Purposeful Academic Classes for Excelling Students Program

(Department of Education, Western Australia)

Mathematics Methods Units 3 & 4

Session 4

General continuous random variables

4.2.1 use relative frequencies and histograms obtained from data to estimate probabilities associated with a continuous random variable

4.2.2 examine the concepts of a probability density function, cumulative distribution function, and probabilities associated with a continuous random variable given by integrals; examine simple types of continuous random variables and use them in appropriate contexts

4.2.3 identify the expected value, variance and standard deviation of a continuous random variable and evaluate them using technology

4.2.4 examine the effects of linear changes of scale and origin on the mean and the standard deviation

Normal distributions

4.2.5 identify contexts, such as naturally occurring variation, that are suitable for modelling by normal random variables

4.2.6 identify features of the graph of the probability density function of the normal distribution with mean μ and standard deviation σ and the use of the standard normal distribution

4.2.7 calculate probabilities and quantiles associated with a given normal distribution using technology, and use these to solve practical problems

Random sampling

4.3.1 examine the concept of a random sample

4.3.2 discuss sources of bias in samples, and procedures to ensure randomness

4.3.3 use graphical displays of simulated data to investigate the variability of random samples from various types of distributions, including uniform, normal and Bernoulli

Sample proportions

4.3.4 examine the concept of the sample proportion $\hat{p}$ as a random variable whose value varies between samples, and the formulas for the mean p and standard deviation $\sqrt{\frac{p(1-p)}{n}}$ of the sample proportion $\hat{p}$

4.3.5 examine the approximate normality of the distribution of $\hat{p}$ for large samples

4.3.6 simulate repeated random sampling, for a variety of values of p and a range of sample sizes, to illustrate the distribution of $\hat{p}$ and the approximate standard normality of $\frac{\hat{p}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$ where the closeness of the approximation depends on both n and p

Confidence intervals for proportions

4.3.7 examine the concept of an interval estimate for a parameter associated with a random variable

4.3.8 use the approximate confidence interval $\left(\hat{p}-z\sqrt{\left(\frac{\hat{p}(1-\hat{p})}{n}\right)}, \hat{p}+z\sqrt{\left(\frac{\hat{p}(1-\hat{p})}{n}\right)}\right)$ as an interval estimate for p, where z is the appropriate quantile for the standard normal distribution

4.3.9 define the approximate margin of error $E = z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ and understand the trade-off between margin of error and level of confidence

4.3.10 use simulation to illustrate variations in confidence intervals between samples and to show that most, but not all, confidence intervals contain p

Continuous Random Variables

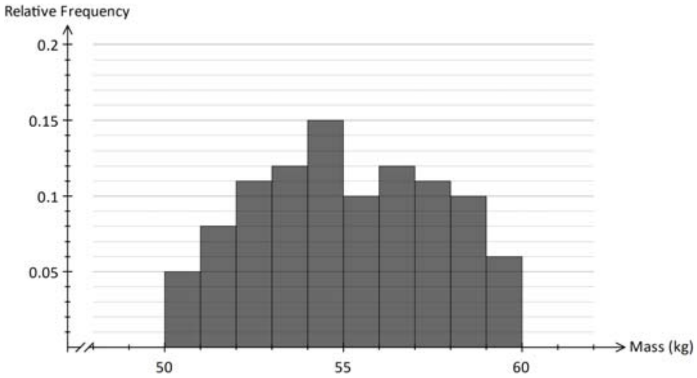
- The function $f(x)$ is defined as the *probability density function* for X where $a \le x \le b$ if:
  - $f(x) \ge 0 \qquad a \le x \le b$
  - $\int_a^b f(x)\,dx = 1$
- If $f(x)$ is the probability density function (pdf) for X then:

$$P(m \le X \le n) = \int_m^n f(x)\,dx \qquad \text{where } a \le m \le n \le b$$

- For a continuous random variable with probability density function $f(x)$ defined for $a \le x \le b$:
  - The mean is given by $\mu = E(X) = \int_a^b x\, f(x)\,dx$
  - The variance is given by $\text{Var}(X) = \int_a^b (x-\mu)^2 f(x)\,dx = \int_a^b x^2 f(x)\,dx - \mu^2$

Worked Example 1 **Calculator Free**

The *continuous* random variable T is defined as the mass of a group of students. The diagram below shows the relative frequency histogram for T for this group of students.



(a) Estimate the probability that a randomly selected student has mass between 55 kg and 58 kg.

$$\boxed{P(55 < T < 58) = 0.1 + 0.12 + 0.11 = 0.33}$$

(b) Estimate the probability that a student randomly selected from those with mass between 55 kg and 58 kg has mass 57 kg.

$$\boxed{P(T = 57 \mid 55 < T < 58) = 0}$$

(c) Estimate the probability that a student with mass exceeding 55kg has mass exceeding 58 kg.

$$\begin{aligned} P(T \geq 58 \mid T \geq 55) &= \frac{P(T \geq 58)}{P(T \geq 55)} \\ &= \frac{0.1 + 0.06}{0.33 + 0.1 + 0.06} \\ &= \frac{0.16}{0.49} = \frac{16}{49} \end{aligned}$$

(d) Given that there are 100 students in this group, calculate the probability that two randomly chosen students will have mass between 55 kg and 56 kg.

Required probability = $\frac{10}{100} \times \frac{9}{99} = \frac{1}{110}$

(e) The mass within each interval of the relative frequency histogram is uniformly distributed. The 90th percentile for T lies between 58 kg and 59 kg.
Estimate the value of k, the 90th percentile for T

$P(T < k) = 0.9$

$P(T < k) = P(T < 58) + P(58 < T < k)$

$0.9 = 0.84 + P(58 < T < k)$

$P(58 < T < k) = 0.06$

But $P(58 < T < 59) = 0.1$

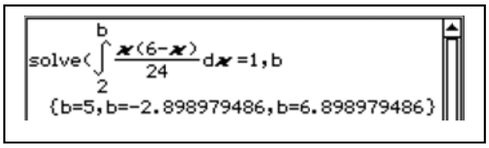
Hence, $\frac{k - 58}{1} = \frac{0.06}{0.1} \Rightarrow k = 58.6$

Worked Example 2 **Calculator Assumed**

The random variable X has probability density function $f(x) = \dfrac{x(6-x)}{24}$ for $2 \le x \le b$.

(a) Determine the value of *b*.

$$\int_2^b \frac{x(6-x)}{24}\,dx = 1$$



```
        b
solve(∫ x(6-x)  dx=1,b
         -----
          24
        2
{b=5,b=-2.898979486,b=6.898979486}
```

Use “solve command:

$b = -2.8990,\ 5,\ 6.8990$

Reject $b = -2.8990$ as $b > 2$. Reject $b = 6.8990$ as $f(6.8990) < 0$.

Hence, $b = 5$

(b) Calculate the μ and σ, respectively the mean and standard deviation for X.

$$\text{E(X)} = \int_2^5 \frac{x(6-x)}{24} \times x\ dx = \frac{109}{32}$$

$$\text{VAR(X)} = \int_2^5 \frac{x(6-x)}{24} \times \left(x - \frac{109}{32}\right)^2\ dx = \frac{3507}{5120}$$

STD(X) ≈ 0.8276

(c) Given that $Y = \dfrac{X-\mu}{\sigma}$, determine the mean and standard deviation for *Y*.

$$Y = \frac{1}{\sigma}X - \frac{\mu}{\sigma}$$

$$\text{E}(Y) = \text{E}\left(\frac{1}{\sigma}X - \frac{\mu}{\sigma}\right) = \frac{1}{\sigma}\text{E}(X) - \frac{\mu}{\sigma}$$

$$= \frac{\mu}{\sigma} - \frac{\mu}{\sigma} = 0.$$

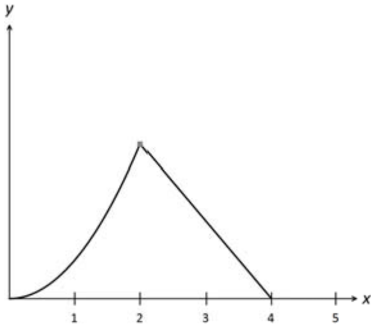
$$\text{STD}(Y) = \text{STD}\left(\frac{1}{\sigma}X - \frac{\mu}{\sigma}\right) = \left|\frac{1}{\sigma}\right|\text{STD}(X)$$

$$= \frac{1}{\sigma} \times \sigma = 1$$

Worked Example 3 **Calculator Assumed**

The graph of the probability density function of the random variable X is shown in the accompanying diagram with equation:

$$f(x) = \begin{cases} kx^2 & 0 \le x < 2 \\ -0.3x + 1.2 & 2 \le x \le 4 \end{cases}.$$



(a) Calculate the value of *k*.

$$k\int_0^2 x^2 \; dx + \int_2^4 -0.3x + 1.2 \; dx = 1$$

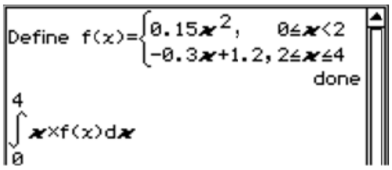
$$\frac{8k}{3} + \frac{3}{5} = 1$$

$$\Rightarrow \quad k = 0.15$$

(b) Determine the mean for X.

$$\text{E(X)} = 0.15\int_0^2 x^3 \; dx + \int_2^4 -0.3x^2 + 1.2x \; dx$$

$$= \frac{3}{5} + \frac{8}{5} = \frac{11}{5}$$



Define $f(x) = \begin{cases} 0.15x^2, & 0 \le x < 2 \\ -0.3x+1.2, & 2 \le x \le 4 \end{cases}$

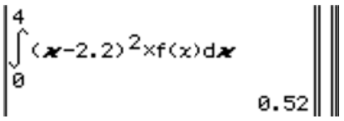
done

$\int_0^4 x \times f(x)dx$

(c) Determine the variance for X.

$$\text{VAR(X)} = 0.15\int_0^2 x^2 \times \left(x - \frac{11}{5}\right)^2 \; dx + \int_2^4 (-0.3x + 1.2) \times \left(x - \frac{11}{5}\right)^2 \; dx$$

$$= \frac{32}{125} + \frac{33}{125} = \frac{13}{25}$$



$\int_0^4 (x-2.2)^2 \times f(x)dx$

0.52

(d) Determine with reasons, if the mean is equal to, less than or greater than the median.

$$0.15\int_0^2 x^2 \; dx = 0.4$$

$$\int_2^m -0.3x^2 + 1.2x \; dx = 0.5 - 0.4$$

$$\Rightarrow \quad m \approx 2.17$$

Hence, the mean > median.

Worked Example 4 **Calculator Assumed**

The cumulative density function for a random variable X is given by:

$$P(X \le x) = 0.005x^2 + 0.05x \qquad \text{for } 0 \le x \le 10.$$

(a) Calculate $P(2 \le X \le 5)$.

$$\begin{aligned} P(2 \le X \le 5) &= P(X \le 5) - P(X \le 2) \\ &= 0.375 - 0.12 \\ &= 0.255 \end{aligned}$$

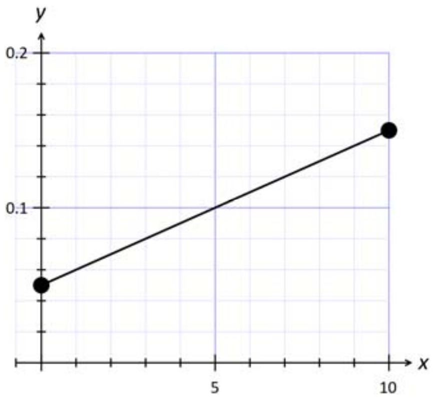
(b) Calculate the exact value of the median for X.

$$0.005x^2 + 0.05x = 0.5$$
$$x = -5 + 5\sqrt{5}$$

(c) Determine the probability density function for X.

$$f(x) = \frac{d}{dx}(0.005x^2 + 0.05x)$$
$$= 0.01x + 0.05 \text{ for } 0 \le x \le 10$$

(d) Sketch the graph of the probability density function for X.
Use this graph to determine with reasons whether the mean is more likely to be greater than, equal to or less than the median for X.



Mean < Median
as graph is skewed to the left.

(e) 25 observations on X was taken. Write a mathematical expression for the probability that exactly 10 of these observations have values of X less than the median.
Do not evaluate this expression.

Define Y: No. of observations with $X \le$ median.

$P(X \le \text{median}) = 0.5$

$$P(Y = 10) = \binom{25}{10} \times 0.5^{25}$$

Worked Example 5 **Calculator Assumed**

A random variable X has probability distribution given by $f(x) = \dfrac{1+x^2}{6}$ for $-1 \le x \le 2$.

(a) Determine the *exact* expected value of X.

$$\text{E(X)} = \frac{1}{6}\int_{-1}^{2} (1+x^2)\,x\,dx = \frac{7}{8}$$

(b) Calculate the *exact* variance for X.

$$\text{Var(X)} = \frac{1}{6}\int_{-1}^{2} (1+x^2)\left(x-\frac{7}{8}\right)^2 dx = \frac{267}{320}$$

The probability distribution for random variable T is given in the table below.

t	–1	0	1
P(T = t)	P(X ≤ 0) $\frac{1}{6}\int_{-1}^{0} (1+x^2)\,dx = \frac{2}{9}$	P(0 < X ≤ 1) $\frac{1}{6}\int_{0}^{-1} (1+x^2)\,dx = \frac{2}{9}$	P(1 < X ≤ 2) $= \frac{5}{9}$

(c) Complete the table above, giving the numerical values for the probability distribution for T.

(d) The variables T_1 and T_2 each have the same distribution as T.
Calculate the probability that the sum of the values for T_1 and T_2 is –1.

$$\begin{aligned}\text{P(Sum} = -1) &= \text{P}(T_1 = 0 \cap T_2 = -1) \\ &\quad + \text{P}(T_1 = -1 \cap T_2 = 0) \\ &= \frac{2}{9} \times \frac{2}{9} \times 2 = \frac{8}{81}\end{aligned}$$

(e) The variables T_1, T_2 and T_3 each have the same distribution as T.
Calculate the probability that the sum of the values for T_1, T_2 and T_3 is 1.

P(Sum = –1) = P(two “0” and one “1”) + P(two “1” and one “–1”)

$$= \frac{2}{9} \times \frac{2}{9} \times \frac{5}{9} \times 3 + \frac{5}{9} \times \frac{5}{9} \times \frac{2}{9} \times 3$$

$$= \frac{210}{729} = \frac{70}{243}$$

Uniform Distribution

- If the continuous random variable *X* is uniformly distributed in the interval $a \le x \le b$, then the probability density function of *X* is given by $f(x) = \dfrac{1}{(b-a)}$ for $a \le x \le b$.

  - The mean or expected value of X is $\dfrac{a+b}{2}$.
  - The variance for X is $\dfrac{(b-a)^2}{12}$.

Worked Example 6 **Calculator Assumed**

A continuous random variable X has probability density function $f(x) = 0.1$ for $a \le x \le b$, and $P(X \le 5) = 0.4$. Find:

(a) the values of *a* and *b*

$$\boxed{\begin{aligned} &P(X \le 5) = 0.4 \;\Rightarrow\; (5 - a) \times 0.1 = 04 \\ &\qquad\qquad a = 1 \;\Rightarrow\; b = 11 \end{aligned}}$$

(b) Q1 and Q3 respectively the lower and upper quartiles

$$\boxed{\begin{aligned} &P(X \le Q1) = 0.25 \;\Rightarrow\; (Q1 - 1) \times 0.1 = 0.25 \\ &\qquad\qquad Q1 = 3.5 \\ &P(X \ge Q3) = 0.25 \;\Rightarrow\; (11 - Q3) \times 0.1 = 0.25 \\ &\qquad\qquad Q3 = 8.5 \end{aligned}}$$

Worked Example 7 **Calculator Free**

The random variable X is uniformly distributed in the interval $a \le x \le 200$ where $a < 200$.

(a) State the probability density function for X.

$$\boxed{f(x) = \frac{1}{200-a} \quad \text{where } a \le x \le 200}$$

(b) Calculate the mean for X if $P(X < 50) = 0.7$.

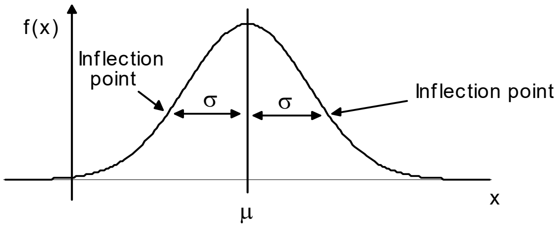
$$\boxed{\begin{aligned} &P(X < 50) = 0.7 \;\Rightarrow\; P(50 \le X \le 200) = 0.3 \\ &\qquad 150 \times \frac{1}{200-a} = 0.3 \;\Rightarrow\; a = -300 \\ &\text{Hence, mean for X} = \frac{-300+200}{2} = -50 \end{aligned}}$$

(c) Given that $P(X > 4a) = 4 \times P(X < 5a)$, calculate $P(X > 6a)$.

$$\boxed{\begin{aligned} &P(X > 4a) = 4 \times P(X < 5a) \\ &\frac{200-4a}{200-a} = 4 \times \frac{5a-a}{200-a} \;\Rightarrow\; a = 10 \\ &P(X > 6a) = P(X > 60) = \frac{14}{19} \end{aligned}}$$

The Normal Distribution

- The graph of the pdf of a normal variable X with mean μ and standard deviation σ is bell-shaped and is symmetrical about the mean μ. The inflection points are located at a distance of σ on either side of the mean. Clearly the larger the value of σ, the "wider" the curve.



- If $\mu = 0$ and $\sigma = 1$, then, X is a *standard* normal variable (represented by the letter Z)
- If $X \sim N(\mu,\sigma^2)$ then $\dfrac{X-\mu}{\sigma} \sim N(0, 1)$.
- The empirical rule for normal distributions.
  In a normal distribution approximately:
  68% of values lie within one standard deviation of the mean
  95% of values lie within two standard deviations of the mean
  99.7% of values lie within three standard deviations of the mean.

Worked Example 8 **Calculator Free**

In a normal distribution approximately:

68% of values lie within one standard deviation of the mean
95% of values lie within two standard deviations of the mean
99.7% of values lie within three standard deviations of the mean.

The mass of chocolate bars produced in a factory is normally distributed with mean 4.5 kg and standard deviation 20 g.

(a) Calculate the probability that a randomly chosen chocolate bar produced in this factory has mass in excess of 4540 g.

$$P(X > 4540) = P(Z > 2) = 0.025$$

(b) Determine the probability that a chocolate bar from this factory with mass less than 4540 g has mass in excess of 4520 g.

$$P(X > 4520 \mid X < 4540) = P(Z > 1 \mid Z < 2) = \frac{P(1<Z<2)}{P(Z<2)}$$
$$= \frac{0.135}{0.975} = \frac{135}{975}$$

(c) Determine the 84th percentile mass of these chocolate bars.

$P(X < k) = 0.84$
$P(Z < 1) = 0.5 + 0.34$
Hence $k = 4500 + 20 = 4520$ g

Worked Example 9 **Calculator Assumed**

The packed cell volume (PCV) or haematocrit (HCT) is the percentage of blood volume that is occupied by red blood cells. The PCV of average human males is normally distributed with mean 45 and standard deviation 1.5. A measured level of PCV above 50 would indicate that the male has undergone blood doping, that is, he tests positive for blood doping.

(a) Find the probability that a randomly selected male would test positive for blood doping.

X: PCV of adult male

$X \sim N(45, 1.5^2)$.

$P(X > 50) = 0.0004291$

(b) Find the probability that in two separate tests, a randomly selected male would test positive twice.

Prob. $= 0.0004291 \times 0.0004291 = 0.0000001841$

(c) John, a male athlete practices blood doping so that his PCV has a mean of 48 with standard deviation 1.2. Find the probability that in two tests, John will test positive for blood doping twice.

Y: John's PCV

$Y \sim N(48, 1.2^2)$.

$P(Y > 50) = 0.04779$

Prob. tests positive twice $= 0.04779^2 = 0.002284$

(d) Bans are imposed if an athlete tests positive in two tests.
Use your answers in (b) and (c) to comment on the appropriateness of these bans.

There is about a 2 in 1000 chance of being caught for practicing blood doping.
There is a less than 2 in 10 000 000 chance of being incorrectly tested.
Hence, it is 10 000 times more likely for a cheat to be caught.
Therefore, the ban seems appropriate.

(e) An athlete practices blood doping and raises his PCV to μ with standard deviation 1.2. Find μ if the probability of him testing positive to blood doping is to be less than 1%.

W: Athlete's PCV

$W \sim N(\mu, 1.2^2)$.

$$P(Y > 50) < 0.01 \;\Rightarrow\; P(Z > \frac{50-\mu}{1.2}) < 0.01$$

$$\frac{50-\mu}{1.2} = 2.32635 \;\Rightarrow\; \mu = 47.21$$

Worked Example 10 Calculator Assumed

An organic farm grows Brussels sprouts on a commercial scale. The diameter of Brussels sprouts is assumed to be normally distributed with mean 30 mm and standard deviation 4 mm. Sprouts with diameters over 35 mm are sold to a composting factory at \$0.20 per kg. Sprouts with diameters between 25 mm and 35 mm inclusive are sold to a supermarket chain at \$4.00 per kg. Sprouts with diameters less than 25 mm are sold to a frozen food factory for \$2.00 per kg.

(a) Estimate the total income received by the farm for a crop of Brussel sprouts of 30 tonnes.

X: diameter of Brussels sprouts

$X \sim N(30, 4^2)$

$P(X > 35) = 0.10565$

$P(25 \le X \le 35) = 0.7887$

$P(X < 25) = 0.10565$

Total income $= 0.10565 \times 30\,000 \times 0.20 + 0.7887 \times 30\,000 \times 4 + 0.10565 \times 30\,000 \times 2$
$= \$101\,616.90$

(b) To reduce the amount of sprouts having diameters greater than 35 mm, the sprouts may be harvested earlier so that the mean diameter is now *a* mm with the standard deviation remaining constant. Find the value of *a* if 5% of sprouts are to have diameters exceeding 35 mm.

Y: diameter of Brussels sprouts

$Y \sim N(a, 4^2)$

$P(Y > 35) = 0.05 \quad \Rightarrow \quad P(Z > \frac{35-a}{4}) = 0.05$

$\frac{35-a}{4} = 1.64485 \quad \Rightarrow \quad a = 28.4206 \approx 28.4$ mm

(c) Use your answer in (b) and the distribution details described at the start of this question, to determine if this will bring greater income to the farm. Assume that the prices remain the same.

W: diameter of Brussels sprouts

$W \sim N(32.4, 4^2)$

$P(W > 35) = 0.049472$

$P(25 \le W \le 35) = 0.752866$

$P(W < 25) = 0.197663$

Total income $= 0.049472 \times 30\,000 \times 0.20 + 0.752866 \times 30\,000 \times 4 + 0.197663 \times 30\,000 \times 2$
$= \$102\,500.53.$

Hence, Yes!

Worked Example 11 **Calculator Assumed**

The amount of sodium consumed daily is a normal variable with mean μ mg and standard deviation σ mg. 79.8% of adults consume at least 1900 mg of sodium daily.
25.4% of those who consume at least 1900 mg of sodium daily consume at least 2100 mg of sodium daily. Calculate the values of μ and σ.

Define X: Amount of sodium consumed

$$X \sim N(\mu, \sigma^2)$$

Given: $P(X \geq 1900) = 0.798$

$$\Rightarrow \quad P(Z \geq \frac{1900-\mu}{\sigma}) = 0.798$$

$$\frac{1900-\mu}{\sigma} = -0.834499 \qquad \text{I}$$

Given:

$$P(X \geq 2100 \mid X \geq 1900) = 0.254$$

$$P(X \geq 2100) = 0.254 \times 0.798$$

$$= 0.202692$$

$$\Rightarrow \quad P(Z \geq \frac{2100-\mu}{\sigma}) = 0.202692$$

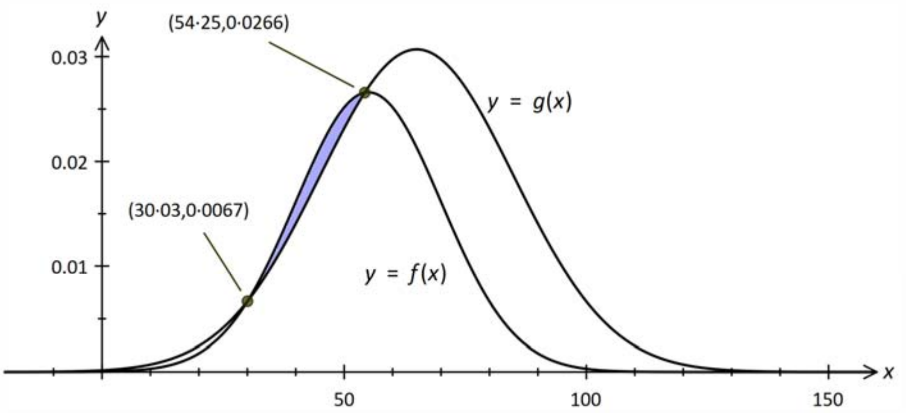
$$\frac{2100-\mu}{\sigma} = 0.832044 \qquad \text{II}$$

Solve I and II simultaneously:

$$\mu \approx 2000.1 \quad \sigma = 120.0$$

Worked Example 12 Calculator Assumed

The diagram below shows the graph of $y = f(x)$ and $y = g(\mathrm{x})$ where $f(x)$ is the probability density function of a normal random variable with mean 55 and standard deviation 15 and $g(x)$ is the probability density function of a normal random variable with mean 65 and standard deviation 13. The two graphs intersect at the points (30.03, 0.0067) and (54.25, 0.0266). The region R as shaded is trapped between these two curves.



(a) Determine with reasons if each of the following functions may be used as probability density functions of random variables.

(i) $y = f(2x)$

Yes.
$y = f(2x)$ is dilated along the x-axis by a factor or 0.5.

Hence, $y \geq 0$ for all x. Also $y = f(2x)$ will still be a normal curve and $\int_{-\infty}^{\infty} f(2x)\,dx = 1$.

(ii) $y = g(x) + 1$

No.

$$\int_{-\infty}^{\infty} g(x)+1\,dx = \int_{-\infty}^{\infty} g(x)\,dx + \int_{-\infty}^{\infty} 1\,dx \to \infty$$

(b) Calculate the area of region R.

Let X ~ N(55, 15^2) and Y ~ N(65, 13^2).

Area of R = P(30.03 ≤ X ≤ 54.25) – P(30.03 ≤ Y ≤ 54.25)
= 0.432072 – 0.200568
= 0.231504

The Exponential Distribution

- A random variable X with probability density function $f(x) = ke^{-kx}$ for $x > 0$ is said to be exponentially distributed with parameter *k*.
  - The mean of X is $\frac{1}{k}$ and the standard deviation for X is $\frac{1}{k}$.

Worked Example 13 **Calculator Assumed**

The time interval between two consecutive patients arriving at the Emergency Department of a Hospital between 10 pm and 3 am over a Friday evening may be modelled by the variable T with probability density function $f(t) = 0.2e^{-0.2t}$ for $t > 0$ minutes.

(a) Find the probability that the inter-patient arrival time is more than 3 minutes.

$$\mathrm{P}(T > 3) = \int_3^{\infty} 0.2e^{-0.2t}\ dt = 0.5488$$

(b) Find the probability that there is a wait of at least another 2 minutes before the next patient arrives given that it has been more than 1 minute since the arrival of the previous patient.

$$\mathrm{P}(T > 3 \mid T > 1) = \frac{\int_3^{\infty} 0.2e^{-0.2t}\ dt}{\int_1^{\infty} 0.2e^{-0.2t}\ dt}$$

$$= 0.6703$$

(c) Calculate the mean time between the arrivals of 2 consecutive patients and the associated standard deviation.

$$\mathrm{E}(T) = \int_0^{\infty} 0.2e^{-0.2t} \times t\ dt = 5 \text{ minutes}$$

$$\mathrm{Var}(T) = \int_0^{\infty} 0.2e^{-0.2t} \times (t-5)^2\ dt = 25$$

STD(T) = 5 minutes

(d) Calculate the mean and standard deviation for W = 60T and give a possible explanation of what these values represents.

E(W) = 60 × E(T) = 300
STD(W) = 60 X STD(T) = 300

E(W) and STD(W) represent the mean inter-patient time and its standard deviation measured in seconds.

Sampling Distributions of sample proportions

- Let p be the proportion of a population with a particular attribute A.
- The sampling distribution of sample proportions of sample size n has a mean of p and standard deviation $\sqrt{\dfrac{p(1-p)}{n}}$.
- As the sample size n increases, the sampling distribution of sample proportions $\hat{p}$ tends towards a <u>*normal distribution*</u> with:

  mean = population proportion p and standard deviation = $\sqrt{\dfrac{p(1-p)}{n}}$.
- For practical purposes, it is acceptable to treat the sampling distribution as normally distributed as long as the size of the sample $n \geq 30$ where $n\,p \geq 10$ and $n(1-p) \geq 10$.
- For a given sample size n, the sampling distribution has standard deviation no larger than $\dfrac{1}{2\sqrt{n}}$.

Worked Example 14 **Calculator Free**

Determine with reasons if it is possible to have a sampling distribution for proportion p of size 100 with a standard deviation of 0.5

$0.5 = \sqrt{\dfrac{p(1-p)}{100}} \Rightarrow 0.25 = \dfrac{p(1-p)}{100}$

$p^2 - p + 25 = 0$

$\Delta = 1 - 100 = -99 < 0$

Hence, there are no real values for p.

Therefore, impossible!

Worked Example 15 **Calculator Assumed**

Assume that students who are right handed are in the majority.

The percentage of year one students who are right handed is 100p%.

500 samples of 40 year one students were formed. The proportion of right handed students in each sample was calculated. The standard deviation for the set of 500 sample proportions calculated is 0.05. Use this result to calculate a reasonable estimate for p.

$\sqrt{\dfrac{p \times (1-p)}{40}} = 0.05$

$p = 0.1127$ or 0.8873

Hence, reasonable estimate for $p = 0.8873$

Worked Example 16 Calculator Assumed

Renal agenesis is a birth defect where a person is born with only one kidney. Let *p* be the proportion of persons in the community with renal agenesis. Sample A consists of 2 000 randomly chosen people and 3 were found to suffer from renal agenesis.

(a) Use sample A to calculate:

(i) a point (single-value) estimate for *p*.

Estimate for $p \approx \dfrac{3}{2000} = 0.0015$

(ii) correct to 4 significant figures, the standard deviation for the sampling distribution of $\hat{p}$, the mean of all sample proportions of sample size 2000.

$$\text{STD} = \sqrt{\frac{0.0015\times(1-0.0015)}{2000}}$$
$$\approx 0.000\,865\,4$$

(b) Use the sampling distribution for $\hat{p}$ suggested by sample A to estimate the probability that a randomly chosen sample of 2000 persons will have at least two persons who have this birth defect.

$\hat{p} \sim \text{N}(\mu = 0.0015, \sigma^2 = 0.000\,865\,4^2)$

$\text{P}(\hat{p} \geq \dfrac{2}{2000}) = 0.71828$

(c) Use an appropriate discrete variable suggested by sample A to calculate the probability that a randomly chosen sample of 2000 persons will have at least two persons who have this birth defect.

X: No. in sample of 2000 with renal agenesis.
$X \sim \text{B}(2000, 0.0015)$
$\text{P}(X \geq 2) = 0.8011$

(d) Determine with reasons if the answer in (b) or the answer in (c) would give a more accurate answer to the probability that a randomly chosen sample of 2000 persons will have at least two persons who have this birth defect.

Answer in (c) is more accurate.
Answer in (b) is uses a continuous distribution to estimate a discrete probability.

Worked Example 17 Calculator Assumed

X is a binomial variable with n = 10 and $p = \frac{2}{3}$. Samples of 500 observations on X are taken.

(a) Calculate P(X ≥ 8).

$\boxed{P(X \geq 4) = 0.299\,141}$

(b) Describe $\hat{p}$ the sampling distribution of sample proportions of observations on X with values of at least 8, for samples of size 500.

As sample size $n = 500 \geq 30$
and $np \approx 150 \geq 10$ and $n(1 - p) \approx 350 \geq 10$;
Sampling distribution is approximately normal.
Mean for sample proportions = 0.299 141

Standard deviation = $\sqrt{\dfrac{0.299141(1-0.299141)}{500}}$ = 0.020 477

(c) Calculate the probability that for a randomly chosen sample of size 500, the sample proportion of observations of X with values of at least 8, is no more than 0.3.

$\boxed{P(\hat{p} \leq 0.3) = 0.5167}$

(d) 2000 samples each with 500 observations on X was taken. How many of these samples are expected to have sample proportions of observations of X with values of at least 8, that is no more than 0.3?

$\boxed{\text{Expected number} = 2000 \times 0.5167 \approx 1033}$

Worked Example 18 **Calculator Assumed**

To determine the proportion of adults that own dogs, a sample of 100 adults were interviewed and 37 of these adults own dogs.

(a) Use the sample provided to determine an estimate for the proportion of adults that own dogs.

Estimate = 0.37

(b) A second sample of *N* adults were interviewed. Use your answer in (a) to calculate the value of *N* if the sampling distribution of the proportion of adults that own dogs (of sample size *N*) has a standard deviation not exceeding 0.05.

$$\sqrt{\frac{0.37\times0.63}{N}} \le 0.05$$

$$N \ge 93.2$$

Integer $N \ge 94$

(c) A third sample of 100 adults found that 8 own dogs. The adults in this sample were selected from residents in a 15 storey apartment block. Discuss the possible sources of bias in this sample.

Exclusion bias: No representatives from non-residents.
Selection bias: Sample taken from environment not conducive to owning dogs/rules against owing dogs
Response bias: Residents "refusing" to provide answer for fear of "getting into trouble" if they did own a dog assuming dogs are not permitted.

Point & Interval Estimates

- If the proportion *p* of attribute A in a population is unknown, then the calculated sample proportion of attribute A, $\hat{p}_0$, could be used to estimate *p*. When $\hat{p}_0$ is used in this sense, it is termed a *point estimate* of the population proportion *p*.
- The table below summarises the different confidence intervals for μ.

Confidence Level	Confidence interval
90%	$\hat{p} \pm 1.645\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$
95%	$\hat{p} \pm 1.960\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$
99%	$\hat{p} \pm 2.576\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$
100*c* %	$\hat{p} \pm z_c \times \sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$

Worked Example 19 **Calculator Free**

In a sample of 400 students from Perth, 320 indicated that they had never visited Geraldton, a town 420 km north of Perth. Let *p* represent the proportion of Perth students that have never visited Geraldton.

(a) Given that $P(-2 \le Z \le 2) = 0.954$ where $Z \sim N(0, 1)$, calculate a 95.4% confidence interval for the proportion of students in Perth that have never visited Geraldton.

Sample proportion = 0.8

CI: $0.8 \pm 2 \times \sqrt{\dfrac{0.8 \times 0.2}{400}}$

$0.8 \pm 2 \times \dfrac{0.4}{20} \Rightarrow 0.76 \le p \le 0.84$

(b) A second sample of 200 students from Perth, 165 students indicated that they had never visited Geraldton. Use your answer in (a) to determine if the second sample was statistically different from the first sample.

Sample proportion = 0.825 falls within the 95.4% CI.
Hence, there is no evidence to suggest that they are statistically different.

(c) A third sample of 200 students from Perth, yielded a 95.4% confidence interval of $0.80 \le p \le 0.90$. Use your answer in (a) to determine if the third sample was statistically different from the first sample.

Sample proportion = 0.85 falls outside the 95.4% CI.
Hence, there is evidence to suggest that they are statistically different.

Worked Example 20 **Calculator Free**

In a normal distribution approximately:

68% of values lie within one standard deviation of the mean
95% of values lie within two standard deviations of the mean
99.7% of values lie within three standard deviations of the mean.

Let *p* be the true proportion of students that spend more than 2 hours each day travelling between their homes and school.

(a) Of 36 students, 12 students spend more than 2 hours each day travelling between their homes and school. Use the empirical rule to determine the margin of error for a 99.7% confidence interval for *p*.

Sample proportion = $\frac{1}{3}$

Margin of error = $3 \times \sqrt{\dfrac{\frac{1}{3} \times \frac{2}{3}}{36}} = \dfrac{\sqrt{2}}{6}$

(b) Using the empirical rule, a 95% confidence interval for *p* was found to be $\frac{1}{4} - \frac{\sqrt{3}}{20} \leq p \leq \frac{1}{4} + \frac{\sqrt{3}}{20}$. Determine the size of the sample used.

Sample proportion = $\frac{1}{4}$

Margin of error = $2 \times \sqrt{\dfrac{\frac{1}{4} \times \frac{3}{4}}{n}} = \dfrac{\sqrt{3}}{20}$

$$\sqrt{\frac{3}{16n}} = \frac{\sqrt{3}}{40}$$

$$\sqrt{16n} = 40$$

$$n = 100$$

(c) Twenty samples of 400 students were taken. From each sample, a 90% confidence interval for *p* is calculated so that a set of twenty 90% confidence intervals is obtained.

(i) How many of these twenty 90% confidence intervals are expected to contain *p*?

Expected number = 20 × 0.9 = 18

(ii) Calculate the probability that all twenty of these intervals will contain *p*.

Probability = 0.9^{20}

Worked Example 21 Calculator Assumed

Let *p* represent the true proportion of an attribute in a population.

(a) A confidence interval for *p* is given by $\frac{13}{75} \le p \le \frac{17}{75}$. Determine the confidence level of this interval if the sample used to calculate this interval has size 1125.

Sample proportion = $\frac{15}{75} = \frac{3}{15}$

Margin of error = $\frac{2}{75}$

Hence: $z \times \sqrt{\frac{\frac{3}{15}\times(1-\frac{3}{15})}{1125}} = \frac{2}{75} \quad \Rightarrow \quad z = \sqrt{5}$

$\text{P}(-\sqrt{5} \le Z \le \sqrt{5}) = 0.9746$

Hence, confidence level ≈ 97.5%

(b) Using the same confidence level and sample proportion as in question (a), determine the percentage change in width of the confidence interval for *p* if the sample size is quadrupled (4 times larger)

Margin of error = $\sqrt{5} \times \sqrt{\frac{\frac{3}{15}\times(1-\frac{3}{15})}{1125\times 4}} = \frac{1}{75}$

Hence, the width of the interval is reduced by 50%.

(c) Determine if the result in part (b) describing the width of a confidence interval of a specific level and the size of the sample used may be applied to all confidence intervals for *p*.

Ratio of margin of errors = $\dfrac{z\times\sqrt{\frac{p\times(1-p)}{n}}}{z\times\sqrt{\frac{p\times(1-p)}{4n}}} = \frac{1}{2}$

⇒ The width of the interval will always be reduced by 50% if the sample size is quadrupled.

Worked Example 22 **Calculator Assumed**

Let p be the proportion of households that have been burgled at least once.

(a) A sample of 400 households was randomly selected and the 100c% confidence interval for p was $0.68 \le p \le 0.78$.

(i) How many households in this sample had been burgled at least once?

Sample proportion = $\dfrac{0.68+0.78}{2}$ = 0.73

$\Rightarrow$ No. of households = 292

(ii) Calculate the confidence level for the confidence interval calculated.

Margin of error $z \times \sqrt{\dfrac{0.73(1-0.73)}{400}} = 0.05$

$z = 2.252458$

$\text{P}(-2.252458 \le Z \le 2.252458) = 0.9757$

Hence, confidence level ≈ 97.6%.

(b) Assume that p = 0. 8. How large should a sample be, if the lower limit of a 99% confidence interval for p is to be at least 0.7?

Margin of error $2.576 \times \sqrt{\dfrac{0.8(1-0.8)}{N}} \le 0.1$

$N \ge 106.2$

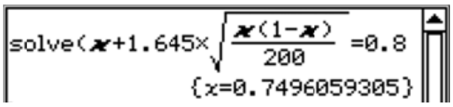
Integer $N \ge 107$

(c) In a sample of 200 households, the proportion of households that have been burgled at least once was p_0. For this sample, the upper limit of the 90% confidence interval for p was 0.8. Calculate the value of p_0.

Margin of error = $1.645 \times \sqrt{\dfrac{p_0 \times (1-p_0)}{200}}$

Hence: $p_0 + 1.645 \times \sqrt{\dfrac{p_0 \times (1-p_0)}{200}} = 0.8$

$p_0 = 0.7496 \approx 0.75$



```
solve(x+1.645×√(x(1-x)/200)=0.8
                    {x=0.7496059305}
```

Worked Example 23 **Calculator Assumed**

Let p be the proportion of teachers that are left handed.

(a) A sample of n (where $n > 100$) teachers were surveyed. A 99 % confidence interval for p from this sample was $0.2624 \le p \le 0.3696$.

(i) Use this confidence interval to estimate the mean and standard deviation of the sampling distribution of sample size n for the proportions of left handed teachers.

Mean of sample proportion = $\dfrac{0.2624 + 0.3696}{2}$ = 0.316

Margin of Error = 0.3696 – 0.316 = 0.0536

$\Rightarrow \quad 2.576 \times sd = 0.0536$

$\qquad\qquad sd = 0.0208$

(ii) Calculate the probability that in a randomly chosen sample of n teachers more than 35% are left-handed.

As $n > 100$, sampling distribution is approximately normal.

$\hat{p} \sim \text{N}(0.316, 0.0208^2)$

Hence, $\text{P}(\hat{p} > 0.35) = 0.05106$

(b) A sample of n (where $n > 100$) teachers were surveyed. A 100c % confidence interval for p from this sample was $0.3161 \le p \le 0.4039$. The probability that a randomly chosen sample of size n has no more than 185 teachers who are left handed is 0.6793. Calculate n.

sample proportion = $\dfrac{0.3161 + 0.4039}{2}$ = 0.36

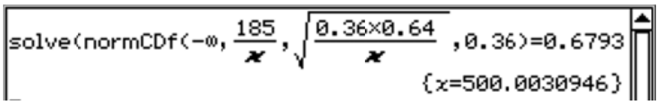
Margin of error = $\sqrt{\dfrac{0.36 \times 0.64}{n}}$

As $n > 100$, sampling distribution is approximately normal.

$\hat{p} \sim \text{N}\left(0.36, \dfrac{0.36 \times 0.64}{n}\right)$

$\text{P}\left(\hat{p} < \dfrac{185}{n}\right) = 0.6793$

$n = 500$



$\text{solve}\left(\text{normCDf}\left(-\infty, \dfrac{185}{x}, \sqrt{\dfrac{0.36\times0.64}{x}}, 0.36\right)=0.6793\right)$

$\{x=500.0030946\}$