Statistical and Mathematical Methods

Statistical and Mathematical Methods for Data Science
DS5003

Dr. Nasir Touheed

Statistics

Data Collection

- to visualize data, understand the patterns, and make quick statements about the system's behavior;
- to characterize this behavior in simple terms and quantities;
- to estimate the distribution parameters;
- to assess reliability of our estimates
- to test statements about parameters and the entire system;
- to understand relations among variables;
- to fit suitable models and use them to make forecasts.

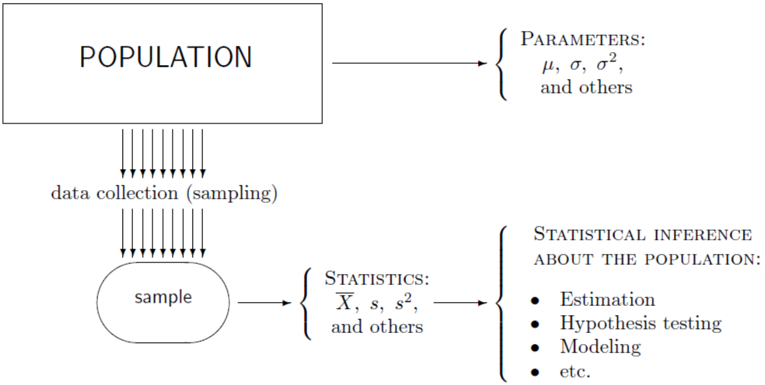
Population

- 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.
- Instead of a census, we may collect data in a form of a random sample from a population
- We can measure them, perform calculations, and estimate the unknown parameters of the population up to a certain measurable degree of accuracy.

$$\theta = populationparameter$$

$$\widehat{\theta} = itsestimator, obtainedfromasample$$

Population


POPULATION
PARAMETERS:
μ, σ, σ^2,
and others
data collection (sampling)
sample
STATISTICS:
$\overline{X}$, s, s^2,
and others
STATISTICAL INFERENCE
ABOUT THE POPULATION:
Estimation
Hypothesis testing
Modeling
etc.

Sampling and non-sampling errors

- Sampling errors are caused by the mere fact that only a sample, a portion of a population, is observed.
- For most of reasonable statistical procedures, sampling errors decrease (and converge to zero) as the sample size increases.
- Non-sampling errors are caused by inappropriate sampling schemes or wrong statistical techniques. Often no wise statistical techniques can rescue a poorly collected sample of data.
- To evaluate the work of a Windows help desk, a survey of social science students of some university is conducted. This sample poorly represents the whole population of all Windows users. For example, computer science students and especially computer professionals may have a totally different opinion about the Windows help desk.
- Simple random sampling is a sampling design where units are collected from the entire population independently of each other, all being equally likely to be sampled.

Simple descriptive statistics

- Simple descriptive statistics measuring the location, spread, variability, and other characteristics can be computed immediately.
  - mean, measuring the average value of a sample;
  - median, measuring the central value;
  - quantiles and quartiles, showing where certain portions of a sample are located;
  - variance, standard deviation, and interquartile range, measuring variability and spread of data.
- Each statistic is a random variable because it is computed from random data. It has a so-called sampling distribution.
- Each statistic estimates the corresponding population parameter and adds certain information about the distribution of X, the variable of interest.

Mean Median

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

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

Median means a “central” value.

Sample median $\widehat{M}$ 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

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

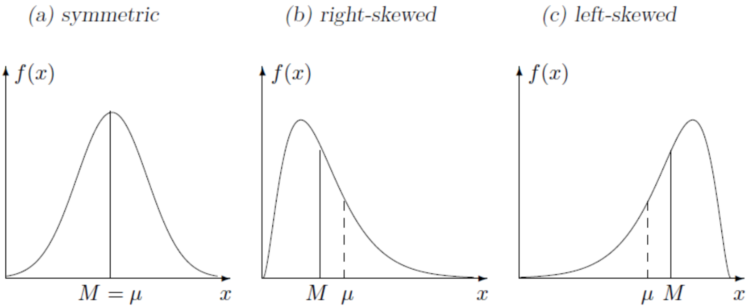
Mean

- For example, to evaluate effectiveness of a processor for a certain type of tasks, we recorded the CPU time in seconds for n $=$ 30 randomly chosen jobs (data set CPU),

70	36	43	69	82	48	34	62	35	15
59	139	46	37	42	30	55	56	36	82
38	89	54	25	35	24	22	9	56	19

- we estimate the average (expected) CPU time μ
- compute the median of n $=$ 30 CPU times from the data

Median



(a) *symmetric*

(b) *right-skewed*

(c) *left-skewed*

A mean μ and a median M for distributions of different shapes.

Percentile

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

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

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.

q_p	$=$	population p-quantile
$\widehat{q}_p$	$=$	sample p-quantile, estimator of q_p
π_γ	$=$	population γ-percentile
$\widehat{\pi}_\gamma$	$=$	sample γ-percentile, estimator of π_γ
$Q_1,\ Q_2,\ Q_3$	$=$	population quartiles
$\widehat{Q}_1,\ \widehat{Q}_2,\ \widehat{Q}_3$	$=$	sample quartiles, estimators of $Q_1,\ Q_2,$ and Q_3
M	$=$	population median
$\widehat{M}$	$=$	sample median, estimator of M

$$q_p = \pi_{100p}$$

$$Q_1 = \pi_{25} = q_{1/4} \qquad Q_3 = \pi_{75} = q_{3/4}$$

$$M = Q_2 = \pi_{50} = q_{1/2}$$

Variance and standard deviation

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

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

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 $\sigma = \mathrm{Std}(X)$.

$$s^2 = \frac{\sum\limits_{i=1}^{n} X_i^2 - n\overline{X}^2}{n-1}. \qquad s^2 = \frac{70^2 + \ldots + 19^2 - (30)(48.2333)^2}{30-1} = \frac{90,185 - 69,794}{29} = 703.1506 \text{ (sec}^2\text{)}.$$

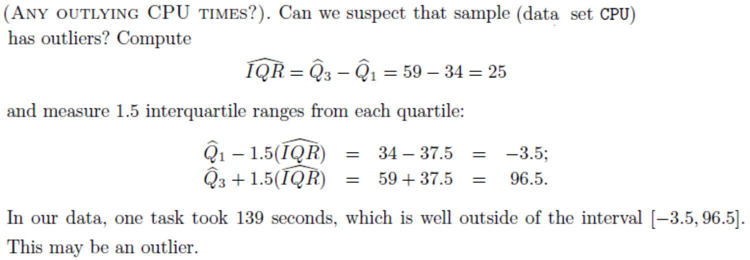
interquartile range

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*

$$\widehat{IQR} = \widehat{Q}_3 - \widehat{Q}_1.$$



(ANY OUTLYING CPU TIMES?). Can we suspect that sample (data set `CPU`) has outliers? Compute

$$\widehat{IQR} = \widehat{Q}_3 - \widehat{Q}_1 = 59 - 34 = 25$$

and measure 1.5 interquartile ranges from each quartile:

$$\begin{array}{rcccl} \widehat{Q}_1 - 1.5(\widehat{IQR}) & = & 34 - 37.5 & = & -3.5; \\ \widehat{Q}_3 + 1.5(\widehat{IQR}) & = & 59 + 37.5 & = & 96.5. \end{array}$$

In our data, one task took 139 seconds, which is well outside of the interval $[-3.5, 96.5]$. This may be an outlier.

Graphical statistics

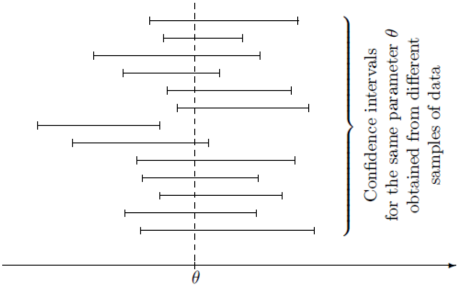
- A quick look at a sample may clearly suggest
  - a probability model, i.e., a family of distributions to be used;
  - statistical methods suitable for the given data;
  - presence or absence of outliers;
  - presence or absence of heterogeneity;
  - existence of time trends and other patterns;
  - relation between two or several variables.
- There is a number of simple and advanced ways to visualize data.
  - histograms,
  - stem-and-leaf plots,
  - boxplots,
  - time plot
  - scatter plots

Confidence Interval

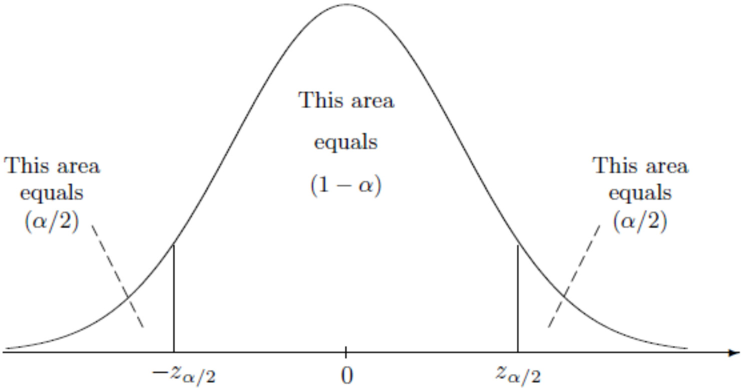
An interval $[a, b]$ is a $(1-\alpha)100\%$ **confidence interval** for the parameter θ if it contains the parameter with probability $(1-\alpha)$,

$$\boldsymbol{P}\{a \leq \theta \leq b\} = 1-\alpha.$$

The **coverage probability** $(1-\alpha)$ is also called a **confidence level.**

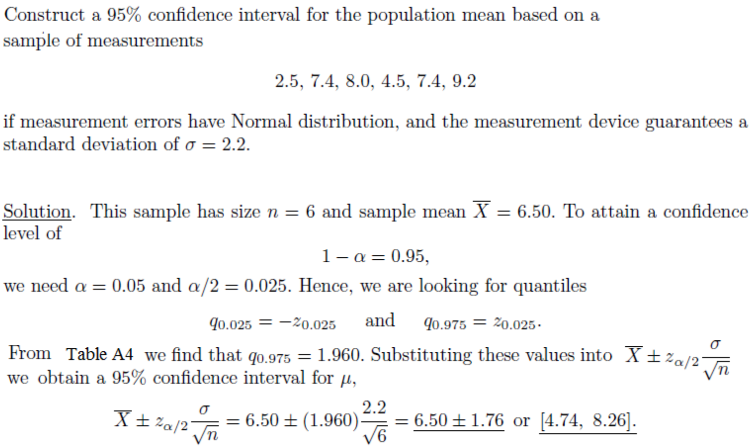


Construction of confidence intervals



$$P\{a \leq \theta \leq| b\} = 1 - \alpha \qquad Z = \frac{\widehat{\theta} - \mathrm{E}(\widehat{\theta})}{\sigma(\widehat{\theta})} = \frac{\widehat{\theta} - \theta}{\sigma(\widehat{\theta})},$$

Construction of confidence intervals



Construct a 95% confidence interval for the population mean based on a sample of measurements

$$2.5,\ 7.4,\ 8.0,\ 4.5,\ 7.4,\ 9.2$$

if measurement errors have Normal distribution, and the measurement device guarantees a standard deviation of $\sigma = 2.2$.

Solution. This sample has size $n = 6$ and sample mean $\overline{X} = 6.50$. To attain a confidence level of

$$1 - \alpha = 0.95,$$

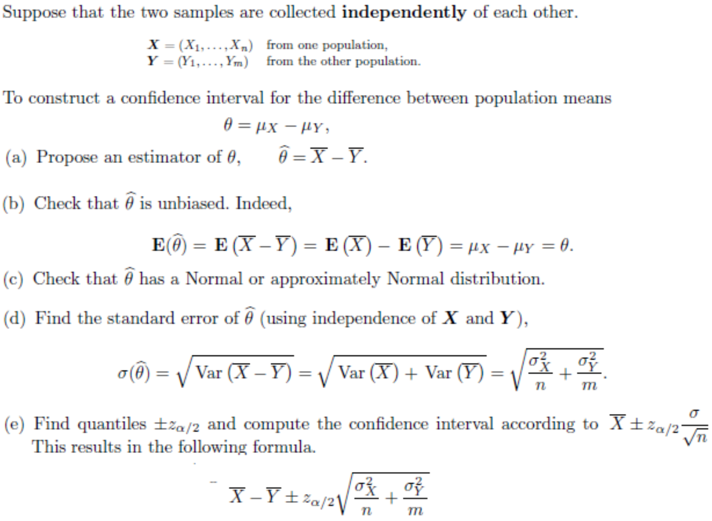
we need $\alpha = 0.05$ and $\alpha/2 = 0.025$. Hence, we are looking for quantiles

$$q_{0.025} = -z_{0.025} \quad \text{and} \quad q_{0.975} = z_{0.025}.$$

From Table A4 we find that $q_{0.975} = 1.960$. Substituting these values into $\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ we obtain a 95% confidence interval for μ,

$$\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 6.50 \pm (1.960)\frac{2.2}{\sqrt{6}} = \underline{6.50 \pm 1.76} \text{ or } \underline{[4.74,\ 8.26]}.$$

Confidence interval for the difference between two means



Suppose that the two samples are collected **independently** of each other.

$$\begin{array}{ll} \boldsymbol{X} = (X_1, \ldots, X_n) & \text{from one population,} \\ \boldsymbol{Y} = (Y_1, \ldots, Y_m) & \text{from the other population.} \end{array}$$

To construct a confidence interval for the difference between population means

$$\theta = \mu_X - \mu_Y,$$

(a) Propose an estimator of θ, $\quad \widehat{\theta} = \overline{X} - \overline{Y}$.

(b) Check that $\widehat{\theta}$ is unbiased. Indeed,

$$\mathbf{E}(\widehat{\theta}) = \mathbf{E}\left(\overline{X} - \overline{Y}\right) = \mathbf{E}\left(\overline{X}\right) - \mathbf{E}\left(\overline{Y}\right) = \mu_X - \mu_Y = \theta.$$

(c) Check that $\widehat{\theta}$ has a Normal or approximately Normal distribution.

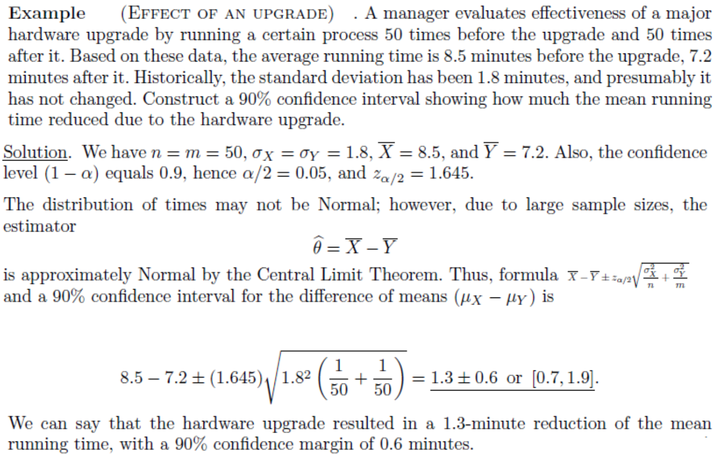
(d) Find the standard error of $\widehat{\theta}$ (using independence of $\boldsymbol{X}$ and $\boldsymbol{Y}$),

$$\sigma(\widehat{\theta}) = \sqrt{\operatorname{Var}\left(\overline{X} - \overline{Y}\right)} = \sqrt{\operatorname{Var}\left(\overline{X}\right) + \operatorname{Var}\left(\overline{Y}\right)} = \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}.$$

(e) Find quantiles $\pm z_{\alpha/2}$ and compute the confidence interval according to $\overline{X} \pm z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$
This results in the following formula.

$$\overline{X} - \overline{Y} \pm z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}$$

Confidence interval for the difference between two means



Example (EFFECT OF AN UPGRADE) . A manager evaluates effectiveness of a major hardware upgrade by running a certain process 50 times before the upgrade and 50 times after it. Based on these data, the average running time is 8.5 minutes before the upgrade, 7.2 minutes after it. Historically, the standard deviation has been 1.8 minutes, and presumably it has not changed. Construct a 90% confidence interval showing how much the mean running time reduced due to the hardware upgrade.

Solution. We have $n = m = 50$, $\sigma_X = \sigma_Y = 1.8$, $\overline{X} = 8.5$, and $\overline{Y} = 7.2$. Also, the confidence level $(1-\alpha)$ equals 0.9, hence $\alpha/2 = 0.05$, and $z_{\alpha/2} = 1.645$.

The distribution of times may not be Normal; however, due to large sample sizes, the estimator

$$\widehat{\theta} = \overline{X} - \overline{Y}$$

is approximately Normal by the Central Limit Theorem. Thus, formula $\overline{X} - \overline{Y} \pm z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}$ and a 90% confidence interval for the difference of means $(\mu_X - \mu_Y)$ is

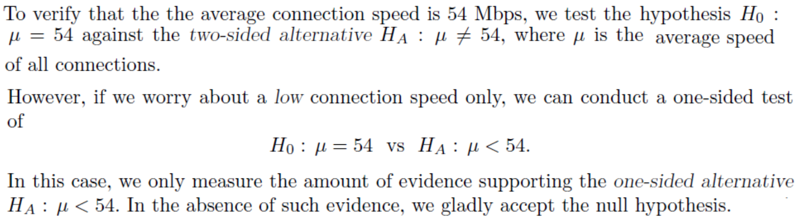
$$8.5 - 7.2 \pm (1.645)\sqrt{1.8^2\left(\frac{1}{50} + \frac{1}{50}\right)} = \underline{1.3 \pm 0.6 \ \text{ or } \ [0.7, 1.9]}.$$

We can say that the hardware upgrade resulted in a 1.3-minute reduction of the mean running time, with a 90% confidence margin of 0.6 minutes.

Hypothesis testing

- Based on a random sample, we can use Statistics to verify whether
    - a system has not been infected,
    - a hardware upgrade was efficient,
    - the average number of concurrent users increased by 2000 this year,
    - the average connection speed is 54 Mbps, as claimed by the internet service provider,
    - the proportion of defective products is at most 3%, as promised by the manufacturer.
- We need to state exactly what we are testing. These are hypothesis and alternative.
- $H_0 =$ hypothesis (the null hypothesis)
- $H_A =$ alternative (the alternative hypothesis)
- H_0 and H_A are simply two mutually exclusive statements.
- Each test results either in acceptance of H_0 or its rejection in favor of H_A.

Hypothesis testing



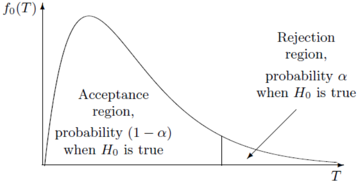
To verify that the the average connection speed is 54 Mbps, we test the hypothesis $H_0 : \mu = 54$ against the *two-sided alternative* $H_A : \mu \neq 54$, where μ is the average speed of all connections.

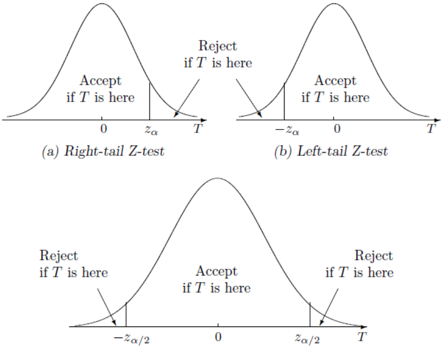
However, if we worry about a *low* connection speed only, we can conduct a one-sided test of

$$H_0 : \mu = 54 \quad \text{vs} \quad H_A : \mu < 54.$$

In this case, we only measure the amount of evidence supporting the *one-sided alternative* $H_A : \mu < 54$. In the absence of such evidence, we gladly accept the null hypothesis.

Acceptance region and rejection region


$f_0(T)$
Acceptance region, probability $(1-\alpha)$ when H_0 is true
Rejection region, probability α when H_0 is true
T



(a) *Right-tail Z-test*

(b) *Left-tail Z-test*

Hypothesis testing