Statistics: Collection, Analysis and Interpretation of Data

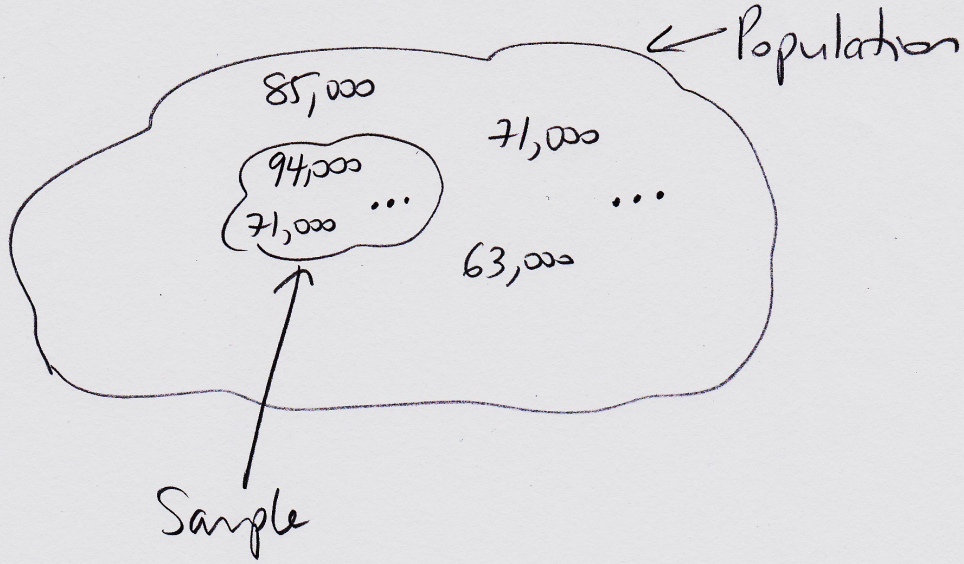
1. Collection and Representation of Data

Population: set of all measurements of interest

Sample: a subset of the population

Ex: Population: salaries of all engineers in Canada ($)

Sample: Salaries of 30 chosen engineers



Samples should be <u>representative</u> of the population.

<u>Ex</u>: Want to know if Canadians want lower taxes

a) Poll 10 people in 1 neighbourhood

not a representative sample

b) Poll 1000 people across the country, in urban and rural areas

more representative of all Canadians

<u>Ex</u>: Population = {A, B, C, D}
Find all samples of size 2

{A,B}, {A,C}, {A,D}, {B,C}, {B,D}, {C,D}

The number of different samples of size r that can be chosen from a population of size n is written nCr "n choose r"

On calculator: 4 2nd F nCr 2 = 6 ✓

Ex: How many samples of size 10 can be chosen from a population of 100 measurements?

$100C10 \simeq 1.7 \times 10^{13}$

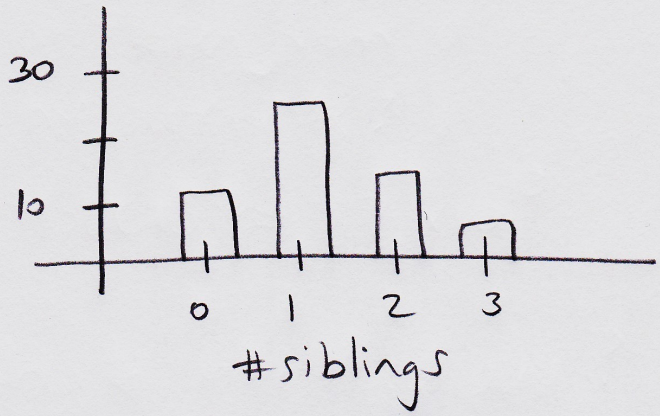
How to select samples?
Handout

Ex: Sample 50 people
How many siblings do they have?

#siblings	frequency
0	11
1	23
2	12
3	4

← # of times the measurement occurs

Frequency histogram:



#siblings	relative frequency
0	$\frac{11}{50} = 0.22$
1	$\frac{23}{50} = 0.46$
2	0.24
3	0.08

← % of measurements

Total # of measurements $n = 50$

Relative Frequency Histogram:

For large samples, data is grouped into <u>classes</u>.

<u>Ex</u>: Loudness of jet engines at takeoff (decibels)

102, 115, 93, 105, 108, 110, 120, 94, 101, 103,
92, 110, 109, 101, 115, 119, 95, 108, 98, 114

a) Create a frequency table with 6 classes

Min: 92

Max: 120

values in range = $120 - 92 \underline{\underline{+1}} = 29$

Bump up to 30 by adding "91" to data set

values/class = $\frac{30}{6} = 5$

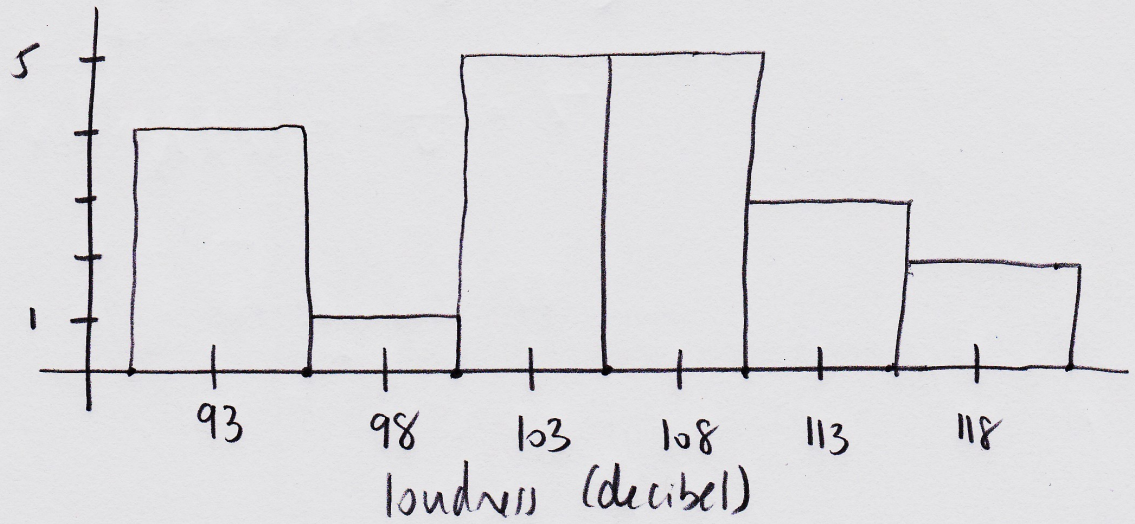
decibel	frequency	
91-95	IIII	4
96-100	I	1
101-105	IIIII	5
106-110	IIIII	5
111-115	III	3
116-120	II	2

b) Draw histogram

Use middle of class as the "class mark"

$\frac{91+95}{2} = 93$

$\frac{96+100}{2} = 98$ etc.



Appendix A in Suggested Problems:

instructions for making histograms in Excel (won't be tested)

SAMPLING METHODS

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

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

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

Ex: 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 $0.7(50) = 35$

3) 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.

Ex: 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.

4) 1-in-k SYSTEMATIC SAMPLE: Randomly select one of the first k measurements in the population and every k-th measurement thereafter.

Ex: 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.

Ex: 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.

stratified random sample

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.

cluster sample

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

simple random sample

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

1-in-25 systematic sample