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] $\rightarrow$ [7] *to select RAND* $\rightarrow$

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

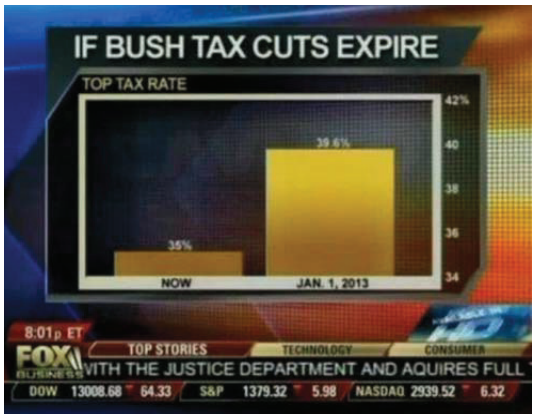
$$102, 115, 93, 105, 108, 110, 120, 94, 101, 103, 90, 110, 109, 101, 115, 119, 95, 108, 98, 114$$

(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