TDAB01 Probability and Statistics

Maryna Prus
IDA, Linköping University

Lecture 7: Introduction to Statistics

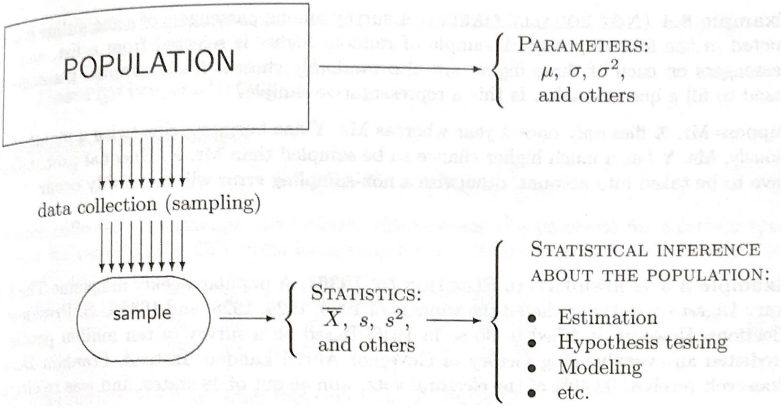
Overview

- **Population and sample, parameters and statistics**
- **Descriptive statistics**
- **Introduction to parameter estimation and sampling distributions**
- **Graphical methods**

Basic concepts

- **Population** = **all** units of interest
  - Sweden's population
  - All units produced at a factory
- **Parameter** = numerical characteristic of **population**
  - Average income (μ) or income dispersion (σ^2)
  - Proportion of broken products (p)
- **Sample** = observed units collected from **population**
  - 1000 randomly selected people
  - 10 selected boxes of products
- **Statistic** = function of **sample**
  - Sample mean $\bar{X}$, sample variance s^2, proportion of defect products $\hat{p}$
- **Simple random sampling** - units are chosen **independently** of each other, **equally likely** to be sampled

Probability theory and statistical inference



Estimator

- **Population parameter:** θ, unknown

  Inference: Learning about θ from data (sample)
- $\hat{\theta}$ - **estimator** of θ, function of sample

  For a given sample $X_1, \ldots, X_n$, we get an **estimate** (a value) of $\hat{\theta}$ representing our **best "guess"** of θ based on information in this sample
- Example: $\theta = p$, success probability for Bernoulli trials

$$\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n} = \text{proportion of success}$$

- $\hat{p}$ is **correct on average** over all possible samples of size n

$$\mathbb{E}(\hat{p}) = \mathbb{E}\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{\sum_{i=1}^{n} \mathbb{E}(X_i)}{n} = \frac{\sum_{i=1}^{n} p}{n} = \frac{np}{n} = p$$

- Estimator $\hat{\theta}$ of θ is **unbiased** if

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

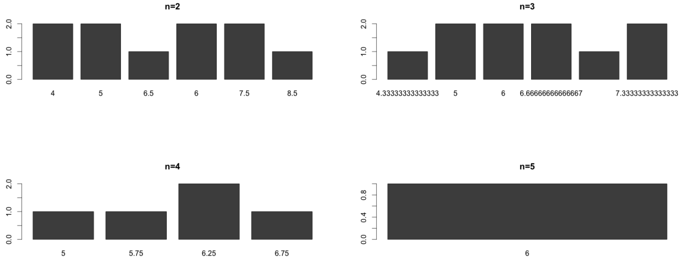
- **Bias**:

$$Bias(\hat{\theta}) = \mathbb{E}\left(\hat{\theta}\right) - \theta$$

Sampling distribution

- **Sampling distribution** for $\hat{\theta}$

  $\rightarrow$ describes how $\hat{\theta}$ can vary from sample to sample
- Example: Population $\{3, 5, 5, 7, 10\}$, θ - mean $\rightarrow \theta = \frac{3+5+5+7+10}{5} = 6$
- Random sample of size $n = 3$:
  - Random sample 1: $\{3, 5, 5\}$ with $\bar{x} = 4.333$
  - Random sample 2: $\{3, 5, 7\}$ with $\bar{x} = 5.000$
  - $\vdots$
  - Random sample 10: $\{5, 7, 10\}$ with $\bar{x} = 7.333$
- Sampling distribution for $\bar{X}$ with $n = 2, 3, 4, 5$:



Mean

- Sample: $X_1, \ldots, X_n$ with $\mathbb{E}(X_i) = \mu$, $Var(X_i) = \sigma^2$, $i = 1, \ldots, n$
- **Sample mean** - arithmetic average

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

- Sample mean is unbiased estimator of μ: $\mathbb{E}(\hat{\mu}) = \mu$
- Simple random sampling
  $\rightarrow$ $X_1, \ldots, X_n$ independent and identically distributed (or **iid**)
- **Variance** for $\hat{\mu}$:

$$Var(\hat{\mu}) = Var\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

- **Standard deviation** for $\hat{\mu}$:

$$Std(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

Consistency

- $\hat{\theta}$ is **consistent** estimator for θ if its distribution becomes increasingly concentrated around θ as the sample size n is increasing
- Formally: Estimator $\hat{\theta}$ is consistent for θ if for all $\varepsilon > 0$

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

- For iid $X_1, \ldots, X_n$, $\bar{X}$ is consistent estimator for μ
- **Proof** via Chebyshev's inequality:

$$\boldsymbol{P}\left\{\left|\bar{X} - \mu\right| > \varepsilon\right\} \le \frac{Var(\bar{X})}{\varepsilon^2} = \frac{\sigma^2/n}{\varepsilon^2} \to 0 \text{ when } n \to \infty.$$

- From **Central Limit Theorem**:
  Distribution for $\bar{X}$ is approximately $N\left(\mu, \sigma^2/n\right)$ for large n
- Formally: Cdf of

$$Z = \frac{\bar{X} - \mathbb{E}(\bar{X})}{Std(\bar{X})} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

  converges to cdf of standard normal distribution

Normal distribution

- If $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$ independent, then

$$aX + bY \sim N(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2)$$

- If X and Y dependent, $aX + bY$ still normally distributed, but with different variance
- This result also holds for multiple variables

  Especially for $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ we obtain

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

  In this case $\bar{X}$ is not *approximately* normally distributed but **exactly** normally distributed

Median and quantiles

- Sample mean is sensitive to extreme measurement values, called **outliers**
- **Median** M is more **robust**:

$$\boldsymbol{P}\,(X < M) \le 0.5$$
$$\boldsymbol{P}(X > M) \le 0.5$$

- (Population) median = half of the probability on the left, half on the right
- **Sample median**:

$$\hat{M} = \begin{cases} \left(\frac{n+1}{2}\right)\text{-th smallest observation} & \text{if } n \text{ odd} \\ \text{the mean of } \left(\frac{n}{2}\right)\text{-th and } \left(\frac{n+2}{2}\right)\text{-th observations} & \text{if } n \text{ even} \end{cases}$$

- Generalization of median: p-**quantile** is a number c which solves

$$\boldsymbol{P}\,(X < c) \le p$$
$$\boldsymbol{P}(X > c) \le 1 - p$$

- **Percentiles**: 5%, 37%, etc. - 0.05-, 0.37, etc.-quantlites
- **Quartiles**: 25%-Q_1, 50%-Q_2, 75%-Q_3; $IQR = Q_3 - Q_1$
- R code for 0.05-quantile for $N(1,2)$: `qnorm(p=0.05,mean=1,sd =2)`

Sampling Variance

- **Sample variance**:
$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$
- s^2 is unbiased estimator for population variance σ^2, i. e. $\mathbb{E}(s^2) = \sigma^2$
- Proof: Rewrite s^2 as
$$s^2 = \frac{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}{n-1}$$
$$\Rightarrow \quad \mathbb{E}(s^2) = \frac{\sum_{i=1}^{n} \mathbb{E}(X_i^2) - n\mathbb{E}\left(\bar{X}^2\right)}{n-1}$$
From
$$Var(X_i) = \mathbb{E}(X_i^2) - \mu^2 = \sigma^2 \text{ and } Var(\bar{X}) = \mathbb{E}\left(\bar{X}^2\right) - \mathbb{E}(\bar{X})^2 = \mathbb{E}\left(\bar{X}^2\right) - \mu^2 = \frac{\sigma^2}{n}$$
$$\Rightarrow \quad \sum_{i=1}^{n} \mathbb{E}(X_i^2) - n\mathbb{E}\left(\bar{X}^2\right) = n\left(\sigma^2 + \mu^2\right) - n\left(\frac{\sigma^2}{n} + \mu^2\right) = \sigma^2(n-1)$$
- **Sample standard deviation**: $s = \sqrt{s^2}$ - estimator of σ

Graphical methods - demo

- See `SS7GraferDemo.R`.

Thank you for your attention!