6. CONTINUOUS RANDOM VARIABLES, NORMAL, AND SAMPLING DISTRIBUTIONS

Continuous random variables

- As mentioned in previous chapters, quantitative data can be classified as discrete or continuous.

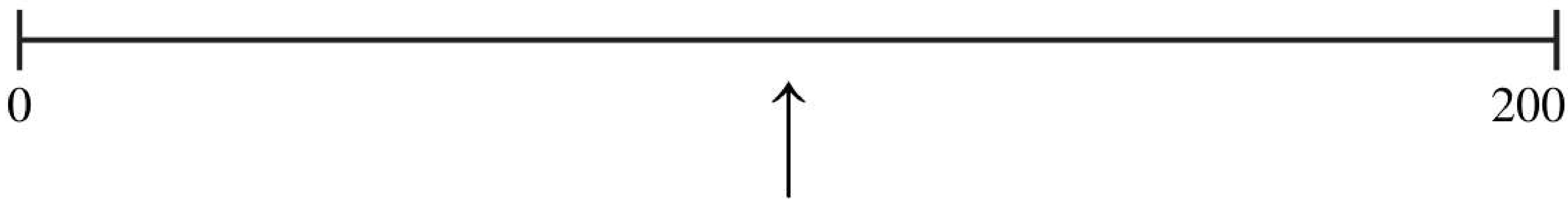
- In the previous chapter, we mentioned that discrete random variables are random variables whose values are countable.

- Continuous random variables are random variables whose values are not countable.

- For continuous random variables, their values can assume any value over an interval or intervals.

Continuous random variables

UNIVERSITI KEBANGSAAN MALAYSIA
The National University of Malaysia



Every point on this line represents a possible value of x that denotes the life of a battery. There is an infinite number of points on this line. The values represented by points on this line are uncountable.

- Examples of continuous random variables:
  - The length of a room
  - The time taken to commute from home to work
  - The weight of a student
  - The price of a house

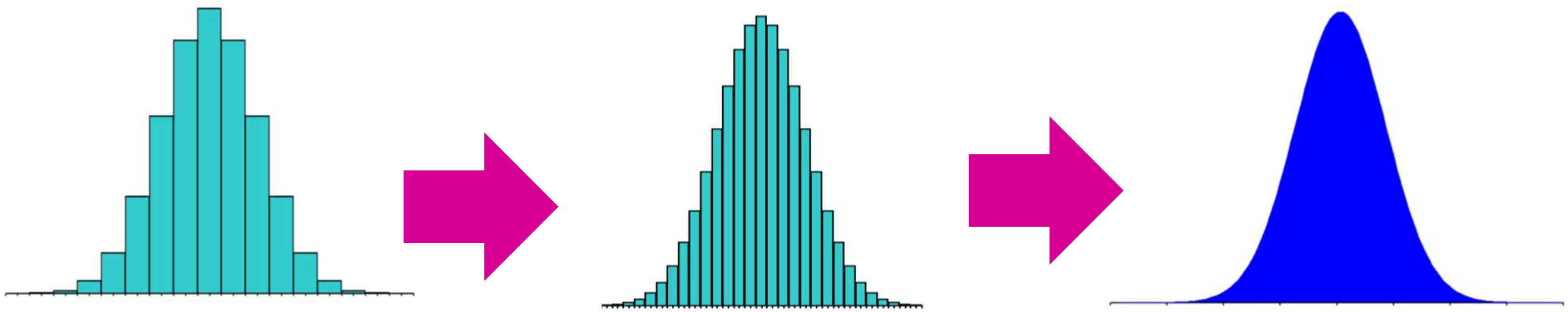
Probability distribution for continuous random variables

- Suppose we have a continuous data, and we construct a frequency table (like we did in Chapter 2).

- From the frequency table, we plot the histogram of the data.

- Now, suppose we modify the frequency table such that the class interval is lower and the number of class is bigger.

Probability distribution for continuous random variables



Decreasing class intervals or increasing number of classes

- Then we will get something like the figure above.

- As the class intervals get smaller, the histogram will converge to a curve.

Probability distribution for continuous random variables

- The probability distribution for a continuous random variable X is a function $f(x)$ that describes the distribution of the random variable X.

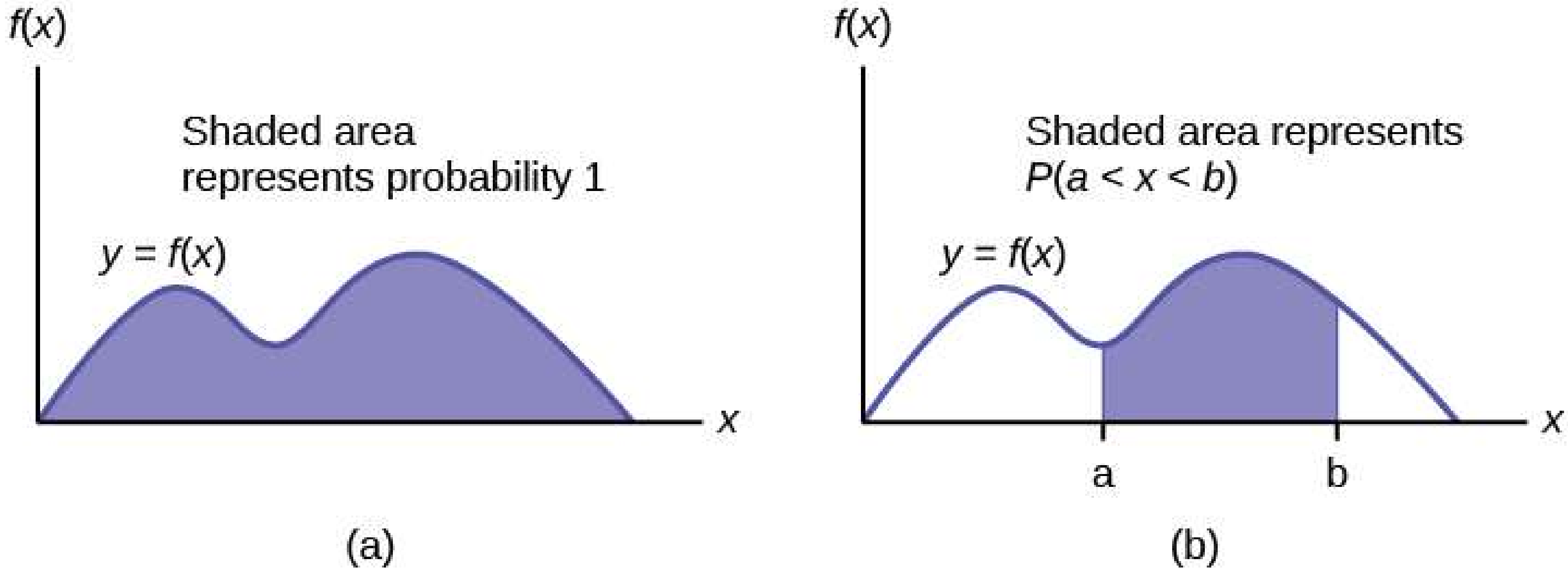
- For continuous random variables, the probability distribution is also called the probability density function (pdf).

- Two properties for probability distribution for continuous random variables:
  - $f(x) \geq 0$
  - The total area under the curve $f(x)$ is one. $\left(\int_{-\infty}^{\infty} f(x)\, dx = 1\right)$

- The probability of X in an interval is the area under the density curve over the interval.

$$P(a < X < b) = \int_a^b f(x)\, dx$$

Example



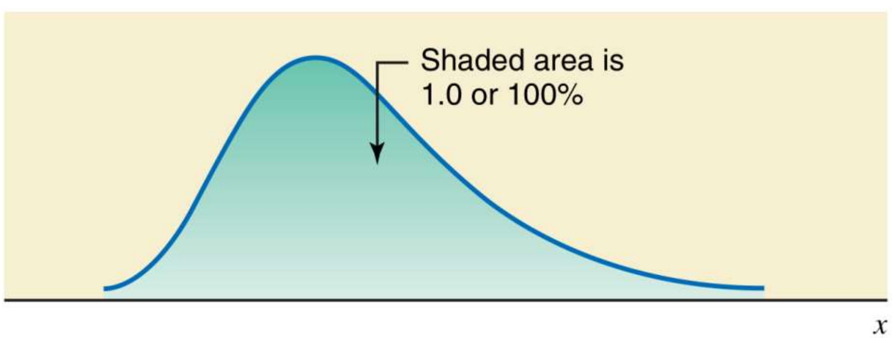


Figure 6.4 Total area under a probability distribution curve.

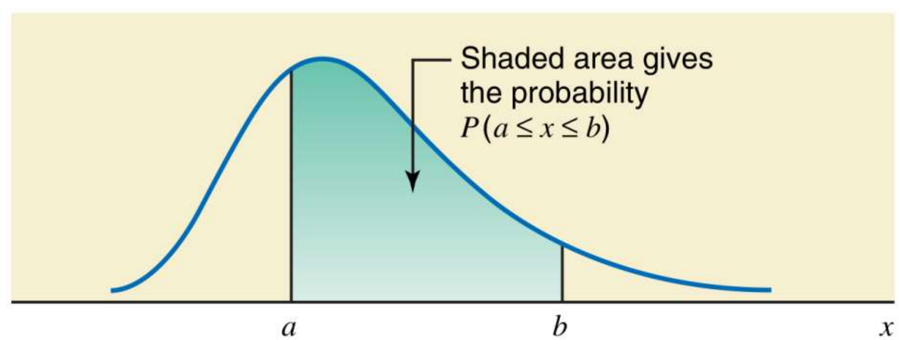


Figure 6.5 Area under the curve as probability.

Additional notes

- $P(X = a) = 0$ for any values of a
  - Because area under the curve will be 0

- $P(a \leq X \leq b) = P(a < X < b)$
  - Probability is the same whether a or b is or is not included in the interval.

- $P(a \leq X \leq b) = P(X \leq b) - P(X \leq a)$
  - Area under the curve between a and b is the same as area under the curve from negative infinity to b minus area from negative infinity to a.

Example

- Show that $f(x) = 3x^2$ for $0 < x < 1$ represents a probability distribution function. Then, find $P(0.1 < X \leq 0.5)$.

- $f(x) \geq 0$
- $\int_{-\infty}^{\infty} f(x)\,dx = \int_0^1 3x^2\,dx = \left[\frac{3x^3}{3}\right]_0^1 = 1 - 0 = 1$
- Therefore, this is a probability distribution function.

$$P(0.1 < X \leq 0.5) = \int_{0.1}^{0.5} f(x)\,dx = \int_{0.1}^{0.5} 3x^2\,dx$$

$$= \left[\frac{3x^3}{3}\right]_{0.1}^{0.5} = 0.5^3 - 0.1^3$$

$$= 0.124$$

Cumulative distribution function for continuous random variables

- If X is a continuous random variable and the value of its probability density at x is $f(x)$, then its distribution function or cumulative distribution function (cdf) is given by

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\,dt$$

for $-\infty < x < \infty$.

- Some properties of the cdf:
  - $F(x)$ is an increasing function, that is $F(a) \leq F(b)$ for any $a < b$.
  - $F(-\infty) = 0$ and $F(\infty) = 1$
  - We can get the pdf by differentiating the cdf:

$$f(x) = \frac{d}{dx}F(x)$$

Example

- The probability density function for a random variable X is given by $f(x) = 3x^2$ for $0 < x < 1$, and 0 elsewhere.
  - Find the cumulative distribution function of X.

For $0 < x < 1$:

$$F(x) = \int_{-\infty}^{x} f(t)\,dt = \int_{0}^{x} 3t^2\,dt = \left[\frac{3t^3}{3}\right]_0^x = x^3$$

Therefore,

$$F(x) = \begin{cases} 0, & \text{for } x \leq 0 \\ x^3, & \text{for } 0 < x < 1 \\ 1, & \text{for } x \geq 1 \end{cases}$$

Example

- Find $P(X < 0.5)$.

$$P(X < 0.5) = F(0.5) = 0.5^3 = 0.125$$

- Find $P(0.2 \leq X < 0.6)$.

$$\begin{aligned} P(0.2 \leq X < 0.6) &= P(X < 0.6) - P(X < 0.2) \\ &= F(0.6) - F(0.2) = 0.6^3 - 0.2^3 \\ &= 0.208 \end{aligned}$$

Mean and variance

- Mean for continuous random variable:

$$\mu = E[X] = \int_{-\infty}^{\infty} x f(x)\, dx$$

- Variance for continuous random variable:

$$\sigma^2 = E[X^2] - E[X]^2$$

$$\sigma^2 = \int_{-\infty}^{\infty} x^2 f(x)\, dx - \mu^2$$

Example

- The probability density function for a random variable X is given by $f(x) = 3x^2$ for $0 < x < 1$, and 0 elsewhere. Find the mean and variance of X.

- Mean:

$$\mu = \int_{-\infty}^{\infty} x f(x)\, dx = \int_0^1 3x^3\, dx = \left[\frac{3x^4}{4}\right]_0^1$$

$$= \frac{3}{4} - 0 = \frac{3}{4}$$

- Variance:

$$\sigma^2 = \int_{-\infty}^{\infty} x^2 f(x)\, dx - \mu^2 = \int_0^1 3x^4\, dx - \left(\frac{3}{4}\right)^2$$

$$= \left[\frac{3x^5}{5}\right]_0^1 - \frac{9}{16} = \frac{3}{5} - \frac{9}{16} = \frac{3}{80} = 0.0375$$

Exercise

- Given the probability density $f(x) = \frac{c}{\sqrt{x}}$ for $0 < x < 4$.

  a) Find the value of c.

  b) Find the distribution function $F(x)$.

  c) Calculate $P\left(X < \frac{1}{4}\right)$.

  d) Calculate the mean and standard deviation.

Exercise

- The length of time to failure (in hundreds of hours) for a transistor is a random variable Y with distribution function given as:

$$F(y) = \begin{cases} 0, & y < 0 \\ 1 - e^{-y^2}, & y \geq 0 \end{cases}$$

a) Find the density function $f(y)$.

b) Find the probability that the transistor operates for at least 200 hours.

c) Find the probability that the transistor operates for about 150 to 250 hours.

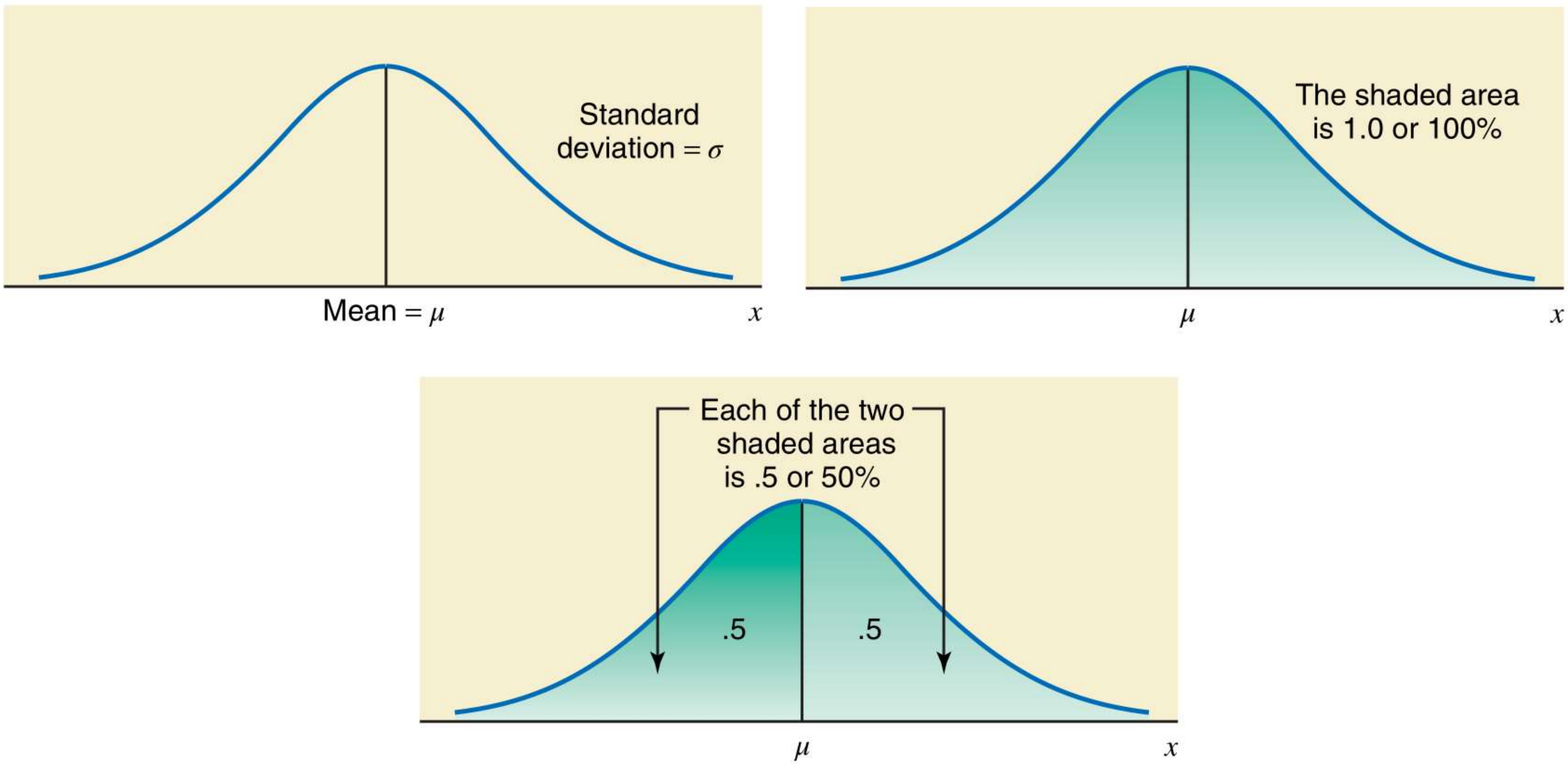
Normal distribution

Normal distribution

- Normal distribution is the most popular and used continuous probability distribution.
  - Popular because many real-world phenomena are normally distributed.

- A normal probability distribution, when plotted, gives a bell-shaped curve such that:
  - The total area under the curve is 1.0.
  - The curve is symmetric about the mean.
  - The two tails of the curve extend indefinitely.

Normal distribution



Normal distribution

- A normal distribution has two parameters:
  - Mean, μ
  - Standard deviation, σ

- The probability distribution function for a normal distribution :

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

- The value μ determines the centre of the normal distribution, and the value σ determines the spread of the curve

- If X is normally distributed with mean μ and variance σ^2, we usually write:

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

Normal distribution

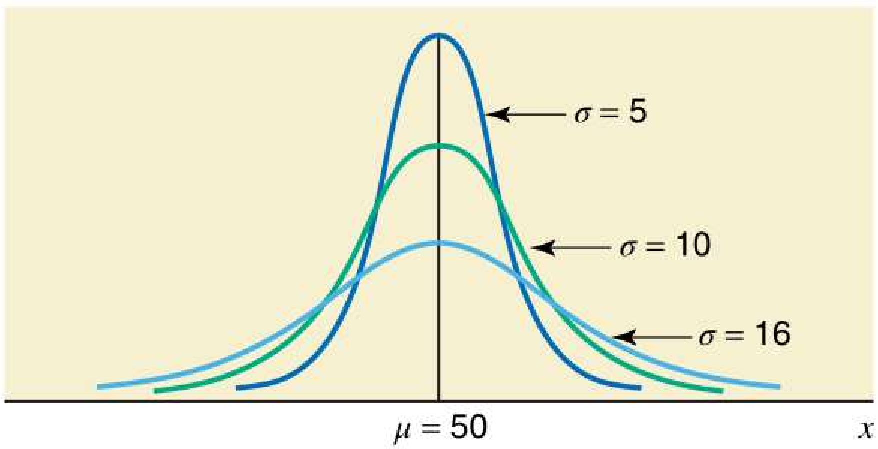


Figure 6.15 Three normal distribution curves with the same mean but different standard deviations.

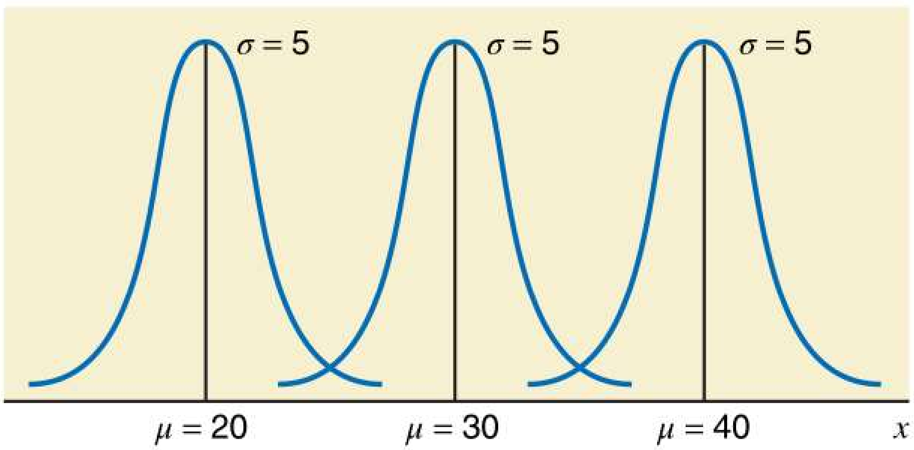


Figure 6.16 Three normal distribution curves with different means but the same standard deviation.

Standard normal distribution

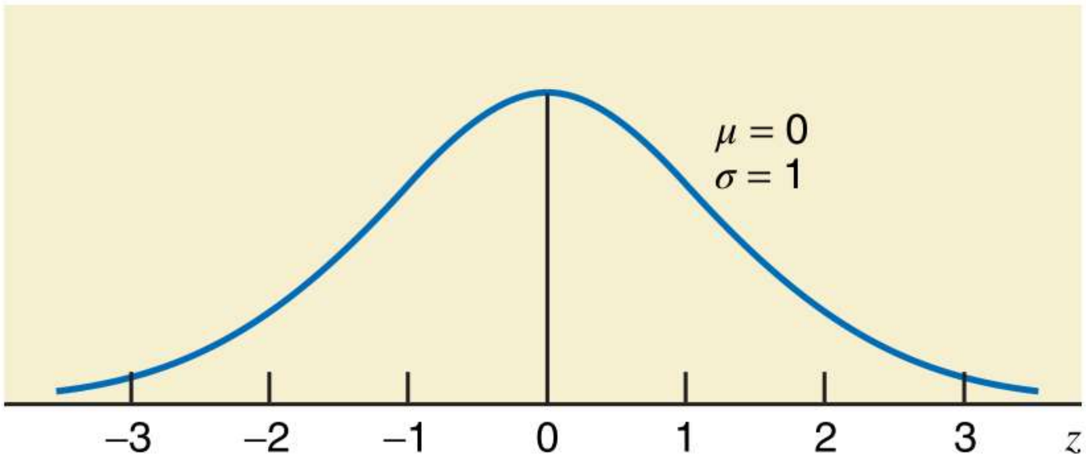
Standard normal distribution

- The standard normal distribution is a normal distribution with $\mu = 0$ and $\sigma = 1$.

- We normally denote standard normal distribution as Z and the values of Z as z-values or z-scores.

- The standard normal distribution table gives the probability $P(Z > z)$.

Standard normal distribution



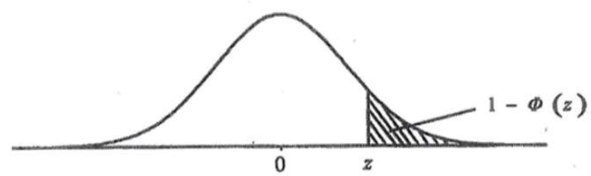
- Some properties of standard normal distribution:
  - The probability distribution is symmetric around 0. Eg $f(-2) = f(2)$.
  - $P(Z < -a) = P(Z > a)$ because of the symmetric probability distribution.

Using the table

Table 3 Areas in Upper Tail of the Normal Distribution

The function tabulated is $1-\Phi(z)$ where $\Phi(z)$ is the cumulative distribution function of a standardised Normal variable, z.

Thus $1-\Phi(z)=\frac{1}{\sqrt{2\pi}}\int_z^{\infty} e^{-z^2/2}$ is the probability that a standardised Normal variate selected at random will be greater than a value of z $\left(=\frac{x-\mu}{\sigma}\right)$



$\frac{x-\mu}{\sigma}$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681

Using the table

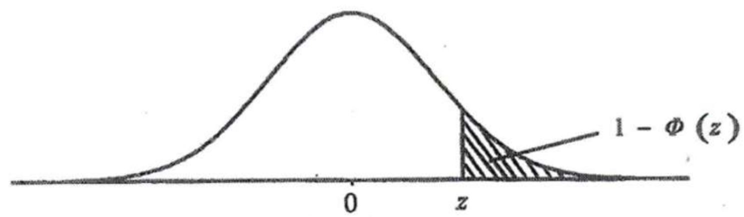
- Steps:
  1. Know the value of z. Suppose $z = 1.12$ and we want to find $P(Z > 1.12)$.
  2. Draw the normal distribution and shade the required area.
  3. On the table, the z-values are divided into two portion:
     - The number before decimal and one digit after decimal (1.1).
     - The second digit after decimal (0.02).

     Note that $z = 1.1 + 0.02$.
  3. To find $z = 1.12$ in the table, locate 1.1 in the column for z, and 0.02 in the row for z at the top of the table.
  4. The entry where the row and column intersect is the probability $P(Z > z)$.

Table 3 Areas in Upper Tail of the Normal Distribution

The function tabulated is $1-\Phi(z)$ where $\Phi(z)$ is the cumulative distribution function of a standardised Normal variable, z.

Thus $1-\Phi(z)=\frac{1}{\sqrt{2\pi}}\int_z^{\infty} e^{-z^2/2}$ is the probability that a standardised Normal variate selected at random will be greater than a value of z $\left(=\frac{x-\mu}{\sigma}\right)$



Second digit after decimal

Digit before decimal and after one decimal

$\frac{x-\mu}{\sigma}$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681

$P(Z>1.12)=0.1314$

Additional notes

- Some table gives $P(Z < z)$ instead of $P(Z > z)$. So make sure you properly read the table information first.

- Remember that $P(Z < -z) = P(Z > z)$.
  - Example: If I want to find $P(Z < -1.12)$, this is the same as $P(Z > 1.12)$

- Remember that $P(Z < z) = 1 - P(Z > z)$.
  - Example: If I want to find $P(Z < 1.12)$, all I need to find in the table is $P(Z > 1.12)$ and calculate $1 - P(Z > 1.12)$

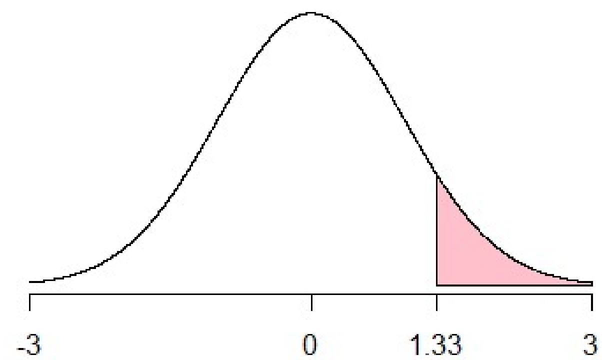
- As always for continuous distribution, $P(Z \leq z) = P(Z < z)$.

Examples

- Using the standard normal distribution table, find these values:
  a) $P(Z > 1.33)$
  b) $P(Z > 2)$
  c) $P(Z < -1.33)$
  d) $P(Z > -2)$
  e) $P(1.33 < Z \leq 2)$
  f) $P(-2 < Z < -1.33)$
  g) $P(-2 < Z < 1.33)$
  h) $P(-2 \leq Z \leq 2)$

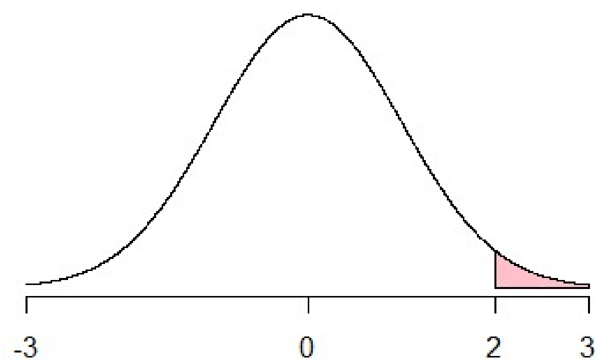
Example

a) $P(Z > 1.33)$:



From the table, $P(Z > 1.33) = 0.0918$

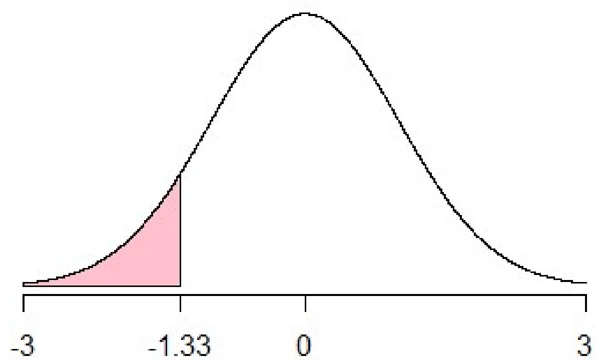
b) $P(Z > 2)$:



From the table, $P(Z > 2) = 0.02275$

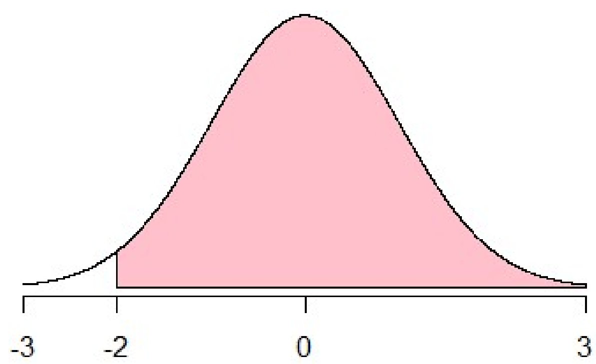
Example

c) $P(Z < -1.33)$:



$$P(Z < -1.33) = P(Z > 1.33) = 0.0918$$

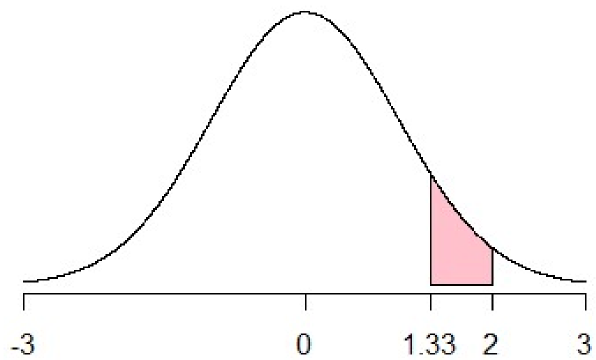
d) $P(Z > -2)$:



$$\begin{aligned} P(Z > -2) &= 1 - P(Z < -2) \\ &= 1 - P(Z > 2) \\ &= 1 - 0.02275 = 0.97725 \end{aligned}$$

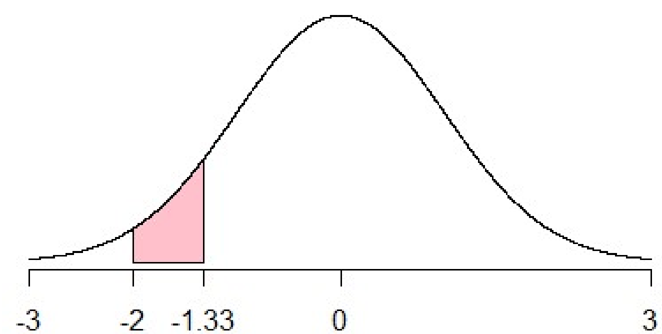
Example

e) $P(1.33 < Z \le 2)$:



$$P(1.33 < Z \le 2) = P(Z > 1.33) - P(Z > 2)$$
$$= 0.0918 - 0.02275 = 0.06905$$

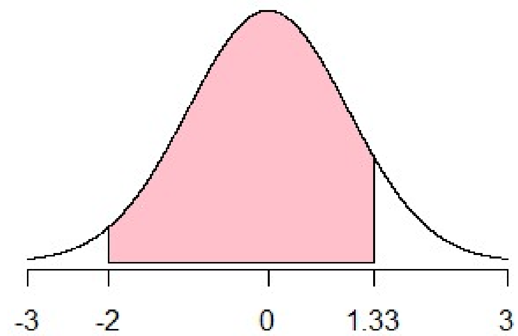
f) $P(-2 < Z < -1.33)$:



$$P(-2 < Z < -1.33) = P(Z < -1.33) - P(Z < -2)$$
$$= P(Z > 1.33) - P(Z > 2)$$
$$= 0.0918 - 0.02275 = 0.06905$$

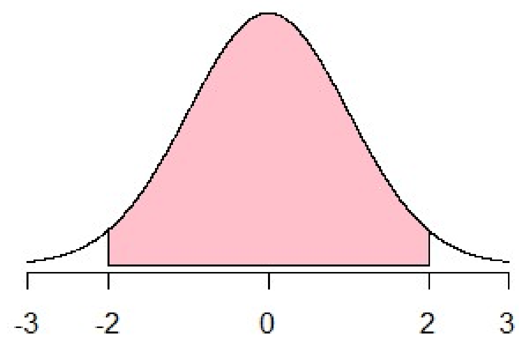
Example

g) $P(-2 < Z < 1.33)$:



$$
\begin{aligned}
P(-2 < Z < 1.33) &= 1 - P(Z < -2) - P(Z > 1.33) \\
&= 1 - P(Z > 2) - P(Z > 1.33) \\
&= 1 - 0.02275 - 0.0918 = 0.88545
\end{aligned}
$$

h) $P(-2 \leq Z \leq 2)$:



$$
\begin{aligned}
P(-2 \leq Z \leq 2) &= 1 - P(Z < -2) - P(Z > 2) \\
&= 1 - 2 \times P(Z > 2) \\
&= 1 - 2(0.02275) \\
&= 0.9545
\end{aligned}
$$

Exercise

- Determine the following probabilities for the standard normal distribution.
  a) $P(-1.83 \leq Z \leq 2.57)$
  b) $P(0 \leq Z \leq 2.02)$
  c) $P(-1.99 \leq Z \leq 0)$
  d) $P(Z \geq 1.48)$

Standardizing a normal distribution

Why standardize normal distribution

- For a general normal distribution, μ and σ can take any values (as long as $\sigma > 0$).

- Finding probability using the probability distribution for a normal distribution with μ and σ is difficult.
  - We will have to integrate the probability distribution over the interval.
  - Or use computer to calculate the probability.

- Alternatively, we can use the standard normal distribution table to find the probability.

Standardizing a normal distribution

$$X \sim N(\mu, \sigma^2) \Rightarrow Z \sim N(0,1)$$

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

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

has the standard normal distribution.

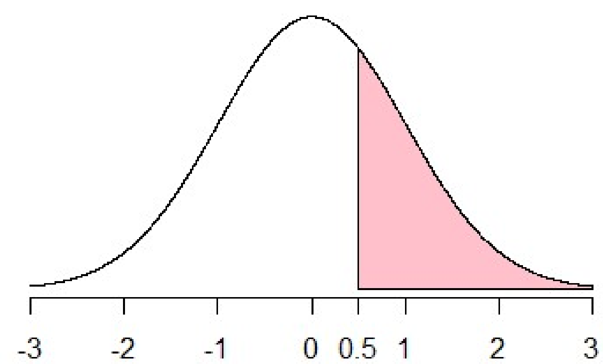
- Therefore, for any value of x, we can standardize it using the above formula and use the standard normal distribution table.

Example

- Let X be a continuous random variable that has a normal distribution with a mean of 50 and a standard deviation of 10. Find $P(X > 55)$.

- $X \sim N(50, 10^2)$

$$P(X > 55) = P\left(Z > \frac{55-\mu}{\sigma}\right)$$
$$= P\left(Z > \frac{55-50}{10}\right)$$
$$= P(Z > 0.5)$$



- From the table, $P(Z > 0.5) = 0.3085$.

$$P(X > 55) = 0.3085$$

Exercise

- Given that X is normally distributed with mean 10 and standard deviation 5, find the following probabilities:
  a) $P(X > 7)$
  b) $P(12 < X \leq 15)$
  c) $P(X < 6)$

Exercise

- The average number of calories in a 40 grams chocolate bar is 225. Suppose that the distribution of calories is approximately normal with $\sigma = 10$. Find the probability that a randomly selected chocolate bar will have
  a) Between 200 and 220 calories
  b) Less than 200 calories

Determining x or z values from probability

From probability to z-values

- Previously, we find the probability of normally distributed random variables within an interval.

- But suppose we were given the probability first, and we would like to find the interval with the given probability.

- Example: Suppose X is normally distributed.
  - What is the value of x such that $P(X > x) = 0.3$?

- To find the x values, we use the standard normal distribution table and find the z-values correspond to the probability. Then calculate x using

$$x = \mu + z\sigma$$

Example

- Find the value of z such that the area under the standard normal curve to the right of z is 0.4052.

- We want to find z such that $P(Z > z) = 0.4052$
- We use the table and find the probability $P(Z > z) = 0.4052$ and found that the z-value is 0.24.
- So $z = 0.24$

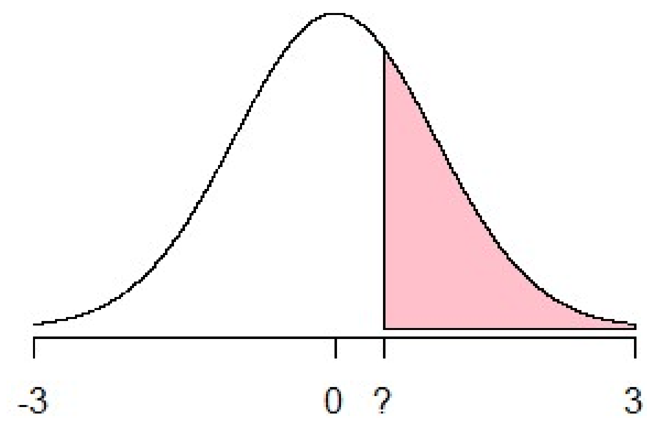
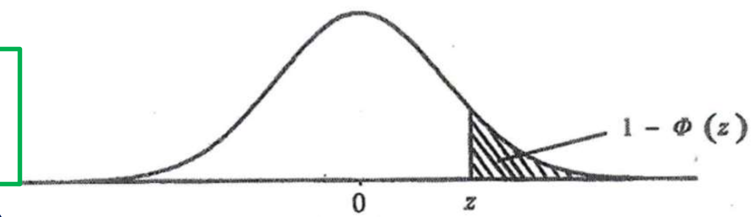


Table 3 Areas in Upper Tail of the Normal Distribution

The function tabulated is $1-\Phi(z)$ where $\Phi(z)$ is the cumulative distribution function of a standardised Normal variable, z.

Thus $1-\Phi(z)=\frac{1}{\sqrt{2\pi}}\int_z^\infty e^{-z^2/2}$ is the probability that a standardised Normal variate selected at random will be greater than a value of $z\left(=\frac{x-\mu}{\sigma}\right)$



Digit before decimal and after one decimal

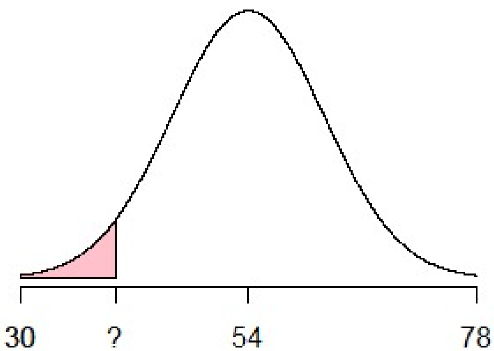
Second digit after decimal

$\frac{x-\mu}{\sigma}$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4090	$\boxed{.4052}$	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
0.8	.2119	.2090	.2061	.2033	.2005	.1977	[illegible]	[illegible]	[illegible]	.1867
0.9	.1841	.1814	.1788	.1762	.1736	.1711	[illegible]	[illegible]	[illegible]	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681

$P(Z>z)=0.4052$

Example

- The life span of a calculator has a normal distribution with a mean of 54 months and a standard deviation of 8 months. What should the warranty period be to replace a malfunctioning calculator if the company does not want to replace more than 1% of all the calculators sold?



- Let X = life span of a calculator in months.
- $X \sim N(54, 8^2)$
- Want to find x such that $P(X < x) = 0.01$
- Need to find z such that $P(Z < z) = 0.01$
- From the standard normal distribution table, $P(Z > 2.33) = 0.00990$ gives the closest probability to 0.01, and we choose $z = -2.33$

$$x = \mu + z\sigma = 54 - 2.33(8) = 35.36$$

- Warranty period should be within 35.36 months.

Exercise

- To qualify for a police academy, applicants are given a test of physical fitness. The scores are normally distributed with a mean of 64 and a standard deviation of 9. If only the top 20% of the applicants are selected, find the cutoff score.

Sampling distribution of sample mean

Sampling distribution of $\bar{X}$

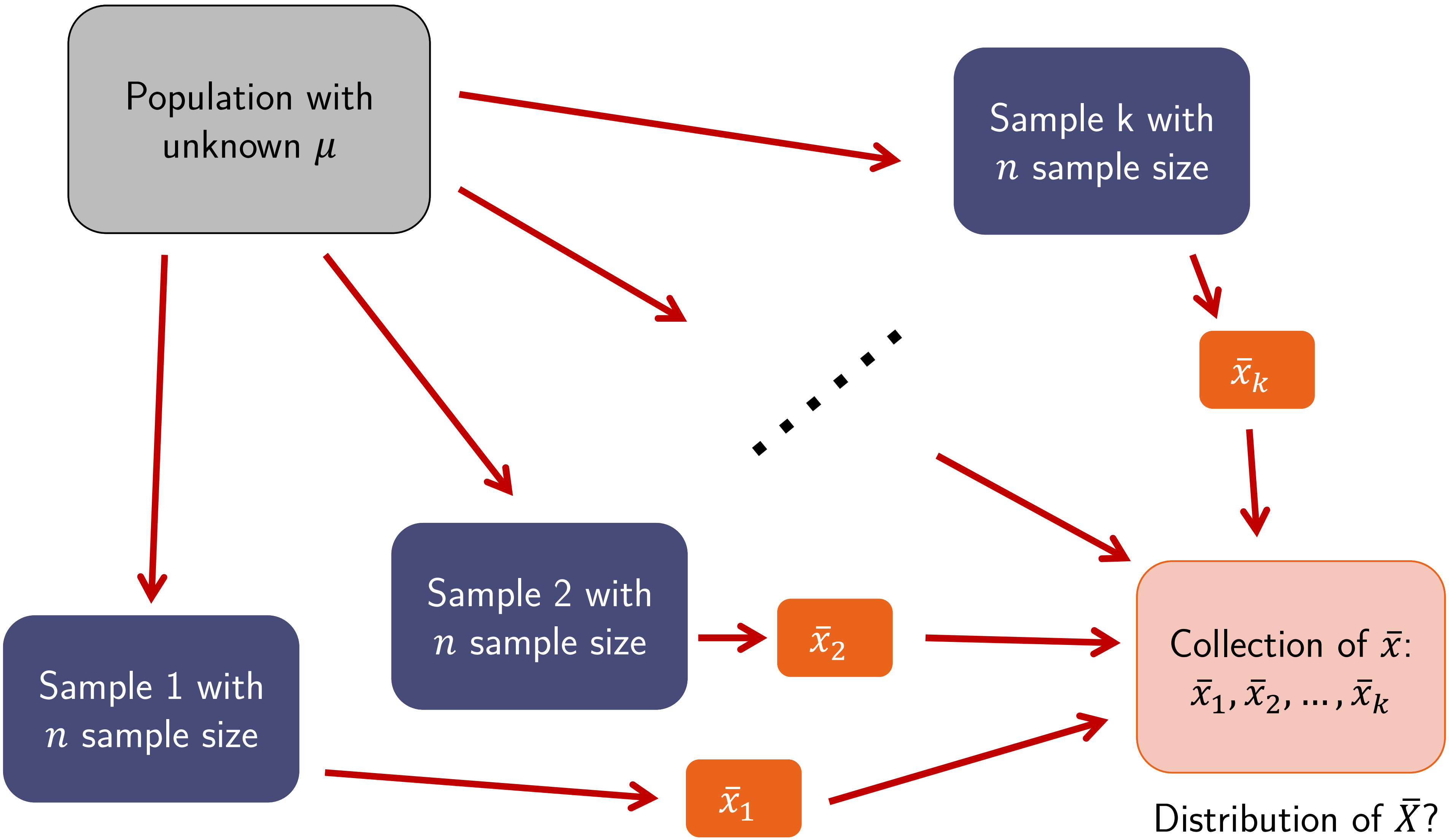
- Suppose I have a population, then I can collect a sample from the population.

- Now suppose I collect multiple samples, and for each sample I calculate its sample mean, $\bar{x}$.

- Then $\bar{X}$ will be a random variable with its own probability distribution.

- We are interested in the distribution of the sample mean, $\bar{X}$.

Sampling distribution of $\bar{X}$



Mean and standard deviation of $\bar{X}$

- Mean and standard deviation of $\bar{X}$ with n sample size:

$$\mu_{\bar{X}} = \mu, \qquad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

- Also the variance of $\bar{X}$:

$$\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}$$

- The mean of $\bar{X}$ is the same as the mean of X.

- The variance or spread the distribution for $\bar{X}$ decreases as n increases.

Example

- Assume the training heart rates of all 20-year-old athletes are distributed with a mean of 135 beats per minute and a standard deviation of 18 beats per minute. Find the mean and standard deviation of $\bar{X}$ for a sample size of
  a) 4
  b) 9
  c) 16

- It is given that $\mu = 135$ and $\sigma = 18$.
  a) When $n = 4$, $\mu_{\bar{X}} = \mu = 135$, and $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{18}{\sqrt{4}} = 9$
  b) When $n = 9$, $\mu_{\bar{X}} = 135$, and $\sigma_{\bar{X}} = \frac{18}{\sqrt{9}} = 6$
  c) When $n = 16$, $\mu_{\bar{X}} = 135$, and $\sigma_{\bar{X}} = \frac{18}{\sqrt{16}} = 4.5$

Shape of sampling distribution of $\bar{X}$

- Two cases:
  - Samples are drawn from a population that has a normal distribution.
  - Samples are drawn from a population that does not have a normal distribution.

Case 1: Population has a normal distribution

- If the population has a normal distribution, then the sample mean $\bar{X}$ will also have a normal distribution, no matter what the value of n is.

- The mean and standard deviation is as mentioned before:

$$\mu_{\bar{X}} = \mu, \qquad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

- In this case, if $X \sim N(\mu, \sigma^2)$, then

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

Case 2: Population does not have a normal distribution

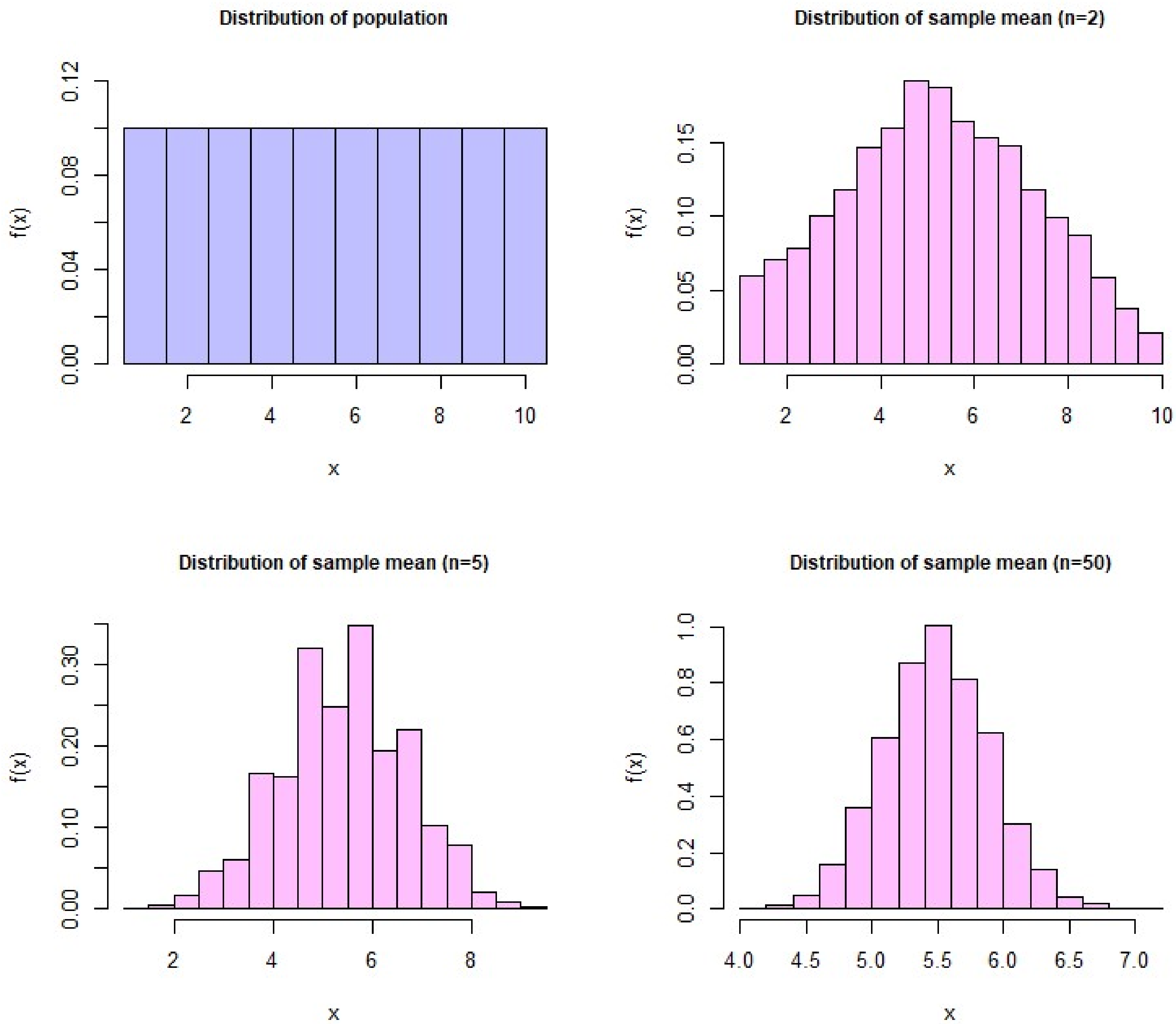
- If the population does not have a normal distribution, we will depend on the Central Limit Theorem.

- Central Limit Theorem (CLT):
  - For a large sample size, the sampling distribution of $\bar{X}$ is approximately normal, irrespective of the shape of the population distribution.

- What does CLT mean?
  - When n is large, $\bar{X}$ is approximately normal.
  - Distribution of population does not matter.
  - General rule: n is large enough to use CLT when $n \geq 30$.

Simulation example

- Let X be a random number from 1 to 10 with equal probability.
- Then take 5,000 samples each with sample size n from X, and calculate the sample mean, $\bar{X}$.
- Then draw the histogram of the sample mean $\bar{X}$.
- For this simulation, I use $n = 2$, 5 and 50.


Distribution of population
Distribution of sample mean (n=2)
Distribution of sample mean (n=5)
Distribution of sample mean (n=50)
f(x)
x

Shape of sampling distribution of $\bar{X}$

- In both cases,

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

- For population that has a normal distribution, the above is true for all value of n.

- But for population that does not have a normal distribution, the above is true when n is large $(n \geq 30)$ using CLT.

Example

- The average time spent by construction workers who work on weekends is 7.93 hours (over 2 days). Assume the distribution is approximately normal and has a standard deviation of 0.8 hour.
  a) If a sample of 2 construction workers is randomly selected, find the probability that the mean of the sample is less than 8 hours.
  b) If a sample of 40 construction workers is randomly selected, find the probability that the mean of the sample will be less than 8 hours.

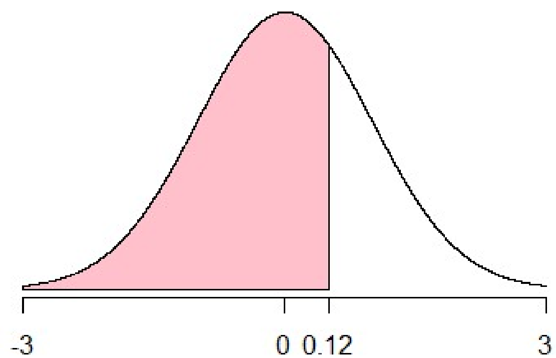
- Let X = time spend by construction workers on weekends.
- Given in the question $X \sim N(7.93, 0.8^2)$.

Example

a) Suppose $n = 2$. Since X is normally distributed, then even when n is small, $\bar{X}$ is normally distributed. In this case, $\mu_{\bar{X}} = \mu = 7.93$ and $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{0.8}{\sqrt{2}} = 0.5657$.

$$\bar{X} \sim N(7.93, 0.5657^2)$$

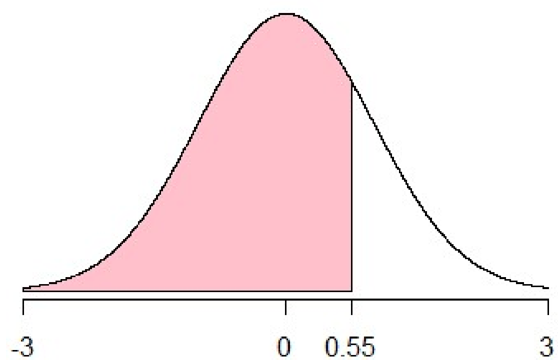
$$P(\bar{X} < 8) = P\left(Z < \frac{8 - 7.93}{0.5657}\right) = P(Z < 0.1237)$$
$$= 1 - P(Z > 0.12)$$
$$= 1 - 0.4522 = 0.5478$$



b) Suppose $n = 40$, then $\mu_{\bar{X}} = 7.93$ and $\sigma_{\bar{X}} = \frac{0.8}{\sqrt{40}} = 0.1265$.

$$\bar{X} \sim N(7.93, 0.1265^2)$$

$$P(\bar{X} < 8) = P\left(Z < \frac{8 - 7.93}{0.1265}\right) = P(Z < 0.5534)$$
$$= 1 - P(Z > 0.55)$$
$$= 1 - 0.2912 = 0.7088$$



Another example

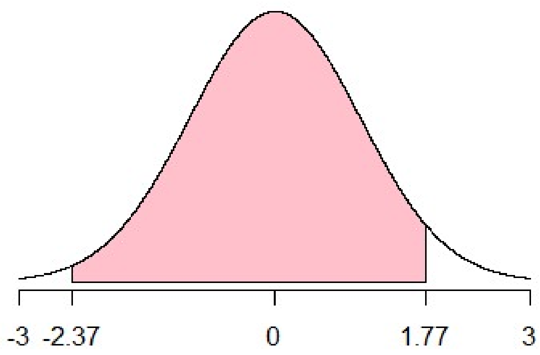
- The average number of earthquakes that occur in Los Angeles over one month is 36. (Most are undetectable.) Assume the standard deviation is 5. If a random sample of 35 months is selected, find the probability that the mean of the sample is between 34 and 37.5.

- Let X be the number of earthquakes in a month. Given that $\mu = 36$, $\sigma = 5$.
- If $n = 35$, then we can use CLT and

$$\mu_{\bar{X}} = \mu = 36, \qquad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{5}{\sqrt{35}} = 0.8452$$

- $\bar{X} \sim N(36, 0.8452^2)$

$$P(34 < \bar{X} < 37.5) = P\left(\frac{34-36}{0.8452} < Z < \frac{37.5-36}{0.8452}\right)$$
$$= P(-2.37 < Z < 1.77)$$
$$= 1 - P(Z > 2.37) - P(Z > 1.77)$$
$$= 1 - 0.00889 - 0.0384 = 0.9527$$



Exercise

- The GPAs of all students enrolled at a large university have an approximately normal distribution with a mean of 3.02 and a standard deviation of 0.29. Find the probability that the mean GPA of a random sample of 20 students selected from this university is
  a) 3.10 or higher
  b) 2.90 or lower
  c) 2.95 to 3.11

Exercise

- Suppose that the current distribution of times spent watching television per day by all Americans age 15 and over has a mean of 168 minutes and a standard deviation of 20 minutes. Find the probability that the average time spent per day watching television by a random sample of 400 Americans age 15 and over is
  a) at most 165 minutes
  b) more than 169.8 minutes

Summary

- Continuous random variable
  - Probability density function
  - Cumulative distribution function
  - Mean and variance

- Normal distribution
  - Read the probability using standard normal distribution table
  - Standardizing normal distribution

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

  - Solving problems involving normal distribution

Summary

- Distribution for sample mean:

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

- When population has a normal distribution, then $\bar{X}$ is also normally distributed no matter what n is.

- When population does not have a normal distribution, then we rely on CLT and check if $n \geq 30$. If it is, $\bar{X}$ is approximately normal.