Conrad

School of Entrepreneurship and Business

Agenda

01

Recap

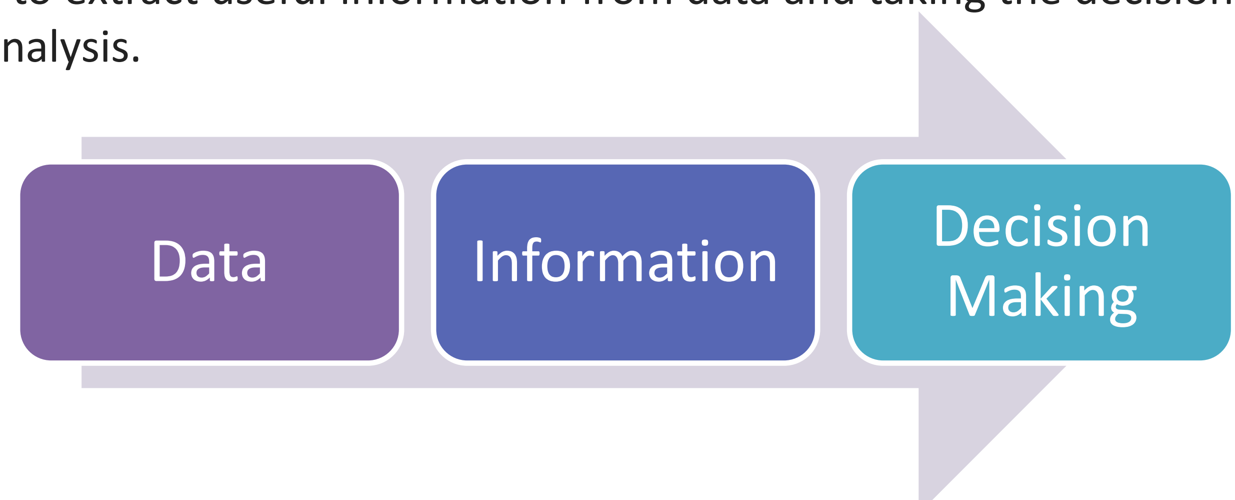
02

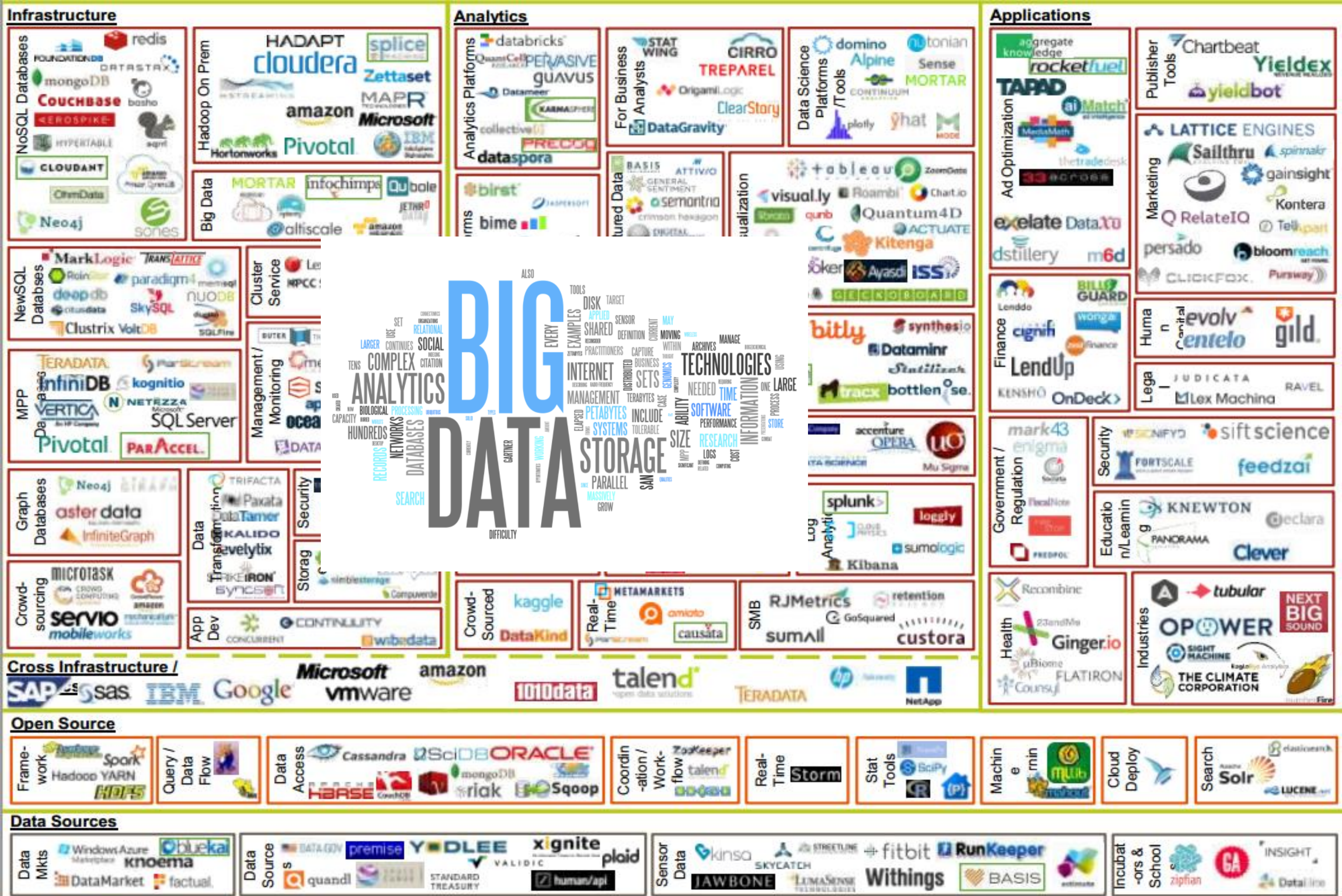
Data Fundamentals


DATA
ANALYSIS
STRUCTURE
PHASE
QUALITY
CHECKED
ANALYTICS
WAYS
ONE
STATISTICS
MODELS
ASSESSED
IMPORTANT
MEASUREMENT
FOCUSES
POSSIBLE
FREQUENCY
INITIAL
RESEARCH
CLEAR
STRUCTURAL
EDA
TECHNIQUES
MODELING
CHECK
QUESTION
HYPOTHESES
INSTRUMENTS
RESULTS
EXPLORATORY
ORIGINAL
DESCRIPTIVE
DIFFERENT
PREDICTIVE
ANALYSES
CDA
SAMPLE
SEVERAL
PLOTS
SCIENCE
MAIN
CONFIRMATORY
STATISTICAL
APPROACH
BUSINESS
INFORMATION

What is Data Analysis?

Data analysis is defined as a process of cleaning, transforming, and modeling data to discover useful information for business decision-making. The purpose of Data Analysis is to extract useful information from data and taking the decision based upon the data analysis.




Infrastructure
NoSQL Databases
Hadoop On Prem
Big Data
NewSQL Databases
Cluster Service
MPP Databases
Management / Monitoring
Graph Databases
Data Transformation
Security
Storage
Crowd-sourcing
App Dev
Analytics
Analytics Platforms
For Business Analysts
Data Science Platforms /Tools
Structured Data
Visualization
Crowd-Sourced
Real-Time
SMB
Log Analytics
Applications
Ad Optimization
Publisher Tools
Marketing
Finance
Human Capital
Legal
Government / Regulation
Security
Education/Learning
Health
Industries
Cross Infrastructure / Analytics
Open Source
Framework
Query / Data Flow
Data Access
Coordination / Workflow
Real-Time
Stat Tools
Machine Learning
Cloud Deployment
Search
Data Sources
Data Mkts
Data Sources
Sensor Data
Incubators & Schools
BIG
DATA
ANALYTICS
TECHNOLOGIES
STORAGE
INTERNET
INFORMATION
SOFTWARE

What is your definition of Big Data? Researchers' understanding of the phenomenon of the decade

Maddalena Favaretto, Eva De Clercq, Christophe Olivier Schneble, Bernice Simone Elger

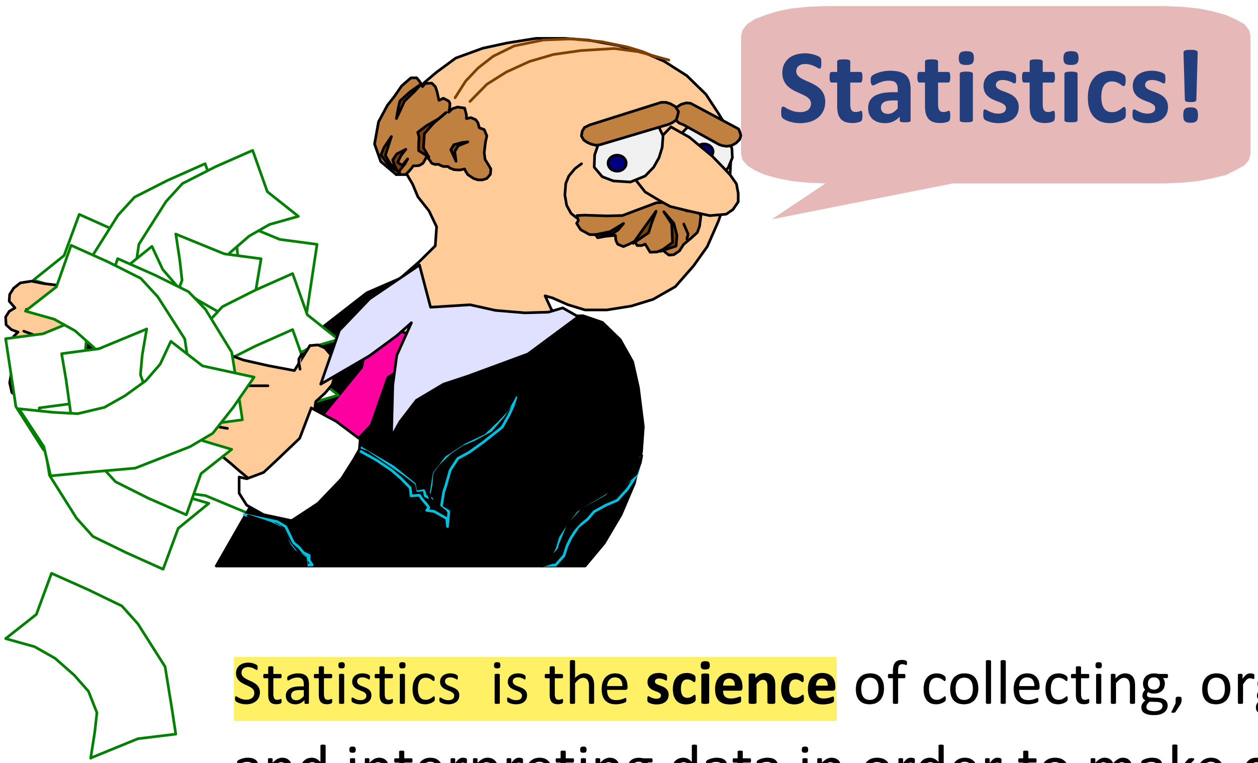
Published: February 25, 2020 • https://doi.org/10.1371/journal.pone.0228987

Conclusion

The study identified an overall uncertainty or uneasiness among researchers towards the use of the term Big Data which might derive from the tendency to recognize Big Data as a shifting and evolving cultural phenomenon. Moreover, the currently enacted use of the term as a hyped-up buzzword might further aggravate the conceptual vagueness of Big Data.

What is Data Analysis?

Data analysis is defined as a **process** of cleaning, transforming, and modeling data to discover useful information for business decision-making. The purpose of Data Analysis is to extract useful information from data and taking the decision based upon the data analysis.



Statistics is the **science** of collecting, organizing, analyzing, and interpreting data in order to make decisions.

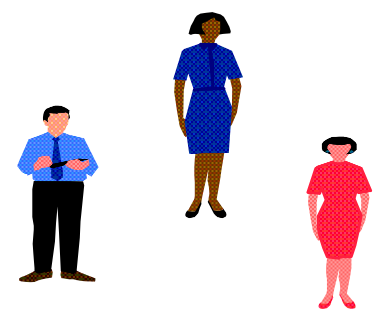
Important Terms

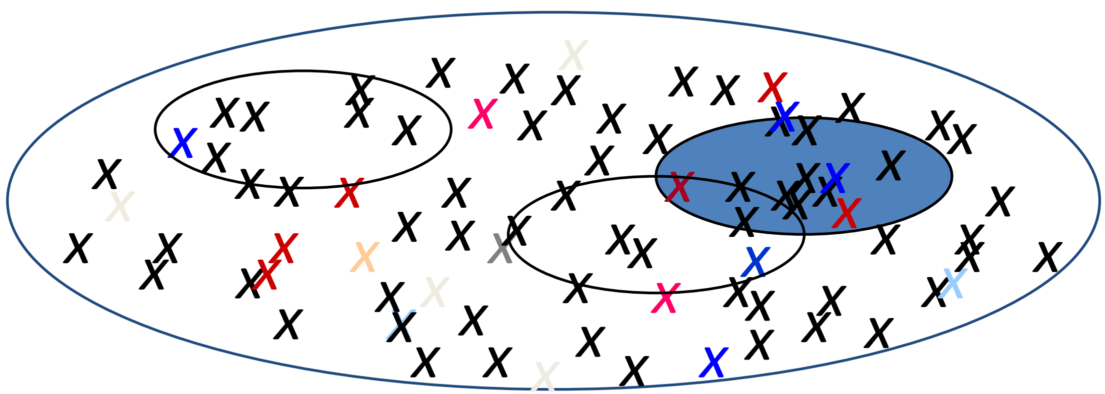
- **Population**

The collection of *all* responses, measurements, or counts that are of interest.

- **Sample**

A portion or subset of the population.





Important Terms

- **Parameter:**

A number that describes a population characteristic.

Average gross income of all Canadian residents in 2021

- **Statistic:**

A number that describes a sample characteristic.

2021 gross income of residents from a sample of two provinces.

Measurement and Scaling

- **Measurement**

  Standardized process of assigning numbers or other symbols to certain characteristics of objects of interests according to pre-specified rules

- **Characteristics for Standardization**

  •One-to-one correspondence between the symbol and the characteristic in the object that is being measured

  •Rules for assignment should be invariant over time* and the objects being measured

The taste of Honey Munch Cereal is

Example: Divide 100 points among the following characteristic of a delivery service according to how important each characteristic is to you when selecting a delivery company

Accurate Invoicing ________
Delivery as Promised ________
Lower Price ________

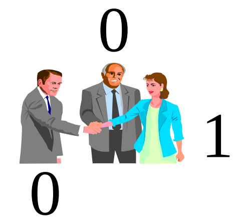
Measurement and Scaling

Scaling

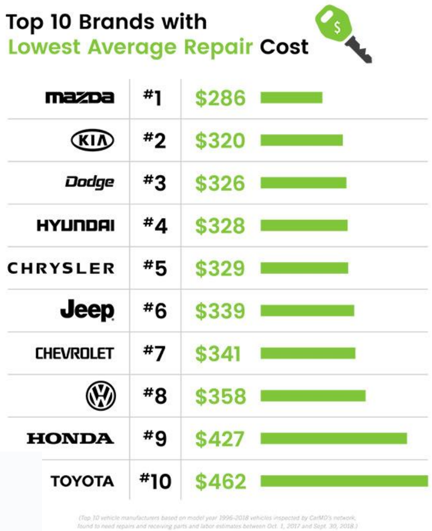
Process of creating a continuum on which objects are located according to the amount of the measured characteristic that the object possesses

Scale Types

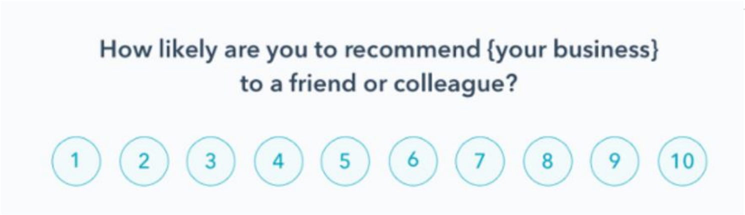
- **Nominal Scales**
  - Ex. Type of car you drive, gender



- **Ordinal Scales**
  - Ex. TV ratings, product ranking



- **Interval Scale**
  - Ex. Most market research



- **Ratio Scale**
  - Ex. Monthly Income

How to Identify a Scale

Question	*Answer*	
	If Yes	If No
Does it have a true zero?	Ratio Scale. Stop	Not Ratio
Are the distances between the scale points equal?	Interval Scale. Stop	Not Interval
Are the scale points greater than or less than one another?	Ordinal Scale. Stop	Not Ordinal
Are the scale points unique categories that you can't say are greater than or less than one another?	Nominal Scale. Stop	Not nominal
If you arrive here, you have made a mistake.	Repeat from the top	Repeat from the top

Why Worry About The Scale?

- Low level scales (nominal and ordinal) are “coarse” measures, while high level scales are “fine” measures

- ***So What?***

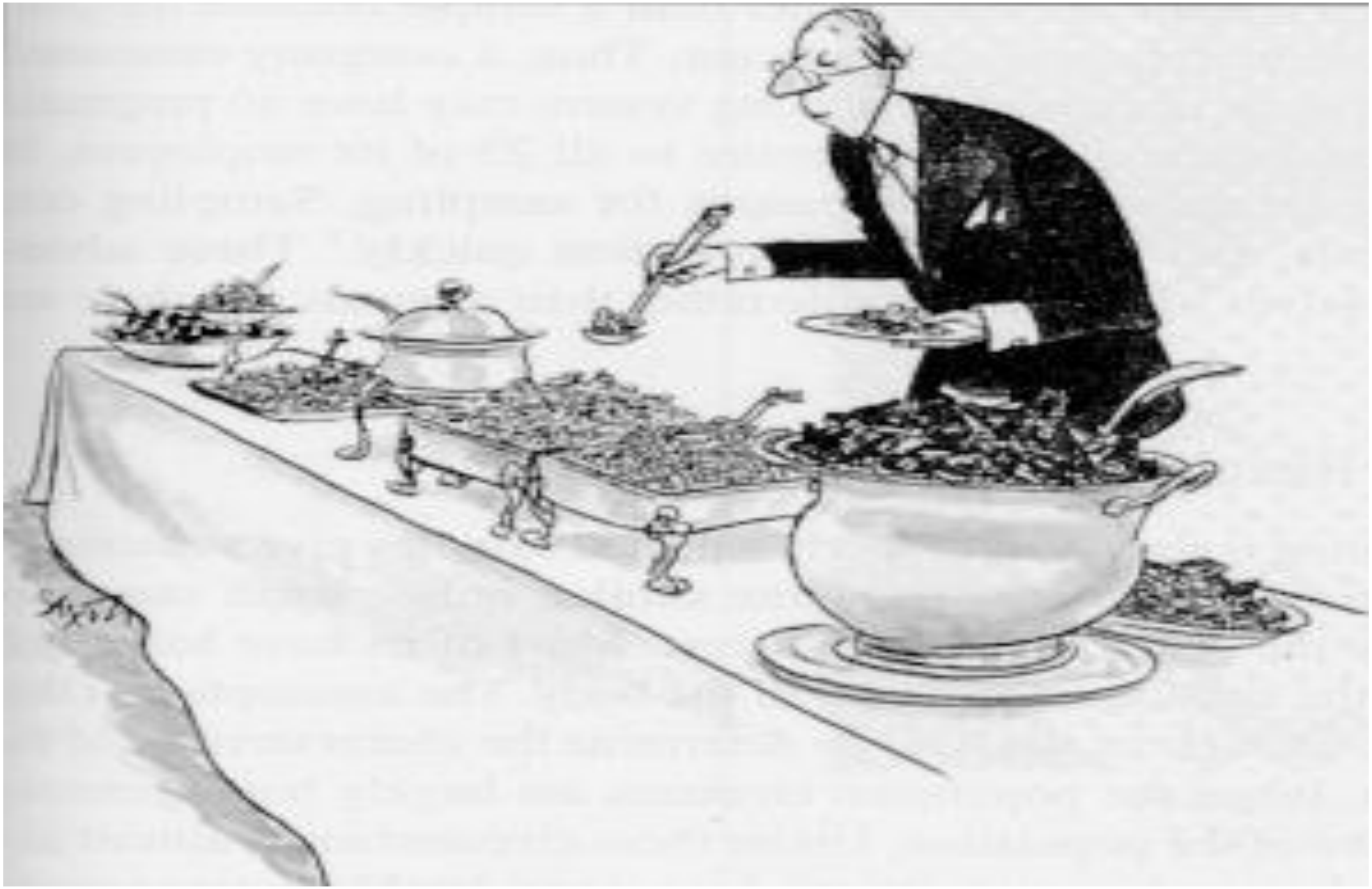
Statistical procedures have scale level assumptions

- ***And…***

Higher scales can be collapsed to lower ones, but the reverse is not true

Q. How satisfied were you with the food quality at our restaurant today?				
sample =100				
Extremely Satisfied	**Satisfied**	**Neutral**	**Dissatisfied**	**Extermely Dissatisfied**
32%	10%	25%	15%	18%
Satisfied			**Dissatfied**	

Sampling



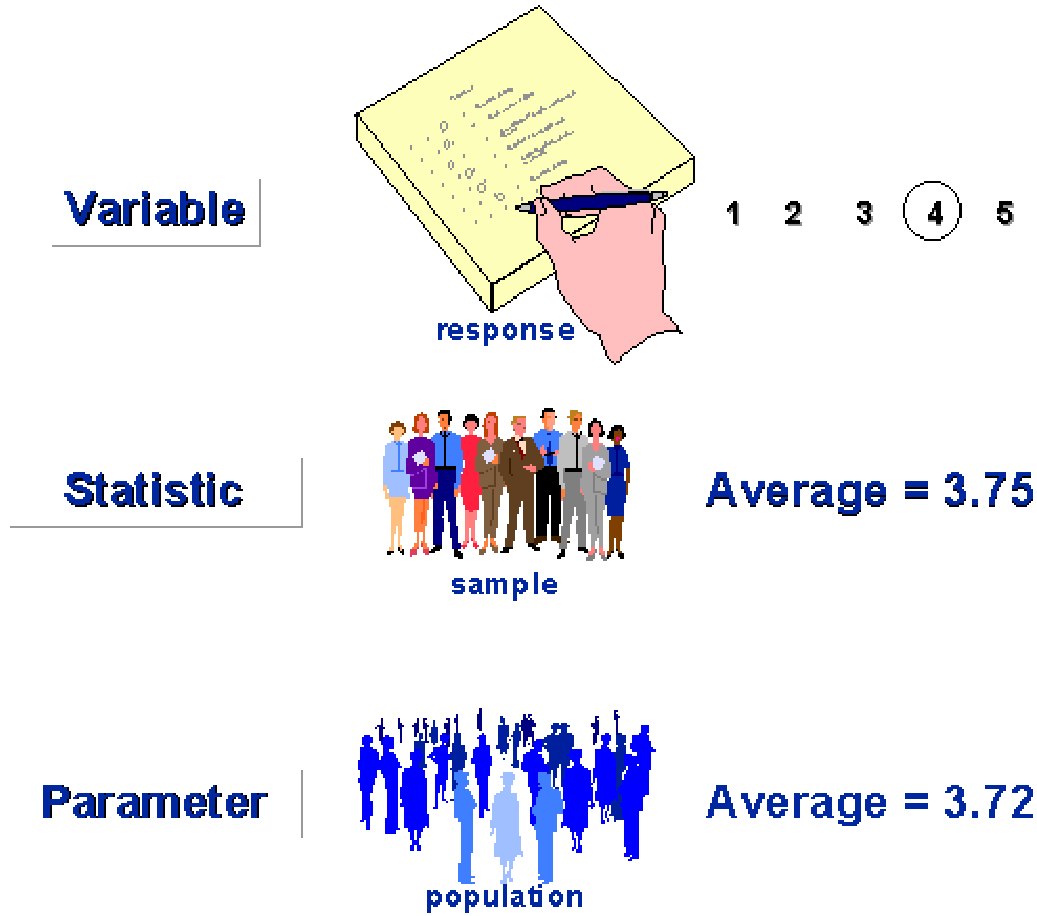
Sampling Fundamentals

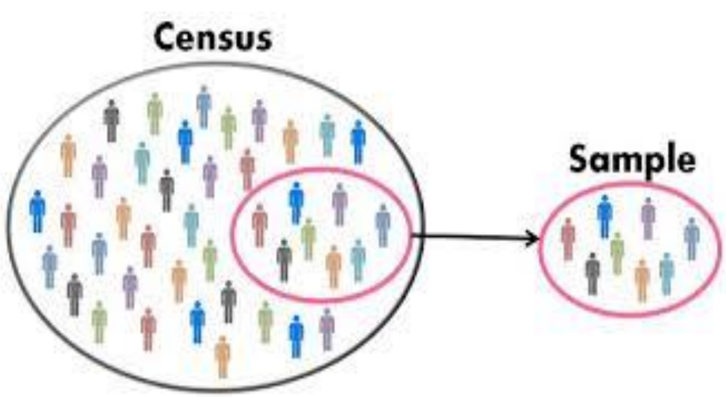
Marketing Research usually involves collecting information about some characteristic of the population be it usage, satisfaction levels, attitudes etc.

Two Ways

Population -> Parameter

Sample -> Statistic





Sampling Fundamentals

When Is Census Appropriate?

- Population size itself is quite small
- Information is needed from every individual in the population
- Cost of making an incorrect decision is high
- Sampling errors are high

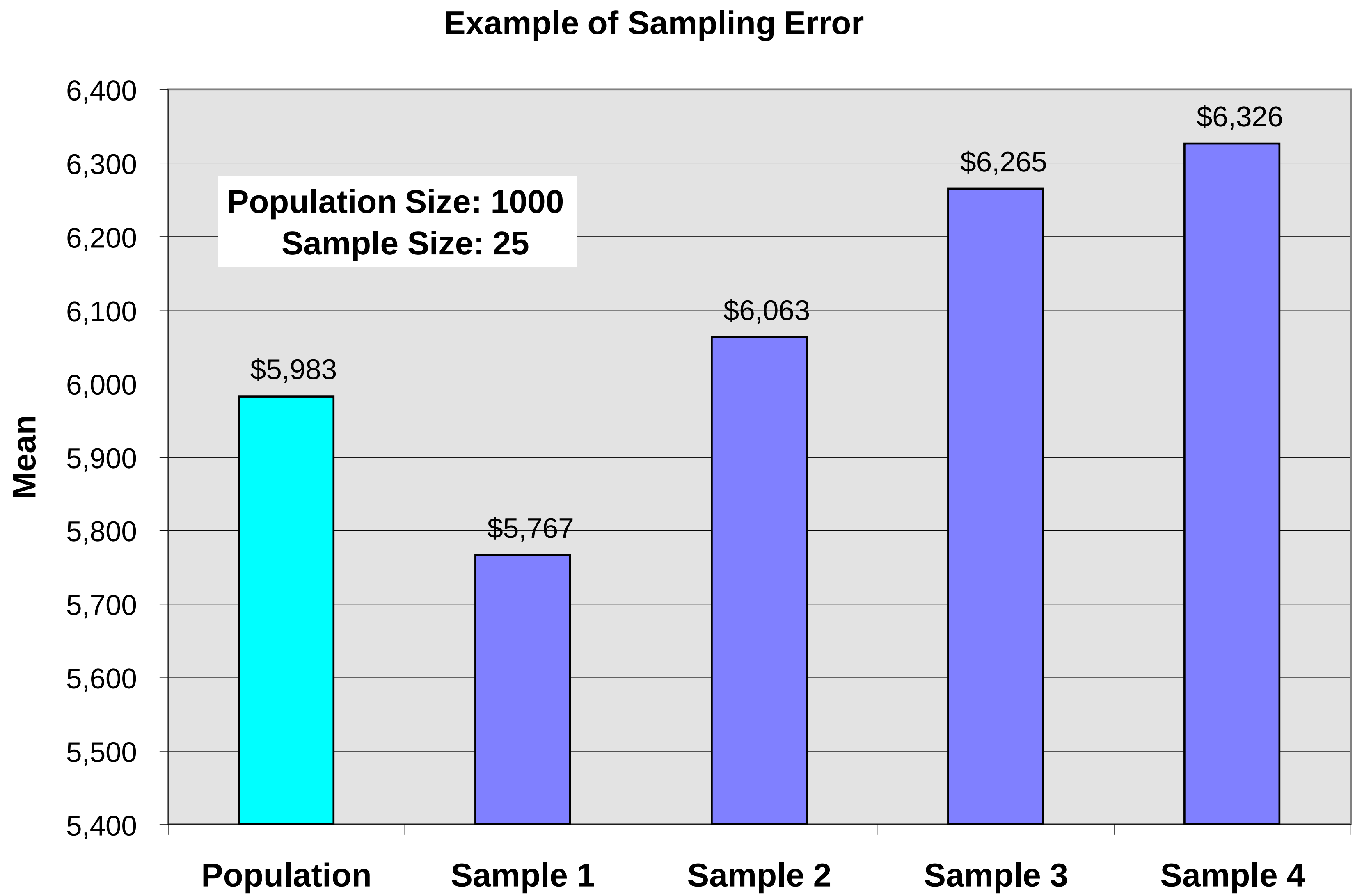
When Is Sample Appropriate?

- Population size is large
- Both cost and time associated with obtaining information from the population is high
- Quick decision is needed
- Population being dealt with is homogeneous*

Error between population and sample values

- **Total Error**
  - Difference between the true value and the observed value of a variable
- **Sampling Error** ← Can be estimated
  - Error is due to sampling
- **Non-sampling Error** ← Can be controlled
  - Error is observed in both census and sample
    - » Measurement Error
    - » Data Recording Error
    - » Data Analysis Error
    - » Non-response Error

Measure: Monthly Household Income



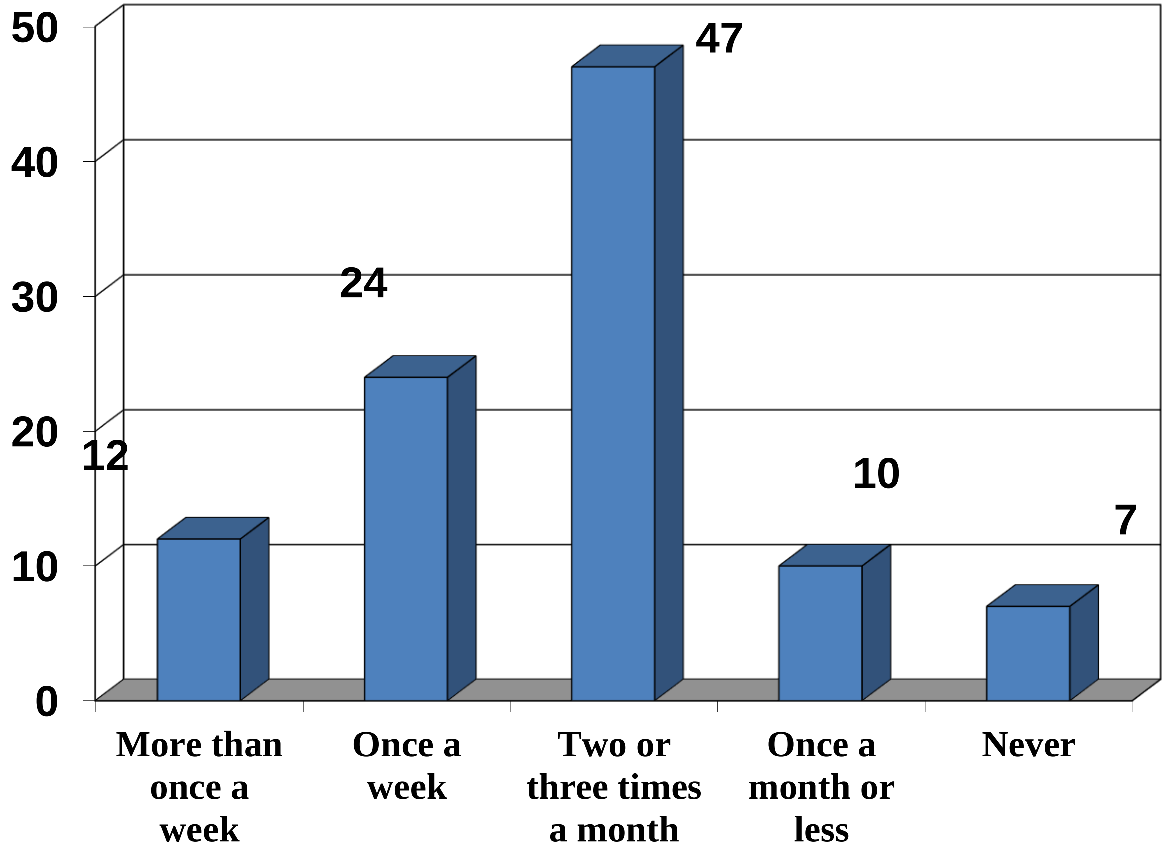
The One and Only Goal in Sampling!!

Select a sample that is as representative as possible.


STATISTICS
101

In the past three months, how often have you ordered a delivery from Burger King?

Coding:	-2	-1	0	1	2



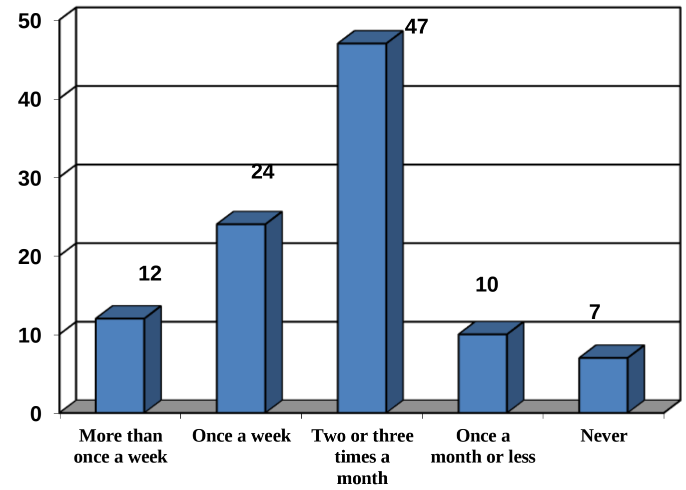
Sample Size: 100

Basic Measures

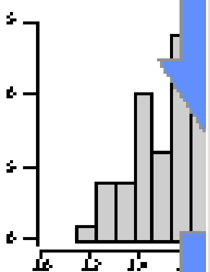
	Population	Sample
Mean	μ	$\bar{X}$
Variance	σ^2	s^2
Standard Deviation	σ	s
Sample Size	N	n

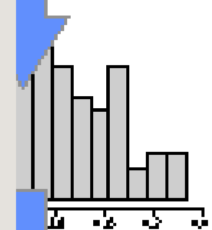
$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = -0.24$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = 1.067$$



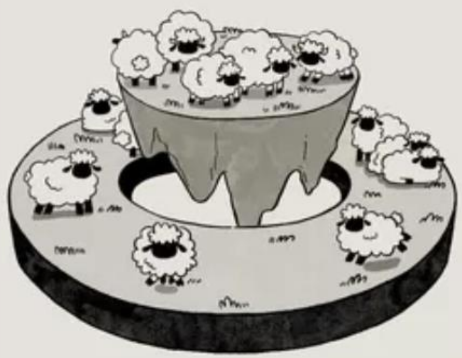
Sampling Distribution





Aver

rage

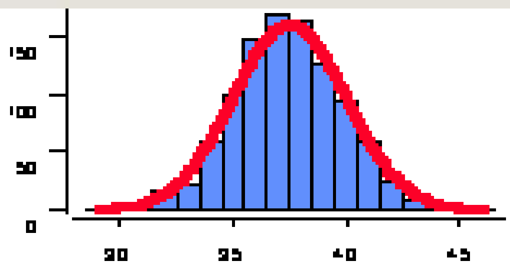


Central Limit Theorem (CLT)

['sen-trəl 'li-mət 'thē-ə-rəm]

The principle that the distribution of sample means approximates a normal distribution as the sample size gets larger, regardless of the population's distribution.

The Sampling Distribution...



...is the distribution of a statistic across an infinite number of samples

Standard Error

- **Variation of $\bar{X}$**
  - Standard error depends on sample size and population standard deviation
  - Assume that variation of $\bar{X}$ follows normal distribution
  - Sampling distribution

$$\boxed{\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}}}$$

= Variation in $\bar{X}$

Where:

$\sigma_{\bar{X}}$ = standard error

σ_X = standard deviation of population

n = sample size

If standard deviation of population (σ) not known use s

Interval Estimation

- $\bar{X}$ varies from sample to sample
- The difference between the sample mean ($\bar{X}$) and the population mean is the sampling error
- $\bar{X} \pm$ sampling error = interval estimate of population mean

$\bar{X} \pm Z\, \sigma_{\bar{x}}$ (sampling error)

Or

$\bar{x} \pm z\, \sigma_x / \sqrt{n}$

n - sample size

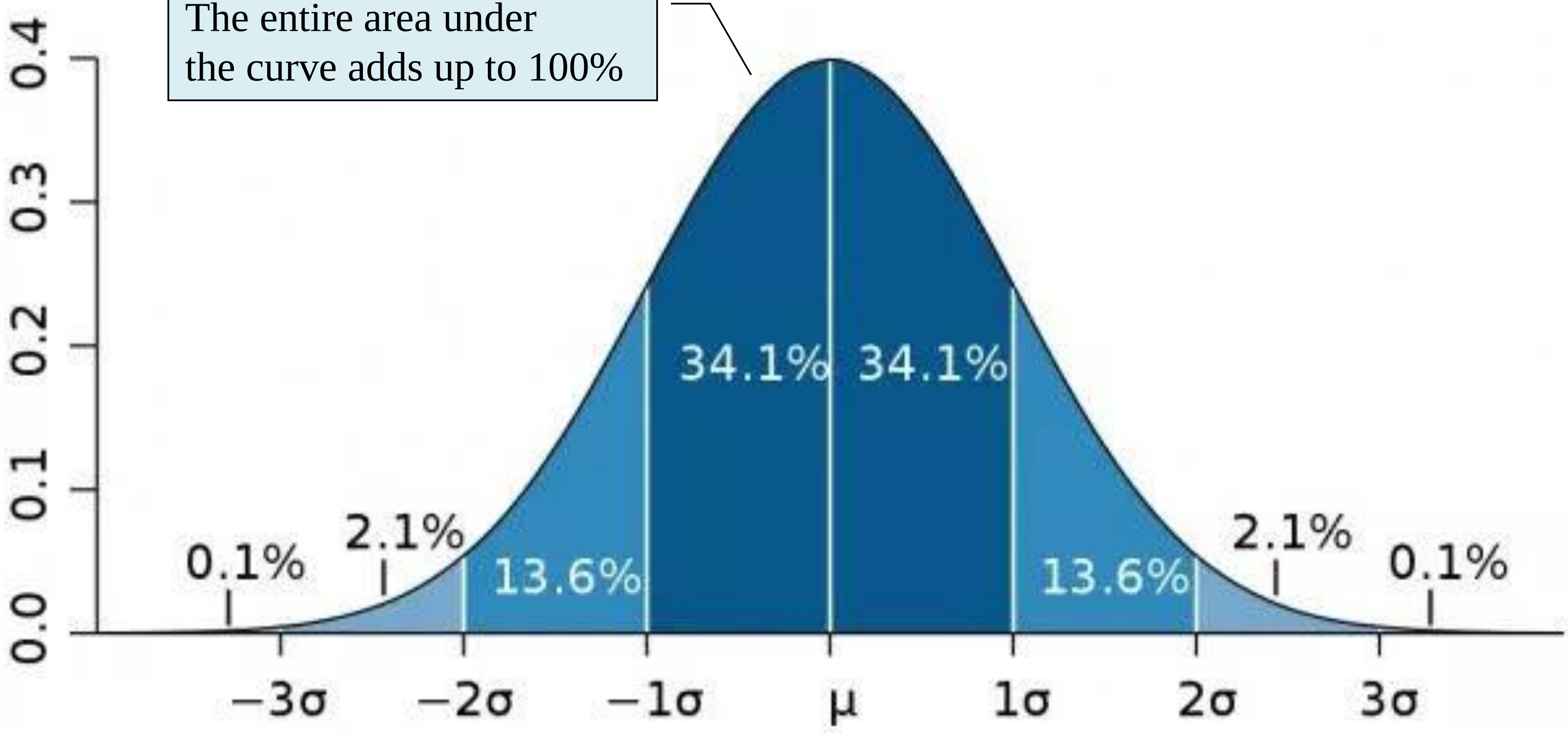
σ_x - population standard deviation

z - confidence coefficient

(Value = 2.00 for 95% and 2.58 for 99%)

Confidence Interval

The entire area under the curve adds up to 100%



Confidence Interval

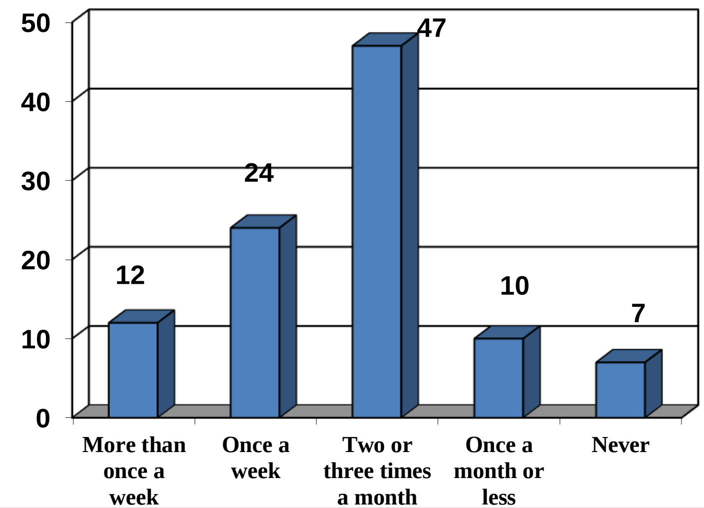
$\bar{X} \pm$ sampling error = interval estimate of μ

95% confidence interval:

$\bar{X} \pm 2\sigma_{\bar{X}} =$ 95% interval estimate of μ

$$-0.24 \pm 2(0.103) = (\underbrace{-0.446, -0.034}_{\text{Confidence Interval}})$$

Coding: -2 -1 0 1 2



Sample Size Dependencies

$\bar{X} \pm Z\, \sigma_{\bar{x}}$ (sampling error)

Or

$\bar{x} \pm z\, \sigma_x / \sqrt{n}$

Hence, $n = Z^2 \sigma^2_x / (\text{sampling error})^2$

Sample Size Depends Upon:

- Confidence level
- Population standard deviation
- Sampling Error

Find anything odd?


HYPOTHESIS
TESTING

Hypothesis Testing

- The general goal of a hypothesis test is to rule out chance (sampling error) as a plausible explanation for the results from a research study.
- Is also called *significance testing*
- Tests a claim about a parameter using evidence (data in a sample)
- Let’s consider a one-sample z test (test used to test means when population variance is known)

Key Terms

A. Null and alternative hypotheses

B. Test statistic

C. P-value and interpretation

Null and Alternative Hypotheses

- Convert the research question to null and alternative hypotheses
- The **null hypothesis (H_0)** is a claim of "no difference in the population"
- The **alternative hypothesis (H_a)** claims "H_0 is false"
- Collect data and seek evidence against H_0 as a way of bolstering H_a (deduction)

Illustrative Example: “Body Weight”

- **The problem:** In the 1970s, 20–29-year-old men in Canada had a mean μ body weight of 170 pounds. Standard deviation σ was 40 pounds. We test whether mean body weight in the population is still 170 pounds.
- **Null hypothesis** $H_{0:}$ $\mu = 170$ (no difference)
- The **alternative hypothesis** can be either $H_{a:}$ $\mu > 170$ (**one-sided test**) or
$H_{a:}$ $\mu \neq 170$ (**two-sided test**)

Test Statistic

This is an example of a one-sample test of a mean when σ is known. The test statistic to solve the problem:

$$z_{\text{stat}} = \frac{\bar{x} - \mu_0}{SE_{\bar{x}}}$$

where $\mu_0 \equiv$ population mean assuming H_0 is true

and $SE_{\bar{x}} =$

z statistic

- For the illustrative example, $\mu_0 = 170$
- We know $\sigma = 40$
- Take a random sample of $n = 64$. Therefore

$$SE_{\bar{x}} =$$

$$SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

- If we found a sample mean of 173, then

$$z_{\text{stat}} =$$

$$z_{\text{stat}} = \frac{\bar{x} - \mu_0}{SE_{\bar{x}}}$$

z statistic

If we found a sample mean of 185 for another sample of 64 , then

$$z_{\text{stat}} = \frac{\bar{x} - \mu_0}{SE_{\bar{x}}}$$

$$z_{\text{stat}} =$$

What about standard error?

$$SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \qquad SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{40}{\sqrt{64}} = 5$$

Why does standard error stay consistent across the two samples?

P-value

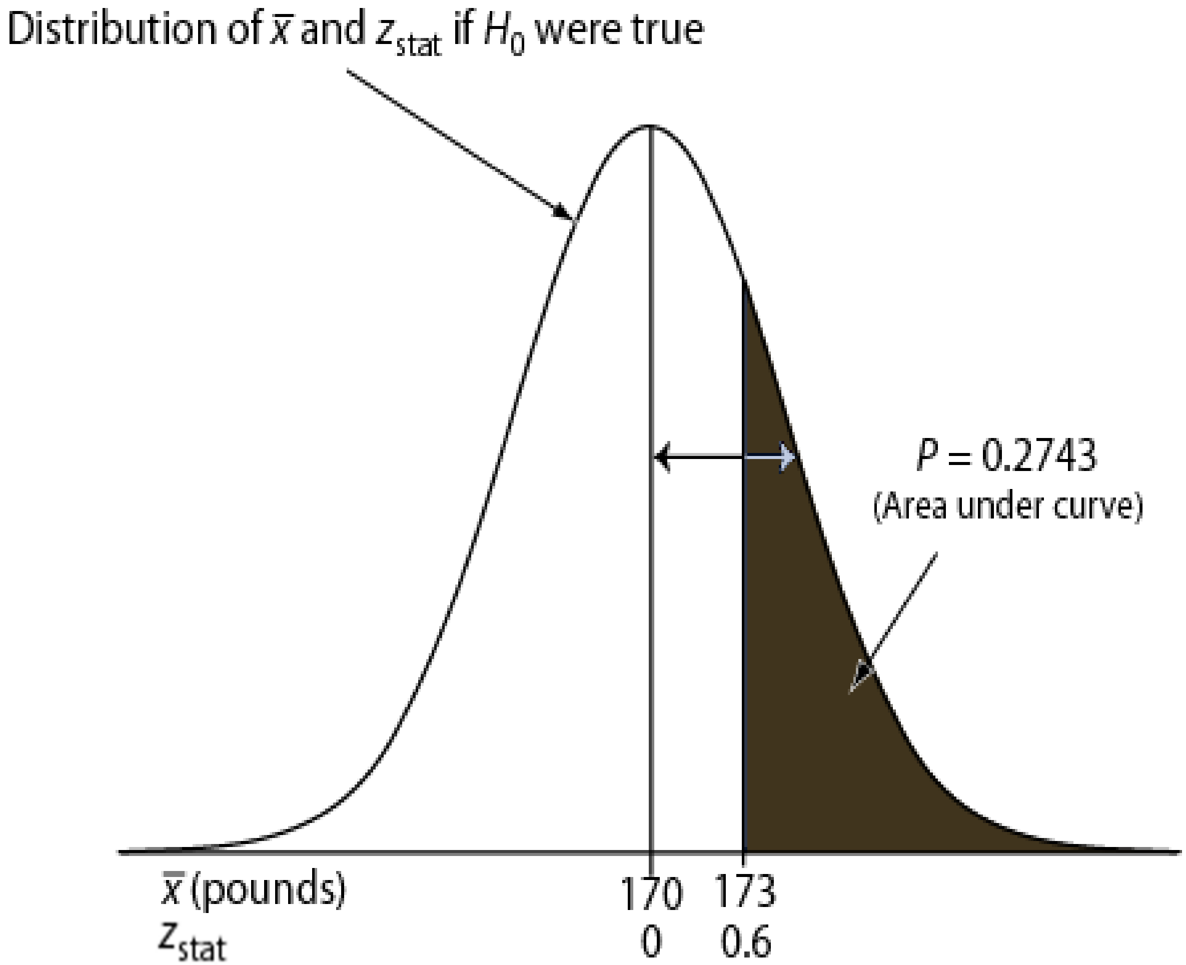
- The *P*-value answers the question: What is the probability of the observed test statistic **when H_0 *is true***?
- This corresponds to the area under the curve in the tail of the Standard Normal distribution beyond the $z_{stat.}$
- Convert *z* statistics to *P*-value :

  For H_a: $\mu > \mu_0 \Rightarrow P$ = Pr(right-tail beyond z_{stat})

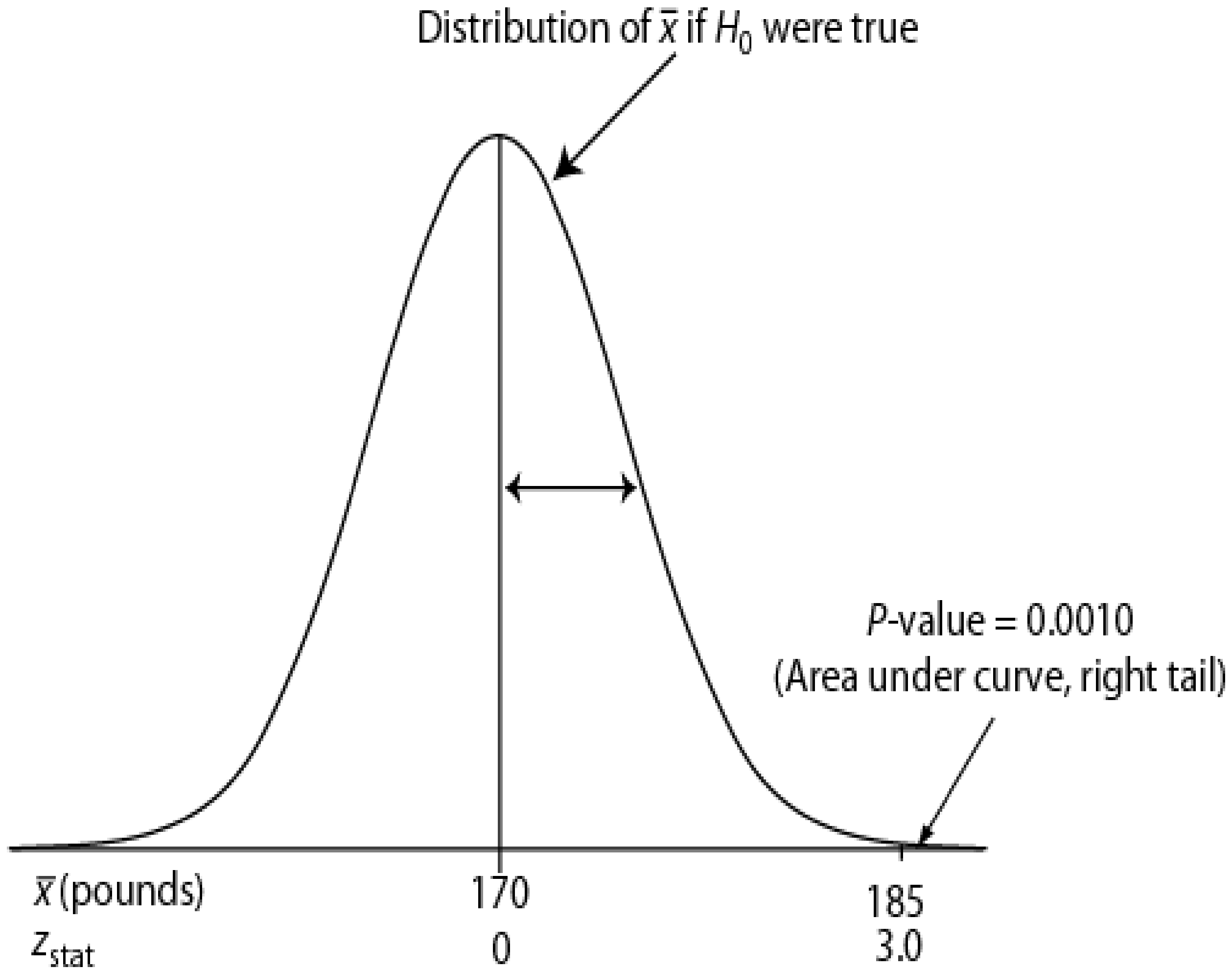
  For H_a: $\mu < \mu_0 \Rightarrow P$ = Pr(left tail beyond z_{stat}) (negative z_{stat})

  For H_a: $\mu \neq \mu_0 \Rightarrow P = 2 \times$ one-tailed *P*-value

One-sided *P*-value for z_{stat} of 0.6

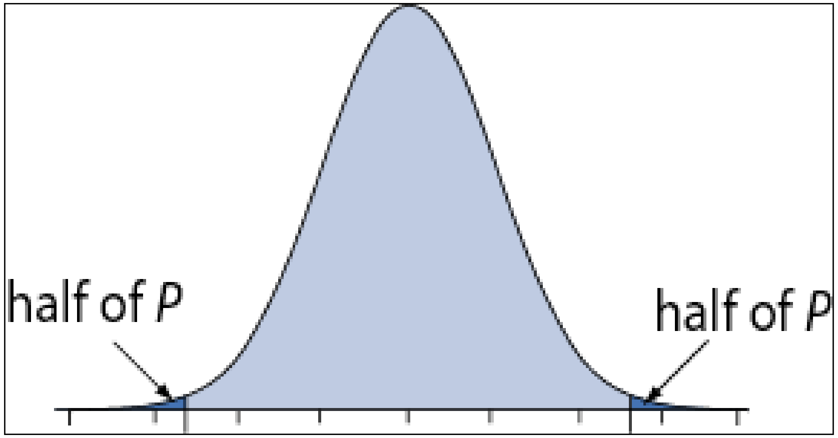


One-sided *P*-value for z_{stat} of 3.0



Two-Sided *P*-Value

- One-sided $H_a \Rightarrow$ AUC in tail beyond z_{stat}
- Two-sided $H_a \Rightarrow$ consider potential deviations in both directions $\Rightarrow$ double the one-sided *P*-value



Examples:

- ❑ If one-sided $P = 0.2743$, then two-sided $P = 2 \times 0.2743 = 0.549$.
- ❑ If one-sided $P = 0.0010$, then two-sided $P = 2 \times 0.0010 = 0.002$.

Interpretation

- *P*-value answer the question: What is the probability of the observed test statistic … **when H_0 *is true*?**
- Thus, smaller and smaller *P*-values provide stronger and stronger evidence against H_0
- Small *P*-value $\Rightarrow$ strong evidence

Interpretation

Conventions*

$P > 0.10 \Rightarrow$ non-significant evidence against H_0

$0.05 < P \leq 0.10 \Rightarrow$ marginally significant evidence

$0.01 < P \leq 0.05 \Rightarrow$ significant evidence against H_0

$P \leq 0.01 \Rightarrow$ highly significant evidence against H_0

Examples

$P = .549 \Rightarrow$ non-significant evidence against H_0

$P = .002 \Rightarrow$ highly significant evidence against H_0

TILL NEXT TIME