Introductory Statistics Refresher

Dr. Julia L. Sharp

Short Course on Introductory Statistics
Part I

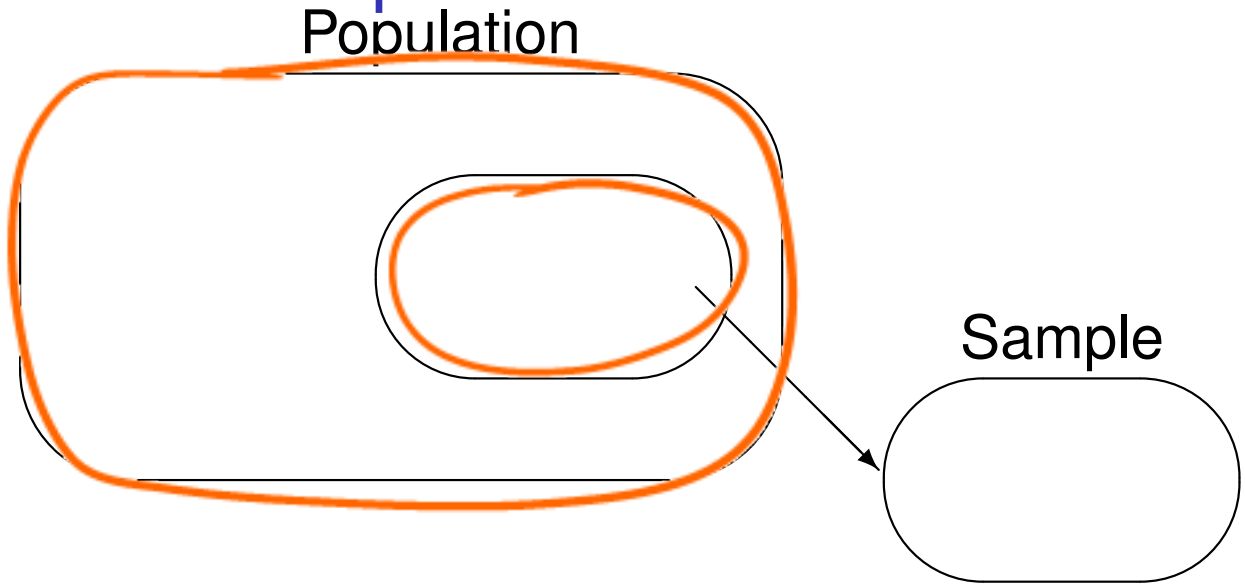
Definition of Statistics

- Statistics: Area of science concerned with the extraction of information from numerical data and its use in making inference about a population using data obtained from a sample.

- Statistical Inference: An “informed guess" about a parameter based on a statistic.

main goal of statistics

Population and Sample



- Population: The set of all individuals (sometimes defined as measurements) of interest to the researcher.
  - Parameter: An unknown population characteristic of interest.

unknown: mean – μ; variance – σ^2; standard deviation – σ

- Sample: The observed set of individuals (sometimes defined as measurements) from the population.
  - Statistic: A sample characteristic of interest.

compute: mean – $\bar{y}$; variance – s^2; standard deviation – s

Example

Selecting the proper diet for shrimp or other sea animals is an important aspect of sea farming. A researcher wishes to estimate the mean weight of shrimp maintained on a specific diet for a period of 6 months. One hundred shrimp are randomly selected from an artificial pond and each is weighed.

- Population: the weights of all shrimp on a specific diet
- Parameter(s) of interest: mean weight of all shrimp on a specific diet - μ
- Sample: the weights of 100 shrimp on a specific diet
- Statistic(s) of interest: mean weight of 100 shrimp on a specific diet - $\bar{y}$

Five Steps in a Statistical Study – Scientific Method

objectives

1. Stating the problem. – specifically identify pop to be sampled, all measurable variables and identify the parameter(s) of interest
2. Gathering the data. observational studies experiments
3. Summarizing the data. descriptive statistics graphical displays
4. Analyzing the data. estimation hypothesis testing
5. Reporting the results.

Gathering the data.

- Observational studies: (passive)
  - A study where measurements are observed but there is no attempt to influence the outcome.
  - Existing conditions or outcomes are studied.

survey - collecting opinions
blood pressure without attempting to influence the response

- Scientific studies: (experiments - active)
  - A study where a treatment is deliberately imposed to observe the effects on a measurement.
  - Change the conditions for one or more factors to study the effects of these changes.

blood pressure - individuals run around the building and then measure blood pressure
apply several different fertilizers to measure growth of several varieties of apple trees

Sampling Techniques

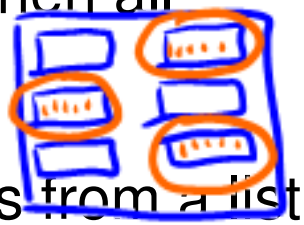
sampling frame - list of all companies
→ randomly select companies

- Simple Random Sample (SRS): A group of n units is selected in such a way that each sample of size n has the same chance of being selected.

most common

- Stratified Random Sample: the group of individuals is divided into separate groups, or strata. An SRS is selected from each strata.

- Cluster Sample: An SRS of groups is selected and then all individuals within the selected groups are sampled.



- Systematic Sample: Systematically select individuals from a list of all individuals.

phone book - systematically select every 10th person

20 m^2
net plot 12 m^2

Experimental Design Lingo

- Experimental Units: the object or individual to which a treatment is randomly and independently assigned.
- Observational Units: the object or individual on which the response of interest is measured.

can differ

- Response variable: the characteristic that is measured.
- Factor: a variable used to explain variation in the response variable that takes on two or more values.
- Levels: the values that a factor can take on.
- Treatments: The levels of a single factor or the combination of the levels of several factors.

generic label – fertilizer, variety of tree

the "thing" that is imposed

fertilizer 1,2,3

variety A, B

fertilizer 1, variety A
fertilizer 2, variety A

6 treatments
all of the combinations of fertilizer and variet

Main Experimental Design Concepts

- Control: control influential variables on the response
  - Extraneous factors that we have some control over
  - Other factors that we may not have control over can be treated as blocks.
- Randomization: Assign treatments to subjects randomly.

reduces experimenter bias

- Replication: Assess the same treatment on multiple subjects to assess subject-to-subject variation.

several experimental units receive the treatment

used to obtain reliable estimates of the effect of each treatment

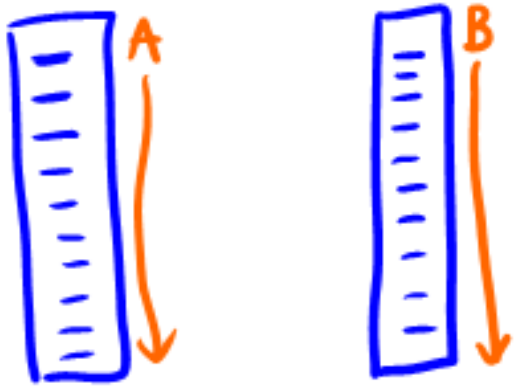
  - Pesticide A is compared with Pesticide B. Two benches are set up in a greenhouse and 10 plants are placed on each bench.

(1) →
    - Pesticide A is used for the 10 plants on one bench and Pesticide B is used for the 10 plants on the other bench.

(2) →
    - Pesticide A is used for 10 randomly selected plants on either bench. Pesticide B is used for the remaining 10 plants.

Replication Continued

factor:
pesticide
levels: A, B



① experimental units – bench of plants
observational unit – plant

no replication of observations of experimental units

confounding between trt. and bench

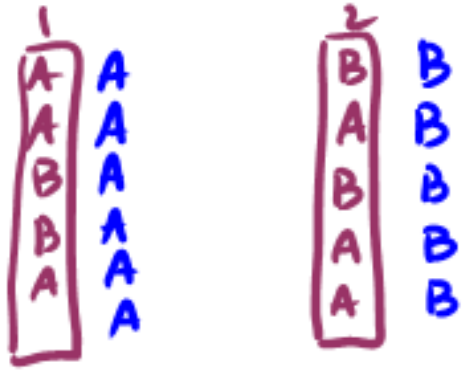
"pseudo-replication" – technical reps (repeated measurements of the same sample)

② experimental units: plant

observational unit: plant

biological reps – measurements are taken from several independent individuals

Types of Scientific Studies



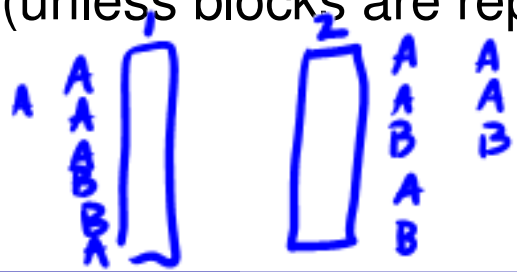
- Completely Randomized Design.
  - Used to compare t treatments.
  - All experimental units are allocated at random among all of the treatments.
  - A disadvantage of this type of experiment is that if differences exist among the experimental units that affect the response, differences among the treatments will be difficult to detect.

If the bench is coming in to play, it may be difficult to detect differences

- Randomized Block Design.
  - Treatments are randomly assigned to units within each block.
  - Every treatment occurs in every block.
  - A disadvantage of this type of experiment is that block-to-block comparisons cannot be made (unless blocks are replicated and selected at random).

10 A
10 B

Avoids the comparison of treatments distorted by differences in blocks.



Definition of Variables

- Variables: Characteristics of individuals or objects that may vary for different observations.

- Qualitative Variables: The variable represents categorical information.

  - gender
  - race
  - pesticide A,B

  variety A,B
  fertilizer 1,2,3

- Quantitative Variables: The variable represents a numerical scale.

  have actual units of measure

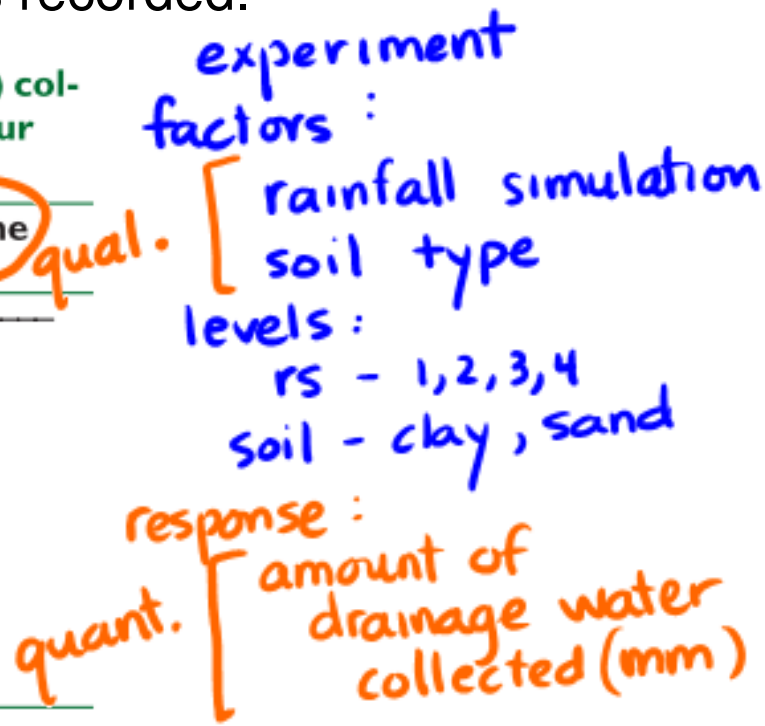
  - age
  - height
  - number of nematodes
  - growth of plants

Example

Riddle and Bergström (2013) describe several experiments to examine Phosphorus leaching from two soils. A table of results from one of the experiments is reproduced below. There were four different rain simulations used and two soil types (clay and sand). The amount of drainage water collected from lysimeters was recorded.

Table 4. The mean amount of drainage water (±SE, n = 1–5) collected from the lysimeters from each of the two soils and four rainfall simulations. The desired application was 50 mm.

Rainfall simulation	Soil	Mean	SE of the mean
		mm	
1	clay	37.6	2.5
	sand	27.1	1.1
2	clay	42.1	2.6
	sand	46.4	1.0
3	clay	47.3	2.9
	sand	46.2	1.2
4	clay	49.5	1.3
	sand	52.0	1.6

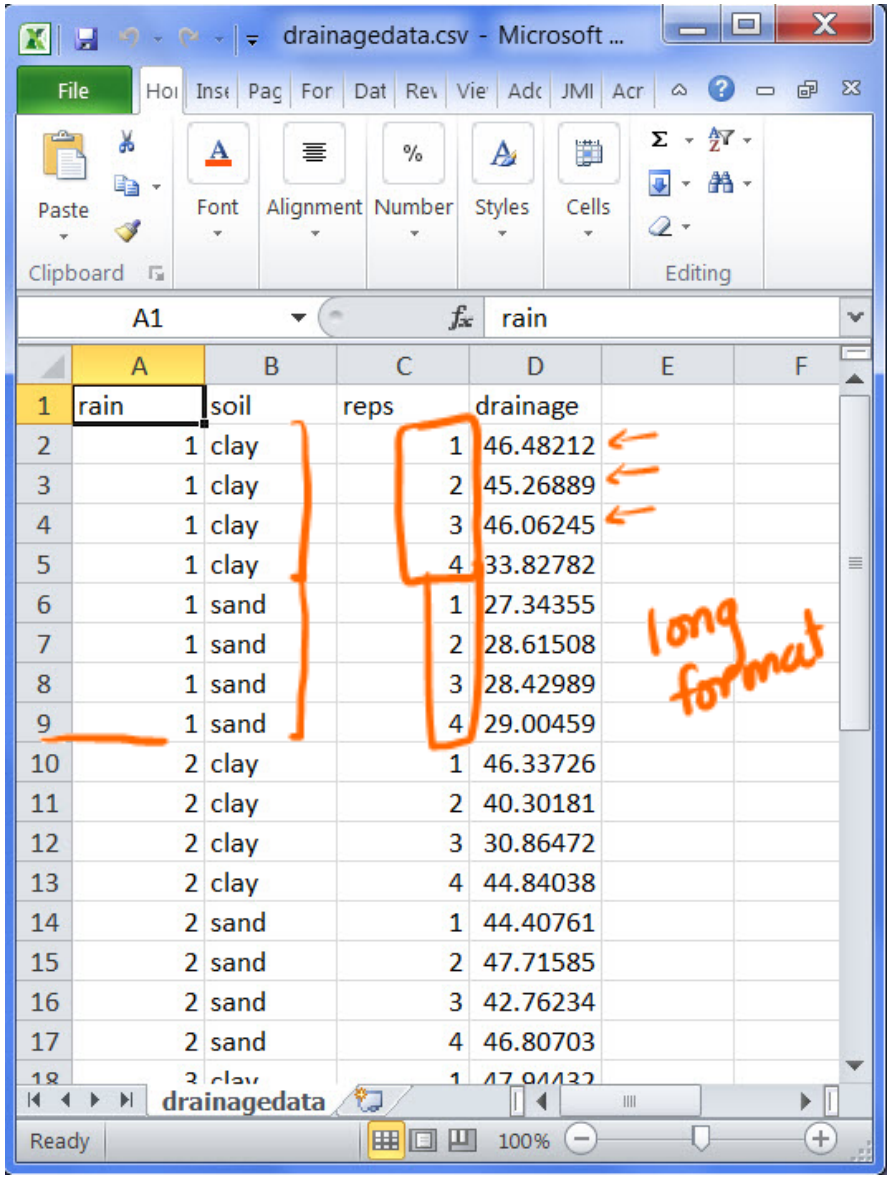


Riddle, M. U. and Bergström, L. (2013). "Phosphorus leaching from two soils with catch crops exposed to freeze-thaw cycles," *Agronomy Journal*, 105(3): 803-811.

Example Data in Excel

Data was simulated based on the results of Riddle and Bergström (2013).

Rain Simulation	Soil Type Clay	Soil Type Sand
1	46.482	27.344
	45.269	28.615
	46.063	28.430
	33.828	29.005
2	46.337	44.408
	40.302	47.716
	40.865	42.762
	44.840	46.807
3	54.698	47.558
	47.461	43.279
	51.457	43.202
	54.698	41.812
4	53.186	53.097
	52.124	56.796
	48.477	52.054
	48.904	49.813



	A	B	C	D
1	rain	soil	reps	drainage
2	1	clay	1	46.48212
3	1	clay	2	45.26889
4	1	clay	3	46.06245
5	1	clay	4	33.82782
6	1	sand	1	27.34355
7	1	sand	2	28.61508
8	1	sand	3	28.42989
9	1	sand	4	29.00459
10	2	clay	1	46.33726
11	2	clay	2	40.30181
12	2	clay	3	30.86472
13	2	clay	4	44.84038
14	2	sand	1	44.40761
15	2	sand	2	47.71585
16	2	sand	3	42.76234
17	2	sand	4	46.80703

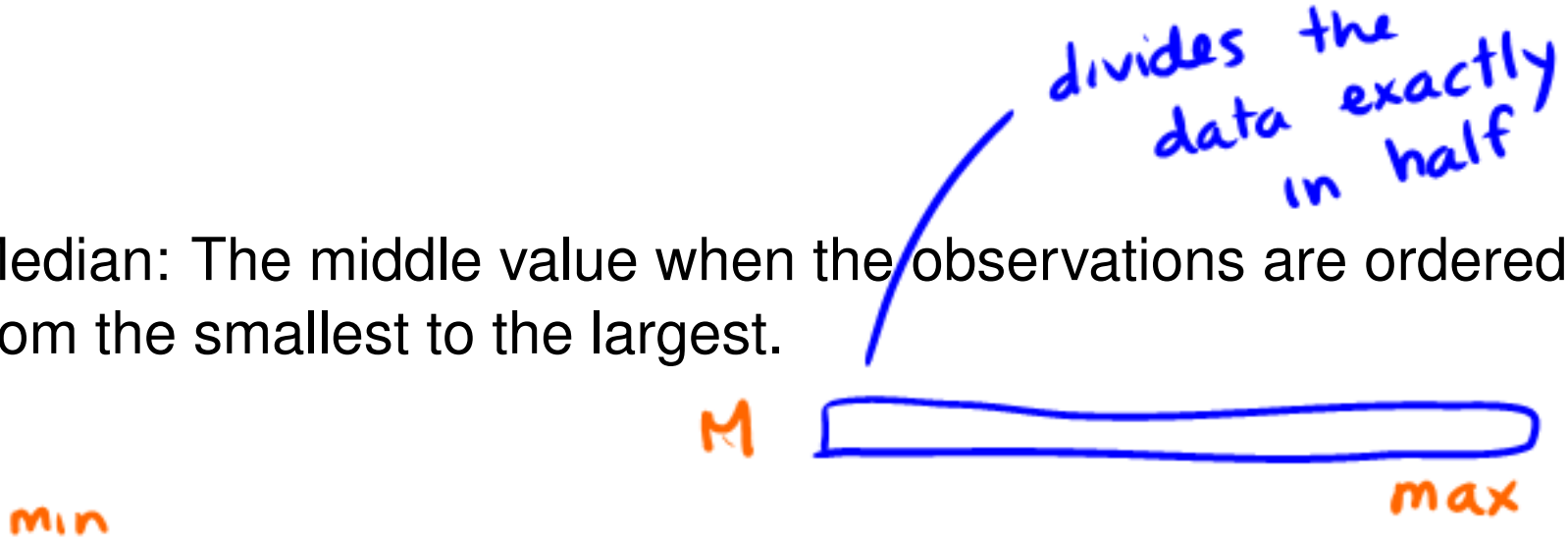
Descriptive Measures

count

proportions or percentages

- Qualitative variables: Frequencies, relative frequencies
- Quantitative variables
  - Measures of Center: Median, mean
  - Measures of Other Locations: Minimum, maximum, first quartile, third quartile
  - Measures of Spread: Variance, standard deviation, interquartile range, range

Measures of Center



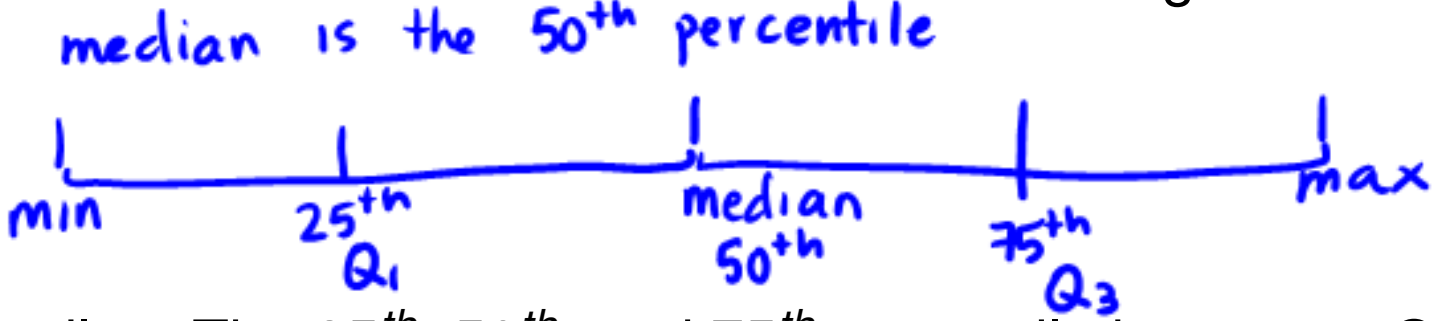
- Median: The middle value when the observations are ordered from the smallest to the largest.

- Mean: The 'balance point' of the data. Found by summing all of the observations and dividing by the number of observations.

Measures of Other Locations

- Minimum and maximum: The smallest and largest values in a data set.

- Percentile: The value, x_p, that has p% of the measurements below it and $(1 - p)$% of the measurements above it when the measurements are ordered from smallest to largest.



- Quartiles: The 25^{th}, 50^{th}, and 75^{th} percentile known as Q_1, the median, and Q_3, respectively.

divide data into 4 parts

Q_3 - third quartile - 75% of the obs. below it

Q_1 - first quartile - 25% of the obs. below it

Measures of Spread: Variance and Standard Deviation

Sample variance:

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$$

sum

a measure that is similar to the average of the squared deviations (observations minus the sample mean)

divide by n-1: because if we know the mean and we know n-1 observations, then we can find the n^{th} obs. n-1 of the observations are allowed to freely vary.

$$= \frac{\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}}{n-1}$$

computational formula

Sample standard deviation: $s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}}{n-1}}$

same units as the observations

Measures of Spread: Range and IQR

- Range is the maximum value minus the minimum value

R

- Interquartile range (IQR) is the range of the middle 50% of the data: $IQR = Q_3 - Q_1$

beneficial when we're looking at observations from data with extreme measurements

- Identifying outlying values:

$$\begin{bmatrix} Q_3 + 1.5\,IQR \\ Q_1 - 1.5\,IQR \end{bmatrix}$$

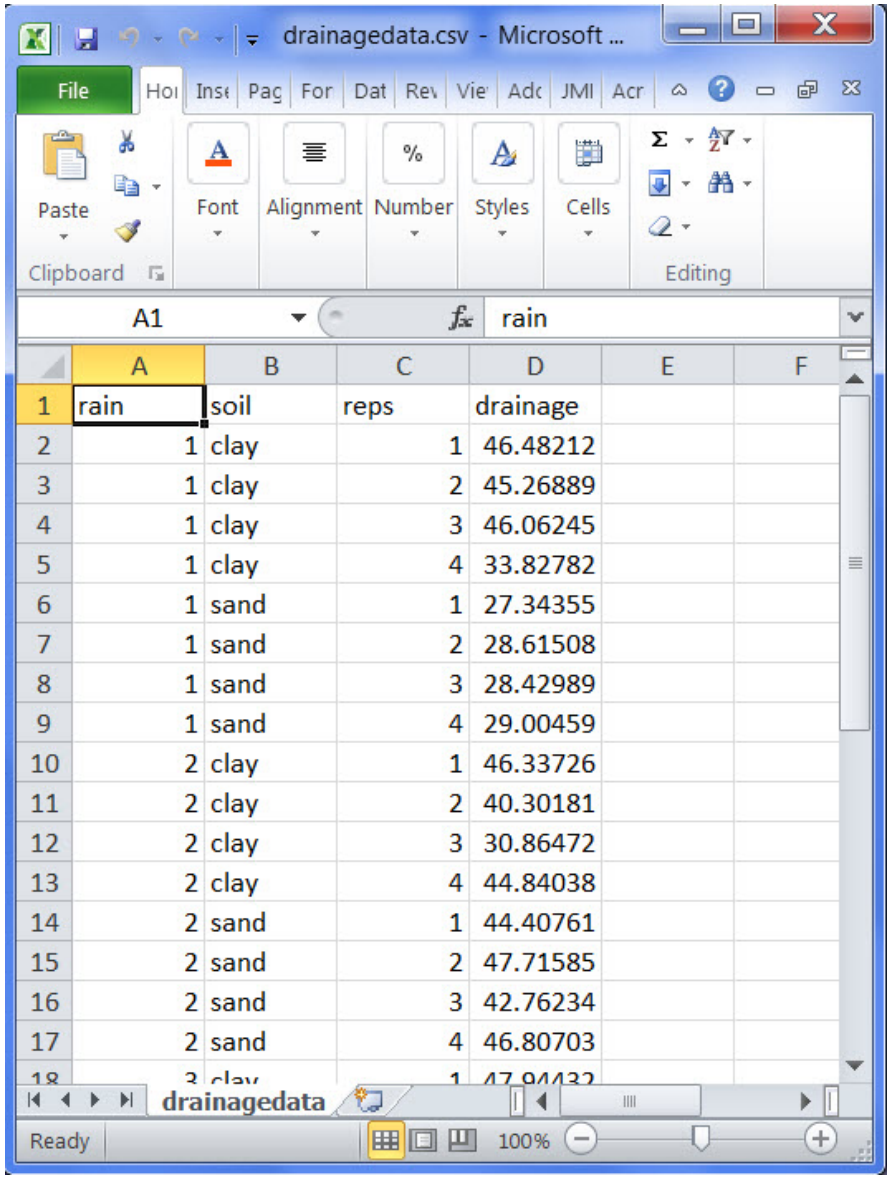
If observations are above or below, these observations may be considered extreme

mean $\pm$ 3 s.d.

Example Data in Excel

Data was simulated based on the results of Riddle and Bergström (2013).

Rain Simulation	Soil Type Clay	Soil Type Sand
1	46.482	27.344
	45.269	28.615
	46.063	28.430
	33.828	29.005
2	46.337	44.408
	40.302	47.716
	40.865	42.762
	44.840	46.807
3	54.698	47.558
	47.461	43.279
	51.457	43.202
	54.698	41.812
4	53.186	53.097
	52.124	56.796
	48.477	52.054
	48.904	49.813



	A	B	C	D
1	rain	soil	reps	drainage
2	1	clay	1	46.48212
3	1	clay	2	45.26889
4	1	clay	3	46.06245
5	1	clay	4	33.82782
6	1	sand	1	27.34355
7	1	sand	2	28.61508
8	1	sand	3	28.42989
9	1	sand	4	29.00459
10	2	clay	1	46.33726
11	2	clay	2	40.30181
12	2	clay	3	30.86472
13	2	clay	4	44.84038
14	2	sand	1	44.40761
15	2	sand	2	47.71585
16	2	sand	3	42.76234
17	2	sand	4	46.80703

Descriptive Statistics: Qualitative Variables

The SAS System

The FREQ Procedure

rain	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	8	25.00	8	25.00
2	8	25.00	16	50.00
3	8	25.00	24	75.00
4	8	25.00	32	100.00

De le Quantitative

De

The SAS System

The UNIVARIATE Procedure
Variable: drainage

Moments			
N	32	Sum Weights	32
Mean	44.7166384	Sum Observations	1430.93243
Std Deviation	7.80685659	Variance	60.9470098
Skewness	-0.9867255	Kurtosis	0.43003464
Uncorrected SS	65875.8452	Corrected SS	1889.3573
Coeff Variation	17.4585051	Std Error Mean	1.38007031

Basic Statistical Measures			
Location		Variability	
Mean	44.71664	Std Deviation	7.80686
Median	46.40969	Variance	60.94701
Mode	.	Range	29.45237

Descriptive Statistics: Quantitative Variable
Descriptive Statistics

The SAS System

The MEANS Procedure

Analysis Variable : drainage				
N	Mean	Std Dev	Minimum	Maximum
32	44.7166384	7.8068566	27.3435549	56.7959287

Descriptive Statistics: Quantitative Descriptive Statistics by Qualitative Variables

The MEANS Procedure

soil=clay

amt of drainage water

Analysis Variable : drainage								
N	Mean	Median	Std Dev	Variance	Minimum	Maximum	25th Pctl	75th Pctl
16	46.7648008	46.9714428	5.2595361	27.6627197	33.8278208	54.6983761	45.0546332	50.1805586

soil=sand

Analysis Variable : drainage								
N	Mean	Median	Std Dev	Variance	Minimum	Maximum	25th Pctl	75th Pctl
16	42.6684759	43.8432317	9.4522572	89.3451655	27.3435549	56.7959287	35.4084242	48.7645161

Graphical Displays

- Qualitative Variables
  - Bar Graphs (Frequency or Relative Frequency)
- Quantitative Variables – single variable
  - Histograms, Stem-and-Leaf Plots, Box-and-whisker Plots
- Combinations of Variables
  - Qualitative, Quantitative: Box-and-whisker Plots, Bar Graphs
  - Two Quantitative: Scatterplots

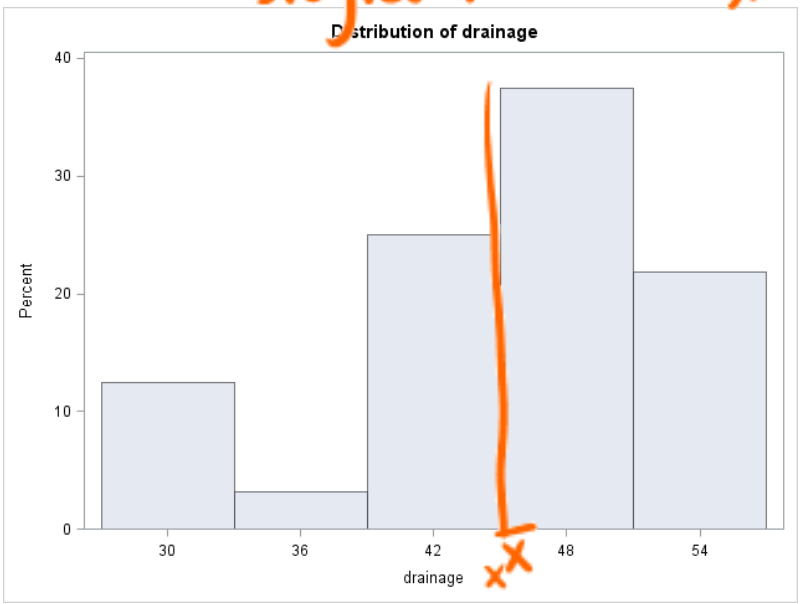
mean

Single Variable Plots

Quantitative : drainage water collected

histogram

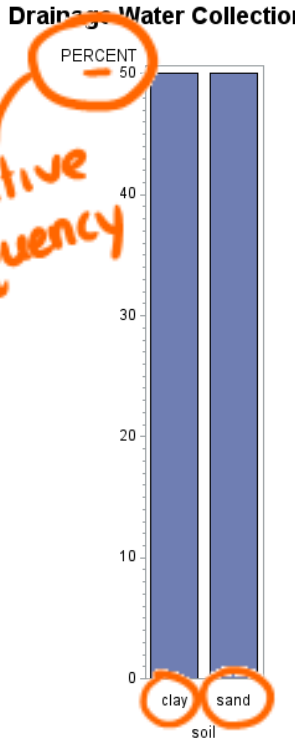
$\bar{x}$ - average drainage water collected



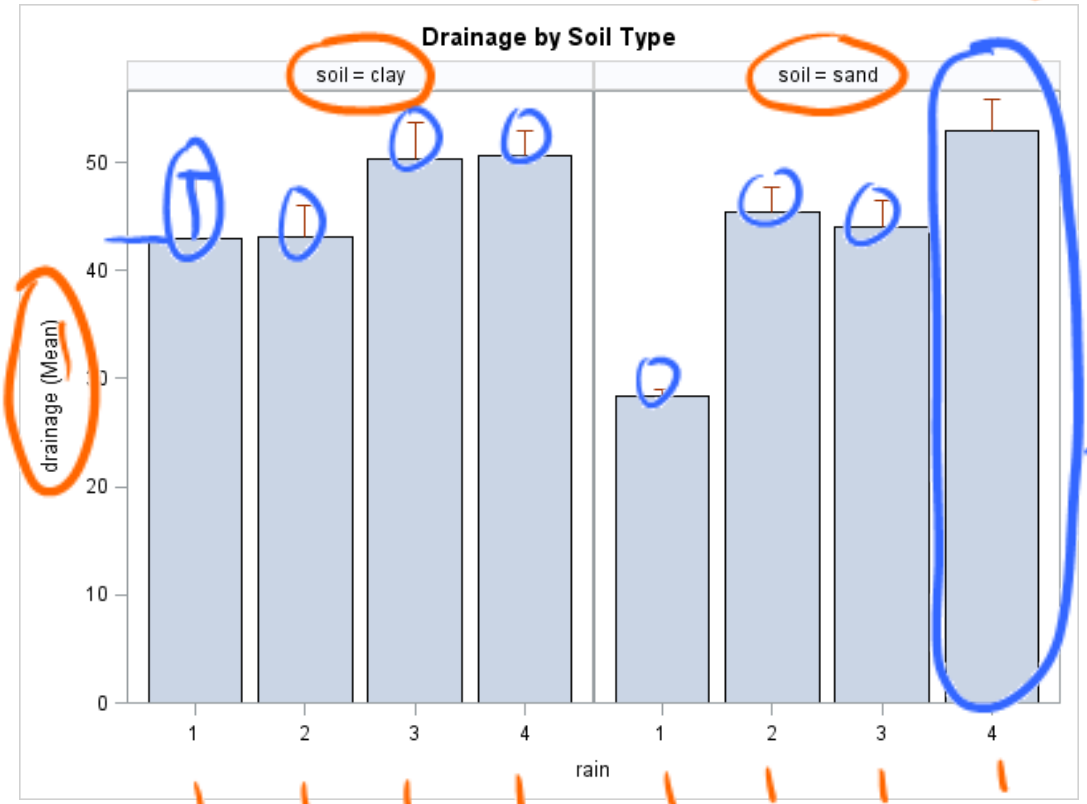
bar graph

qual. var : soil type

relative frequency



Panel of Bar Graphs



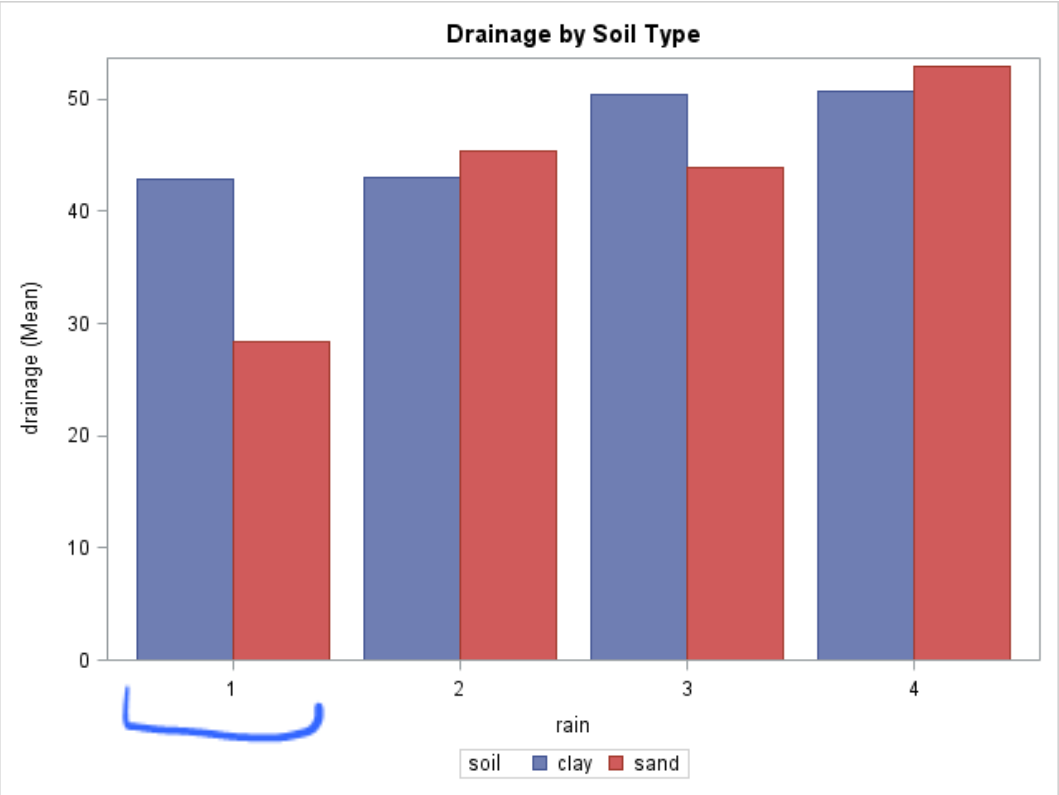
qual. [soil type
rain sim

quant - drainage

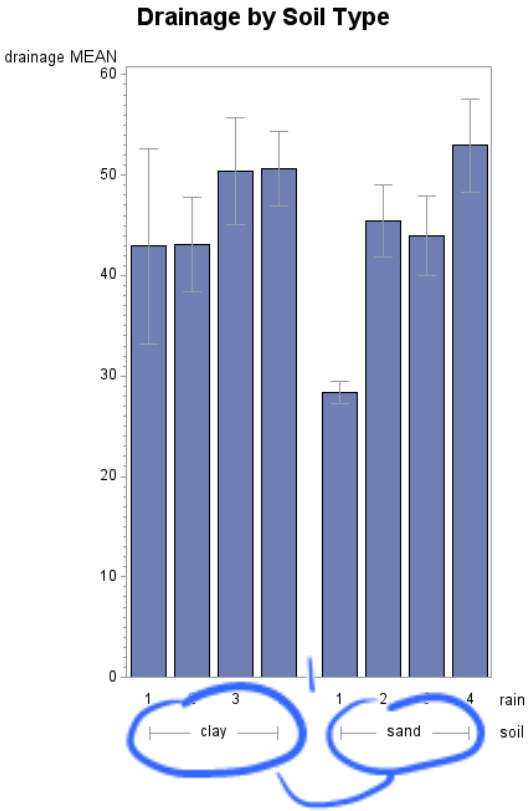
average drainage water collected for rain simulation 4 in sandy soil

Figure X. Mean (SD) drainage water collected

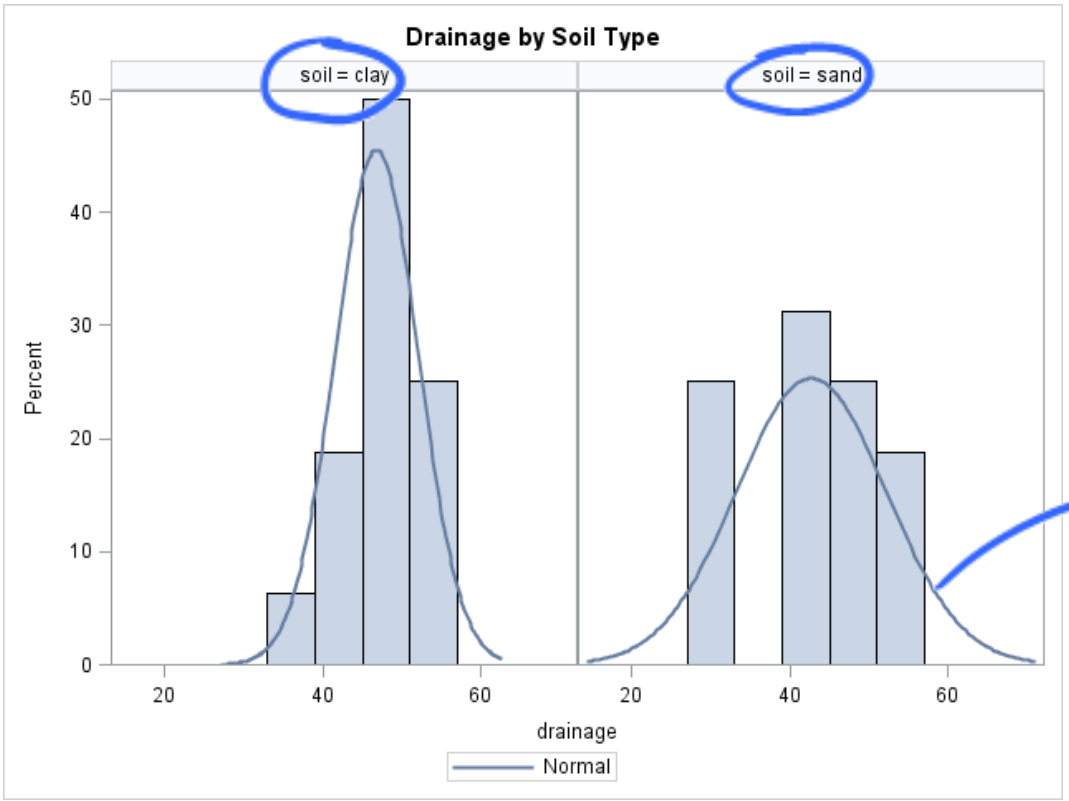
Side-by-side Bar Graphs



Side-by-side Bar Graphs



Panel of Histograms

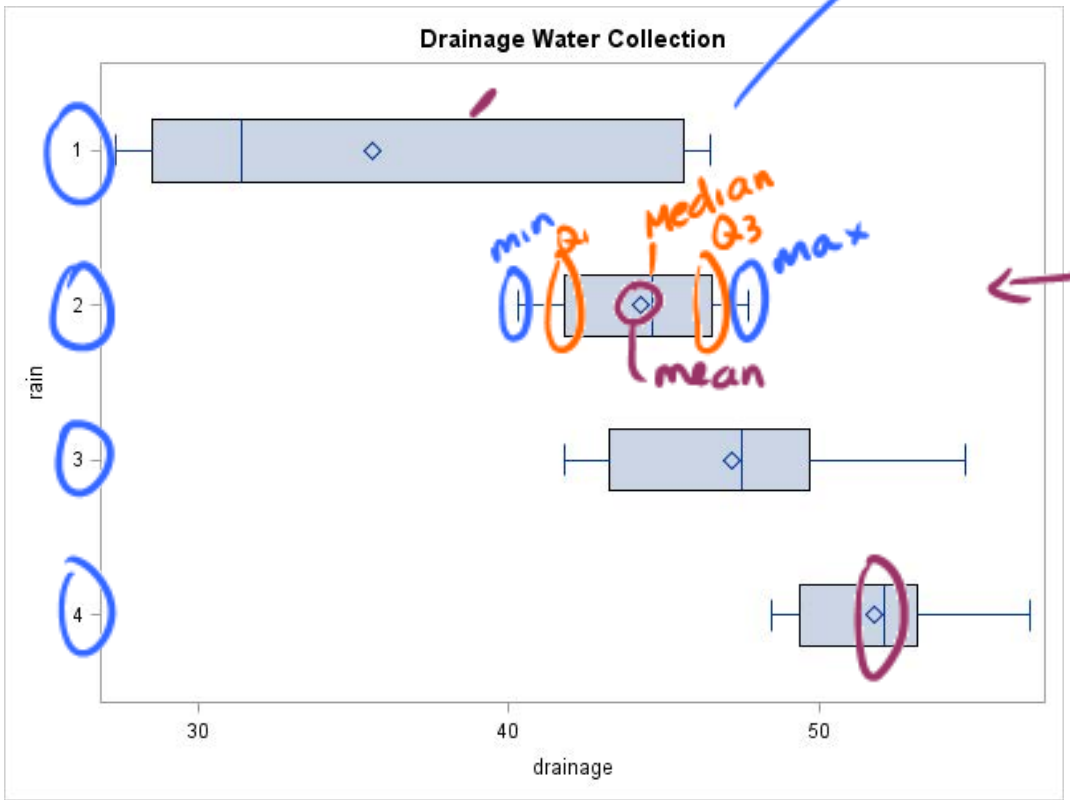


density curve - smooth curve to summarize the distributions

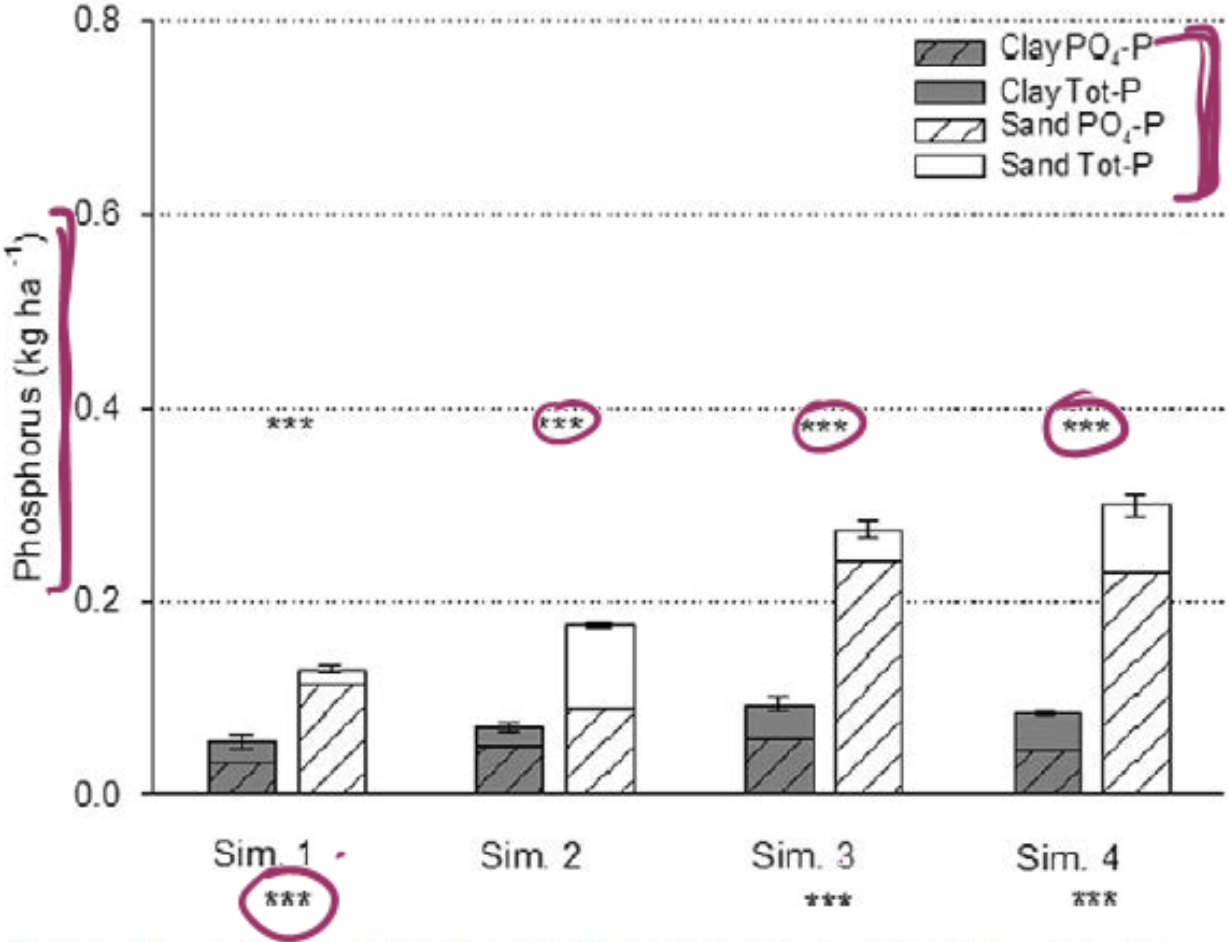
Box-and-whisker Plots

plot of the 5-number summary

min, Q_1, median, Q_3, max



Graphics From Riddle and Bergström (2013)



multiple variables

bar graph

Fig. 1. Mean P leaching loads from all control treatments for the four simulations. Error bars are standard errors of the mean. Top asterisks indicate total P (Tot-P) significance and bottom asterisks indicate PO_4–P significance between soil types within the same simulation: ***significant at $P < 0.001$

next week

– two weeks

two weeks