Statistics

"The art and science of gathering, analyzing, and making inferences from data"

Mosteller, Rourke and Thomas ("MRT2")
Probability with statistical applications 2nd Edition, p2

"The art of making numerical conjectures about puzzling questions"

Freedman, Pisani, Purves and Adhikari ("FPPA2")
Statistics 2nd Edition, pxiii

"The science of collecting, organizing, and interpreting numerical facts, which we call *data*"

Moore and McCabe ("M&M3")
Introduction to the Practice of Statistics 3rd Edition, p xxv

"A branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data"

"Methods for drawing conclusions from results of experiments or processes"

"The entire science of decision making in the face of uncertanty"

cited by Wackerly, Mendenhall and Scheaffer ("WMS5")
Mathematical Statistics with Applications 5th Edition ,p 1

"The field of statistical inference **leans heavily upon the theory of probability**, but supplements it... some parts of ststistics are not mathematical, while other parts are"

Mosteller, Rourke and Thomas , p2

Inference

"Population"	"Sample"	Ref.
"The large body of data that is the target of interest"	"The subset selected from it"	*WMS5, p2*
A **parameter** is a number that describes the population. A parameter is a fixed number, but in practice we do not know its value"	A **statistic** is a number that describes a sample. Its value is known when we have taken a sample, but it can change from (one possible) sample to (another possible) sample	*M&M3, p 268*
"population" of all possible sets of measurements obtainable or imaginable under comparable experimental conditions"	Any set of measure-ments can be thought of as a "sample" from the "population" ..	*MRT2, p223*
"A whole class of individuals an investigator wants to generalize about.."	"part of it.."	*FPPA2, p305*
Numerical fact about the population is called a PARAMETER	parameter is estimated by a STATISTIC, a number which can be computed from a sample	
Information useful in inferring some characteristic of a population (either **existing** or **conceptual**) is purchased an a specified quantity and results in an inference (**estimation** or **decision**) with an associated degree of goodness.		*WMS5, pp 2-3*

"THE OBJECTIVE OF STATISTICS IS TO MAKE AN INFERENCE ABOUT A POPULATION BASED ON INFORMATION CONTAINED IN A SAMPLE AND TO PROVIDE AN ASSOCIATED MEASURE OF GOODNESS FOR THE INFERENCE."

Variables (from M&M3, p 4)

Any set of data contains information about some group of **individuals**. The information is organized in **variables**.

Individuals : the objects (people, animals, things) described by the data (the 'rows' in a spreadsheet)

A **variable** is any characteristic of an individual. A variable can take different values for different individuals.

Might set it up as a spreadsheet or database with individuals (cases, observations) in rows, and variables (data items) in columns.

Categorical and Quantitative Variables (from M&M3, p 5)

Categorical: represented as one of several adjectives or categories

Quantitative: takes numerical values ("discete" / "continuous")

Distribution of a Variable

- **what values** a variable takes and
- **how often** it takes these values

See also:

NOTES FOR COURSE 607: M&M Ch 1.1 and 1.2 Displaying / Describing Distributions

Distribution of a Quantitative Variable (from M&M3)

Stemplot (also known as stem-and-leaf-plot)

- Gives shape of distribution
- Includes the actual numerical values in the graph
- Good for small numbers of positive values

Steps...

1 separate each observation into

stem (all but rightmost digit)

leaf (the final digit)

2 write the stems in vertical column with the smallest at the top, and

draw a vertical line at the right of this column

3 write each leaf in a row to right of its stem, in increasing order out from the stem

Back-to-back Stemplot (with common stems)

- useful for comparing 2 related distributions

Splitting the stems (e.g. 1 with leaves 0-4, other 5-9)

Rounding values before making stemplot

if observed values have too many digits

Histogram (see WMS5)

frequency [count] or relative frequency [fraction or percent]) falling in each class interval

NB: Areas of rectangles over class intervals proportional to frequencies
(total area = 1 or 100% or number (n) of individuals)

Descriptive Statistics

Central Tendency / Location

Mean

Average

- arithmetic (strict meaning of "average")
- sometimes "weighted"

Geometric

Harmonic

(Minimum value + Maximum value) / 2

Median (Y_{50})

Middlemost value (1/2 above it, 1/2 below)

Mode

Most frequent value

Spread / Dispersion / Scatter

Range: maximum - minimum

Interquartile Range ("IQR" : Y_{25} and Y_{75})

Other Percentiles (e.g., Y_5 to Y_{95} ; Y_3 to Y_{97})

Probable Error ($Y_{75} – Y_{25}$) / 2 [Old, before SD]

Average (absolute) Deviation from Mean

Square Root of Average Squared Deviation from Mean

= $\sqrt{\text{Average* Squared Deviation from Mean}}$

= $\sqrt{\text{"Variance"}}$ = "**Standard" Deviation** (SD)

= Root Mean Square (RMS)

* used slightly loosely [divisor is "n–1" rather than "n"]

Notation

Measure	PARAMETER	STATISTIC
	"P"="Population"	"S"="Statistic"
Mean	μ read "mew" spelled "mu"	$\bar{y}$ read ("y-bar")
Standard Deviation	σ read "sigma"	s
Variance	σ^2 read "sigma-squared"	s^2
Others	No special symbols; distinguished by words (e.g., "Population" median" vs "sample" median) or by context	

Empirical Rule"normal"distribution

Statistically Correct Terminology

Gaussian Distribution or
"Bell-shaped" Distribution

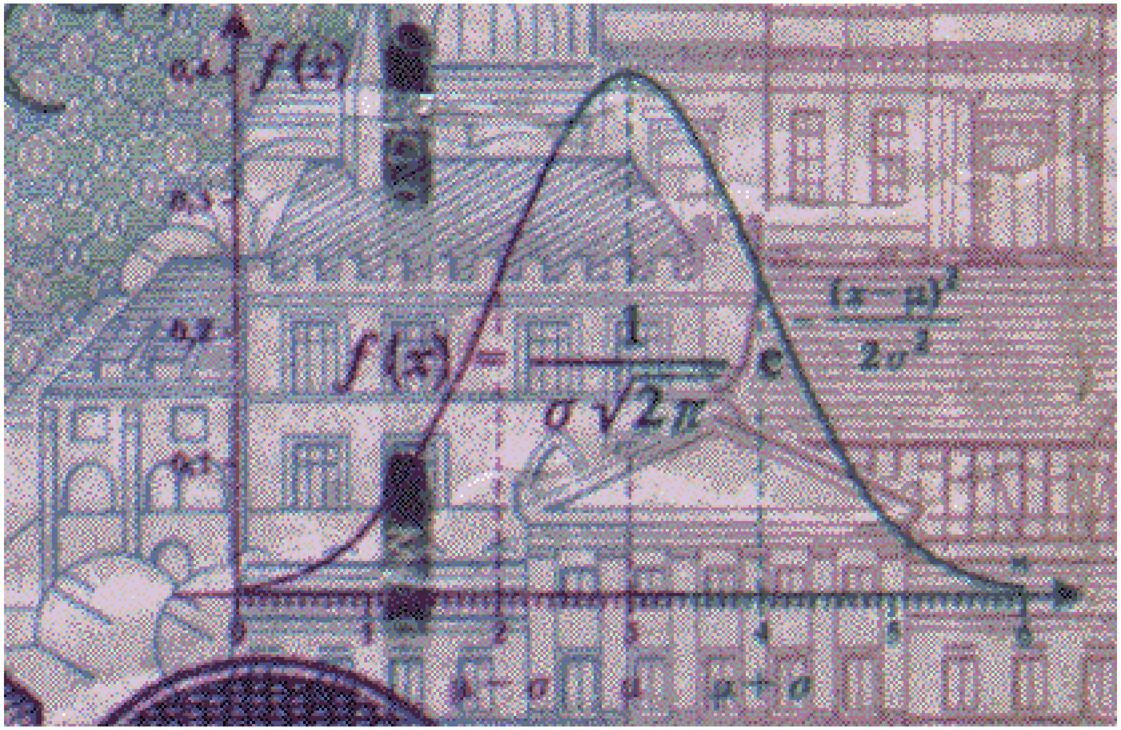
The Gaussian distribution in nature

Many more distributions of individual values (even "in nature") are Non-Gaussian than Gaussian

A standard deviation (SD) that is large relative to (or larger than) the mean can be a tip-off that the distribution of values is markedly Non-Gaussian

The Gaussian distribution of (man-made) statistics

The variability of many statistics (computed by summing or averaging or otherwise aggregating individual values) is often near-Gaussian, especially if the statistic is based on reasonable sample sizes


$f(x)$
0,4
0,3
0,2
0,1
$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
0
1
2
3
4
5
6
x
$\mu-\sigma$
μ
$\mu+\sigma$