

I01 - Statistics

STAT 587 (Engineering)
Iowa State University

September 7, 2020

Statistics

The **field of statistics** is the study of the collection, analysis, interpretation, presentation, and organization of data.

<https://en.wikipedia.org/wiki/Statistics>

There are two different phases of statistics:

- descriptive statistics
 - statistics
 - graphical statistics
- inferential statistics
 - uses a sample to make statements about a population.

Convenience sample

The **population** consists of all units of interest. Any numerical characteristic of a population is a **parameter**. The **sample** consists of observed units collected from the population. Any function of a sample is called a **statistic**.

Population: in-use routers by graduate students at Iowa State University.

Parameter: proportion of those routers that have Gigabit speed.

Sample: students in STAT 587-2

Statistics: proportion of those students that have Gigabit routers.

Simple random sampling

A **simple random sample** is a sample from the population where all subsets of the same size are equally likely to be sampled. Random samples ensure that statistical conclusions will be valid.

Population: in-use routers by graduate students at Iowa State University.

Parameter: proportion of those routers that have Gigabit speed.

Sample: a pseudo-random number generator gives each graduate student a $\text{Unif}(0,1)$ number and the lowest 100 are contacted

Statistics: proportion that have Gigabit routers.

Sampling and non-sampling errors

Sampling errors are caused by the mere fact that only a sample, a portion of a population, is observed. Fortunately,

error \downarrow as sample size (n) \uparrow

Non-sampling errors are caused by inappropriate sampling schemes and wrong statistical techniques. Often, no statistical technique can rescue a poorly collected sample of data.

Sample: students in STAT 587-2

Statistics and estimators

A **statistic** is any function of the data.

Descriptive statistics:

- Sample mean, median, mode
- Sample quantiles
- Sample variance, standard deviation

When a statistic is meant to estimate a corresponding population parameter, we call that statistic an **estimator**.

Sample mean

Let X_1, \dots, X_n be a random sample from a distribution with

$$E[X_i] = \mu \quad \text{and} \quad \text{Var}[X_i] = \sigma^2$$

where we assume independence between the X_i .

The sample mean is

$$\hat{\mu} = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

and estimates the population mean μ .

Sample variance

Let X_1, \dots, X_n be a random sample from a distribution with

$$E[X_i] = \mu \quad \text{and} \quad \text{Var}[X_i] = \sigma^2$$

where we assume independence between the X_i .

The sample variance is

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

and estimates the population variance σ^2 .

The sample standard deviation is $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$ and estimates the population standard deviation.

Quantiles

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

$$P(X < x) \leq p \quad \text{and} \quad P(X > x) \leq 1 - p.$$

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 **$100p$ -percentile** is a p -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 a 0.5-quantile, 50th percentile, and 2nd quartile.

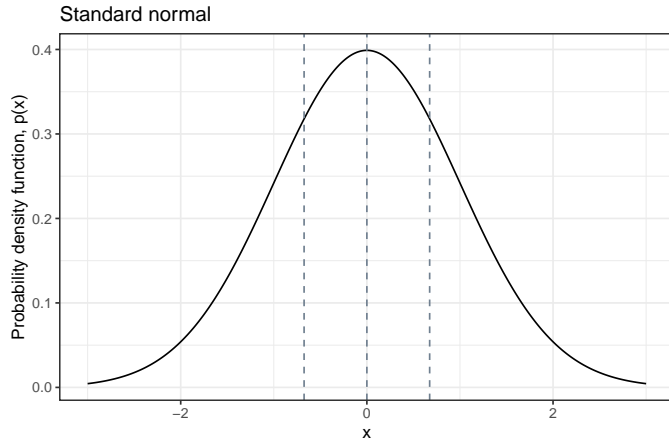
The **interquartile range** is the third quartile minus the first quartile, i.e.

$$IQR = Q_3 - Q_1$$

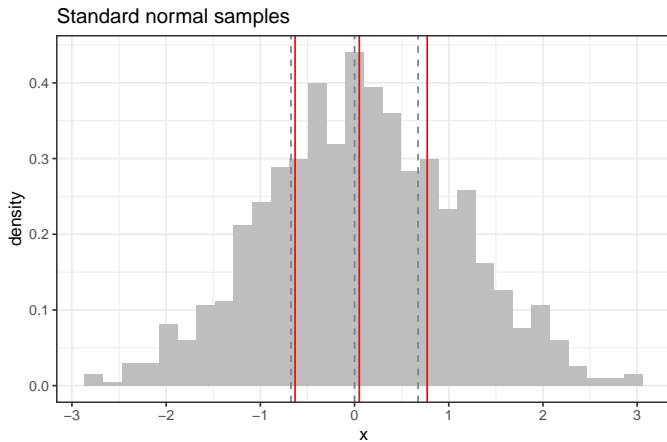
and the **sample interquartile range** is the third sample quartile minus the first sample quartile, i.e.

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

Standard normal quartiles



Sample quartiles from a standard normal



Properties of statistics and estimators

Statistics can have properties, e.g.

- standard error

Estimators can have properties, e.g.

- unbiased
- consistent

Standard error

The **standard error** of a statistic $\hat{\theta}$ is the standard deviation of that statistic (when the data are considered random).

If X_i are independent and have $Var[X_i] = \sigma^2$, then

$$\begin{aligned} Var[\bar{X}] &= Var\left[\frac{1}{n} \sum_{i=1}^n X_i\right] \\ &= \frac{1}{n^2} \sum_{i=1}^n Var[X_i] = \frac{1}{n^2} \sum_{i=1}^n \sigma^2 = \frac{\sigma^2}{n} \end{aligned}$$

and thus

$$SD[\bar{X}] = \sqrt{Var[\bar{X}]} = \sigma/\sqrt{n}.$$

Thus the standard error of the sample mean is σ/\sqrt{n} .

Unbiased

An estimator $\hat{\theta}$ is **unbiased** for a parameter θ if its expectation (when the data are considered random) equals the parameter, i.e.

$$E[\hat{\theta}] = \theta.$$

The sample mean is unbiased for the population mean μ since

$$E[\bar{X}] = E\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \frac{1}{n} \sum_{i=1}^n E[X_i] = \mu.$$

and the sample variance is unbiased for the population variance σ^2 .

Consistent

An estimator $\hat{\theta}$, or $\hat{\theta}_n(x)$, is **consistent** for a parameter θ if the probability of its sampling error of any magnitude converges to 0 as the sample size n increases to infinity, i.e.

$$P \left(\left| \hat{\theta}_n(X) - \theta \right| > \epsilon \right) \rightarrow 0 \text{ as } n \rightarrow \infty$$

for any $\epsilon > 0$.

The sample mean is consistent for μ since

$Var [\bar{X}] = \sigma^2/n$ and

$$P (|\bar{X} - \mu| > \epsilon) \leq \frac{Var [\bar{X}]}{\epsilon^2} = \frac{\sigma^2/n}{\epsilon^2} \rightarrow 0$$

where the inequality is from Chebyshev's inequality.

Binomial example

Suppose $Y \sim \text{Bin}(n, \theta)$ where θ is the probability of success. The statistic $\hat{\theta} = Y/n$ is an estimator of θ .

Since

$$E \left[\hat{\theta} \right] = E \left[\frac{Y}{n} \right] = \frac{1}{n} E[Y] = \frac{1}{n} n\theta = \theta$$

the estimator is **unbiased**.

Binomial example

Suppose $Y \sim \text{Bin}(n, \theta)$ where θ is the probability of success. The statistic $\hat{\theta} = Y/n$ is an estimator of θ .

The variance of the estimator is

$$\text{Var} \left[\hat{\theta} \right] = \text{Var} \left[\frac{Y}{n} \right] = \frac{1}{n^2} \text{Var}[Y] = \frac{1}{n^2} n\theta(1 - \theta) = \frac{\theta(1 - \theta)}{n}.$$

Thus the **standard error** is

$$SE(\hat{\theta}) = \sqrt{\text{Var}[\hat{\theta}]} = \sqrt{\frac{\theta(1 - \theta)}{n}}.$$

By Chebychev's inequality, this estimator is **consistent** for θ .

Summary

- Statistics are functions of data.
- Statistics have some properties:
 - Standard error
- Estimators are statistics that estimate population parameters.
- Estimators may have properties:
 - Unbiased
 - Consistent

Look at it!

Before you do anything with a data set,
LOOK AT IT!

Why should you look at your data?

1. Find errors
 - Do variables have the correct range, e.g. positive?
 - How are Not Available encoded?
 - Are there outliers?
2. Do known or suspected relationships exist?
 - Is X linearly associated with Y?
 - Is X quadratically associated with Y?
3. Are there new relationships?
 - What is associated with Y and how?
4. Do variables adhere to distributional assumptions?
 - Does Y have an approximately normal distribution?
 - Right/left skew
 - Heavy tails

Principles of professional statistical graphics

<https://moz.com/blog/data-visualization-principles-lessons-from-tufte>

- Show the data
 - Avoid distorting the data, e.g. pie charts, 3d pie charts, exploding wedge 3d pie charts, bar charts that do not start at zero
- Plots should be self-explanatory
 - Use informative caption, legend
 - Use normative colors, shapes, etc
- Have a high information to ink ratio
 - Avoid bar charts
- Encourage eyes to compare
 - Use size, shape, and color to highlight differences

Stock market return

http://www.nytimes.com/interactive/2011/01/02/business/20110102-metrics-graphic.html?_r=0

