Set08 - Statistics

STAT 401 (Engineering) - Iowa State University

February 13, 2017

Statistics

Definition

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.

Population and sample

Definition

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.

Example

Consider the population of all in-use routers by undergraduate students at Iowa State University. We are interested in what proportion have Gigabit speed. We collect data from students in STAT 401 (our sample) and record the proportion (a statistic) that have Gigabit routers.

Simple random sampling

Definition

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

Example

Consider the population of all in-use routers by undergraduate students at Iowa State University. We are interested in what proportion have Gigabit speed. A pseudo-random number generator gives each student a Unif(0,1) number and the lowest 100 are contacted (our sample) and the proportion (a statistic) of these students who have Gigabit routers is recorded.

Sampling and non-sampling errors

Definition

Sampling errors are caused by the mere fact that only a sample, a portion of a population, is observed. For most reasonable statistical procedures, sampling errors decrease (and converge to zero) as the sample size increases.

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.

Example

In our example, no statistical technique can help us estimate the proportion of students at ISU who have Gigabit routers based on our convenience sample of STAT 401 students who have a Gigabit router.

Descriptive statistics

Definition

A statistic is any function of the data.

Example

Statistics:

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

Definition

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

Sample mean

Let $X_1, \ldots, X_n$ be a sample from a distribution with

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

where we assume independence between the X_i.

The sample mean is

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

and estimates the population mean μ.

The sample variance is

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

and estimates the population variance σ^2. The sample standard deviation is $S = \sqrt{S^2}$.

Unbiased

Definition

An estimator is unbiased for a parameter if its expectation (when the data are considered random) equals the parameter.

Example

The sample mean is unbiased for μ since

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

and the sample variance is unbiased for σ^2.

Consistent

Definition

An estimator $\hat{\theta}$, or $\hat{\theta}(x)$, 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, i.e.

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

for any $\epsilon > 0$.

Example

The sample mean is consistent for μ since $Var\left[\overline{X}\right] = \sigma^2/n$ and

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

where the inequality is from Chebyshev's inequality.

Quantiles

Definition

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

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

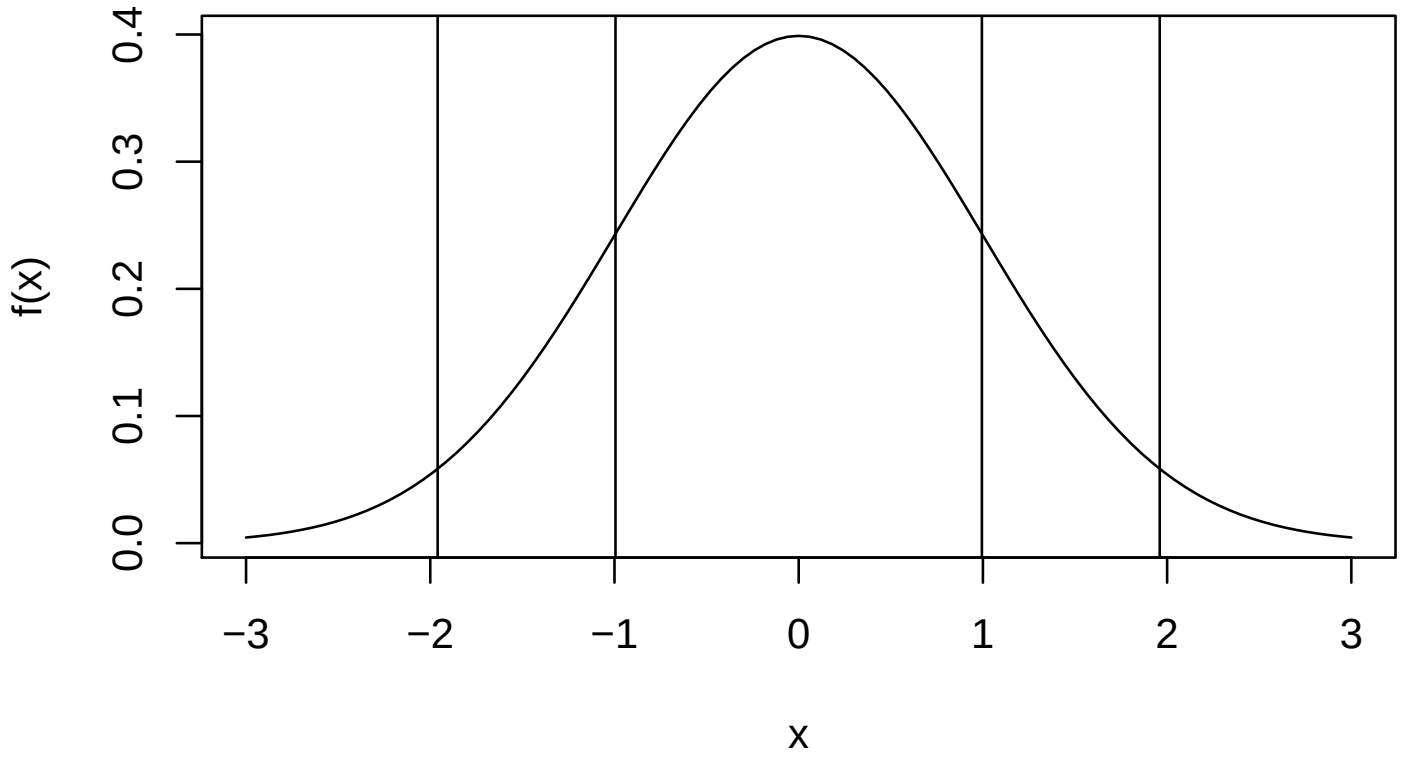
A sample p-quantile is any number that exceeds at most $100p\%$ of the smample, and is exceeded by at most $100(1-p)\%$ of the sample. A $100p$-percentile is a p-quantile. First, second, and third quartiles are the 25th, 50th, and 75th percentiles. They split a population or a sample into four equal 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.

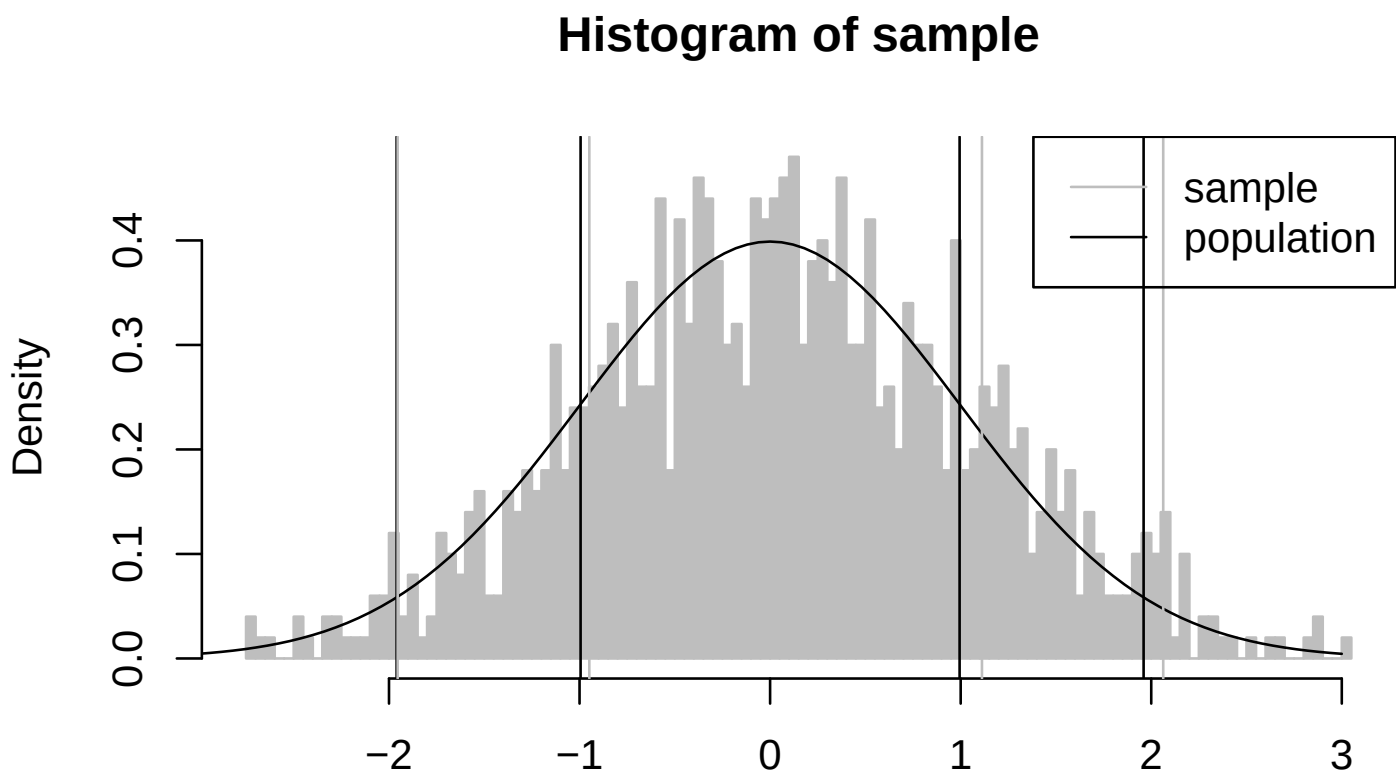
Standard normal quartiles

```
curve(expr = dnorm, from = -3, to = 3, ylab = "f(x)")
quantiles = c(.025,.16,.84,.975)
abline(v = qnorm(p = quantiles)) # default is standard normal
```



Sample quartiles from a standard normal

```
n = 1000
sample = rnorm(n)
hist(x = sample, breaks = 101, probability = TRUE, border = "gray", col = "gray")
curve(expr = dnorm, from = -3, to = 3, ylab = "f(x)", col = "black", add = TRUE)
abline(v = qnorm(p = quantiles), col = "black")
abline(v = quantile(sample, prob = quantiles), col = "gray")
legend("topright", c("sample","population"), lty=1, col=c("gray","black"))
```



Standard error

Definition

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

Example

The standard error of the sample mean is $\sigma/\sqrt{n}$ since (if the X_i are independent) we have

$$Var\left[\overline{X}\right] = Var\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} Var[X_i] = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \sigma^2/n$$

and thus

$$SD\left[\overline{X}\right] = \sqrt{Var\left[\overline{X}\right]} = \sigma/\sqrt{n}.$$

Binomial example

Suppose $Y \sim Bin(n, \theta)$ 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.

The variance of the estimator is

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

Thus the standard error is

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

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

Summary

- Statistics are functions of data.
- Statistics have some properties:
  - Standard error
- Statistics often try to estimate population parameters and are then called estimators.
- Estimators may have these properties relative to the population parameter they are trying to estimate:
  - Unbiased
  - Consistent

Look at it!

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

Why should you look at your data?

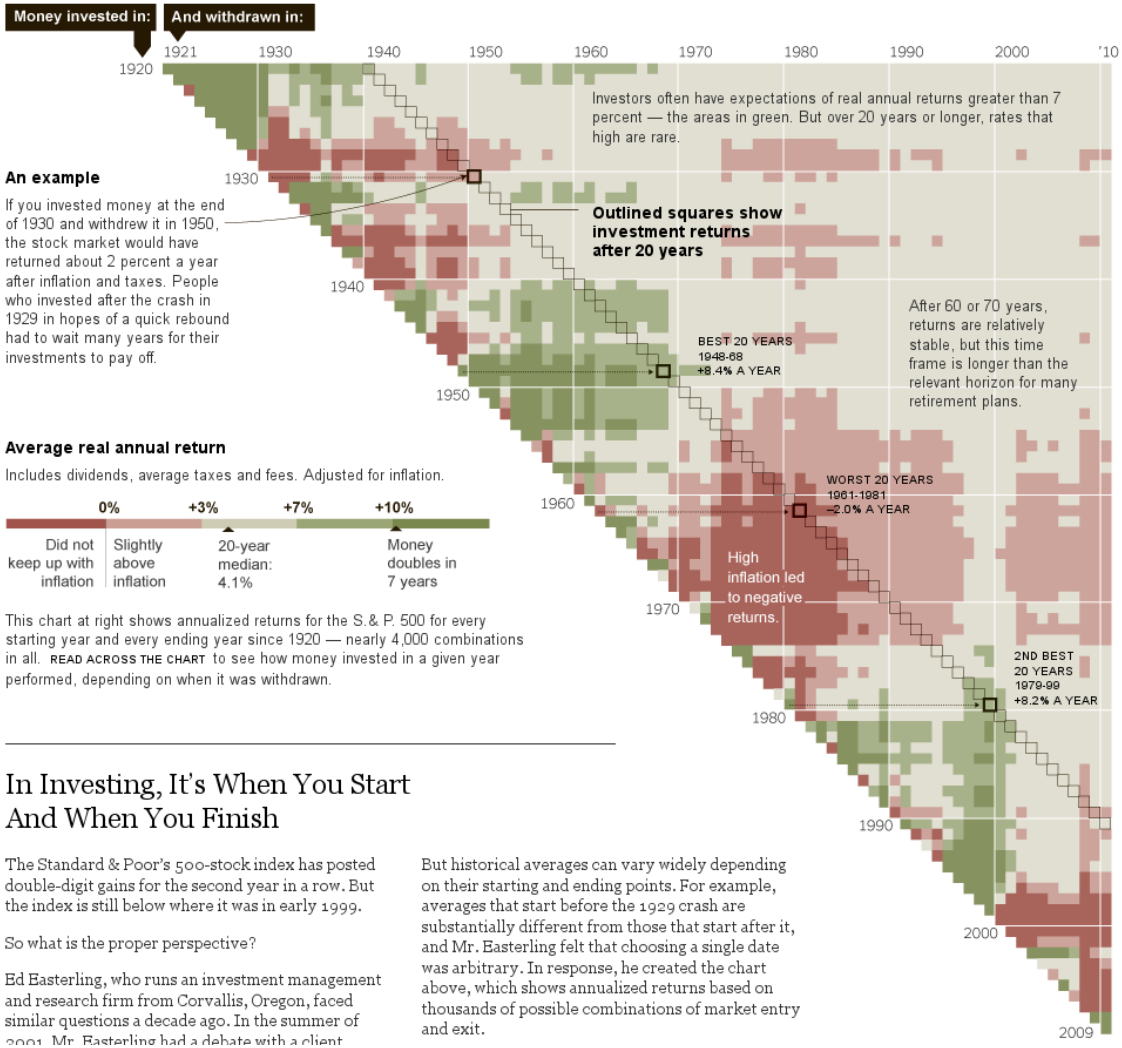
1. Find errors
   - Do variables have the correct range, e.g. positive?
   - How are Not Available encoded?
   - Are there outliers?
2. Do known or suspected relationships exist?
   - Is X linearly associated with Y?
   - Is X quadratically associated with Y?
3. Are there new relationships?
   - What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

Principles of professional statistical graphics

- 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

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

Stock market return



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