MM104/MM106/BM110
Statistics and Data Presentation

Lecture 5:
Sampling distributions

The importance of sampling
Statistics as (random) variables

Chris Robertson

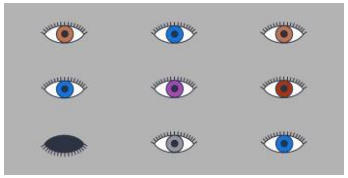
DEPARTMENT OF MATHEMATICS & STATISTICS

University of Strathclyde Science

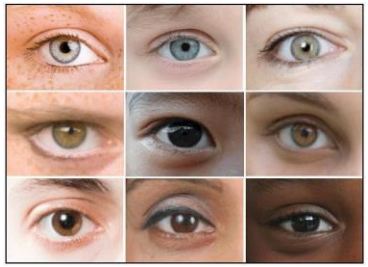
Sampling and Bias

Sampling

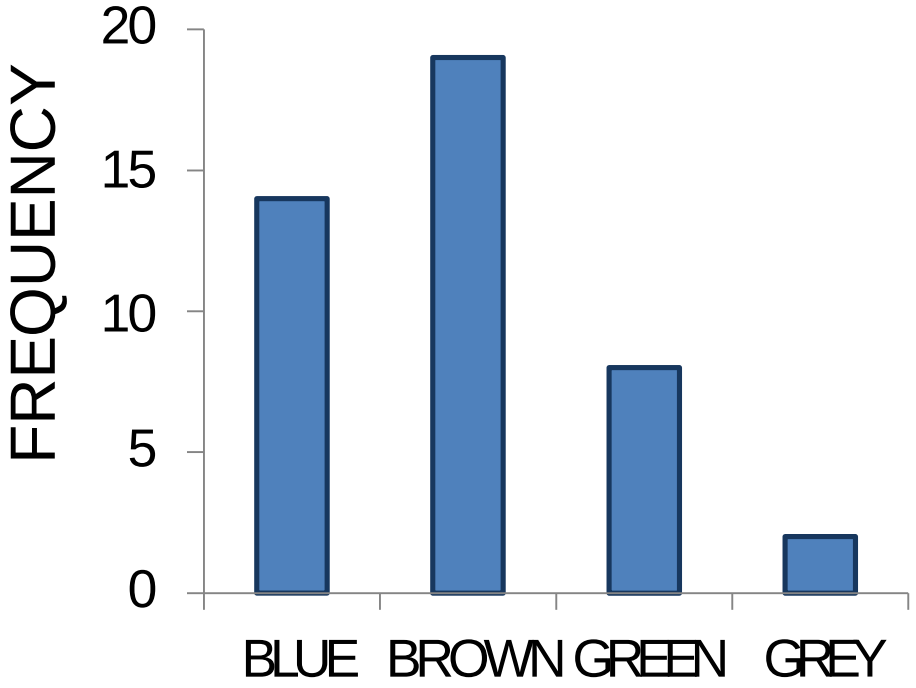
- Statistics allows us to use small samples to learn about the population:



Sampling



Sample size=43



- Bad sampling will influence conclusions.
- If a sample misrepresents the population you have bias, conclusions necessarily constrained to that sample!!

Common types of random samples.

- Simple random sample: all samples in population equally probable.
- Stratified random sample: samples organized into subpopulations or categories (*strata*).
  - Age group, gender, region could be strata
- Cluster random sample: samples taken from available clusters in population.
  - Schools, hospitals, geographic regions
- Systematic random sample: from ordered population, takes k elements then 1 every k.

Bias in Response

- Data collection needs to be done carefully!!

The way in which a question is phrased can influence the result:

"Would you favor or oppose a new US space program that would send astronauts to the moon?"

Favour: 53%; Oppose: 45%; No opinion: 2%

"Would you favor or oppose the US government spending billions of dollars to send astronauts to the moon?"

Favour: 31%; Oppose: 67%; No opinion: 2%

(CNN/USA Today/Gallup poll)

Sources of Bias in Samples

- Data collection needs to be done carefully!!
  Possible problems in sampling:

  — Wording bias, if using a questionnaire.

  "Have you ever committed a crime/been convicted?"

  — Undercoverage, if not all the possible outcomes/sectors are equally represented.

  — Bias due to a lack of response from uninterested (or overwhelming from opinionated) sectors.

  Example: TV polls in which viewers call to vote.

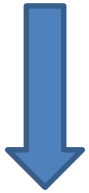
Key Points

- Taking a sample from a population is an efficient way of collecting data
- Samples are generally much smaller than the whole population
- Any sample of data which is not collected using RANDOM selection of the subjects to be in the sample will be biased
- You can minimise bias by selecting a RANDOM sample
  - cluster random or stratified random or simple random or systematic random
- Even random samples can be biased if there is non response

Statistics as (random) variables

Statistics as random variables

- Statistics from samples provide information about a population.



Statistics from samples are *estimators* for population *parameters*.

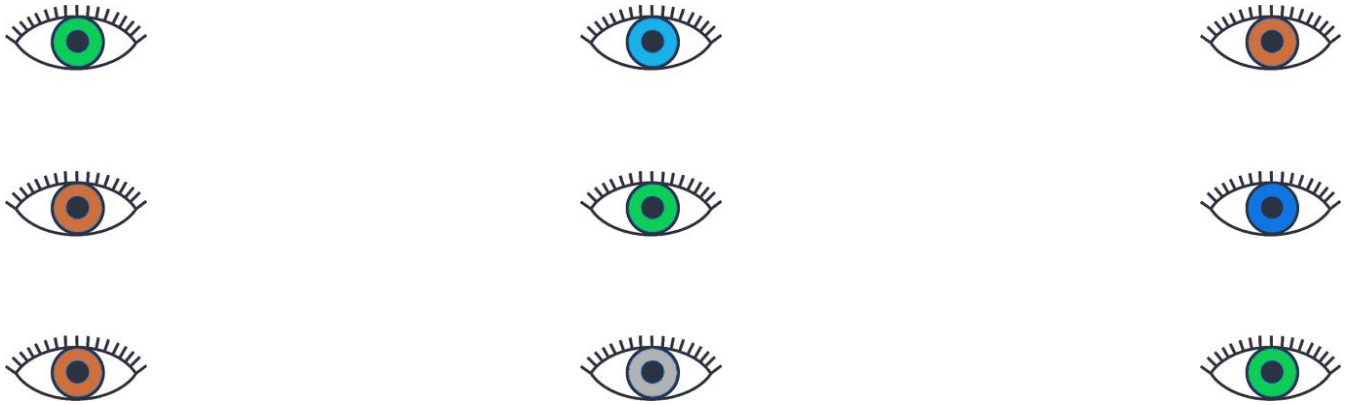
Sample mean, $\bar{x}$	ESTIMATES →	Population mean, μ
Sample standard deviation, s	→	Population standard deviation, σ

- Different samples provide different values for these statistics.

Repeated Samples

- Back to eye color: let's choose several 3-people samples.

Sample 1 | Sample 2 | Sample 3



- Different samples provided different values for statistics.
- Let's focus now on numerical data.

Sampling distributions

- Different samples → different value statistics → mean, median, etc. become random variables!!
- Example: population of 5 different numbers

3, 6, 9, 12, 15

Population mean value: $\mu = \frac{3+6+9+12+15}{5} = 9.$

One random sample of 3 numbers:

3, 6, 12  $\bar{x} = \frac{3+6+12}{3} = 7$

How many more combinations/samples can you find?

Sampling distributions

- Different samples → different value statistics → mean, median, etc. become random variables!!
- population of 5 different numbers 3, 6, 9, 12, 15
- How many possible samples? $5^3 = 125$

Sample	Values	Sample $\bar{x}$
1	3, 6, 9	6
2	3, 6, 12	7
3	3, 6, 15	8
4	3, 9, 12	8
5	3, 9, 15	9
6	3, 12, 15	10
7	6, 9, 12	9
8	6, 9, 15	10
9	6, 12, 15	11
10	9, 12, 15	12

Some Examples

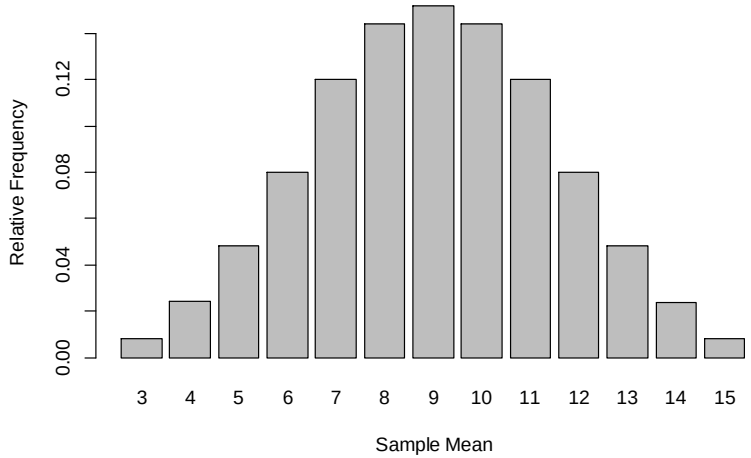
FREQUENCY TABLE

Mean	Frequency
3	1
4	3
5	6
6	10
7	15
8	18
9	19
10	18
11	15
12	10
13	6
14	3
15	1

Sampling distributions

The distribution of the sample mean over all the possible samples from the population is known as the

Sampling Distribution of the sample mean



Sampling Distribution of the sample mean

Mean	Frequency
3	1
4	3
5	6
6	10
7	15
8	18
9	19
10	18
11	15
12	10
13	6
14	3
15	1

Note that the shape is very similar to a normal distribution

Mean of sampling distribution

Population 3, 6, 9, 12, 15

Population mean value: $\mu = \frac{3+6+9+12+15}{5} = 9.$

$$\boldsymbol{Mean} = \frac{\sum_{i=1}^{n} \bar{x}_i}{n} =$$

$$= \frac{3 + 4*3 + 5*6 + \ldots + 13*6+14*3+15}{125};$$

Mean = 9

(Mean of the distribution of the sample means)

Mean	Frequency	F*M
3	1	3
4	3	12
5	6	30
6	10	60
7	15	105
8	18	144
9	19	171
10	18	180
11	15	165
12	10	120
13	6	78
14	3	42
15	1	15
	Sum	1125
	Mean	9

Mean of sampling distribution

The mean of the sampling distribution of the sample mean is the population mean

$$\mu_{\bar{X}} = \mu$$

Standard Deviation of the sampling distribution

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{(3-9)^2 + (6-9)^2 + (9-9)^2 + (12-9)^2 + (15-9)^2}{5}$$

$$\sigma^2 = \frac{36 + 9 + 0 + 9 + 36}{5}$$

$$= \frac{90}{5} = 18$$

Population Size

Dividing by N as this is the variance in the population

Not a sample

Standard Deviation of the sampling distribution

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{(3-9)^2 + (6-9)^2 + (9-9)^2 + (12-9)^2 + (15-9)^2}{5}$$

$$\sigma^2 = \frac{36 + 9 + 0 + 9 + 36}{5}$$

$$= \frac{90}{5} = 18$$

The standard deviation of the sampling distribution of the mean is called the standard error of the mean, denoted $\sigma_{\bar{X}}^2$

$$\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n} = \frac{18}{3} = 6$$

Sample Size

$\bar{x}$	F	$\bar{x} - \mu$	$(\bar{x} - \mu)^2$	$F(\bar{x} - \mu)^2$
3	1	-6	36	36
4	3	-5	25	75
5	6	-4	16	96
6	10	-3	9	90
7	15	-2	4	60
8	18	-1	1	18
9	19	0	0	0
10	18	1	1	18
11	15	2	4	60
12	10	3	9	90
13	6	4	16	96
14	3	5	25	75
15	1	6	36	36
			Sum	750
			$\sigma_{\bar{X}}^2$	6

Standard Error of sampling distribution of sample mean

The standard deviation of the sampling distribution of the sample mean is known as the standard error of the sample mean

It is equal to the square root of the variance of the population divided by the sample size

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Key Points

- Any statistic that you calculate on a sample has a sampling distribution
- Sampling distribution of the sample mean, the sample proportion
  - These are the two commonest statistics to use and we will look at these in the next section
  - It turns out that the normal distribution can be used for both of these sampling distributions
- Could have the sampling distribution of the sample median or the sample variance or the sample standard deviation
  - These are less common to use and also mathematically harder to derive the probability distribution
  - We will not look at these in this class