Introduction to statistics

January 27, 2023

Outline

- Descriptive and Inference Statistics
- Population and sample
- Statistics: a function of random sample, which is used to make inference about population from information observed on sample
- Descriptive Statistics: summarize information of sample by statistics, histogram, boxplot ...

Statistic is the art of learning from data

1. Descriptive statistic: collect, summarize, report, store information (**data**)
2. Inferential statistic: interprete data, make decision/prediction, draw conclusion from data *in the face of **uncertainty** and **variation***

Example

- In a biomedical study of a new drug that reduces hypertension, 85% of patients experienced relief
- the current drug, or “old” drug, brings relief to 80% of patients that have chronic hypertension.
- Should the new drug be adopted?

Example (cont)

- The "85%" value is based on a certain number of patients chosen for the study.
- Perhaps if the study were repeated with new patients the observed number of "successes" would be 75%!

Example (cont)

- The “85%” value is based on a certain number of patients chosen for the study.
- Perhaps if the study were repeated with new patients the observed number of “successes” would be 75%!

Population - Sample

Population

- **A population** consists all the observation with which we concerne
- **Size of population**: the number of observations in the population (*finite or infinite*)
- Ex: If there are 600 students in the school whom we classified according to blood type, we say that we have a population of size 600.

Example - Biomedical study

- Population: set of results of new drug for all patient
- 2 possible results for each patient: new drug reduces hypertension or not which can be indicated by 1 (success) and 0 (failure)
- A result can be considered as an observed value of Bernouilli distribution $Ber(p)$
- p: probability that new drug reduces hypertension of a patient

Observation vs RV

Each *observation* in a population is *a value of a random variable* with distribution $f(x)$

- *Binomial population, normal population or population* $f(x)$: value of observations has binomial distribution, normal distribution ...
- Mean, variance of RV are referred to mean, variance of population

Why

- To determine the average length of life of a certain brand of light bulb
- **Impossible** or impractical to **test** all such bulbs - population
- we must depend on a **subset of observations from the population** (all bulbs of the brand) to help us make inferences.

Sample

Sample is a subset of population

Why

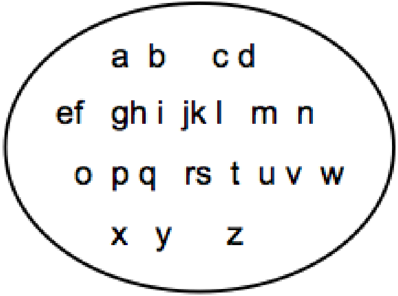
- To determine the average length of life of a certain brand of light bulb
- **Impossible** or impractical to **test** all such bulbs - population
- we must depend on a **subset of observations from the population** (all bulbs of the brand) to help us make inferences.

Sample

Sample is a subset of population

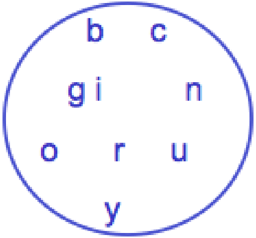
Parameters vs Statistics

Population



Values calculated using population data are called parameters

Sample



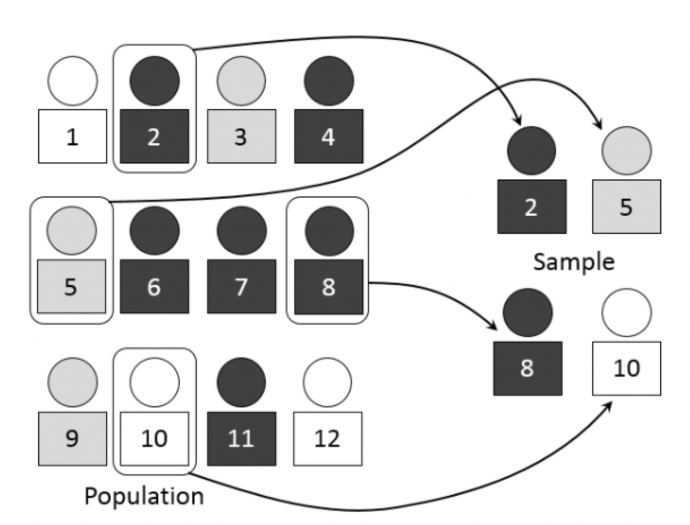
Values computed from sample data are called statistics

- Statistics are the values can be calculated from Data
- Assuming that Data are samples from a random quantity
- Statistics gives us properties of the random quantity
- Can use Statistics to predict value or compare different data

Random Sample

- Sample must be representative of the population $\Longrightarrow$ valid inferences from the sample to the population
- **Random sample**: observations when select randomly n individual from the population
- *Random sample of size n* consists of n r.v $X_1, X_2, \ldots, X_n$ independent and identically distributed $(i.i.d)$.

Example



- Select with replacement randomly two individuals from the population to get a random sample X_1, X_2
- 2, 5 are observed values of X_1, X_2
- 8, 10 are observed values of X_1, X_2
- both X_1 and X_2 can take any value from 1 to 12

$$P(X_i = k) = \frac{1}{12}, \quad \text{for } i = 1, 2 \text{ and } k = 1, \ldots 12$$

- X_1, X_2 have the same distribution as population

Statistics

- Use random samples to elicit information about the unknown population parameters
- **Statistics** is a function of the random variables constituting a random sample

- Use random samples to elicit information about the unknown population parameters
- **Statistics** is a function of the random variables constituting a random sample

Descriptive Statistics

Numerical Summaries of Data

- Measure of central tendency
  - Sample mean
  - Sample mode
  - Sample median
- Measure of dispersion of sample data
  - Sample variance
  - Sample standard deviation

Sample mean

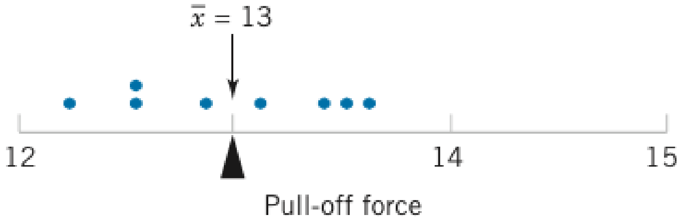
- Random sample $X_1, X_2, \ldots, X_n$
- Data: $x_1, x_2, \ldots, x_n$ (observed value of random sample)
- n is sample size
- Sample mean:
    - Statistics $\bar{X} = \frac{X_1 + \ldots + X_n}{n}$
    - Observed value $\bar{x} = \frac{x_1 + \ldots + x_n}{n}$

Example

Let's consider the eight observations on pull-off force collected from the prototype engine connectors. The eight observations are $x_1 = 12.6$, $x_2 = 12.9$, $x_3 = 13.4$, $x_4 = 12.3$, $x_5 = 13.6$, $x_6 = 13.5$, $x_7 = 12.6$, and $x_8 = 13.1$. The sample mean is

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_8}{8} = 13.0$$

pounds


$\bar{x}$ = 13
12
14
15
Pull-off force

Weighted mean

- Data in frequency table

x	x_1	x_2	$\ldots$	x_k	sample size
f	f_1	f_2	$\ldots$	f_k	$n = \sum f_i$

- $\bar{x} = \dfrac{1}{n} \displaystyle\sum_{i=1}^{k} x_i f_i$

Sample mode

- Mode = number has largest frequency
- Mode = most

Example

Suppose a data set consists of the following observations:
0.32, 0.53, 0.28, 0.37, 0.47, 0.43, 0.36, 0.42, 0.38, 0.43.
The sample mode is 0.43, since this value occurs more than any other value.

Sample median

The median is the value which divides the observations into two equal parts such that at least 50% of the values are greater than or equal to the median and at least 50% of the values are less than or equal to the median.

Denote: Med or $q_{0.5}$

Find sample median

- Sorted data: $x_1 \leq x_2 \leq \cdots \leq x_n$
- If n is odd, median is $x_{(n+1)/2}$
- If n is even, median is $\frac{1}{2}(x_{n/2} + x_{n/2+1})$
- Median is the middle number

Example

Suppose the data set is the following: 1.7, 2.2, 3.9, 3.11, and 14.7. The sample median is 3.9.

Sample variance - Sample standard deviation

Sample variance

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

Sample standard deviation, $S = \sqrt{S^2}$.

$n-1$: *degree of freedom associated with the variance* estimate since $\sum(x_i - \bar{x}) = 0$ and then the last $x - \bar{x}$ is determined by the $n-1$ intial of them, called by "pieces of information" that produces variance

Computation of observed sample variance

i	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	12.6	−0.4	0.16
2	12.9	−0.1	0.01
3	13.4	0.4	0.16
4	12.3	−0.7	0.49
5	13.6	0.6	0.36
6	13.5	0.5	0.25
7	12.6	−0.4	0.16
8	13.1	0.1	0.01
	104.0	0.0	1.60

$$s^2 = \frac{1.60}{8-1} = 0.2286\,(pounds^2),\, s = \sqrt{s^2} = \sqrt{0.2286} = 0.48\,p$$

How data is distributed

Table 6-2 Compressive Strength (in psi) of 80 Aluminum-Lithium Alloy Specimens

105	221	183	186	121	181	180	143
97	154	153	174	120	168	167	141
245	228	174	199	181	158	176	110
163	131	154	115	160	208	158	133
207	180	190	193	194	133	156	123
134	178	76	167	184	135	229	146
218	157	101	171	165	172	158	169
199	151	142	163	145	171	148	158
160	175	149	87	160	237	150	135
196	201	200	176	150	170	118	149

Frequency distribution

Class	$70 \le x < 90$	$90 \le x < 110$	$110 \le x < 130$	$130 \le x < 150$	$150 \le x < 170$	$170 \le x < 190$	$190 \le x < 210$	$210 \le x < 230$	$230 \le x < 250$
Frequency	2	3	6	14	22	17	10	4	2
Relative frequency	0.0250	0.0375	0.0750	0.1750	0.2750	0.2125	0.1250	0.0500	0.0250
Cumulative relative frequency	0.0250	0.0625	0.1375	0.3125	0.5875	0.8000	0.9250	0.9750	1.0000

Construct frequency distribution

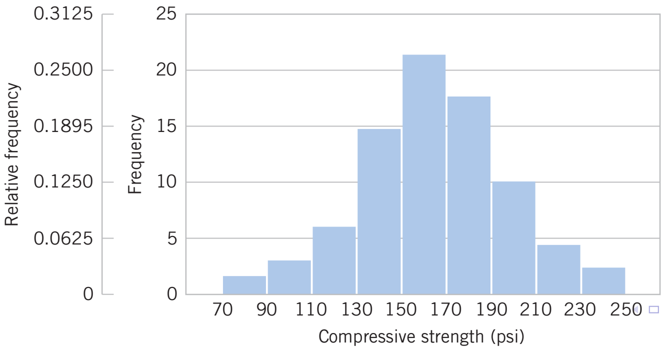
- divide the range of the data into intervals, called **class intervals, cells, or bins**
- Relative frequency $= \dfrac{\text{observed frequency in each bin}}{\text{total number of observation}}$ *empirical probability*
- Cumulative relative frequency - *empirical distribution*

Construct frequency distribution

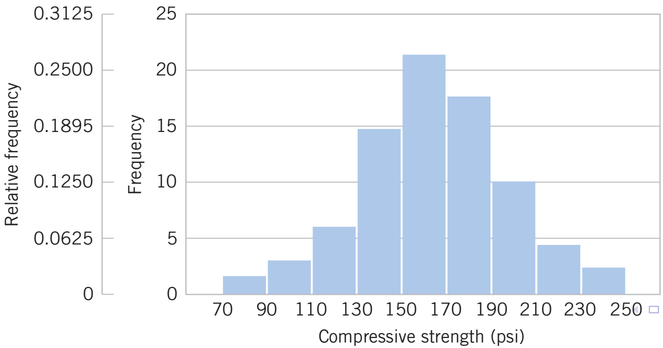
- divide the range of the data into intervals, called **class intervals, cells, or bins**
- Relative frequency $= \dfrac{\text{observed frequency in each bin}}{\text{total number of observation}}$ *empirical probability*
- Cumulative relative frequency - *empirical distribution*

Histogram

a reasonably reliable indicator of the general **shape** of the distribution

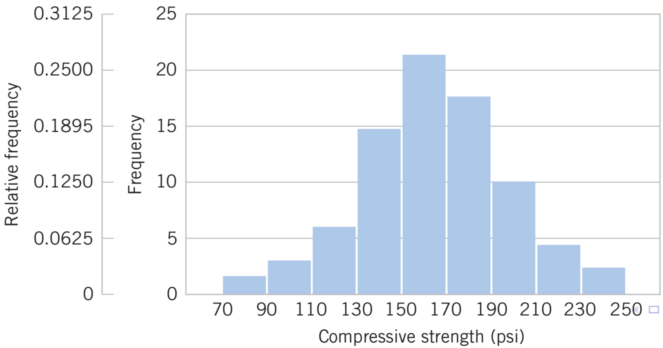


a reasonably reliable indicator of the general **shape** of the distribution



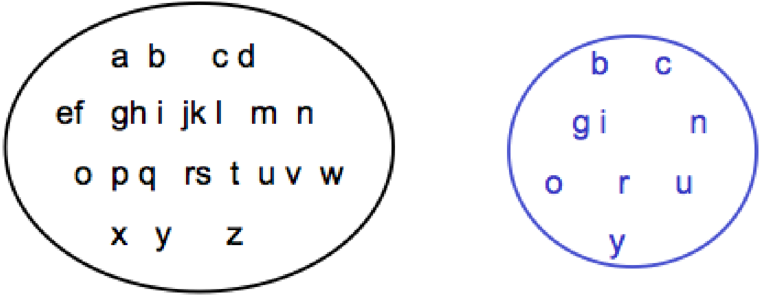
Histogram

a reasonably reliable indicator of the general **shape** of the distribution



Summary

Population **Sample**



Values calculated using population data are called parameters

Values computed from sample data are called statistics

Important statistics

- Sample mean

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n}$$

- Sample median
- Sample mode
- Sample variance

$$S^2 = \frac{(X_1 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2}{n - 1}$$

Histogram

- How data is distributed: which distribution
- Verify assumption about distribution of data

Keywords

- Population
- Sample
- Statistics: function of observed data
  - Sample mean, mode, median
  - Sample variance, sample standard deviation