Announcements

Topics:

In the Probability and Statistics module:

- **Section 7:** The Centre of a Distribution (Mean)
- **Section 8:** The Spread of a Distribution (Variance and Standard Deviation)
- **Section 10:** The Binomial Distribution
- **Section 13:** Continuous Random Variables
- **Section 14:** The Normal Distribution

To Do:

- Work on Assignments and Suggested Practice Problems assigned on the webpage under the SCHEDULE + HOMEWORK link

Statistics on a Distribution

- Often, important information about a distribution is realized by studying its centre and spread.

- One way to do this is to determine the mean, variance, and standard deviation of the distribution.

Mean or Expected Value

Example: Marks
Consider the following set of marks for 10 students:

Test 1, out of 40:

20, 24, 24, 27, 28, 29, 36, 36, 36, 40

Mark							
Frequency							
Relative Frequency							

What is the average of these marks?

Mean or Expected Value

Example: Marks
Consider the following set of marks for 10 students:

Test 1, out of 40:

20, 24, 24, 27, 28, 29, 36, 36, 36, 40

Mark	20	24	27	28	29	36	40
Frequency	1	2	1	1	1	3	1
Relative Frequency	1/10	2/10	1/10	1/10	1/10	3/10	1/10

What is the average of these marks?

Mean or Expected Value

Definition:

Let X be a discrete random variable. The mean or the expected value of X is the number

$$E(X) = \sum_x xP(X = x) = \sum_x xp(x)$$

where the sum goes over all values x for which p(x)=P(X=x) is not zero.

Mean or Expected Value

Example: Leopard Population with Immigration

Consider a population of leopards p_t modelled by

$$p_{t+1} = p_t + I_t \quad \text{where} \quad I_t = \begin{cases} 10 & \text{with a 90\% chance} \\ -100 & \text{with a 10\% chance} \end{cases}$$

Suppose that initially there are 300 leopards. Determine the expected number of leopards after **2** years.

Mean or Expected Value of a Function of a Random Variable

Definition:

Assume that X is a discrete random variable and that p(x)=P(X=x) is its probability mass function. Let g(x) be a function of x. The expected value of the random variable g(X) is

$$E(g(X)) = \sum_x g(x)P(X = x) = \sum_x g(x)p(x)$$

where the sum goes over all values x for which p(x) is not zero.

Properties of the Expected Value

Theorem:

Let X and Y be discrete random variables and a and b be real numbers. Then

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

(2) $X \pm Y$ is a discrete random variable and

$E(X \pm Y)=E(X) \pm E(Y)$

The Spread of a Distribution

Example: Marks
Consider the following sets of marks for 10 students:

Test 1, out of 40:

20, 20, 20, 20, 20, 40, 40, 40, 40, 40

Test 2, out of 40:

30, 30, 30, 30, 30, 30, 30, 30, 30, 30

Compare the spreads of Test 1 and Test 2 scores.

Variance

Definition:

Assume that X is a random variable with mean $\mu=E(X)$. The variance of X is the real number

$$\sigma^2 = \mathrm{var}(X) = E\left[(X-\mu)^2\right] = E\left[(X-E(X))^2\right]$$

Variance

In words:

The variance of a random variable *X* is the expected value of the difference (squared) between *X* and its mean.

The larger the variance, the larger the spread of a distribution.

Standard Deviation

Definition:

Let X be a random variable whose variance is $\sigma^2 = \mathrm{var}(X)$. The standard deviation of X is the number

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

The standard deviation is measured in the same units as the random variable X.

Variance and Standard Deviation

Example: Leopard Population with Immigration
Determine the variance and standard deviation for the number of leopards after **2** years.

x	p(x)
100	0.01
210	0.18
320	0.81

Properties of the Variance

Let X be a random variable and a and b be real numbers. Then

(1) $\mathrm{var}(aX+b)=a^2\mathrm{var}(X)$

(2) $\mathrm{var}(X)=E(X^2)-[E(X)]^2$

Bernoulli Experiment and Bernoulli Random Variable

A *Bernoulli experiment* is a simple random experiment with only two possible outcomes.

Definition:
A discrete random variable that takes on the value 1 (“success”) with probability p and the value 0 (“no-success”) with probability $1-p$ is called a *Bernoulli random variable*.

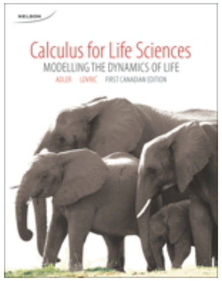
Bernoulli Experiment and Bernoulli Random Variable

Example: <u>Elephant Population with Immigration</u>

Consider a population of elephants p_t modelled by

$$p_{t+1} = p_t + I_t \quad \text{where} \quad I_t = \begin{cases} 10 & \text{with a } 90\% \text{ chance} \\ 0 & \text{with a } 10\% \text{ chance} \end{cases}$$

Define a Bernoulli experiment and determine the probability mass function for the corresponding Bernoulli random variable.



The Binomial Distribution

Assume that we repeat the same Bernoulli experiment n times and that the outcomes are independent.

Let N denote the random variable that counts the number of successes in n repetitions of the experiment, where p is the probability of success in a single experiment.

Then N is a binomially distributed random variable and we write

$$N \sim B(n,p)$$

The Binomial Distribution

Define the *binomial probability distribution* by

$$b(k, n; p) = P(N = k)$$

where $b(k, n; p)$ is the probability of exactly k successes in n repetitions of the same experiment, where p is the probability of success in a single experiment.

The Binomial Distribution

Example: Elephant Population with Immigration
Consider the population of elephants p_t modelled by

$$p_{t+1} = p_t + I_t \quad \text{where} \quad I_t = \begin{cases} 10 & \text{with a 90\% chance} \\ 0 & \text{with a 10\% chance} \end{cases}$$

where *t*=0, 1, 2, … is measured in years.

Let *N* count the number of times immigration occurs over the next 3 years. Determine the probability mass function for *N*.

The Binomial Distribution

The probability of *k* successes in *n* experiments is (number of ways of obtaining *k* successes in *n* experiments)*(probability of success)^{k}*(probability of no-success)$^{n-k}$

i.e., $$b(k, n; p) = C(n,k)\, p^{k} (1 - p)^{n-k}$$

The Probability Distribution of the Binomial Variable

Theorem:

The probability distribution of the binomial variable *N* is given by

$$P(N = k) = b(k, n; p) = \binom{n}{k} p^k (1 - p)^{n-k}$$

where *N* counts the number of successes in *n* independent repetitions of the same Bernoulli experiment and *p* is the probability of success.

The Probability Distribution of the Binomial Variable

Example: <u>Coin Toss</u>

What is the probability of exactly 7 tails in 10 tosses?

The Probability Distribution of the Binomial Variable

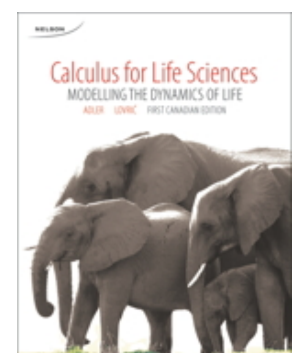
Example: Elephant Population with Immigration

Consider a population of elephants p_t modelled by

$$p_{t+1} = p_t + I_t \quad \text{where } I_t = \begin{cases} 10 & \text{with a 90\% chance} \\ 0 & \text{with a 10\% chance} \end{cases}$$

Suppose that initially there are 80 elephants.

What is the probability that there will be more than 300 elephants after 25 years?



The Mean and Variance of the Binomial Distribution

Mean and Variance of the Binomial Random Variable *N*:

$$E(N) = np$$

$$Var(N) = np(1 - p)$$

The Mean and Variance of the Binomial Distribution

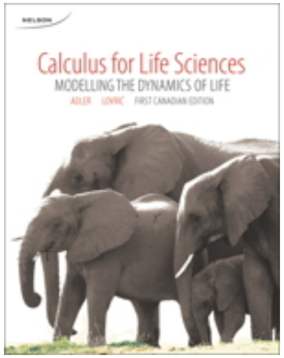
Example: Elephant Population with Immigration

Consider a population of elephants p_t modelled by

$$p_{t+1} = p_t + I_t \quad \text{where } I_t = \begin{cases} 10 & \text{with a } 90\% \text{ chance} \\ 0 & \text{with a } 10\% \text{ chance} \end{cases}$$

Suppose that initially there are 80 elephants.

What is the expected value of the population after 25 years? What is the standard deviation?



Continuous Random Variables

Definition:

A random variable that takes on a *continuum* of values is called a continuous random variable.

Continuous Random Variables

Example:

<u>Distributions of Lengths of Boa Constrictors</u>

The boa constrictor is a large species of snake that can grow to anywhere between 1 m and 4 m in length.

Let *L* be the continuous random variable that measures the length of a snake.

$$L : S \rightarrow [1,4]$$

Continuous Random Variables

The lengths of 500 boas are recorded below:

Table 13.1

Length range (m)	Frequency	Relative frequency
$[1, 1.5)$	20	0.04
$[1.5, 2)$	58	0.116
$[2, 2.5)$	122	0.244
$[2.5, 3)$	180	0.36
$[3, 3.5)$	86	0.172
$[3.5, 4)$	34	0.068

Note: relative frequency = frequency/500 = probability

Continuous Random Variables

<u>Histogram for Probability Mass</u>:

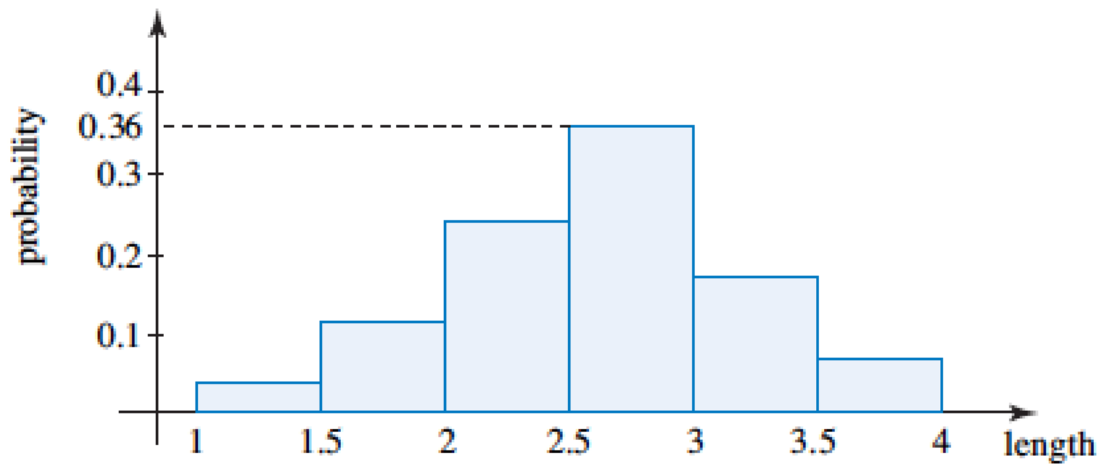


FIGURE 13.1

Histogram: the heights represent the probability

The probability that a randomly selected boa is between 2.5 m and 3 m in length is the <u>height</u> of the rectangle over [2.5, 3), i.e., 0.36.

Continuous Random Variables

To draw a histogram representing probability **density**, we re-label the vertical axis so that the probability that *L* belongs to an interval is the area of the rectangle above that interval.

Continuous Random Variables

WANT: area = probability

HAVE: area = base length * height

NEED: height = probability/base length

Continuous Random Variables

For example, consider the interval [2.5, 3). The probability that L falls in this range is 0.36.

Now, we want this value to be the <u>area</u> of the rectangle over [2.5, 3), so

$$\text{height} = 0.36/(3-2.5) = 0.72$$

Continuous Random Variables

Note:

$$\frac{\text{probability mass}}{\text{length of interval}} = \text{probability density}$$

Continuous Random Variables

<u>Histogram for Probability Density</u>:

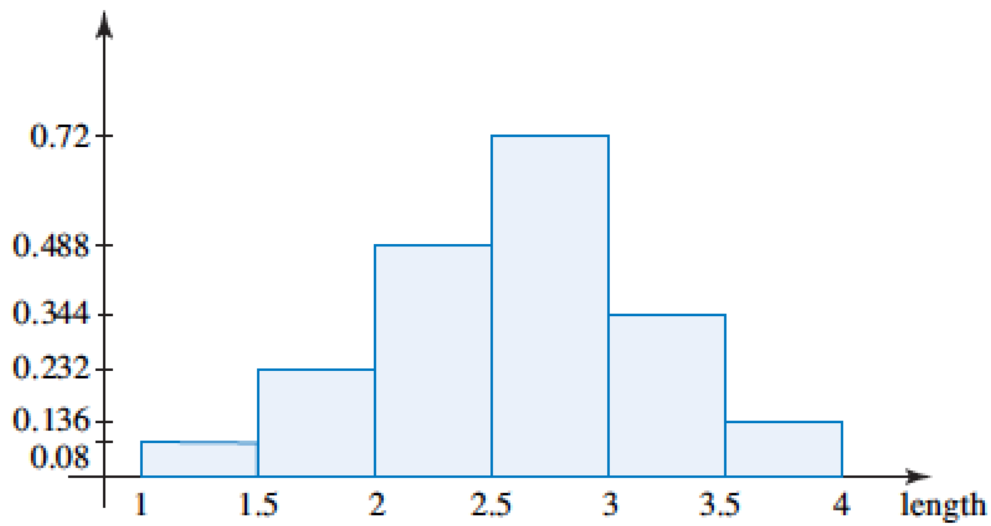


FIGURE 13.2

Histogram: the areas represent the probability

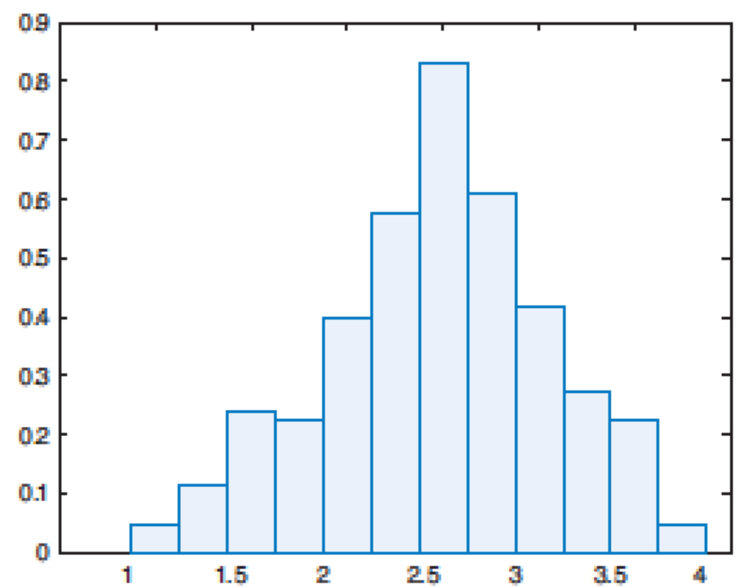
The probability that a randomly selected boa is between 2.5 m and 3 m in length is the <u>area</u> of the rectangle above [2.5, 3), i.e. 0.36.

Continuous Random Variables

To get a more precise probability mass (or density) function, we divide [1,4] into smaller subintervals:

Table 13.2

Length range (m)	Frequency	Relative frequency
$[1, 1.25)$	6	0.012
$[1.25, 1.5)$	14	0.028
$[1.5, 1.75)$	30	0.06
$[1.75, 2)$	28	0.056
$[2, 2.25)$	50	0.1
$[2.25, 2.5)$	72	0.144
$[2.5, 2.75)$	104	0.208
$[2.75, 3)$	76	0.152
$[3, 3.25)$	52	0.104
$[3.25, 3.5)$	34	0.068
$[3.5, 3.75)$	28	0.056
$[3.75, 4)$	6	0.012



Continuous Random Variables

As we continue to increase the number of subintervals, we obtain a more and more refined histogram.

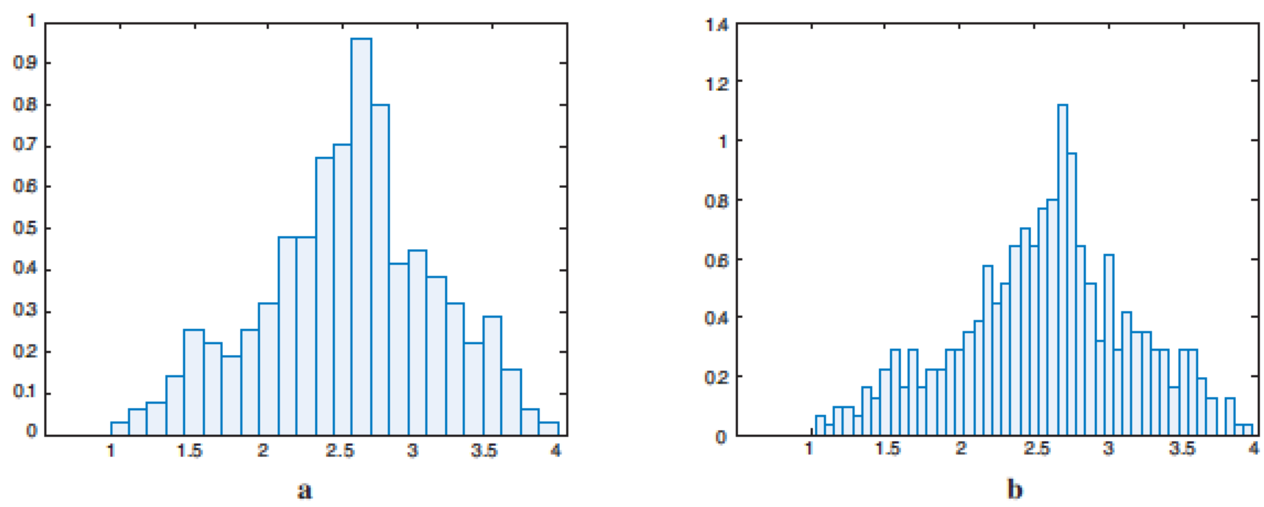


FIGURE 13.4
Histograms based on 24 and 48 subintervals

Continuous Random Variables

<u>Riemann Sum</u>:

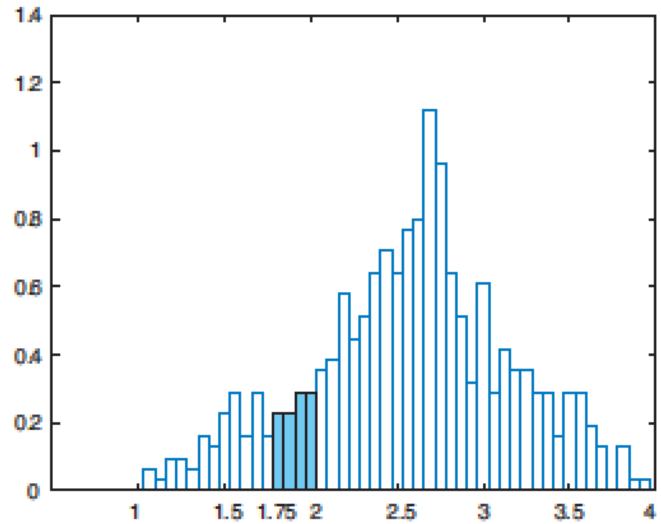


FIGURE 13.5

Probability of boa length between 1.75 m and 2 m

The probability that a randomly chosen boa is between 1.75 m and 2 m in length is the sum of the areas of the rectangles over the interval [1.75, 2).

Continuous Random Variables

To obtain the probability density **function**, we let the length of the intervals approach 0 and the number of rectangles approach ∞.

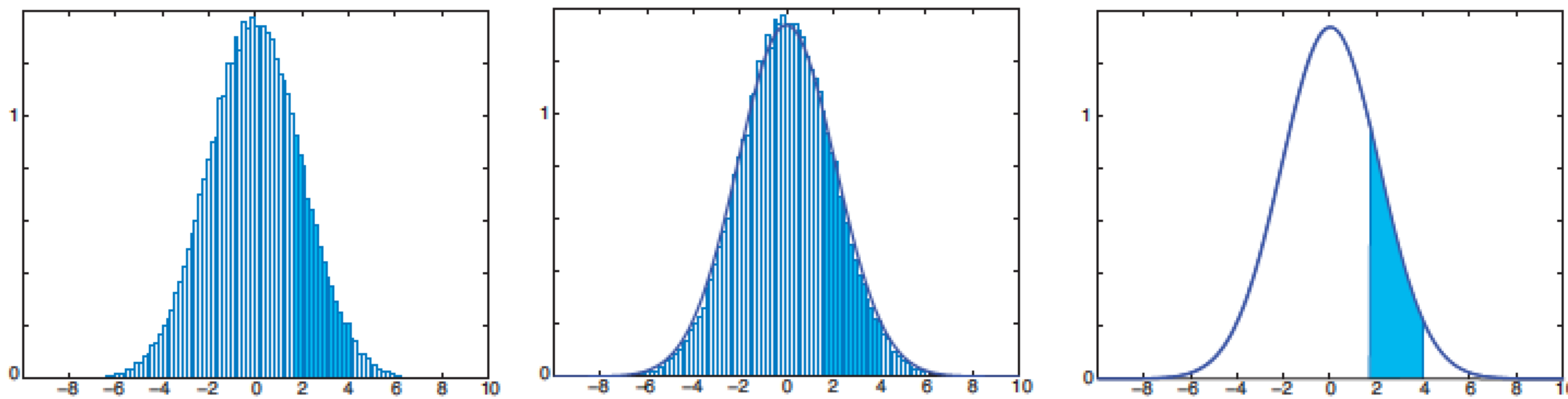


FIGURE 13.6

From a histogram to a density function

Probability Density Functions

Definition: Defining Properties of a PDF

Assume that the interval *I* represents the sample space of an experiment (thus, a simple event is represented by a real number from *I*). A function *f(x)* can be a probability density function if

(1) $f(x) \geq 0 \quad \text{for all } x \in I.$

(2) $\int_I f(x)dx = 1.$

Probability Density Functions

Example:

Show that $f(x) = \dfrac{2}{\pi(1+x^2)}$

could be a probability density function for some continuous random variable on $[0, \infty)$.

Calculating Probabilities

For a continuous random variable, we calculate the probability that a random variable belongs to an *interval* of real numbers.

The probability that an outcome *X* is between *a* and *b* is the area under the graph of *f(x)* on *[a,b]*:

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

Calculating Probabilities

The probability that an outcome is *equal* to a particular value is zero.

$$P(a \le X \le a) = \int_a^a f(x)dx = 0$$

For this reason, **including or excluding the endpoints of an interval does not affect the probability**, i.e.,

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

Calculating Probabilities

Example #32:
The distance between a seed and the plant it came from is modelled by the density function

$$f(x) = \frac{2}{\pi(1 + x^2)}$$

where *x* represents the distance (in metres), $x \in [0,\infty)$.

What is the probability that a seed will be found farther than 5 m from the plant?

Cumulative Distribution Function

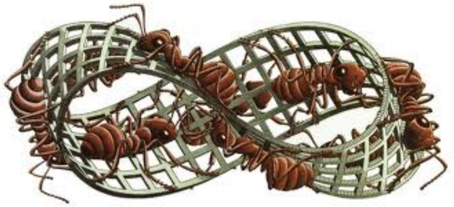
Definition:
Suppose that *f(x)* is a probability density function defined on an interval *[a,b]*. The function *F(x)* defined by

$$F(x) = P(X \leq x) = \int_a^x f(t)\,dt$$

for all *x* in *[a,b]* is called a cumulative distribution function of *f(x)*.

Cumulative Distribution Function

Example #30 (modified):
Suppose that the lifetime of an insect is given by the probability density function



$$f(t) = 0.2e^{-0.2t}$$

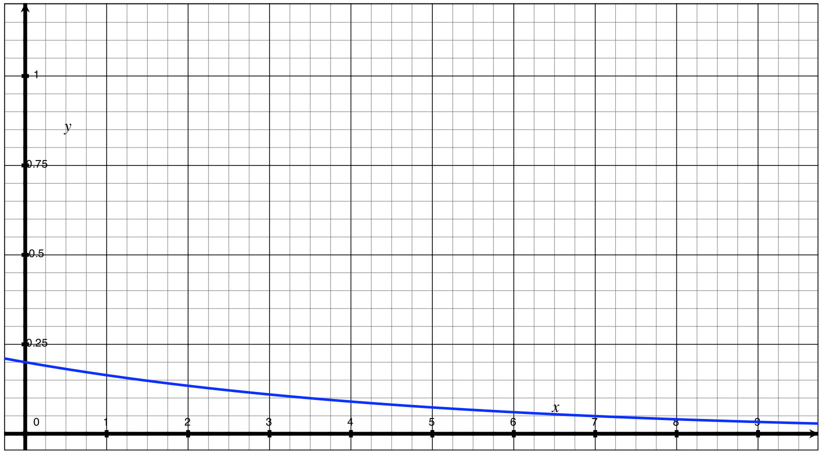
where *t* is measured in days, $t \in [0,\infty)$.

(a) Determine the corresponding cumulative distribution function, *F(t)*.
(b) Find the probability that the insect will live between 5-7 days.

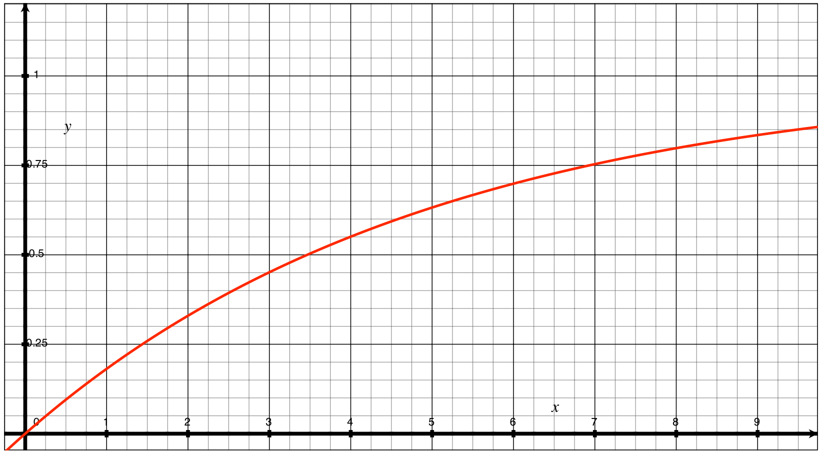
Cumulative Distribution Function

Example #30 (modified):

probability density function f(t)



cumulative distribution function F(t)



Cumulative Distribution Function

Properties of the CDF:

Assume that *f* is a probability density function, defined and continuous on an interval *[a,b]*. The left end *a* could be a real number or negative infinity; the right end *b* could be a real number or infinity. Denote by *F* the associated cumulative distribution function. Then

(1) $0 \le F(x) \le 1$ for all $x \in [a,b]$.

(2) $F(x)$ is continuous and non-decreasing.

(3) $\lim_{x \to a} F(x) = 0$ and $\lim_{x \to b} F(x) = 1$.

The Mean and the Variance

Definition:

Let *X* be a continuous random variable with probability density function *f(x)*, defined on an interval *[a,b]*.

The mean (or the expected value) of *X* is given by

$$\mu = E(X) = \int_a^b x\, f(x)\,dx$$

The variance of *X* is

$$\operatorname{var}(X) = E\left[(X-\mu)^2\right] = \int_a^b (x-\mu)^2 f(x)\,dx$$

The Mean and the Variance

Example #24:
Consider the continuous random variable *X* given by the probability density function

$$f(x) = 0.3 + 0.2x \quad \text{for} \quad 0 \le x \le 2.$$

Find the probability that the values of *X* are at least one standard deviation above the mean.

The Normal Distribution

The normal distribution is the most important continuous distribution as it can be used to model many phenomena in a variety of fields.

Many measurements for large sample sizes are said to be ‘normally distributed’.

For example, heights of trees, IQ scores, and duration of pregnancy are all normally distributed measurements.

The Normal Distribution

Definition:

A continuous random variable *X* has a normal distribution (or is distributed normally) with mean μ and variance σ^2, denoted by $X \sim N(\mu, \sigma^2)$, if its probability density function is

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

where $x \in (-\infty, \infty)$.

The Normal Distribution

The graph of the probability density function of the normal distribution (also known as the Gaussian distribution) is a bell-shaped curve.

Properties of the Normal Distribution Density Function

Theorem:

The probability density function *f(x)* of the normal distribution satisfies the following properties:

(a) f(x) is symmetric with respect to the vertical line $x = \mu$.

(b) *f(x)* is increasing for $x < \mu$ and decreasing for $x > \mu$. It has a local (also global) maximum value $1/\sigma\sqrt{2\pi}$ at $x = \mu$.

(c) The inflection points of *f(x)* are $x = \mu \pm \sigma$.

(d) $\lim_{x \to -\infty} f(x) = \lim_{x \to \infty} f(x) = 0$

Calculating Probabilities

If *X* is a normally distributed continuous random variable with mean μ and variance σ^2, then

$$P(a \le X \le b) = \int_a^b f(x)dx = \int_a^b \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

Calculating Probabilities

This integral cannot be evaluated without estimation techniques, such as using a Taylor polynomial to approximate $f(x)$.

To evaluate this integral, we reduce a general normal distribution to a special normal distribution, called the standard normal distribution, and then use tables of estimated values.

Standard Normal Distribution

Definition:

The standard normal distribution is the normal distribution with mean 0 and variance 1; in symbols, it is *N(0,1)*. Its probability density function is given by

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

for all $x \in (-\infty, \infty)$.

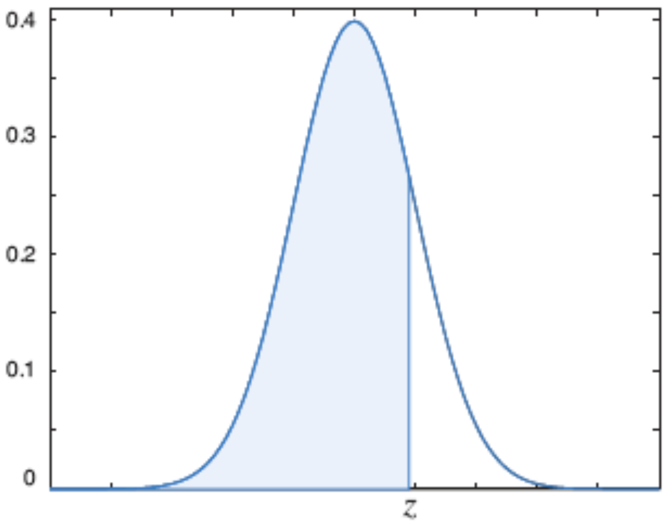
Standard Normal Distribution

We use the symbol Z to denote the continuous random variable that has the standard normal distribution; i.e., $Z \sim N(0,1)$.

Standard Normal Distribution

The cumulative distribution function of *Z* is given by

$$F(z) = \int_{-\infty}^{z} f(x)dx = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$



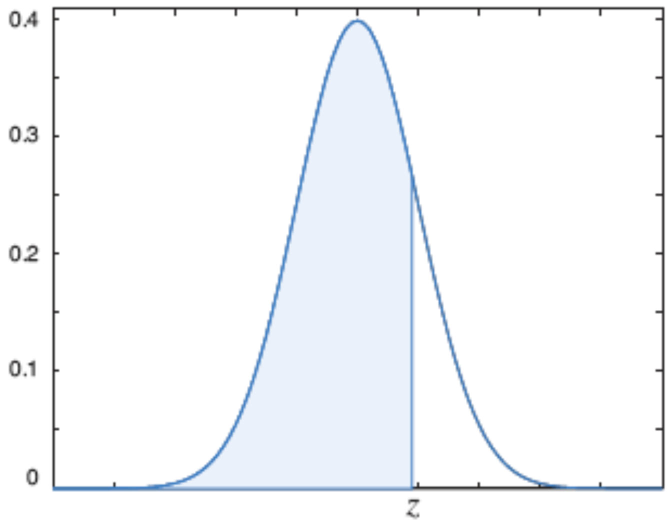
$$F(-z) = 1 - F(z)$$

Standard Normal Distribution

Partial Table of Values for *F(z)*:

Table 14.2

z	$F(z)$	z	$F(z)$
-4	0.000032	1	0.841345
-3	0.001350	2	0.977250
-2	0.022750	3	0.998650
-1	0.158655	4	0.999968
0	0.500000	5	0.999999



The Normal and the Standard Normal Distributions

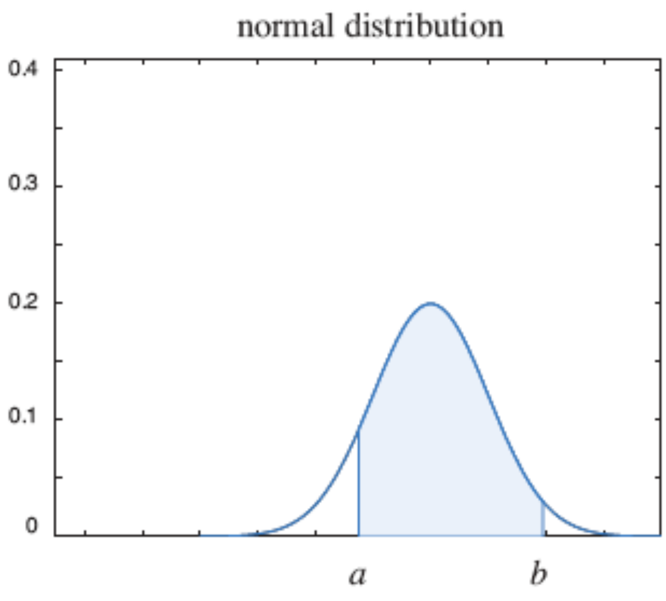
Theorem:

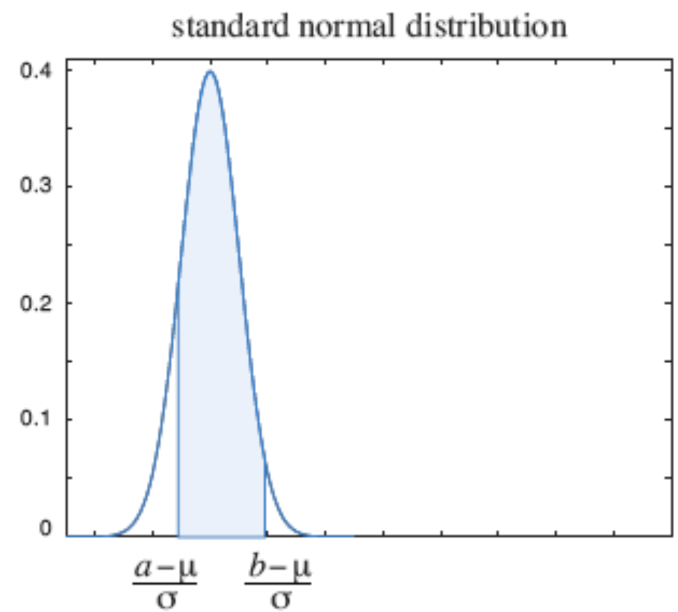
Assume that $X \sim N(\mu, \sigma^2)$. The random variable $Z = (X - \mu)/\sigma$ has the standard normal distribution, i.e., $Z \sim N(0,1)$.

So then $P(a \leq X \leq b) = P(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma})$.

The Normal and the Standard Normal Distributions

In words, the area under the normal distribution density function between *a* and *b* is equal to the area under the standard normal density function between $(a-\mu)/\sigma$ and $(b-\mu)/\sigma$.





The Normal and the Standard Normal Distributions

Example #10:

Let $X \sim N(-2,4)$; find $P(-3 \leq X \leq 1)$.

Example #30:

Let $X \sim N(2,144)$; find a value of x that satisfies $P(X > x) = 0.3$.

z	$F(z)$	z	$F(z)$	z	$F(z)$	z	$F(z)$
0	0.500000	1	0.841345	2	0.977250	3	0.998650
0.05	0.519938	1.05	0.853141	2.05	0.979818	3.05	0.998856
0.1	0.539828	1.1	0.864334	2.1	0.982136	3.1	0.999032
0.15	0.559618	1.15	0.874928	2.15	0.984222	3.15	0.999184
0.2	0.579260	1.2	0.884930	2.2	0.986097	3.2	0.999313
0.25	0.598706	1.25	0.894350	2.25	0.987776	3.25	0.999423
0.3	0.617911	1.3	0.903200	2.3	0.989276	3.3	0.999517
0.35	0.636831	1.35	0.911492	2.35	0.990613	3.35	0.999596
0.4	0.655422	1.4	0.919243	2.4	0.991802	3.4	0.999663
0.45	0.673645	1.45	0.926471	2.45	0.992857	3.45	0.999720
0.5	0.691462	1.5	0.933193	2.5	0.993790	3.5	0.999767
0.55	0.708840	1.55	0.939429	2.55	0.994614	3.55	0.999807
0.6	0.725747	1.6	0.945201	2.6	0.995339	3.6	0.999840
0.65	0.742154	1.65	0.950529	2.65	0.995975	3.65	0.999869
0.7	0.758036	1.7	0.955435	2.7	0.996533	3.7	0.999892
0.75	0.773373	1.75	0.959941	2.75	0.997020	3.75	0.999912
0.8	0.788145	1.8	0.964070	2.8	0.997445	3.8	0.999928
0.85	0.802337	1.85	0.967843	2.85	0.997814	3.85	0.999941
0.9	0.815940	1.9	0.971283	2.9	0.998134	3.9	0.999952
0.95	0.828944	1.95	0.974412	2.95	0.998411	3.95	0.999961
						4	0.999968

Application

Example:
Intelligence quotient (IQ) scores are distributed normally with mean 100 and standard deviation 15.

(a) What percentage of the population has an IQ score between 85 and 115?

(b) What percentage of the population has an IQ above 140?

(c) What IQ score do 90% of people fall under?

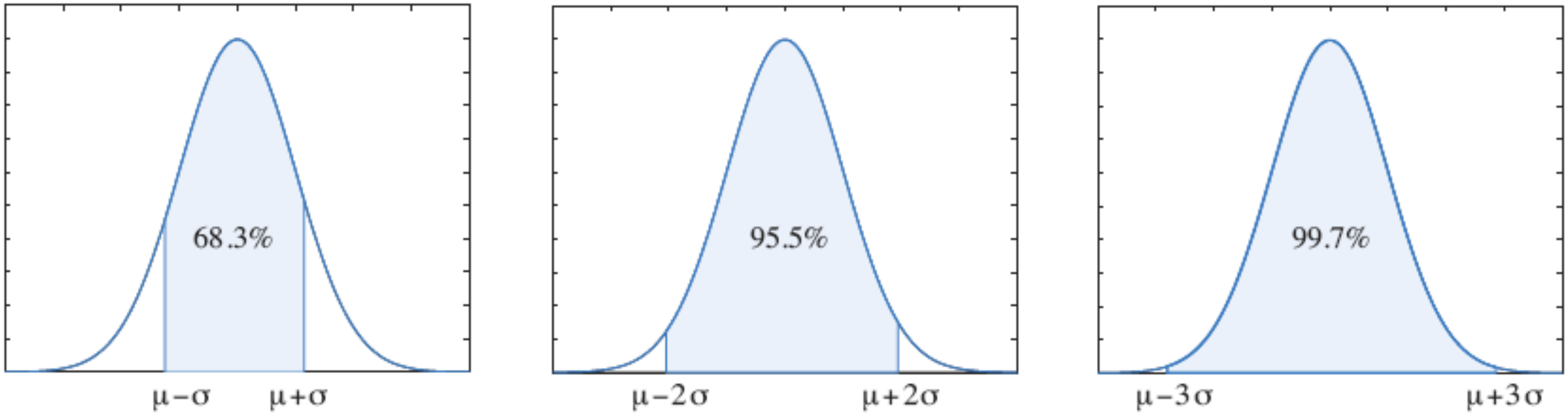
68-95-99.7 Rule

If *X* is a continuous random variable distributed normally with mean μ and standard deviation σ, then

$$P(\mu - \sigma \leq X \leq \mu + \sigma) = 0.683$$

$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = 0.955$$

$$P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) = 0.997$$



68-95-99.7 Rule

In words, for a normally distributed random variable:

68.3% of the values fall within one standard deviation of the mean.

95.5% of the values fall within two standard deviations of the mean.

99.7% of the values fall within three standard deviations of the mean.