Random Variables

Outline

Discrete Distributions

Definition (Discrete Distribution)

For any random variable X with a discrete distribution, we have a sample space Ω with values $x = \{x_1, x_2, \ldots\}$ and associated probabilities $\{p_1, p_2 \ldots\}$, where $\{p_i = P(X = x_i)\}$.

Properties:

- there is a countable number of possible values;
- $\sum_i p_i = 1$

Definition (Probability Distribution Function)

The probability distribution function (or probability distribution) of X is the set of $\{x, P(X = x)\}$.

Definition (Cumulative Distribution Function (CDF))

The cumulative distribution function (CDF) of X is

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

This is a step function.

The Mean of a Discrete Distribution

Definition (Mean or Expectation)

The mean of X is

$$\mu = E(X) = \sum_{\text{all } x} xP(X = x)$$

Definition (Expectation of a Function)

The expectation of $g(X)$ is

$$E(g(X)) = \sum_{\text{all } x} g(x)P(X = x)$$

For example: $E(X^2) = \sum_{\text{all } x} x^2 P(X = x)$.

The Variance of a Discrete Distribution

Definition (Variance)

The variance of X is

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

Example

Mean and variance of 5 tosses of a coin

Let $X =$ the number of heads in 5 tosses of a coin, $x = 0, 1, \ldots, 5$, with probability distribution function:

x	0	1	2	3	4	5
$P(X = x)$	$\frac{1}{32}$	$\frac{5}{32}$	$\frac{10}{32}$	$\frac{10}{32}$	$\frac{5}{32}$	$\frac{1}{32}$

Find the mean and variance of X.

$$\mu = E(X) = \sum_x xP(X = x) = 0 \times \frac{1}{32} + 1 \times \frac{5}{32} + \ldots 5 \times \frac{1}{32} = 2.5$$

$$E(X^2) = \sum_i x^2 P(X = x) = 0^2 \times \frac{1}{32} + 1^2 \times \frac{5}{32} + \ldots 5^2 \times \frac{1}{32} = 7.5$$

Hence

$$Var(X) = E(X^2) - E(X)^2 = 7.5 - (2.5)^2 = 1.25$$

Common Types of Discrete Distributions

There are an infinite number of discrete distributions.

We will concentrate on 2 special examples:

- The Binomial Distribution in this lecture;
- The Hypergeometric Distribution (in the extra material section and in the tutorial);

Binomial Distribution

Definition (Binomial Distribution)

The Binomial distribution models a context in which we have:

- a fixed number n of independent Binary trials;
- a fixed likelihood of a success at each trial $p = P(\text{success})$.

If $X =$ the number of successes in n trials, then $X \sim Bin(n, p)$ with

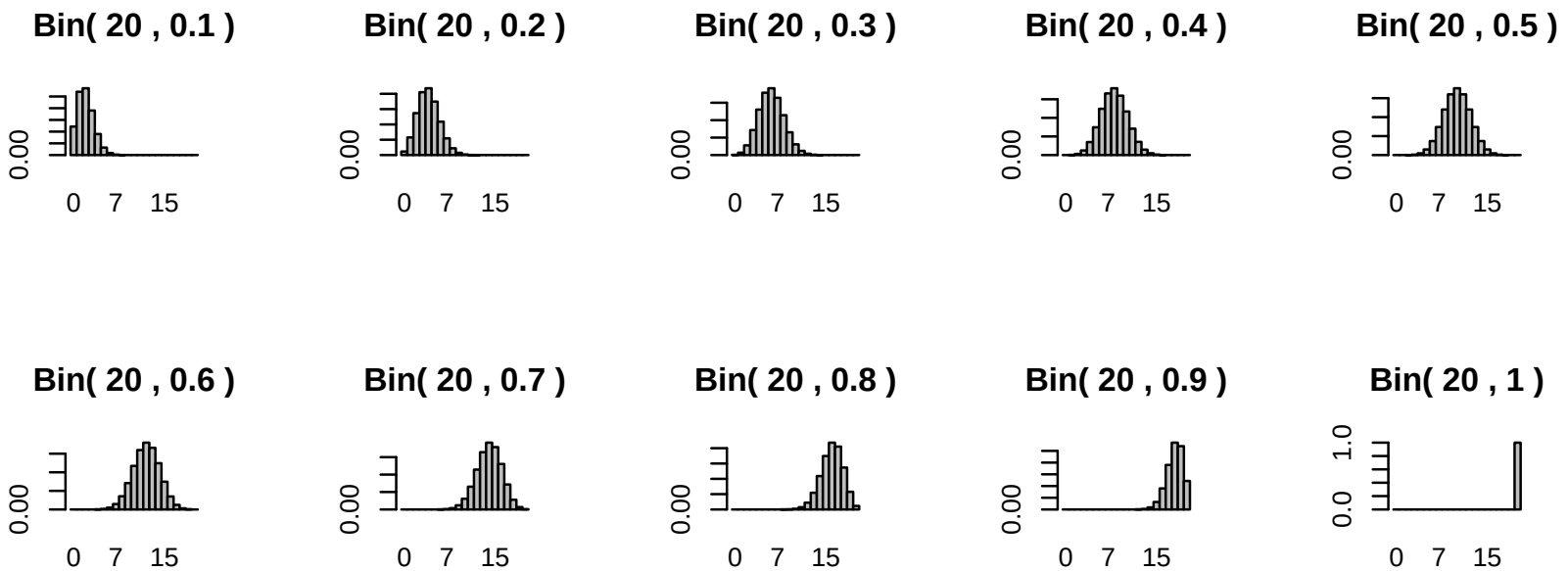
$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \qquad \text{for } x = 0, 1, 2, \ldots, n.$$

Notes:

- A *Binary* (or Bernoulli) trial is an event where there can only be 2 options: success or failure. For example, 1 fruit fly is buzzing around a fruit tree: will it land or not?
- *Success* designates the event we are interested in counting, which may not be good. For example, $p = P(\text{fruit fly lands on fruit tree})$.
- The Binomial distribution has 2 parameters: n and p. Parameters represent the numerical inputs needed for the model.
- It can be shown (by algebra) that the Binomial distribution has mean $E(X) = np$ and variance $Var(X) = np(1-p)$.

Example of Binomial Distribution with changing p

$X \sim Bin(n = 20, p)$, for different $p = 0.1, 0.2, \ldots, 1$.



Example

Fruit Flies

There are 10 fruit flies buzzing around a lime tree. The flies have a 20% chance on landing on the tree and act independently. If X represents the number of flies that land on the tree, what is the distribution of X? What is the chance that no flies land on the tree? What is the chance that less than 2 flies land on the tree?

1. Identify the model:
   $X =$ the number of flies that land $\sim Bin(n, p)$, where
   $n =$ number of flies $= 10$; $p = P(\text{fruit fly lands}) = 0.2$.
2. Calculate probability:
   $P(\text{no flies land}) = P(X = 0) = \binom{10}{0}(0.2)^0(0.8)^{10} = \frac{10!}{0!(10-0)!}(0.8)^{10} = (0.8)^{10} \approx 0.11$.

Calculate probability:
$P(\text{less than 2 flies land}) = P(X \leq 1) = P(X = 0) + P(X = 1) = 0.1073742 + \binom{10}{1}(0.2)^1(0.8)^9 \approx 0.38.$

```
# dbinom(x,n,p) calculates P(X=x) for Bin(n,p)
dbinom(0,10,0.2)

## [1] 0.1073742

dbinom(1,10,0.2)

## [1] 0.2684355
```

```
# pbinom(x,n,p) calculates P(X<=x) for Bin(n,p)
pbinom(1,10,0.2)

## [1] 0.3758096
```

Hypergeometric Distribution

If we randomly select *n* items without replacement from a set of *N* items of which:

- N_1 of the items are of one type and
- N_2 of the items are of a second type,

then the probability mass function of the discrete random variable *X* is called the **hypergeometric distribution** and is of the form:

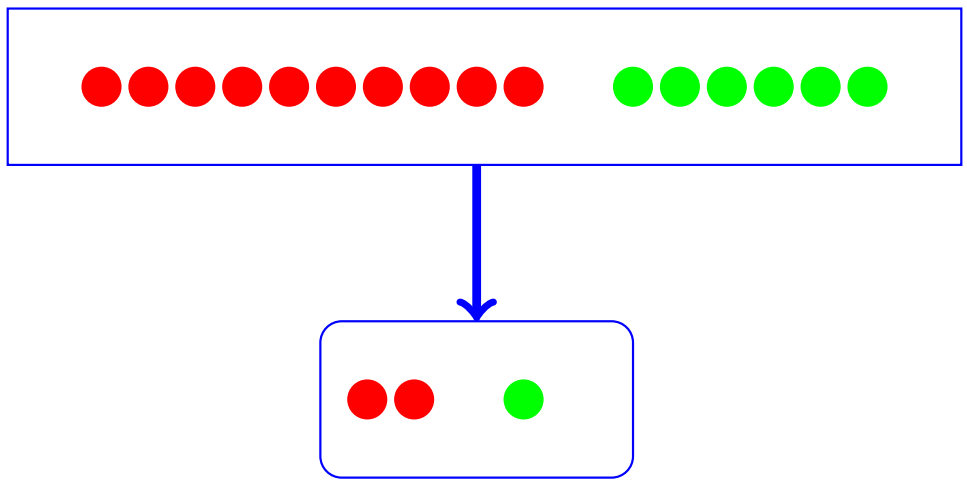
$$P(X = x) = \frac{\binom{N_1}{x}\binom{N_2}{n-x}}{\binom{N}{n}}$$

where *x* that satisfies the inequalities:

- $x \leq N_1$
- $x \leq n$
- $n - x \leq n$

Example (Packet of M & Ms)

Suppose a small Christmas packet of M & Ms contains 16 chocolates, of which 10 are red and 6 are green. We select a random sample of 3 chocolates. What is the probability of selecting exactly 2 red ones?



The probability that we select exactly $n = 2$ red balls is

$$\frac{\binom{10}{2}\binom{6}{1}}{\binom{16}{3}} \approx 0.48$$

```
choose(10,2)*choose(6,1)/choose(16,3)
```

```
## [1] 0.4821429
```

```
dhyper(2,10,6,3)    # dhyper(x,N1,N2,n)
```

```
## [1] 0.4821429
```

Examples

A box contains 6 red balls and 4 white balls.

a. If 3 balls are selected one at a time with replacement, what is the probability of getting exactly 2 red balls?

b. If 3 balls are selected one at a time without replacement, what is the probability of getting exactly 2 red balls?

Continuous Random Variables

- We introduced random variables to model real data.
- But keep in mind, we have 2 main types of data, categorical data and numerical data.
- The discrete random variables we just talked about was used to model categorical data.
- The numerical data on the other hand, is usually modelled by **continuous random variables**.

Comparing Discrete and Continuous Distributions

There are 5 fundamental differences between discrete and continuous distributions:

	Discrete	Continuous
Values	Countable	Infinite
Plot	Histogram $P(X = x)$ probability distribution function	Smooth curve $f(x)$ probability density function (pdf)
$P(X = x)$	$0 \leq P(X = x) \leq 1 \ \ \forall x$	$P(X = x) = 0 \ \ \forall x$
Sum of Probabilities	$\sum_x P(X = x) = 1$ Area of histogram	$\int_x f(x)dx = 1$ Area under density
$F(x) = P(X \leq x)$	$\sum_{y=min(x)}^{x} P(X = y)$	$\int_{-\infty}^{x} f(y)dy$

For any continuous distribution:

- there is an infinite number of possible values;
- these values may be within a fixed interval. For example, male human heights (in cm) belong to [54.6,272].
- the total of all the probabilites, represented by the area under the probability density function (pdf), must be 1.
- For a continuous distribution

$$P(a < X < b) = P(a \leq X \leq b)$$

This is not generally true for a discrete distribution.

Normal Distribution

Definition (Normal Distribution)

The Normal distribution models a symmetric, bell-shaped variable with 2 parameters mean μ and variance σ^2 and points of inflection at $\mu \pm \sigma$. We say the variable $X \sim N(\mu, \sigma^2)$.

The probability density function (pdf) is:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad \text{for } x \in (-\infty, \infty)$$

The cumulative distribution function (CDF) is

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(y)dy$$

The Normal Distribution is very important because it can approximate many natural phenomemon, like annual rainful (sometimes skewed), humidity, evapotranspiration, heights/weights/length of animals, intelligence, and measurement errors.

It can also approximate sums of random variables, via the Central Limit Theorem (stay tuned).

Does the Normal Distribution approximate the distribution of human heights?

Normal Probabilities

What is the probability of finding an Australian woman of 'goal player' height or taller?

If $X \sim N(161.8, 6^2)$, what is $P(X > 189)$?

Method1: Integrate the pdf to calculate area under the curve

$$P(X > 189) = \int_{189}^{\infty} \frac{1}{\sqrt{2\pi(6^2)}} e^{-\frac{(y-161.8)^2}{2(6^2)}} dy$$

There is no closed form, but we could Numerical Integration.

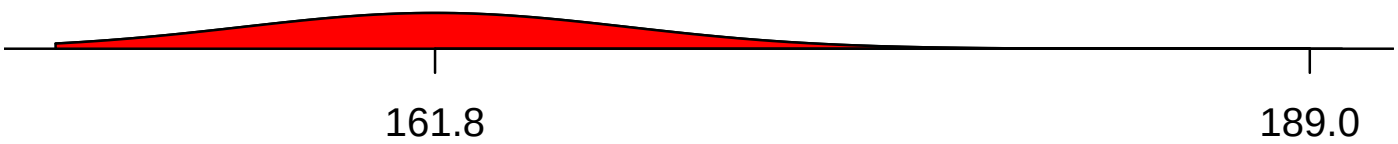
```
f <- function(x) {dnorm(x,161.8,6)}
integrate(f,189,200)

## 2.902907e-06 with absolute error < 3.2e-20
```

Method2: Use R

```
pnorm(189,161.8,6)   #pnorm(x,mean,sd)

## [1] 0.9999971
```



Height of Australian women in cms (x)

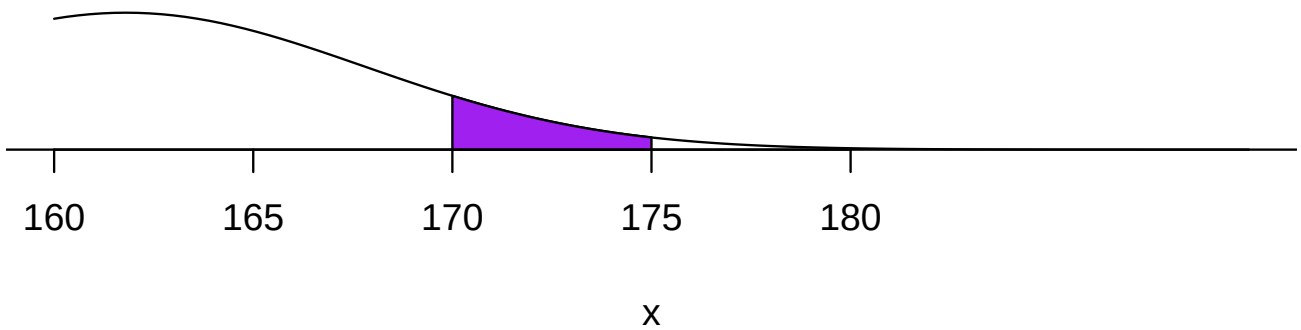
Upper Tail probabilities are found from Lower Tail probabilties:

$$P(X > 189) = 1 - P(X \leq 189) = 2.9e - 06$$

Notes on using R:

- Interval probabilities are found by subtraction:

$$P(170 < X \leq 175) = P(X \leq 175) - P(X \leq 170) = 0.07196146$$

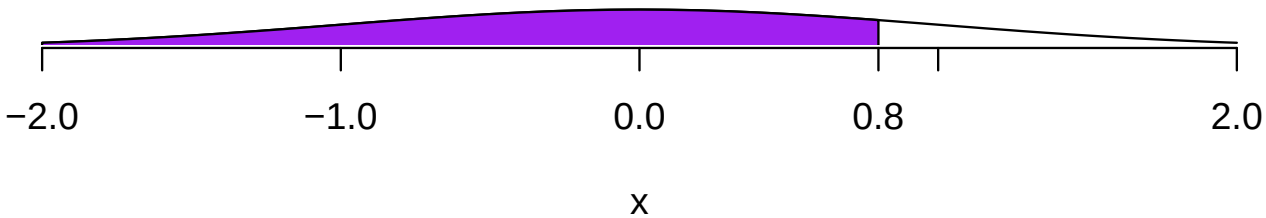



```
pnorm(175,161.8,6)-pnorm(170,161.8,6)
```

```
## [1] 0.07196146
```

- For the standard Normal $Z \sim N(0, 1)$, we can leave the mean and standard deviation unspecified.

$$P(Z \leq 0.8) = 0.7881446$$




```
pnorm(0.8,0,1)
```

```
## [1] 0.7881446
```

```
pnorm(0.8)
```

```
## [1] 0.7881446
```

How do we standardize a general Normal random variable?

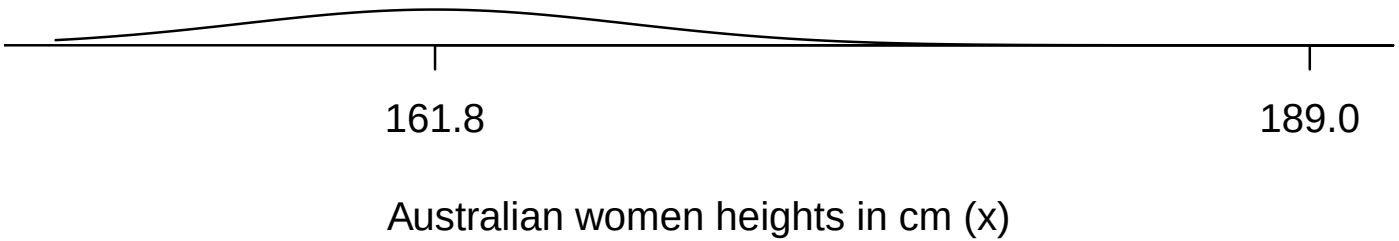
Every General Normal $X \sim N(\mu, \sigma^2)$ can be transformed into the Standard Normal $Z \sim N(0, 1)$.

Definition (Standardardising a Normal)

If $X \sim N(\mu, \sigma^2)$ and $Z \sim N(0, 1)$, then
$P(X \leq x) = P\left(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right) = P\left(Z \leq \frac{x-\mu}{\sigma}\right)$

$$\begin{aligned} P(X > 189) &= P(\frac{X - 161.8}{6} > \frac{189 - 161.8}{6}) \\ &= P(Z > 4.533333) \end{aligned}$$

Effectively we have found that



is equivalent to

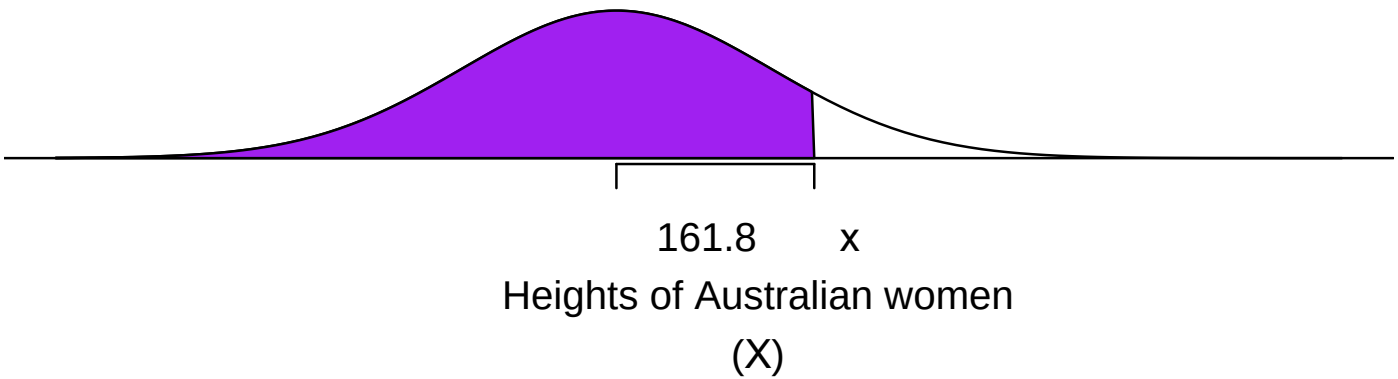

```
1-pnorm(4.533333)
```

```
## [1] 2.903009e-06
```

Normal Percentiles (Inverse Probabilities)

Given $X \sim N(161.8, 6^2)$, what is the 90% percentile for heights of Australian women.

We need to find x such that $P(X \leq \mathrm{x}) = 0.9$.




```
qnorm(0.9,161.8,6)   #qnorm(%,mean,sd)

## [1] 169.4893
```

Summary Examples

Have a go

a) Dharshani Sivalingam (208.3 cm) is the tallest netball player in the world. What is the probability of finding an Australian woman taller than Dharshani given that the average height of a woman is 161.8 with standard deviation 6?
b) Madison Robinson (168 cm) is the shortest Australian International player. What percentage of Australian women are between Madison and Dharshani's heights?
c) If 60% of Australian women are below a certain height, what is that height?

Summary Examples

```
#Check your answers
d=(208.3 - 161.8)/6
pnorm(d)

## [1] 1

m=(168 - 161.8)/6
pnorm(d)-pnorm(m)

## [1] 0.150724

qnorm(0.6,161.8,6)

## [1] 163.3201
```

Linear Function of a Random Variable

Definition (Linear Function of Random Variable)

Given a random variable X, then $Y = a + bX$ has moments

$$\boxed{E(Y) = a + bE(X)}$$

and

$$\boxed{Var(Y) = b^2 Var(X)}$$

for all 2 constants a and b.

Special Case: If $X \sim N(\mu, \sigma^2)$, then $Y \sim N(a + b\mu, b^2\sigma^2)$.

Notes:
(1) Expectation retains linearity.
(2) 'A linear function of a Normal is a Normal'. This is the reason that we can standardise a Normal.

Example: Linear Function

Suppose the weight of an Australian women in kg, $W \sim N(71.1, 12^2)$.

▸ AustralianWeights

Find the distribution of the weight of an Australian women in pounds, given 1kg = 1 pound/2.2046.

Let $P =$ Weight of an Australian women in pounds $= 2.2406W$.
This is a linear function where $a = 0$ and $b = 2.2406$.
Hence

$$E(P) = 0 + 2.2406E(W) = 2.2406 \times 71.1 = 159.3067$$

$$Var(P) = 2.2406^2 Var(W) = 2.2406^2 \times 12^2 = 722.9215$$

So $P \sim N(159.3067, 26.8872^2)$

Definition (Independence for Random Variables)

For any random variables X and Y, we say that X and Y are independent iff

$$P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y)$$

i.e. the joint CDF splits into the 2 individual CDFs.

Notes:
(1) You will not need to justify independence. Rather it will be assumed in any question requiring it.
(2) Symbolically, it is the same as the set independence we introduced last week, with $A = \{X \leq x\}$ and $B = \{Y \leq y\}$.
(3) It follows that if X and Y are independent, then $Cov(X, Y) = E(XY) - E(X)E(Y) = 0$ (which relates to correlation). However the inverse is not true; i.e., Cov(X, Y) can be 0 for variables that are not independent.

Sums of Random Variables

Definition (Total of Random Variables)

Given any sequence of random variables $X_1, X_2, \ldots, X_n$, the total $T = \sum_{i=1}^{n} X_i$ has moments

$$\boxed{E(T) = \sum_{i=1}^{n} E(X_i)}$$

and assuming independence,

$$\boxed{Var(T) = \sum_{i=1}^{n} Var(X_i)}$$

Definition (Sample Mean of Random Variables)

Given any sequence of random variables $X_1, X_2, \ldots, X_n$, the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ has moments

$$\boxed{E(\bar{X}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i)}$$

and assuming independence,

$$\boxed{Var(\bar{X}) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i)}$$

Sums of Normal Random Variables

Definition (Total and Sample Mean of Normal RVs)

Given a sequence of random variables $X_i \sim N(\mu_i, \sigma_i^2)$ (for $i = 1, 2 \ldots, n$)
then

$$T = \sum_{i=1}^{n} X_i \sim N(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2)$$

and

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \sim N(\frac{1}{n} \sum_{i=1}^{n} \mu_i, \frac{1}{n^2} \sum_{i=1}^{n} \sigma_i^2)$$

Summary: for constants a_i,

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

Example: Total

Suppose the weight of an Australian women is $W \sim N(71.1, 12^2)$, a carry on bag is $C \sim N(6.9, 0.5^2)$ and a handbag is $H \sim N(1.3, 0.4^2)$. ▸ QantasCarryonLuggage

Find the probability that the total weight of an Australian woman with carry on luggage is more than 100kg.

Let $T =$ Total Weight of a woman with carry on luggage.
This is a sum of 3 random variables, $T = W + C + H$.
Hence

$$E(T) = E(W) + E(C) + E(H) = 71.1 + 6.9 + 1.3 = 79.3$$

$$Var(T) = Var(W) + Var(C) + Var(H) = 12^2 + .5^2 + .4^2 = 144.41$$

So $T \sim N(79.3, 144.41) = T \sim N(79.3, 12.01707^2)$

So using standardising,

$$P(T > 100) = P(\frac{T - 79.3}{12.01707} > \frac{100 - 79.3}{12.01707}) = P(Z > 1.72255) \approx 0.04$$

```
1-pnorm(100,79.3,12.01707)
```

```
## [1] 0.042485
```

```
1-pnorm(1.72255)
```

```
## [1] 0.04248497
```

Sum of iid normal

Definition (Total and Sample Mean of iid Normal RVs)

Given a sequence of iid random variables $X_i \sim N(\mu, \sigma^2)$ (for $i = 1, 2 \ldots, n$)
then

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

and

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

Example: Sample Mean

Find the probability that the average weight of 10 Australian women with carry on luggage is more than 100kg.

We have already worked out that 1 woman has a total carry on weight of $T \sim N(79.3, 12.01707^2)$.

Now change the notation and consider a sequence of 10 women: $X_1, X_2, \ldots, X_{10}$ where
$X_i =$ carry on weight $\sim N(79.3, 12.01707^2)$.

Assuming the women are independent, let
$\bar{X} =$ Average Weight of 10 women with carry on luggage.

We have

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}) = N(79.3, \frac{12.01707^2}{10}) = N(79.3, 3.80013^2)$$

So using standardising,

$$P(\bar{X} > 100) = P(\frac{\bar{X} - 79.3}{3.80013} > \frac{100 - 79.3}{3.80013}) = P(Z > 5.447182) \approx 0$$

```
1-pnorm(100,79.3,3.80013)
```

```
## [1] 2.558704e-08
```

```
1-pnorm(5.447182)
```

```
## [1] 2.558705e-08
```

What is the Distribution of the Sample Mean for Any Population?

If the population has distribution $X \sim ?(\mu, \sigma^2)$, what is the distribution of $\bar{X}$? We introduce a miracle theorem, which effectively allows us to use the results on Sums from the previous section, even when the random variable are not Normal!

Definition (Central Limit Theorem (CLT))

If $X_i \sim (\mu, \sigma^2)$ for $i = 1, 2, \ldots, n$ then

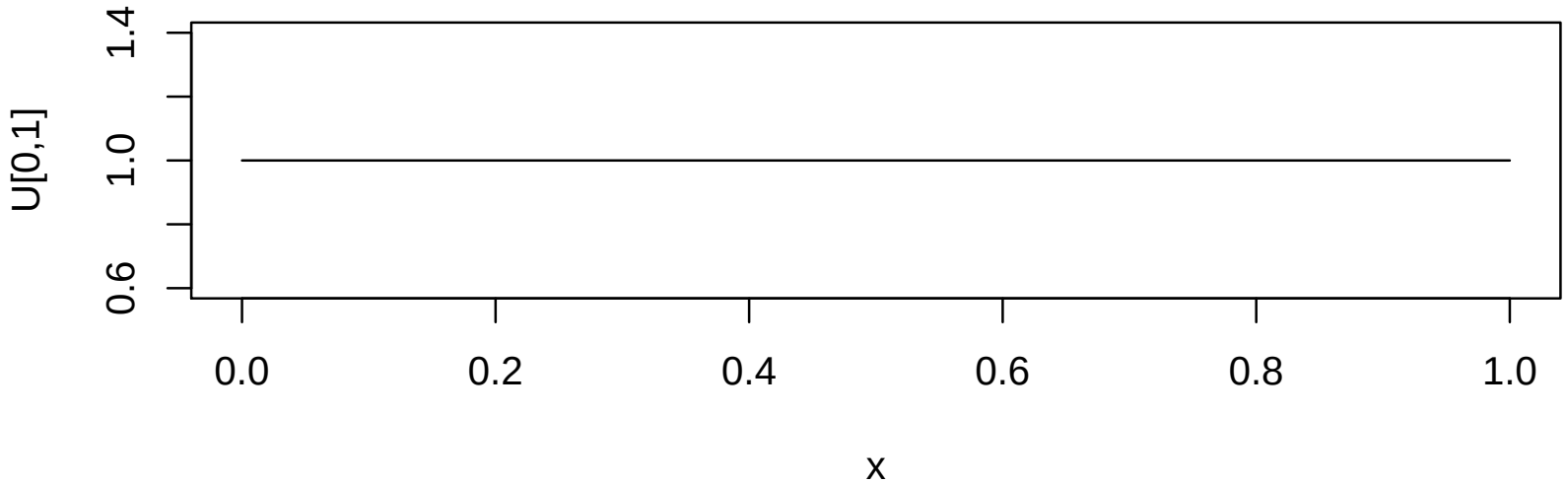
$$\bar{X} \approx N(\mu, \frac{\sigma^2}{n})$$

Notes:

- The CLT is the most important result in this course, and in much of statistical theory.
- The CLT requires few assumptions:
  - We must have a 'big enough' sample size n;
  - We must have finite variance $\sigma^2 < \infty$.
- What is a 'big enough' sample size? Some textbooks give a rule of thumb (eg $n > 25$), but it all depends on the type of distribution. If X is fairly symmetric, then n could be small; if X is highly asymmetric, then n could be larger.
- To visualise the CLT ▸ Lock5 Stat Key ▸ App

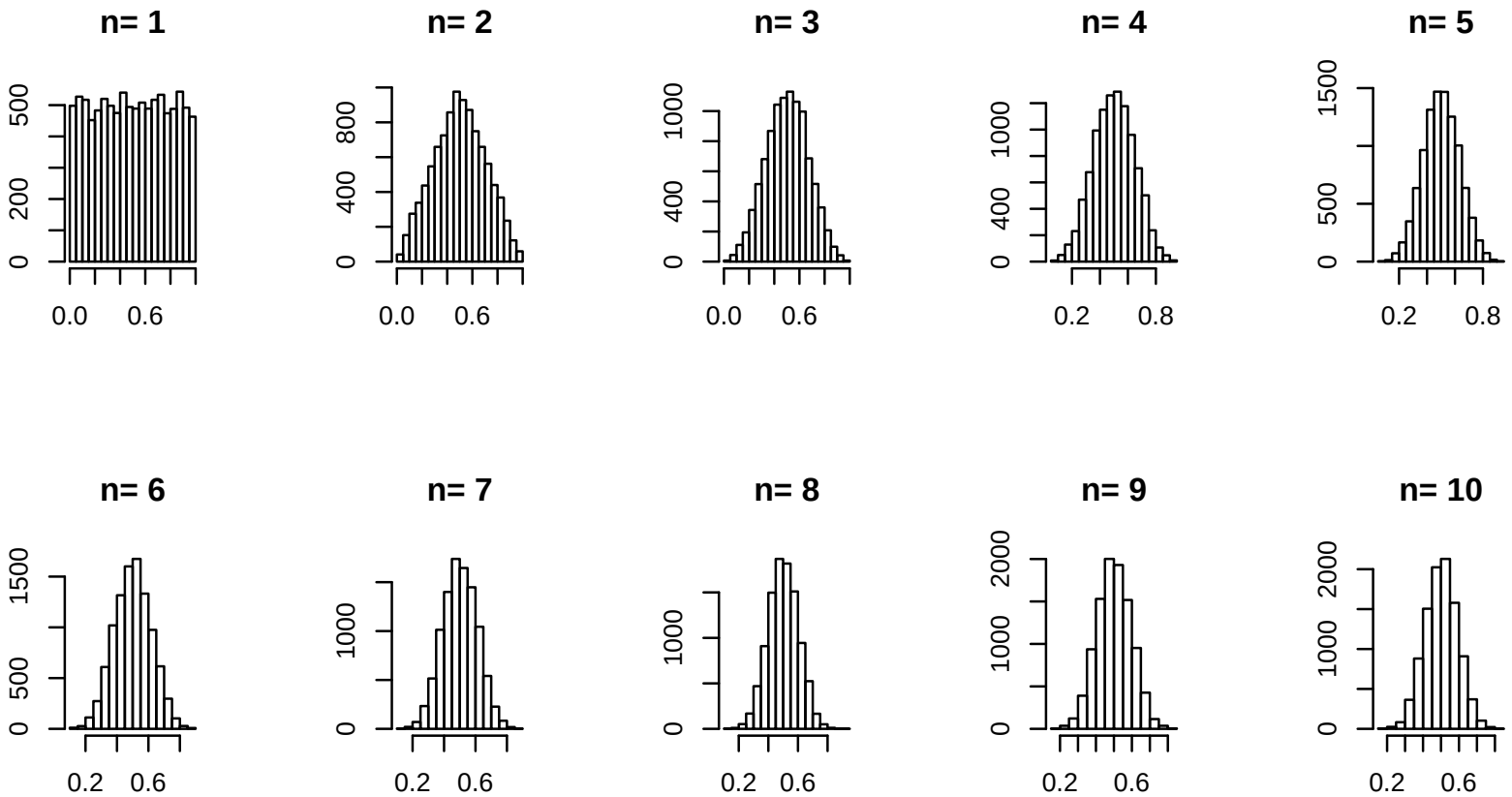
Examples of the CLT

(1) Uniform Distribution: $X \sim U(0,1)$, with $\mu = \frac{1}{2}$ and $\sigma^2 = \frac{1}{12}$.



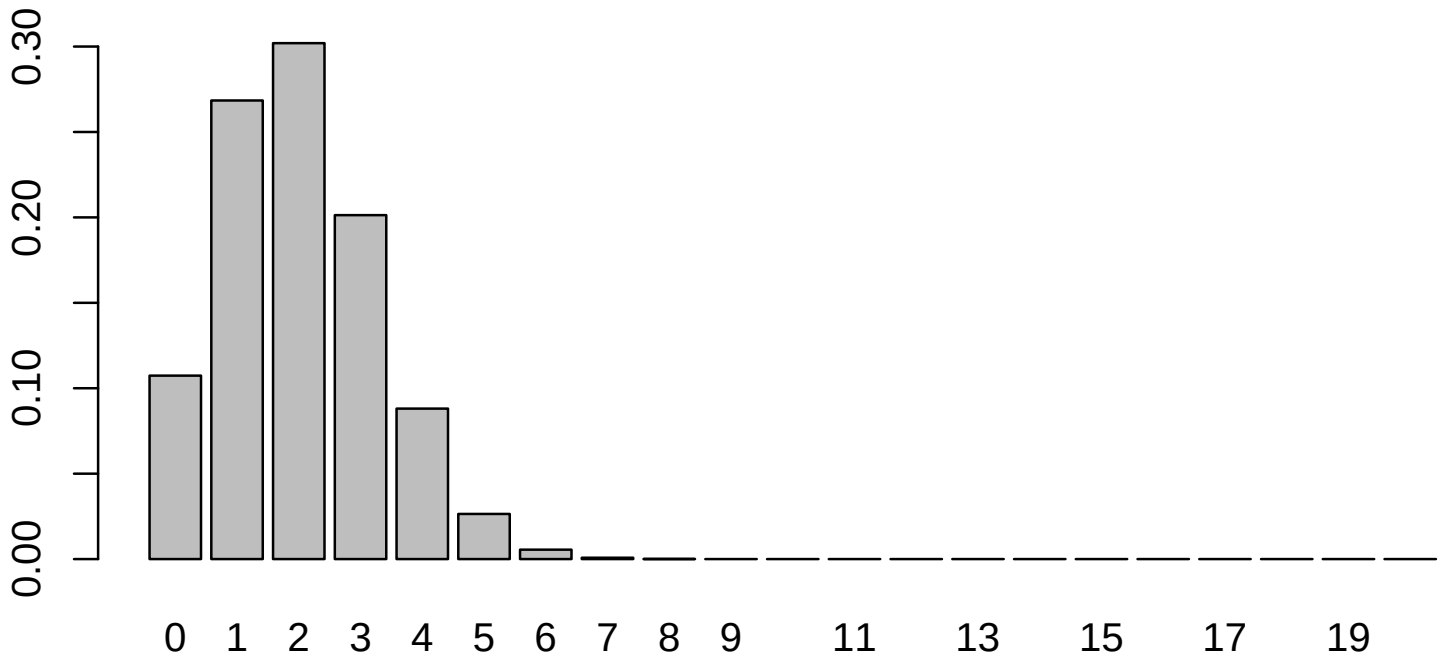
Clearly, this is a symmetric distribution.

Simulation of Sample Mean for $n = 1, 2, \ldots 10$:
$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \approx N(\mu, \frac{\sigma^2}{n}) = N(\frac{1}{2}, \frac{1}{12n})$



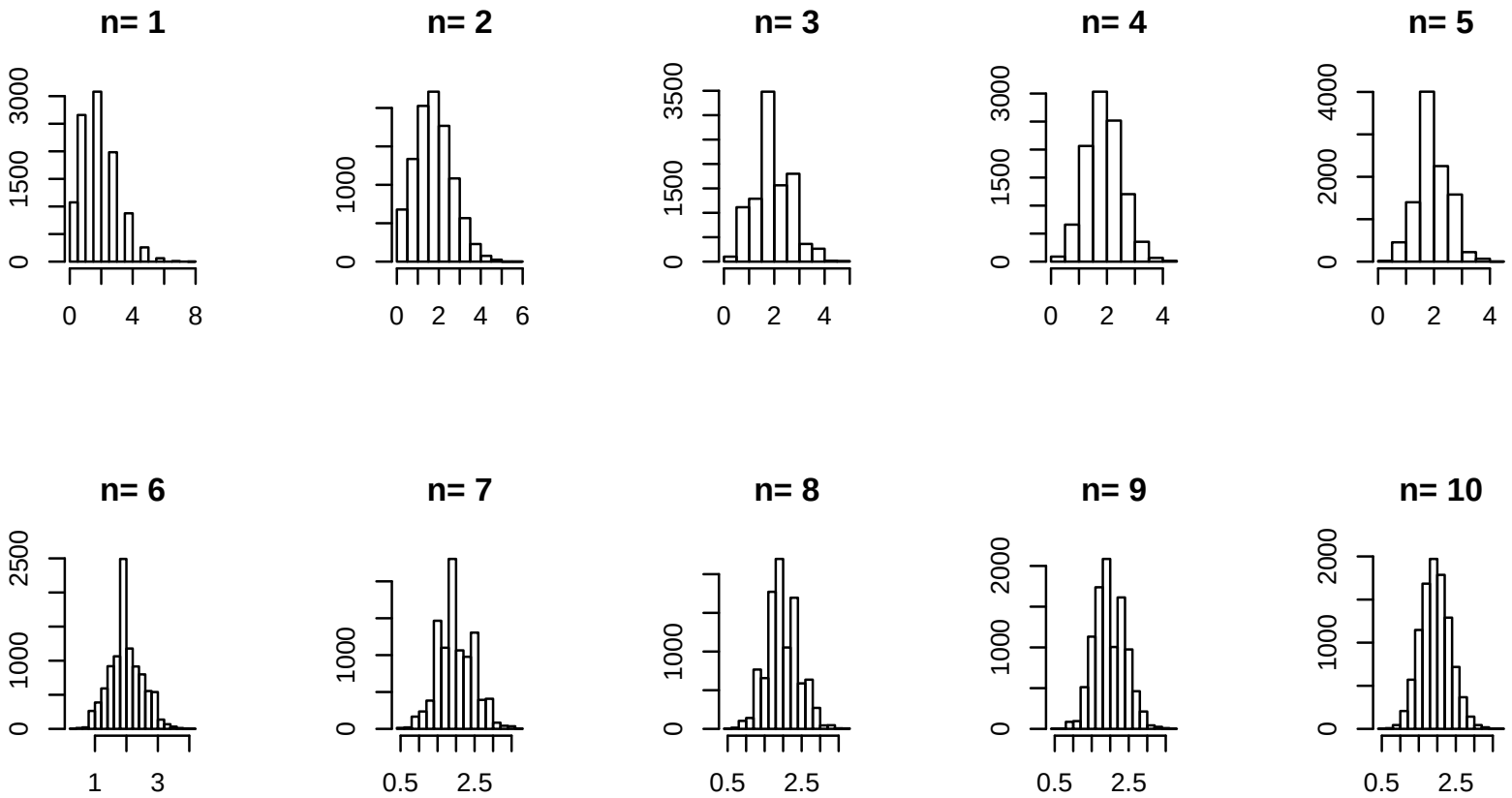
Given symmetry of X, $\bar{X}$ looks Normal for even $n = 5$.

(2) Binomial Distribution: $X \sim Bin(10, 0.2)$, with $\mu = 2$ and $\sigma^2 = 1.6$.



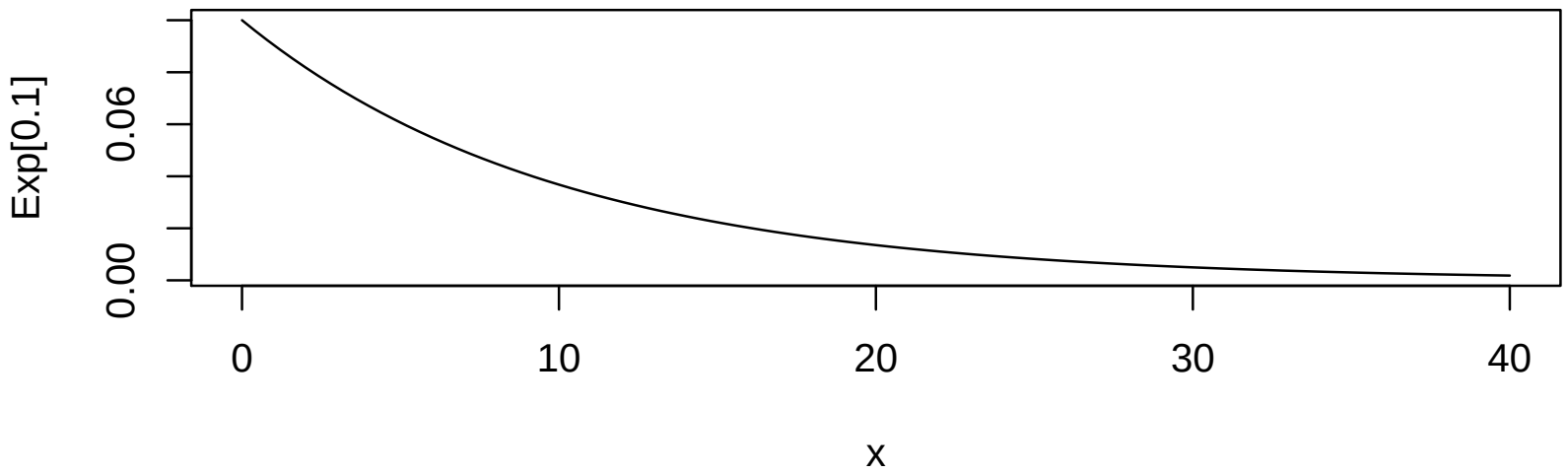
Clearly, this is a skewed distribution, as $p = 0.2$.

Simulation of Sample Mean for $n = 1, 2 \ldots, 10$:
$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \approx N(\mu, \frac{\sigma^2}{n}) = N(2, \frac{1.6}{n})$



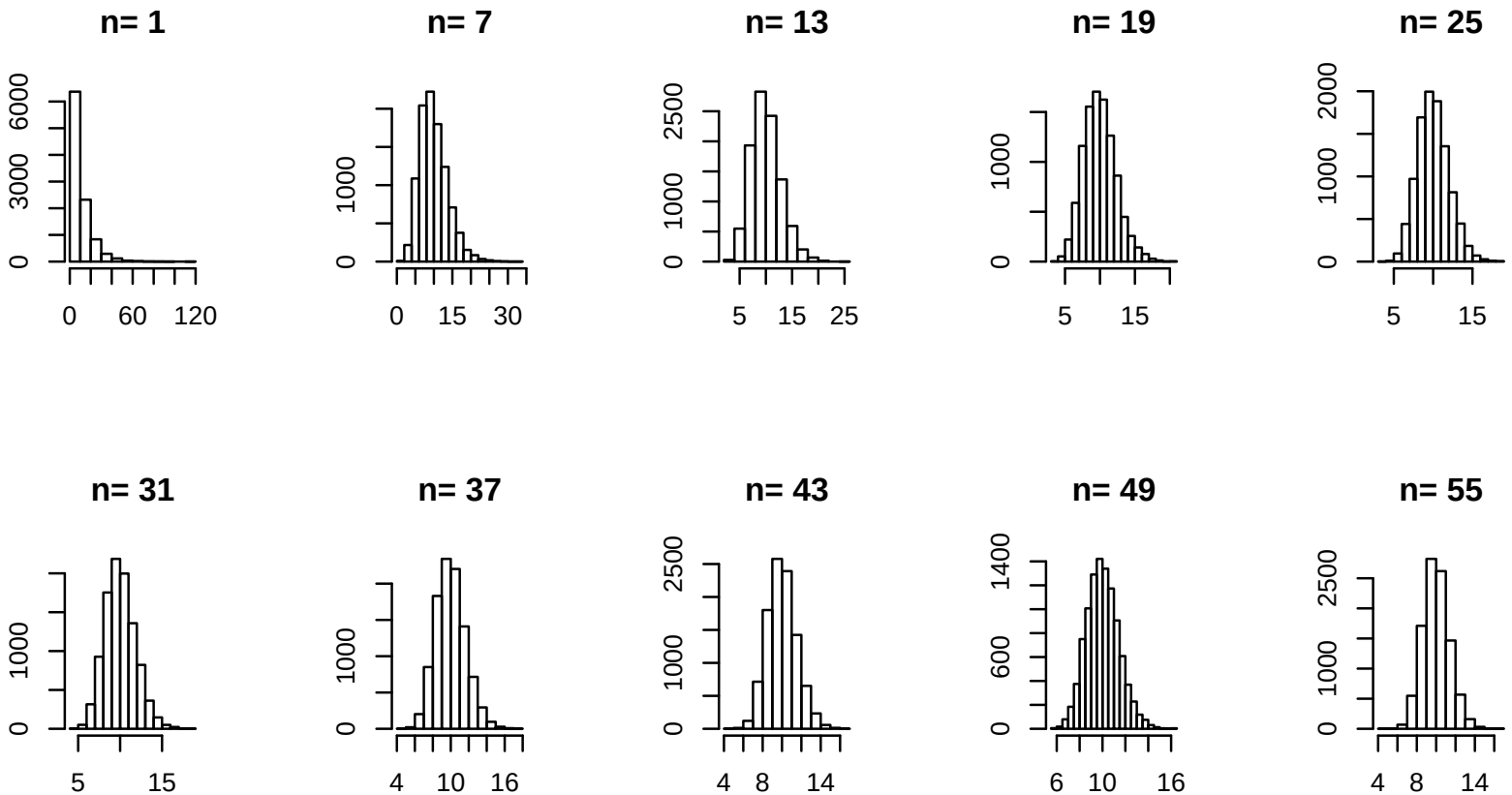
Notice, the approximation to Normal distribution, for about $n = 10$.

(3) Exponential Distribution: $X \sim Exp(0.1)$, with $\mu = 10$ and $\sigma^2 = 100$.



This is a highly skewed distribution.

Simulation of Sample Mean for $n = 50$:
$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \approx N(\mu, \frac{\sigma^2}{n}) = N(10, \frac{100}{n})$



Notice, the approximation to Normal distribution, for $n = 50+$.

Example: CLT

Assume that checked in luggage is highly skewed, as most people pack to the limit of 32kg, so $L \sim (31.9, 1^2)$. Find the probability that the average weight of the checked in luggage of 555 passengers and crew (independent) is over 32 kg.

Consider the 555 bags: $L_1, L_2, \ldots, L_{555}$ where
$L_i =$ checked in luggage $\sim (31.9, 1^2)$.

Assuming independence, let
$\bar{L} =$ Average Weight of the checked in luggage.

Using the CLT,

$$\bar{L} \approx N(\mu, \frac{\sigma^2}{n}) = N(31.9, \frac{1^2}{555}) = N(31.9, 0.04244764^2)$$

So using standardising,

$$P(\bar{L} > 32) = P(\frac{\bar{L} - 31.9}{0.04244764} > \frac{32 - 31.9}{0.04244764}) = P(Z > 2.355844) \approx 0$$

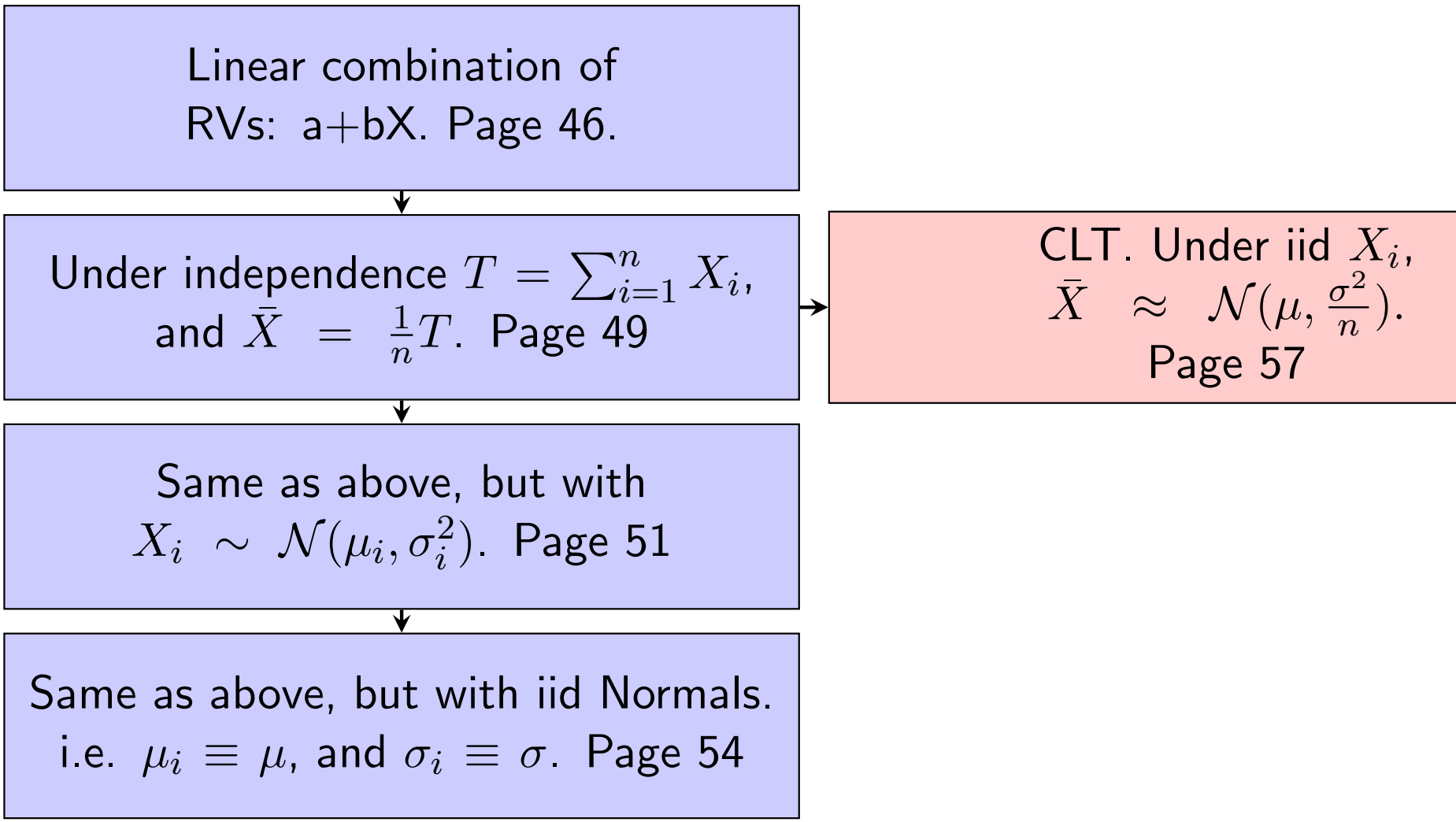
```
1-pnorm(32,31.9,0.04244764)
```

```
## [1] 0.009240349
```

```
1-pnorm(2.355844)
```

```
## [1] 0.009240338
```

Flowchart


Linear combination of RVs: a+bX. Page 46.
Under independence $T = \sum_{i=1}^{n} X_i$, and $\bar{X} = \frac{1}{n}T$. Page 49
CLT. Under iid X_i, $\bar{X} \approx \mathcal{N}(\mu, \frac{\sigma^2}{n})$. Page 57
Same as above, but with $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$. Page 51
Same as above, but with iid Normals. i.e. $\mu_i \equiv \mu$, and $\sigma_i \equiv \sigma$. Page 54