ST221 Introduction to Statistics

Dr Rafael Moral

Dept of Mathematics and Statistics
Maynooth University

rafael.deandrademoral@mu.ie

5.1 Sampling distributions

Parameter: Population characteristic.
For example, μ, σ, σ^2, π.

Sample statistic: Any quantity computed from values in a sample.
For example $\bar{x}$, s, s^2, p.

The value of a population characteristic is fixed, but if you take, for example, 10 samples from a population and compute $\bar{x}$ for each sample, would you expect each $\bar{x}$ to be the same?

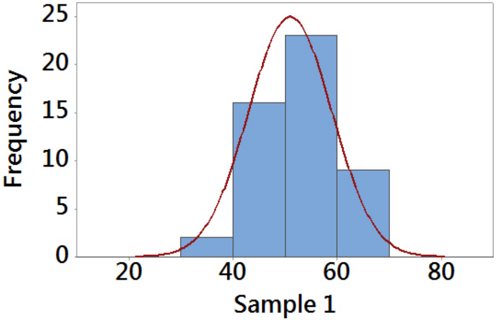
A sample statistic (considered in the context of any possible sample from the population) is a random variable and it has a probability distribution called the 'sampling distribution'.

NB: The 'sampling distribution' of a sample statistic is NOT the same as the 'sample distribution' which is the distribution of the raw data in the sample.

The sampling distribution of the mean

Example 1

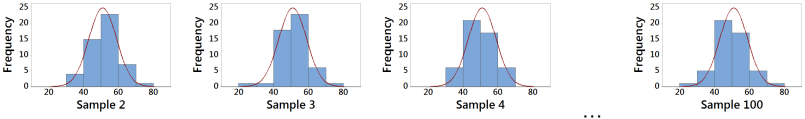
Suppose exam results in a population are normally distributed with a mean of 51 and a variance of 64. If we take a sample of size 50 from this population, what would we expect a histogram of the data to look like?



$\bar{x}_1 = 53.03$

Example 1 contd

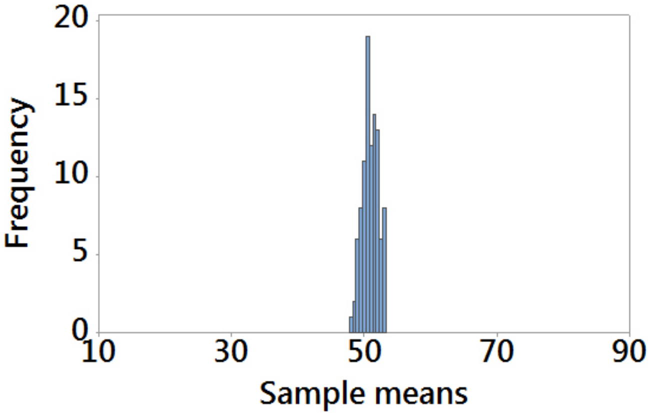
Suppose we repeat this process 100 times, i.e. take 100 samples each of size 50 from the N(51, 64) population and record $\bar{x}$ in each case.



$\bar{x}_2$ =53.03 $\bar{x}_3$=51.81 $\bar{x}_4$=51.32 ... $\bar{x}_{100}$=49.31.

What would we expect a histogram of the means from the 100 samples to look like?

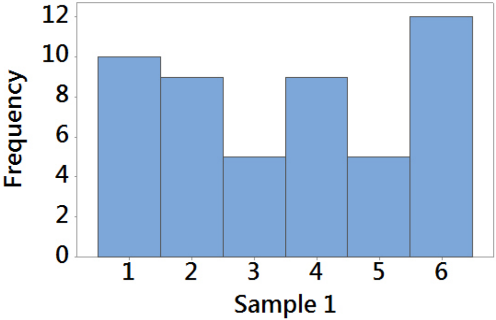
Example 1 contd



The sampling distribution of the mean

Example 2

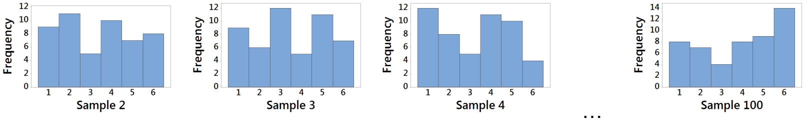
If you roll a fair die, you are equally likely to get any number from 1 to 6, i.e. the distribution of this population is uniform. Suppose we take a sample of size 50 from this population, i.e. roll a die 50 times and record the values.



$\bar{x}_1 = 3.52$

Example 2 contd

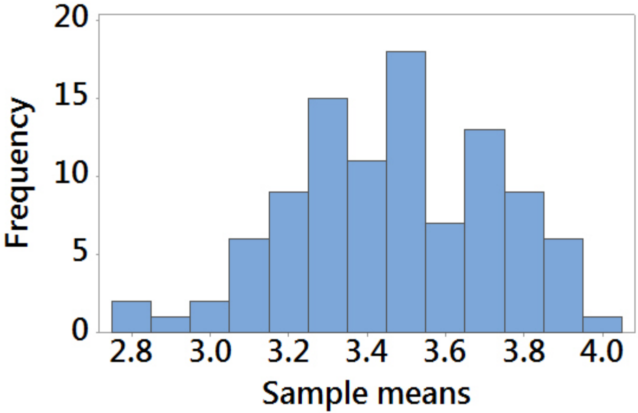
Suppose we repeat this process 100 times, i.e. take 100 samples each of size 50 from the Uniform[1,6] population and record $\bar{x}$ in each case.



$\bar{x}_2$=3.38 $\bar{x}_3$=3.48 $\bar{x}_4$=3.22 ... $\bar{x}_{100}$=3.90.

What would we expect a histogram of the means from the 100 samples to look like?

Example 2 contd



How would you describe the shape of this data?

Recap

A **sample statistic** (considered in the context of any possible sample from the population) is a random variable and it has a probability distribution called the '**sampling distribution**'.

NB: The 'sampling distribution' of a sample statistic is NOT the same as the 'sample distribution' which is the distribution of the raw data in the sample.

5.2 The Central Limit Theorem

Any linear combination of n independent random variables has an approximate normal distribution for large n.

What does this mean? In practice...

What does this NOT mean? In practice...

What does this mean? Algebraically:

Suppose $X_1, ..., X_n$ are independent random variables with mean $\mathrm{E}[X_i] = \mu_i$ and variance $\mathrm{Var}(X_i) = \sigma_i^2$. Then for all constants $a_1, ..., a_n$

$$\sum_{i=1}^{n} a_i X_i \sim N(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2)$$

approximately for large n (usually the approximation works well for $n \geq 30$). We can see

$$E[\sum_{i=1}^{n} a_i X_i] = \sum_{i=1}^{n} E[a_i X_i] = \sum_{i=1}^{n} a_i E[X_i] = \sum_{i=1}^{n} a_i \mu_i$$

and

$$Var(\sum_{i=1}^{n} a_i X_i) = \sum_{i=1}^{n} Var(a_i X_i) = \sum_{i=1}^{n} a_i^2 Var(X_i) = \sum_{i=1}^{n} a_i^2 \sigma_i^2$$

1. Special case of the CLT

Let $X_1, \ldots, X_n$ be independent random variables. Let $\mu_1 = \mu_2 = \ldots = \mu_n = \mu$
$\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_n^2 = \sigma^2$
$a_1 = a_2 = \ldots = a_n = 1$.
Let $S = \sum_{i=1}^n X_i$. Then

$$S \underset{approx}{\sim} N(n\mu, n\sigma^2)$$

We can see that

$$E[\sum_{i=1}^n X_i] = \sum_{i=1}^n E[X_i] = \sum_{i=1}^n \mu = n\mu$$

and

$$Var(\sum_{i=1}^n X_i) = \sum_{i=1}^n Var(X_i) = \sum_{i=1}^n \sigma^2 = n\sigma^2$$

2. Special case of the CLT

Let $X_1, ..., X_n$ be independent random variables. Let $\mu_1 = \mu_2 = ... = \mu_n = \mu$
$\sigma_1^2 = \sigma_2^2 = ... = \sigma_n^2 = \sigma^2$
$a_1 = a_2 = ... = a_n = \frac{1}{n}$. Let $\bar{X} = \dfrac{\sum_{i=1}^n X_i}{n}$. Then

$$\bar{X} \underset{approx}{\sim} N(\mu, \frac{\sigma^2}{n})$$

We can see that

$$E[\frac{\sum_{i=1}^n X_i}{n}] = \frac{1}{n}\sum_{i=1}^n E[X_i] = \frac{1}{n}\sum_{i=1}^n \mu = \frac{n\mu}{n} = \mu$$

and

$$Var(\frac{\sum_{i=1}^n X_i}{n}) = \frac{1}{n^2}\sum_{i=1}^n Var(X_i) = \frac{1}{n^2}\sum_{i=1}^n \sigma^2 = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

CLT Illustration (separate handout)

Example

Suppose that scores from a test have a distribution with mean 50 and standard deviation 10.

Suppose that a class of 35 students represents a random sample from the distribution. What is the probability that their average test result is great than 55?

Example

An online shop sells t-shirt merchandise from four different television shows. The proportion of t-shirts for each show and the profit per show is

	A	B	C	D
Profit €	1	0.5	3	2
Proportion	0.2	0.4	0.2	0.2

If the store sells 100 t-shirts in a day, what is the probability they make less than €150 profit?

5.3 Normal approximation to the binomial

Let $X \sim \text{Binomial}(n, p)$. Remember that $\text{E}[X]{=}np$ and $\text{Var}(X){=}np(1-p)$.

Then, for large n

$$X \underset{approx}{\sim} N(np, np(1-p))$$

To see this, suppose that $X = \#$ heads in n tosses of a coin, $p{=}\text{P(Head)}$.

We can write $X = X_1 + ... + X_n$, where $X_i = 1$ if the i^{th} toss is a head and 0 if it is a tail and each $X_i \sim \text{Binomial}(n = 1, p)$.

Since X is a sum of independent random variables $X_1, ..., X_n$ and each X_i has $\text{E}[X_i]{=}p$ and $\text{Var}(X_i){=}p(1-p)$, then, by the CLT

$$X \underset{approx}{\sim} N(np, np(1-p))$$

Example

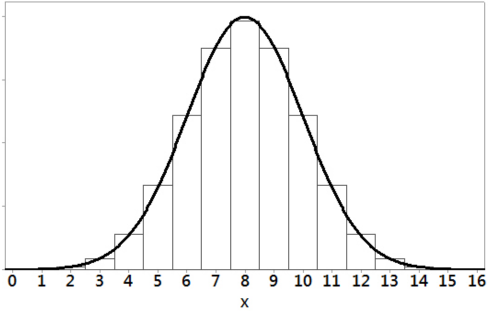
O-negative blood can be given to anyone, regardless of the recipient's blood type. Suppose that 6% of people from a particular population have O-negative blood. A blood donor unit is set up and it samples randomly from this population. It is estimated that it will need at least 1850 units of O-negative blood this year and that it will sample 32000 donors over the course of the year.

What is the probability that the unit will fall short of its O-negative requirement?

Note about continuity correction

When approximating a discrete distribution by a continuous distribution, we are using a curve instead of a histogram to calculate probabilities.

Example - graphical



Note about continuity correction contd

More generally, suppose X is a discrete random variable (takes values on the integer) which can be approximated by a normal random variable X^*. Then, $\mathrm{P}(X \leq x) \approx \mathrm{P}(X^* \leq x + \frac{1}{2})$. If follows that

$$
\begin{aligned}
P(X < x) &\approx P(X^* \leq x - \frac{1}{2}) \\
P(X \geq x) &\approx P(X^* \geq x - \frac{1}{2}) \\
P(X > x) &\approx P(X^* \geq x + \frac{1}{2}) \\
P(X = x) &\approx P(x - \frac{1}{2} \leq X^* \geq x + \frac{1}{2})
\end{aligned}
$$