Lecture 6: samples and populations

Today's lecture

- Look at fundamental concepts of samples and populations
- Intended to reinforce similar material in MAS2901
- Adopt a different perspective to MAS2901: use simulation rather than analytic calculation

Example

Type of problem looked at in MAS2901:

- Mercury waste dumped in a river
- Affects prawns which live in the river
- Max permitted level is one part per million on average
- A sample of prawns is collected and mercury content measured in these
- Attempt to infer the population mean mercury content from the sample

Example

Type of problem looked at in MAS2901:

- Mercury waste dumped in a river
- Affects prawns which live in the river
- Max permitted level is one part per million on average
- A sample of prawns is collected and mercury content measured in these
- Attempt to infer the population mean mercury content from the sample

Use a hypothesis test to decide whether population mean is greater than max allowed level – see MAS2901 for details

Populations

Suppose we measure some random quantity X

- X can adopt a range of possible values: some values are more likely than others
- This is the distribution of X
- Usually we do not know this distibution exactly
- The unknown distribution is called the population distribution

In the example:

- the population consists of the prawns in the estuary;
- the random quantity X is the mercury concentration in a randomly selected prawn; and
- the population distribution is the distribution of X.

Learning about populations

We are usually interested in key properties of the population distribution such as:

- the expectation of X – usually called the population mean;
- the variance of X – usually called the population variance; or
- the 95th percentile of X (for example).

Often we make some simplifying assumptions about the population distribution. For example, we might assume:

(a) X is normally distributed with unknown mean and variance;

(b) X is exponentially distributed with rate parameter λ, where λ is uknown but lies on the interval $(0, 1)$;

(c) X is normally distributed with unknown mean and variance $\sigma^2 = 5$.

A set of assumptions like this is referred to as a model.

Fully-specified population distributions

In some situations – usually rather artificial ones – we know the population distribution exactly.

For example:

- let X be the score obtained from rolling a fair die; or
- let X be the number on a card drawn at random from a full deck. (Assume Jack, Queen, King numbered 11,12,13 respectively.)

Samples

- We do not know everything about the population distribution
- We learn about the population distribution by drawing a sample
- A sample of size n corresponds to taking n independent measurements from the distribution
- Each measurement is a random variable with the same distribution as X: the sample measurements denoted $X_1, X_2, \ldots, X_n$
- The actual measurements obtained are denoted $x_1, x_2, \ldots, x_n$

Samples

- We do not know everything about the population distribution
- We learn about the population distribution by drawing a sample
- A sample of size n corresponds to taking n independent measurements from the distribution
- Each measurement is a random variable with the same distribution as X: the sample measurements denoted $X_1, X_2, \ldots, X_n$
- The actual measurements obtained are denoted $x_1, x_2, \ldots, x_n$

The distinction between the population distribution and how we learn about the population from limited samples is probably the most important concept in statistics

Estimators

- Suppose we wish to learn about some aspect of the population distribution e.g. population mean or population variance
- We construct an estimator for the quantity of interest
- For example, for population mean, a good estimator is the sample mean

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Estimators

- Suppose we wish to learn about some aspect of the population distribution e.g. population mean or population variance
- We construct an estimator for the quantity of interest
- For example, for population mean, a good estimator is the sample mean

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

- Formally, an estimator is defined to be some function of the sample:

$$S = g(X_1, X_2, \ldots, X_n)$$

  for some function g
- When we observe some measurements $X_1 = x_1, \ldots, X_n = x_n$ then we can compute an estimate $s = g(x_1, x_2, \ldots, x_n)$.

Simulation study of estimators

Since any estimator S is a random variable it makes sense to talk about its distribution – we can use simulation to do this

Example 6.2: Suppose the population distribution is normal, and we wish to estimate the population mean. Suppose the sample size is $n = 4$ and our estimator is $\bar{X} = (X_1 + X_2 + X_3 + X_4)/4$.

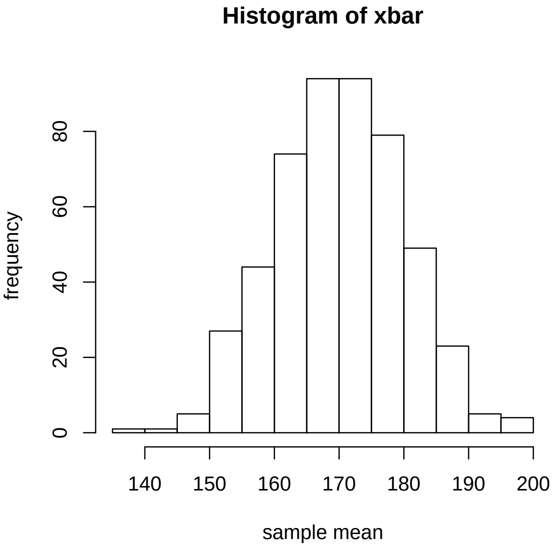
What is the distribution of $\bar{X}$ when the population distribution is $N(170, 20^2)$?

Example 6.2 – R code

```
simulate.sample.mean = function(n) {
    xbar = vector(mode="numeric",length=n)
    for (i in 1:n) {
        x = rnorm(4,170,20) # Generate a sample of size 4
        xbar[i] = 0.25*sum(x)
    }
    xbar
}

xbar=simulate.sample.mean(500)
hist(xbar,xlab="sample mean",ylab="frequency")
```

Example 6.2 – plot



Example 6.3

Suppose the population distribution is normal, and we wish to estimate the 90th percentile using a sample of size 10.

A sensible estimator is to define S to be the second largest value in the sample (i.e. the 9th value when the samples are ordered from smallest to largest).

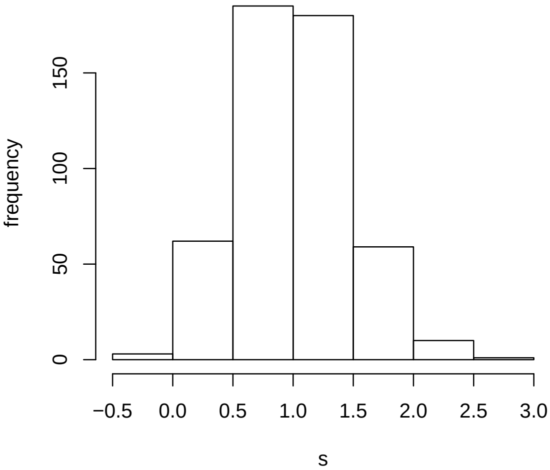
What is the distribution of S when the population distribution is $N(0, 1)$?

Example 6.3 – R code

```
simulate.percentile = function(n) {
    s = vector(mode="numeric",length=n)
    for (i in 1:n) {
        x = rnorm(10,0,1) # Generate a sample of size 10
        x = sort(x)
        s[i] = x[9] # Get 9th value on sorted list
    }
    s
}

s=simulate.percentile(500)
hist(s,xlab="s",ylab="frequency",main="")
```

Example 6.3 – plot

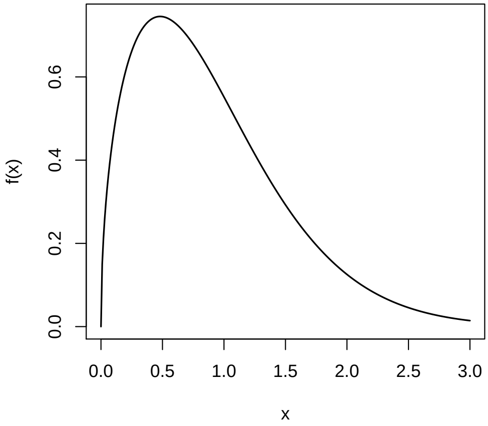


What does the distribution of $\bar{X}$ look like?

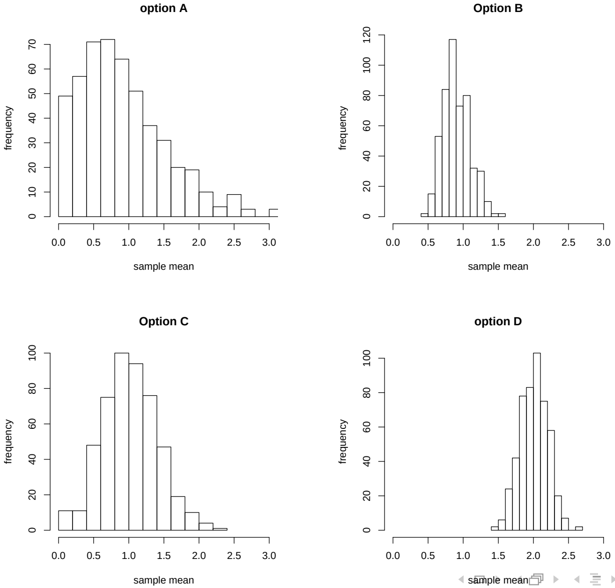
Consider the following two examples for the density of the population distribution.

For each example, decide which histogram on the slides (A, B, C or D) is most likely to represent the distribution of the sample mean $\bar{X}$ when the sample size is 10...

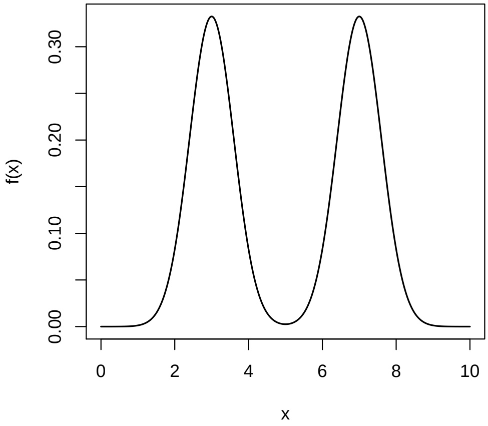
Example 6.4



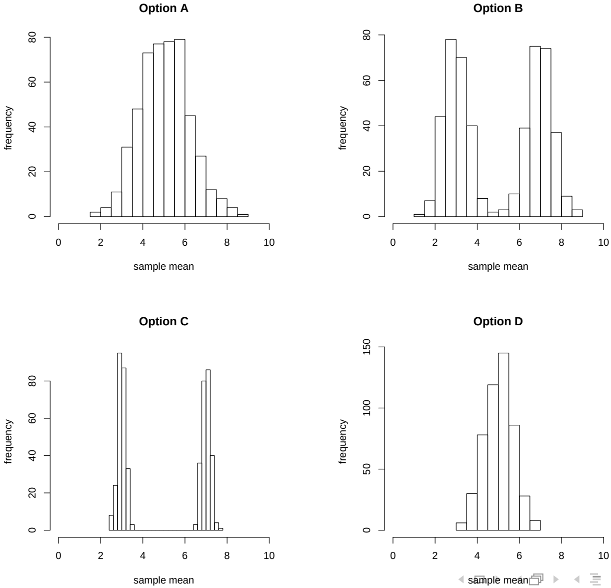
Options A–D


option A
Option B
Option C
option D
frequency
sample mean
0
10
20
30
40
50
60
70
80
100
120
0.0
0.5
1.0
1.5
2.0
2.5
3.0

Example 6.5



Options A–D



Option A
frequency
0
20
40
60
80
0
2
4
6
8
10
sample mean
Option B
frequency
0
20
40
60
80
0
2
4
6
8
10
sample mean
Option C
frequency
0
20
40
60
80
0
2
4
6
8
10
sample mean
Option D
frequency
0
50
100
150
0
2
4
6
8
10
sample mean

Answers

Example 6.4: option B

Example 6.5: option D

Conclusions

- The sample mean is distributed around the population mean.
- The distribution of sample mean values ‘forgets’ the underlying shape of the population distrubition.
- As n increases we expect the distribution of $\bar{X}$ to become more clustered around the true value.

The central limit theorem

Suppose $X_1, X_2, \ldots, X_n$ are independent and identically distributed random variables with common mean μ and variance σ^2 which are both finite.

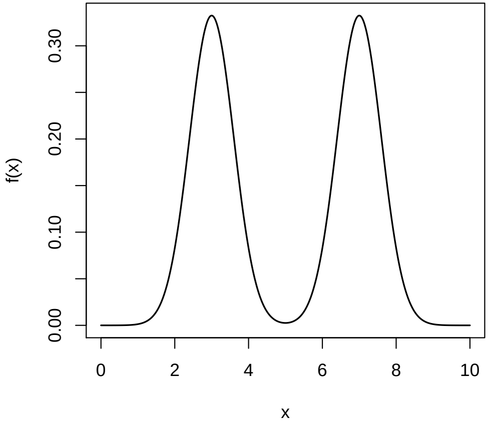
Define

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}.$$

Then as $n \to \infty$ the distribution of Z tends to $N(0, 1)$.

CLT via simulation

Population distribution: normal mixture with two components

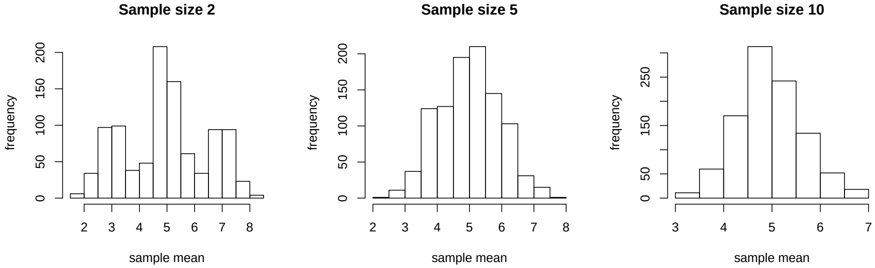


The population mean is $\mu = 5$ and variance is $\sigma^2 = 4.3$.

R code for sampling $\bar{X}$

```
simulate.bimod = function(k,n) {
    # Generate k samples of size n
    s = vector(mode="numeric",length=k)
    for (i in 1:k) {
        u = rnorm(n,3,0.6)
        v = rnorm(n,7,0.6)
        r = runif(n)
        x = c(u[r>0.5],v[r<=0.5])
        s[i] = mean(x)
    }
    s
}
```

Histograms from simulations of $\bar{X}$



Mean and variance for simulated $\bar{X}$

Sample size n	μ	σ^2/n	Simulated mean of $\bar{X}$	Variance of $\bar{X}$
2	5.0	2.15	4.94	2.27
5	5.0	0.86	4.98	0.862
10	5.0	0.43	4.96	0.443