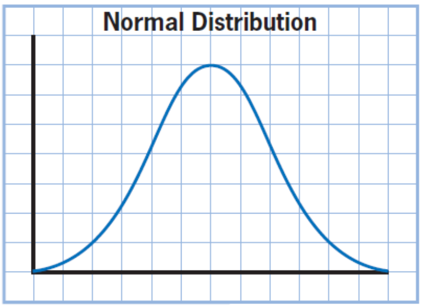
Review

Probability Distributions for Continuous Variables

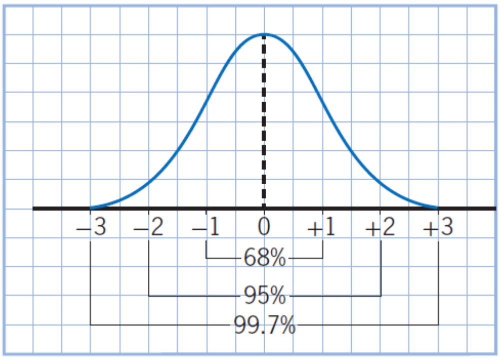
Page 438 #s 1 - 5

- Some situations result in discrete data. These are often whole numbers.
- Some situations result in continuous data over a range. Continuous data include fractional or decimal values. Continuous distributions are often the result of measurements.
- The probability that a variable falls within a range of values is equal to the area under the probability density graph for that range of values.
- You can represent a sample of values for a continuous random variable using a frequency table, a frequency histogram, or a frequency polygon.
- The frequency polygon approximates the shape of the probability density distribution.

- The frequency polygon approximates the shape of the frequency distribution.
- You must use a range of values to determine the theoretical probability for a continuous random variable.
- The probability that a continuous random variable takes on any single value is zero.
- A normal distribution is a probability distribution around a central value, dropping off symmetrically to the right and left, forming a bell-like shape.
- You can determine the probability that a variable will lie within a range of values by finding an area under the normal distribution.
- You can use z-scores to determine probabilities, either from a table or by using technology.



- You can use the normal distribution to model the frequency and probability density distributions of continuous random variables.
- The normal distribution has a central peak, and is symmetric about the mean.
- The mean and median are equal.
- About 68% of the data values are within one standard deviation of the mean, about 95% of the data values are within two standard deviations of the mean, and about 99.7% are within three standard deviations of the mean.



- The confidence level is the probability that a particular statistic is within the range indicated by the margin of error.
- Commonly used confidence levels are 90%, 95%, and 99%. These are related to the z-scores of the distribution.
- A margin of error is the range of values that a particular statistic is said to be within. For a statistic with probability p, the margin of error is $E = z\sqrt{\frac{p(1-p)}{n}}$.
- The greater the sample size, the smaller the margin of error. The smaller the margin of error, the greater the accuracy of the measurement.
- The confidence interval is the range of possible values of the measured statistic.
- For repeated samples of the same size taken from the same population with a normal distribution, the standard deviation of the sample means is $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ and the margin of error is $E = z\frac{\sigma}{\sqrt{n}}$.

- As the number of trials increases, a binomial distribution takes on the characteristics of a normal distribution.
- If the values of np and nq are both greater than 5, you can approximate the binomial distribution using a normal distribution.
- If the sample size is small compared to the population size, a hypergeometric distribution takes on the characteristics of a normal distribution.
- If the sample size is less than one-tenth of the population size, $n < \frac{1}{10}NP$, you can approximate the hypergeometric distribution using a normal distribution.
- You must use a continuity correction when approximating a discrete distribution with a normal distribution. For example, if you want the probability of rolling a 6 exactly 3 times, you must calculate $P(2.5 \leq X \leq 3.5)$. If you want the probability of rolling at least 3 sixes, you must calculate $P(X \geq 2.5)$. However, if you want the probability of rolling more than 3 sixes, you must calculate $P(X \geq 3.5)$.

Solutions

1. The chart shows systolic blood pressure measurements for a sample of 30 athletes.

Systolic Blood Pressure (mmHg)					
120	118	118	115	120	119
124	120	115	122	121	117
121	124	116	123	118	121
117	123	116	116	123	122
119	115	124	117	122	119

a) Is the distribution uniform?

b) What is the height of the probability distribution?

c) What is the probability that an athlete's blood pressure is 120 mmHg or less?

d) How would you expect the distribution to differ if the general population was measured instead of a select group? Give reasons for your answer.

Pressure (mmHg)	Frequency
115	3
116	3
117	3
118	3
119	3
120	3
121	3
122	3
123	3
124	3
125	3

a) Putting the data into a frequency table reveals that the distribution is uniform. This is because each pressure has the same frequency.

b) The area of a probability distribution is always 1. The area of the rectangle equals the area. Therefore the base x height = 1. The base is from 115 to 125 which is 10.

Base x Height = Area

$10(h) = 1$

$h = 0.1$

The height of the probability distribution is 0.1

c) $P(115 \leq x \leq 120)$ = area under the graph

$= (120 - 115)(0.1)$

$= 0.5$

The probability that an athlete's blood pressure is 120 mmHg or less is 0.5

d) I would expect there to a larger range of pressures and that the distribution would no longer be uniform due to different levels of health.

2. At a provincial high school track meet, 40 contestants participated in the discus event. The table shows the results.

Distance (m)	Frequency
25–30	1
30–35	3
35–40	5
40–45	9
45–50	6
50–55	6
55–60	5
60–65	3
65–70	1
70–75	1

a) Sketch a frequency histogram for these data.

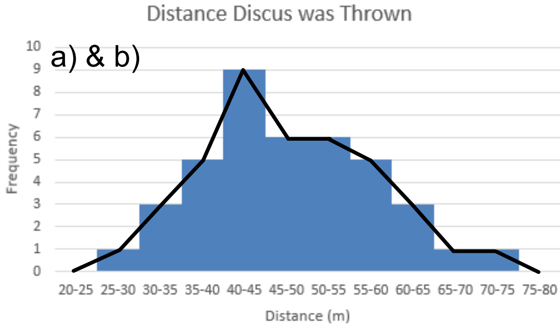
b) Sketch a frequency polygon.

c) Estimate the mean distance thrown.

d) Add a relative frequency column to the table.

e) What is the probability that a contestant throws the discus more than 60 m?

f) How would you expect the distribution to differ if the data were taken from students in a physical education class instead of those who were participating in a track meet? Give reasons for your answer.



c) Total of (frequency x midpoints) ÷ 40

Estimate is about 47.8 metres.

Distance (m)	Frequency	Relative Freq
20-25	0	0.000
25-30	1	0.025
30-35	3	0.075
35-40	5	0.125
40-45	9	0.225
45-50	6	0.150
50-55	6	0.150
55-60	5	0.125
60-65	3	0.075
65-70	1	0.025
70-75	1	0.025
75-80	0	0.000

e) P(more than 60 metres) = 0.075 + 0.025 + 0.025

= 0.125

The probability of an athlete throwing more than 60 metres is 0.125

f) I would expect the distribution to shift to the left as the skill level is likely to be lower. The data would move to the left on the histogram. The distribution would also have a lower mean.

3. A large number of airplanes in Canada are built from kits. A company advertises that building its kit takes a mean time of 300 h, with a standard deviation of 40 h. The company has sold 120 such kits.

a) What percent of the builders are expected to take longer than 400 h to complete their kits?

b) How many builders are expected to finish their kits in less than 250 h?

c) If 120 kits were given to people selected at random from the community, how would the distribution of times change? Give reasons for your answer.

x = 400, $\overline{x}$ = 300, s = 40

a) Find the z-score and then use the table

$$z = \frac{x - \overline{x}}{s}$$

z = (400 - 300)/40

z = 2.5

Using the table gives 0.9938

Therefore, the percentage of builders taking LONGER than 400 hours is 1 - 0.9938 = 0.0062 = 0.62%

b) x = 250, $\overline{x}$ = 300, s = 40

$$z = \frac{x - \overline{x}}{s}$$

z = (250 - 300)/40

z = -1.25

Using the table gives 0.1056

Therefore, the percentage of builders taking less than 250 hours is 10.56%

c) If the kits were given out at random I would expect the distribution to shift to the right. The reason for this is that the skill level to assemble the kits will likely be less than those who ordered them. There would be an increase in the mean time taken as well as the standard deviation.

4. An industrial wind turbine is designed to produce a maximum of 1.8 MW of power in a 60 km/h wind. Lab tests of 120 turbines showed that seven did not meet this standard.

a) Determine a 99% confidence interval for the proportion of turbines from the plant that do not meet the standard.

b) Use a standard statistical format suitable for a report to state the results of the lab tests.

$p = 7 \div 120$ $z = 2.576$

$p = 0.0583...$ $n = 120$

Confidence Level	z-Score
90%	1.645
95%	1.960
99%	2.576

a) $E = z\sqrt{\dfrac{p(1-p)}{n}}$

$$E = 2.576\sqrt{\frac{0.0583(1 - 0.0583)}{120}}$$

$E = 0.055099...$

The margin of error at a 99% confidence level is about 5.5%

Lower limit = 5.8% - 5.5%
= 0.3%

Upper limit = 5.8% + 5.5%
= 11.3%

The 99% confidence interval is 0.3% to 11.3%

b) 5.8% of industrial wind turbines will not meet the standard. This estimate is considered correct within ±5.5%, 99 times out of 100.

5. A drug prescribed to prevent gout produces side effects in 2% of patients who take it. A new formulation is tested on 1500 patients, and 20 suffer side effects.

a) Could this distribution be reasonably modelled using a normal approximation? Give reasons for your answer.

b) Determine the mean and standard deviation of the normal approximation.

c) What is the probability that the new formulation is no better than the original? Use the normal approximation to determine the probability that at most 20 patients suffered side effects.

d) Is the company justified in claiming that the new formulation results in side effects in about 1.3% of the patients who take it? Give reasons for your answer.

a) To use the normal approximation both np and nq need to be greater than 5.

np = 1500(0.02) nq = 1500(0.98)

np = 30 nq = 1470

Yes, it is reasonable to use the normal approximation.

b) $\mu = np$
= 1500(0.02)
= 30

$\sigma = \sqrt{npq}$
$= \sqrt{(1500)(0.02)(.98)}$
= 5.422...

The mean is 30, and the standard deviation is about 5.422.

c) Applying the continuity correction we need to find P(-0.5 ≤ x ≤ 20.5).
We need the mean, standard deviation, and the z-scores for -0.5 and 20.5.

x = -0.5, $\bar{x}$ = 30, s = 5.422

$$z = \frac{x - \bar{x}}{s}$$

z = (-0.5 - 30) / 5.422

z = -5.625... ⟶ P(x < -0.5) = 0

x = 20.5, $\bar{x}$ = 30, s = 5.422

$$z = \frac{x - \bar{x}}{s}$$

z = (20.5 - 30) / 5.422

z = -1.844... ⟶ P(x < 20.5) = 0.0329

P(-0.5 < x < 20.5) = 0.0329 - 0
= 0.0329

The probability that at most 20 patients suffered side effects is about 3.3%

d) No the company are not justified in making this claim, since the probability of side effects is higher than the old formulation (3.3% vs 2%).