Statistical Analysis

Sampling distributions

Ana Cristina Costa

ccosta@novaims.unl.pt

Topics

- **LU3 – Sampling distributions**
  - Introduction and concepts
  - Sample statistics and sampling distributions
  - Distribution of the sample mean
  - Distribution of the sample proportion

Objectives

- **At the end of this learning unit students should be able to**
  - Understand the concept of sampling distribution
  - Describe the Central Limit Theorem
  - Identify the distribution of the sample mean and apply it
  - Identify the distribution of the sample proportion and apply it
  - Explain the impact of the sample size on the sampling distribution

Suggested reading

- Newbold, P., Carlson, W. L., Thorne, B. (2013). Statistics for Business and Economics. 8th Edition, Boston: Pearson, pages 244-260 (ch. 6), 265-270 (ch. 6).

- Mariappan, P. (2019). Statistics for Business. New York: Chapman and Hall/CRC, pages 189-206 (ch. 12).

Introduction and concepts

- **General objectives**

A population is the entire set of elements having one or more common characteristics

- All cows in India
- All customers shopping at a department store on a chosen day
- All computer chips produced this month at a semiconductor plant
- All families in Lisbon, Portugal

Often, we are interested in estimating a population parameter

- Average weight of all cows in India
- The standard deviation of the amount spent by a department store customer
- Proportion of all computer chips that are defective
- Median income of families in Lisbon, Portugal

Introduction and concepts

- **General objectives**
  - In practical situations, you may be able to decide which type of probability distribution to use as a model, but the values of the parameters that specify its exact form are unknown
    - A pollster is sure that the responses to his “agree/disagree” questions will follow a binomial distribution, but p, the proportion of those who “agree” in the population, is unknown
    - An agronomist believes that the yield per acre of a variety of wheat is approximately normally distributed, but the mean μ and standard deviation σ of the yields are unknown.
  - In these cases, you must rely on the sample to learn about these parameters. The proportion of those who “agree” in the pollster’s sample provides information about the actual value of p. The mean and standard deviation of the agronomist’s sample approximate the actual values of μ and σ. If you want the sample to provide reliable information about the population, however, you must select your sample in a certain way*!*

Introduction and concepts

- **Sample**

A sample is a subset of the population that we observe to glean insights about the population

Sampling reduces

- Costs
- Implementation time
- Measurement error

A sample rather than a census when

- The population is large or infinite
- Collecting information may involve destroying elements of the population
- The scope is limited but precise
- The population is 'hard-to-reach', hidden or elusive (e.g. unregulated workers, child labour, illegal immigrants, homeless people, drug users, sex workers)

Introduction and concepts

- **Potential sources of error**

SAMPLING ERROR	NON-SAMPLING ERROR
Occurs because the sample is not the whole population	Occurs because of factors that are independent of the survey plan
	Can occur at any stage of the survey or census
	Can not be calculated, although it can be controlled and minimised

Introduction and concepts

- **Sampling error**

Discrepancy between the true population parameter and the sample statistic

- It is a random error because the estimates behave randomly around the true value of the parameter
- The sampling error can be estimated from the sample observations for a given confidence level
- The sampling error tends to decrease when the sample size increases

Introduction and concepts

Sources of non-sampling error

Instrument	Respondent	Interviewer / processor
• Insufficient or defective specification of units, scales, etc. • Defective questionnaire • Defective measurement tools • Incorrect or outdated secondary information (e.g. demographic or administrative data)	• Respondents bias answers in order to influence a particular outcome • Respondents are forced to value attributes with which they have little or no experience • Non-response	• Lack of training • Poor coding and editing of the questionnaire • Mistakes in data entry • Programming errors

Sample statistics and sampling distributions

- **Definitions**

Population	Set of elements with one attribute of interest The population is represented by a random variable ***X***
Random sample	Set of independent and identically distributed (**iid**) random variables **{X_1, X_2, …, X_n}** with the same probability distribution of *X*
Observed sample	Set of specific values {x_1, x_2, …, x_n}

Sample statistics and sampling distributions

- **Definitions**

Parameter	Numerical characteristic of the population
Statistic	Function of the random variables that form the sample Therefore, it is also a random variable
Estimator	Function of the random variables that form the sample, which is used for estimating an unknown parameter Therefore, it is also a random variable
Estimate	Numerical value assumed by an estimator for a specific sample Therefore, it is a numerical value taken by the estimator

Sample statistics and sampling distributions

- **Notation**

Population parameter	θ
Estimator of θ	$\widehat{\Theta} = g(X_1, X_2, \dots, X_n)$
Estimate of θ	$\hat{\theta} = g(\mathrm{x}_1, \mathrm{x}_2, \dots, \mathrm{x}_n)$
Joint probability distribution of the random sample	$f(x_1, x_2, \dots, x_n) = f(x_1) f(x_2) \cdots f(x_n)$

Sample statistics and sampling distributions

- **Main parameters and sample statistics**

Parameter (θ)			Statistic / Estimator ($\hat{\theta}$)	
μ	Mean	→	$\overline{X}$	Sample mean
σ^2	Variance	→	S^2	Sample variance
σ	Standard deviation	→	S	Sample standard deviation
p	Proportion	→	$\hat{p}$	Sample proportion

Sample statistics and sampling distributions

- **Sample statistics**

  Let X_1, X_2, …, X_n be a random sample of size *n*

  - **Sample mean**: $$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$
  - **Sample variance**: $$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^{\,2} - n\overline{X}^{\,2}\right]$$
  - **Sample standard deviation**: $$S = \sqrt{S^2}$$

Sample statistics and sampling distributions

- **Sample statistics**

Let $X_1, X_2, \ldots, X_n$ be iid random variables with **Bernoulli distribution**

Properties of **$X_i \sim B(p)$**

Probability function: $P(X=x) = p^x(1-p)^{1-x}$, $x = 0, 1;$ $0 \leq p \leq 1$

Mean and variance: $E(X) = p$ $V(X) = p(1-p) = pq$

- **Sample proportion**: indicates the proportion of successes in the sample

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad , X_i = 0, 1$$

Sample statistics and sampling distributions

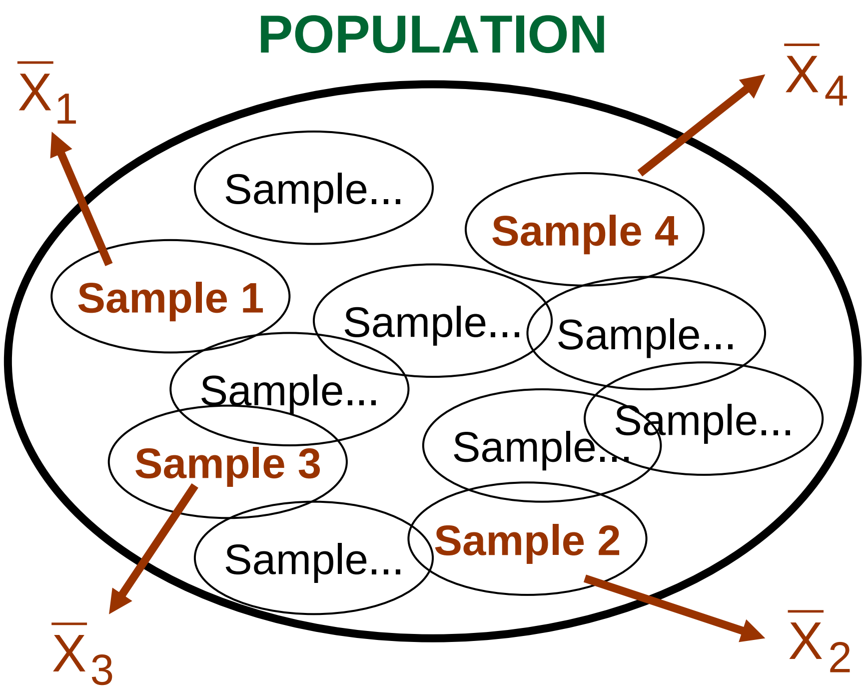
- **Sampling distribution**
  - The sampling distribution is the **probability distribution** of the statistic
    - All statistics have a sampling distribution
    - <u>Different samples of the same size produce different sample statistics values</u>
      - Some statistical values of a statistic are more likely to occur than others
      - The sampling distribution indicates the likelihood (probability) of obtaining certain values
    - The sampling distribution of a statistic can be described by parameters

Sample statistics and sampling distributions

- **Example**
  - Suppose that a sample of size n = 50 is collected from a population and the average of the 50 values is calculated (*sample mean*)
  - Then, suppose we collect a new sample of size n = 50 from this population and calculate the corresponding sample mean
  - Suppose that we repeat this process for all possible samples
  - The distribution of values of the sample mean that are obtained at the end of the process is called the **sampling distribution of the mean**

Sample statistics and sampling distributions

- **Example**



Sample		Mean
Sample 1	→	$\overline{X}_1$
Sample 2	→	$\overline{X}_2$
Sample 3	→	$\overline{X}_3$
⋮		⋮
Sample *i*	→	$\overline{X}_i$
⋮		⋮

Sample statistics and sampling distributions

- **Example 1**
  - Population: {3, 5, 6, 9, 11}
  - Parameters: μ **= 6.8** $\sigma^2 = 8.16$
  - All 10 possible samples (*without replacement*) of size n=2:

    {3,5}, {3,6}, {3,9}, {3,11}, {5,6}, {5,9}, {5,11}, {6,9}, {6,11}, {9,11}
  - Observed values of the sample mean:

    {4}, {4.5}, {6}, {7}, {5.5}, {7}, {8}, {7.5}, {8.5}, {10}
  - Sampling distribution of the sample mean given by its **probability function**:

$\bar{x}$	4	4.5	5.5	6	7	7.5	8	8.5	10
$P(\bar{X} = \bar{x})$	0.1	0.1	0.1	0.1	0.2	0.1	0.1	0.1	0.1

Sample statistics and sampling distributions

- **Sampling error**
  - It is the difference between the estimate obtained from the sample and the corresponding unknown parameter of the population: $\hat{\theta} - \theta$

- **Sampling distribution**
  - It is the probability distribution of the sample statistic
    - Depends on the distribution of the population and the sample size
    - Allows to evaluate and control the sampling error for any sample

Sample statistics and sampling distributions

- **Misconceptions related to distributions**
  - Confuse distribution of a population, a sample from the population, and a sampling distribution of a sample statistic
  - Assume two samples from the same population will be similar
  - Assume the sampling distribution will look like that of the population (for n>1)
  - Believe sampling distributions for small and large sample sizes have same variability

Burrill, G. (2019) Understanding Sampling Distributions: The Role of Interactive Dynamic Technology. In: S Budgett (Ed.), Decision Making Based on Data. Proceedings of the Satellite conference of the International Association for Statistical Education (IASE),August 2019, Kuala Lumpur, Malaysia

Distribution of the sample mean

- **Distribution of the sample mean – case I**
  - Let X_1, X_2, …, X_n be a random sample of **iid** random variables from a Normal population with mean μ and known variance σ^2

- Normal population
- σ^2 known
- Any sample size

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \quad or \quad \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

Distribution of the sample mean

- **Example 2**
  - A manufacturer claims that the duration of the spark plugs produced by him follows a normal distribution with mean 36000 km and standard deviation 4000 km. Suppose that a sample of 16 spark plugs was obtained, and their average duration was 34500 km. If the manufacturer's claim is correct, what is the probability of obtaining a sample mean as low or even lower?

✓ $\bar{X}$~N(36000, 1000) $\quad E(\bar{X}) = \mu = 36000 \quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{4000}{\sqrt{16}} = 1000$

✓ $P(\bar{X} < 34500) = P\left(Z < \frac{34500-36000}{1000}\right) = P(Z < -1.5) = 1 - 0.9332 = 0.0668$

✓ This low probability suggests that the manufacturer's claim may be true, because most of the samples would provide a sample mean greater than 34500 km.

➢ If 1000 samples of size 16 were obtained, only 67 (0.0668×1000) would have a sample mean lower than 34500 km

Distribution of the sample mean

Example 3

- Suppose that, based on historical data, we believe that the annual percentage salary increases for the chief executive officers of all midsize corporations are normally distributed with a mean of 12.2% and a standard deviation of 3.6%. A random sample of 9 observations is obtained from this population, and the sample mean is computed. What is the probability that the sample mean will be greater than 14.4%?

  - ✓ $\bar{X}$ ~N(12.2, 1.2) $\quad E(\bar{X}) = \mu = 12.2 \quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{3.6}{\sqrt{9}} = 1.2$

  - ✓ $P(\bar{X} > 14.4) = P\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} > \frac{14.4-12.2}{1.2}\right) = P(Z > 1.83) = 1 - 0.9664 = 0.0336$

  - ✓ If a sample mean greater than 14.4% *actually occurred*, we might begin to suspect that the population mean is greater than 12.2% or that we do not have a random sample that properly represents the population probability distribution.

Distribution of the sample mean

- **Variance of the sample mean**
  - The variance, or the standard deviation, of the sample statistic describes the spread of the statistic’s values from all possible samples
    - The greater the variance, the greater the difference between the statistic’s values
    - *A large variance is good or bad?*
  - A larger sample provides more information than a smaller sample. Hence, a statistic computed from a large sample should have a smaller *sampling error* than a statistic computed from a small sample.

Distribution of the sample mean

- **Variance of the sample mean**
  - Sampling with replacement from finite populations, or sampling from **infinite** populations (or the sample size is a small fraction of the population size)

$$\sigma^2{}_{\bar{X}} = V(\bar{X}) = \frac{\sigma^2}{n}$$

- The larger the sample size, n, the smaller the variance of the sample mean
  - The distribution of the sample mean becomes more concentrated around μ when the sample size increases. Hence, increasing the sample size increases the precision of the estimates of μ.

- The larger the population variance the larger the variance of the sample mean
  - The variance of the sample mean is proportional to the population variance

- $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$ is called the standard error of $\bar{X}$ because it refers to the precision of $\bar{X}$

Distribution of the sample mean

- **Variance of the sample mean**
  - Sampling without replacement from finite populations, or when the sample size is not a small fraction of the population size

$$\sigma^2{}_{\bar{X}} = \mathrm{V}(\bar{\mathrm{X}}) = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$$

(finite population correction factor)

- When samples are drawn without replacement, the observations are not selected independently
- If the sample size, *n*, is not a small fraction of the population size, N, then the individual sample members are not distributed independently of one another
  - Thus, the observations are not selected independently

Distribution of the sample mean

- **Example 4**
  - The duration of light bulbs produced by a plant follows a normal distribution with mean 450 hours and standard deviation 10 hours. Suppose that samples of size **n = 10** lamps are taken.
  - $\bar{X}$ ~N(450, 3.16) $E(\bar{X}) = \mu = 450$ $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{10}} = 3.16$

$$P\left(449 < \bar{X} < 451\right) = 0.251$$
$$P\left(448 < \bar{X} < 452\right) = 0.4713$$
$$P\left(447 < \bar{X} < 453\right) = 0.6579$$

✓ **25.1%** of the samples of size **n=10** will have lamps with a mean duration between 449 and 451 hours;

✓ **47.13%** of the samples of size **n=10** will have lamps with a mean duration between 448 and 452 hours;

✓ **65.79%** of the samples of size **n=10** will have lamps with a mean duration between 447 and 453 hours

Distribution of the sample mean

- **Example 4** (continued)
  - Suppose now that samples of size **n = 100** lamps are taken.
  - $\bar{X}$ ~N(450, 1) $E(\bar{X}) = \mu = 450$ $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{100}} = 1$

$$P(449 < \bar{X} < 451) = 0.6827$$
$$P(448 < \bar{X} < 452) = 0.9545$$
$$P(447 < \bar{X} < 453) = 0.9973$$

✓ **68.27%** of the samples of size **n=100** will have lamps with a mean duration between 449 and 451 hours;

✓ **95.45%** of the samples of size **n=100** will have lamps with a mean duration between 448 and 452 hours;

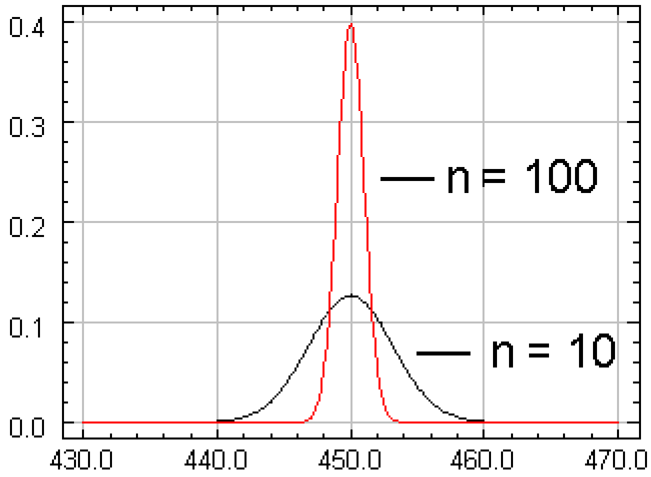
✓ **99.73%** of the samples of size **n=100** will have lamps with a mean duration between 447 and 453 hours

Distribution of the sample mean

- **Example 4** (continued)
  - Graphically

$$\overline{X}_{10} \sim N(450, 3.16)$$

$$\overline{X}_{100} \sim N(450, 1)$$



Distribution of the sample mean

- **Central Limit Theorem** (CLT)
  - Let $X_1, X_2, \ldots, X_n$ be a random sample of **iid** random variables from a population with mean μ and finite variance σ^2
    - Any population
    - σ^2 known
    - *Large* sample size

$$\bar{X} \overset{a}{\sim} N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \quad or \quad Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \overset{a}{\sim} N(0,1)$$

Z approaches the distribution N(0,1) when $n \to \infty$

Hogg, R., & Craig, A. (1995, 246) *Mathematical Statistics*. 5th ed., Englewood Cliffs, NJ: Prentice Hall

Distribution of the sample mean

- **Central Limit Theorem** (CLT)
  - The more asymmetric and away from the normal shape the population is, the larger must be the sample size
    - If the distributions are symmetric, then the means from samples of $n = 20$ to 25 are well approximated by the normal distribution
    - For skewed distributions, the required sample sizes are generally somewhat larger ($n \geq 50$)
    - The normal approximation is generally *satisfactory* if **$n \geq 30$**
  - The CLT can be applied to both discrete and continuous random variables, but only if the population has a finite variance
    - Happens in most situations
    - Counter-example: Cauchy distribution

Distribution of the sample mean

- **Example 5**
  - Antelope Coffee, Inc., is considering the possibility of opening a gourmet coffee shop in Big Rock, Montana. Previous research has indicated that its shops will be successful in cities of this size if the mean annual family income is above $70,000. It is also assumed that the standard deviation of income is $5,000 in Big Rock, Montana. A random sample of 36 people was obtained, and the mean income was $72,300. Does this sample provide evidence to conclude that a shop should be opened?
    - Assuming $\mu = 70000$ and $\sigma = 5000$, $\mathrm{P}(\bar{X} \geq 72300) = ?$
    - ✓ The distribution of incomes is known to be skewed, but the CLT enables us to conclude that $\bar{X} \overset{a}{\sim}$N(70000, 833.33)
    - ✓ $P(\bar{X} \geq 72300) \cong P\left(Z > \frac{72300-70000}{833.33}\right) \approx P(Z > 2.76) = 1 - 0.9971 = 0.0029$
    - ✓ Most of the samples would provide a sample mean lower than $72,300. *It is likely that the population mean income is higher than $70,000. The coffee shop is likely to be a success* (Newbold et al., 2013, pp. 259-260).

Distribution of the sample mean

- **Distribution of the sample mean – case II**
  - Let $X_1, X_2, \ldots, X_n$ be a random sample of **iid** random variables from a population with mean μ and unknown variance σ^2
    - Any population
    - σ^2 unknown
    - *Large* sample size
  - We must use the statistic S^2 to estimate σ^2

$$\bar{X} \overset{a}{\sim} N\left(\mu, \frac{S}{\sqrt{n}}\right) \quad or \quad Z = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \overset{a}{\sim} N(0,1)$$

Z approaches the distribution N(0,1) when $n \to \infty$

Distribution of the sample mean

- **Distribution of the sample mean – case III**
  - Let $X_1, X_2, \ldots, X_n$ be a random sample of **iid** random variables from a Normal population with mean μ and unknown variance σ^2
    - Normal population
    - σ^2 unknown
    - Any sample size

$$T = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t_{(n-1)}$$

  - We must use the statistic S^2 to estimate σ^2
  - To the variability of $\bar{X}$ adds the variability of S, and therefore the distribution of this r.v. is the Student's t distribution

Distribution of the sample mean

- **Example 6**
  - Suppose that, based on historical data, we believe that the annual percentage salary increases for the chief executive officers of all midsize corporations are normally distributed with a mean of 12.2%. A random sample of 9 observations is obtained from this population. The sample standard deviation is 3.6%, and the sample mean is also computed. What is the probability that the sample mean will be greater than 14.43%?

✓ $T = \frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}} \sim t_{(n-1)} \qquad \mu = 12.2 \qquad s = 3.6 \qquad n = 9$

✓ $P(\bar{X} > 14.43) = P\left(\frac{\bar{X}-12.2}{\frac{3.6}{\sqrt{9}}} > \frac{14.43-12.2}{3.6/\sqrt{9}}\right) = P\left(t_{(8)} > 1.86\right) = 1 - 0.95 = 0.05$

Distribution of the sample proportion

- **Bernoulli population**
  - X is a random variable with exactly two possible outcomes, "success“ (X=1) and "failure“ (X=0), where
    - *Success* occurs with probability p
    - *Failure* occurs with probability $q = 1 - p$

$$X \begin{cases} 1 & 0 \\ p & 1-p \end{cases}$$

$E(X) = p$ and $V(X) = pq$

  - We use the sample proportion $\hat{p}$ to estimate *p*, where $\hat{p}$ is the proportion of successes in a random sample drawn from a population with Bernoulli(*p*) distribution

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad , X_i = 0, 1$$

Distribution of the sample proportion

- **Distribution of the sample proportion**
  - $E(\hat{p}) = p \qquad V(\hat{p}) = \frac{p(1-p)}{n} = \frac{pq}{n}$
  - Based on the Central Limit Theorem:

$$\hat{p} \overset{a}{\sim} N\left(p, \sqrt{\frac{pq}{n}}\right) \quad or \quad Z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} \overset{a}{\sim} N(0,1)$$

$\hat{p}$ converges to the normal distribution when $n \to \infty$

- *Satisfactory* approximation if $n > 20$ and $0.1 < p < 0.9$,
  or ***$np \geq 5$ and $n(1-p) \geq 5$***

Distribution of the sample proportion

- **Example 7**
  - Suppose that, in the Country of Discommodity, the cable television In-Sult-Channel claims that 10% of the households subscribe it, which is true. However, given their dubious reputation, a marketing company decided to estimate this proportion from a sample of 100 households, before renewing their advertising contracts with the In-Sult-Channel. Assuming that contracts are renewed only if the sample proportion is greater than 8.5%, determine the probability of that happening.

✓ Since $np = 10 \geq 5$ and $n(1-p) = 90 \geq 5$, using the CLT we have

$$\hat{p} \overset{a}{\sim} N\left(0.1, \sqrt{\frac{0.1 \times 0.9}{100}}\right) \Leftrightarrow \hat{p} \overset{a}{\sim} N(0.1, 0.03)$$

✓ $P(\hat{p} > 0.085) \cong P\left(Z > \frac{0.085-0.1}{0.03}\right) = P(Z > -0.5) = P(Z < 0.5) = 0.691$

Distribution of the sample proportion

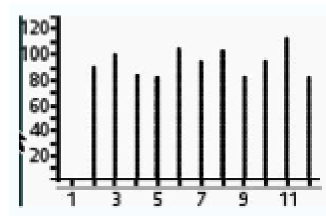
- **Example 8**
  - A random sample of 270 homes was taken from a large population of older homes to estimate the proportion of homes with unsafe wiring. If, in fact, 20% of the homes have unsafe wiring, what is the probability that the sample proportion will be between 16% and 24%?
    - ✓ Since *np* = *54* $\geq$ *5* and *n(1–p)* = *216* $\geq$ *5*, using the CLT we have

$$\hat{p} \overset{a}{\sim} N\left(0.2, \sqrt{\frac{0.2 \times 0.8}{270}}\right) \Leftrightarrow \hat{p} \overset{a}{\sim} N(0.2, 0.024)$$

    - ✓ $P(0.16 < \hat{p} < 0.24) \cong P\left(\frac{0.16-0.2}{0.024} < Z < \frac{0.24-0.2}{0.024}\right) = P(-1.67 < Z < 1.67) = 0.9050$

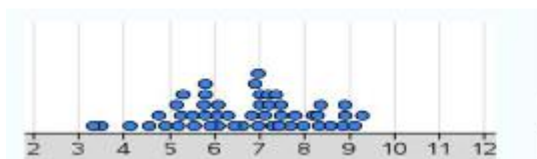
Review Quiz

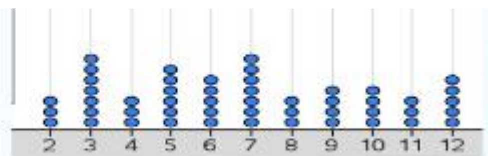
- Consider the following POPULATION



  - Which of the two frequency distributions is more likely to be a random SAMPLE (n=50)?

a

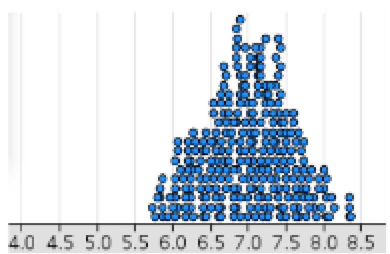


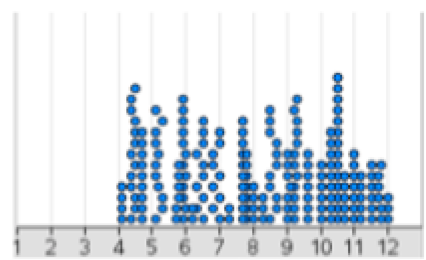


b

  - Which of the distributions could be a simulated DISTRIBUTION OF SAMPLE MEANS (n=200)?

a





b

Burrill, G. (2019) Understanding Sampling Distributions: The Role of Interactive Dynamic Technology. In: S Budgett (Ed.), Decision Making Based on Data. Proceedings of the Satellite conference of the International Association for Statistical Education (IASE),August 2019, Kuala Lumpur, Malaysia

Sampling distributions

Do the homework*!*

Ana Cristina Costa, ccosta@novaims.unl.pt