Statistics Introduction

Rohit Singh Chauhan

Book

- Statistical Sleuth 3rd Edition
- R has a package for this book, called Sleuth3. Contains the datasets for examples and exercises in the book.
- Location for datasets

  Calance Vision Team > Data Science channel > Files section > DataScienceLearning folder > Statistics > Datasets_Sleuth3.zip

  R Code to extract data

```
install.packages('Sleuth3')
library('Sleuth3')
install.packages('vcdExtra')
Sleuth3_datasets <- vcdExtra::datasets("Sleuth3")
# setwd('C:\\Users\\RohitSinghChauhan\\Downloads\\Study\\Statistics\\StatisticalSleuth\\Datasets\\')
setwd('path\\to\\data\\directory')

Sleuth3_datasets$Write_csv <- 'write.csv('
Sleuth3_datasets$comma_with_single_quote <- ",'"
Sleuth3_datasets$df_name_with_csv <- ".csv')"

Sleuth3_datasets$full_write_csv <- paste(Sleuth3_datasets$Write_csv, Sleuth3_datasets$Item, Sleuth3_datasets$comma_with_single_quote,
Sleuth3_datasets$Item,Sleuth3_datasets$df_name_with_csv, sep="")

write.table(Sleuth3_datasets$full_write_csv, file = "full_write.R", sep = "\n",
row.names = FALSE, col.names = FALSE, quote = FALSE)
```

Objective of this lecture

- Understand statistical way of thinking
- Drawing statistical inference from studies
- Study design basics – random sampling and randomization
- Focusing on two sample problems – mostly around t-test
- Example studies – Motivation study and sex-discrimination study

Example of Randomized experiment – Motivation and Creativity study

DISPLAY 1.1 Creativity scores in two motivation groups, and their summary statistics

	Intrinsic group		Extrinsic group	
	12.0	20.5	5.0	17.4
	12.0	20.6	5.4	17.5
	12.9	21.3	6.1	18.5
	13.6	21.6	10.9	18.7
	16.6	22.1	11.8	18.7
	17.2	22.2	12.0	19.2
	17.5	22.6	12.3	19.5
	18.2	23.1	14.8	20.7
	19.1	24.0	15.0	21.2
	19.3	24.3	16.8	22.1
	19.8	26.7	17.2	24.0
	20.3	29.7	17.2	
Sample Size:	24		23	
Average:	19.88		15.74	
Sample Standard Deviation:	4.44		5.25	

Study Objective : Do ranking systems and incentive awards increase productivity among employees? Do rewards and praise stimulate children to learn?

Study design

- Subjects with considerable experience in creative writing
- Randomly assigned to one of the treatment groups
- Intrinsic group : motivation was satisfaction
- Extrinsic group : motivation was reward
- Evaluation of Haiku poems done by 12 poets for each poem by each subject
- Score was average of all 12 evaluations by each judge(poet) for each poem
- Judges not informed about study's purpose (why?)

Questionnaire - Motivation and Creativity study

DISPLAY 1.2 Questionnaires given creative writers, to rank intrinsic and extrinsic reasons for writing

INSTRUCTIONS: Please rank the following list of reasons for writing, in order of personal importance to you (1 = highest, 7 = lowest).

__ You get a lot of pleasure out of reading something good that you have written.
__ You enjoy the opportunity for self-expression.
__ You achieve new insights through your writing.
__ You derive satisfaction from expressing yourself clearly and eloquently.
__ You feel relaxed when writing.
__ You like to play with words.
__ You enjoy becoming involved with ideas, characters, events, and images in your writing.

*List of **extrinsic** reasons for writing.*

*List of **intrinsic** reasons for writing.*

INSTRUCTIONS: Please rank the following list of reasons for writing, in order of personal importance to you (1 = highest, 7 = lowest).

__ You realize that, with the introduction of dozens of magazines every year, the market for free-lance writing is constantly expanding.
__ You want your writing teachers to be favorably impressed with your writing talent.
__ You have heard of cases where one best-selling novel or collection of poems has made the author financially secure.
__ You enjoy public recognition of your work.
__ You know that many of the best jobs available require good writing skills.
__ You know that writing ability is one of the major criteria for acceptance into graduate school.
__ Your teachers and parents have encouraged you to go into writing.

Motivation and Creativity study – Statistical Conclusion

- Only for the volunteer participants of the two groups, there is strong statistical evidence that creative writers given “intrinsic” motivation caused writers to score higher compared to “extrinsic” motivation
- two-sided p-value = 0.005 from a two-sample t-test as an approximation to a randomization test.
- The estimated treatment effect—the increase in score attributed to the “intrinsic” questionnaire—is 4.1 points (95% confidence interval: 1.3 to 7.0 points) on a 0–40-point scale.
- Because the subjects were not selected randomly from any population, extending this inference to any other group is speculative. – Very important
- This deficiency, however, is minor; the causal conclusion is strong even if it applies only to the recruited subjects.

Example of Observational study – Sex discrimination in Employment

DISPLAY 1.3 Starting salaries ($U.S.) for 32 male and 61 female clerical hires at a bank

Males			Females					
4,620	5,700	6,000	3,900	4,500	4,800	5,220	5,400	5,640
5,040	6,000	6,000	4,020	4,620	4,800	5,220	5,400	5,700
5,100	6,000	6,000	4,290	4,800	4,980	5,280	5,400	5,700
5,100	6,000	6,300	4,380	4,800	5,100	5,280	5,400	5,700
5,220	6,000	6,600	4,380	4,800	5,100	5,280	5,400	5,700
5,400	6,000	6,600	4,380	4,800	5,100	5,400	5,400	5,700
5,400	6,000	6,600	4,380	4,800	5,100	5,400	5,400	6,000
5,400	6,000	6,840	4,380	4,800	5,100	5,400	5,520	6,000
5,400	6,000	6,900	4,440	4,800	5,100	5,400	5,520	6,120
5,400	6,000	6,900	4,500	4,800	5,160	5,400	5,580	6,300
	6,000	8,100						6,300

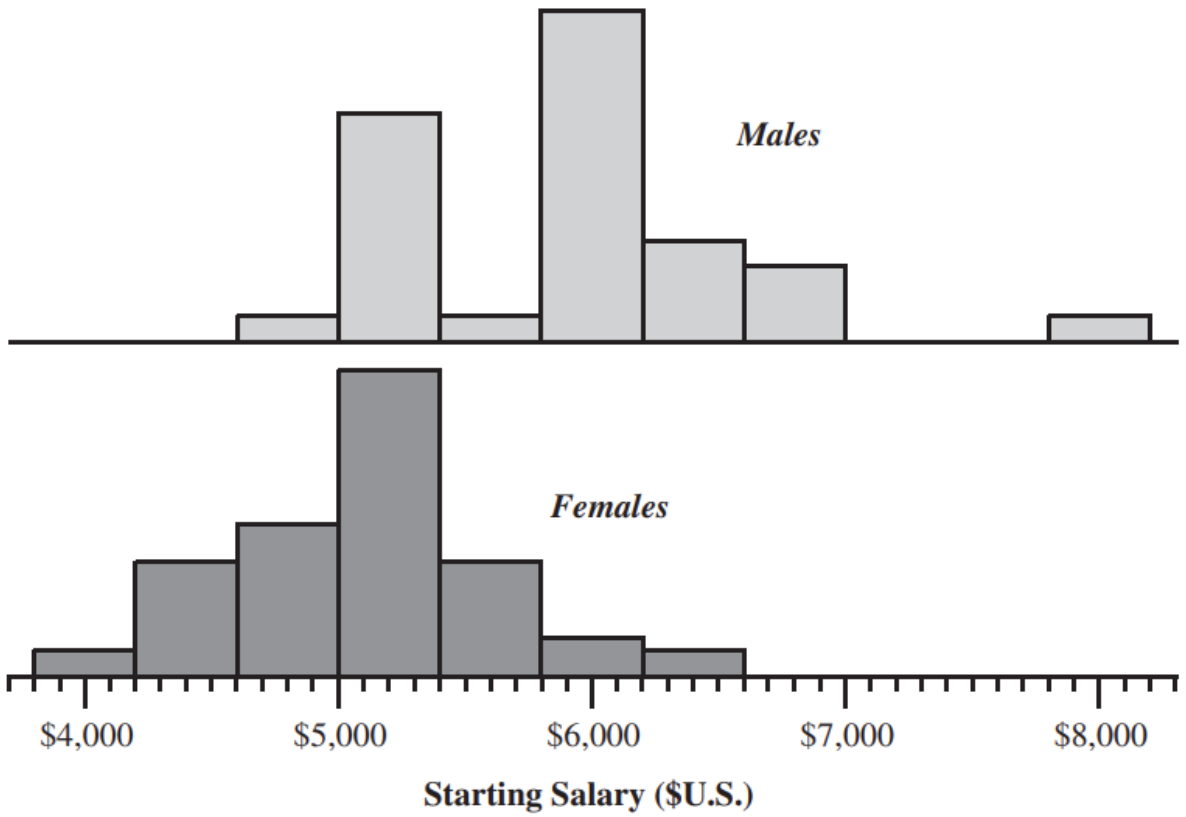
Study Objective : Did a bank discriminatorily pay higher starting salaries to men than to women?

Study design

- Data on left shows starting salaries for 32 males and 61 females
- Note that there is no treatment group in the study, i.e., sex of the employee is fixed.
- If a treatment group were to be allocated, a study design might have looked like this:
  - 100 subjects (50 male and 50 female) would have been randomly selected from a job portal. The resumes would have their names omitted from them
  - 10 different employers would have been given the 100 resumes once with their sex mentioned and once with their sex not mentioned, with a gap of a few days so the employers would forget the resumes
  - Based on the resumes, the employers would have been asked to assign salary, once without seeing the sex and once after seeing the sex
  - If the employers would have assigned higher salary to males one seeing sex compared to not on seeing sex, then there would have been a bias given the p value of the test was low enough to reject the null hypothesis of no difference in salary based on sex

Histogram - Sex discrimination in Employment

DISPLAY 1.4 Histograms for male and female starting salaries



Sex discrimination in Employment – Statistical Conclusion

- Convincing evidence that males received higher salaries
- Statistics alone cannot attribute this difference to sex discrimination
- Confounding variable : Experience in job

Statistical inference and study design

- Causal inference :
  - **Causal inference** is the process of determining the independent, actual effect of a particular phenomenon that is a component of a larger system. Causal inference analyzes the response of an effect variable when a cause of the effect variable is changed.
  - Statistical inferences of cause-and-effect relationships can be drawn from randomized experiments, but not from observational studies.
    - Motivation study was a randomized experiment, while sex discrimination study was an observational study
  - Randomization ensures that subjects with different, and possibly relevant, features are mixed up between the two groups.
  - For Motivation study, even though there is a chance that highly creative writers got placed in intrinsic group, BUT, due to randomization, each subject had the same chance of being placed in any group.

Statistical inference and study design

- In an observational study, it is impossible to draw a causal conclusion from the statistical analysis alone.
- One cannot rule out the possibility that confounding variables are responsible for group differences in the measured outcome.
- **Confounding variable** is related both to group membership and to the outcome. Its presence makes it hard to establish the outcome as being a direct consequence of group membership.

Why do observational studies

- The goal of observational studies might not be just to establish causation. Example – Blood Pressure in Asians vs Americans
- If confounding variables not present, then observational studies can establish causation. Example – Japanese atomic bombing chromosomal abberations
- Analysis of observational data may lend evidence toward causal theories and suggest the direction of future research. Example – Smoking and lung cancer early research was observational, then became randomized trials on animals, humans etc

Statistical inference and study design

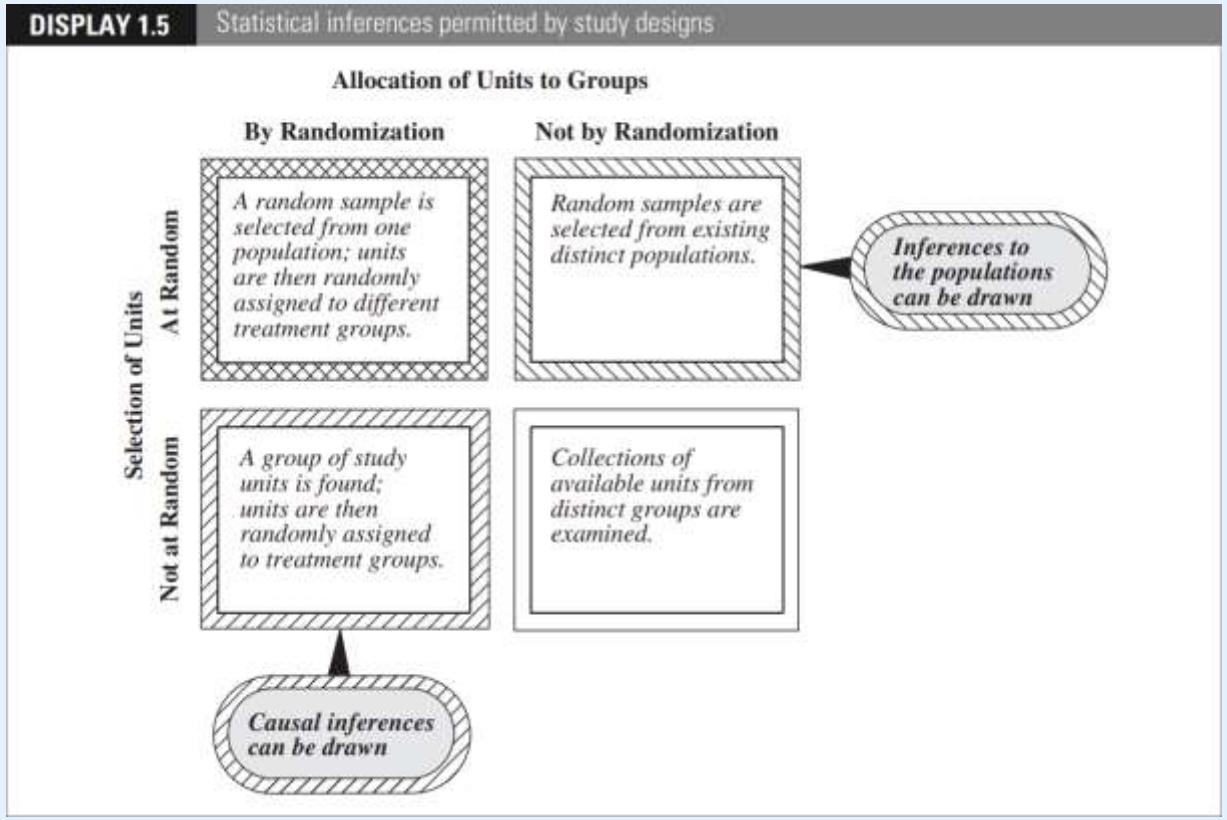
Inference to Populations :

- Inferences to populations can be drawn from random sampling studies, but not otherwise.
- Random sampling ensures that all subpopulations are represented in the sample in roughly the same mix as in the overall population.

Statistical Inference and Chance Mechanisms

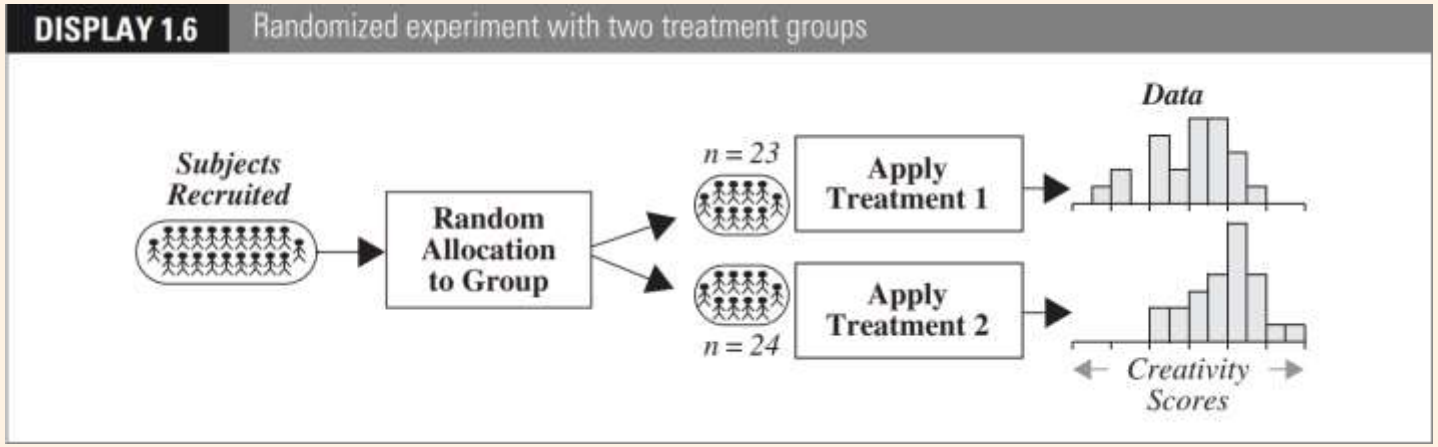
- An inference is a conclusion that patterns in the data are present in some broader context.
- A statistical inference is an inference justified by a probability model linking the data to the broader context.

Statistical Inferences Based on Chance Mechanisms



- Example: We wish to study the effect of pollution on residents of NCR
- We pick 1000 residents in all of NCR randomly, and ask 500 of them to remain in air purifier rooms and 500 to continue life as it is – Both inference to population and causal inference can be drawn
- We invite 1000 volunteers to take part in study. Keep 500 in air purifier rooms and 500 to continue life – Causal inference can be drawn, but inference to population cannot be drawn
- We pick 500 people from hospitals being treated for lung related ailments, and 500 people from healthy population. Now randomize these 1000 patients into 500 for air purifier rooms and 500 for continuing life – Inference to population can be drawn, but causal inference cannot be drawn
- We examine 500 NCR resident (who volunteered) and 500 non-NCR residents (who also volunteered). See if NCR volunteers are have more lung ailments. – Neither Causal inference can be drawn, nor inference to general population can be drawn

A Probability Model for Randomized Experiments



- Additive Treatment model:

Y : Extrinsic (Money motivation – lesser score) creativity score

Y_ : Intrinsic(self satisfaction – higher score) creativity score

Assumption of additive treatment model is

Y_ = Y + $

Where $ is a parameter

A Probability Model for Randomized Experiments

- Question to ask

Did something happen by chance (if we translate it to the motivation study, then was it because subjects with higher creativity were assigned to intrinsic group, which had higher score, BY CHANCE) – NULL HYPOTHESIS, i.e. $ = 0

OR

Was there a treatment effect (i.e., no matter whichever subject with whichever creativity skill got assigned to either intrinsic or extrinsic group, the intrinsic group score would have been always higher. So it would NOT BE BY CHANCE, but due to TREATMENT EFFECT) – ALTERNATE HYPOTHESIS,

i.e., $ > 0 or $ < 0

Refer ch1.ipynb file for t-test function

Different group assignment for Creativity study

DISPLAY 1.7 A different group assignment for the creativity study, and a different result

Creativity score	Actual grouping	Another grouping	Creativity score	Actual grouping	Another grouping
12.0	Intrinsic(2)	1	5.0	Extrinsic(1)	2
12.0	Intrinsic	2	5.4	Extrinsic	2
12.9	Intrinsic	1	6.1	Extrinsic	1
13.6	Intrinsic	2	10.9	Extrinsic	2
16.6	Intrinsic	2	11.8	Extrinsic	1
17.2	Intrinsic	1	12.0	Extrinsic	1
17.5	Intrinsic	2	12.3	Extrinsic	1
18.2	Intrinsic	2	14.8	Extrinsic	2
19.1	Intrinsic	1	15.0	Extrinsic	2
19.3	Intrinsic	2	16.8	Extrinsic	2
19.8	Intrinsic	2	17.2	Extrinsic	2
20.3	Intrinsic	2	17.2	Extrinsic	1
20.5	Intrinsic	1	17.4	Extrinsic	2
20.6	Intrinsic	2	17.5	Extrinsic	2
21.3	Intrinsic	1	18.5	Extrinsic	2
21.6	Intrinsic	2	18.7	Extrinsic	1
22.1	Intrinsic	1	18.7	Extrinsic	1
22.2	Intrinsic	2	19.2	Extrinsic	1
22.6	Intrinsic	1	19.5	Extrinsic	1
23.1	Intrinsic	1	20.7	Extrinsic	1
24.0	Intrinsic	1	21.2	Extrinsic	1
24.3	Intrinsic	1	22.1	Extrinsic	2
26.7	Intrinsic	1	24.0	Extrinsic	2
29.7	Intrinsic	1			

Averages from actual grouping

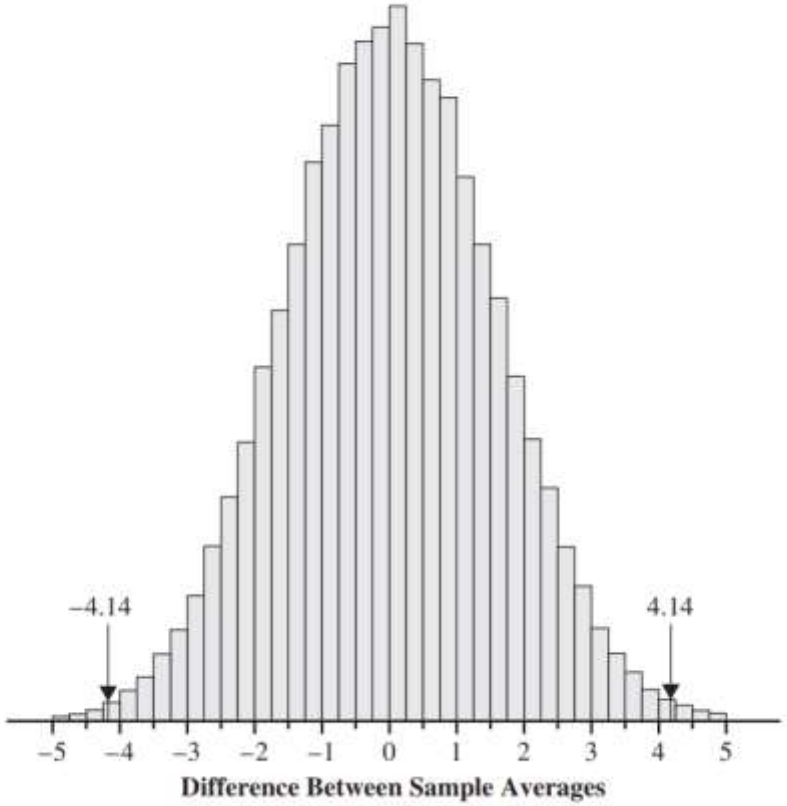
Group	Average	Difference
Intrinsic (2)	19.88	4.14
Extrinsic (1)	15.74	

Averages from another grouping

Group	Average	Difference
Group 1	18.87	2.07
Group 2	16.80	

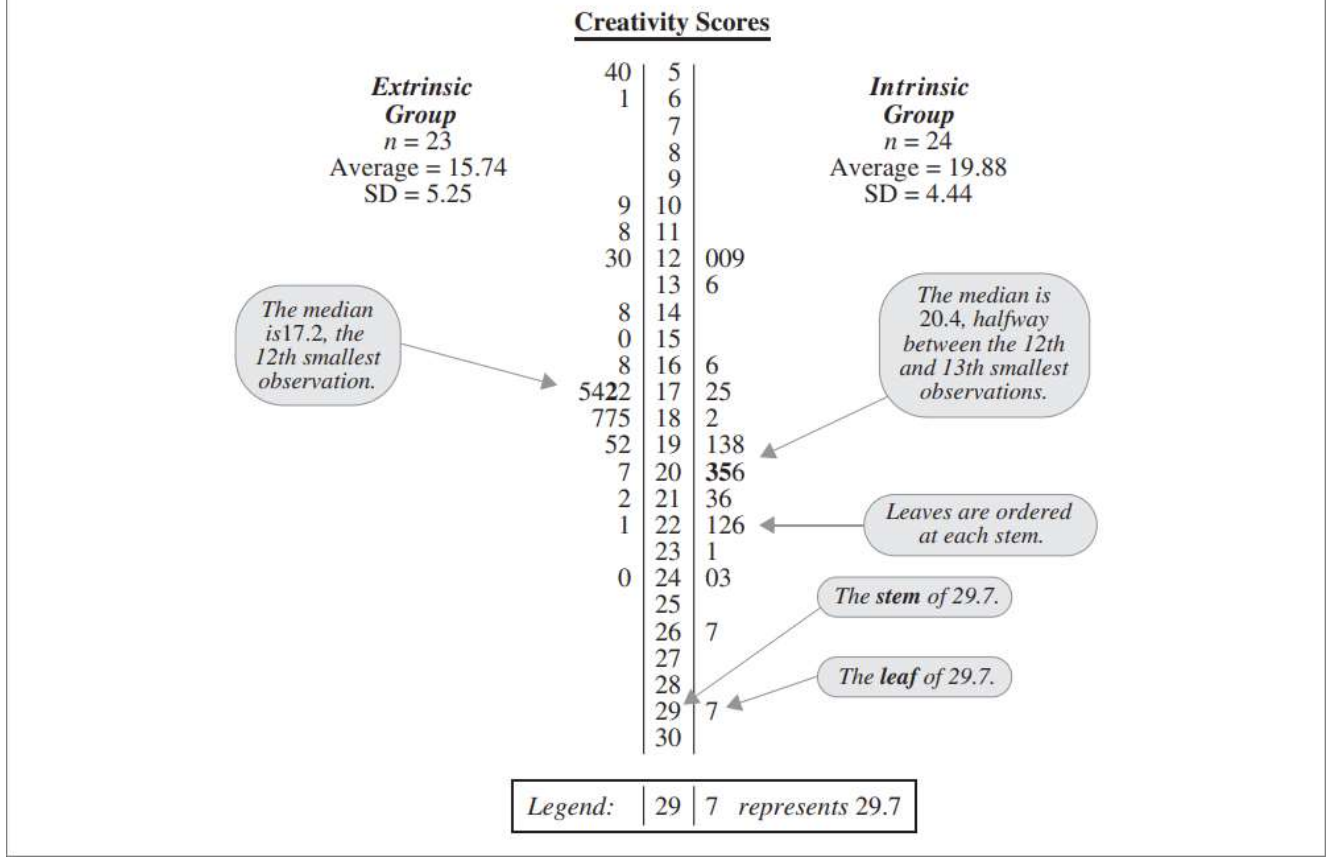
Histogram of student t-test distribution

DISPLAY 1.8 Histogram of differences between group averages, from 500,000 regroupings of the creativity study data

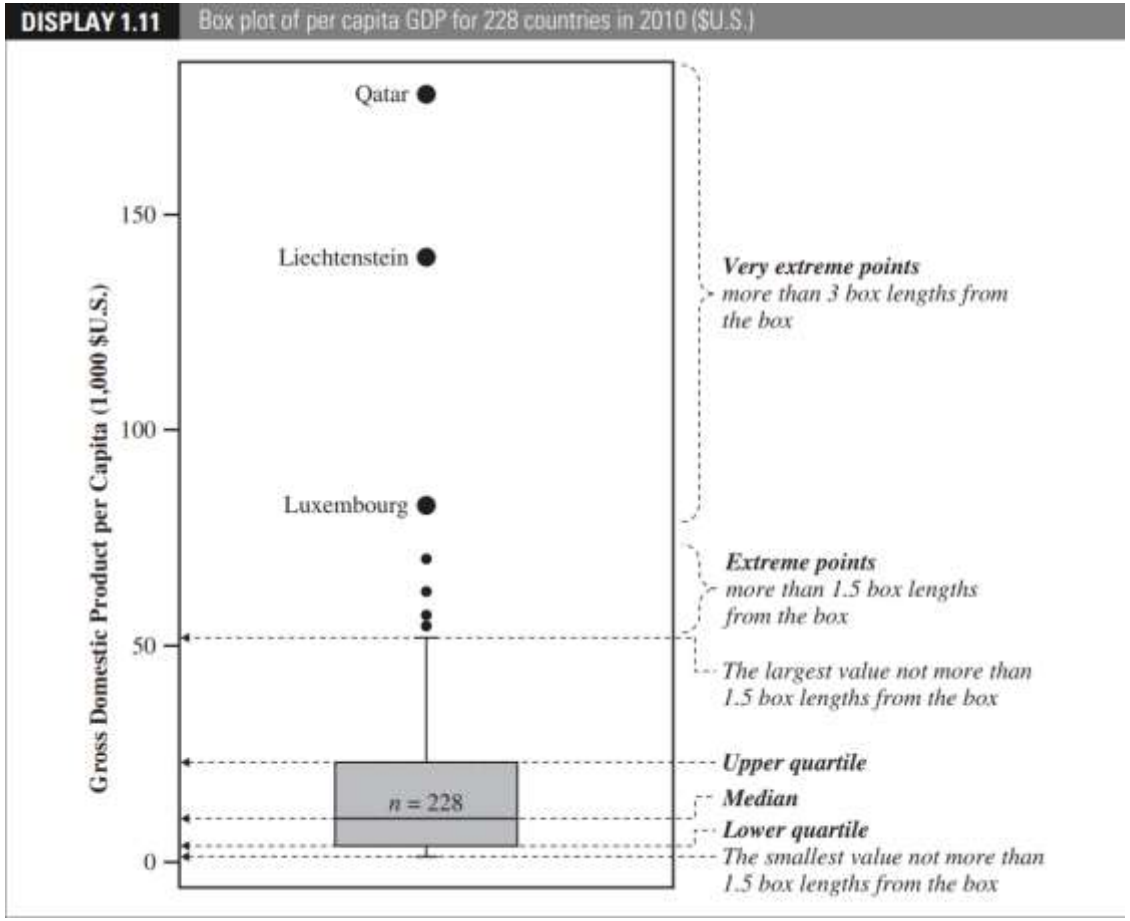


Stem and Leaf plot

DISPLAY 1.10 Back-to-back stem-and-leaf diagrams for the creativity study data



Box-Plot


DISPLAY 1.11
Box plot of per capita GDP for 228 countries in 2010 ($U.S.)
Qatar
Liechtenstein
Luxembourg
150
100
50
0
Gross Domestic Product per Capita (1,000 $U.S.)
n = 228
Very extreme points
more than 3 box lengths from the box
Extreme points
more than 1.5 box lengths from the box
The largest value not more than 1.5 box lengths from the box
Upper quartile
Median
Lower quartile
The smallest value not more than 1.5 box lengths from the box

Long tail, short tail, normal and skewed distribution based on box plot

DISPLAY 1.13 Histograms and box plots for 100 observations from four distributions