

Index of 1-L

Page	Title
1	List of participants
2	Practical information
3	Computing resources
4	Course exam
5	What's new? (vs. earlier years)
6	What is statistics? and why statistics?
7	Data
8	Data example: Parasites
9	Statistical concepts and procedures
10	Graphs of categorical variables
11	Bar graph and pie chart
12	Graphs of quantitative data: stemplots
13	Stemplots produced by software
14	Graphs of quantitative data: histograms
15	Graphs for Breaking strength data
16–17	Mean and median
18	Simple measures of spread
19	Boxplot for parasite data
20	Standard deviation
21	Appendix: Time plots
22	Summary notes

LIST OF PARTICIPANTS

Name	Department	Discuss?		Text- book
		schedule	regist.	
Anna Potter	Health Management		✓	
Brodie Reinhart	Pathology and Microbiol.			
Devon Lynn	Biology		✓	
Elizah McFarland	Health Management		✓	
Hassan Afzaal	Engineering		✓	
Humna Khan	Engineering			
Joao Romero	Health Management		✓	
Landon Warder	Health Management			
Madeleine Stein	Companion Animals		✓	
Mariana Fonseca			✓	
Natasha Boyle	Environmental Sciences			
Nicolas Decelles	Pathology and Microbiol.			
Rimsha Khan	Engineering		✓	
Santiago Uribe Diaz	Chemistry		✓	

Instructor/Professor:

Henrik Stryhn, AVC biostatistician, MSc and PhD in (mathematical) statistics, Dep. of Health Management, room 412S, phone 894-2847, e-mail: hstryhn@upei.ca, homepage: people.upei.ca/hstryhn.

PRACTICAL INFORMATION

WELCOME!!

Major news:

- to find latest course information... → web page for VHM 801: `people.upei.ca/hstryhn/vhm801`
- to “connect” yourself to the course (information, discussion...) → log into Moodle account for VHM 801 (`moodle.upei.ca`),
- to follow the course efficiently → *recommended* that you decide about a textbook pretty soon.

Today’s lecture: Introduction and Gentle Start

- lots of practical details about the course, incl. discussion of schedule, textbook and exam,
- descriptive statistics for distributions, in two forms:
 - * graphical displays, e.g. histograms and box plots,¹
 - * numerical summaries, e.g. mean and median,²
- software demonstrations included, but software practice for you will be in the lab session (tomorrow).

¹ Textbook coverage: Stephens: Chapter 3; Baldi & Moore: Chapter 1.

² Textbooks: Stephens: Chapter 2; Baldi & Moore: Chapter 2.

COMPUTING RESOURCES

Calculators:

- the traditional learning tool for statistics, and may improve understanding of formulas and methods,
- in this course: handy during labs, essential for exams; a calculator with basic calculus³ should suffice.

Computers and statistical packages

- today we cannot imagine statistics without computers,
 - * easier — avoids tedious calculations, and vastly widens the feasible range of models and analyses,
 - * increases also the risk of errors . . . ,
- in this course we use primarily⁴ Minitab (version 18) but support also Stata and R; all software packages:
 - * are well-documented and updated programs, and available at UPEI (Stata: personal license needed),
 - * have good graphing facilities,
 - * have both menus and commands,
- you choose between Minitab and others — a trade-off:
 - * Minitab: is easier to use, has better help facilities,
 - * Stata/R: are used in other AVC/UPEI courses, have wider range of statistical methods, but also steeper learning curves.

³ Including memory and logarithm etc., possibly also “1-variable statistics”.

⁴ Virtually all demonstrations in lectures and labs will be based on Minitab.

COURSE EXAM

The course exam is made up by:

- 4 home assignments (two for 10% and two for 15%),
 - * tentative dates: 26/9, 14/10, 31/10 and 14/11 (deadlines one week later),
 - * will you have “own data” by mid-November?
(to replace the last assignment by a small project)
- final exam (50%):
 - * *tentative* date: Monday 9/12,
 - * 3 hours, in-class, open book,
 - * no computers (instead: computer listings),
- mid-term exam (*optional* \sim 15%):
 - * date: around 24/10, duration: 1 hour,
 - * covers Sessions 1–8,
 - * same conditions as final exam \Rightarrow training session.

As a general rule, students who follow the course seriously should pass the exam rather easily...

Course marks in previous years:

Year	2011	2012	2013	2014	2015	2016	2018w	2018f
Avg. mark (%)	76.6	80.7	81.8	80.3	80.3	79.4	80.7	80.0

WHAT'S NEW? (VS. EARLIER YEARS)

Changes in course content and organization:

- new textbooks (2014):
 - * Baldi & Moore (4th ed.) replaces previous main text (Moore, McCabe & Craig),
 - * optional use of other texts (e.g., Stephens vers. 4.1),
- new software or software versions (2019):
 - * Minitab vers. 18: minor improvements over previous versions only, incl. some added analyses,
 - * Stata vers. 16: few major changes in commands used in course from earlier versions,
 - * R vers. 3.0+ (support): selected exercise solution files,
- new topics (2006 onwards):
 - * reporting of statistical analysis in papers,
 - * Bayesian methods (hard to fit into course schedule),
- open consultation hours (2006): possible to convert one of these into lab review covering selected problems,
- optional mid-term exam (2007),
- inclusion of summary notes in lecture handouts (2013),
- inclusion of review problems in lectures/labs (2014),
- inclusion of optional Moodle quizzes/reflections (2018),
- inclusion of interactive polls in lectures/labs (2019).

WHAT IS STATISTICS? AND WHY STATISTICS?

- 2 branches of statistics:
 - * “official statistics” — the collection and display of figures in statistical yearbooks etc.,
 - * “inferential statistics” — the science of analysing and interpreting data, from experiments or databases.
- “With statistics one can prove everything” — not true,
 - * cannot really *prove* anything,
 - * can separate random variation from systematic effects (differences, associations ...),
 - * can (strongly) indicate certain tendencies in data,
 - * statistical results do *not* imply causation ..., (nor guarantee biological significance...).
- Statistics is undergoing major changes, e.g. “big data science”, “critique of *p*-values”, Bayesian methods.

Why statistics (in particular, Biostats 801)?

- mandatory⁵, unless you’ve had “statistics” before,
- useful (indispensable) for data analysis,
- helps to develop *critical sense* for data and the results of data analysis,
- basics/building block for more advanced methods.

⁵ UPEI Calendar: “All [AVC] students are expected to complete VHM 801 [...] unless comparable training has been completed prior to entry into the program.”

DATA

Give an example of data (real or hypothetical) related to your project!

Organization of data:

- individuals (units of measurement/observation, experimental units, subjects) = the objects described by a set of data (people, animals, things),
- variables = characteristics (measurements, recordings) of the individuals,
- organized in computer programs in spreadsheet format with individuals as rows and variables as columns.

Types of variables:

- quantitative⁶ (either continuous or discrete):
 - * takes numerical values for which arithmetic operations such as adding and averaging make sense,
 - * values often have units or are counts,
- categorical (also grouped/qualitative):
 - * places individuals into one of several categories,
 - * categories often have labels,
 - * categories may be unordered (nominal) or ordinal,
 - * a quantitative variable may be split into categories.

⁶ S further distinguishes between interval and ratio measurements.

DATA EXAMPLE: PARASITES

“Natural Trichostrongylid exposure of calves in Lithuania”

(study in parasitology):

- 19 calves, first all put at a naturally infected pasture in late spring; after 8 weeks, 9 calves moved to “safe” pasture (hay production),
- consider here faecal nematode eggs counts⁷ at 10 weeks,
- 2 possible data layouts: 10 and 9 rows in 2 separate columns for each group of calves, or 19 rows with all calves:

			egg counts		
calves	infect.	safe	calves	pasture	egg counts
1	52	8	1	infect.	52
2	30	34	2	infect.	30
3	70	46	3	infect.	70
4	36	0
5	100	38	10	infect.	30
6	70	26	11	safe	8
7	50	8	12	safe	34
8	54	10	13	safe	46
9	20	44
10	30		19	safe	44

⁷ scaled to: per 0.1g of faeces.

STATISTICAL CONCEPTS AND PROCEDURES

Distributions:

- tell us what values a variable takes, and how often,
- features: shape, center, spread, and deviations from overall shape,
- distributions of continuous and categorical variables are the same thing, but displayed in different ways,
 - * categorical distributions as a list of values and how often each value occurs,
 - * quantitative distributions often displayed in groups,
- 2 types of distributions:
 - * data (observed, or empirical, distributions),
 - * theoretical (that we use for modelling data).

Descriptive statistical analysis:

- descriptive statistics: plots, tables, simple statistics,
— choose methods appropriate for data at hand. . . ,
- purpose:
 - * provide overview of the data,
 - * detect errors / “different observations” (outliers⁸)
 - * focus attention on what’s relevant,
 - * aid subsequent modelling of the data.

⁸ Outlier (informal def.): observation that does not belong with the other values.

GRAPHS OF CATEGORICAL VARIABLES

Another data example: Migration to/from PEI July 2017 – June 2018:⁹

Province	Immigrants to PEI		Emigrants from PEI	
	Count	Proportion	Count	Proportion
NL	258	7.4%	183	4.6%
NS	451	12.9%	556	14.1%
NB	486	13.9%	374	9.5%
QC	199	5.7%	160	4.1%
ON	1266	36.2%	1674	42.5%
MB	31	0.9%	35	0.9%
SK	93	2.7%	41	1.0%
AB	460	13.2%	459	11.6%
BC	219	6.3%	426	10.8%
terr.	32	0.9%	33	0.8%
total	3495	100.1%	3941	99.9%

Bar graph:

- displays number in each group as a bar of corresponding height,
- generated in Minitab using the menu Graph–Bar chart (using “Values from a table”, and province as a categorical variable).

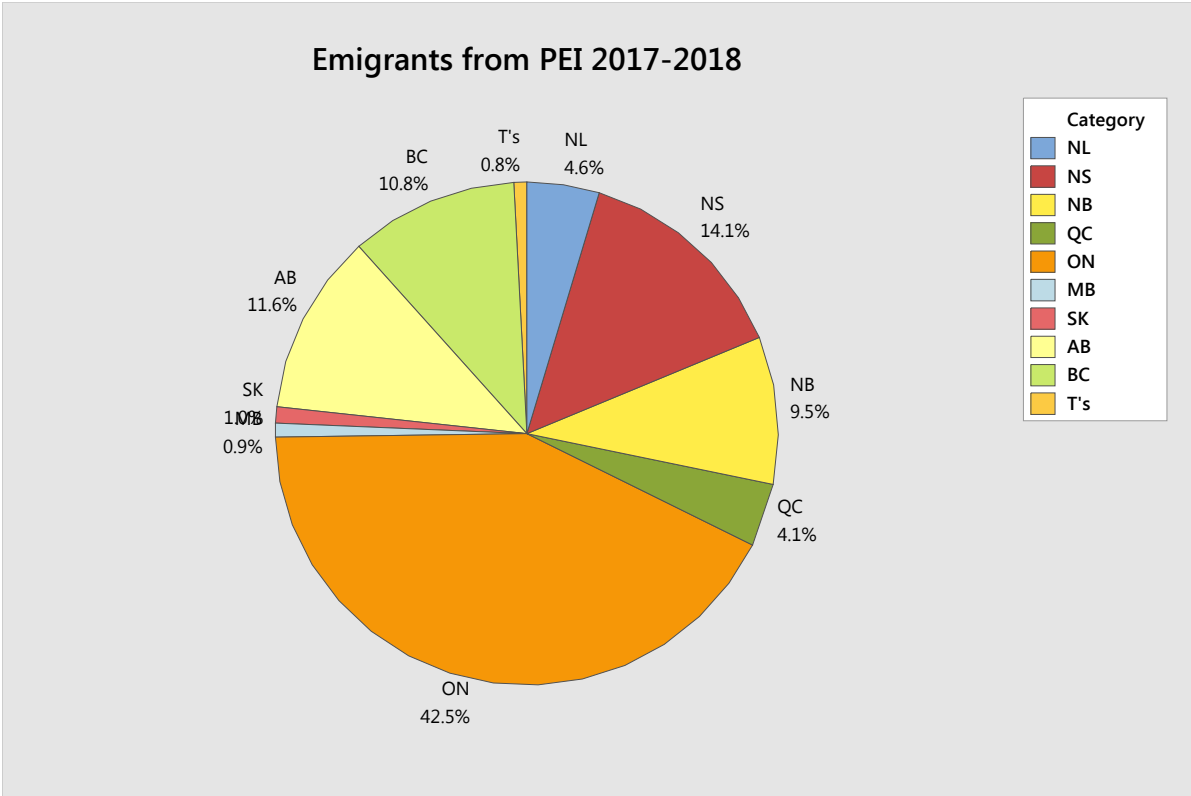
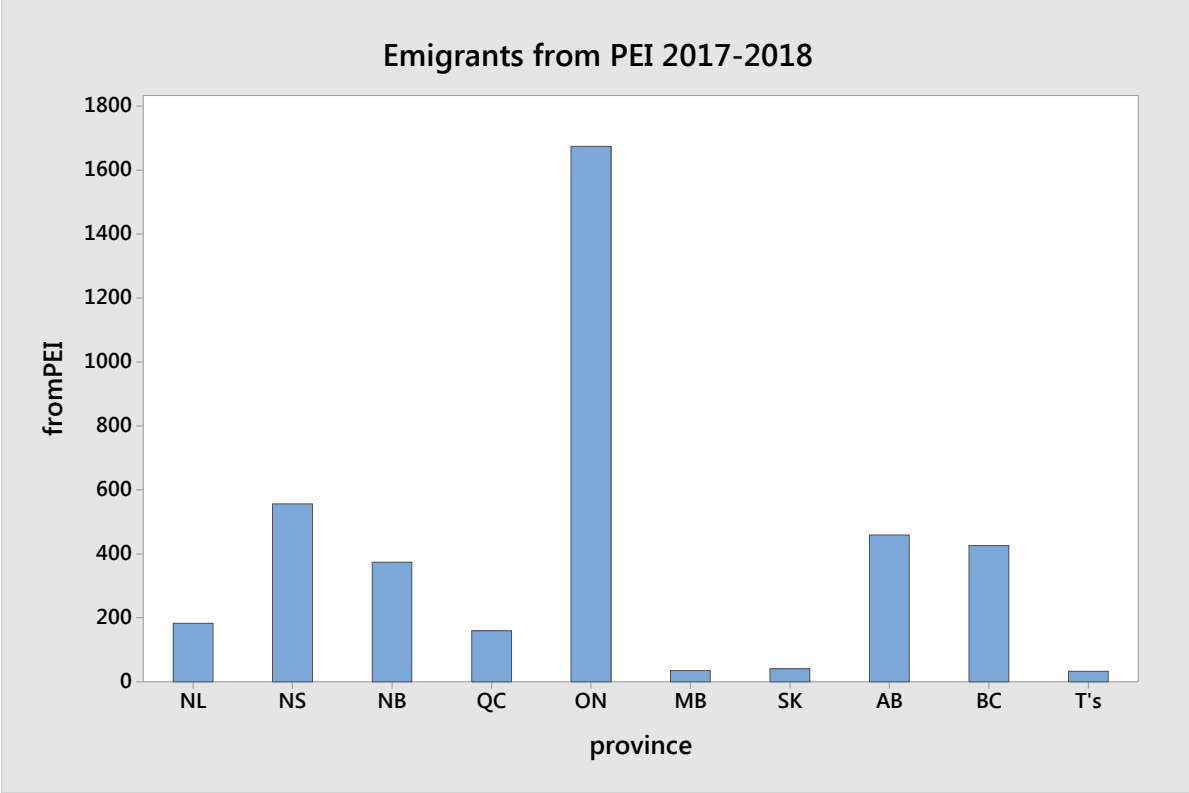
Pie chart:

- displays the proportions, or percentages (summing up to 100%), as corresponding parts of a pie,
- generated in Minitab using the menu Graph–Pie Chart (as above).

Note: pie charts are only for proportions; in bar graphs, the numbers need not to be seen relative to their total.

⁹ www.princeedwardisland.ca/sites/default/files/publications/web_2018_stat_review.pdf

BAR GRAPH AND PIE CHART



GRAPHS OF QUANTITATIVE DATA: STEMPLOTS

Stem plot (or, stem and leaf plot):

- more a (vertical) listing than a plot, displaying a distribution's pattern — shape, center, spread,
- requires first the separation of each data value into “leaf” and “stem” parts,
 - * for two-digit numbers (e.g. 15), the “stem” is usually the first digit (1), and the “leaf” is usually the last digit (5),
 - * for multi-digit numbers (e.g. 156), may truncate least important digits to retain two digits (i.e., $156 \rightarrow 150$),
 - * may split stems further before separating into stems/leaves,
- method of display:
 - * write stems in a vertical column sorted (increasingly) from top to bottom,
 - * write leaves to the right of stems (and separated by vertical line), and sorted out from the stem,
- back-to-back stem plots for two groups using same stems.

Stem plot (done manually) for parasite data, safe group:

Safe group (0, 8, 8, 10, 26, 34, 38, 44, 46)

```
0 | 088
1 | 0
2 | 6
3 | 48
4 | 46
```

STEMPLOTS PRODUCED BY SOFTWARE

Minitab plots (cannot be done side-by-side):

Infected			Safe		
1	2	0	3	0	088
4	3	006	4	1	0
4	4		(1)	2	6
(3)	5	024	4	3	48
3	6		2	4	46
3	7	00			
1	8				
1	9				
1	10	0			

Stata plots (cannot be done side-by-side):

Infected		Safe	
2*	0	0*	088
3*	006	1*	0
4*		2*	6
5*	024	3*	48
6*		4*	46
7*	00		
8*			
9*			
10*	0		

GRAPHS OF QUANTITATIVE DATA: HISTOGRAMS

Stem plots are good for small datasets, for larger datasets we often use histograms:

- divide the data range into classes of *equal* width,
- for each class, count the number of observations and draw corresponding bar/bin (no space between bars),
- may also plot the proportions (relative frequencies) by dividing with the total number of observations,
- procedures exist also for histograms of unequal width, where the bar area and not the height reflects the numbers (Exercise 1.22).

Demonstration by another data set: Weight in g of 162 crabs.

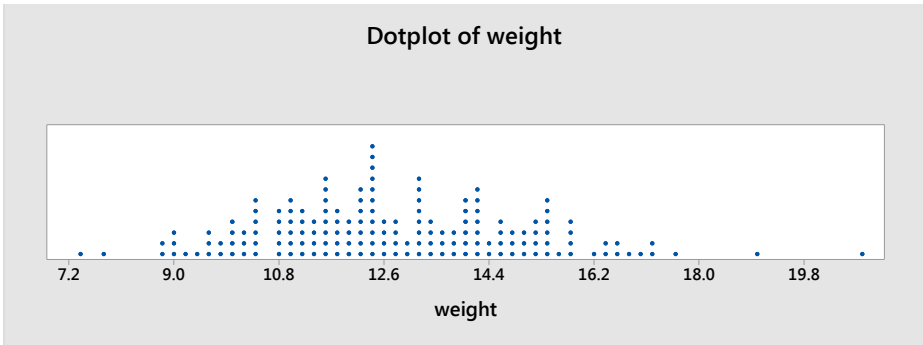
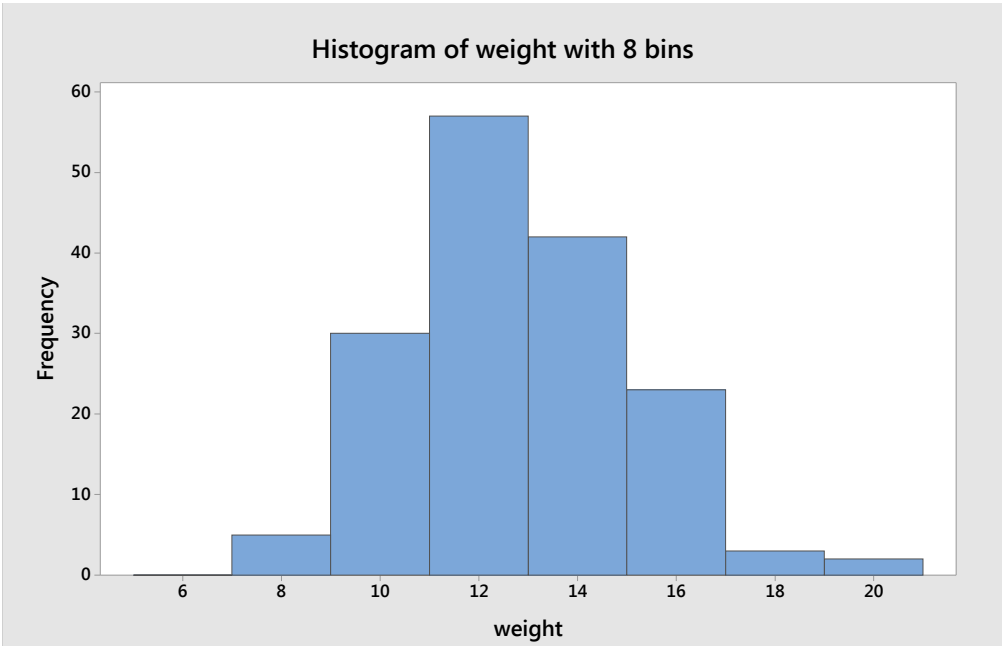
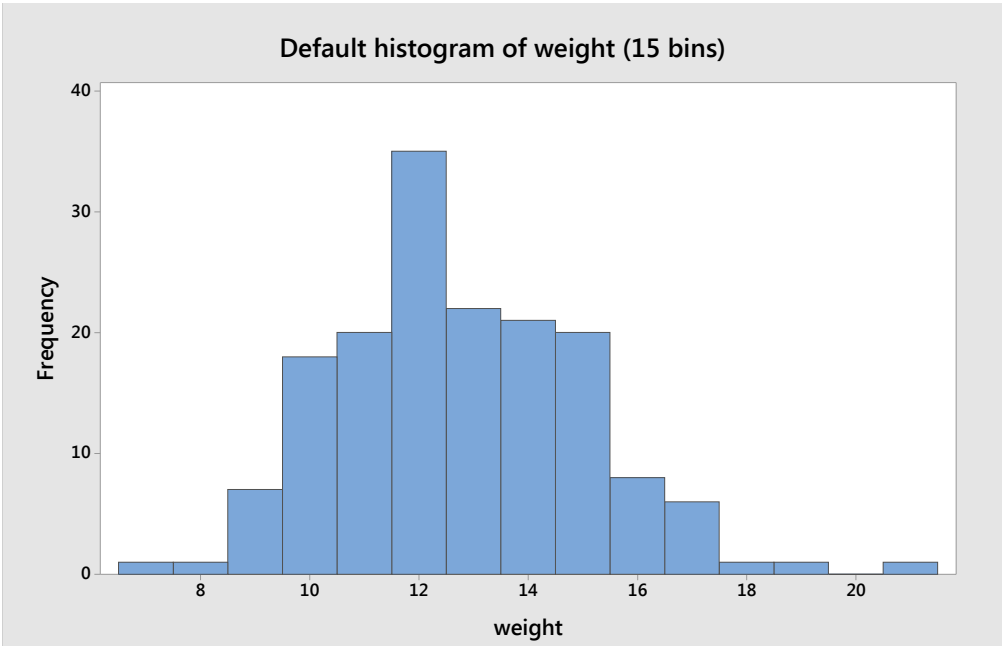
Histograms in practice:

- generated by computer program, here Minitab using command Graph–Histogram,
- displays the distribution's shape, and shows additional features such as mode(s) and symmetry/skewness,
- the number and location of bars affect the shape of the histogram; rules of thumb exist.¹⁰

In a dotplot, data points are marked above a horizontal axis: most useful for small n .

¹⁰ For n observations, the Stata default is \sqrt{n} bars for $n \leq 900$, and $10 \times \log_{10}(n)$ bars for $n > 900$, whereas the R default (Sturges' formula) is $(1 + \log_2(n))$ bars.

GRAPHS FOR CRAB WEIGHT DATA



MEAN AND MEDIAN

Mean (average) and median (middle value) are numeric quantities related to the center of a distribution, but different in definition and interpretation. We will first define them for an observed distribution (data).

Data: x_1, \dots, x_n , a total of n obs. in arbitrary order.

Mean/average:

$$\bar{x} = \frac{x_1 + \dots + x_n}{n} \quad \text{or} \quad \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i.$$

- most commonly used and perhaps most intuitive single value reported from a sample/dataset,
- can be appreciably affected by a few extreme obs.

Median:

- order the observations from smallest to largest,
- next step depends on whether n is odd or even:
 - * odd: median = observation in the middle, that is, number $(n+1)/2$ from either end,
 - * even: median = average of two middle observations, that is, numbers $n/2$ and $(n+2)/2$.

- has 50% of distribution to either side,
- is little affected by a few extreme observations
 \Rightarrow resistant (or robust).

Statistics for parasite data:

Pasture	n	mean	s	min	Q_1	median	Q_3	max
infected	10	51.2	24.0	20	30	51	70	100
safe	9	23.8	17.6	0	8	26	41	46

Symmetry/skewness and measures of center:

- symmetric distribution: mean = median,
- right-skewed distribution: mean $>$ median,
- left-skewed distribution: mean $<$ median.

Percentiles:

- other divisions of a distribution than 50:50,
- p th percentile:
 - * has p % below and $100-p$ % above,
 - * determined as $(p/100) \times (n+1)$ largest obs.
if value not number then either roundoff or interpolate between nearest obs. (software differences exist),
- special names: median M ($p = 50$), first quartile Q_1 ($p=25$), third quartile Q_3 ($p=75$).

SIMPLE MEASURES OF SPREAD

Spread (width) of an observed distrib. we can measure by:

- range:
 - * formula: $\max - \min$,
 - * simple and easy to compute, but not resistant,
 - * no theoretical counterpart in unbounded distributions (\Rightarrow tends to increase with sample size),
- interquartile range (IQR):
 - * formula: $Q_3 - Q_1$,
 - * more difficult to calculate, but resistant,
- 5-number summary of a distribution:
 - min, Q_1 , median, Q_3 , max,
 - * gives a fair overview of the distribution's shape,
 - * graphical representation = boxplot, available in Minitab using menu Graph–Boxplot.

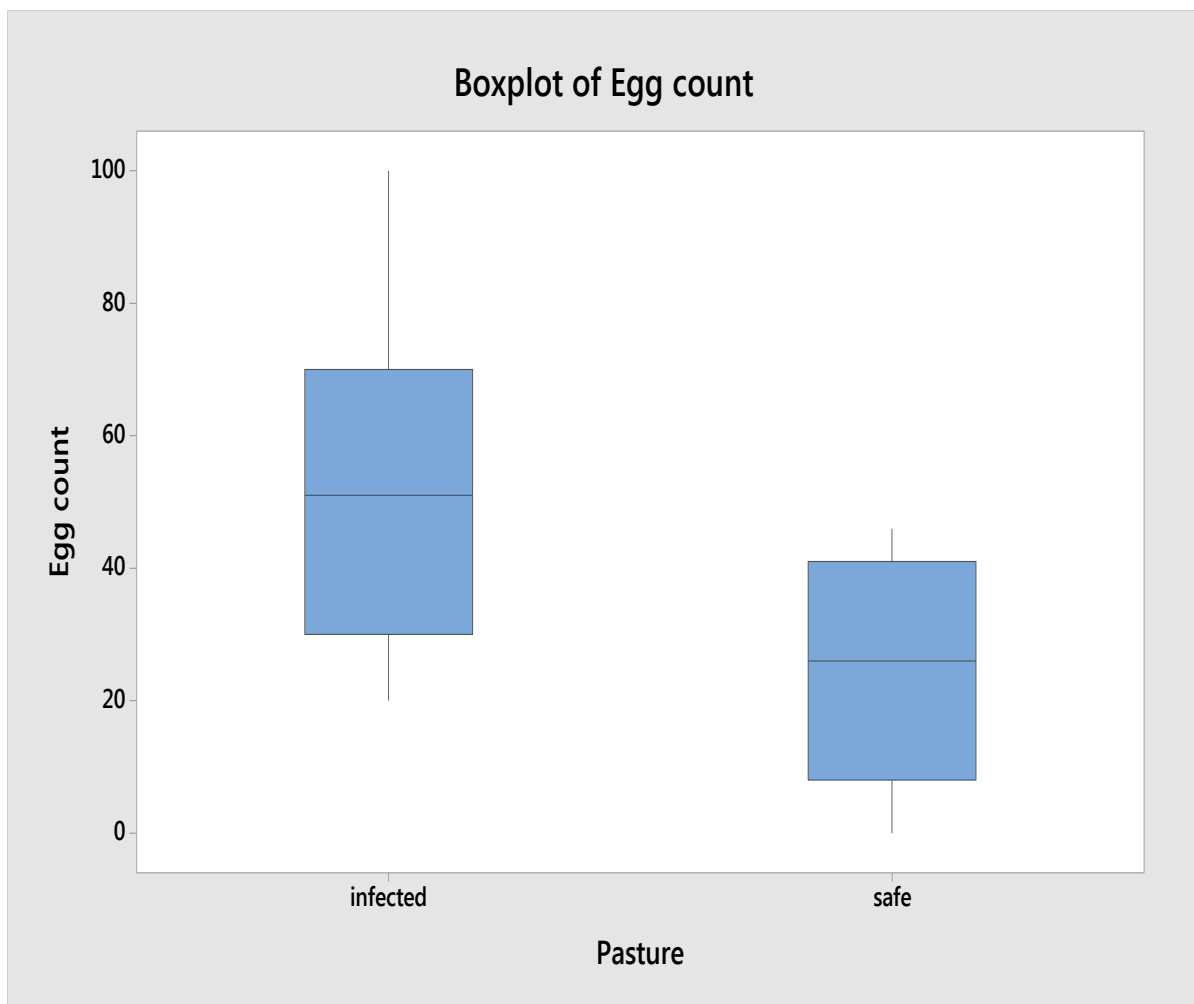
Interquartile criterion for suspected outliers:

- rule: “*suspected outlier*” if more than $1.5 \times \text{IQR}$ beyond Q_1 or Q_3 (\Rightarrow observations worth looking at...),
- a rough guideline — often indicates too many outliers,
- based on symmetric distribution; in normal distrib.: $1.5 \text{ IQR beyond } Q_3 (Q_1) \sim 99.65\% (0.35\%)$ percentile.

BOXPLOT FOR PARASITE DATA

Also called box and whisker plot:

- the box is formed by Q_1 and Q_3 ,
- the box is divided by the median (in some software, the mean is also indicated),
- some software allows to adjust the width of the box (e.g. proportional to the square-root of the number of obs.),
- the whiskers extend at most 1.5 IQR beyond the box,
- observations beyond the whiskers \sim asterisks.



STANDARD DEVIATION

Definition of variance s^2 and standard deviation s of an observed distribution x_1, \dots, x_n :

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2, \quad \text{and} \quad s = \sqrt{s^2}.$$

— that is, we average the *squared deviations from \bar{x}* .

Properties of s and s^2 :

- s most natural: same scale as observations themselves (s^2 on scale of squared observations)
- s^2 mathematically simplest (but less important to us),
- s measures spread about the mean,
- $s = 0$ means no dispersion (all observations equal), otherwise $s > 0$,
- s most commonly used measure of spread, and justifiably so (in my opinion), despite its pitfalls:
 - * s certainly not resistant (more sensitive to extremes than \bar{x}),
 - * as an overall measure, s does not take into account skewness in the distribution (with different spreads left and right of the center),
 - * never take for granted that a distrib. with a certain mean and spread looks like a nice (normal) distrib.,
- coefficient of variation (cv) = s/\bar{x} (“relative spread”).

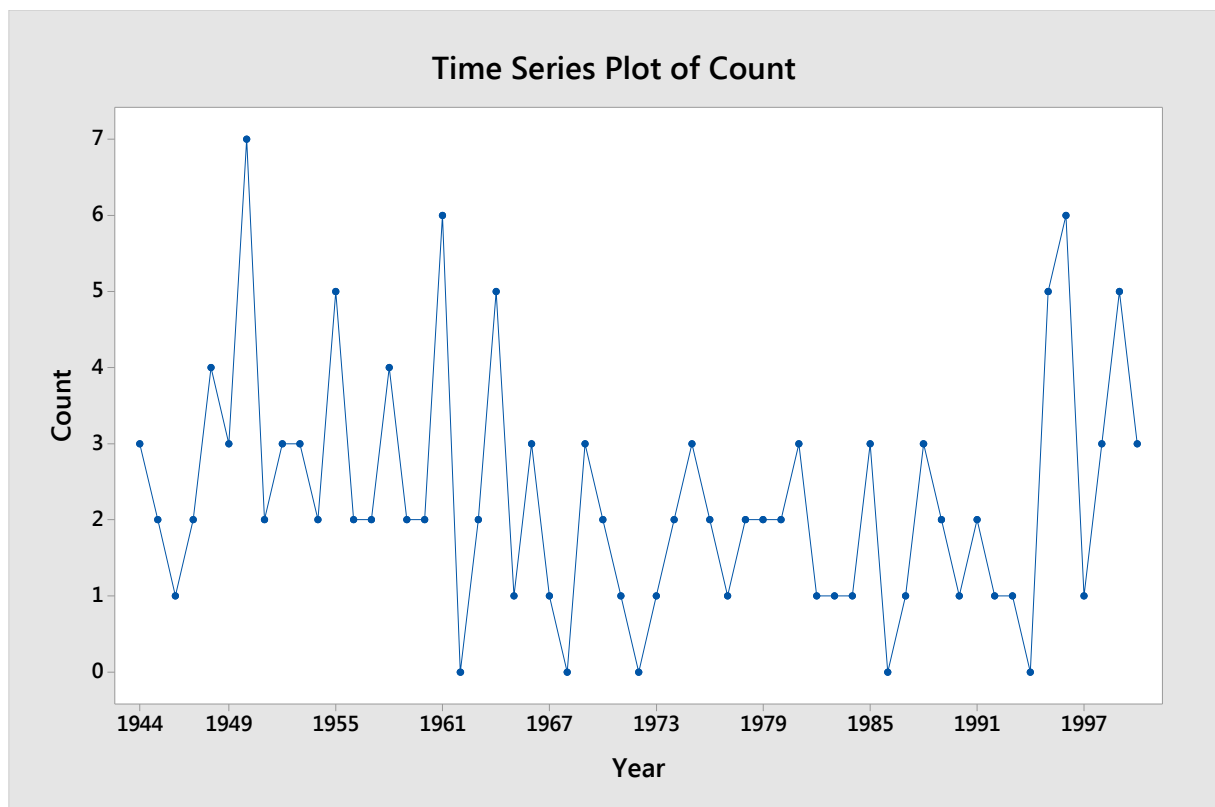
APPENDIX: TIME PLOTS

Time plots = plots against time (time as x -axis):
may be useful for descriptive purposes, e.g. to show,

- trend: persistent, long-term rise or fall,
- seasonal (periodic) variation: repeating patterns, not necessarily related to yearly seasons,
- data errors or unstable conditions developing over time,

Example: Hurricanes (Atlantic) in the United States
(Exercise 1.23):

- plot generated using Minitab command Graph—Time Series Plot (with Year as “Stamp” variable).



SUMMARY NOTES

Course content and organization:

- * applied statistics course based on use of statistical software and with focus on critical thinking about data and results,
- * course offers a wealth of options enabling each student to select the most effective learning tools,
- * class participation is not required but strongly recommended (both lectures and labs),
- * textbook reading is strongly recommended (and may be helpful as preparation for lectures and/or labs),
- * marks are based on submitted home assignments and exams.

Descriptive analysis involves graphical and numerical representations of data,

- * different techniques apply to quantitative and categorical data (and may need to be customized to actual data),
- * two primary features of distributions for quantitative data are: location (center) and spread.

Key words and concepts:

- data structured by cases (individuals) and variables,
- distributions (of observed data),
- outlier (outlying observation), and “suspected outlier”,
- mean, median, percentiles,
- interquartile range, standard deviation, 5-number summary,
- resistant statistic.