Lec 1. Introduction

Zhang W.P.

Contents

1.1 Brief history of statistics 1
1.2 Statistical terms 8

Learning objectives

After this section, you should be able to ...

- Understand the meaning of "statistics"
- Understand population, sample and statistical model
- Understand the commonly-used statistics
- Distinguish between population distribution, sampling distribution, and the distribution of sample data

1.1 Brief history of statistics

Origin of the word "Statistics"

- Derived from Latin *statisticum collegium* ("council of state")
- Italian word *statista*("statesman" or "politician")
- German book *Statistik*, published in 1749, described the analysis of demographic and economic data about the state (political arithmetic in English)
- Was broadened in 1800s to include the collection, summary, and analysis of data of any type; also was conjoined with probability for the purpose of statistical inference.
- Recently, some believe this word is holding back our field

Probability

- By 1700 the mathematics of many games of chance was well understood; major contributors were Fermat, Pascal, Huygens, Leibniz.

- But this early work did not consider the inferential problem: How, from one or several outcomes of the game, could one learn about the properties of the game and how could one quantify the uncertainty of the inferred knowledge of these properties?

- It was Jacob Bernoulli (1654-1705) who began down this road. His weak law of large numbers was published in 1713 (posthumously), and he put a lower bound on the probability that X/N (the proportion of successes in N repeated trials) was within a specified distance from the true probability of success.

- De Moivre (1667-1754) refined Bernoulli's bound and stumbled upon a discrete version of the normal curve as an approximation to binomial probabilities

Error distributions

- In 1755 Thomas Simpson introduced the notion of error curves, including what are now called probability density functions. His pdf of choice was triangular.

- Others proposed error curves of different shapes: semicircular (Lambert 1765), exponential (Laplace 1774), parabolic (Lagrange 1776), normal (Laplace 1778 , although it was not called the normal distribution until 1873)

- In 1809, Gauss showed that the (sample) mean maximizes the likelihood of the errors only when the error curve is of the form

$$\phi(\Delta) = \frac{h}{\sqrt{\pi}} e^{-h^2 \Delta^2}$$

  for some positive constant h

- In early 1810, Laplace came out with the Central Limit Theorem, which showed that the aggregate of a large number of errors will be approximately normally distributed.

- Later in 1810 , Laplace read Gauss's work and made the revolutionary connection between the central limit theorem and least squares estimation : the error curve that had the most compelling rationale was also the one that led to the simplest statistical analysis!
- The Gauss-Laplace synthesis is regarded as one of the major milestones in the history of science, and it became a staple in astronomy and other physical sciences by the mid-1800s. It took until about 1900 before it had fully diffused into other scientific disciplines
- Francis Galton (1822-1911) - fitted normal curves to data, discovered reversion (later called regression) to the mean, and correlation
- Francis Edgeworth (1845-1926)- the first to compare means of two populations, using a precursor to the two-sample t test
- Karl Pearson (1857-1936) - Introduced moments, Pearson's correlation coefficient, P-value, Pearson's chi-square test, principal component analysis, among many other things
- Ronald A. Fisher (1890-1962)- Introduced randomization test, named and promoted the analysis of variance and the design of experiments

Qualities of a Good STATISTICIAN

- **S**cientific
- **T**alented
- **A**ctive
- **T**enacious
- **I**nnovative
- **S**killful
- **T**errific
- **I**nventive
- **C**reative
- **I**nterpretative
- **A**ccurate
- **N**oble





Statistics Definition

Statistic is the art and science of collecting, analyzing, presenting and interpreting data.

Def.

This definition includes four terms:

- Collection of data
- Presentation of data
- Analysis of data and
- Interpretation of data

Major branch of Statistics

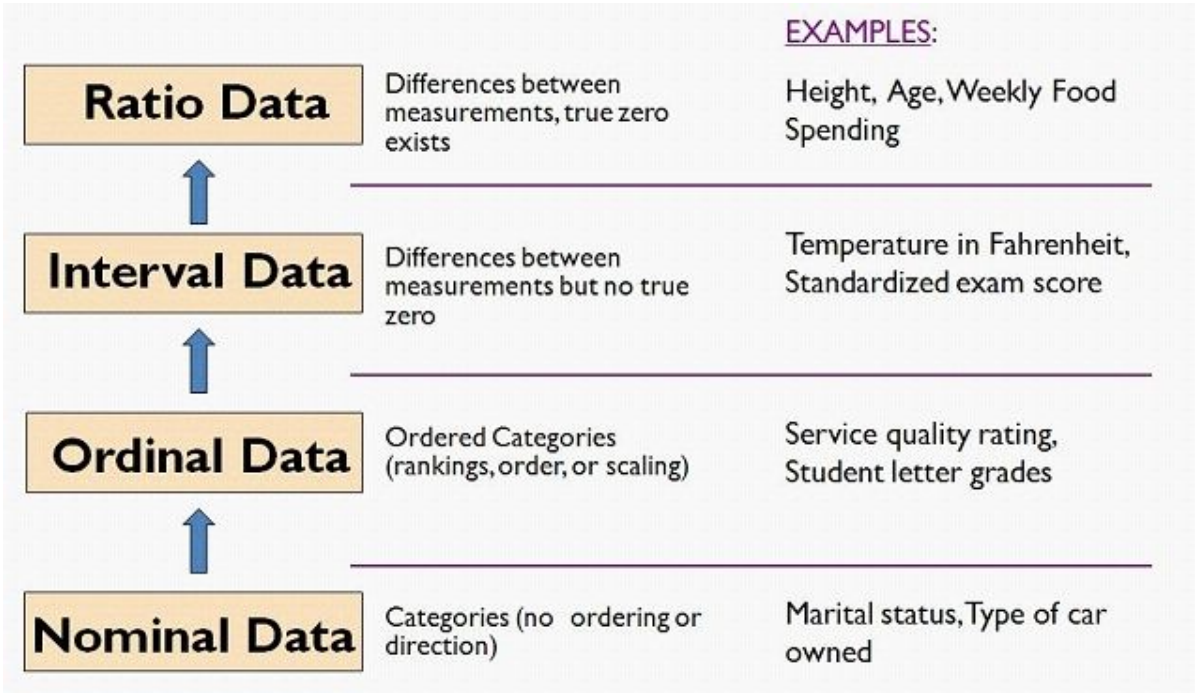
- Descriptive statistics— a statistical method concerned with describing the properties and characteristics of a set of data.
  - measures of central tendency: mode, mean, median
  - measures of variability (spread): standard deviation, variance, minimum and maximum variables, and kurtosis and skewness.
- Inferential statistics— concerned with the analysis of a sample data leading to a prediction, inferences or conclusion about the entire population.
  - Estimation
  - Hypothesis tests.

1.2 Statistical terms

Data

— any quantitative or qualitative information.

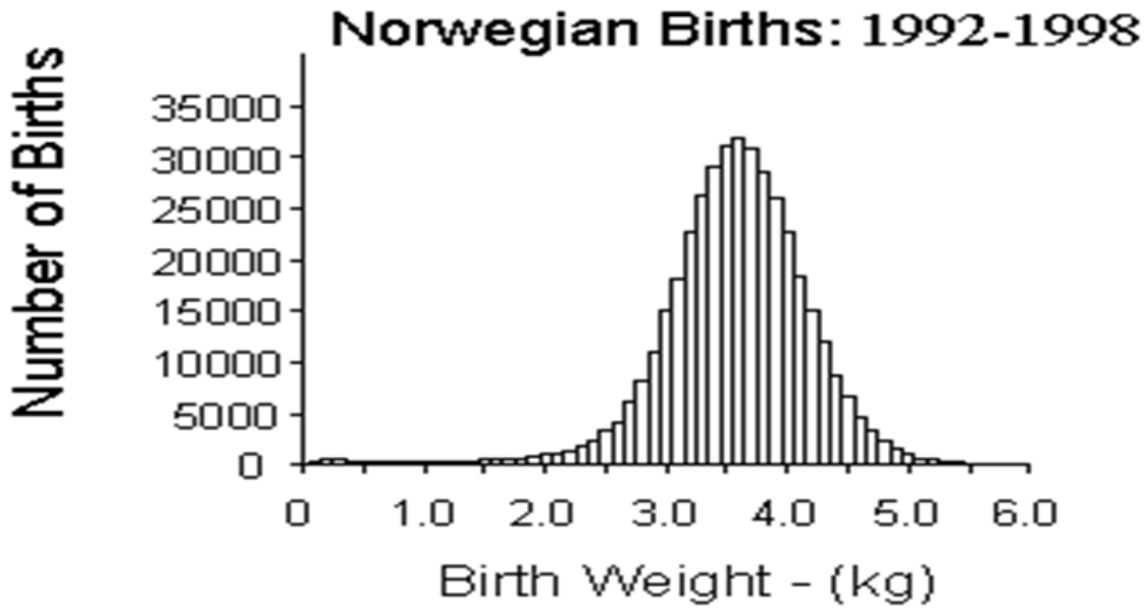
Level of data



Population

总体 — the totality of all elements or persons for which one has interest at a particular time.

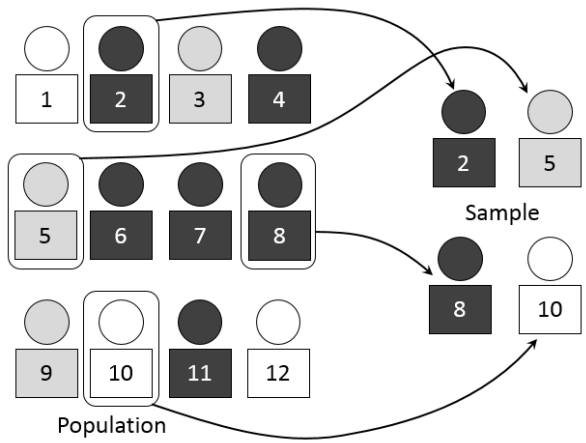
— is the collection of all possible observations of a random variable X.



population distribution(总体分布): $X \sim F$, e.g., normal population, binomial population, poisson population, etc.

Sample

样本 — is a set of individuals or objects collected or selected from a statistical population by a defined procedure



- sample $\mathbf{X} = (X_1, \ldots, X_n)'$, n is sample size
- (simple) random sample(简单随机样本): $X_1, \ldots, X_n, i.i.d \sim X$
- sample distribution(样本分布): $\mathbf{X} \sim f(x_1, \ldots, x_n)$

Statistical model

—A statistical model is a mathematical model that embodies a set of statistical assumptions concerning the generation of sample data. A statistical model represents, often in considerably idealized form, the data-generating process.

- Sample $\mathbf{X} = (X_1, X_2, \ldots, X_n)'$
- Probability distribution of $\mathbf{X}$
- Statistical Model $\mathcal{P} = \{$ distribution family of the sample $\}$
- **Parametric statistical model:** Parameter θ identifies/specifies distribution in $\mathcal{P}$.
  — $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$, $\Theta = \{\theta\}$ is the Parameter Space
- **Non-Parametric statistical model:** $\mathcal{P}$ cannot be defined in terms of such a finite set of parameters.

Statistic

—A statistic (统计量) (singular) or sample statistic is any quantity computed from values in a sample that is used for a statistical purpose. Statistical purposes include estimating a population parameter, describing a sample, or evaluating a hypothesis.

- sample mean (样本均值)
- sample variance/deviation (样本方差/偏差)
- sample median/sample quantile/order statistics (样本中位数/分位数/次序统计量)
- sample moments (样本矩)
- empirical distribution function (ecdf) (经验分布函数)
- sample skewness/kurtosis (样本偏度/峰度)

Convergence of Random Variables:

- Convergence in Probability: $\{Z_n\}$ converges in probability to Z.

$$|Z_n - Z| \xrightarrow{P} 0, \quad or \quad Z_n \xrightarrow{P} Z$$

Definition: For every $\epsilon > 0 : \lim_{n\to\infty} P\left(|Z_n - Z| > \epsilon\right) \to 0$

- Convergence in Distribution: $\{Z_n\}$ converges in law to Z.

$$Z_n \xrightarrow{\mathcal{L}} Z$$

Definition: for every $\mathbf{t} \in R^d$, where the distribution function F_Z of Z is continuous:

$$\lim_{n\to\infty} F_{z_n}(t) = F_z(t)$$

- Almost sure convergence: X_n converges almost surely or almost everywhere or with probability 1.

Definition: $P\left(\lim_{n\to\infty} X_n = X\right) = 1$

- (Continuous mapping theorem) If $Z_n \xrightarrow{P} z_0$(constant), and g is continuous at z_0, then

$$g\left(Z_n\right) \xrightarrow{P} g\left(z_0\right)$$

- (Continuous mapping theorem)If $Z_n \xrightarrow{\mathcal{L}} Z$, and g is continuous, then

$$g\left(Z_n\right) \xrightarrow{\mathcal{L}} g(Z)$$

- (Slutsky's Theorem) If $Z_n \xrightarrow{\mathcal{L}} Z$, and $U_n \xrightarrow{P} u_0$, a constant, then

$$\text{(a). } Z_n + U_n \xrightarrow{\mathcal{L}} Z + u_0$$
$$\text{(b). } U_n Z_n \xrightarrow{\mathcal{L}} u_0 Z$$

Theorem 1. *Let $X_i, i = 1, \ldots, n$ be an i.i.d. sequence of random variables with distribution function F on $\mathbf{R}$. The empirical distribution function is the function of x defined by*

$$F_n(x) = \frac{1}{n} \sum_{1 \le i \le n} I\{X_i \le x\}$$

has the following asymptotic properties:

(1) For any fixed $x \in \mathbf{R}$, $nF_n(x) \sim B(n, F(x))$ and

$$\frac{\sqrt{n}(F_n(x) - F(x))}{\sqrt{F(x)(1 - F(x))}} \xrightarrow{\mathcal{L}} N(0, 1)$$

(2) Glivenko-Cantelli Theorem (fundamental theorem of statistics):

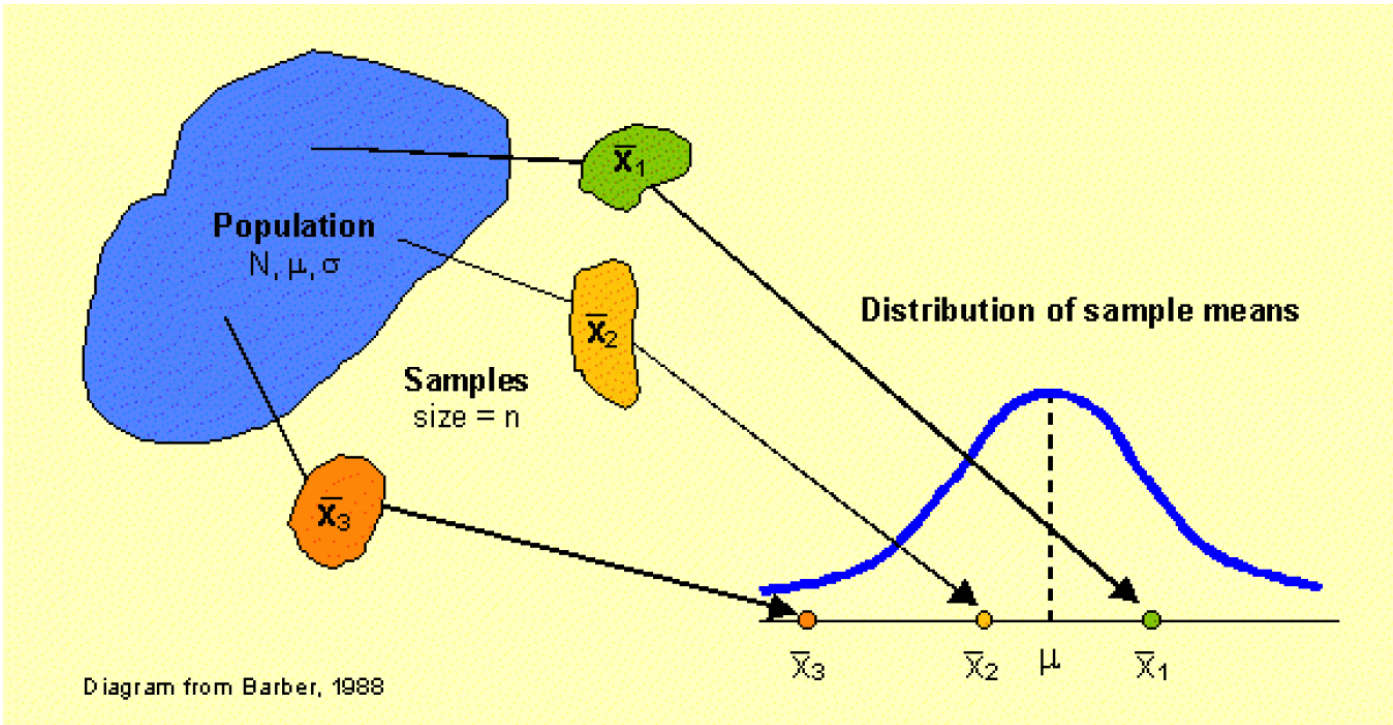
$$\sup_{x \in \mathbf{R}} |F_n(x) - F(x)| \to 0 \ a.s.$$

(3) Dvoretzky-Kiefer-Wolfowith(DKW) Inequality: for any $\epsilon > 0$, and n,

$$P\left(\sup_{x \in R} |F_n(x) - F(x)| > \epsilon\right) \le 2e^{-2nc^2}$$

Sampling distribution

— The sampling distribution (抽样分布) of a statistic is the distribution of values taken by the statistic in all possible samples of the same size from the same population.



Population Distributions vs. Sampling Distributions

There are actually three distinct distributions involved when we sample repeatedly and measure a variable of interest.

- The population distribution gives the values of the variable for all the individuals in the population.
- The distribution of sample data shows the values of the variable for all the individuals in the sample.
- The sampling distribution shows the statistic values from all the possible samples of the same size from the population.