Camosun College
Department of Mathematics & Statistics

Coursepack for the
STATISTICS
Portion of MATH 193

Contents

Acknowledgements

Many of the examples and exercises in this coursepack were contributed by Leah Howard.

The *Standard Normal Distribution Table* and the *t-Distribution Table* were typeset by Gilles Cazelais.

1 Collection and Representation of Data

The field of statistics deals with the collection, analysis and interpretation of numerical data. In this section we introduce common methods of collecting and presenting this data.

Definition: A population is the set of all measurements of interest and a sample is a subset of the population.

For example, city officials might want to know whether the level of bacteria in the water supply is within safety standards. The entire water supply is the population in this case. Because not all of the water can be checked, answers must be based on the partial information from samples of water that are collected and tested for this purpose.

Because of constraints on time and money, conclusions about a population are usually drawn after observing only a sample. To make accurate conclusions, great care must be taken to choose a sample that is representative of the population and in designing the method by which measurements/observations are to be made on that sample.

Example 1.1. A politician wants to know how Canadians feel about the Employment Insurance (EI) program. He decides to poll 10 of his neighbors and asks them the question: *"Do you want the government to give your tax dollars to people who don't want to work?"* All 10 people answered "No", so he concludes that Canadians would prefer to scrap the EI program.

(a) Why is this an inaccurate conclusion?

(b) What are some ways to redesign this survey so that a more accurate conclusion could be found?

1.1 Combinations and Randomness

There are many possible samples that can be taken from a given population.

Example 1.2. Consider the small population A, B, C, D, E.

(a) Find all samples of size 2 that can be chosen from this population.

(b) Count the samples of size 2 found in part (a).

Definition: A combination is an unordered selection of a subset from a collection of objects.
Notation: nCr is the number of ways to select r objects from a collection of n objects

For example, 5C2 is the number of possible samples of size 2 from a population of size 5. As we saw in Example 1.2, 5C2 = 10.

On the **SHARP EL-531X**: $\boxed{5} \rightarrow \boxed{\text{2nd F}} \rightarrow \boxed{5}$ *to select* $nCr \rightarrow \boxed{2} \rightarrow \boxed{=}$

The proper definition is $n\text{C}r = \dfrac{n!}{(n-r)!r!}$, but we'll simply use the calculator.

Example 1.3. How many samples of size 10 are possible from a population of size 100?

Demonstration: Real vs Fake Coin Flips

Random number generators are often used to ensure true randomness.

On the **SHARP EL-531X**: [2nd F] → [7] *to select RAND* →

- [0] → [=] → [=] → ⋯
  will produce a sequence of random 3-decimal numbers between 0 and 1
- [1] → [=] → [=] → ⋯
  will produce a sequence of random numbers from the set $\{1, 2, 3, 4, 5, 6\}$
- [2] → [=] → [=] → ⋯
  will produce a sequence of random numbers from the set $\{0, 1\}$
- [3] → [=] → [=] → ⋯
  will produce a sequence of random numbers from the set $\{0, 1, 2, 3, \cdots, 100\}$

Excel has two useful functions when it comes to generating random numbers. Typing [=RAND()] produces a number between 0 and 1, and typing [=RANDBETWEEN(a, b)] produces an integer between a and b.

There are also many easy to find random number generators available online.

1.2 Sampling Methods

SIMPLE RANDOM SAMPLE: Every measurement in the population has equal probability of being chosen.

Example: To form a random student committee, assign each student a number and use a calculator's random number generator to select students.

STRATIFIED RANDOM SAMPLE: The population is divided into subpopulations, then a random sample is selected from each subpopulation.

Example: Thirty percent of ball bearings at a factory have 5mm radius and the other 70% have 10mm radius. Say we want a random sample of 50 ball bearings. Take a random sample of 15 of the 5mm ball bearings and a random sample of 35 of the 10mm ball bearings.

Comment: $0.3(50) = 15$ and $0.7(50) = 35$

CLUSTER SAMPLE: Divide the population into clusters and take a random sample of the clusters. ALL measurements in the chosen clusters are included in the sample.

Example: To form a sample of buildings in Victoria, let the city blocks represent the clusters. Take a random sample of the city blocks; all buildings in the chosen blocks are included in the sample.

1-in-k SYSTEMATIC SAMPLE: Randomly select one of the first k measurements in the population and every k-th measurement thereafter.

Example: Ball bearings #3,23,43,63,... from a production line form a 1-in-20 systematic sample.

Comments: The random starting point makes this a random sample. Avoid patterns when choosing k, e.g. all ball bearings produced by same machine.

Example 1.4. Identify the sampling method:

(a) A lightbulb company makes 60W and 100W bulbs; 80% are 60W and the rest are 100W. A random sample of 40 of the 60W bulbs is selected, together with a random sample of 10 of the 100W bulbs.

(b) Engineers in a large city want to perform a random check on red-light cameras in 85 different neighbourhoods. A random sample of 10 neighbourhoods is selected and every red-light camera in the chosen neighbourhoods is inspected.

(c) A random number generator is used to select 12 of 100 shipments for quality-control testing.

(d) Starting with the 11th part, every 25th part coming off the production line is selected for further inspection.

1.3 Representation of Data Sets

Once a sample has been selected from a population, and the measurements made on that sample, the result is usually a large list of numbers.

A useful tabular representation of a data set is a "frequency distribution table" and the most commonly used graphical representation is a "histogram".

Definition: The frequency of a measurement is the number of times it occurs in the data set.

Demonstration: How large is your family

Frequency distribution table: *Histogram:*

Average:

The most recently stated fertility rate in Canada was 1.6. Compare this with our class average family size.

Note: The Appendix contains instructions for making histograms in Excel 2013.

Definition: The relative frequency of a measurement is its frequency divided by the total number of measurements in the data set.

Example 1.5. Find the relative frequency distribution for the demonstration *How large is your family.*

Demonstration: An experiment that looks like a survey

If the data set contains many distinct values, the data is grouped into classes. A class is an interval of data values; classes must be mutually exclusive and all classes must have the same width. In the process of grouping, the detail of the raw data is lost, but the advantage is that a much clearer overall pattern of the data can be obtained. On a histogram, each class is denoted by a representative value along the x-axis. This value is called the class mark, and is usually found by taking the average of the lower and upper class limits.

Note: Excel defines the class mark as the upper class limit.

Example 1.6. A test station measured the loudness of the sound of jets taking off from a certain airport. The decibel (dB) readings measured to the nearest integer for the first 20 jets were as follows:

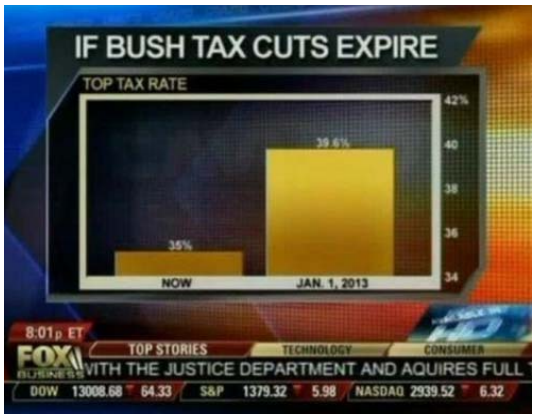
$$102, 115, 93, 105, 108, 110, 120, 94, 101, 103, 90, 110, 109, 101, 115, 119, 95, 108, 98, 114$$

(a) Create a frequency distribution table with 6 classes.

(b) Find the relative frequencies for each class in part (a).

(c) Draw a histogram for the data in part (a).

A common way to use histograms to give a misleading representation of the data is to start the y-axis with a value other than 0.



Additional Notes

Additional Notes

Exercises

1. Consider the following population: $1, 2, 4, 8, 9$.

   (a) How many possible samples of $n = 3$ measurements can be chosen from this population?

   (b) Write out all the possible samples of $n = 3$ measurements.

2. For quality control purposes, 40 machines need to be split randomly into two groups of size 20. Explain how you could use a simple random sample to accomplish this.

3. A beverage company produces one brand of pop, in regular and diet versions. Today the company produced 10,650 cans of regular pop and 4,350 cans of diet pop. A random sample of 600 cans is required from today's production. How many of each type of pop should be included in the sample?

4. Name each sampling method described below:

   (a) Starting with a random ball bearing, every 50th ball bearing coming off the manufacturing line is selected for further inspection.

   (b) Twenty soil samples are in test tubes labelled $1, 2, \ldots, 20$. A random number generator is used to select 5 soil samples for analysis.

   (c) A manufacturing company's drill bits are divided into two sizes: 40% are long and 60% are short. A random sample of 12 long drill bits and a random sample of 18 short drill bits are selected for further inspection.

   (d) A mining company's operations are divided over 23 sites, each containing several mines. A random sample of 3 sites is selected and every mine at the selected sites is investigated.

5. [3, p. 619] A car company tested a new engine and found the following results in 20 tests of the number of litres of fuel used by a certain model for each 100 km travelled:

   $$5.3, 5.8, 5.6, 5.4, 5.9, 5.4, 6.0, 5.8, 5.8, 5.4, 6.3, 5.6, 5.7, 5.6, 5.7, 5.9, 5.5, 6.1, 5.9, 5.8$$

   (a) Make a frequency distribution table with 5 classes.

   (b) Find the relative frequencies for each of the classes in part (a).

   (c) Draw a histogram for the data in part (a).

2 Summarizing Data

A common way to summarize a data set is with a single value considered to be the centre of the data, along with a measure of how spread out the data is from that centre. Two ways to define the centre of a data set are *mean* and *median*, and the most widely used measure of spread is the *standard deviation*.

Decimal place convention: mean, median and standard deviation are usually rounded off to one more decimal place than was present in the original data.

Units: mean, median and standard deviation all have the same unit as the original data.

2.1 Mean

The mean is the number most people simply call the "average". It is sometimes referred to as the *arithmetic mean*.

Definition: The mean of a data set $x_1, x_2, \cdots, x_n$ is the average value

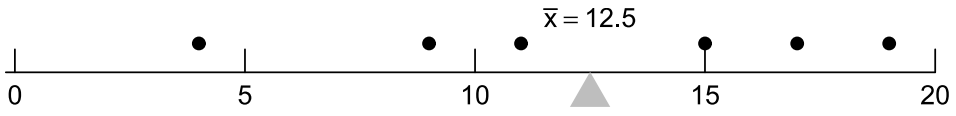
$$\text{mean} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum x}{n}.$$

Notation: μ is used to denote a population mean and $\bar{x}$ is used for a sample mean. If a data set is not labelled as a sample, we will assume that it is a population.

The SHARP EL-531X has a function to calculate the mean of a data set. Instructions are in the Appendix.

Example 2.1. Find the mean of the sample $11, 9, 17, 19, 4, 15$.

Physical Interpretation of the Mean: If each value in a data set is represented along a weightless horizontal axis by a ball of equal weight, then the mean corresponds to the centre of inertia or balance point of the data. For the previous example, this looks like [2, p. 30]:



Example 2.2. A student has test marks 58, 63, 71. What mark on his 4th test gives him an average of 70?

Example 2.3. If the following two samples are combined into one sample, find the mean.

	sample size	$\bar{x}$
Sample 1	43	71
Sample 2	26	68

If a data set is given using a frequency distribution table instead of a list of numbers,

value	frequency
x_1	f_1
x_2	f_2
$\vdots$	$\vdots$
x_n	f_n

then

$$\boxed{\text{mean} = \frac{x_1 f_1 + x_2 f_2 + \cdots + x_n f_n}{f_1 + f_2 + \cdots + f_n} = \frac{\sum xf}{f}}.$$

Example 2.4. Find the mean for the following sample:

Temperature (°C)	Frequency
22	11
23	6
25	3

If a data set is given using relative frequencies,

value	relative frequency
x_1	r_1
x_2	r_2
$\vdots$	$\vdots$
x_n	r_n

then

$$\boxed{\text{mean} = x_1r_1 + x_2r_2 + \cdots + x_nr_n = \sum xr}.$$

Example 2.5. Find the mean for the following sample:

mass (g)	relative frequency
82	0.1
85	0.35
86	0.5
88	0.05

2.2 Median

The second way to define the centre of a data set is to order the values from smallest to largest and then simply select the middle value. This way the data set contains the same number of values that are larger than the median and that are smaller than the median. One convenient property of the median is that it eliminates the effect of outliers (values that are much larger or smaller than the rest of the data).

Definition: If the data set $x_1, x_2, \cdots, x_n$ is listed in order from smallest to largest, then the median is defined as

$$\text{median} = \begin{cases} \text{middle value} & \text{if } n \text{ is odd} \\ \text{average of the 2 middle values} & \text{if } n \text{ is even} \end{cases}.$$

Specifically, if n is odd then the median is the value in position $\dfrac{n+1}{2}$ in the list. And if n is even, then the median is the average of the values in position $\dfrac{n}{2}$ and $\dfrac{n}{2}+1$.
There is no standard notation for the median.

Example 2.6. Find the median of the following data sets:

(a) 2, 9, 11, 5, 6.

(b) 2, 9, 11, 5, 6, 10.

Example 2.7. Find the median for the sample in Example 2.4:

Temperature (°C)	Frequency
22	11
23	6
25	3

If a data set is given using relative frequencies,

value	relative frequency
x_1	r_1
x_2	r_2
$\vdots$	$\vdots$
x_n	r_n

where the values $x_1, x_2, \cdots, x_n$ are listed in order from smallest to largest, then the median is x_i where i is the smallest index for which $r_1 + r_2 + \cdots + r_i \geq 0.5$. That is, we add the relative frequencies in order until 0.5 is first reached or exceeded, and the median is the corresponding value.

Example 2.8. Find the median for the sample in Example 2.5:

mass (g)	relative frequency
82	0.1
85	0.35
86	0.5
88	0.05

2.3 Standard Deviation

Now we turn our attention to measure how spread out a data set is. The spread is small if the values are all bunched close to the mean, and it is large if the values are scattered widely. An intuitive way to do this would be to take the average of the difference of each value from the mean, but this average is always 0.

Definition: The population $x_1, x_2, \cdots, x_n$ has a population variance

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}$$

and a standard deviation (SD)

$$\sigma = \sqrt{\sigma^2}.$$

The SHARP EL-531X has a function to calculate population SD. Instructions are in the Appendix.

Example 2.9. Find the population SD of 2, 5, 8, 9.

Definition: The sample $x_1, x_2, \cdots, x_n$ has a sample variance

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

and a standard deviation (SD)

$$s = \sqrt{s^2}.$$

For the purpose of calculating SD, if a given data set is not labelled as a sample, we assume that it is a population.

Example 2.10. Find the sample SD of 12, 13, 17.

Example 2.11. Which sample is more spread out?

(a) 1, 4, 10

(b) 31, 36, 38

A data set is considered to be accurate if the mean is close to the target value, and it is precise if the variance or SD is small.

Example 2.12. Two machines are filling 355 mL cans of pop. A sample of volumes has the following means and variances (in mL).

	Machine 1	Machine 2
$\overline{x}$	355.8	355.2
s^2	0.3	1.4

(a) Which machine is more accurate?

(b) Which machine is more precise?

Example 2.13. Let a population consist of the salaries at a small engineering firm, where the second highest salary is \$50,000 less than the highest salary. What happens to the mean, median and SD in each situation:

(a) Each employee get a \$2,000 raise.

(b) Each employee's salary is doubled.

(c) The highest salary is decreased by \$10,000.

Example 2.14. Compare the mean, median and SD for the results from the demonstration *An experiment that looks like a survey* in Section 1.

Additional Notes

Additional Notes

Exercises

1. Calculate the mean and median for the following population of test scores:

$$49, 61, 67, 68, 74, 77, 79, 82, 91$$

2. Calculate the mean and median for the following sample of masses (in grams):

$$22, 25, 28, 23, 22, 27, 27, 29$$

3. A math class has four tests. A student has earned the following marks on the first three tests: 52, 69, 73. What mark does the student need on the fourth test in order to have an average of 70 on the four tests?

4. Leah's class of 29 students has a mean test score of 78. Pat's class of 36 students has a mean test score of 72. Find the mean test score if the two sets of test scores are combined into one population.

5. Calculate the mean and median for the sample of temperature readings:

Temperature (°C)	Frequency
36.8	4
37.1	6
37.2	2
37.4	5
37.7	8

6. Find the mean and median for the following sample of house prices:

Price ($)	Relative Frequency
600,000	0.1
800,000	0.45
1,000,000	0.3
1,200,000	0.1
1,400,000	0.05

7. Calculate the variance and standard deviation for the following population of test scores: $71, 76, 76, 79, 83$.

8. Calculate the variance and standard deviation for the following sample of masses (in grams): $22, 25, 27, 28$.

9. Which data set is more spread out, or are they equally spread out?

   (a) Set 1: $5, 7, 11, 13, 19$ Set 2: $15, 17, 21, 23, 29$.

   (b) Set 1: $5, 7, 11, 13, 19$ Set 2: $10, 14, 22, 26, 38$.

10. A small engineering firm has three employees with the following salaries: $\$35,000$, $\$60,000$ and $\$100,000$. State what happens to the mean, median and standard deviation of the salaries in each situation below (i.e. do they increase, decrease or stay the same?)

    (a) Each employee gets a $\$5,000$ raise.

    (b) Each employee gets a 10% raise.

    (c) The lowest salary is bumped up to $\$50,000$.

3 Probability

In statistics, an *experiment* is any procedure than can be repeated and has a well-defined set of possible outcomes. An experiment can be something as simple as rolling a die and noting the number, or something as complicated as finding the mass of an electron.

Definition: The sample space of an experiment is the set of all possible outcomes.
An event is a subset of a sample space.
Notation: A sample space is usually denoted by S.
If E is an event, then $n(E)$ denotes the number of outcomes in E.

Example 3.1. An experiment consists of flipping a coin 3 times and noting if it's heads (H) or tails (T).

(a) Find the sample space.

(b) Let E be the event that exactly two heads occur. Find $n(E)$.

3.1 Probabilities Involving Equally Likely Outcomes

There are a number of ways to define the probability of an event. In this section we consider the simplest case.

Definition: If each outcome of an experiment is *equally likely* then the probability of an event E is the ratio

$$P(E) = \frac{n(E)}{n(S)},$$

where S is the sample space of the experiment.

Since E is a subset of S, we always have $0 \leq P(E) \leq 1$. For a finite sample space, $P(E) = 0$ if and only if $E = \varnothing$; and $P(E) = 1$ if and only if $E = S$.

Example 3.2. An experiment consists of flipping a coin 3 times and noting if it's heads (H) or tails (T). What is the probability of getting exactly two heads?

Example 3.3. An experiment consists of randomly selecting a number between 1 and 40 (inclusive). Find the probability of getting a multiple of 5 or 7.

Example 3.4. An experiment consists of rolling a pair of standard 6-sided dice and noting the number. Find the probability of getting a sum of at most 5.

Example 3.5. Four sidewalk squares have the following number of cracks: 4, 6, 7, 9. Pick two of the squares at random (order doesn't matter). Find the probability that they have at least 15 cracks in total.

3.2 Venn Diagrams

Demonstration: Coffee and Spicy Food

	like coffee	don't like coffee
like spicy food		
don't like spicy food		

(a) Represent this data in a Venn diagram.

(b) Find the probability that a randomly chosen student likes coffee.

(c) Find the probability that a randomly chosen student likes coffee but doesn't like spicy food.

(d) Find the probability that a randomly chosen student likes coffee or spicy food.

Example 3.6.

In a class of 45 students, 26 have jobs and 17 have cars. Of those who don't have a car, 10 have jobs. Find the probability that a student has:

(a) a car or a job.

(b) a car but not a job.

Example 3.7. On any given day, the probability that Machine I breaks down is 4%, the probability that Machine II breaks down is 7%, and the probability that both machines break down is 2%. Find the probability that Machine II breaks down and Machine I doesn't.

3.3 Probability Rules

- $n(A \text{ or } B) = n(A) + n(B) - n(A \text{ and } B)$

  $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

- $n(\text{not } A) = n(S) - n(A)$

  $P(\text{not } A) = 1 - P(A)$

- If Experiment 1 has n possible outcomes and Experiment 2 has m possible outcomes, then the experiment consisting of Experiment 1 followed by Experiment 2 has

  $$n \cdot m$$

  possible outcomes.

Example 3.8. Find the number of possible outcomes when a standard 6-sided die is rolled

(a) twice.

(b) 5 times.

Example 3.9. A password consists of 7 digits, each chosen from 0,1,2, $\cdots$, 9. Find the

(a) total number of passwords possible.

(b) number of passwords that end with 3.

(c) number of passwords that don't end with 3.

(d) the probability that a password starts with 4.

(e) the probability that a password doesn't start with 4.

(f) the probability that a password doesn't start with 32.

(g) the probability that a password contains a least one 4.

(h) the probability that a password starts with 29 or ends with 1.

Additional Notes

Additional Notes

Exercises

1. An experiment consists of flipping a coin three times. Find the probability that you will get:

   (a) at most one head.

   (b) exactly two tails.

2. Five employees have the following years of experience: $1, 3, 7, 11$ and 13. If two of these employees are randomly selected for a project, what is the probability that they have at least 15 years of experience in total?

3. An experiment consists of rolling a pair of 4-sided dice. Find the probability that the two rolls:

   (a) sum to 4.

   (b) sum to 3 or 4.

   (c) don't sum to 6.

4. You will be assigned two of four different products to analyze. Call the products A,B,C and D. Find the probability that:

   (a) you are assigned products A and C.

   (b) product B is assigned to you.

   (c) product D is not assigned to you.

5. An experiment consists of randomly selecting a number from 1 to 30 (inclusive). What is the probability that the number is divisible by 3 or 5?

6. Below is the make-up of employees at an engineering firm.

	Male	Female
Contract	37	41
Permanent	98	55

   Find the probability that an employee is:

   (a) female.

   (b) male or on contract.

   (c) female and permanent.

7. The probability that Device A fails is 2.3%. The probability that Device B fails is 3.1%. The probability that both devices fail is 0.3%. Find the probability that neither device fails.

8. Out of 62 job applicants, 35 have their P.Eng. qualification and 23 are fluent in French. Of those who are fluent in French, 17 have their P.Eng. qualification. What is the probability that an applicant has their P.Eng. but does not speak French?

9. Canadian postal codes have the following format:
letter-number-letter number-letter-number, where numbers 0-9 are used. The letters D,F,I,O,Q,U are never used; the letters W and Z cannot be used in the first position.

   (a) How many postal codes are possible?

   (b) How many postal codes begin with A?

   (c) How many postal codes don't end with 0?

   (d) How many postal codes begin with B or end with 9?

10. Your employer's computer network requires a case-sensitive alphanumeric password that is four symbols long. Find the probability that a password:

   (a) contains at least one number.

   (b) starts or ends with d.

4 Discrete Random Variables

Demonstration: Surprize Quiz

correct answers	frequency	relative frequency
0		
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

Relative frequency histogram:

Definition: A discrete random variable is a function that assigns a number to each outcome of an experiment with a finite sample space.
Notation: A random variable is usually denoted X, and a generic number assigned by X is denoted x.

In the *Surprize Quiz* demonstration, we were interested in the number of correct answers on each quiz. If C stands for correct and I stands for incorrect, an example of an outcome in this experiment is an ordered list like ICIICIIIC. The random variable should assign to each outcome of the experiment the number that is of interest, namely the number of correct answers. That is, we let $X =$ the number of correct answers. Then $X = 3$ for the outcome ICIICIIIC.

4.1 Probability Distributions

Discrete random variables are best represented using a table.

Definition: A probability distribution of a discrete random variable X is a table listing the probability of each possible value of X:

x	$P(x)$

Notation: $P(x) = P(X = x)$ is the probability that a randomly chosen outcome will have the value x.

Note that for experiments with equally likely outcomes, $P(x)$ is simply the relative frequency of x.

All discrete random variables have the property

$$\sum P(x) = 1.$$

Example 4.1. Find the probability distribution of $X =$ the number of correct answers in the *Surprize Quiz* demonstration.

Example 4.2. Find the probability distribution of X = the number of heads in 3 coin tosses.

4.2 Expected Value and Standard Deviation

In Section 2.1 we looked at the mean of a data set given as a relative frequency table. The mean of a discrete random variable is defined in the same manner.

Definition: For a discrete random variable, X,

- the mean or expected value of X is

$$\mu = E(X) = \sum xP(x)$$

- the variance of X is

$$\sigma^2 = E(X^2) - \mu^2, \text{ where } E(X^2) = \sum x^2 P(x)$$

- the standard deviation (SD) of X is

$$\sigma = \sqrt{\sigma^2}$$

The SHARP EL-531X calculator can calculate the mean and SD for a probability distribution. Instructions can be found in the Appendix.

Example 4.3. Find the mean and SD of X = the number of correct answers in the *Surprize Quiz* demonstration. Use the probability distribution from Example 4.1.

Since each outcome of an experiment is assigned exactly one number, $P(a \leq X \leq b)$ is simply the sum of all the probabilities for x-values that fall within $[a, b]$.

Example 4.4. Given the following probability distribution

x	$P(x)$
-5	0.15
-2	0.2
1	0.4
6	0.25

find:

(a) $P(-2.5 \leq X \leq 2.5)$.

(b) the mean of X.

(c) the variance of X.

(d) the standard deviation of X.

(e) the probability that an x-value lies within one standard deviation of the mean; i.e. $P(\mu - \sigma \leq X \leq \mu + \sigma)$.

Example 4.5. Project 1 has a 35% chance of earning \$0, a 50% chance of earning \$300,000, and a 15% chance of earning \$800,000.

Project 2 has a 60% chance of earning \$0 and a 40% chance of earning \$1,000,000.

(a) Find the probability distributions of the earnings for each project.

(b) Find the expected earnings for each project.

(c) Find the standard deviation of earnings for each project.

(d) Which project has higher expected earnings?

(e) In terms of earnings, which project is riskier?

Example 4.6. Suppose you want to insure a \$2,000 tablet against theft for one year by paying a premium m, and that the probability of theft is 4.7%.

(a) Find the probability distribution of the insurance company's gain.

(b) Find the premium (i.e. the value of m) if the insurance company expects to gain \$40.

Additional Notes

Additional Notes

Exercises

1. Consider the probability distribution of a random variable X:

x	$P(x)$
9.12	0.41
10.89	0.28
12.31	0.13
14.22	0.11
15.06	0.07

   Find:

   (a) $P(10 \leq X \leq 15)$.

   (b) the mean (or expected value) of X.

   (c) the variance of X.

   (d) the standard deviation of X.

   (e) the probability that a value of X lies within two standard deviations of the mean.

2. Consider the probability distribution of a random variable Y:

y	$P(y)$
-4	0.4
-3	0.2
2	0.1
7	0.3

   Find:

   (a) $P(Y < 2)$.

   (b) the mean (or expected value) of Y.

   (c) the variance of Y.

   (d) the standard deviation of Y.

   (e) the probability that a value of Y lies within one standard deviation of the mean.

3. Project A has a 60% probability of earning $\$10,000$, a 30% probability of earning $\$5,000$ and a 10% probability of earning nothing. Project B has an 80% probability of earning $\$25,000$ and a 20% probability of earning nothing.

   (a) Find the expected earnings of Project A.

   (b) Find the expected earnings of Project B.

   (c) Find the variance of earnings for Project A.

   (d) Find the variance of earnings for Project B.

(e) In terms of earnings, which project is more risky? (Which project's earnings has a larger variance?)

4. Your engineering firm is considering competing for Project Alpha. The cost of competing for Project Alpha is $\$10,000$. The firm has a 35% probability of success, which will mean revenue of $\$80,000$.

   (a) Find the probability distribution of the earnings from Project Alpha, where earnings=revenue−cost.

   (b) Find the expected earnings.

   (c) Find the standard deviation of the earnings.

   (d) Project Beta's earnings has a standard deviation of $\$25,000$. In terms of earnings, which project is more risky: Project Alpha, or Project Beta?

5. You want to insure a used car worth $\$4,000$ against theft (not damage) for one year by paying a premium m. The probability of theft during the year is 1.3%.

   (a) Find the probability distribution of the insurance company's gain.

   (b) Find the premium if the insurance company expects to gain \$60.

5 Binomial and Poisson Distributions

In this section we will look at the probability distributions of two useful discrete random variables.

5.1 The Binomial Distribution

Many statistical experiments involve repeated trials. Here we are interested in a series of "identical success/failure trials", by which we mean:

1. each trial has two possible outcomes, which are called "success" and "failure" (a success is not necessarily the desirable outcome, it is simply the outcome of interest in the experiment),
2. the probability of success is the same for each trial, and
3. the outcome of a trial is independent of the outcome of any other trial.

A series of trials with these features is often referred to as a *Bernoulli trial*.

Definition: Let $X =$ the number of successes in n identical success/failure trials. Then X has the binomial distribution with

$$P(x) = nCx \cdot p^x \cdot q^{n-x},$$

where p is the probability of success on a single trial, and $q = 1-p$ is the probability of failure on a single trial.

For example, if we consider *all* possible ways to answer the *Surprize Quiz* from Section 4 with a success being a correct answer, then

$$X = \text{ the number of correct answers}$$

has a binomial distribution with $n = 10$, $p = \frac{1}{3}$ and $q = \frac{2}{3}$. So that the probability of getting 4 correct answers is

$$P(4) = 10C4 \cdot \left(\frac{1}{3}\right)^4 \cdot \left(\frac{2}{3}\right)^6 \approx 0.23.$$

Why is this probability different from the $P(4)$ we calculated for the demonstration?

Example 5.1. [1, p. 53] In genetics, a *dihybrid cross* is a cross between two varieties of the same species that differ in two observed traits. A 1923 genetics article reported a cross between yellow-wrinkled and green-round peas. The variable of interest was

$$X = \text{ the number of YR (yellow and round) peas in a pod.}$$

Mendelian laws of inheritance imply that the probability of an individual pea being YR is $p = \frac{9}{16}$.

(a) Find the probability distribution for pods with 6 peas.

(b) Draw the histogram for the distribution in part (a).

Note: Histograms for binomial distributions are symmetric if and only if $p = \frac{1}{2}$.

Example 5.2. An experiment consists of rolling a standard 6-sided die 13 times. Find the probability of rolling at most two 5s or 6s.

Example 5.3. A solar panel installation company makes the claim that in 90% of their installations, the utility bill is reduced by at least one third. If we interpret this to mean that for any installation, the probability of the utility bill being reduced by at least one third is 0.9, find the probability that the utility bill will be reduced by at least one third in

(a) exactly 9 of the next 10 installations.

(b) at least 8 of the next 10 installations.

A convenient online tool for calculating and graphing binomial distributions is available at `http://homepage.divms.uiowa.edu/~mbognar/applets/bin.html`

5.2 The Poisson Distribution

One of the main features of the binomial distribution is that there are a fixed number of trials. For counts that do not have a natural upper bound given by the number of trials, the Poisson distribution is often used as a model.

Definition: Let X = the number of occurrences of an event in a unit of time or space, with λ = the average number of occurrences of that event in that unit of time or space. Then the Poisson distribution has probabilities given by

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!}.$$

The SHARP EL-531X has a button for the factorial function. To calculate $5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1$, use $\boxed{5} \rightarrow \boxed{\text{2nd F}} \rightarrow \boxed{4}$ *to select* $n!$ $\rightarrow \boxed{=}$.

Example 5.4. Uncle Tom has been crabbing from the Sidney Pier every Sunday for the past 5 years. He has been diligently recording the number of crabs in his trap each week, and found that on average, he catches 4 crabs in an outing. Let X be the number of crabs caught in an outing. We want to find the probability distribution of X.

(a) Why is the binomial distribution not appropriate in this example?

(b) Find the Poisson distribution for $X = 0, 1, 2, \cdots, 10$. Round off each probability to 2 decimal places.

(c) Draw the histogram for the distribution in part (b) .

(d) Find $P(X \leq 10)$. Why is it not 1?

A convenient online tool for calculating and graphing Poisson distributions is available at `http://homepage.divms.uiowa.edu/~mbognar/applets/pois.html`

Example 5.5. It has been found that the average daily traffic volume on a certain quiet street is 7, and can be modelled using the Poisson distribution.

(a) What is the probability that the traffic volume will be at most 5 on a given day?

(b) What is the probability that the traffic volume will at least 5 on a given day?

Example 5.6. Over the past 30 years, there have been an average of 8 days in January with rainfall above 5mm in Victoria. Using a Poisson distribution, find the probability that next January there will be

(a) Exactly 8 days of rainfall above 5mm in Victoria.

(b) Less than 8 days of rainfall above 5mm in Victoria.

(c) More than 8 days of rainfall above 5mm in Victoria.

(d) Repeat (a) to (c) using a binomial distribution and assuming that each January day has a 8/31 chance of having rainfall above 5mm.

Example 5.7. Suppose that the concentration of bacteria in the inner harbour is 3 per 100 mL of water. Use an appropriate distribution to find the probability that there are at most 2 bacteria in a 50 mL sample of water.

Example 5.8. There are an average of 1.8 accidents per week on a certain highway. Use an appropriate distribution to find the probability that there will be at least 4 accidents in the next 2 weeks.

Additional Notes

Additional Notes

Exercises

1. A basketball player makes 72% of his free throws. He does not improve with practice. Find the probability that in his next six free throw attempts:

   (a) he makes exactly five of them.

   (b) he makes at least four of them.

   (c) he makes at least two of them.

2. A multiple choice test has 20 questions, each of which has 3 possible answers. A student guesses randomly on each question. What is the probability that the student gets:

   (a) exactly 6 questions right?

   (b) between 5 and 7 (inclusive) questions right ?

   (c) at most 3 questions right?

3. For a particular cement mix, the average number of cracks per cubic metre of concrete is 1.7. Find the probability that a randomly-chosen cubic metre of concrete has:

   (a) at least one crack.

   (b) at most three cracks.

4. Suppose 400 typos are distributed randomly throughout a textbook that is 1000 pages long. Find the probability that a given page contains:

   (a) exactly two typos.

   (b) more than one typo.

5. A web server receives an average of 3 requests per 15-minute interval. What is the probability that the server receives at most 4 requests in the next hour?

6 Continuous Random Variables

Recall that discrete random variables were used to assign numbers (usually counts of events) to the outcomes in a finite or countably infinite sample space.

If an experiment consists of measuring some quantity that can have any real value, e.g. mass, time, length, ..., then that quantity is a value of a continuous random variable X.

6.1 Probability Density Functions

For a continuous random variable, X, we are interested in the probability that X will take values on an *interval* rather than one specific value.

Definition: For a continuous random variable, X, the probability distribution is described by a probability density function (p.d.f.), $f(x)$, satisfying:

1. $f(x) \geq 0$ for all x in $\mathbb{R}$
2. $\int_{-\infty}^{\infty} f(x)\, dx = 1$
3. $P(a \leq X \leq b) = \int_a^b f(x)\, dx$

We note that

$$P(x) = 0$$

for any constant x, and as a result,

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

Example 6.1. The p.d.f. for a continuous random variable, X, is

$$f(x) = \begin{cases} \frac{1}{8}x & \text{if } 0 < x \leq 2 \\ \frac{1}{4} & \text{if } 2 < x \leq 5 \\ 0 & \text{otherwise} \end{cases}$$

Find:

(a) $P(X = 2.2)$

(b) $P(1 \leq X \leq 3)$

(c) $P(1 < X < 3)$

(d) $P(X > 1.2)$

(e) $P(X < 0.6)$

Example 6.2. Find the value of k that makes $f(x)$ a valid p.d.f.:

$$f(x) = \begin{cases} kx^7 & 1 \leq x \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

Example 6.3. [2, p. 153] The mileage (in thousands of miles) that car owners get with a certain kind of tire is a continuous random variable having the p.d.f.

$$f(x) = \begin{cases} \frac{1}{20}e^{-x/20} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

Find the probabilities that one of these tires will last

(a) Exactly 10,000 miles.

(b) At most 10,000 miles.

(c) Anywhere from 16,000 to 24,000 miles.

(d) At least 30,000 miles.

Example 6.4. [2, p. 153] In a certain city, the daily consumption of electric power (in millions of kilowatt hours) is a continuous random variable with the p.d.f.

$$f(x) = \begin{cases} \frac{1}{9}xe^{-x/3} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

If the city's power plant has a daily capacity of 12 million kilowatt-hours, what is the probability that this power supply will be inadequate on any given day?

6.2 Mean and Standard Deviation

For discrete random variables, we used sums to calculate the mean and SD. Summing over a continuous domain requires integration.

Definition: For a continuous random variable, X, with p.d.f. $f(x)$,

- the mean or expected value of X is

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

- the variance of X is

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

- and the standard deviation (SD) of X is

$$\sigma = \sqrt{\sigma^2}$$

Example 6.5. Find the expected value (aka the average) and SD of the daily consumption of electric power in Example 6.4.

Example 6.6. [2, p. 172] In a certain municipality, the proportion of highway sections requiring repairs in any given year is a continuous random variable with p.d.f.

$$f(x) = \begin{cases} 12x^2(1-x) & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

(a) Graph $f(x)$.

(b) Find the expected proportion of highway sections requiring repairs in any given year.

(c) Find the probability that at most half of the highway sections will require repairs in any given year.

6.3 Two Special Distributions

In this section we will define two special forms of p.d.f.s. In the next section we will study the *Normal Distribution*, which has the most commonly used p.d.f.

Definition: A continuous random variable is a uniform random variable if its p.d.f. has the form

$$f(x) = \begin{cases} \dfrac{1}{b-a} & a < x < b \\ 0 & \text{otherwise} \end{cases}$$

Example 6.7. [2, p. 165] Suppose that the wheel of a train has radius r and that x is the location of a point on its outer edge measured (in cm) along the circumference from some reference point 0. When the brakes are applied, some point on the wheel will make sliding contact with the rail, and heavy wear will occur at that point. For repeated applications of the brakes, it would seem reasonable to assume that x is a value of a uniform random variable with p.d.f.

$$f(x) = \begin{cases} \dfrac{1}{2\pi r} & 0 < x < 2\pi r \\ 0 & \text{otherwise} \end{cases}$$

Otherwise the wheel would eventually have flat spots.

(a) Find the mean of this random variable.

(b) Find the variance of this random variable.

Definition: A continuous random variable is a exponential random variable if its p.d.f. has the form

$$f(x) = \begin{cases} ke^{-kx} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

for some constant k.

Example 6.8. [2, p. 171] The Poisson distribution has many important applications in queing problems, e.g. The number of aircraft arriving at an airport. If in a Poisson process the average number of arrivals per unit of time is λ, then the waiting time between successive arrivals is an exponential random variable with

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

Suppose that on average, 3 trucks arrive per hour to be unloaded at a warehouse, and the number of arrivals per hour is described by the Poisson distribution. Therefore the time (in hours) between the arrival of successive trucks to be unloaded at a warehouse is given by the p.d.f.

$$f(x) = \begin{cases} 3e^{-3x} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

Find the probability that the time between the arrival of successive trucks will be

(a) less than 5 minutes.

(b) at least 45 minutes.

Additional Notes

Additional Notes

Exercises

1. The probability density function for X is

$$f(x) = \begin{cases} x & 0 < x \leq 1 \\ \frac{1}{4} & 1 < x \leq 3 \\ 0 & \text{otherwise} \end{cases}$$

Find:

(a) $P(X = 1.5)$.

(b) $P(0.5 < X < 1.5)$.

(c) $P(0.5 \leq X \leq 1.5)$.

(d) $P(X > 1.5)$.

(e) $P(X < 0.5)$.

(f) the mean and standard deviation of X .

2. Consider a continuous random variable, X, with probability density function

$$f(x) = \begin{cases} kx^4 & 0 \leq x \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

(a) Find the value of k that makes $f(x)$ a valid probability density function.

(b) Find the mean of X.

(c) Find the standard deviation of X.

3. Let X represent how often a student studies alone, as a proportion of their total study time. (For example, $X = 0.35$ indicates that a student spends 35% of their total study time alone.) The probability density function of X is:

$$f(x) = \begin{cases} \frac{1}{(\ln 2)(x+1)} & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Find the probability that a student studies alone:

(a) exactly 20% of the time.

(b) less than 20% of the time.

(c) at least 30% of the time.

(d) between 50% and 75% of the time.

4. The time it takes students to complete a certain project is a uniform continuous random variable with values between 2 and 11 hours. Find:

   (a) the probability density function for the completion time.

   (b) the probability that a student takes between 3 and 8 hours to complete their project.

   (c) the probability that a student takes more than 7 hours to complete their project.

   (d) the probability that a student takes less than 4 hours to complete their project.

5. The lifetime of a certain machine part (in years) has probability density function

$$f(x) = \begin{cases} 2e^{-2x}, & x > 0 \\ 0 & x \leq 0 \end{cases}$$

   (a) Find the probability that a part lasts less than 0.1 years.

   (b) Using part (a), find the probability that a part lasts at least 0.1 years.

6. The shelf life of a brand of tomato soup (in months) has probability density function

$$f(x) = \begin{cases} 0.1e^{-0.1x} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

   (a) Find the probability that a can of soup has a shelf life between two and five months.

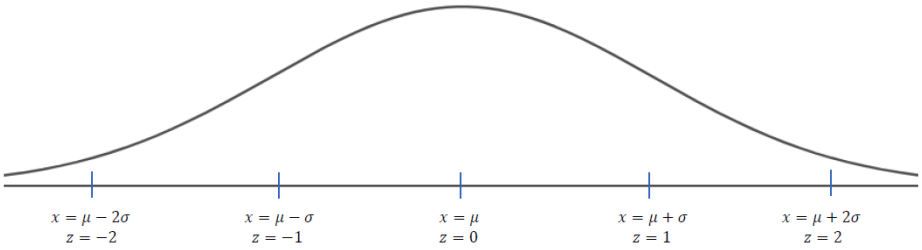
   (b) Find the average shelf life given that $\int_0^\infty xe^{-kx}\,dx = \frac{1}{k^2}$ for $k > 0$.

7 The Normal Distribution

The most important and most widely used distribution in statistics is the normal distribution.

Definition: A continuous random variable, X, with mean μ and SD σ has a normal distribution if its p.d.f. is

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



By use of a change of variable called the z-score,

$$z = \frac{x - \mu}{\sigma},$$

we can standardize any normal distribution. The standard normal distribution has the p.d.f.

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

The graph of a normal distribution is often referred to as a *bell curve*. Since the p.d.f. is very difficult to integrate, we use the *Standard Normal Distribution Table* to find the area under the bell curve corresponding to

$$\int_a^b f(x)\,dx = P(a \le X \le b) = P\left(\frac{a-\mu}{\sigma} \le z \le \frac{b-\mu}{\sigma}\right).$$

Example 7.1. Find the proportion of x-values that are within one standard deviation of the mean for a normal distribution; that is, find $P(\mu - \sigma \le X \le \mu + \sigma)$.

Example 7.2. The volume in bottles of gingerale is normally distributed with a mean of 2.01 L and a SD of 0.13 L. Find the probability that a bottle has a volume

(a) between 1.77 and 2.29 L.

(b) more than 1.91 L.

Example 7.3. [1, p. 45] Mopeds are popular in Europe because of their mobility, ease of operation, and low cost. A 2008 article described a rolling bench test for determining maximum vehicle speed. A normal distribution with $\mu = 46.8$ km/h and $\sigma = 1.75$ km/h is proposed.

(a) What proportion of mopeds have a maximum speed that is at most 50 km/h?

(b) What proportion of mopeds have a maximum speed that is at least 48 km/h?

(c) What speed separates the fastest 75% of all mopeds from the others?

Example 7.4. [1, p. 45] Suppose the flow of current (in milliamps) in wire strips of a certain type under specified conditions can be modeled with a normal distribution having $\mu = 20$ and $\sigma = 1$. How large must a current flow be to be among the largest 5% of all flows?

Example 7.5. [1, p. 41] The time that it takes a driver to react to the brake on a decelerating vehicle is critical in avoiding rear-end collisions. A 1993 article from *Ergonomics* suggests that reaction time for an in-traffic response to a brake signal from standard brake lights can be modeled with a normal distribution having parameters $\mu = 1.25$ sec and $\sigma = 0.46$ sec. In the long run, what proportion of reaction times will

(a) be between 1.00 sec and 1.75 sec?

(b) exceed 2 sec?

Example 7.6. The time it takes to inspect a certain type of production component is normally distributed with a mean of 6.8 s. Find the SD of the inspection times if 26.62% of inspection times are between 6.2 and 7.4 s.

Additional Notes

Additional Notes

Exercises

1. Let z be the standard normal random variable. Find:

   (a) $P(0 \leq z \leq 1.20)$

   (b) $P(-2.81 \leq z \leq 0)$

   (c) $P(-1.98 \leq z \leq 3.41)$

   (d) $P(1.35 \leq z \leq 1.85)$

   (e) $P(-2.93 \leq z \leq -1.90)$

   (f) $P(z \geq 2.46)$

   (g) $P(z \leq -1.34)$

2. Find the proportion of x-values that are within three standard deviations of the mean for a normal distribution.

3. [1, p. 45] Spray drift is a constant concern for pesticide applicators and agricultural producers. The inverse relationship between droplet size and drift potential is well known. The normal distribution with $\mu = 1050$ μm and $\sigma = 150$ μm can be used as a model for droplet size for water sprayed through a 760 ml/min nozzle.

   (a) What proportion of all droplets have a size that is less than 1500 μm?

   (b) What proportion of all droplets have a size that is between 1000 and 1500 μm?

   (c) How would you characterize the smallest 2% of all droplets?

4. [1, p. 60] The bursting strength of wine bottles of a certain type is normally distributed with parameters $\mu = 250$ psi and $\sigma = 30$ psi. What proportion of these bottles have a bursting strength greater than 300 psi?

5. [1, p. 46] Based on extensive data from an urban freeway near Toronto, it is assumed that vehicle speeds can best be represented by a normal distribution. The values of $\mu = 119$ km/h and $\sigma = 13.1$ km/h were reported.

   (a) What percentage of vehicles have speeds that are between 100 and 120 km/h?

   (b) What speed characterizes the fastest 10% of all speeds?

   (c) The posted speed limit was 100 km/h. What percentage of vehicles were traveling at speeds exceeding this posted limit?

   (d) Find the value a such that 90% of all vehicle speeds are between $\mu - a$ and $\mu + a$ km/h.

6. [2, p. 158] The actual amount of instant coffee that a filling maching puts into "4-ounce" jars can be modeled as a random variable having a normal distribution with $\sigma = 0.04$ ounces. If only 2% of the jars are to contain less than 4 ounces, what should be the mean fill of these jars?

8 Central Limit Theorem

Demonstration: The Penny Experiment

We calculated sample means in *The Penny Experiment*, and we saw that the shape of the "histogram" changed dramatically when we went from the population of individual penny years to the population of average penny years in samples of size 5.

For large sample sizes, the sample means have a very convenient property:

> **The Central Limit Theorem**
> For a fixed sample size n, the set of all possible $\bar{x}$-values can be approximated by a normal distribution when n is sufficiently large, regardless of the shape of the population distribution.

Rule of thumb: the Central Limit Theorem can be used if the sample size is $n \geq 30$.

The mean of all $\bar{x}$-values for a fixed sample size is the same as the mean of the population that the samples were taken from. However, the standard deviation of the $\bar{x}$-values is smaller than the standard deviation of the original population.

> **Definition:** For a fixed sample size n, the standard deviation of the population of all $\bar{x}$-values is called the standard error of the mean, and is given by
>
> $$SE = \frac{\sigma}{\sqrt{n}},$$
>
> where σ is the SD of the population that the samples were taken from.

As a result, for samples of size $n \geq 30$, the z-score corresponding to a given $\bar{x}$-value is

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}},$$

and we use the *Standard Normal Distribution Table* to find probabilities involving sample means.

Example 8.1. A large class has a test average of 72 with a SD of 8. Take a random sample of n tests. Find the probability that the average of the n tests is more than 75 if:

(a) $n = 30$.

(b) $n = 80$.

Example 8.2. Suppose that checked baggage has a mean mass of 21 kg with a SD of 4 kg.

(a) If 40 bags are randomly selected, find the probability that their average mass is between 20 and 23 kg.

(b) Find the probability that the total mass of 50 randomly selected bags is greater than 1130 kg.

Example 8.3. [1, p. 241] Suppose that the sediment density (in g/cm^3) of specimens from a certain region is normally distributed with a mean of 2.65 and a standard deviation of 0.85.

(a) If a single specimen is randomly selected, what is the probability that its sediment density is at most 2.75?

(b) If a random sample of 35 such specimens is selected, what is the probability that the sample average sediment density is at most 2.75?

(c) How large a sample would be required to ensure that $P(\bar{x} \leq 2.75) \geq 0.99$?

(d) For a sample size of $n = 35$, find c such that $P(\bar{x} \geq c) = 0.95$?

Additional Notes

Additional Notes

Exercises

1. At a walk-in clinic the average amount of time patients spend with a doctor is 7.1 minutes. The standard deviation is 5.2 minutes. Find the probability that for a random sample of 60 patients, the mean visit time is between 6 and 8 minutes.

2. The test scores in a large Calculus class have a mean of 71 and a standard deviation of 9. Find the probability that a random sample of n tests have an average score of less than 70 if:

   (a) $n = 40$.

   (b) $n = 250$.

3. [2, p. 213] If a 1-gallon can of paint covers an average area of 513.3 square feet with a standard deviation of 31.5 square feet, what is the probability that the sample mean area covered by a sample of 40 of these cans of paint will be

   (a) less than 510 sq ft?

   (b) more than 520 sq ft?

   (c) between 510 and 520 sq ft?

4. At a large engineering firm, employees worked a mean of 45.5 hours last week, with a standard deviation of 6 hours. One hundred and sixty employees are selected at random. What is the probability that their work hours totalled more than 7344 hours?

5. A machine is filling cans of pop. The volume per can has a standard deviation of 1.9 mL. What should the volume be set to on the machine (this is μ) in order to ensure that in a random sample of 30 cans, there is a 99% probability that the mean is at least 355.0 mL?

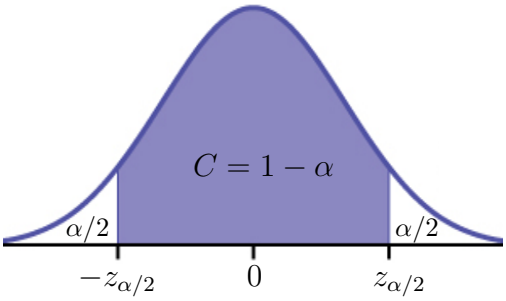
6. [2, p. 216] If the distribution of the weights of all passengers traveling by air on a certain route has a mean of 145 pounds with a standard deviation of 18 pounds, what is the probability that the combined weight of 40 passengers traveling on that route exceeds 6000 pounds?

9 Confidence Intervals

In this section, we want to use the size, mean and SD of a *sample* to find an interval estimate for the mean of the *population* that the sample came from. All interval estimates are calculated by first selecting a *confidence level*, which measures the degree of certainty that the interval estimate produced using a normal distribution will contain the true population mean. The most common values for the confidence level are 90%, 95%, 98% and 99%. A confidence level of 95% means that, of all possible samples of size n taken from a population, 95% of them will give an interval estimate that contains the true population mean and 5% will not. It does not mean that the probability that the population mean is in a certain interval is 95%.

9.1 Large Samples

For random samples of size $n \geq 30$ taken from any population, the Central Limit Theorem tells us that $\bar{x}$-values are normally distributed. As a result, given a confidence level, we can use a reverse look-up on the *Standard Normal Distribution Table* to find an interval estimate for the population mean.



Since confidence levels are usually given as 90%, 95%, 98% or 99%, it is quicker to use this reverse look-up table:

$1-\alpha$	0.9	0.95	0.98	0.99
$z_{\alpha/2}$	1.645	1.960	2.326	2.576
z_{α}	1.282	1.645	2.054	2.326

Definition: Given a confidence level $C = 1 - \alpha$ and a random sample of size $n \geq 30$ from a population with standard deviation σ, a confidence interval (CI) for the population mean, μ, is given by

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}},$$

where σ may be approximated using the sample standard deviation s.
The quantity $z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$ is called the margin of error (ME).

Example 9.1. A random sample of 60 cans of Coke had an average volume of 355.3 mL and a standard deviation of 2.5 mL.

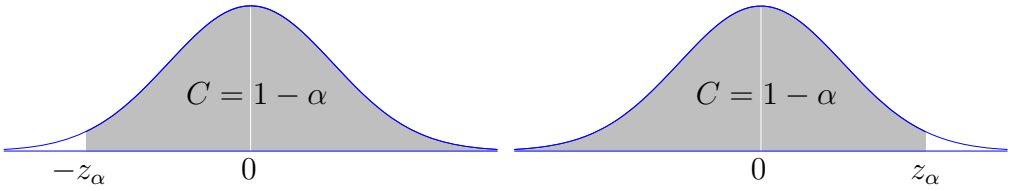
(a) Find a 95% confidence interval for the average volume among all cans of Coke.

(b) Would a 99% confidence interval be wider or narrower than the 95% confidence interval in part (a)?

Example 9.2. [2, p. 233] 80 readings of daily emission (in tons) of sulfur oxides from an industrial plant had an average of 18.85 tons and a standard deviation of 5.55 tons. Use this data to construct a 99% confidence interval for the plant's true average daily emission of sulfur oxides.

Example 9.3. [3, p. 638] The thickness of a certain type of sheet metal has a known standard deviation of 0.27 mm. We want to estimate μ with a 95% margin of error of less than 0.01 mm. What is the minimum sample size n required?

Sometimes we are interested only in the lower or the upper bound on an estimate of the population mean, rather than an interval.



Definition: Given a confidence level $C = 1-\alpha$ and a random sample of size $n \geq 30$ from a population with standard deviation σ, a lower confidence bound (LCB) for the population mean, μ, is given as

$$\mu > \bar{x} - z_\alpha \frac{\sigma}{\sqrt{n}}$$

and an upper confidence bound (UCB) for the population mean, μ, is given as

$$\mu < \bar{x} + z_\alpha \frac{\sigma}{\sqrt{n}}$$

where σ may be approximated using the sample standard deviation s.

As with confidence intervals, we use the handy table given at the beginning of this section to find the z_α-values for the common confidence levels rather than doing a reverse look-up on the *Standard Normal Distribution Table*.

Example 9.4. 30 randomly selected water samples have a mean pollution concentration of 48.1 ppm with a standard deviation of 6.2 ppm. Find a 99% UCB for the mean pollution concentration in the body of water.

Example 9.5. In a large class, test marks have a SD of 10.3. A random sample of 40 tests has an average mark of 69.1. Find a 98% LCB for the class average.

9.2 Small Samples

If the sample size is $n < 30$, we need to know that the population has a normal distribution to be able to calculate confidence intervals. Moreover, the sample size affects the shape of the bell curve for small values of n, so a new parameter called the *degrees of freedom*,

$$df = n - 1,$$

is needed as well. As a result, the probabilities are not associated with the standard normal distribution, and we use the *t-Distribution Table* instead.

Definition: Given a confidence level $C = 1 - \alpha$ and a random sample with size $n < 30$ and standard deviation s from a normally distributed population, a confidence interval (CI) for the population mean, μ, is given by

$$\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}.$$

The quantity $t_{\alpha/2}\frac{s}{\sqrt{n}}$ is called the margin of error (ME).
A lower confidence bound (LCB) for the population mean, μ, is given as

$$\mu > \bar{x} - t_{\alpha}\frac{s}{\sqrt{n}}$$

and an upper confidence bound (UCB) for the population mean, μ, is given as

$$\mu < \bar{x} + t_{\alpha}\frac{s}{\sqrt{n}}.$$

Demonstration: Confidence Intervals

Example 9.6. [2, p. 237] Ten bearings made by a certain process have a mean diameter of 0.5060 cm and a standard deviation of 0.0040 cm. Assuming that the data is a random sample from a normal population, construct a 95% confidence interval for the actual average diameter of bearings made by this process.

Example 9.7. [3, p. 640] A sample of 15 washing machines of a certain brand had a mean replacement time of 9.1 years, with a standard deviation of 2.7 years. Find a 90% lower confidence bound for the mean replacement time of all washing machines of this brand.

Example 9.8. [3, p. 640] A test station measured the loudness of a random sample of 22 jets taking off from a certain airport. The mean was found to be 107.2 dB, with a standard deviation of 9.2 dB. Find a 95% upper confidence bound for the mean loudness of all jets taking off from this airport.

Additional Notes

Additional Notes

Exercises

1. A random sample of 40 test marks from a large Calculus class had an average of 78.0 and a standard deviation of 5.0. Find a 95% confidence interval for the average test mark.

2. For a random sample of 50 room temperature readings taken in a lab, the mean temperature is 28.00 °C and the standard deviation is 3.32 °C. Find an upper confidence bound for the true mean temperature in the lab with:

   (a) 95% confidence.

   (b) 98% confidence.

3. At a paper factory, the paper length is known to have a standard deviation of 0.08 inches. In a random sample of 100 sheets, the mean length is found to be 11.00 inches. Find a lower confidence bound for the mean length among all sheets of paper produced at the factory with:

   (a) 90% confidence.

   (b) 99% confidence.

4. [3, p. 640] A toy manufacturer wishes to estimate the average time it takes an adult to assemble a certain "easy to assemble" toy. How many adults must be sampled so that a 99% confidence interval for the true mean assembly time μ will have a maximum margin of error of 2.0 minutes? Use the known standard deviation of 5.9 minutes for a similar model.

5. At ABC Cereal Company, the mass of cereal boxes has a standard deviation of 13 grams. We want a 95% confidence interval for the mean mass among all their cereal boxes with a margin of error less than 2 grams. Find the minimum sample size.

6. Consider a large-sample confidence interval for the population mean. Describe the effect of the following on the margin of error, assuming that the other quantities remain unchanged:

   (a) the sample size increases.

   (b) the standard deviation increases.

   (c) the confidence level increases.

   (d) the sample mean increases.

7. Ten random water samples taken from the inner harbour yield a mean nitrate ion concentration of 25.0 ppm with a standard deviation of 5.1 ppm. Assuming that the ion concentrations are normally distributed throughout the inner harbour, find a 95% confidence interval for the mean ion concentration in the inner harbour.

8. Fifteen randomly selected ropes had a mean breaking strength of 69.1 pounds, with a standard deviation of 3.5 pounds. The breaking strengths of this brand of rope are known to be normally distributed. Find a 99% confidence interval for the mean breaking strength of this brand of rope.

9. Fuel efficiencies for city driving are measured for a random sample of twelve 2012 Prius cars. The mean fuel efficiency was 51 miles per gallon, with a standard deviation of 2 miles per gallon. The fuel efficiencies of 2012 Prius cars are normally distributed. Find an upper confidence bound for the mean fuel efficiency among all 2012 Prius cars with:

   (a) 90% confidence.

   (b) 97.5% confidence.

10. At a bottling plant, volumes are measured for a random sample of 20 cans of pop. The mean volume was 356.1 mL, with a standard deviation of 1.9 mL. The volumes among all cans at the bottling plant are normally distributed. Find a lower confidence bound for the mean volume among all cans at the bottling plant with:

    (a) 95% confidence.

    (b) 99% confidence.

11. [2, p. 231] Measurement errors in scientific experiments are normally distributed. In fact, "the normal curve of errors" was first studied in the eighteenth century when scientists observed an astonishing degree of regularity in errors of measurement.

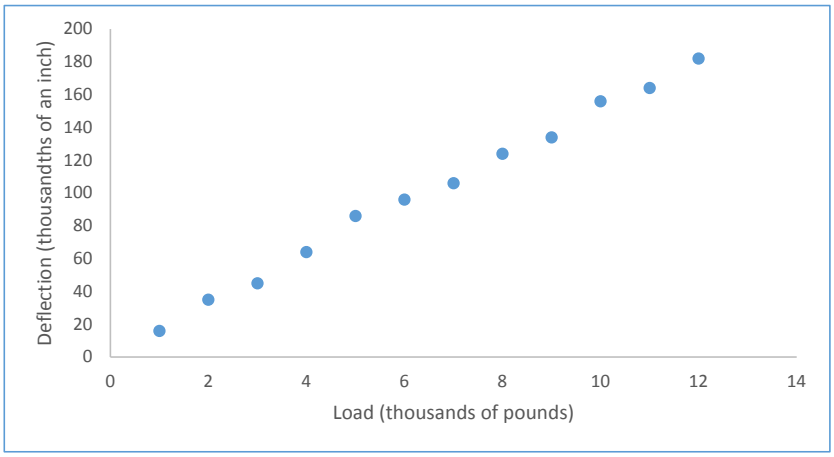
    In six determinations of the melting point of tin, a chemist obtained a mean of 232.26 °C with a standard deviation of 0.14 °C. If he uses this mean to estimate the actual melting point of tin, what is the 98% margin of error?

10 Linear Regression

Given a *bivariate* data set (i.e. a set of points in two variables), it is often desirable to express a concise relationship between the variables. The best way to start is to graph the data set using a *scatterplot* to visually assess the pattern of the points.

Consider, for example, the following table listing the deflection of a tensile ring at various loads [2, p. 338]. The x's are the load forces in thousands of pounds and the y values are the corresponding deflections in thousandths of an inch:

x	y
1	16
2	35
3	45
4	64
5	86
6	96
7	106
8	124
9	134
10	156
11	164
12	182



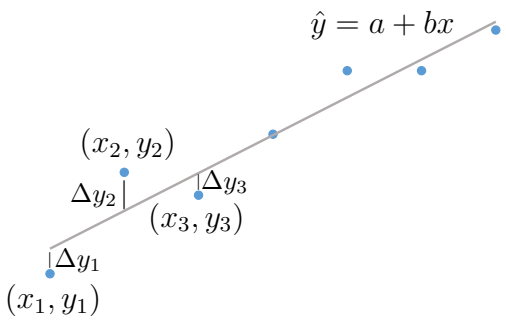
From the accompanying scatterplot, we can tell that this data set has a strong linear association. Bivariate data sets can also have non-linear correlation, but in this course we will only look at the linear case.

10.1 The Least Squares Regression Line

The *least squares regression line* is also commonly referred to as the *best-fit line*. It is denoted $\hat{y}$ and is the unique line that minimizes

$$\sum (\Delta y_i)^2,$$

where $\Delta y_i = y_i - \hat{y}_i$ is called the *residual* of the point (x_i, y_i).



If a scatterplot for a bivariate data set with n points reveals a linear association, we can find the equation of the least squares regression line as follows:

Step 1: Calculate the means $\bar{x}$ and $\bar{y}$

Step 2: Calculate

$$S_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n} \qquad \text{and} \qquad S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$$

To find these two values by hand, it is useful to make a table:

x	y	xy	x^2
$\sum x$	$\sum y$	$\sum xy$	$\sum x^2$

Step 3: Calculate

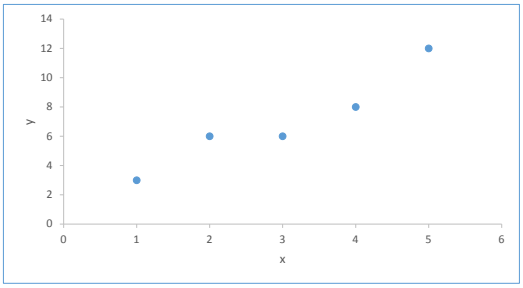
$$b = \frac{S_{xy}}{S_{xx}} \qquad \text{and} \qquad a = \overline{y} - b\overline{x}$$

Step 4: The equation is $\hat{y} = a + bx$

The SHARP EL-531X calculator has functions to calculate a and b for $\hat{y}$. Instructions are available in the Appendix.

Example 10.1. Consider the following data set:

x	y	xy	x^2
1	3		
2	6		
3	6		
4	8		
5	12		



(a) Find the equation of the least squares regression line.

(b) Graph the least squares regression line for $1 \leq x \leq 5$ on the given scatterplot.

(c) Find the residual for the point $(2, 6)$.

If a least squares regression line is to be used to predict values that are not in the given data set, it is important to do so only for values in the given ranges.

Example 10.2. For the data set in the previous example, use $\hat{y}$ to predict:

(a) the y-value for $x = 2.4$.

(b) the y-value for $x = 6$.

(c) the x-value for $y = 10$.

10.2 The Coefficients of Correlation and Determination

There are two useful coefficients that describe how well a linear regression fits a data set, but they should only be used in conjunction with scatterplots.

Definition The coefficient of correlation is a measure of the strength of the linear relationship of two variables, and is defined by

$$r = b\frac{s_x}{s_y},$$

where s_x and s_y are the standard deviations of the x-values and y-values, respectively, and b is the slope of $\hat{y}$. The values of r lie in the the range $-1 \leq r \leq 1$. If r is near 1, the variables are positively correlated, and if r is near -1, the variables are negatively correlated. For the values of r between -0.5 and 0.5, the linear correlation is poor.

The coefficient of determination is r^2 and indicates what percentage of the variation in y is accounted for by the best-fit line:

$$r^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}.$$

The SHARP EL-531X calculator has a function to calculate r. Instructions are available in the Appendix.

An alternative way to find r is to use

$$r^2 = \frac{(S_{xy})^2}{S_{xx}S_{yy}},$$

where S_{xy} and S_{xx} were defined along with the instructions for finding the equation for $\hat{y}$, and

$$S_{yy} = \sum y^2 - \frac{(\sum y)^2}{n}.$$

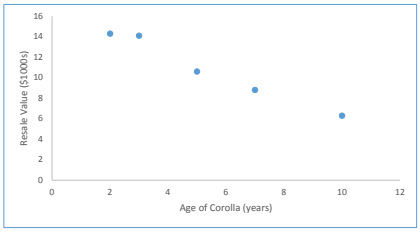
Demonstration: Anscombe's Quartet

Example 10.3. The following bivariate data set has $\hat{y} = 16.54 - 1.06x$ and a coefficient of determination of 0.9746:

x = age of Corolla (years)

y = resale value ($1000s)

x	y
2	14.3
3	14.1
5	10.6
7	8.8
10	6.3



(a) Is the linear association positive or negative?

(b) Find the correlation coefficient.

(c) What % of the variation in y is accounted for by the best-fit line?

(d) What resale value is predicted for a 4-year-old Corolla?

(e) Why should we not predict the resale value for a 1-year-old Corolla?

(f) What age corresponds to a resale value of $4,500?

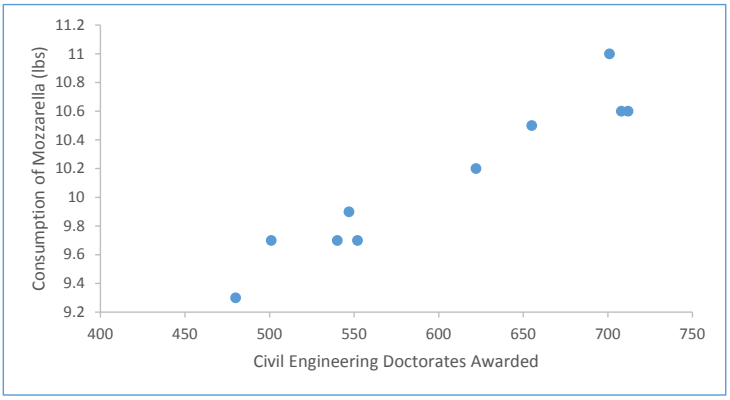
Correlation does NOT imply causation!

The following data set was taken from
`http://www.tylervigen.com/spurious-correlations`

$x =$ the number of civil engineering doctorates awarded in the US, and

$y =$ the per capita consumption of mozzeralla cheese in the US (in lbs):

year	x	y
2000	480	9.3
2001	501	9.7
2002	540	9.7
2003	552	9.7
2004	547	9.9
2005	622	10.2
2006	655	10.5
2007	701	11
2008	712	10.6
2009	708	10.6



Surprisingly, $r = 0.9586$. The best-fit line is $\hat{y} = 6.600 + 0.006x$, but this does NOT mean that an increase of 1 doctorate **causes** the mozzarella consumption to increase by 0.006 lbs per person (nor can we assume that mozzarella consumption causes an increase in Civil Engineering doctorates). The accurate interpretation of the slope is: *as the number of doctorates increases by 1, the mozzarella consumption increases on average by 0.006 lbs per person.*

Additional Notes

Additional Notes

Exercises

1. For each data set below, use a scatterplot to decide whether the linear association is best described as positive, negative, zero, or nonlinear.

(a)

x	10	15	25	30	40	50
y	78	71	45	41	20	3

(b)

x	1	5	7	11	13	14
y	33	18	2	5	19	24

(c)

x	1.1	1.4	1.6	1.6	1.8	1.9
y	1.3	1.7	2.1	2.3	2.6	2.9

(d)

x	5	10	15	20	25	30	35
y	7	9	6	8	7	9	6

2. Consider a bivariate data set. What percentage of the variation in y is accounted for by the best-fit line if the correlation coefficient is:

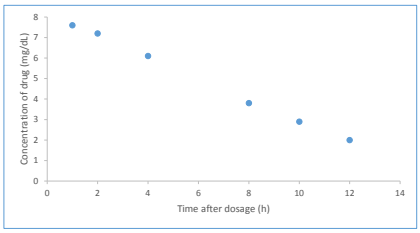
   (a) 0.8?

   (b) -0.9?

3. Bivariate Data Set A has a correlation coefficient of 0.84. Bivariate Data Set B has a correlation coefficient of -0.96. Which data set has a stronger linear association? Explain.

4. [3, p. 650] In a research project to determine the amount of a drug that remains in the bloodstream after a given dosage, the concentrations y (in mg of drug / dL of blood) were recorded after t hours, as shown:

t	y
1.0	7.6
2.0	7.2
4.0	6.1
8.0	3.8
10.0	2.9
12.0	2.0



   (a) Find the equation of the best-fit line.

   (b) Find the residual for the point $(4.0, 6.1)$.

   (c) Find the correlation coefficient and the coefficient of determination.

   (d) What percentage of the variation in y is accounted for by the best-fit line?

   (e) What drug concentration does the best-fit line predict at 6 hours?

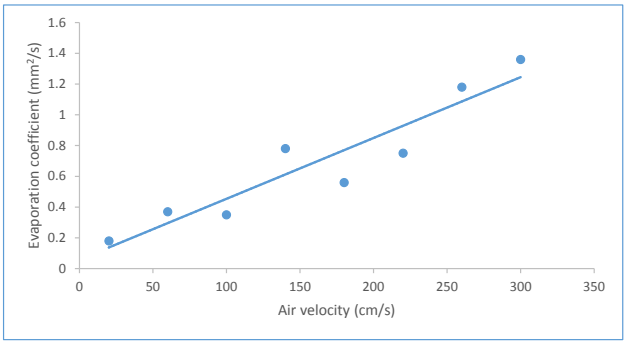
   (f) Why should we not use the data to predict the drug concentration 24 hours after the dosage is given?

(g) According to the best-fit line, what value of t corresponds to $y = 5.0$ mg/dL?

5. [2, p. 341] The following are measurements of the air velocity and evaporation coefficient of burning fuel droplets in an impulse engine:

$$x = \text{air velocity (cm/s)}$$
$$y = \text{evaporation coefficient (mm}^2\text{/s)}$$

x	y
20	0.18
60	0.37
100	0.35
140	0.78
180	0.56
220	0.75
260	1.18
300	1.36



The equation of the best-fit line is $\hat{y} = 0.058 + 0.004t$ and the coefficient of determination is 0.8793.

(a) Is the linear association positive or negative?

(b) Find the correlation coefficient.

(c) What percentage of the variation in y is accounted for by the best-fit line?

(d) Interpret the slope of $\hat{y}$.

Appendix A Histograms in Excel 2013

To create a histogram for a data set like

$$90, 93, 94, 95, 98, 100, 101, 102, 103, 105, 108, 108, 109, 110, 110, 114, 115, 115, 119, 120$$

Step 1 Set up Excel for Data Analysis

FILE → Options → Add-Ins

At the bottom of the window, under *Manage:* select Excel Add-ins and click Go...

A new window will open. Select Analysis ToolPak and click OK

Step 2 Enter your data points in column A

Step 3 Enter your class endpoints in column B

Excel calls classes "bins" and uses the right endpoint of each class as the class mark. Your classes will be $(-\infty, B1]$, $(B1, B2]$, $(B2, B3]$, ...

For the given example, a good choice of classes is $(-\infty, 90]$, $(90, 95]$, $(95, 100]$, $\cdots$, $(115, 120]$ so column B should consist of the list $90, 95, 100, 105, 110, 115, 120$

Step 4 Create the histogram

DATA → Data Analysis → Histogram → OK

A new window will open.

- For *Input Range:* enter the cell range of your data set. For this example it is A1:A20
- For *Bin Range:* enter the cell range of your class endpoints. For this example it is B1:B7
- Select New Workbook
- Select Chart Output
- Click OK

A new Excel workbook will open, displaying the frequency distribution table and the histogram. Click on the histogram to activate a menu for editing options.

Note The *More* class includes any data points greater than your largest class endpoint.

Appendix B Statistics on the SHARP EL-531X

To find the **mean** and **standard deviation** of a list of numbers like

$$5, 2, 6, 4, 7$$

Step 1 Set up the calculator

$\boxed{\text{MODE}} \rightarrow \boxed{1}$ *to select STAT* $\rightarrow \boxed{0}$ *to select SD*

Step 2 Enter your data points

For the given example, this looks like:

$\boxed{5} \rightarrow \boxed{\text{M+}} \rightarrow \boxed{2} \rightarrow \boxed{\text{M+}} \rightarrow \boxed{6} \rightarrow \boxed{\text{M+}} \rightarrow \boxed{4} \rightarrow \boxed{\text{M+}} \rightarrow \boxed{7} \rightarrow \boxed{\text{M+}}$

At the end of this sequence, the screen should read *DATA SET = 5* to indicate that you've entered 5 data points.

Step 3 Calculate statistics on your data set

You do not need to re-enter the data set between calculations.

- For population mean, μ, or sample mean, $\bar{x}$:
  $\boxed{\text{ALPHA}} \rightarrow \boxed{4}$ *to select* $\bar{x}$
  For the given example, your answer should be 4.8
- For sample standard deviation, s:
  $\boxed{\text{ALPHA}} \rightarrow \boxed{5}$ *to select* sx
  For the given example, your answer should be $1.9234\cdots$
- For population standard deviation, σ:
  $\boxed{\text{ALPHA}} \rightarrow \boxed{6}$ *to select* σx
  For the given example, your answer should be $1.7204\cdots$

Step 4 Clear the data set

$\boxed{\text{2nd F}} \rightarrow \boxed{\text{ALPHA}} \rightarrow \boxed{0}$ *to select MEM* $\rightarrow \boxed{0}$ *to select CLR_MEMORY?*

To find the **mean** and **standard deviation** of a probability distribution like

x	$P(x)$
0	0.1
1	0.5
2	0.4

Step 1 Set up the calculator

$\boxed{\text{MODE}} \rightarrow \boxed{1}$ *to select STAT* $\rightarrow \boxed{0}$ *to select SD*

Step 2 Enter your data points

For the given example, this looks like:

$\boxed{0} \rightarrow \boxed{\text{STO}}$ to select $(x, y) \rightarrow \boxed{\cdot} \rightarrow \boxed{1} \rightarrow \boxed{\text{M+}} \rightarrow$

$\boxed{1} \rightarrow \boxed{\text{STO}}$ to select $(x, y) \rightarrow \boxed{\cdot} \rightarrow \boxed{5} \rightarrow \boxed{\text{M+}} \rightarrow$

$\boxed{2} \rightarrow \boxed{\text{STO}}$ to select $(x, y) \rightarrow \boxed{\cdot} \rightarrow \boxed{4} \rightarrow \boxed{\text{M+}}$

At the end of this sequence, the screen should read *DATA SET = 3* to indicate that you've entered 3 data points.

Step 3 Calculate statistics on your probability distribution

You do not need to re-enter the data set between calculations.

- For mean, μ, aka expected value:
  $\boxed{\text{ALPHA}} \rightarrow \boxed{4}$ *to select* $\bar{x}$
  For the given example, your answer should be 1.3
- For standard deviation, σ:
  $\boxed{\text{ALPHA}} \rightarrow \boxed{6}$ *to select* σx
  For the given example, your answer should be $0.64031\cdots$

Step 4 Clear the data set

$\boxed{\text{2nd F}} \rightarrow \boxed{\text{ALPHA}} \rightarrow \boxed{0}$ *to select MEM* $\rightarrow \boxed{0}$ *to select CLR_MEMORY?*

Note: This process can also be used to find the mean and SD of a data set given as a frequency distribution table.

To find the **linear regression line** and **correlation coefficient** of a bivariate data set like

x	y
1	3
2	6
3	6
4	8
5	12

Step 1 Set up the calculator

[MODE] → [1] *to select STAT* → [1] *to select LINE*

Step 2 Enter your data points

For the given example, this looks like:

[1] → [STO] *to select (x,y)* → [3] → [M+] →

[2] → [STO] *to select (x,y)* → [6] → [M+] →

[3] → [STO] *to select (x,y)* → [6] → [M+] →

[4] → [STO] *to select (x,y)* → [8] → [M+] →

[5] → [STO] *to select (x,y)* → [1] → [2] → [M+]

At the end of this sequence, the screen should read *DATA SET = 5* to indicate that you've entered 5 data points.

Step 3 Calculate statistics on your data set

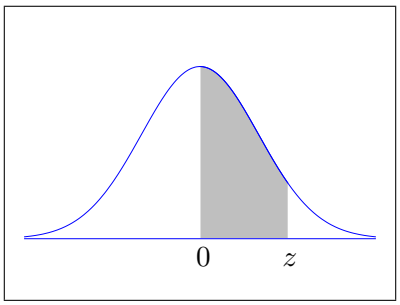
You do not need to re-enter the data set between calculations.

- For the linear regression equation, $\hat{y} = a + bx$:
  [ALPHA] → [(] *to select a* → [=]
  For the given example, your answer should be 1
  [ALPHA] → [)] *to select b* → [=]
  For the given example, your answer should be 2
  Together, the answer for this example is $\hat{y} = 1 + 2x$.
- For the correlation coefficient, r:
  [ALPHA] → [÷] *to select r* → [=]
  For the given example, your answer should be $0.9534\cdots$
- You can also calculate quantities like $\bar{x}$, $\bar{y}$, $\sum x$, $\sum x^2$, $\sum xy, \cdots$

Step 4 Clear the data set

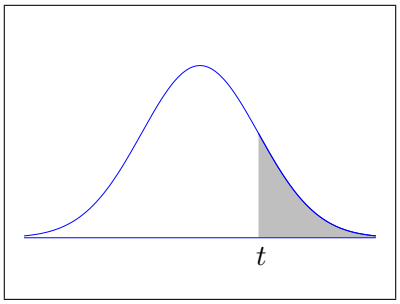
[2nd F] → [ALPHA] → [0] *to select MEM* → [0] *to select CLR_MEMORY?*

Standard Normal Distribution Table



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.4990	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995
3.3	.4995	.4995	.4995	.4996	.4996	.4996	.4996	.4996	.4996	.4997
3.4	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4998
3.5	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998

t-Distribution Table



The shaded area is equal to α for $t = t_\alpha$.

df	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
32	1.309	1.694	2.037	2.449	2.738
34	1.307	1.691	2.032	2.441	2.728
36	1.306	1.688	2.028	2.434	2.719
38	1.304	1.686	2.024	2.429	2.712
∞	1.282	1.645	1.960	2.326	2.576

Answers to Exercises

1 Collection and Representation of Data

1. (a) 5C3=10

   (b) {1,2,4}, {1,2,8}, {1,2,9}, {1,4,8}, {1,4,9},{1,8,9}, {2,4,8}, {2,4,9}, {2,8,9}, {4,8,9}

2. Select a simple random sample of 20 machines. The selected machines will be Group 1; the unselected machines will be Group 2.

3. 426 cans of regular pop and 174 cans of diet pop

4. (a) 1-in-50 systematic sample

   (b) simple random sample

   (c) stratified random sample

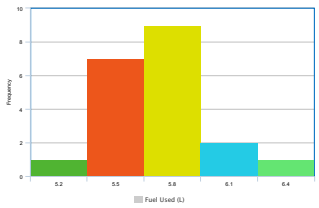
   (d) cluster sample

5. (a)

fuel used (L)	frequency
5.1-5.3	1
5.4-5.6	7
5.7-5.9	9
6.0-6.2	2
6.3-6.5	1

   (b)

fuel used (L)	relative frequency
5.1-5.3	$1/20 = 0.05$
5.4-5.6	$7/20 = 0.35$
5.7-5.9	$9/20 = 0.45$
6.0-6.2	$2/20 = 0.1$
6.3-6.5	$1/20 = 0.05$

   (c) 

2 Summarizing Data

1. $\mu = 72$; median = 74

2. $\overline{x} = 25.4$ g ; median = 26 g

3. 86

4. $\mu = 74.7$

5. $\overline{x} = 37.31$ °C ; median = 37.4 °C

6. $\overline{x} = \$910,000$; median = $\$800,000$

7. $\sigma^2 = 15.6$; $\sigma = 3.9$

8. $s^2 = 7$ g^2; $s = 2.6$ g

9. (a) Data sets are equally spread out.

   (b) Set 2 is more spread out.

10. (a) Mean and median both increase (by $\$5,000$); SD stays the same.

    (b) Mean, median and SD all increase (by 10%)

    (c) Mean increases. Median stays the same. SD decreases.

3 Probability

1. (a) $\frac{4}{8}$

   (b) $\frac{3}{8}$

2. $\frac{4}{10}$

3. (a) $\frac{3}{16}$

   (b) $\frac{5}{16}$

   (c) $\frac{13}{16}$

4. (a) $\frac{1}{6}$

   (b) $\frac{3}{6}$

   (c) $\frac{3}{6}$

5. $\frac{14}{30}$

6. (a) $\frac{96}{231}$

   (b) $\frac{176}{231}$

   (c) $\frac{55}{231}$

7. 0.949 or 94.9%

8. $\frac{18}{62}$

9. (a) 7.2 million

   (b) 400,000

(c) 6.48 million

(d) 1.08 million

10. (a) 0.505

(b) 0.032

4 Discrete Random Variables

1. (a) 0.52

(b) $\mu = E(X) = 11.007$

(c) $\sigma^2 = 3.972$

(d) $\sigma = 1.993$

(e) 0.93

2. (a) 0.6

(b) $\mu = E(Y) = 0.1$

(c) $\sigma^2 = 23.3$

(d) $\sigma = 4.83$

(e) 0.7

3. (a) \$7,500

(b) \$20,000

(c) 11,250,000 $\$^2$

(d) 100,000,000 $\$^2$

(e) Project B

4. (a)

x	$P(x)$
70,000	0.35
-10,000	0.65

(b) \$18,000

(c) \$38,000

(d) Project Alpha

5. (a)

x	$P(x)$
$m - 4000$	0.013
m	0.987

(b) \$112

5 Binomial and Poisson Distributions

1. (a) 0.33

(b) 0.78

(c) 0.99

2. (a) 0.18

(b) 0.51

(c) 0.06

3. (a) 0.82

(b) 0.91

4. (a) 0.05

(b) 0.06

5. 0.008

6 Continuous Random Variables

1. (a) 0

(b) 0.5

(c) 0.5

(d) 0.375

(e) 0.125

(f) $\mu = \frac{4}{3}$, $\sigma = 0.80$

2. (a) $k = \frac{5}{32}$

(b) $\frac{5}{3}$

(c) 0.28

3. (a) 0

(b) 0.26

(c) 0.62

(d) 0.22

4. (a) $f(x) = \begin{cases} \frac{1}{9} & 2 < x < 11 \\ 0 & \text{otherwise} \end{cases}$

(b) $\frac{5}{9}$

(c) $\frac{4}{9}$

(d) $\frac{2}{9}$

5. (a) 0.18

(b) 0.82

6. (a) 0.21

(b) 10 months

7 The Normal Distribution

1. (a) 0.3849

(b) 0.4975

(c) 0.9758
(d) 0.0563
(e) 0.0270
(f) 0.0069
(g) 0.0901

2. 0.9974
3. (a) 0.9987
   (b) 0.6280
   (c) The smallest 2% are under 742.5 μm
4. 0.0475
5. (a) 45.84%
   (b) Fastest 10% are above 135.8 km/h
   (c) 92.65%
   (d) $a = 21.484$
6. 4.082 ounces

8 Central Limit Theorem

1. 0.8594
2. (a) 0.2420
   (b) 0.0392
3. (a) 0.2546
   (b) 0.0885
   (c) 0.6569
4. 0.2005
5. 355.8 mL
6. 0.0392

9 Confidence Intervals

1. $76.5 < \mu < 79.5$
2. (a) $\mu < 28.77$ °C
   (b) $\mu < 28.96$ °C
3. (a) $\mu > 10.99$ inches
   (b) $\mu > 10.98$ inches
4. 58
5. 163
6. (a) margin of error decreases
   (b) margin of error increases
   (c) margin of error increases
   (d) margin of error doesn't change
7. $21.4 < \mu < 28.6$ ppm
8. $66.4 < \mu < 71.8$ pounds
9. (a) $\mu < 52$ miles per gallon
   (b) $\mu < 52$ miles per gallon
10. (a) $\mu > 355.4$ mL
    (b) $\mu > 355.0$ mL
11. 0.19 °C

10 Linear Regression

1. (a) negative
   (b) nonlinear
   (c) positive
   (d) zero
2. (a) 64%
   (b) 81%
3. Data Set B because $|r_B| > |r_A|$
4. (a) $\hat{y} = 8.16 - 0.52t$
   (b) 0.02
   (c) $r = -0.999$, $r^2 = 0.998$
   (d) 99.8%
   (e) 5.04 mg/dL
   (f) 24 is outside the data set range $1.0 \leq t \leq 12.0$
   (g) 6.1 h
5. (a) positive
   (b) 0.9377
   (c) 87.93%
   (d) As the air velocity increases by 1 cm/s, the evaporation coefficient increases on average by 0.004 mm^2/s

References

[1] Devore, J., Farnum, N., & Doi, J. (2014). *Applied Statistics For Engineers and Scientists* (3rd ed.). Stamford, CT: Cengage Learning.

[2] Johnson, R. (2005) *Miller and Freund's Probability and Statistics for Engineers* (7th ed.) Upper Saddle River, NJ: Pearson Prentice Hall.

[3] Washington, A. & Boue, M. (2010) *Basic Technical Mathematics with Calculus: SI Version* (10th ed.) Toronto: Pearson.