SBWL Service & Digital Marketing

Marketing Research

Daniel Winkler

Summer Term 2019

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

EFMD EQUIS ACCREDITED AACSB ACCREDITED ASSOCIATION OF MBAS AMBA ACCREDITED

Welcome to Marketing Research!

Contact

Daniel Winkler
Teaching and Research Associate @ IMSM
Room: D2.1.548
daniel.winkler@wu.ac.at
+43 1 31336 4888

Please feel free to contact me if you have any questions!

Join us!

Facebook

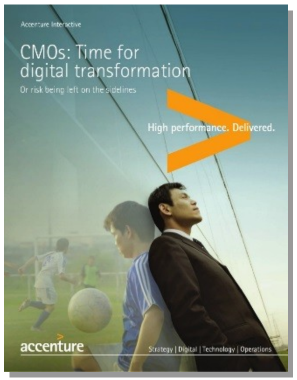
Twitter

Course Goals

After taking this course, you will…

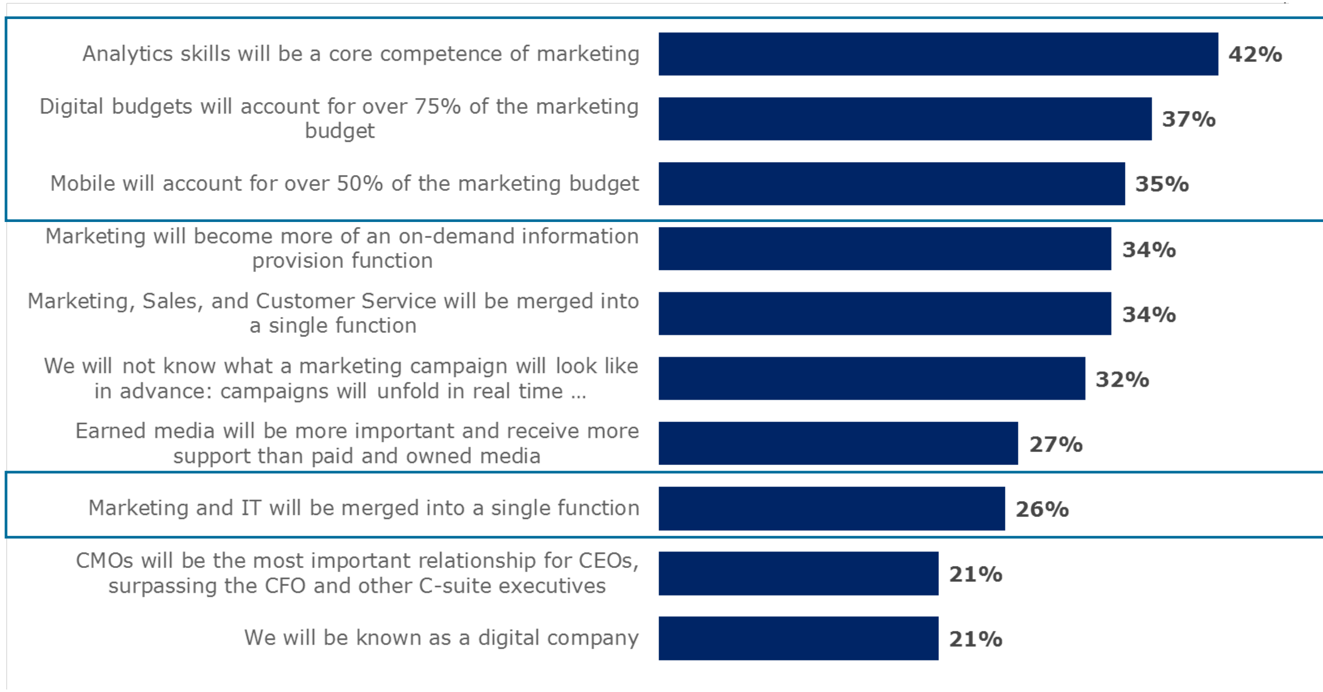
- …know why analytical skills are important
- …know what you can and cannot achieve with data analytics
- …have a good vocabulary for data analysis
- …know when to use which technique of data analysis
- …have confidence in your analyses
- …know how to implement important techniques of data analysis in R
- …be able to conduct your own research projects
- …be able to manually make important calculations
- …not be afraid of numbers & statistics any more (in case you were)
- …feel the desire to learn more about analytics

Analytics skills become more important for marketing managers



Areas of fundamental change for marketing over the next 5 years

% of 581 senior marketers around the world (2014)

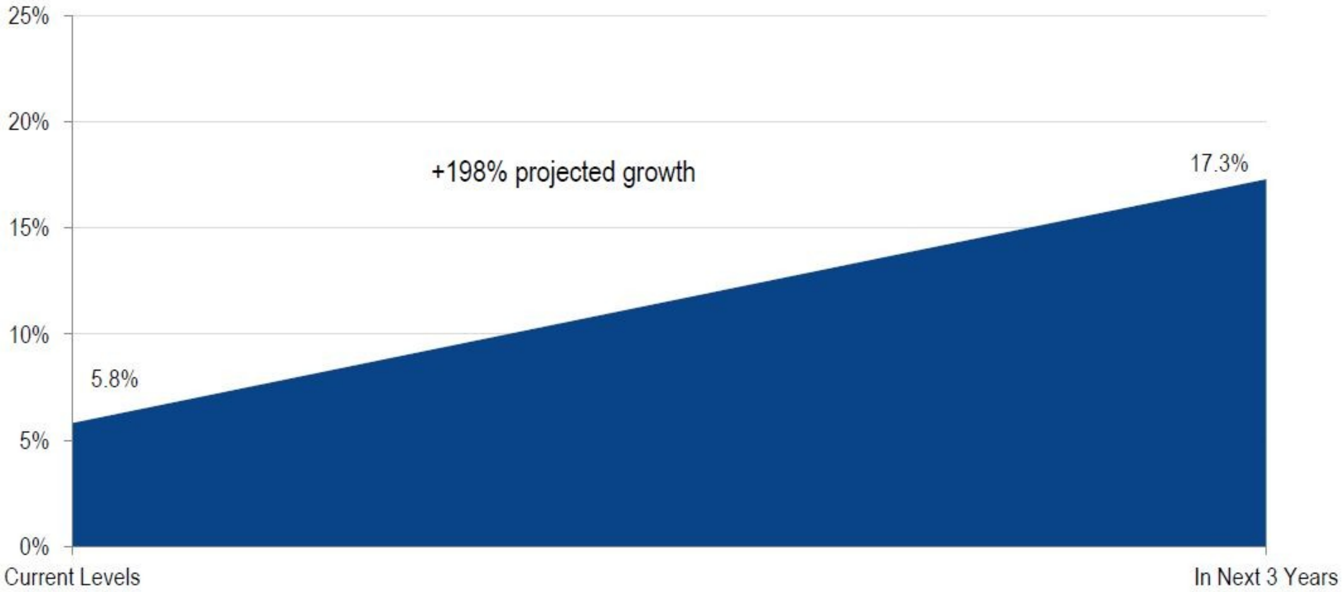


Source: 'CMOs: Time for digital transformation', Accenture Interactive (2014)

Increasing investments in marketing analytics

What percent of your marketing budget do you spend on marketing analytics (…today, … three years from now)?

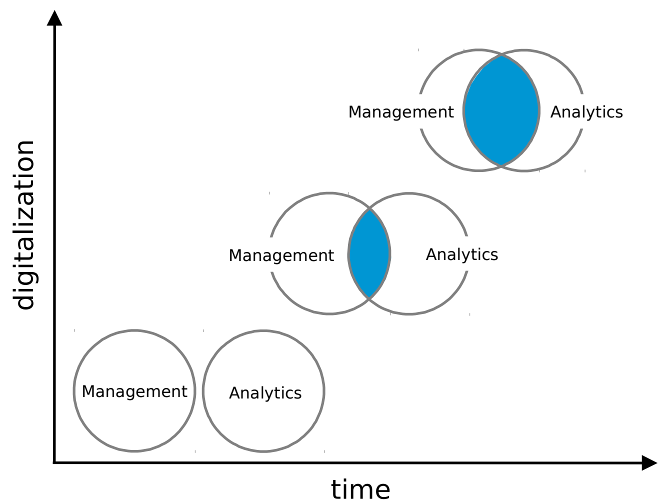
324 senior marketers around the world (2018)



https://cmosurvey.org/wp-content/uploads/sites/15/2018/08/The_CMO_Survey-Highlights_and_Insights_Report-Aug-2018.pdf

Reasons for the increasing relevance of data analytics

"Marketing researchers are becoming more involved in the decision process, whereas marketing managers are becoming more involved with research." (Malhotra 2010, p.44)



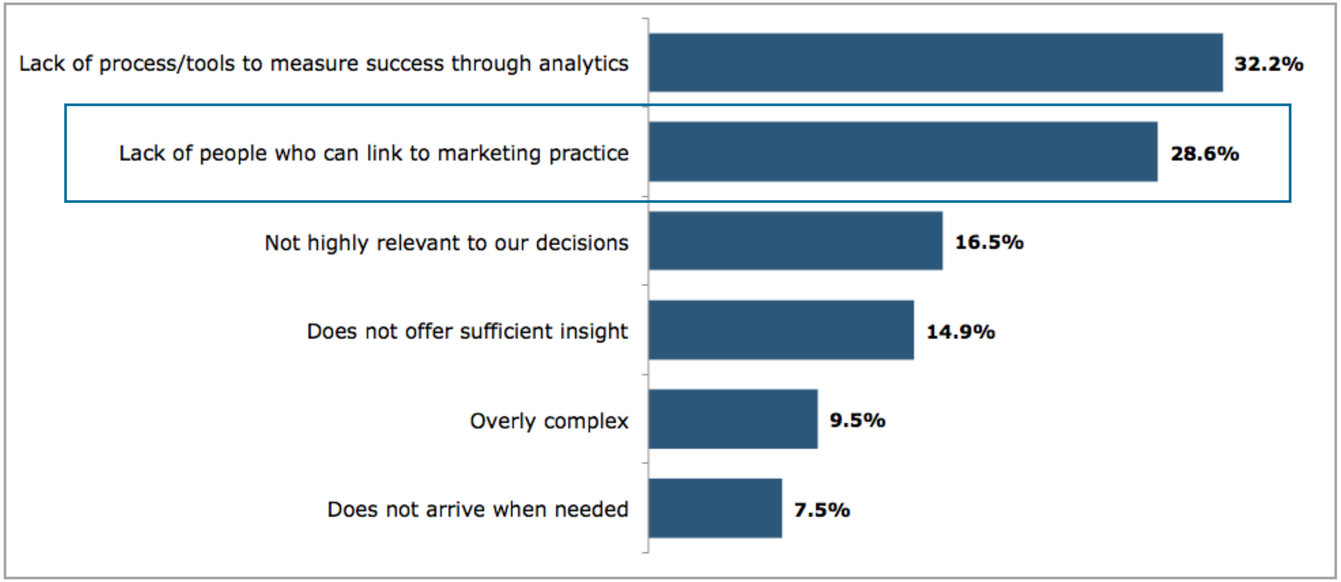
Reasons for this trend:

- Better training of marketing managers
- The Internet and other advances in technology
- Shift in marketing research paradigm in which marketing research is being undertaken on an ongoing basis

Data analytics skills are a valuable asset on the job market

Factors preventing companies from using more marketing analytics

% of 388 top marketers in the US (March 2017)



Source: The CMO survey 2017

Course structure & contents

Theory	Practice
▪ Introduce the concepts and methods necessary for marketing research	▪ Introduce the statistical software necessary to conduct marketing research analyses ▪ Training exercises ▪ Presentations

Course structure & contents

Theory	Practice
Slides & class	Data, R-code, online tutorial

1. Introduction
2. Foundations of inferential statistics
3. The R environment
4. Hypothesis testing
5. Analysis of variance
6. chi-square test
7. Correlation and regression

- DataCamp exercise
- R-code available at Learn@WU
- Online tutorial available at:
  short.wu.ac.at/MRDA

Course dates

Date & Time	Room	Content	Reading (Script)
04.03.2019 08:30 AM – 01:00 PM	D2.0.330	• Introduction • Statistical inference	Getting Started (1.) Summarizing d. (3.1-3.2) Intro. Stat. Inference (4.)
11.03.2019 08:30 AM – 01:00 PM	**D4.0.047**	• Introduction to R • Hypothesis testing	Hypothesis t. (5.1 – 5.3.1)
18.03.2019 08:30 AM – 01:00 PM	D2.0.330	• ANOVA • χ-square test	ANOVA (5.4.3.1) Chi-sq. test (5.6.2)
25.03.2019 08:30 AM – 01:00 PM	D2.0.330	• Correlation • Regression (1)	Correlation (6.1) Regression (6.2-6.2.1)
01.04.2019 08:30 AM – 01:00 PM	D2.0.330	• Regression (2)	Regression (6.2.2, 6.4)
29.04.2019 **09:00 AM** – 01:00 PM	**TC.2.01**	**Exam**	

Reading

If anything is unclear in the online script please contact me!

The script was developed for YOU so its only effective if its clear to you and we are actively working on improvements.

Since it was originally stated as required literature in the syllabus you can also use Andy Field's Discovering Statistics using R!

- The suggested chapters are:
  - 1-7
  - 9-10
  - 18
- However, this is **NOT** required to master the exam

Grading

Component	Weight
DataCamp Exercises	4 x 2.5% = 10%
Assignments (best 3 count): 1) Statistical Inference 2) Introduction to R & Hypothesis testing 3) ANOVA & χ-square test 4) Correlation and Regression	3 x 10% = 30%
Exam (min. 30%)	60%

- Overall minimum for a passing grade is 60%
- Please solve the assignments individually (have a look at WU guidelines on plagiarism - they apply to all coursework!)

Schedule

- This course is worth 4 ECTS-credits
- Which means 4x25h = 100h of work
- ~25h you will spend in class
- The exam is 56 days (or 40 weekdays) from today
- So you should spend 75/56 ~ 1h 20min per day (75/40 ~ 1h 50min per Weekday) working on this course outside of class
- I spend this much time preparing & teaching
- So I ask you to spend this much time learning about marketing research!

Introduction

- **What is data?**
- **Categorical data**
- **Continuous data**

The marketing research process

Problem definition
What is the problem?

- Identify and clarify information needs
- Identify problems that can be solved with information from market research
- Define research problem, questions, and objectives
- Confirm information value

Research design
How can we solve the problem?

Decide on:
- Information needed
- Data sources (primary vs. secondary)
- Research approaches (qual. vs. quant.)
- Measurement & scaling procedures
- Questionnaire/ experimental design
- Sampling plan
- Plan of data analysis

Data collection
How can we obtain the required data?

- Collect data according to the research design (survey, observation, experiment)
- Internal execution or employ an external firm
- Definition of timelines (in particular for external service providers)

Analyze data
What are the results & implications?

- Analyze data statistically or subjectively
- Definition of the final products of data analysis (in particular for external service providers)
- Infer answers and implications

Report findings
What are the implications?

- Identify recipients
- Define the type of result processing (e.g., presentation, report)
- Define communication contents
- Identify the level of detail (e.g., methods)
- Formulate conclusions
- Prepare finalized report
- Monitoring of the use of results

Levels of measurement

Categorical (Non-metric | Qualitative)

- **Nominal** (e.g., ID numbers, eye color)
- **Ordinal** (e.g., education, height {tall, medium, short})

Continuous (Metric | Quantitative)

- **Interval** (e.g., calendar dates, temperature in Celsius and Fahrenheit)
- **Ratio** (e.g., age, mass, length, temperature in Kelvin)

Levels of measurement: Categorical (non-metric) variables

Nominal

- Numbers only serve as labels for identification and categorization
- Numbers do not reflect the amount of the characteristic possessed by the objects
- Called “binary” for two categories
- Only permissible operation is counting

➢ E.g., Starting numbers in a race

7 11 3

Ordinal

- Numbers indicate the relative position of objects
- But not the magnitude of difference between them
- Besides counting, less than/greater than relations are possible
- Also statistics based on centiles, e.g. percentile, quartile, median

➢ E.g., Order of boats in finish

1st place 2nd place 3rd place

http://wwwa.jura.uni-tuebingen.de/~s-krp2/Rennen2013/CIMG9891.htm

Levels of measurement: Continuous (metric) variables

Interval

- Differences between objects can be compared (equal intervals)
- But zero point is arbitrary
- Not meaningful to take ratios of scale values
- In addition, statistics such as range, mean, and standard deviation can be computed

➢ E.g., Performance rating on a 0 to 10 scale

9.6

8.4

4.2

Ratio

- Possesses all properties of nominal, ordinal and interval scales
- Has an absolute zero point
- Meaningful to compute ratios of scale values
- All statistical techniques can be applied to ratio data

➢ E.g., Time to finish in minutes

7.1

14.2

15.2

http://wwwa.jura.uni-tuebingen.de/~s-krp2/Rennen2013/CIMG9891.htm

Scales types and permissible statistics

Scale	Level of information [allowed operators]	Marketing example	Permissible statistics: Descriptive	Permissible statistics: Inferential
Nominal	Description (counting) [=,≠]	Gender, classification of retail outlet types	Cell count, mode, percentages	**Chi-square test**, binomial test
Ordinal	Order among categories (ranking) [=,≠,<,>]	Rank order of favorite TV program	Percentile, median	Rank order correlation, non-parametric tests (e.g., Wilcoxon)
Interval	Distance (equal intervals) [=,≠,<,>,+,-]	Attitudes, opinions	Mean, standard deviation	**Correlations**, parametric tests (e.g., ***t*-tests**), **ANOVA**, **regression**, factor analysis
Ratio	Origin (meaningful zero) [=,≠,<,>,+,-,*,/]	Income, sales, market share, willingness-to-pay		

Categorical (Nominal, Ordinal) · Continuous (Interval, Ratio)

- Permissible inferential statistics depend on whether the scale is used as the dependent or as the independent variable
- For a detailed overview see: https://stats.idre.ucla.edu/other/mult-pkg/whatstat/

Introduction

- ✓ What is data?
- ➔ Categorical data
- Continuous data

Frequency Distribution – The Data

```
> test_data[1:30, c("ipaddress", "experience", "overall_knowledge", "group")]
         ipaddress experience overall_knowledge group
1   188.23.190.194          3                 2     1
2      93.10.250.2          3                 2     1
3       91.0.18.86          3                 2     2
4     37.116.255.55         1                 3     1
5  178.165.131.208          1                 3     1
6      212.28.68.81         3                 1     1
7       83.7.80.55          1                 3     1
8  213.147.160.138          3                 1     2
9      80.110.91.14         3                 2     1
10     93.229.85.17         1                 2     1
11 178.115.131.225          1                 2     2
12    62.178.118.68         1                 2     2
13     62.46.39.117         1                 2     1
14    92.231.231.64         2                 2     2
15    77.119.129.58         4                 2     2
16    80.110.92.227         1                 2     1
17      93.44.84.39         1                 2     1
18   121.102.95.180         3                 2     2
19     89.104.11.34         3                 2     1
20    80.109.54.154         1                 1     2
21    82.218.163.42         1                 2     1
22    88.128.80.105         1                 4     2
23    213.225.0.210         1                 2     1
24   90.110.104.106         4                 2     1
25     213.225.8.16         1                 2     2
26     62.46.232.14         1                 2     2
27   80.110.104.238         3                 2     2
28   178.190.73.145         1                 2     1
29       91.141.0.1         1                 2     1
30    46.125.250.96         2                 4     2
```

- Arranged in a grid ("Matrix")
- Each row is an observation
- Each column a variable ("vector")
- Value in cell [i,j] is the answer to the i-th question given by the j-th participant of the survey

  e.g. test_data[6, 3] = 1

  test_data[1, 2] = 3
- In R we can use both indices and names of rows/columns
- Matrices have 2 indices and vectors have one index

  e.g. x[i]: i-th element of vector x

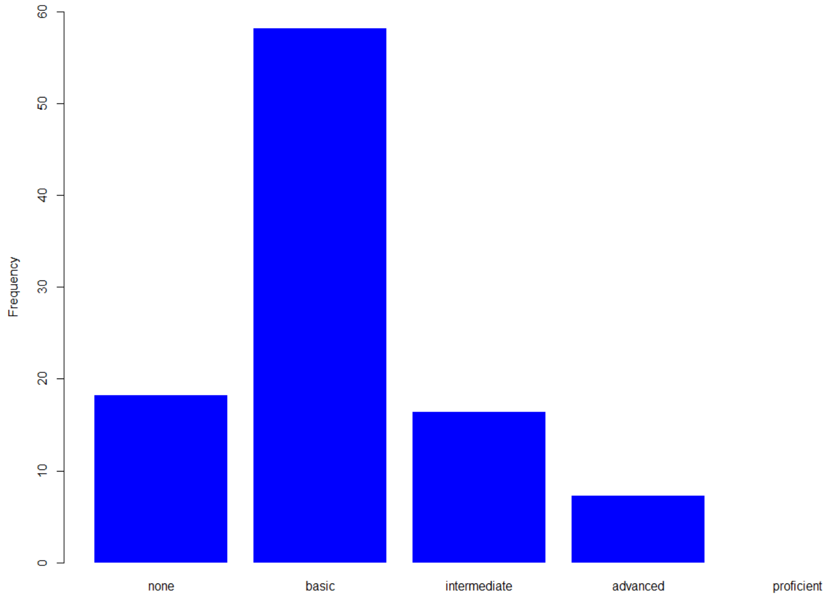
Frequency distribution

Value label	Value	Frequency (N)	Percentage	Cumulative percentage
None	1	10	18.18	18.18
Basic	2	32	58.18	76.36
Intermediate	3	9	16.36	92.73
Advanced	4	4	7.27	100.00
Proficient	5	0	0	100.00
Total		55	100.0	

The frequency distribution shows how many time each score occurs

Bar charts for categorical data

Value label	Value	Frequency (N)
None	1	10
Basic	2	32
Intermediate	3	9
Advanced	4	4
Proficient	5	0
Total		55



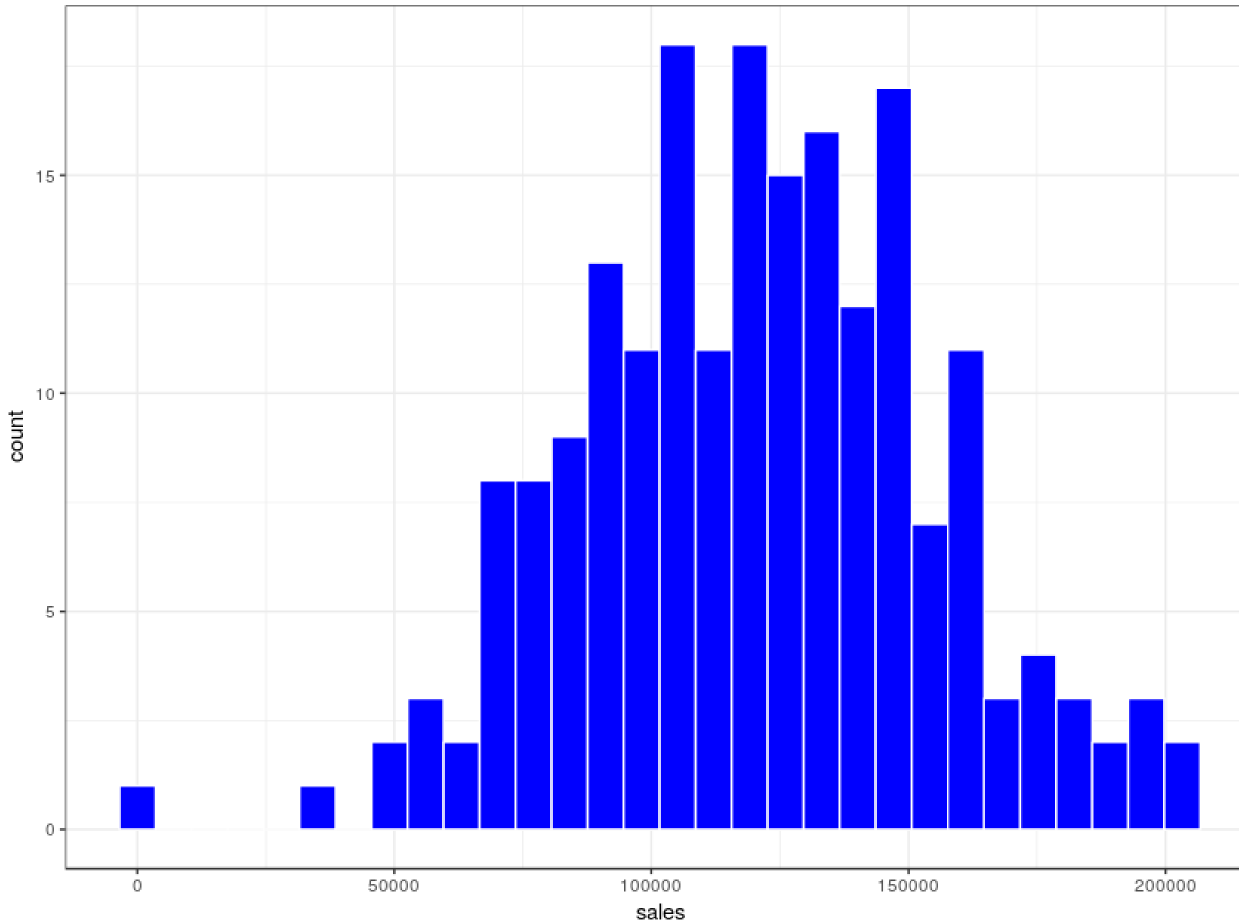
Continuous variables

Introduction

- ✓ What is data?
- ✓ Categorical data
- ➔ Continuous data

Visualization in a histogram

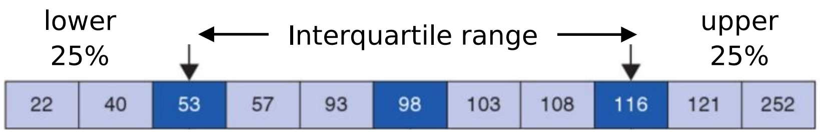
```
> head(adv_data, 30)
    sales advertising
1  133553        1841
2  119778        1782
3  147097        1782
4   55589         573
5  109886        1877
6  150285        1907
7  108779        1719
8  142993        2069
9   87877         938
10 106663        1544
11 105976        1096
12  96569        1087
13 128616        1853
14 104664        1072
15 122913        1263
16 125959        1481
17 140737        1743
18 126234        1405
19  95547        1337
20 163220        2126
21 123908        1558
22 117765        1344
23 117213        1391
24 125994        2299
25  96809        1433
26  82472         981
27 164035        1962
28 136468        1707
29 132531        1588
30  87896         412
```



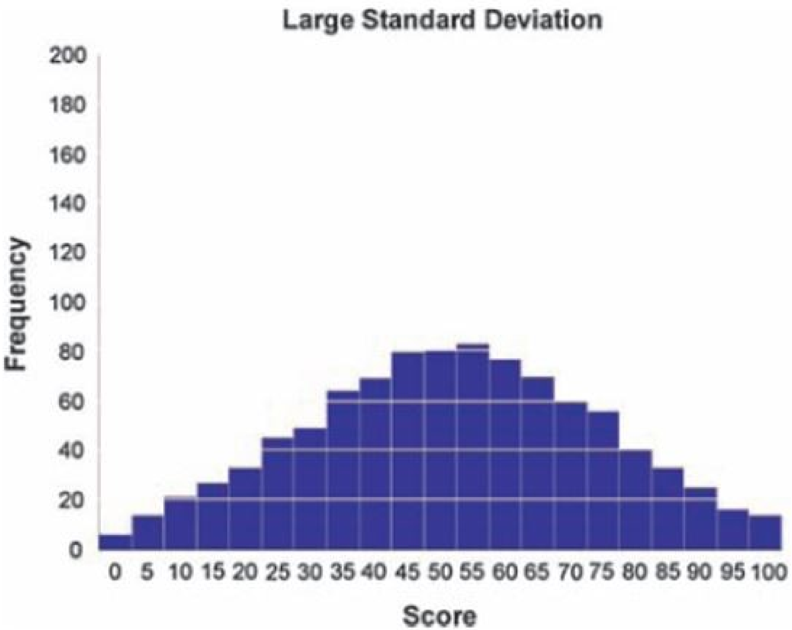
Measures of location

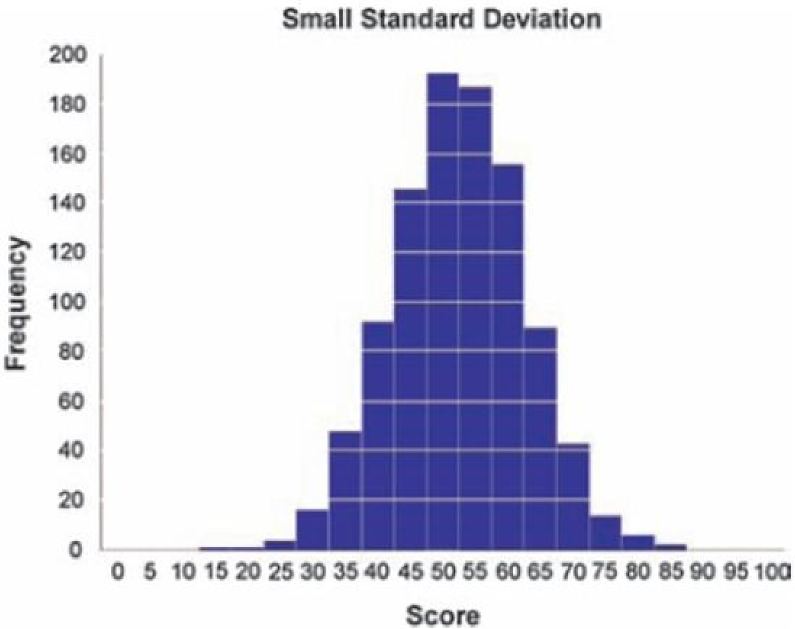
Statistic	Description	Definition
Mean	• The average • Sum up all elements and divide by the number of elements	$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ X_i : Observed value of the variable X n : number of observations
Mode	• Value that occurs most frequently • Highest peak of the distribution	
Median	• Middle value when the data are arranged in ascending or descending order • 50th percentile	

Measures of dispersion

Statistic	Description	Definition
Range	▪ The difference between the largest and smallest values in the sample	$Range = X_{largest} - X_{smallest}$
Interquartile range	▪ The range of the middle 50% of scores	
Variance	▪ The mean squared deviation of all the values of the mean	$s^2 = \frac{1}{n-1} * \sum_{i=1}^{n} (X_i - \bar{X})^2$
Standard deviation	▪ The square root of the variance	$s_x = \sqrt{s^2}$

Measures of dispersion





In-class exercise

A sample of 10 adults was asked to report the number of hours they spent on the Internet the previous month.

Results are given:

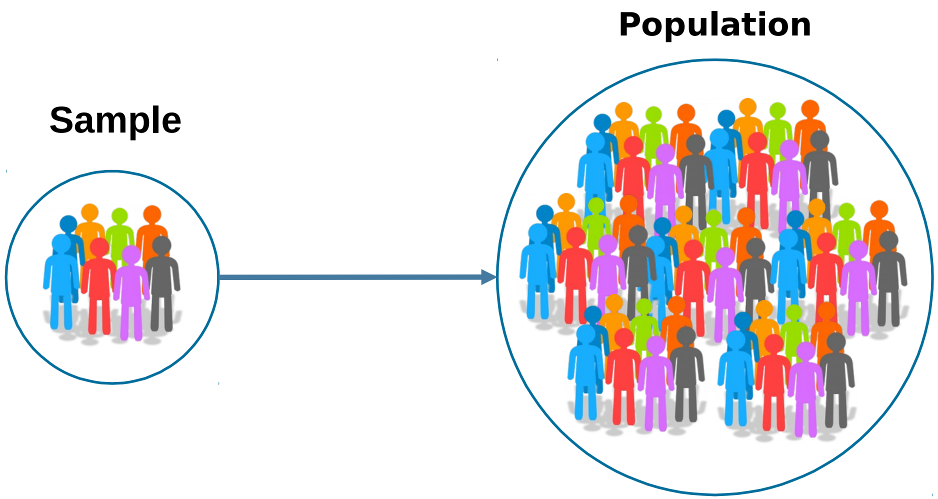
Observations	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Time spent on Internet	0	7	12	0	33	14	8	0	9	**22**

Statistical Inference

- ➔ Samples vs. Population
- Randomness & Probability
- Confidence Intervals

Statistical inference

- One important goal in marketing is to calculate statistics (e.g., mean, proportion) and use them to **estimate** the corresponding **true population values**.
- **Statistical inference** is the process of **generalizing** the sample results **to the population** results.



Samples vs. Populations

Variable	Sample statistic	Population parameter
Size	n	N
Mean	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$	$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$
Variance	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$	$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$
Standard deviation	$s = \sqrt{s^2}$	$\sigma = \sqrt{\sigma^2}$
Standard error	$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

Samples vs. populations

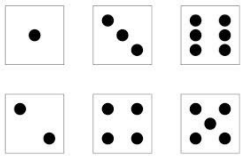
- **Sample statistics**
  - Mean and SD describe only the sample from which they were calculated
- **Population parameters**
  - Mean and SD are intended to describe the entire population (very rare in marketing)
- **Sample to Population**
  - Mean and SD are obtained from a sample, but are used to estimate the mean and SD of the population (very common in marketing)

Statistical Inference

- ✓ Samples vs. Population
- ➔ Randomness & Probability
- Confidence Intervals

Random =
An event with an unknown outcome

- Even though you may know all **possible** outcomes for an event (e.g., dice rolls: 1 through 6, coins: heads or tails), you never know which of those outcomes will occur
- **Random** = An event with an unknown outcome
- Lots of random events over time = **Probability**
- Probability is the **long-run relative frequency** that an event occurs
- Outcomes from a multitude of random events **will converge on some expected value** - even when the individual outcomes are random (**Law of Large Numbers**)



Radziwill, N.M. (2015): *Statistics (The Easier Way) With R*, Lapis Lucera.

Random Variables & Probability Distributions

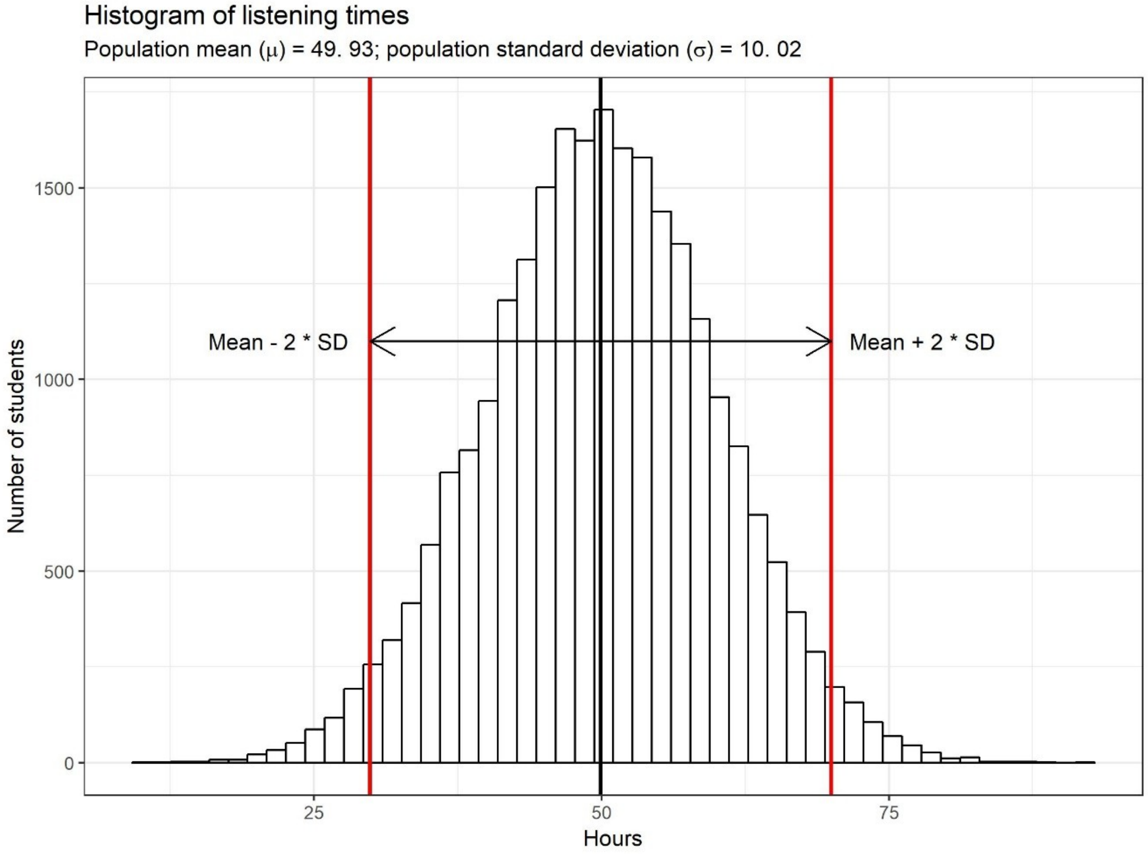
- **Random Variable**: Outcome we express as a number
  - e.g. Number of heads when flipping a coin
  - e.g. Avg. Music listening time per week
- **Probability distribution**:

  Assigns probability to each possible value of a random variable
  - e.g. p(heads) = 50% and p(tails) = 50% for a fair coin
  - e.g. p(listening hours <10) = 70% and p(listening hours ≧ 10) = 30%
- Many random variables (approx.) follow **known** probability distributions
  - e.g. coin toss → binomial distribution
  - e.g. listening time → gamma distribution (?)
    - Technically gamma is defined from 0 to infinity but infinity hours?
- If you are very interested

An example:

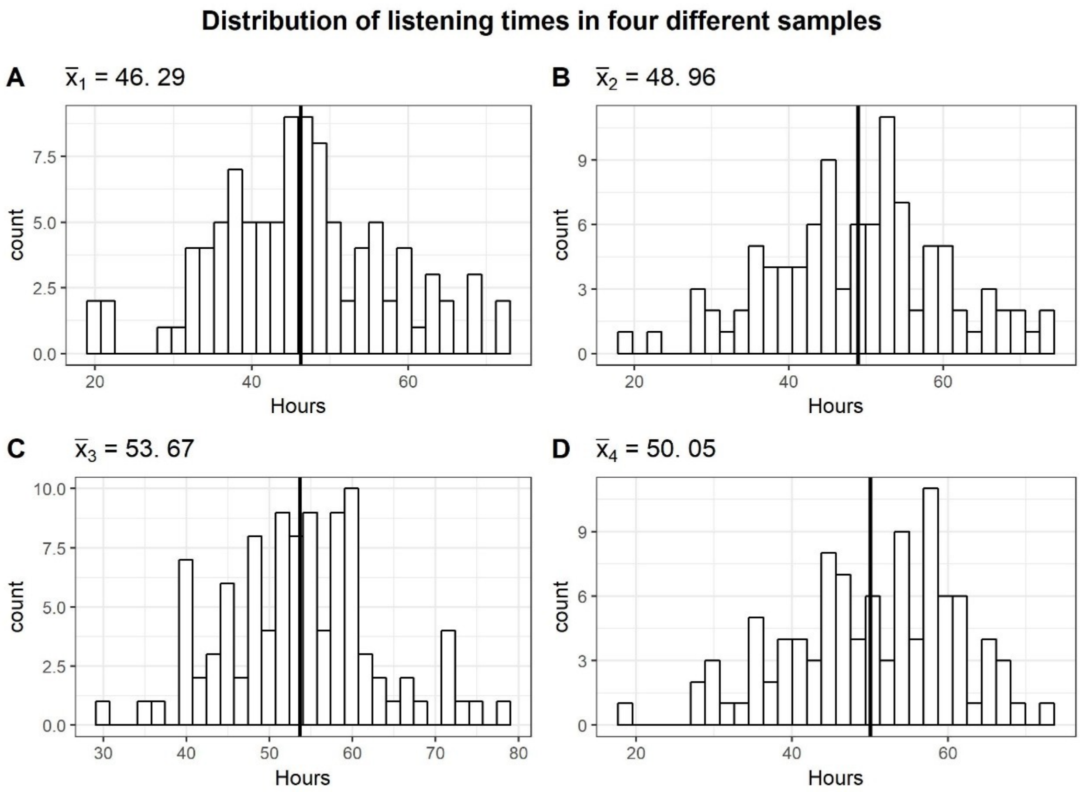
You own a marketing firm in Vienna specialized in online media. A new streaming service, Pear Music, wants to enter the Austrian market. In order to do so they want to assess the market of cool kids so naturally they look at WU. Initially they just want to assess the market and know **how much music is consumed at WU?**

The population



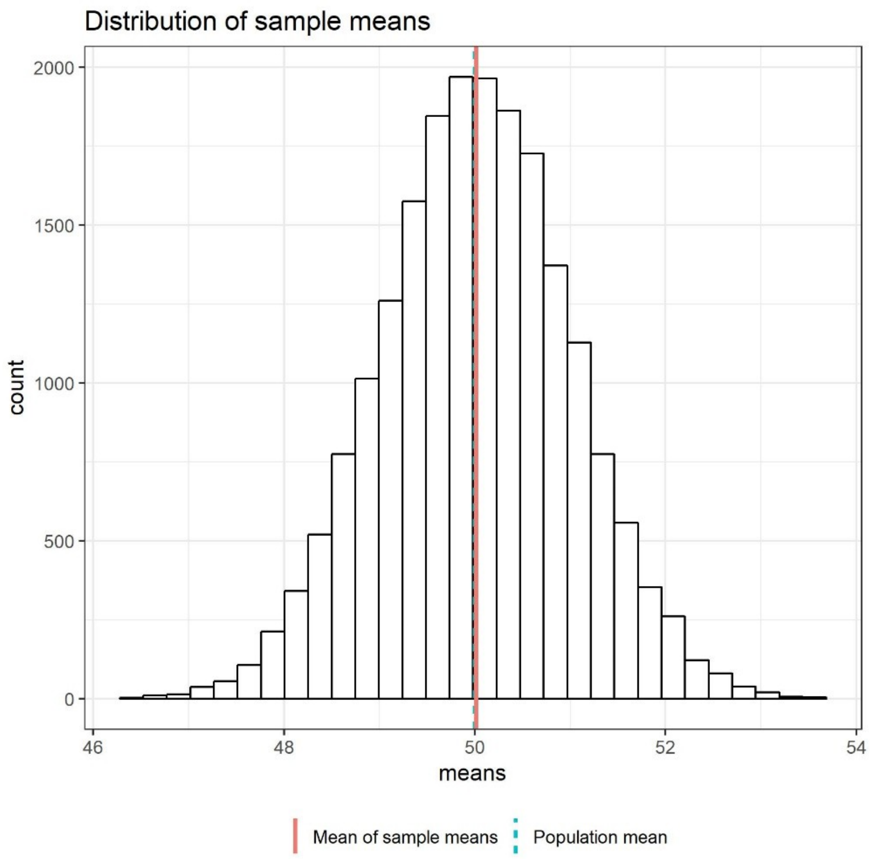
- Example: music listening times of students
- Assume we have information of every student (population)
- The population mean and standard deviation are known

Taking samples from the population



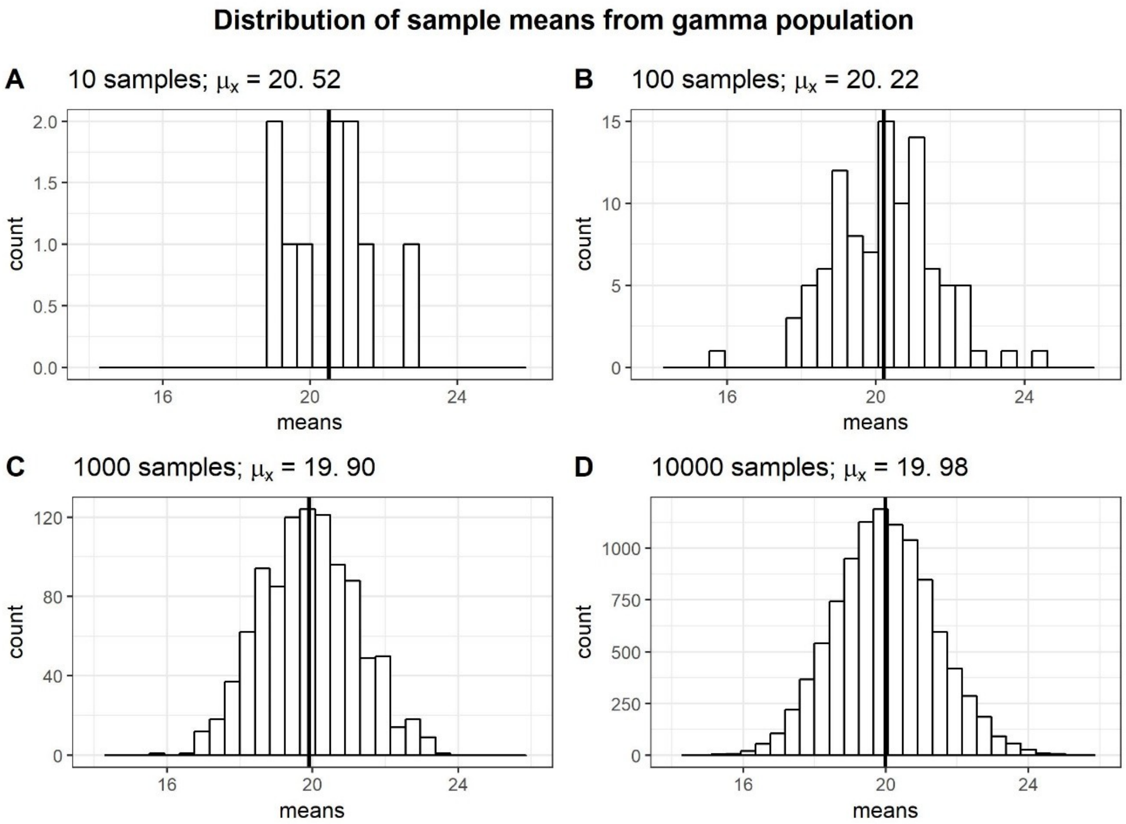
- In reality, we only have access to one sample (say, 100 students)
- We use this sample to generalize to the population
- However, each sample will be different, so there is uncertainty about how close our sample statistic actually is to the population parameter of interest

Sampling distribution



- Assume that we would be able to **repeatedly take random samples,** consisting of 100 students each, from the population
- When we plot a histogram of all sample means that we would get, this tells us something about the **distribution of means that we could potentially get**
- The mean of this distribution is (in the limit) the same as the population mean
- The **distribution of sample means** follows a **normal distribution**
- It follows that the sample mean is a good estimate of the population mean

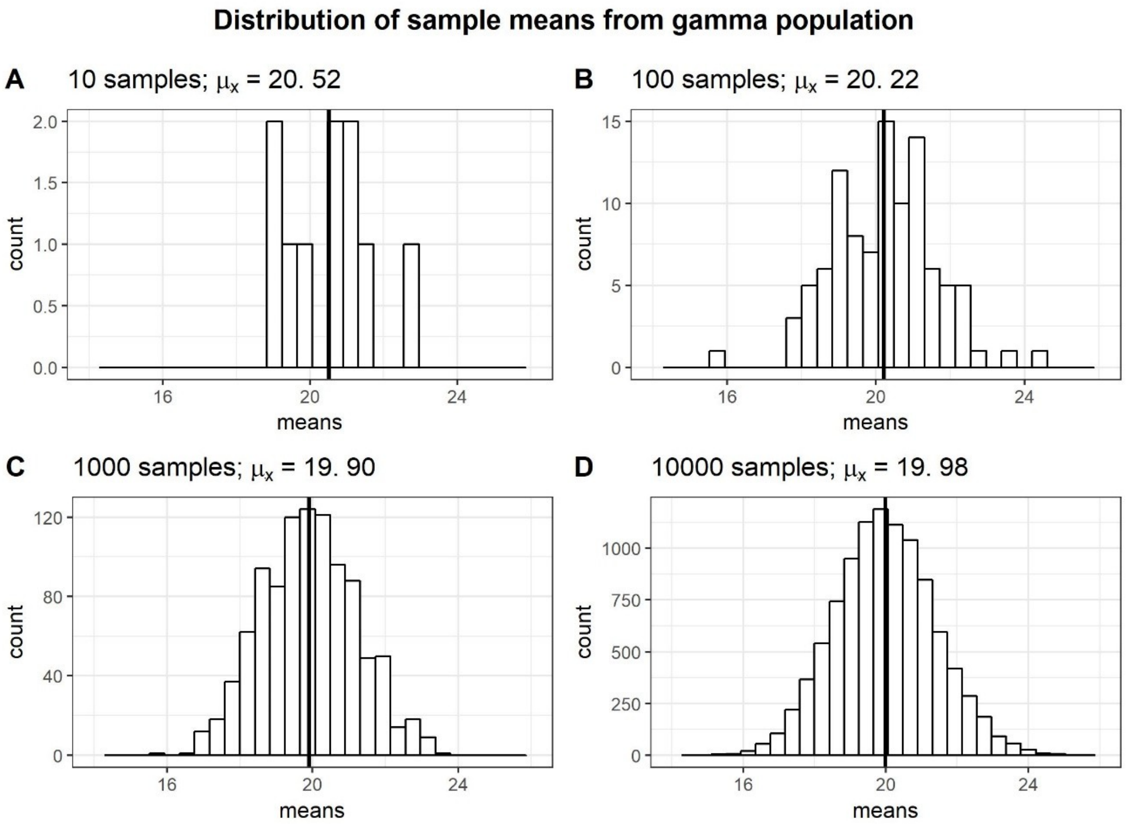
Central limit theorem



Central Limit Theorem: the distribution of a sample mean will be approximately normal, provided the sample size is sufficiently large (e.g., >40)

#DEMO

Standard error of the mean



- Again, the standard error would tell us something about the precision of our estimate

$$\boxed{\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}}$$

$$\sigma_{\bar{x}} = \frac{14.15}{\sqrt{100}} = 1.40$$

Two types of variation

Types of variation

Natural variation that can be observed (e.g., product satisfaction)

Standard deviation (s) is used to describe the variation in a sample

Data may follow a **non-normal** distribution

Variation of study results, when a study is repeated (<u>hypothetically!</u>)

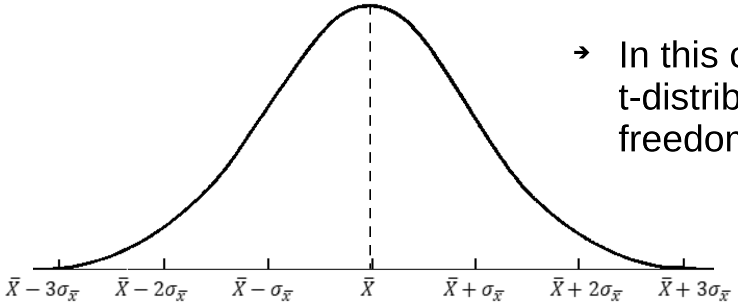
Standard error $(\sigma_{\bar{x}})$ is used to describe the variation of study results if **hypothetically** you were to repeat the study many times

Variation often follows a **normal distribution**, even if data from each individual study are non-normal

Using what we actually know

- So far we have assumed to know the population standard deviation (σ)
- This an unrealistic assumption since we do not know the entire population
- The best guess for the population standard deviation we have is the sample standard deviation (s)
- Thus, the **standard error** ($\sigma_{\dot{x}}$) of the mean is usually estimated from the sample standard deviation:

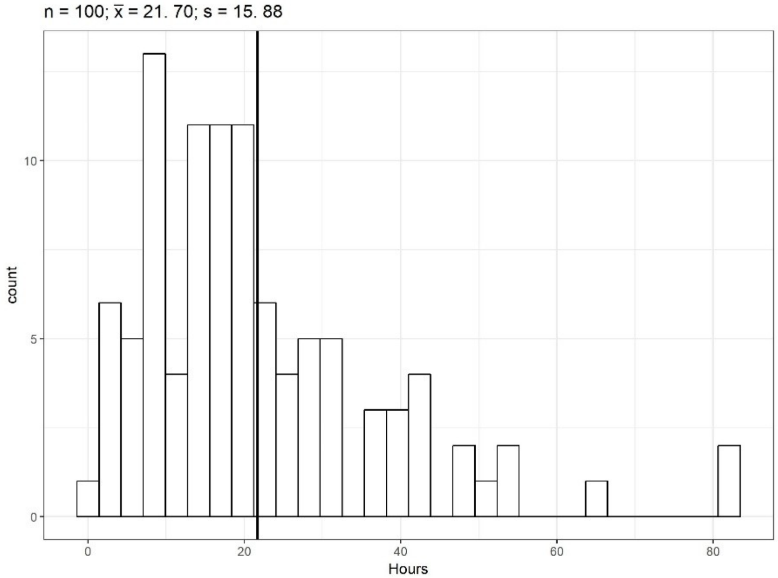
$$\boxed{\sigma_{\bar{x}} \approx SE_{\bar{x}} = \frac{s}{\sqrt{n}}}$$



→ In this case we use the t-distribution with n-1 degrees of freedom to account for uncertainty of s

Question

Natural variation
(random sample)



$\bar{X} = 21.70,\ s = 15.88$

- Say, you wanted to **guess the listening time** of a new student
- **What would be your guess?**
- **How certain are you?**

Statistical Inference

- ✓ Samples vs. Population
- ✓ Randomness & Probability
- ➔ <u>Confidence Intervals</u>

Point estimate vs. interval estimate

- We measure **statistics** from samples in order to determine **parameters** of populations
- The average value from your sample only provides a **guess** of what the real population parameter is
- The next time you collect the **same size sample**, you could get a different average (that’s perfectly normal → **sampling variation**)
- **Point estimate**: Report the average from our sample:The average listening time is 19.79 hours
- **Interval estimate**: We are 95% confident that the true listening time is between 18.59 hours and 24.81 hours.

The Big Question

- **Given our (one) sample** and what we just learned about the theoretical distribution of sample means, **within which range will the population mean likely be?**
- We need:
  - ✔ An estimate for the population mean → Sample mean
  - – A margin of error for the estimate that indicates the range
    - We have:
      - Sample standard deviation (s)
      - Sample size (n)
- Using the t-distribution we can assign the probability of observing a certain range of values
- The range of values around the sample mean that contains the population mean most likely depends on

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

Thanks to the Central Limit Theorem we know that **all sample means combined follow a t-distribution** (Normal distribution with sample estimate for standard deviation)

Confidence Interval

- What does **most likely** mean?
  - By convention 95% → if we choose this range, in 95% of the samples the population mean is captured within it.
  - But you can argue for a different value e.g. 99% (we will discuss the choices later when talking about hypothesis testing)
  - I personally have never seen serious work with a confidence interval below 90%

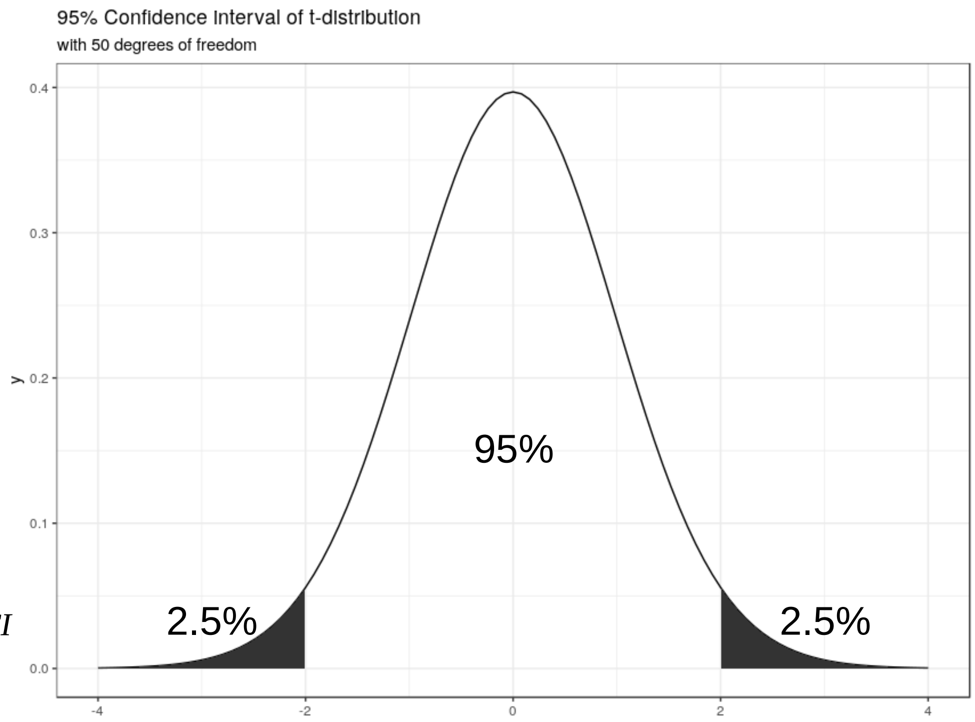
Confidence Interval

- The area under any probability density integrates (really just a fancy sum) to 1
- The graph on the right is scaled s.t. mean = 0 and SE = 1
- That is useful to be able to use the same distribution to calculate the CI all the time
- We just scale it to our sample using mean and SE

$$\bar{x} \pm t \times SE_{\bar{x}} \rightarrow t \approx 2 \text{ for } 95\% \, CI$$

- In R use e.g.
  `qt(c(0.025,0.975), n-1)`
  or go to:

http://www.sjsu.edu/faculty/gerstman/StatPrimer/t-table.pdf



Interpretation of confidence intervals

- For a certain percentage of times (e.g., 95%), the true value of the population mean will fall within these limits (⇒ confidence level).

> “If we’d collected 100 samples, calculated the mean and then calculated a confidence interval for that mean, then for 95 of these samples, the confidence intervals we constructed would contain the true value of the mean in the population.”

Field, A. et al. (2012). *Discovering Statistics Using R*. Sage.

http://rpsychologist.com/d3/CI/