

MATH1905 Statistics

Lecture 9

Lecturer: Marc Raimondo
Carlaw 817

Lectures: Mon.11am (Chem LT1), Tue.8am (Chem LT1)

Student consultations: Tuesday 2-3pm

General information, tutes, solutions etc...

<http://www.maths.usyd.edu.au/u/UG/JM/MATH1905/>

or

First Year Office (FYO), Carlaw 520.

Random variables These are random quantities associated with the outcome of a random experiment. For instance if we toss a coin, we can let the rv X take the value 1 if we get a head and 0 if we get a tail. The *probability distribution* of X is $\{p_0, p_1\}$ with $p_i = \mathbf{P}(X = i)$, $i = 0, 1$ so that p_0 is the probability of getting a tail and $p_1 = 1 - p_0$ is the probability of getting a head. X is an *integer-valued* rv. Other examples:

- Throw a fair die (once)

$X : \Omega \rightarrow \{1, 2, 3, \dots, 5, 6\}$ here $\Omega = \{1, 2, 3, \dots, 5, 6\}$;

$X(w) = w$ and $p_i = P(X = i) = \frac{1}{6}$, $i = 1, 2, \dots, 6$

- Loaded die: $p_1 = \frac{1}{9}$, $p_2 = \frac{2}{9}$, $p_3 = \frac{1}{9}$, \dots , $p_6 = \frac{2}{9}$

1. The binomial variable This is based on a random experiment consisting of n independent trials (n a fixed number), for each of which there are two possible outcomes which we will denote as success or failure. The associated binomial rv, X , denotes the number of successes; X can take any of the values $0, 1, \dots, n$. If the probability of a success is p at each trial then the probability distribution $\{p_0, p_1, \dots, p_n\}$ satisfies

$$p_i = \mathbf{P}(X = i) = \binom{n}{i} p^i (1 - p)^{n-i}, \quad i = 0, 1, \dots, n$$

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

Sketch of proof

Denote $A_i = \text{Head at } i^{\text{th}} \text{ throw}$.

$$\begin{aligned} \{X = i\} &= \{A_1 \cap A_2 \cap \dots \cap A_i \cap A_{i+1}^c \cap A_{i+2}^c \cap \dots \cap A_n^c\} \cup \\ &\{A_1^c \cap A_2 \cap \dots \cap A_i \cap A_{i+1} \cap A_{i+2}^c \cap \dots \cap A_n^c\} \cup \dots \\ &\{A_1^c \cap A_2^c \cap \dots \cap A_{n-i}^c \cap A_{n-i+1} \cap A_{n-i+2} \dots \cap A_n\} \end{aligned}$$

all m.e. events with probability: $p^i \times (1 - p)^{n-i}$

How many ways can we choose i -items (numbers) among n ? $\binom{n}{i}$

So that $P(X = i) = \binom{n}{i} p^i (1 - p)^{n-i}$, $i = 0, 1, \dots, n$

Note that we have the binomial formula:

$$\sum_{i=0}^n \binom{n}{i} p^i (1 - p)^{n-i} = (p + (1 - p))^n = 1.$$

Example Roll a fair die 10 times. Let X be the number of threes obtained. Then $X \sim \mathcal{B}(10, 1/6)$; that is

$$p_i = \mathbf{P}(X = i) = \binom{10}{i} \frac{5^{10-i}}{6^{10}}, \quad i = 0, 1, \dots, 10$$

To 5 d.p., the distribution is

i	0	1	2	3	4	5
p_i	0.16151	0.32301	0.29071	0.15505	0.05427	0.01302
i	6	7	8	9	10	
p_i	0.00217	0.00025	0.00002	0.00000	0.00000	

The geometric variable We now suppose we have an infinite sequence of independent trials, each of which gives a success with probability p and failure with probability $q = 1 - p$. The geometric variable X counts the number of failures before the first success. The probability distribution is $\{p_0, p_1, \dots\}$ with

$$p_i = q^i p, \quad i = 0, 1, 2, \dots$$

$$p_0 = P(A_1), \quad p_1 = P(A_1^c \cap A_2) = (1 - p)p;$$

$$p_2 = P(A_1^c \cap A_2^c \cap A_3) = (1 - p)^2 p, \quad \dots$$

Note that there is no upper limit; p_i is positive for every integer $i \geq 0$. However,

$$\sum_{i \geq 0} p_i = \sum_{i \geq 0} q^i p = p \sum_{i \geq 0} q^i = p \frac{1}{1 - q} = 1.$$

Example A fair die is thrown repeatedly until it shows a six. What is the probability that more than 7 throws are required?

Let X denotes the number of failures before 1st success. X follows a geometric distribution with $p = 1/6$

$$\begin{aligned} P(X \geq 7) &= \sum_{i \geq 7} (5/6)^i (1/6) = (5/6)^7 (1/6) \sum_{i \geq 0} (5/6)^i \\ &= (5/6)^7 (1/6) \frac{1}{1-5/6} = (5/6)^7 \end{aligned}$$

Is it more likely that an odd number of throws is required or an even number?

O = Odd number of throws = Even...failures

$$P(X = 2k) = q^{2k}p, \quad k = 0, 1, 2, \dots$$

E = Even number of throws = Odd...failures

$$P(X = 2k + 1) = q^{2k+1}p, \quad k = 0, 1, 2, \dots$$

m.e. events for different values of k so

$$P(O) = \sum_{k \geq 0} pq^{2k} = p \frac{1}{1-q^2}$$

$$P(E) = \sum_{k \geq 0} pq^{2k+1} = pq \sum_{k \geq 0} q^{2k} = pq \frac{1}{1-q^2}$$

so we have $P(O) > P(E)$.

The Poisson variable This variable is used to model the frequency of 'rare events' such as the number of misprints in a book or the number of accidents per week at an intersection. The probability distribution is $\{p_i\}$, where

$$P(X = i) = p_i = \frac{\lambda^i e^{-\lambda}}{i!}, \quad i = 0, 1, 2, \dots$$

(again, there is no upper limit on i). The probability distribution depends on the single parameter λ , which can take any positive value.

$$\sum_{i \geq 0} p_i = e^{-\lambda} \sum_{i \geq 0} \frac{\lambda^i}{i!} = e^{-\lambda} e^{\lambda} = 1.$$

Example: modelling radioactive decay Dataset 2 (P&Q pp. 131–132) gives the number of α -particles released in each of 2608 intervals of 7.5 sec.

i	0	1	2	3	4	5	6	7	8	9	10
f_i	57	203	383	525	532	408	273	139	45	27	16

The mean number of particles released per interval is **3.8673** to 4 d.p. The probabilities p_i for a Poisson variable with $\lambda = 3.8673$ are (to 3 d.p.) $p_0 = 0.021$, $p_1 = 0.081$, $p_2 = 0.156$, $p_3 = 0.202$, $p_4 = 0.195$, $p_5 = 0.151$, $p_6 = 0.097$, $p_7 = 0.054$, $p_8 = 0.026$, $p_9 = 0.011$ and $p_{10} + p_{11} + \dots = 0.006$. Compare the observed frequencies f_i and the expected numbers under the Poisson model, i.e. $E_i = 2608p_i$, $i = 0, \dots, 9$; $E_{10} = 2608(p_{10} + p_{11} + \dots)$:

	0	1	2	3	4	5	6	7	8	9	≥ 10
f_i	57	203	383	525	532	408	273	139	45	27	16
E_i	55	211	408	526	508	393	253	140	68	29	17

Poisson approximation to Binomial. If $X \sim \mathcal{B}(n, p)$ with n large, p small and np moderate, then letting Y be a Poisson variable with $\lambda = np$: $\mathbf{P}(X = k) \approx \mathbf{P}(Y = k)$

Proof. $\lambda = np$ gives $p = \lambda/n$

$$P(X = k) = \frac{n!}{k!(n-k)!} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k}$$

$$P(X = k) = \left(\frac{\lambda}{k!}\right)^k \times \frac{n!}{(n-k)!n^k} \times \left(1 - \frac{\lambda}{n}\right)^n \times \left(1 - \frac{\lambda}{n}\right)^{-k}$$

$$P(X = k) \rightarrow_{n \rightarrow \infty} \left(\frac{\lambda}{k!}\right)^k \times 1 \times e^{-\lambda} \times 1, \text{ hence}$$

$$P(X = k) \approx e^{-\lambda} \left(\frac{\lambda}{k!}\right)^k$$

Example What is the probability that of 500 people, exactly k will have birthdays on New Year's day?

$X \sim \mathcal{B}(500, \frac{1}{365})$, Y is Poisson with

$$\lambda = \frac{500}{365} = 1.3699 \quad \text{to 4 d.p.}$$

k	0	1	2	3	4	5	6
b_k	0.2537	0.3484	0.2388	0.1089	0.0372	0.0101	0.0023
p_k	0.2541	0.3481	0.2385	0.1089	0.0373	0.0102	0.0023

Review problems. P&Q pp. 78–81: 6 (b), 9, 10, 12, 13, 20, 31, 33.

R commands

(discrete) probability distributions

Binomial $X \sim B(n, p), P(X = k) : \text{dbinom}(k, n, p)$

Geometric $X \sim \text{Geometric}(p), P(X = k) : \text{dgeom}(k, p)$

Poisson $X \sim \text{Poisson}(\lambda), P(X = k) : \text{dpois}(k, \lambda)$

(discrete) cumulative probability distributions

Binomial $X \sim B(n, p), P(X \leq k) : \text{pbinom}(k, n, p)$

Geometric $X \sim \text{Geometric}(p), P(X \leq k) : \text{pgeom}(k, p)$

Poisson $X \sim \text{Poisson}(\lambda), P(X \leq k) : \text{ppois}(k, \lambda)$

Lecture 10: Mean and variance of discrete rv's

The **mean** of an integer-valued rv X with distribution $p_i = \mathbf{P}(X = i)$, is $\mu = \sum ip_i$. Idea behind this definition: Suppose that a random experiment, whose outcome is X , is repeated a large number of times. In n replications of the experiment, the relative frequency, f_i/n , of trials which yield the value i will be approximately equal to p_i if n is large, where $p_i = \mathbf{P}(X = i)$. So the sample mean $\bar{x} = \sum if_i/n$ will be approximately equal to $\sum ip_i$ and we define the mean of the rv X to be this 'limiting value'. The mean μ of X is also called the *expected value* of X and written as $\mathbf{E}(X)$ instead of μ . (Other names used for μ include *population mean*, *mathematical expectation* and *first moment* of X .)

The *median* m of a random variable X is any number such that $P(X \leq m) \geq 1/2$ and $P(X \geq m) \geq 1/2$.

Examples X = Number on a die: $P(X = i) = 1/6$,
 $k = 1, 2, \dots, 6$.

$$E(X) = \sum_{i=1}^6 i \times P(X = i) = (1 + 2 + 3 + 4 + 5 + 6)/6 = (6 \times 7)/6 = 42/6 = 3.5$$

- $P(X = k) = pq^k$, $k = 0, 1, 2, \dots$

$$E(X) = \sum_{k \geq 0} k \times pq^k = pq \sum_{k \geq 0} k \times q^{k-1}$$

$$E(X) = pq \sum_{k \geq 0} \frac{\partial}{\partial q} (q^k) = pq \frac{\partial}{\partial q} \sum_{k \geq 0} (q^k) = pq \frac{\partial}{\partial q} \left(\frac{1}{1-q} \right)$$

$$E(X) = pq \frac{1}{(1-q)^2} = \frac{q}{p}$$

- $X \sim \mathcal{B}(n, p)$.

$$E(X) = \sum_{i=0}^n i \binom{n}{i} p^i (1-p)^{n-i}$$

$$E(X) = np \sum_{i=1}^n \frac{(n-1)!}{(i-1)!(n-i)!} p^{i-1} (1-p)^{n-i}$$

changing variable $j = i - 1, i = j + 1$

$$E(X) = np \sum_{j=0}^{n-1} \frac{(n-1)!}{j!(n-j-1)!} p^j (1-p)^{n-j-1}$$

$$E(X) = np(p + (1-p))^{n-1} = np$$

• $P(X = k) = c/k^2$, $k = 1, 2, \dots$ with $c = 6/\pi^2$.

$$E(X) = \sum_{k \geq 1} k \frac{c}{k^2} = c \sum_{k \geq 1} \frac{c}{k} = c \times \lim_{n \rightarrow \infty} \log(n) = +\infty$$

For this distribution the expectation is not defined...However,

$$P(X \leq 1) = \frac{6}{\pi^2} \geq 1/2 \text{ and}$$

$$P(X \geq 1) = 1 \geq 1/2 \text{ so the Median is } m = 1$$

Expected value of a function of a rv

If $Y = g(X)$ for some function g we define its expected value as

$$\mathbf{E}(Y) = \mathbf{E}(g(X)) = \sum_i g(i)P(X = i)$$

Example. This result is often used with $g(x) = (x - \mu)^2$,

$$g(x) = x(x - 1) \quad \text{or}$$

$$g(x) = ax + b.$$

Note: We can also get $\mathbf{E}(Y)$ from its own probability distribution:

$\mathbf{E}(Y) = \sum yP(Y = y)$ but this requires us to know the distribution of Y which may be hard to work out.

Variance. We define the variance of an rv with mean μ as $\text{var}(X) = \mathbf{E}((X - \mu)^2)$.

We sometimes write the variance as σ^2 .

That is, $\sigma^2 = \mathbf{E}(g(X))$ with $g(x) = (x - \mu)^2$.

Interpretation: σ^2 is the *expected squared deviation* of X from its mean.

$\text{var}(X) = \mathbf{E}(X^2) - (\mathbf{E}X)^2$ so for an integer-valued rv,
 $\text{var}(X) = \sum i^2 p_i - \mu^2$.

Proof. $\mathbf{E}(X - \mu)^2 = \mathbf{E}(X^2 + \mu^2 - 2\mu X)$

$\mathbf{E}(X - \mu)^2 = \mathbf{E}(X^2) + \mu^2 - 2\mu\mathbf{E}(X) = \mathbf{E}(X^2) - (\mathbf{E}(X))^2$

Example

- $X \sim \mathcal{B}(1, p)$.

$$\mu = E(X) = np = p \quad (n = 1)$$

$$E(X^2) = \sum_{i=0}^1 i^2 P(X = i) = 0^2(1 - p) + 1^2p = p$$

$$\text{var}(X) = E(X^2) - (EX)^2 = p - p^2 = p(1 - p)$$

A useful identity To get the variances of the binomial, geometric and Poisson variables the following identity is useful:

$$\mathit{var}(X) = \mathbf{E}(X(X - 1)) + \mathbf{E}X - (\mathbf{E}X)^2.$$

Example: $X \sim \mathcal{B}(n, p)$.

$$\mathbf{E}(X(X - 1)) = \sum_{i=0}^n i(i - 1) \binom{n}{i} p^i (1 - p)^{n-i}$$

$$\begin{aligned} \mathbf{E}(X(X - 1)) &= \\ p^2 n(n - 1) \times \sum_{i=0}^n \binom{n-2}{i-2} p^{i-2} (1 - p)^{n-i-2} &= p^2 n(n - 1) \times 1 \end{aligned}$$

$$\text{so } \mathit{var}(X) = p^2 n(n - 1) + np - (np)^2 = (np)^2 - np^2 + np - (np)^2$$

$$\mathit{var}(X) = np(1 - p)$$

Mean & variance of a linear function

If $Y = aX + b$ then $E(Y) = aE(X) + b$ and
 $var(Y) = a^2 var(X)$

Proof. $var(aX + b) = E[aX + b - E(aX + b)]^2$

$$var(aX + b) = E[aX + b - aE(X) + b]^2 = E[aX - aE(X)]^2$$

$$var(aX + b) = E[a(X - E(X))]^2 = a^2 E(X - E(X))^2$$

$$var(aX + b) = a^2 var(X)$$

Example $X \sim \mathcal{B}(n, p), Y = n - X,$

$$E(Y) = E(n - X) = n - E(X) = n - np = n(1 - p)$$

$$var(Y) = var(n - X) = var(X) = np(1 - p)$$

Chebyshev's inequality

If a rv X has mean μ and variance σ^2 , then for any positive number c , $P(|X - \mu| \geq c\sigma) \leq 1/c^2$

Proof $p_i = P(X = i)$

$$S = \{|X - \mu| \geq c\sigma\} = \{i : |i - \mu| \geq c\sigma\}$$

$$S = \{|X - \mu| \geq c\sigma\} = \{i : \frac{|i - \mu|}{c\sigma} \geq 1\}$$

$$P(|X - \mu| \geq c\sigma) = P(\cup_{i \in S} (X = i)) \stackrel{m.e.}{=} \sum_{i \in S} P(X = i)$$

$$\leq \sum_{i \in S} p_i \frac{|i - \mu|}{c\sigma} \leq \sum_{i \in S} p_i \left(\frac{|i - \mu|}{c\sigma}\right)^2 \leq \frac{1}{c^2 \sigma^2} \sum_i p_i (i - \mu)^2 = \frac{1}{c^2}$$

Note that we can re-write the inequality with $c = b/\sigma$

$$P(|X - \mu| \geq b) \leq \frac{\sigma^2}{b^2}$$

Chebyshev's inequality is of theoretical rather than practical importance i.e. (not sharp)**Examples** X = number scored in throwing a die. Mean and variance are $\mu = 3.5$ and $\sigma^2 = 35/12$.

$$P(|X - 3.5| \geq 3) = 0$$

the inequality gives $P(|X - 3.5| \geq 3) \leq (35/12)/9 \approx 0.32$

$$P(|X - 3.5| \geq 2) = P(X = 1) + P(X = 6) = 2/6 = 1/3$$

the inequality gives $P(|X - 3.5| \geq 2) \leq (35/12)/4 \approx 0.72$

Review problems P&Q pp. 80–81: 22 (a), 28, 32, 33.

Review problems P&Q pp. 80–81: 22 (a), 28, 32, 33.

R commands for generating samples of 'independent' discrete 'random' variables

Review problems P&Q pp. 80–81: 22 (a), 28, 32, 33.

R commands for generating samples of 'independent' discrete 'random' variables

Binomial $\mathbf{x}=(x_1, \dots, x_{n_1})$ from a $B(n_2, p)$ -distribution :
`rbinom(n_1, n_2, p)`

Geometric $\mathbf{x}=(x_1, \dots, x_n)$ from a $Geom(p)$ -distribution :
`rgeom(n, p)`

Poisson $\mathbf{x}=(x_1, \dots, x_n)$ from a $Pois(\lambda)$ -distribution :
`rpois(k, λ)`

Review problems P&Q pp. 80–81: 22 (a), 28, 32, 33.

R commands for generating samples of 'independent' discrete 'random' variables

Binomial $\mathbf{x}=(x_1, \dots, x_{n_1})$ from a $B(n_2, p)$ -distribution :
`rbinom(n_1, n_2, p)`

Geometric $\mathbf{x}=(x_1, \dots, x_n)$ from a $Geom(p)$ -distribution :
`rgeom(n, p)`

Poisson $\mathbf{x}=(x_1, \dots, x_n)$ from a $Pois(\lambda)$ -distribution :
`rpois(k, λ)`

R commands for cumulative sums

```
y1=cumsum(x)
```

```
n1=1:n
```

Illustration of statistical regularity: $E(X)$ is the long run behavior of the sample mean:

```
plot(n1,y1/n1)
```