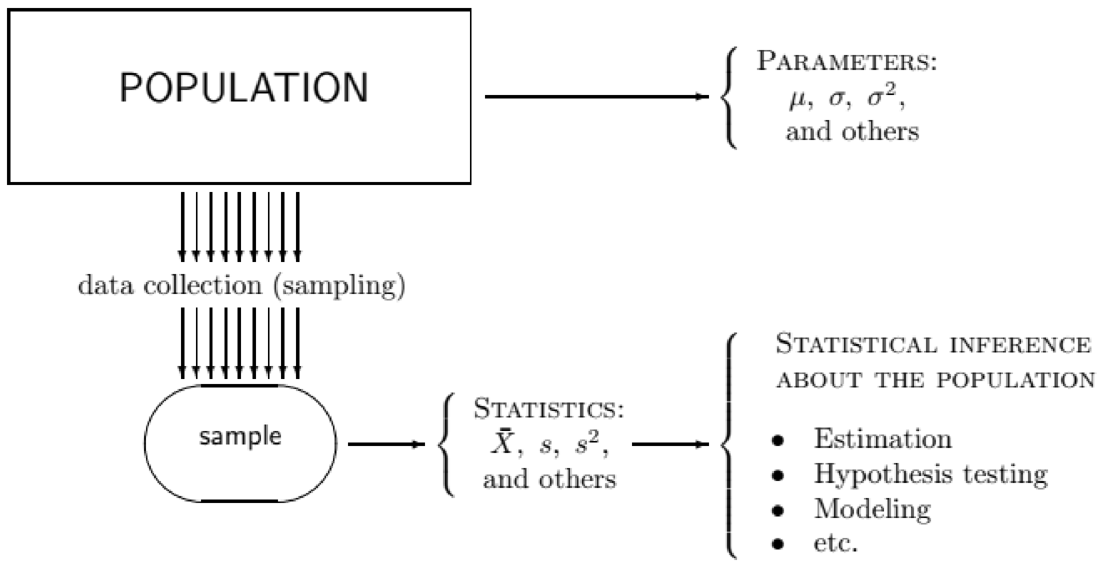
Statistics

DEFINITION 8.1

A **population** consists of all units of interest. Any numerical characteristic of a population is a **parameter**. A **sample** consists of observed units collected from the population. It is used to make statements about the population. Any function of a sample is called **statistic**.



Sample Statistics (Used as estimators for Population Parameters)

Mean

DEFINITION 8.3

Sample mean $\bar{X}$ is the arithmetic average,

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

This is an unbiased, consistent estimator of Population Average.

DEFINITION 8.4

An estimator $\hat{\theta}$ is **unbiased** for a parameter θ if its expectation equals the parameter,

$$\mathbf{E}(\hat{\theta}) = \theta$$

for all possible values of θ.

Bias of $\hat{\theta}$ is defined as $\text{Bias}(\hat{\theta}) = \mathbf{E}(\hat{\theta} - \theta)$.

DEFINITION 8.5

An estimator $\hat{\theta}$ is **consistent** for a parameter θ if the probability of its sampling error of any magnitude converges to 0 as the sample size increases to infinity. Stating it rigorously,

$$\boldsymbol{P}\left\{|\hat{\theta}-\theta|>\varepsilon\right\} \rightarrow 0 \text{ as } n \rightarrow \infty$$

for any $\varepsilon > 0$. That is, when we estimate θ from a large sample, the estimation error $|\hat{\theta}-\theta|$ is unlikely to exceed ε, and it does it with smaller and smaller probabilities as we increase the sample size.

Asymptotic Normality of Sample Mean. By Central Limit Theorem, Sample mean follows a normal distribution as number of samples increases

i.e.

$$Z = \frac{\bar{X} - \mathbf{E}\bar{X}}{\mathrm{Std}\bar{X}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$ converges to a Normal(0,1) as n → ∞

Variance and Std. Deviation

DEFINITION 8.8

For a sample $(X_1, X_2, \ldots, X_n)$, a **sample variance** is defined as

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2. \qquad (8.4)$$

It measures variability among observations and estimates the population variance $\sigma^2 = \mathrm{Var}(X)$.

Sample standard deviation is a square root of a sample variance,

$$s = \sqrt{s^2}.$$

It measures variability in the same units as X and estimates the population standard deviation $\sigma = \mathrm{Std}(X)$.

Alt formula for Variance:

$$\boxed{s^2 = \frac{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}{n-1}.}$$

Median

DEFINITION 8.6

Median means a "central" value.

Sample median $\hat{M}$ is a number that is exceeded by at most a half of observations and is preceded by at most a half of observations.

Population median M is a number that is exceeded with probability no greater than 0.5 and is preceded with probability no greater than 0.5. That is, M is such that

$$\begin{cases} \boldsymbol{P}\{X > M\} & \leq & 0.5 \\ \boldsymbol{P}\{X < M\} & \leq & 0.5 \end{cases}$$

Sample median

If n is odd, the $\left(\frac{n+1}{2}\right)$-th smallest observation is a median.
If n is even, any number between the $\left(\frac{n}{2}\right)$-th smallest and the $\left(\frac{n+2}{2}\right)$-th smallest observations is a median.

Shape of a distribution (comparing mean and median)

$$\begin{array}{lcl} \text{Symmetric distribution} & \Rightarrow & M = \mu \\ \text{Right-skewed distribution} & \Rightarrow & M < \mu \\ \text{Left-skewed distribution} & \Rightarrow & M > \mu \end{array}$$

Quantiles, percentiles and quartiles

DEFINITION 8.7

A p-**quantile** of a population is such a number x that solves equations

$$\begin{cases} \boldsymbol{P}\{X < x\} & \leq & p \\ \boldsymbol{P}\{X > x\} & \leq & 1-p \end{cases}$$

A **sample** p-**quantile** is any number that exceeds at most $100p\%$ of the sample, and is exceeded by at most $100(1-p)\%$ of the sample.

A γ-**percentile** is (0.01γ)-quantile.

First, second, and third **quartiles** are the 25th, 50th, and 75th percentiles. They split a population or a sample into four equal parts.

A **median** is at the same time a 0.5-quantile, 50th percentile, and 2nd quartile.

IQR and outliers.

DEFINITION 8.10

> An **interquartile range** is defined as the difference between the first and the third quartiles,
>
> $$IQR = Q_3 - Q_1.$$
>
> It measures variability of data. Not much affected by outliers, it is often used to detect them. IQR is estimated by the *sample interquartile range*
>
> $$\widehat{IQR} = \hat{Q}_3 - \hat{Q}_1.$$

Any samples that are less than $Q_1 - 1.5(IQR)$ or more than $Q_3 + 1.5(IQR)$ can be treated as potential outliers.

Standard error of any estimator is its std deviation.

$$\begin{array}{rcl} \sigma(\hat{\theta}) & = & \text{standard error of estimator } \hat{\theta} \text{ of parameter } \theta \\ s(\hat{\theta}) & = & \text{estimated standard error } = \hat{\sigma}(\hat{\theta}) \end{array}$$

Graphical Statistics:

Can be used to visualize the given samples to make observations about the nature of the population

- Histograms are bar charts for columns for each bin
  - If height of bin is freq count: Frequency histogram
  - If height of bin is proportion of data: Relative Frequency histogram
  - Each sample value can have its own bin, or you can have multiple nearby values in one bin.
  - You can use histograms to guess shape of distribution.
- Stem and leaf plots
  - Choose the stem such that the values are not all limited to one stem.
  - Distribute values to each stem and sort he leaf values.
  - You can use this to calculate mean, median and guess shape of distribution
  - It can also be used to compare two distributions
- Boxplots
  - Boxplots are based on 5-point summaries
    - $< \min(X_i), \widehat{Q_1}, \widehat{M}, \widehat{Q_3}, \max(X_i) >$
  - Represent sample mean with small cross. Draw a box between sample Q1 and Q3 and draw a line for sample median. Draw whiskers to smallest sample and largest sample that is within the 1.5 IQR range. Draw dots for all samples outside the 1.5 IQR range.
  - Can be used to compare multiple distributions.