Simulation and Modelling

Spring 2023
CS4056

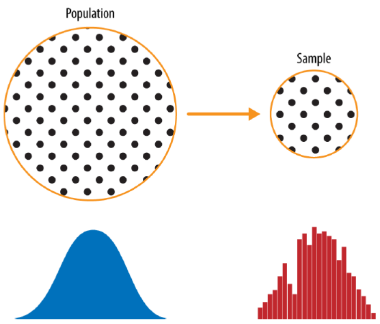
Muhammad Shahid Ashraf

Notes

Statistical Models in Simulation

Notes

Sample vs Population



Notes

Random Sampling

Sample
A subset from a larger data set.

Population
The larger data set or idea of a data set.

N (n)
The size of the population (sample).

Random sampling
Drawing elements into a sample at random.

Notes

Random Sampling

NATIONAL UNIVERSITY of Computer & Emerging Sciences

Stratified sampling
Dividing the population into strata and randomly sampling from each strata.

Stratum (pl., strata)
A homogeneous subgroup of a population with common characteristics.

Simple random sample
The sample that results from random sampling without stratifying the population.

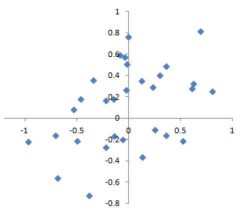
Bias
Systematic error.

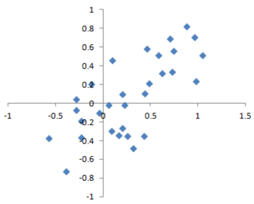
Sample bias
A sample that misrepresents the population.

Notes

Random Error vs Bias Error

NATIONAL UNIVERSITY of Computer & Emerging Sciences





Notes

Bias

NATIONAL UNIVERSITY of Computer & Emerging Sciences

- Selecting a sample to represent the population fairly is actually rather difficult.
- Sampling methods that, by their nature, tend to over- or under-emphasize some characteristics of the population are said to be biased.
- Conclusions based on samples drawn from biased methods are inherently flawed.

Notes

Random Sampling

NATIONAL UNIVERSITY of Computer & Emerging Sciences

- 1936 election: Franklin Delano Roosevelt vs. Alf Landon
- Literary Digest had called the election since 1916
- Sample size: 2.4 million!
- Prediction: Roosevelt 43%
- Actual: Roosevelt: 62%
- (Literary Digest went bankrupt soon after)

Notes

What happened

- Where did the pollsters get their 10 million names?
  - Phone numbers? In 1936, the height of the depression, phones were luxuries. Selecting from a sample of only upper class citizens wouldn't be representative.
  - Driver's licenses and/or memberships in organizations such as country clubs
  - Is this representative of the entire population when the major campaign issue was the economy?

Notes

Gallop Survey

- In 1936, George Gallop used a subsample of only 3000 of the 2.4 million responses that the Literary Digest received to reproduce the wrong prediction of Landon's victory over Roosevelt.
- He then used an entirely different sample of 50,000 and predicted that Roosevelt would get 56% of the vote to Landon's 44%.
- Gallop went of to become one of the leading polling companies.

Notes

Randomize

- The best defense against bias is randomization, in which each individual is given a fair, random chance of selection.
- Randomization also protects us from the influences of all the features of our population, even one we may not have thought about. It does that by making sure that on average the sample looks like the rest of the population.
- Randomization also makes it possible to draw inferences about the population when we see only a sample.
- The fraction of the population that you've sampled does not matter
- Sample size is of key importance in the design of a sample survey because it determines the balance between how well the survey can measure the population and how much the survey costs

Notes

Random Sampling

- Even in the era of big data, random sampling remains an important arrow in the data scientist's quiver.
- Bias occurs when measurements or observations are systematically in error because they are not representative of the full population.
- Data quality is often more important than data quantity, and random sampling can reduce bias and facilitate quality improvement that would otherwise be prohibitively expensive.

Notes

Sampling Distribution

Sample statistic
A metric calculated for a sample of data drawn from a larger population.

Data distribution
The frequency distribution of individual *values* in a data set.

Sampling distribution
The frequency distribution of a *sample statistic* over many samples or resamples.

Central limit theorem
The tendency of the sampling distribution to take on a normal shape as sample size rises.

Standard error
The variability (standard deviation) of a sample *statistic* over many samples (not to be confused with *standard deviation*, which by itself, refers to variability of individual data *values*).

Notes

Random Variables

- A random variable is a function of an outcome,

$$X = f(\omega)$$

- Consider the experiment of tossing two coins. We can define X to be a random variable that measures the number of heads observed in the experiment. For the experiment, the sample space is shown below:

$$S = \{HH, HT, TH, TT\}$$

- There are 4 possible outcomes for the experiment, this is the domain of X.
- For each outcome, the associated value is shown as:

$$X(H,H) = 2$$
$$X(H,T) = 1$$
$$X(T,H) = 1$$
$$X(T,T) = 0$$

Notes

Example

Consider an experiment of tossing 3 fair coins and counting the number of heads. Certainly, the same model suits the number of girls in a family with 3 children, the number of 1's in a random binary string of 3 characters, etc.

$$\begin{aligned}
\boldsymbol{P}\{X=0\} &= \boldsymbol{P}\{\text{three tails}\} = \boldsymbol{P}\{TTT\} = \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{1}{2}\right) = \frac{1}{8} \\
\boldsymbol{P}\{X=1\} &= \boldsymbol{P}\{HTT\} + \boldsymbol{P}\{THT\} + \boldsymbol{P}\{TTH\} = \frac{3}{8} \\
\boldsymbol{P}\{X=2\} &= \boldsymbol{P}\{HHT\} + \boldsymbol{P}\{HTH\} + \boldsymbol{P}\{THH\} = \frac{3}{8} \\
\boldsymbol{P}\{X=3\} &= \boldsymbol{P}\{HHH\} = \frac{1}{8}
\end{aligned}$$

x	$\boldsymbol{P}\{X=x\}$
0	1/8
1	3/8
2	3/8
3	1/8
Total	1

Notes

Distribution of X

Collection of all the probabilities related to X is the **distribution** of X. The function

$$P(x) = \boldsymbol{P}\{X = x\}$$

is the **probability mass function**, or **pmf**. The **cumulative distribution function**, or **cdf** is defined as

$$F(x) = \boldsymbol{P}\{X \le x\} = \sum_{y \le x} \boldsymbol{P}(y).$$

The set of possible values of X is called the **support** of the distribution F.

For every outcome ω, the variable X takes one and only one value x. This makes events $\{X = x\}$ disjoint and exhaustive

$$\sum_x P(x) = \sum_x P\{X = x\} = 1$$

Notes

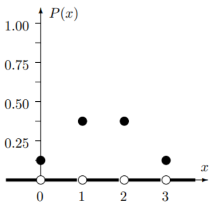
PMF and CMF Distribution of X

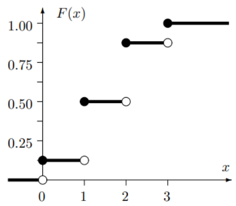
For any set

$$P\{X \in A\} = \sum_{x \in A} P(x)$$

When A is an interval, its probability can be computed directly from the cdf F(x),

$$P\{a < X \leq b\} = F(b) - F(a)$$

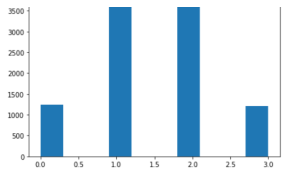




Notes

Histogram

- the two middle columns for X = 1 and X = 2 are about 3 times higher than the columns on each side.for X = 0 and X = 3.
- In a run of 10,000 simulations, values 1 and 2 are attained three times more often than 0 and 3.
- which is our pmf $P(0) = P(3) = 1/8, P(1) = P(2) = 3/8$



Notes

Example

- A program consists of two modules. The number of errors X_1 in the first module has the pmf $P_1(x)$, and the number of errors X_2 in the second module has the pmf $P_2(x)$, independently of X_1, where
- Find the pmf and cdf of $Y = X_1 + X_2$, the total number of errors

x	$P_1(x)$	$P_2(x)$
0	0.5	0.7
1	0.3	0.2
2	0.1	0.1
3	0.1	0

Notes

Types of Random Variables

- Discrete random variables: are random variables, whose range is a countable set. A countable set can be either a finite set or a countably infinite set. For instance, in the above example, X is a discrete variable as its range is a finite set $\{0, 1, 2\}$
- Continuous random variables, have a range in the forms of some interval, bounded or unbounded, of the real line. It can be e a union of several such intervals
- Mixed random variables are ones that are a mixture of both continuous and discrete variables. These variables are more complicated than the other two.

Notes

Examples of Random Variables

- A long jump is formally a continuous random variable because an athlete can jump any distance within some range. Results of a high jump, however, are discrete because the bar can only be placed on a finite number of heights.
- e. Examples of continuous variables include various times (software installation time, code execution time, connection time, waiting time, lifetime), also physical variables like weight, height, voltage.
- A job is sent to a printer.

Notes

Mean of a Discrete Random Variable

- The mean of a discrete random variable, denoted by μ, is actually the mean of its probability Distribution.

$$\mu = \sum xP(x)$$

- The mean of a discrete random variable x is also called its expected value and is denoted by E(x).

$$E(x) = \sum xP(x)$$

Notes

Examples

- Suppose that P(0) = 0.75 and P(1) = 0.25. Then, in a long run, X is equal 1 only 1/4 of times, otherwise it equals 0. Suppose we earn \$1 every time we see X = 1. On the average, we earn \$1 every four times, or \$0.25 per each observation
- Consider a variable that takes values 0 and 1 with probabilities P(0) = P(1) = 0.5.
- Consider two users.One receives either 48 or 52 e-mail messages per day, with a 50-50% chance of each. The other receives either 0 or 100 e-mails, also with a 50-50% chance. Calculate $E(x)$ for both users.

Notes

Variance and Standard Deviation

- Expectation shows where the average value of a random variable is located, or where the variable is expected to be, plus or minus some error.
- How large could this "error" be, and how much can a variable vary around its expectation
- In Previous slide ,consider the first case, the actual number of e-mails is always close to 50, whereas it always differs from it by 50 in the second case.
- The first random variable, X, is more stable; it has low variability. The second variable, Y , has high variability.
- variability of a random variable is measured by its distance from the mean $\mu = E(X)$

Notes

Variance and Standard Deviation

- Variance of a random variable is defined as the expected squared deviation from the mean. For discrete random variables, variance is

$$\sigma^2 = Var(x) = \sum_x (x-\mu)^2 P(x)$$

- Standard deviation is a square root of variance

$$\sigma = Std(X) = \sqrt{Var(X)}$$

Notes

Example

A rowing team consists of four rowers who weigh 152, 156, 160, and 164 pounds. Find all possible random samples with replacement of size two and compute the sample mean for each one. Use them to find the probability distribution, the mean, and the standard deviation of the sample mean $\overline{X}$.

Sample	Mean		Sample	Mean		Sample	Mean		Sample	Mean
152, 152	152		156, 152	154		160, 152	156		164, 152	158
152, 156	154		156, 156	156		160, 156	158		164, 156	160
152, 160	156		156, 160	158		160, 160	160		164, 160	162
152, 164	158		156, 164	160		160, 164	162		164, 164	164

Notes

Example

A rowing team consists of four rowers who weigh 152, 156, 160, and 164 pounds. Find all possible random samples with replacement of size two and compute the sample mean for each one. Use them to find the probability distribution, the mean, and the standard deviation of the sample mean $\overline{X}$.

- The table shows that there are seven possible values of the sample mean $\overline{X}$.
- The value $\overline{x} = 152$ happens only one way (the rower weighing 152 pounds must be selected both times), as does the value $\overline{x} = 164$,
- the other values happen more than one way, hence are more likely to be observed than 152 and 164 are

Notes

Example

- A rowing team consists of four rowers who weigh 152, 156, 160, and 164 pounds. Find all possible random samples with replacement of size two and compute the sample mean for each one. Use them to find the probability distribution, the mean, and the standard deviation of the sample mean $\overline{X}$.
- Since the 16 samples are equally likely, we obtain the probability distribution of the sample mean just by counting:

$\overline{x}$	152	154	156	158	160	162	164
$P(\overline{x})$	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{3}{16}$	$\frac{4}{16}$	$\frac{3}{16}$	$\frac{2}{16}$	$\frac{1}{16}$

Notes

Example

$\overline{x}$	152	154	156	158	160	162	164
$P(\overline{x})$	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{3}{16}$	$\frac{4}{16}$	$\frac{3}{16}$	$\frac{2}{16}$	$\frac{1}{16}$

For the mean and standard deviation of discrete random variable to $\overline{X}$.

For $\mu_{\overline{X}}$ we obtain.

$$\begin{aligned}\mu_X &= \Sigma\overline{x}\,P(\overline{x})\\ &= 152\left(\tfrac{1}{16}\right) + 154\left(\tfrac{2}{16}\right) + 156\left(\tfrac{3}{16}\right) + 158\left(\tfrac{4}{16}\right) + 160\left(\tfrac{3}{16}\right) + 162\left(\tfrac{2}{16}\right) + 164\left(\tfrac{1}{16}\right)\\ &= 158\end{aligned}$$

For $\sigma_{\overline{X}}$ we first compute $\Sigma\overline{x}^2 P\left(\overline{x}\right)$:

$$152^2\left(\frac{1}{16}\right) + 154^2\left(\frac{2}{16}\right) + 156^2\left(\frac{3}{16}\right) + 158^2\left(\frac{4}{16}\right) + 160^2\left(\frac{3}{16}\right) + 162^2\left(\frac{2}{16}\right) + 1$$

which is 24,974, so that

$$\sigma_{\overline{X}} = \sqrt{\Sigma\overline{x}^2 P\left(\overline{x}\right) - \mu_{\overline{x}}^2} = \sqrt{24{,}974 - 158^2} = \sqrt{10}$$

Notes

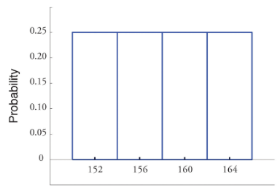
The Mean and Standard Deviation of the Sample Mean

- The mean and standard deviation of the population 152,156,160,164 in the example are $\mu = 158$ and $\sigma = \sqrt{20}$
- The mean of the sample mean $\overline{X}$ that we have just computed is exactly the mean of the population.
- The standard deviation of the sample mean $\overline{X}$that we have just computed is the standard deviation of the population divided by the square root of the sample size: $\sqrt{10} = \frac{\sqrt{20}}{\sqrt{2}}$
- These relationships are not coincidences, but are illustrations of the following formulas.
  - Suppose random samples of size n are drawn from a population with mean μ and standard deviation σ.
  - The mean $\mu_{\overline{X}}$ and standard deviation $\sigma_{\overline{X}}$ of the sample mean $\overline{X}$ satisfy

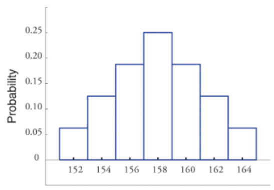
$$\mu_{\overline{X}} = \mu \quad \text{and} \quad \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

Notes

The Mean and Standard Deviation of the Sample Mean



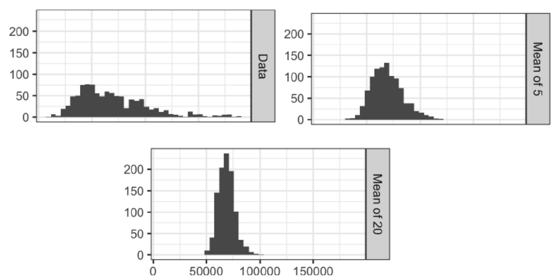
(a) Population



(b) Sample Mean

Notes

Sampling Distribution



Notes

Sampling Distribution

Consider the experiment of tossing a single die. Define X as the number of spots on the up face of the die after a toss. Then $R_X = \{1, 2, 3, 4, 5, 6\}$. Assume the die is loaded so that the probability that a given face lands up is proportional to the number of spots showing. The discrete probability distribution for this random experiment is

Notes

Sampling Distribution

Consider the experiment of tossing a single die. Define X as the number of spots on the up face of the die after a toss. Then $R_X = \{1, 2, 3, 4, 5, 6\}$. Assume the die is loaded so that the probability that a given face lands up is proportional to the number of spots showing. The discrete probability distribution for this random experiment is

x_i	1	2	3	4	5	6
$p(x_i)$	1/21	2/21	3/21	4/21	5/21	6/21

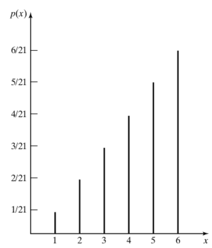
The conditions stated earlier are satisfied—that is,

- $p(x_i) \geq 0$ for $i = 1, 2, \ldots, 6$ and
- $\sum_{i=1}^{\infty} = 1/21 + 2/21 + 3/21 + 4/21 + 5/21 + 6/21 = 1$

Notes

Sampling Distribution

Consider the experiment of tossing a single die. Define X as the number of spots on the up face of the die after a toss. Then $R_X = \{1, 2, 3, 4, 5, 6\}$. Assume the die is loaded so that the probability that a given face lands up is proportional to the number of spots showing. The discrete probability distribution for this random experiment is



Notes

The Mean and Standard Deviation of the Sample Mean

The mean and standard deviation of the tax value of all vehicles registered in a certain state are $\mu = \$13{,}525$ and $\sigma = \$4{,}180$. Suppose random samples of size 100 are drawn from the population of vehicles. What are the mean $\mu_{\overline{X}}$ and standard deviation $\sigma_{\overline{X}}$ of the sample mean $\overline{X}$?

Solution

Since n = 100, the formulas yield

$$\mu_{\overline{X}} = \mu = \$13{,}525 \quad \text{and} \quad \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{\$4180}{\sqrt{100}} = \$418$$

Notes