = \binom{20}{0}\ (0.2)^0\ (0.8)^{20} = 0.0115$. This answer can also be obtained from `R`.

(b) $P(X = 5) = \binom{20}{5}\ (0.2)^5\ (0.8)^{15} = 0.1746$. Compare this with the answer from `R`.

Example (ctd)

(c) at least 10 pumpkins

(d) at most 5 pumpkins

Solution

(c)

(d)

3.6 MEAN AND VARIANCE OF $\mathrm{Bin}(n, p)$

$$X \sim \mathrm{Bin}(n, p)$$

1. $E(X) = \underline{np}$

2. $\mathrm{Var}(X) = np(1 - p)$

3. $\sigma_X = \sqrt{np(1 - p)}$

Proof

For the proof we express the binomial distribution as a sum of n individual $\mathrm{Bern}(p)$ trials. Let the rv X denote the number of successes in n independent and identical Bernoulli trials. Now let $Y_1, Y_2, \ldots, Y_n$ represent the number of successes in the individual trials. Then for $i = 1, 2, \ldots, n$,

$$Y_i = \begin{cases} 1, & \text{if } Y_i \text{ is a success} \\ 0, & \text{otherwise.} \end{cases}$$

Now $Y_i \sim \mathrm{Bern}(p)$, are independent, and $\mathbb{E}(Y_i) = p, \mathrm{Var(Y_i) = p(1-p)}$. Further,

$$X = Y_1 + Y_2 + \ldots + Y_n,$$

so

$$\begin{aligned} \mathbb{E}(X) &= \mathbb{E}(Y_1 + Y_2 + \ldots + Y_n) \\ &= \mathbb{E}(Y_1) + \mathbb{E}(Y_2) + \ldots + \mathbb{E}(Y_n) \\ &= p + p + \ldots + p = np. \end{aligned}$$

Proof (ctd)

$$\begin{aligned}
\mathrm{Var}(\mathrm{X}) &= \mathrm{Var}(\mathrm{Y}_1 + \mathrm{Y}_2 + \ldots + \mathrm{Y}_\mathrm{n}) \\
&= \mathrm{Var}(\mathrm{Y}_1) + \mathrm{Var}(\mathrm{Y}_2) + \ldots + \mathrm{Var}(\mathrm{Y}_\mathrm{n}) \text{ (by independence)} \\
&= \underbrace{p(1-p) + p(1-p) + \ldots + p(1-p)}_{n \text{ times}} \\
&= np(1-p).
\end{aligned}$$

Example

A company employs 5 salespersons, each of whom makes a sale with probability 0.6. In a particular week each salesperson visits 20 clients.

(a) What is the probability that a salesperson make more than 15 sales?

Let the rv X denote the number of sales made by a salesperson. Then $X \sim \mathrm{Binom}(20, 0.6)$, so

$$\begin{aligned} P(X > 15) &= 1 - P(X \leq 15) \\ &= 1 - 0.9490 \\ &= 0.0510. \end{aligned}$$

Example 4.13 (ctd)

(b) What is probability that at least one salesperson makes more than 15 sales?

Let the rv Y denote the *number of salespersons* who make more than 15 sales. The $Y \sim \mathrm{Binom}(5, 0.0510)$, and

$$
\begin{aligned}
P(Y \geq 1) &= 1 - P(Y = 0) \\
&= 1 - \binom{5}{0} (0.0510)^0 \, (0.9490)^5 \\
&= 1 - 0.7697 \\
&= 0.2303.
\end{aligned}
$$

Example (ctd)

(c) What is the expected total number of sales for that week?

3.7 The Poisson distribution

- Models the number of occurrences of a phenomenon in a fixed interval or fixed time period or fixed area or fixed volume.
- This is a counting process, as is the binomial distribution.
- Examples are:
  - The number of telephone calls per hour at a business.
  - The number of cars queued at a traffic light in 1 minute.
  - The number of arrivals at a toll bridge per minute.
  - The number or times a printer breaks down in a month.
  - The number of paint spots on a new car.
  - The number of sewing flaws per pair of jeans during production.

Assumptions of the Poisson distribution

1. The occurrences are independent of each other.
2. Two occurrences cannot happen at the same location.
3. The mean number of occurrences in a specified volume is fixed.

Comparison with Bin(n, p)

1. Binomial distribution has a fixed number of trials. **Poisson does not.**
2. Binomial distribution has two parameters, n and p. **Poisson has only one: the mean number of occurrences.**
3. Binomial has two possible outcomes, Success and Failure, at each trial. **Poisson does not.**
4. Binomial random variable takes the values $0, 1, \ldots, n$. **Poisson takes values** $0, 1, 2, \ldots$**.**

Probability mass function

Let $X \sim \mathrm{Poi}(\lambda)$, where λ is the mean number of occurrences in a fixed volume. Then

$$p_X(k) = P(X = k) = \frac{e^{-\lambda}\ \lambda^k}{k!}, \quad k = 0, 1, 2, \ldots$$

The cumulative distribution function of X is given by

$$P(X \leq k) = F_X(k) = \sum_{x=0}^{k} \frac{e^{-\lambda}\ \lambda^x}{x!}, \quad k = 0, 1, 2, \ldots$$

that is, simply add up the probabilities up to and including k.

Exercise Do the probabilities add up to 1?

Example

A real estate agency sells 1.5 houses per day on average. Assume the sales are Poisson distributed.

(a) What is the probability of selling exactly 2 houses on a day?

(b) What is the probability of selling no houses on a day?

Solution

(a) Let the rv X denote the number of houses sold in a day. then $X \sim \mathrm{Pois}(1.5)$.

$$P(X = 2) = \frac{e^{-1.5}(1.5)^2}{2!} = 0.2510$$

(b)

$$P(X = 0) = \frac{e^{-1.5}(1.5)^0}{0!} = 0.2231$$

Example (ctd)

(c) What is the probability of selling at least 5 houses on a day?

(d) What is the probability of selling more than 5 houses in a day?

Solution

(c)

(d)

Example (ctd)

(e) What is the probability of selling exactly 5 houses in two days?
Solution Let the rv Y denote the number of houses sold in two days. Then

3.8 MEAN AND VARIANCE OF Poi(λ)

$$X \sim \mathrm{Poi}(\lambda)$$

1. $E(X) = \lambda$

2. $\mathrm{Var}(X) = \lambda$

3. $\sigma_X = \sqrt{\lambda}$