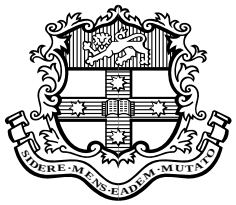


MATH 1015: Life Science Statistics

Lecture Pack for Chapter 2
Weeks 3-7.

Lecturer: Jennifer Chan
Room: Carslaw Room 817
Telephone: 9351 4873.

Text: Phipps, M. and Quine, M. (2001) *A Primer of Statistics* (4th Ed.)

6 Classical probability and counting

6.1 Review

• The mathematical models that underlie statistical analyses all use probability. First we will revise some basic results from the HSC course.

• Suppose we have a random experiment which has exactly c possible *mutually exclusive* (m.e.), *equally likely*, simple outcomes. We can assign a probability to an event A by counting the number of simple outcomes that correspond to A. If the count is a then

$$P(A) = \frac{a}{c}.$$

The set of all possible outcomes is called the **sample space**, Ω.

Examples:

1. Throw a fair six sided die. There are 6 possible outcomes.

$\Omega =$

If A corresponds to observing a multiple of 3 then, in set notation

$A =$

Prob(number is a multiple of 3) $= P(A) =$

2. Toss a fair coin.

$\Omega =$

If A corresponds to observing a head then, in set notation

$A =$

Prob(Head) $= P(A) =$

6.2 Counting Results

1. The number of *ordered* samples of size r we can draw *without replacement* from n objects is

$$n(n-1)\cdots(n-r+1) = \frac{n!}{(n-r)!}.$$

Recall $0! = 1$.

2. The number of samples of size r we can draw without replacement from n objects *where order is not important* is

$${}^nC_r = \left(\begin{array}{c} n \\ r \end{array} \right) = \frac{n!}{r!(n-r)!}.$$

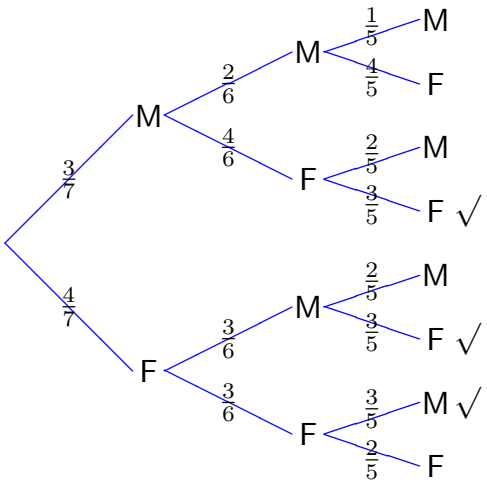
$$\left(\begin{array}{c} n \\ r \end{array} \right) = \frac{n(n-1)\cdots(n-r+1)}{1.2\ldots r} = \left(\begin{array}{c} n \\ n-r \end{array} \right).$$

Note $\left(\begin{array}{c} n \\ 0 \end{array} \right) = 1 = \left(\begin{array}{c} n \\ n \end{array} \right)$.

Sampling without replacement is important in *survey*. We need to determine how many different samples of a given size are possible. A *simple random sampling* procedure is one where each of the possible samples has the same chance of being selected. The possible samples form the list of simple outcomes.

Example 1. A tank contains 3 male and 4 female fish. Three fish are selected at random. What is the probability of getting 1 male and 2 female fish?

Method 1: Tree Diagram.



Prob. is $P(1 \text{ male and } 2 \text{ female}) = \dfrac{3(3 \cdot 4 \cdot 3)}{7 \cdot 6 \cdot 5} = \dfrac{18}{35} = 0.5143$

Method 2. Counting.

There are different samples of size 3 that can be drawn from a set of 7 objects.

There are possible samples of size 1 that can be drawn from the 3 male fish.

These are possible samples of size 2 that can be drawn from the 4 female fish.

Thus $P(1 \text{ male and } 2 \text{ female}) =$

1997 HSC 2 Unit Q9(a)

A bag contains 2 red balls, 1 black ball and 1 white ball. Andrew selects one ball from the bag and keeps it hidden. He then selects a second ball, also keeping it hidden.

(i) Draw a tree diagram to show all possible outcomes.

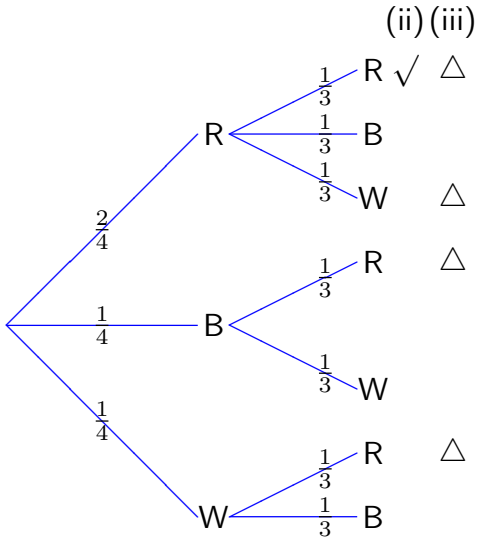
(ii) Find the probability that both the selected balls are red.

(iii) Find the probability that at least one of the selected balls is red.

(iv) Andrew drops one of the selected balls and we can see that it is red. What is the probability that the ball that is still hidden is also red?

Solution:

(i)



(ii) $P(\text{both red}) =$

(iii) $P(\text{at least one red}) =$

7 Frequentist approach to probability

7.1 Probability as long run relative frequency

In many situations we do not have a finite set of equally likely outcomes so the classical approach is not applicable. However the *relative frequency* of a given event does tend to stabilise as the sample size increases. This observable phenomenon is called *statistical regularity.*

We define the probability of an event A as the *long run relative frequency.*

Examples
1. Number of male births for every 100 female births for each of the years 1964-1973 from the Australian Yearbooks.

106.28	105.61	105.95	105.43	105.41
105.22	105.12	105.34	105.47	105.19

2. The age-standardised mortality rate from prostate cancer in NSW is 27.99 per 100,000 and this has been fairly constant over 1971 - 2001.

The frequentist approach also applies in classical situations like coin tossing.

Properties of relative frequencies can be formalised into the *axioms of probability.*

7.2 Axioms

1. For any event A, $P(A) \geq 0$.
2. For the complete sample space Ω, $P(\Omega) = 1$.
3. For two *mutually exclusive* events A and B the probability that A or B occurs is $P(A) + P(B)$.

It follows that for any event A,

$$0 \leq P(A) \leq 1.$$

Example:
Throw a single six sided die.

$$\Omega =$$

A : the number is a multiple of 3

$$A =$$

B : the number is less than 3

$$B =$$

A and B are mutually exclusive. $P(A$ or B occurs$) =$.

7.3 Set notation

• $\bigcup$ is read as 'or'

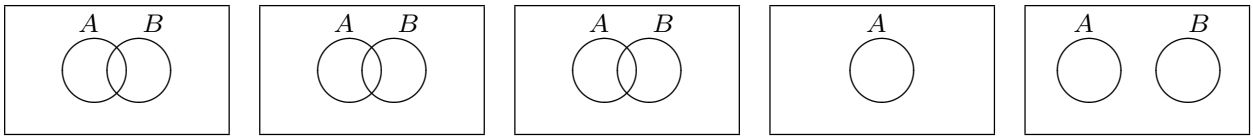
'$A \bigcup B$' is read 'A or B or both occur' (or 'at least one of A and B occur').

• $\bigcap$ is read as 'and'

'$A \bigcap B$' is read as 'A and B both occur'.

• A^c is read as 'A does not occur.'

• $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.



(a) Events A and B are dependent (b) Middle region is $A \cap B$ (c) Region of A or B is $A \cup B$ (d) Region outside A is A' (e) Mutually exclusive events

Example: A lotto type barrel contains 10 balls numbered $1, 2, \ldots, 10$. Three balls are drawn.

Solution:

(i) How many distinct samples can be drawn?

$n =$

(ii) Event A corresponds to all numbers drawn being less than 7.

$P(A) =$

A^c corresponds to at least one number drawn being 7 or more.

$P(A) =$

(iii) Event B corresponds to all numbers drawn being even.

$$P(B) =$$

$$P(A \bigcap B) =$$

$$P(A \bigcup B) = P(A) + P(B) - P(A \bigcap B)$$

$$=$$

7.4 Conditional Probability

In some situations extra information helps us refine the probability of another event occurring. Consider relative frequencies.

Example (Oldfield and Klauer (1980))
A sample of 249 Eskimos was taken and each person classified by blood group and tuberculosis(TB) status.

	O	**A**	**B**	**AB**	Total
TB	34	37	31	11	113
No TB	55	50	24	7	136
Total	89	87	55	18	249

The rel. frequency of people with TB is .

Of those with type **B** blood the rel. frequency of TB is .

If TB incidence did not vary with blood group we would expect the relative frequencies to be roughly the same.

7.5 Independence of events

For probabilities we say two events A and B are *independent* if

$$P(A|B) = P(A),$$

i.e. knowing that B has occurred does not affect the probability of A occurring.

$$P(A|B) = \frac{P(A \bigcap B)}{P(B)}.$$

Thus we have the multiplication rule

$$P(A \bigcap B) = P(A|B) \times P(B).$$

If A and B are independent then

$$P(A \bigcap B) = P(A) \times P(B).$$

Example: A bag contains 4 red and 6 blue balls. Two balls are drawn in sequence *without replacement*.

(i) What is the probability of a red ball on the first draw?

(ii) What is the probability of a red ball on the second draw?

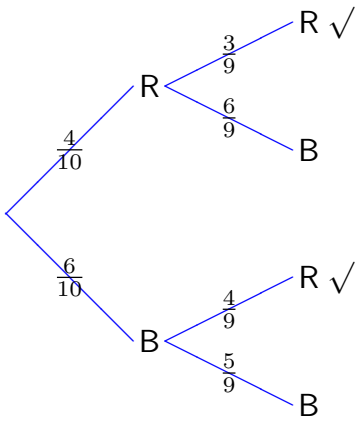
Solution:

(i)

$$P(\text{1st draw is red}) =$$

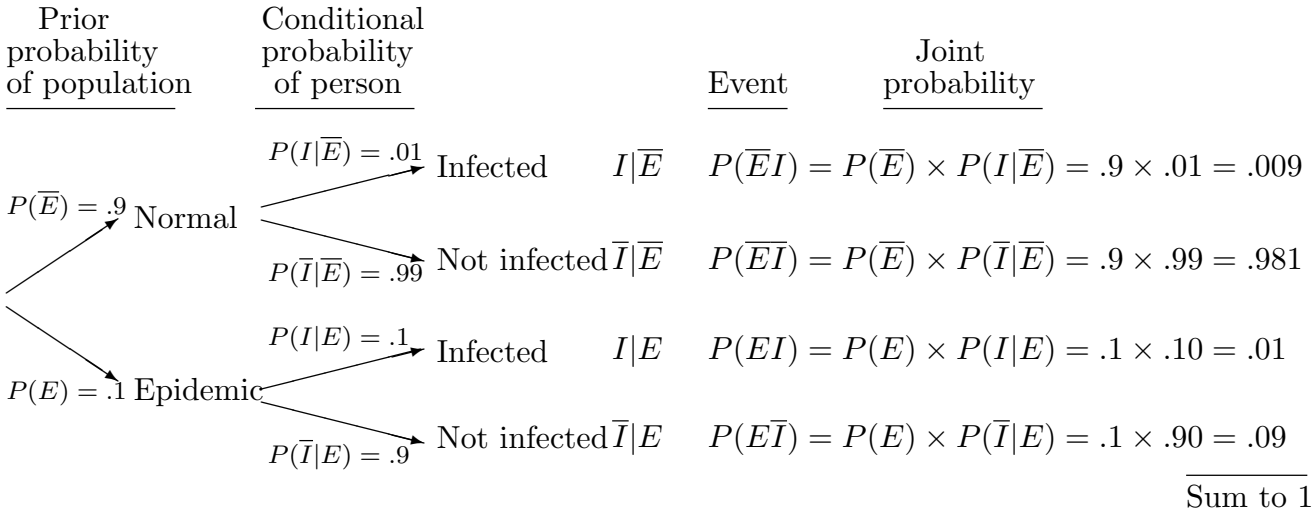
(ii)

$$\begin{aligned} & P(\text{2nd draw is red}) \\ = \; & P(\text{1st draw is red})P(\text{2nd draw is red}|\text{1st draw is red}) + \\ & P(\text{1st draw is blue})P(\text{2nd draw is red}|\text{1st draw is blue}) \\ = \; & \end{aligned}$$



Example: (Epidemic) Under normal condition, 1% of the population will be infected with an influenza epidemic and it becomes 10% when the epidemic prevails. The population is under normal condition 90% of time (*prior probability*). Find the probability of randomly selected person is infected by the epidermic and the probability that the population is under normal condition given that there is a person infected.

Solution:



$$\begin{aligned} P(I) &= P(\overline{E}I) + P(EI) = 0.009 + 0.010 = 0.019 \\ P(\overline{E}|I) &= \frac{P(\overline{E}I)}{P(I)} = \frac{P(\overline{E}I)}{P(\overline{E}I) + P(\overline{E}\overline{I})} = \frac{0.009}{0.019} = 0.47. \end{aligned}$$

Note:

1. Events $\overline{E}$ & I are *dependent* since $0.47 = P(\overline{E}|I) \neq P(\overline{E}) = 0.9$.
2. The prob. $P(I) = 0.019$ is a weighed average of two conditional prob. $P(I|E)$ and $P(I|\overline{E})$ weighed by the prob. $P(E)$ and $P(\overline{E})$ respectively.
3. Given that a person is infected, the prob. $P(\overline{E})$ drops from 0.9 to $P(\overline{E}|I) = 0.47$.

8 Integer value random Variable

8.1 Definition

With any experiment there can be associated a *random variable* (r.v.), X, which corresponds to a count or measurement of some type. For example, if 6 patients are treated with a new drug we can define the r.v. X to be the number cured. Here X is integer valued. It can take the values 0, 1, 2, 3, 4, 5, 6.

8.2 Binomial Variable

Suppose we have n *independent trials* where at each trial there are *two possible outcomes* (commonly labelled a 'success' (S) or 'failure' (F)) and the *probability of a 'success' is the same* at each trial.

Let X denote the number of successes (S) in n trials.

Let p be the probability of a S. We write

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

X can take the values $0, 1, 2, \ldots, n$.

$$P(X = 0) = P(F.F.F...F) = (1-p)^n$$

$$\begin{aligned} P(X = 1) &= P(1 \text{ S and } (n-1) \text{ F's}) \\ &= np(1-p)^{n-1}. \end{aligned}$$

Similarly

$$P(X = i) = \left(\begin{array}{c} n \\ i \end{array} \right) p^i (1-p)^{n-i}.$$

Example: When $n = 4$, the $2^4 = 16$ possible outcomes are

Outcomes and prob. for a binomial experiment with four trials

Outcome	x	Probability	Outcome	x	Probability
SSSS	4	p^4	FSSS	3	$p^3(1-p)$
SSSF	3	$p^3(1-p)$	FSSF	2	$p^2(1-p)^2$
SSFS	3	$p^3(1-p)$	FSFS	2	$p^2(1-p)^2$
SSFF	2	$p^2(1-p)^2$	FSFF	1	$p(1-p)^3$
SFSS	3	$p^3(1-p)$	FFSS	2	$p^2(1-p)^2$
SFSF	2	$p^2(1-p)^2$	FFSF	1	$p(1-p)^3$
SFFS	2	$p^2(1-p)^2$	FFFS	1	$p(1-p)^3$
SFFF	1	$p(1-p)^3$	FFFF	0	$(1-p)^4$

$$\begin{aligned}
P(X=4) &= p^4 &&= \binom{4}{4} p^4 q^{4-4} \\
P(X=3) &= 4p^3q &&= \binom{4}{3} p^3 q^{4-3} \\
P(X=2) &= 6p^2q^2 &&= \binom{4}{2} p^2 q^{4-2} \\
P(X=1) &= 4pq^3 &&= \binom{4}{1} p^1 q^{4-1} \\
P(X=0) &= q^4 &&= \binom{4}{0} p^0 q^{4-0}
\end{aligned}$$

Consider a particular sequence of x S and $n - x$ F,

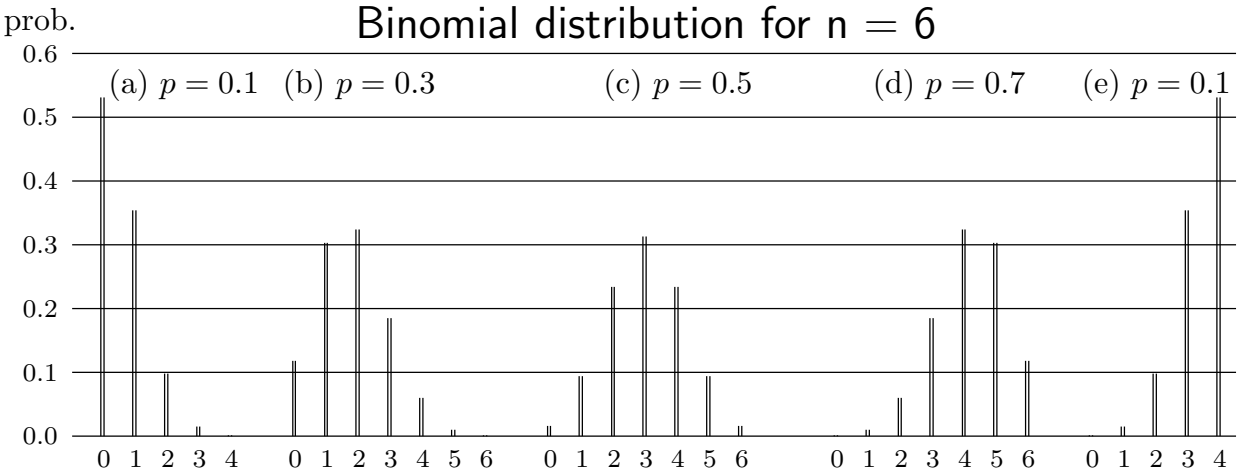
$$\underbrace{S\ S\ \cdots\ S}_{x\ S's}\ \underbrace{F\ F\ \cdots\ F}_{(n-x)\ F's}$$

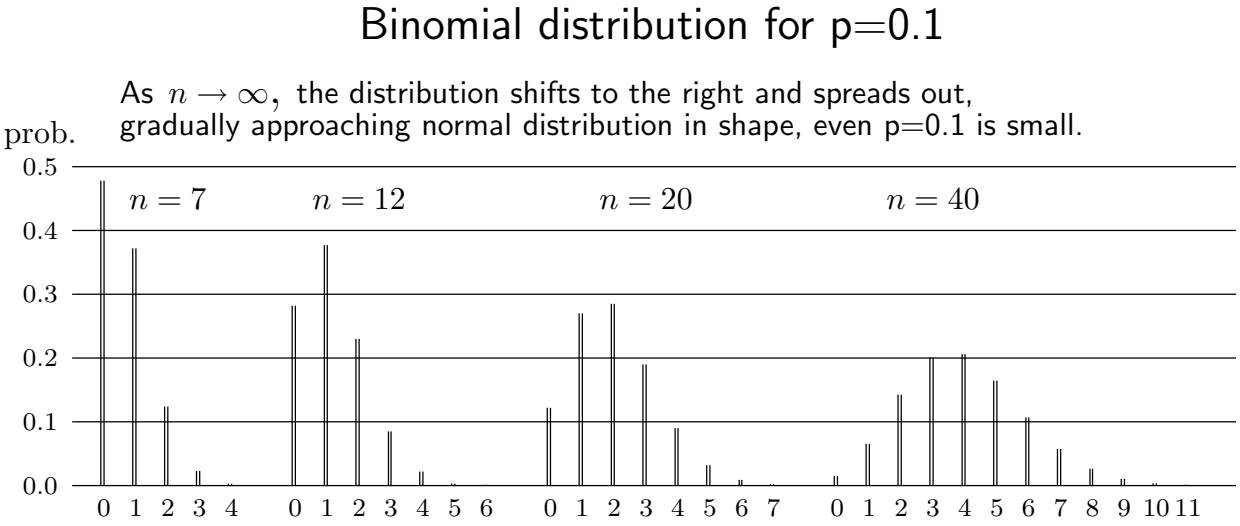
with a probability $p^x(1-p)^{n-x}$, the number of arrangements are $\binom{n}{x}$ (choose x positions from totally n positions for x S). For example, with $n = 4$ and $x = 3$, SSFS corresponds to selecting $\{1, 2, 4\}$ from $\{1, 2, 3, 4\}$.

The binomial probabilities

$$F(x) = P(X \le x) = P(X = 0) + P(X = 1) + .. + P(X = x).$$

are tabulated. When solving problems *always* write the probability required in terms of $\{X \leq r\}$.





8.3 How to recognise a binomial r.v.

- Is it a *count* over a *fixed* number of trials or repetitions?
- Are the trials (or repetitions) *independent*?
- Is the probability of the outcome of interest *constant* across trials?

If the answers are all 'Yes', the count is a binomial r.v..

8.4 R commands for distributions

• p for 'probability', the cumulative distribution function (cdf).

• q for 'quantile', the inverse of cdf.

• d for 'density', the density function (pmf or pdf).

• r for 'random', a random variable having the specified distribution.

• For binomial distribution, the commands are `pbinom`, `qbinom`, `dbinom`, `rbinom` respectively.

Example: (Defective items) Of a large number of mass-produced articles, 20% are defective. Writing X for the number of defective items in a random sample of 8, use the tables to find

(a) $P(X \leq 2)$,

(b) $P(X < 1)$,

(c) $P(X > 4)$,

(d) $P(X \geq 3)$,

(e) $P(X = 3)$,

(f) a such that $P(X \leq a) = 0.5033$.

Solution: We have $n =$ and $p =$.

In R:

```
> pbinom(2,8,0.2,lower.tail = TRUE, log.p = FALSE)
[1] 0.7969178
```

```
> pbinom(0,8,0.2,lower.tail = TRUE, log.p = FALSE)
[1] 0.1677722
> pbinom(4,8,0.2,lower.tail = FALSE, log.p = FALSE)
[1] 0.0104064
> pbinom(2,8,0.2,lower.tail = FALSE, log.p = FALSE)
[1] 0.2030822
> dbinom(3,8,0.2,log = FALSE)
[1] 0.1468006
> qbinom(0.5033,8,0.2,lower.tail = TRUE, log.p = FALSE)
[1] 1
```

By calculation:

(a) $P(X \leq 2) =$

$=$

$=$

(b) $P(X < 1) =$

(c) $P(X > 4) =$

$=$

$=$

(d) $P(X \geq 3) =$

Example: (Soft drinks) Two rival soft drinks, C and P taste the same. In a blindfold test, 12 people are asked (independently) to state their preference for one or the other. What is the probability that the majority prefer P?

Solution: Let X denote the number of people who prefer P. We have $n =$ and $p =$.

$$P(X \geq 7)$$

$$=$$

$$=$$

$$=$$

$$=$$

Example: (Computer disk) It is known that disks produced by a company are defective with probability 0.01 independently of each other. Disks are sold in packs of 10. A money back guarantee is offered if a pack contains more than 1 defective disk. What proportion of sales result in the customer getting their money back?

Solution: Let X denote the number of defective disks. We have $n =$ and $p =$.

$$P(X > 1) =$$

$$=$$

$$=$$

Example: (Question 9, book P.78) A person estimates that the chance of a direct hit on a target is 0.3. He fires 10 arrows. Assuming that his shots are independent, what is the probability that he scores

(a) no direct hit,

(b) just one direct hit,

(c) at least three direct hits?

Solution: Let X denote the number of direct hit. We have $n =$ and $p =$.

(a) $P(X = 0) =$

(b) $P(X = 1) =$

(c) $P(X \geq 3) =$

$=$

$=$

Example: (Fish) In a small pond there are 50 fish, 20 of which have been tagged. Seven fish are caught and X represents the number of tagged fish in the catch. Assume each fish in the pond has the same chance of being caught. Is X binomial

(a) if each fish is *returned* before the next catch?

(b) if the fish are *not returned* once they are caught?

Solution:

(a) Yes. Each fish is caught from the same population independently of the previous catch.

(b) No. Each fish is caught from a different population depending on all fish that are caught previously.

8.5 Geometric variable

Another random variable, Y, is one that counts the number of failures before the first success in a sequence of independent trials. Y is called a *geometric* r.v. which follow $\mathcal{G}(p)$ if

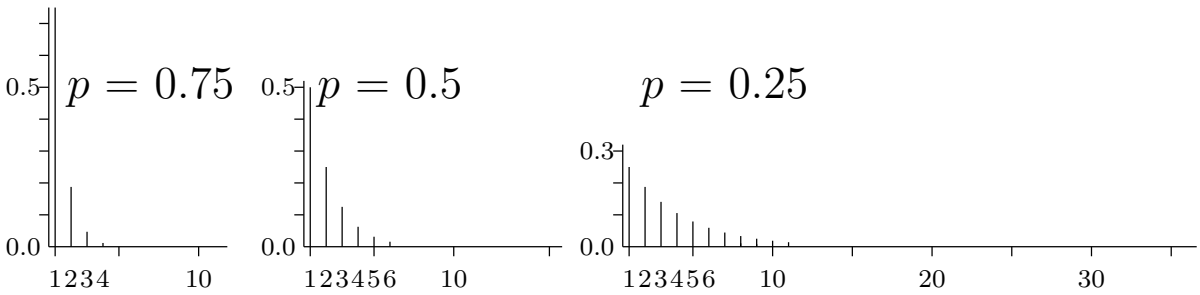
$$P(Y = i) = q^i p, \quad i = 0, 1, 2, 3, \ldots,$$

where $q = 1 - p$.

Note the probabilities sum to 1.

$$\begin{aligned} P(Y = 0) + P(Y = 1) + \ldots &= p + qp + q^2 p + \ldots \\ &= p(1 + q + q^2 + \ldots) \\ &= p \times \frac{1}{1 - q} \\ &= 1. \end{aligned}$$

Different p give different distributions.



9 Mean of random variables

9.1 Definition

Given a data set we can construct a frequency table:

Observation	x_1	x_2	$\cdot$	$\cdot$	x_k
Frequency	f_1	f_2	$\cdot$	$\cdot$	f_k

1. The *cumulative distribution function (cdf)* of X is

$$F(x) = P(X \leq x).$$

   See the DMS Statistical Tables for the cdf $F(x)$ for a binomial r.v..

2. For a random variable X taking values 0,1,2,.. with the *probability mass function (pmf)*

$$P(X = i) = p_i \quad i = 0, 1, 2, ..,$$

   the *mean* or *expected value* of X is

$$E(X) = \mu = \sum_i ip_i.$$

   $E(X)$ is the *long run average* of observations of X.

3. For any function $g(X)$ we define the expected value

$$E[g(X)] = \sum_i g(i)p_i \stackrel{\text{or}}{=} \sum_i g(x_i)p(x_i).$$

   In particular, if $g(x_i) = x_i^2$, $E(X^2) = \sum_i x_i^2 p(x_i)$.

4. For constants a and b,

$$E(aX + b) = aE(X) + b.$$

5. If $X \sim \mathcal{B}(n,p)$ then
$$E(X) = np.$$

6. If $X \sim \mathcal{G}(p)$ then
$$E(X) = \frac{1}{p}.$$

7. The *sample mean* is
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} x_i f_i = \sum_{i=1}^{k} x_i (f_i/n)$$
where f_i/n is the relative frequency of x_i.

Example: Suppose X takes the values 2, 4 and 6 with probabilities

i	2	4	6
p_i	0.1	0.3	0.6

Solution:

$\mu = E(X) =$

$E(X^2) =$

Note $E(X^2) \neq [E(X)]^2$.

Example: (Sale of computer) A computer store has purchased 3 computers at \$5000 each and sell them for \$10000 a piece. The manufacturer agreed to repurchase any computers still unsold after a specified period at \$2000 a piece. Let X be the number of computers sold and suppose that $p(0) = 0.1$, $p(1) = 0.2$, $p(2) = 0.3$ and $p(3) = 0.4$. What is the expected profit?

Solution: With $h(X)$ denoting the profit associated with selling X units, we have

$$\begin{aligned} E(X) &= \sum_i x_i p(x_i) = \\ h(X) &= \text{revenue} - \text{cost} \\ &= \\ E[h(X)] &= \sum_{x=0}^{3} h(x)p(x) \\ &= \\ &\underline{\text{or}} \\ &= \\ &= \end{aligned}$$

Example: (Expect winning) In the American version of roulette the wheel has 18 red numbers, 18 black numbers as well as '0' and '00'.

(i) A player bets \$1 on 'red' to win \$1. What are his expected winnings?

(ii) If a player places a \$1 bet on a single number and wins then he gains \$35. What is the player's expected winnings if he plays this strategy?

Solution:
(i) Let X denote the winnings if the player bets on red.

$X = 1$ with probability

$X = -1$ with probability

Thus $E(X) =$

(ii) Let Y denote the winnings if the player bets on a single number.

$Y = 35$ with probability

$Y = -1$ with probability

$E(Y) =$

Example: (Multiple choice) A multiple choice quiz has 12 questions and each question has 5 possible answers. A student decides to answer the questions by selecting an answer at random.

(a) What is the expected number of correct responses?

(b) What is the probability that the student scores more than 5 on the quiz?

(c) If the student scores 4 for a correct answer but -1 for a wrong response, what is his expected score?

Solution: Let X be the number of correct answer. Then $X \sim \mathcal{B}(12, \frac{1}{5})$.

(a) The expected number of correct responses is

$$E(X) =$$

(b) The probability that the student scores more than 5 on the quiz is

$$P(X > 5) =$$

(c) His expected score

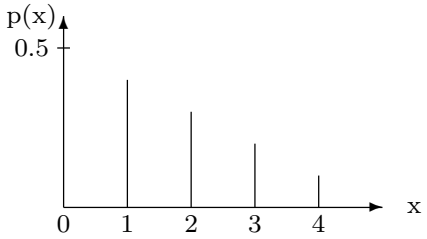
$$E(\text{score}) =$$

Example: (3 tosses of a fair coin) Each equally likely outcome has probability = 1/8. Hence

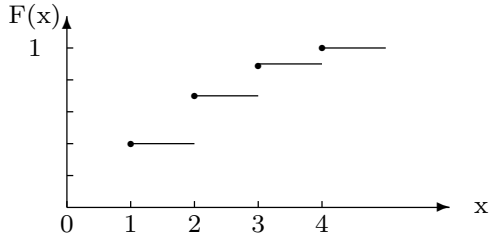
No. of heads, X	0	1	2	3
Probability, $p(x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$
c.d.f., $F(x)$				

Example: (Blood donor) Of the potential blood donors, A, B, C, D and E, only A and B have type O+ blood. Five blood samples, one from each individual, will be typed in random order until an O+ individual is identified.
Let X be the number of typings necessary to identify an O+ individual. Then the p.m.f. of X is

$$\begin{aligned}
P(X=1) &= P(\text{A or B 1st}) = \frac{2}{5} = 0.4 \\
P(X=2) &= P(\text{C, D or E 1st, then A or B}) = \frac{3}{5}\cdot\frac{2}{4} = 0.3 \\
P(X=3) &= P(\text{C, D or E 1st \& 2nd, then A or B}) = \frac{3}{5}\cdot\frac{2}{4}\cdot\frac{2}{3} = 0.2 \\
P(X=4) &= P(\text{C, D \& E all 1st}) = \frac{3}{5}\cdot\frac{2}{4}\cdot\frac{1}{3} = 0.1 \\
P(X=x) &= 0 \qquad\qquad \text{if } x \neq 1,2,3,4.
\end{aligned}$$



A graph of pmf



A graph of the cdf

The c.d.f. for each value of X is given below.

$$\begin{aligned}
F(1) &= P(X \leq 1) = 0.4 \\
F(2) &= P(X \leq 2) = 0.4 + 0.3 = 0.7 \\
F(3) &= P(X \leq 3) = 0.4 + 0.3 + 0.2 = 0.9 \\
F(4) &= P(X \leq 4) = 0.4 + 0.3 + 0.2 + 0.1 = 1
\end{aligned}$$

Note that $F(x) = 0$ for $x < 1$ and $F(x) = 1$ for $x > 4$ in this example.

10 Variance of random variables

10.1 Variance and standard deviation

We define the *variance* of X as a measure of *spread* by

$$\sigma^2 = E(X-\mu)^2,$$

where $\mu = E(X)$. Note that $\sigma^2 = E(X^2) - \mu^2$.

For integer valued random variables

$$\sigma^2 = \Sigma_i (i-\mu)^2 P(X=i).$$

This is like the large sample limit of a sample variance. Given a frequency table:

Observation	x_1	x_2	$\cdot$	$\cdot$	x_k
Frequency	f_1	f_2	$\cdot$	$\cdot$	f_k

The *sample variance* is

$$s^2 = \frac{1}{n-1}\Sigma_i (x_i - \bar{x})^2 f_i.$$

The *standard deviation* of X is

$$\sigma = \sqrt{var(X)}$$

and the *sample standard deviation* is

$$s = \sqrt{s^2}$$

.

Example: (Die) If X represents the number showing when a die is tossed, find the standard deviation of X.

Note $p_i = \dfrac{1}{6}$ for $i = 1, 2, 3, 4, 5, 6$.

$$E(X) = \mu = \sum_i i p_i =$$

$$E(X^2) = \sum_i i^2 p_i =$$

$$\sigma^2 = E(X^2) - \mu^2 =$$

The standard deviation is $\sigma =$.

10.2 Rule

For any constants a and b

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

• If $X \sim \mathcal{B}(n, p)$ then we have that $E(X) = np$ and we can show that

$$var(X) = \sigma^2 = np(1 - p).$$

• If $p = 0$ or 1 then the variance is 0.

• The variance is a maximum when $p = 0.5$. In this case $\sigma^2 = \frac{n}{4}$.

Example: (Sale of computer)

$$\begin{aligned} \mathrm{E}(X^2) &= \\ \mathrm{Var}(X) &= \mathrm{E}(X^2) - [\mathrm{E}(X)]^2 = \\ \mathrm{Var}(8000X - 9000) &= \\ \mathrm{SD}(8000X - 9000) &= \end{aligned}$$

Example: (Drugstore) A small drugstore orders boxes of pain-killing drugs each week. Let X be the demand for the number of boxes of drugs, with p.m.f.

x	1	2	3	4	5	6
$p(x)$	1/15	2 /15	3/15	4/15	3/15	2/15

Suppose the store owner actually pays \$25 for each box of drugs and the selling price to customers is \$40. The delivery charge of the drugs is \$10 disregarding the number of boxes ordered. Unsold boxes after expire date will be returned to the whole-saler without any charge. Find the expected weekly revenue and the standard deviation.

Solution: Let $h(X)$ be the revenue function.

$$\begin{aligned} E(X) &= \sum_i x_i p(x_i) = \\ E(X^2) &= \sum_i x_i^2 p(x_i) = \\ \mathrm{Var}(X) &= E(X^2) - [E(X)]^2 = \\ h(X) &= \\ E[h(X)] &= \\ \mathrm{Var}[h(X)] &= \\ SD[h(X)] &= \sqrt{\mathrm{Var}[h(X)]} = \end{aligned}$$

10.3 Continuous Random Variables.

Consider r.vs. that are not counts but measurements on some random feature, e.g. reaction times, waiting times, stress levels, birthweights, heights of desert plants.

• Construct a relative frequency histogram.

• Use the *shape* to suggest the form of a probability density function (pdf) for a random variable model. Let the r.v. be X.

We develop a mathematical model by working with the distribution function

$$F(x) = P(X \leq x).$$

General properties of F(x).

• $0 \leq F(x) \leq 1$, as F is a probability.

• $F(x)$ is an increasing function of x.

• If $a < b$, $P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a)$.

• If X is a continuous r.v. then $P(X \leq x) = P(X < x)$, i.e. the probability of taking a specific value is 0.

• The mean and variance for continuous r.v.'s are defined as:

$$\begin{aligned} E(X) &= \mu = \int x f(x) dx \\ Var(X) &= \sigma^2 = E(X-\mu)^2 = \int (x-\mu)^2 f(x) dx \end{aligned}$$

compared with $\mu = \Sigma_i i p_i$ for discrete r.v.. The μ and σ^2 are the large sample limits of the sample mean $\bar{x}$ and sample variance σ^2.

Example: A number X is chosen at random from $[0, 1]$.

$$F(x) = P(X \leq x) = \begin{cases} 0, & \text{if } x < 0, \\ x, & \text{if } 0 \leq x \leq 1, \\ 1, & \text{if } x > 1. \end{cases}$$

The probability of selecting a number in a given interval is proportional to the length of the interval. Clearly the end points do not add any length to the interval so $P(X = a) = 0$.

10.4 Probability density function (p.d.f.)

In general the p.d.f. for a continuous r.v. X is

$$f(x) = \frac{d}{dx}\ F(x) = \lim_{h\to 0} \frac{1}{h} P(x \le X \le x+h)$$

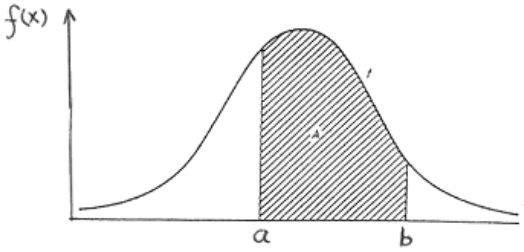
the average 'density of probability' over $[x, x+h]$, for small h.

• $f(x) \ge 0$.

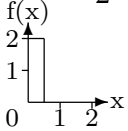
•$P(a < X \le b) = \int_a^b f(x)dx = F(b) - F(a)$.

• The pdf is a curve $y = f(x)$: $f(x) \ge 0$. The **total area** under the curve is 1.

• The area under the curve $y = f(x)$ corresponds to a probability. The pdf plays the same role as p_i does for counting r.v.s. It reflects the shape of a smoothed relative frequency histogram based on a large sample of independent observations of X.



• $f(x)$ is NOT a probability but a density. Hence $f(x) > 1$ for some x is possible.
E.g. for $X \sim U(0, \frac{1}{2})$ (uniform distribution), $f(x) = 2$. Area $= 2 \times \frac{1}{2} = 1$.



11 Continuous random variables

11.1 Standardized random variable

For a r.v. X with mean μ and variance σ^2 we define the standardized r.v. Z by

$$Z = \left(\frac{X-\mu}{\sigma}\right).$$

Then

$$E(Z) = \frac{1}{\sigma}E(X) - \frac{\mu}{\sigma} = 0,$$

$$var(Z) = \left(\frac{1}{\sigma}\right)^2 var(X) = 1.$$

Standardized r.v. has a *mean* 0 and a *variance* 1. This adjustment for *centre* and *spread* enables us to compare different r.v.'s on the basis of the shapes of their distributions.

11.2 The Normal random variable

Many data sets have a symmetric, bell-shaped histogram. These situations can often be modelled by a *normal* or *Gaussian* r.v.

• X is a normal random variable with mean μ and variance σ^2, written $X \sim \mathcal{N}(\mu, \sigma^2)$ if X has pdf

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}, \quad -\infty < x < \infty.$$

• The normal with mean 0 and variance 1 is called the *standard normal r.v.* and is generally denoted by Z. Thus $Z \sim N(0,1)$.

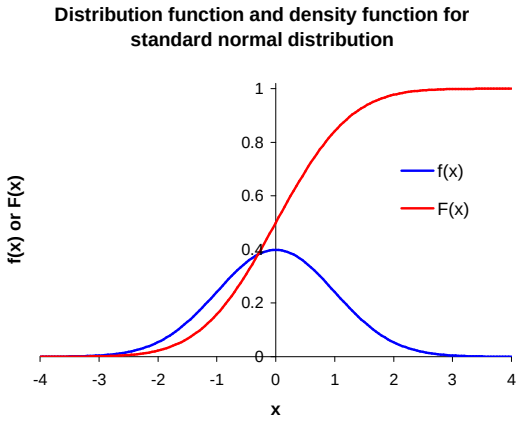
• Z has pdf

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}, \quad -\infty < x < \infty.$$

• Z has distribution function

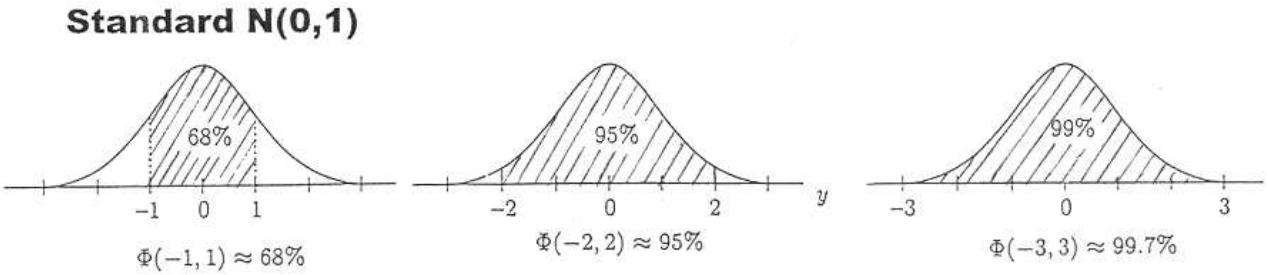
$$\Phi(z) = P(Z \leq z) = \int_{-\infty}^{z} \phi(x)dx.$$

$\Phi(z)$ and $\phi(z)$ is tabulated for $0 \leq z \leq 4$.



• To calculate probabilities recall:

1. the pdf is symmetric about the mean μ and bell-shaped;
2. the total area under the curve is 1;
3. (i) About 68.26% of the area is within the bounds $\mu \pm \sigma$;
   (ii) About 95.44% of the area is within the bounds $\mu \pm 2\sigma$;
   (iii) About 99.74% of the area is within the bounds $\mu \pm 3\sigma$.



4. Z is continuous, so $P(Z \leq z) = P(Z < z)$.

Always SKETCH the curve and mark on the area of interest before trying to use the tables.
Express all probabilities in terms of $P(Z \leq z)$ for some $z \geq 0$.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.99993	0.99993	0.99993	0.99994	0.99994	0.99994	0.99994	0.99995	0.99995	0.99995
3.9	0.99995	0.99995	0.99996	0.99996	0.99996	0.99996	0.99996	0.99996	0.99997	0.99997
4	0.99997									

• **General identities**

1. $P(Z \leq -z) = \Phi(-z) = 1 - \Phi(z)$.

2. $$\begin{aligned} P(|Z| \leq z) &= P(-z \leq Z \leq z) \\ &= \Phi(z) - \Phi(-z) \\ &= 2\Phi(z) - 1. \end{aligned}$$

• **R commands**

The commands `pnorm`, `qnorm`, `dnorm` and `rnorm` are for the cdf, the inverse of cdf, pdf and random variable generation of normal distribution respectively.

Examples. Find

(a) $P(Z \leq 1.6)$,

(b) $P(Z > 1.24)$,

(c) $P(-1.2 < Z < 0.84)$.

(d) Find c and d such that

(i) $P(Z > c) = 0.05$

(ii) $P(|Z| \leq d) = 0.668$.

Solution:

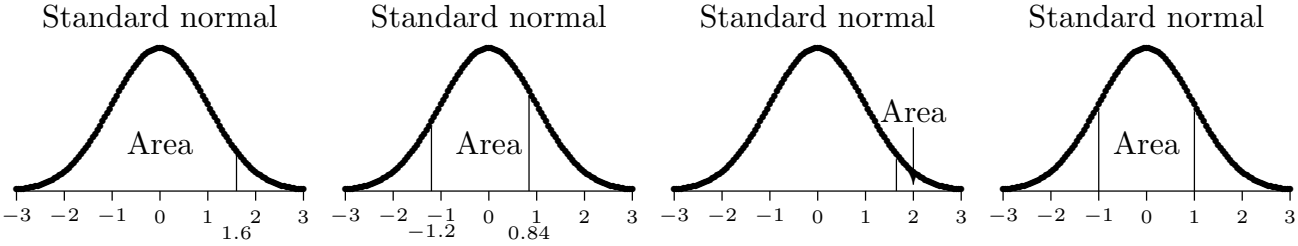
In R,

```
> pnorm(1.6,0,1,lower.tail = TRUE, log.p = FALSE)
[1] 0.9452007
> pnorm(1.24,0,1,lower.tail = FALSE, log.p = FALSE)
```

```
[1] 0.1074877
> pnorm(0.84,0,1,lower.tail = TRUE, log.p = FALSE)-
pnorm(-1.2,0,1,lower.tail = TRUE, log.p = FALSE)
[1] 0.6844761
> qnorm(0.05,0,1,lower.tail = FALSE, log.p = FALSE)
[1] 1.644854
> dd=(1-0.668)/2
> -1*qnorm(dd,0,1,lower.tail = TRUE, log = FALSE)
[1] 0.9700933
```

From table,

$$
\begin{aligned}
P(Z \leq 1.6) &= \\
P(Z > 1.24) &= \\
P(-1.2 < Z < 0.84) &= \\
&= \\
&= \\
P(Z > c) &= 0.05 \\
\Rightarrow & \\
P(|Z| \leq d) &= 0.668 \\
\Rightarrow & \\
\Rightarrow &
\end{aligned}
$$



Let $Z \sim \mathcal{N}(0,1)$. Define $X = \mu + \sigma Z$. From the rules for expectations and variances we have that $E(X) = \mu$ and $var(X) = \sigma^2$. That is $X \sim \mathcal{N}(\mu, \sigma^2)$.

The standardized version of X is

$$Z = \frac{X - \mu}{\sigma}.$$

To calculate any probabilities associated with normal r.v.s we *first express the probability in terms of a standard normal r.v.* then use the tables.

Example: $X \sim \mathcal{N}(3, 2^2)$. Find $P(X \leq 4)$ and $P(X < 1.24)$.

Solution: In R,

```
> pnorm(4,3,2,lower.tail = TRUE, log.p = FALSE)
[1] 0.6914625
> pnorm(1.24,3,2,lower.tail = TRUE, log.p = FALSE)
[1] 0.1894297
```

By calculation,

$$\begin{aligned} P(X \leq 4) &= \\ &= \\ &= \end{aligned}$$

$$\begin{aligned} P(X < 1.24) &= \\ &= \\ &= \\ &= \end{aligned}$$

Example: $X \sim \mathcal{N}(5, 3^2)$. Find c such that

$$P(X > c) = 0.1.$$

Solution: In R,

```
> qnorm(0.1,5,3,lower.tail = FALSE, log.p = FALSE)
[1] 8.844655
```

By calculation,

$$P(X > c) = 0.1$$

$$\Rightarrow$$

$$\Rightarrow$$

$$\Rightarrow$$

$$\Rightarrow$$

12 Central limit theorem

12.1 More on Normal distribution

Example: Birthweights (in grams) of full term babies can be modelled by the normal r.v.

$$W \sim \mathcal{N}(3190, 525^2).$$

(a) What is the probability that a term baby will weigh more than 3.86 kg?

(b) In a sample of 60 term babies born in a hospital in one week how many would you expect to weigh more than 3.86 kg?

(c) Find the weight corresponding to the 75th percentile for term babies. That is, the weight c such that 25% of all term babies are heavier than c gms.

Solution: Let X_1 be the weight of a term baby.

(a) $P(X > 3.86) =$

$=$

$=$

(b) $E(X) =$

$=$

(c) $P(X < c) = 0.75$

$\Rightarrow$

$\Rightarrow$

12.2 Sums of Normal Random Variables.

Many observed phenomena can be modelled as the sum of several random components, e.g. total weight of passengers in a lift. In other situations it is the difference of two r.v.s that is of interest, e.g. weight gain on different diets.

Let $X_1, X_2, \ldots, X_n$ be *independent* r.v.s with

$$X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$$

for $i = 1, 2, \ldots, n$. Normal r.v's have the special property that the sum of independent normal r.v.'s still has a normal distribution.

Let $a_1, a_2, \ldots, a_n$ be constants.

$$\sum_{i=1}^{n} a_i X_i \sim \mathcal{N}\left(\sum_{i=1}^{n} a_i \mu_i,\ \sum_{i=1}^{n} a_i^2 \sigma_i^2\right).$$

If all $X_i \sim \mathcal{N}(\mu, \sigma^2)$ then

$$T = \sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2)$$

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathcal{N}(\mu, \frac{\sigma^2}{n}).$$

• Independence is needed to obtain the above results, particularly the variance formulae.

• If all X_i have the *same variance* σ^2 then

$$var(\bar{X}) = \frac{\sigma^2}{n}.$$

• If X_1 and X_2 are *independent*

$$var(X_1 - X_2) = \sigma_1^2 + (-1)^2\sigma_2^2 = \sigma_1^2 + \sigma_2^2.$$

Example: (Book, P.71) Steel rods, made with diameter

$$R \sim \mathcal{N}(4.90, 0.03^2) \quad \text{(in cm)},$$

are to fit into sockets, made with diameter

$$S \sim \mathcal{N}(5.00, 0.04^2) \quad \text{(in cm)}.$$

For a satisfactory fit the socket diameter should exceed the rod diameter, but by no more than 0.20 cm. If a rod and socket are taken at random, what is the probability that the fit is unsatisfactory?

Solution: Consider $S - R \sim$,
i.e. .

$$\begin{aligned} P(0 < S - R < 0.20) &= \\ &= \\ &= \end{aligned}$$

Example: (Beelte) The tibia length of a certain species of beetle can be modelled by $L \sim \mathcal{N}(7.8, 0.3^2)$ mm.

(i) What is the probability that the average length of 25 independent tibia lengths will be less than 7.6 mm?

(ii) What is the probability that the average will differ from 7.8 by more than 0.1?

Solution: Let X denote the length $L \sim \mathcal{N}(7.8, 0.3^2)$ mm.

(i) Consider $\overline{X}_{25} \sim$ mm.

$$\begin{aligned} P(\overline{X}_{25} < 7.6) &= \\ &= \\ &= \\ &= \end{aligned}$$

(ii) Consider $\overline{X}_{25} - 7.8 \sim$ mm.

$$\begin{aligned} P(\overline{X}_{25} - 7.8 > 0.1) &= \\ &= \\ &= \\ &= \end{aligned}$$

12.3 Central Limit Theorem

When we add two independent normal r.v.'s then the sum still has a normal distribution. This is not true for arbitrary distributions.

Example: (Dice) $X_1, X_2, \ldots, X_6$ are indep. r.v.'s with distribution

x_i:	1	2	3	4	5	6
p_i:	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

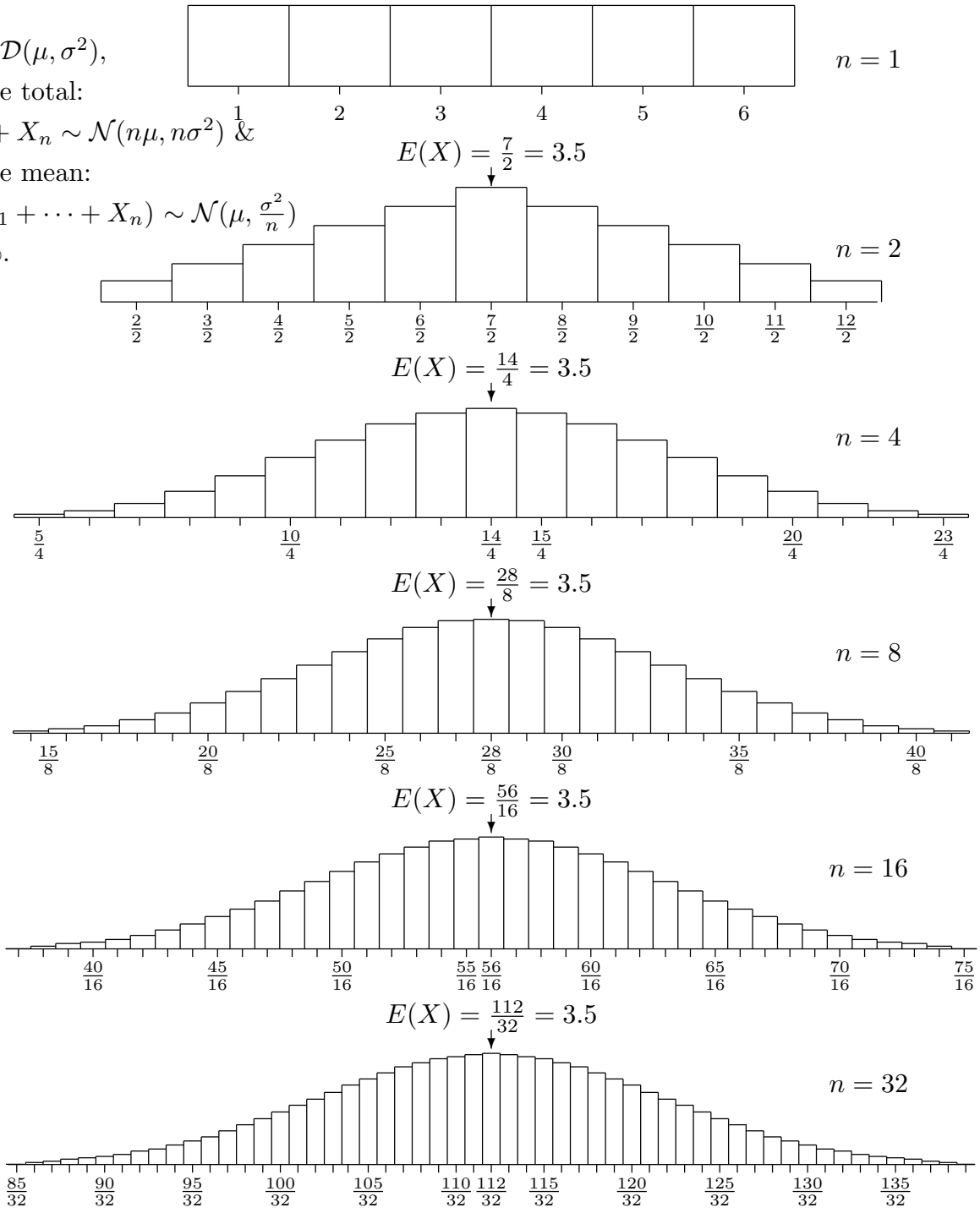
$M_2 = \frac{1}{2}(X_1 + X_2)$ has distribution

m_{2i}:	$\frac{2}{2}$	$\frac{3}{2}$	$\frac{4}{2}$	$\frac{5}{2}$	$\frac{6}{2}$	$\frac{7}{2}$	$\frac{8}{2}$	$\frac{9}{2}$	$\frac{10}{2}$	$\frac{11}{2}$	$\frac{12}{2}$
p_i:	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

$M_3 = \frac{1}{3}(X_1 + X_2 + X_3), \ \ldots$

Note the distribution of M_3 is starting to cluster around the mean $E(M_3) = 3.5$.

For $X_i \sim \mathcal{D}(\mu, \sigma^2)$,
the sample total:
$X_1 + \cdots + X_n \sim \mathcal{N}(n\mu, n\sigma^2)$ &
the sample mean:
$\overline{X} = \frac{1}{n}(X_1 + \cdots + X_n) \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$
as $n \to \infty$.



Distribution of the sum of n die rolls-Illustration of Central Limit Theorem

12.4 Central Limit Theorem (CLT)

If $X_1, X_2, \ldots, X_n$ are independent and identically distributed random variables with mean μ and variance σ^2 $(0 < \sigma^2 < \infty)$ then

$$P\left(\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n\sigma^2}} \leq x\right) \to \Phi(x) = P(Z \leq x)$$

as $n \to \infty$.

Thus for n large ($n \geq 25$) the following are *approximately* true

$$T = X_1 + X_2 + \cdots + X_n \sim \mathcal{N}(n\mu, n\sigma^2)$$

and

$$\bar{X} = (X_1 + \cdots + X_n)/n \sim \mathcal{N}(\mu, \frac{\sigma^2}{n}).$$

The closer the distribution of X_i is to the normal the better the approximation.

Example: The answers to the beetle tibia questions will be approximately correct regardless of the exact distribution of tibia length.

Example: (Blood pressure) Systolic blood pressure (bp) readings for pre-menopausal, non-pregnant women aged 35 -40 have a mean of 122.6 mm Hg and an s.d. of 11 mm Hg. An independent sample of 25 women is drawn from this target population and their bp recorded.

(i) What is the probability that the average bp is greater than 125 mm hg?

(ii) If the sample size was increased to 40 how would the answer to (i) change?

Solution: Let $\overline{X}$ be the average bp. By CLT, $\overline{X} \sim$.

(i) $P(\overline{X} > 125) =$

$=$

$=$

$=$

(ii) $P(\overline{X} > 125) =$

$=$

$=$

$=$

13 Normal approximation to binomial distribution

Let X_i be independent r.v.'s defined as follows.

$$\begin{aligned} X_i &= 1 \text{ if } i\text{th trial is an 'S'} \\ &= 0 \text{ if } i\text{th trial is an 'F'} \end{aligned}$$

Let p denote the probability of an 'S' on the ith trial. Then

$$X = X_1 + X_2 + \cdots + X_n \sim \mathcal{B}(n, p)$$

with

$$E(X) = np$$

$$var(X) = n\ var(X_1) = np(1-p).$$

By CLT, X is approximately $\mathcal{N}(np, np(1-p))$.

The approximation is quite good if $np \geq 5$ *and* $n(1-p) \geq 5$. The closer p is to 0.5 the better the approximation for small n. However when n is large, the calculation of binomial probability becomes complicated.

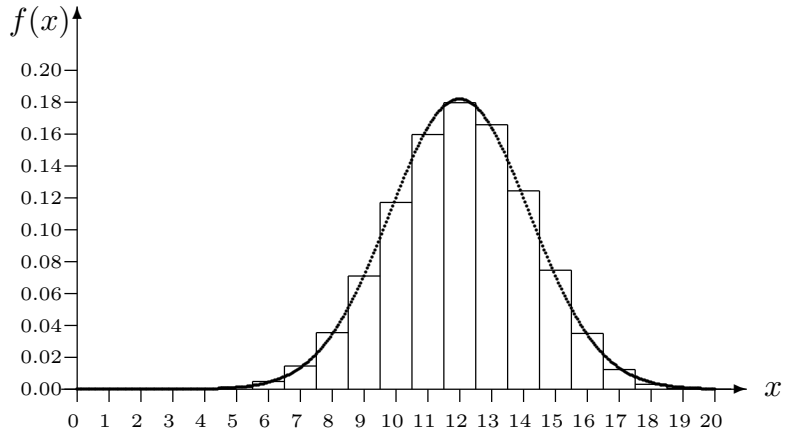
Example: (Normal approximation) $X \sim \mathcal{B}(20, 0.6)$. Find the exact probability $P(X = 10)$.

Solution: We have $n = 20$ and $p = 0.6$. The exact probability is

$$P(X = 10) =$$

$$=$$

The following plot shows that the binomial distribution can be closely approximated by a normal distribution with the same mean and variance.



13.1 Continuity Correction

To approx. binomial probabilities using the normal, we consider the areas of corresponding rectangles with bases adjusted by adding or subtracting 0.5 to X to *increase* the area under the normal curve.

Example: (Normal approximation) Calculate the probability using normal approximation and compare.

Solution: We have $X \sim \mathcal{N}(np,\ np(1-p))$,
i.e., or .

$$P(X = 10) \simeq$$
$$=$$
$$=$$
$$=$$
$$=$$

Comparing the probability to the area under the approximating normal curve, we see a close but over-estimation of the probability.

Example: If $X \sim \mathcal{B}(12, 0.5)$ find
(i) $P(X \leq 4)$
(ii) $P(X > 6)$
(iii) $P(2 \leq X \leq 5)$
using normal approximation.

Solution: We have $X \sim \mathcal{N}(np,\ np(1-p))$,
i.e., or .

$$\begin{aligned}
\text{(i) } P(X \leq 4) &\simeq \\
&= \\
&= \\
\\
\text{(ii) } P(X > 6) &\simeq \\
&= \\
&= \\
\\
\text{(iii) } P(2 \leq X \leq 5) &\simeq \\
&= \\
&= \\
&=
\end{aligned}$$

Example (Book, P.80 Q21) It is known that 80% of patients with a certain disease can be cured with a certain drug. What is the probability that amongst 150 patients with the disease, at most 37 of them cannot be cured with the drug.

Solution: Let X be the number of patient who cannot be cured with the drug. We have $X \sim \mathcal{N}(np,\ np(1-p))$,
i.e., or .

$$\begin{aligned} P(X \leq 37) &\simeq \\ &= \\ &= \end{aligned}$$

Example: The proportion of children having a particular type of birth defect born to Pima Indian women is 0.05. Calculate the probability that in 785 independent births no more than 21 children have the birth defect.

Solution: Let X be the number of patient who cannot be cured with the drug. We have $X \sim \mathcal{N}(np,\ np(1-p))$,
i.e., or .

$$\begin{aligned} P(X \leq 37) &\simeq \\ &= \\ &= \end{aligned}$$

In practice we sample from populations where we do not know the *true* probability distribution. In some cases we may have a proposed or hypothesised model and we want to see if the data are consistent with the model.

Example: In the above Pima Indian case we might know the birth defect rate in another population is 0.05 and we want to check if this model holds in the Pima population. We observe 21 defects in a sample of 785 births. What we have observed is an outcome that is extremely unlikely if $p = 0.05$. What do we conclude?

Solution: Let X be the number of patient who cannot be cured with the drug. We have $X \sim \mathcal{N}(np,\ np(1-p))$,
i.e., or .

$$\begin{aligned} P(X \leq 21) &\simeq \\ &= \\ &= \end{aligned}$$

We conclude that the birth defect rate is unlikely to be 0.05.

13.2 Sampling Distributions

To carry out statistical inference we need to know how statistics vary across samples. For example, suppose we sample 5 adult males, measure their heights and calculate the average $\bar{x}$ and variance s^2. We know that these values will vary from one sample to the next. We call the distribution of a statistic across samples its *sampling distribution.*

If male heights can be modelled by $H \sim N(178, 7^2)$ cm then the

sample averages will vary about 178.

If we have 5 independent readings we can construct a *random variable formula* which gives the recipe for calculating the statistic.

Observations X_1, X_2, X_3, X_4, X_5

Average $\bar{X} = (X_1 + X_2 + X_3 + X_4 + X_5)/5$

$$\bar{X} \sim N(178, 9.8)$$

Knowing the sampling distribution helps identify "unusual" statistic values.