Index of 1-L

Practical Information

WELCOME (again)

Major news (and reminders :-):

- make sure to view the first introductory lecture ("0-L") and post your introduction,
- sorting out the textbook and software is even more urgent than last week...
- first lab session on Friday 4/9 (practical details in e-mail message for Week 1).

Today's lecture — Gentle start:

- aims of data analysis — and our first thoughts about the data,
  - their structure and organization in software,
  - useful terminology for variable types,
- descriptive statistics for distributions, in two forms:
  - graphical displays, e.g. histograms and box plots,[1]
  - numerical summaries, e.g. mean and median,[2]
- a few software (Minitab) demonstrations included, but software practice for you will be in the lab sessions.

[1] Textbook coverage: Baldi & Moore (PSLS 4e): Chapter 1; Moore, McCabe & Craig (IPS 7e): Section 1.1; Stephens (S): Chapter 3.

[2] Textbooks (see previous note for abbreviations): PSLS 4e: Chapter 2; IPS 7e: Section 1.2; S: Chapter 2.

What is Statistics?

2 branches of statistics:

- "official statistics" — the collection and display of figures (numbers) in statistical yearbooks, government publications etc.,
- "inferential statistics" — the science of analysing and interpreting data, from experiments or databases containing data collected for other purposes.

We all know that "With statistics one can prove anything"[3] — 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 biological significance...).

Same old statistics?[4] — actually, statistics is undergoing major changes these years,

- the presence of "big data" and development of "data science",
- the debate and critique of p-values and significance tests,
- the advance of Bayesian statistical methods.

[3] Quotes by (e.g.) George Canning; Homer Simpson.
[4] The development of modern statistics is usually credited to Ronald Fisher and colleagues in the early 20th century as well as, for Bayesian statistics, to Thomas Bayes (18th century).

Data Organization and Variable Types

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,
- typically organized in computer programs in a spreadsheet format with individuals as rows and variables as columns.

Types of variables:

- quantitative[5] (either continuous or discrete, i.e. with natural gaps):
  - 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 ordered (ordinal),
  - a quantitative variable may be split into categories.

[5] S further distinguishes between interval and ratio quantitative measurements.

Simple 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 counts[6] 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	infected	safe
1	52	8
2	30	34
3	70	46
4	36	0
5	100	38
6	70	26
7	50	8
8	54	10
9	20	44
10	30	

calves	pasture	egg counts
1	infected	52
2	infected	30
3	infected	70
...	...	...
10	infected	30
11	safe	8
12	safe	34
13	safe	46
...	...	...
19	safe	44

[6] 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,
- additionally, there are 2 types of distributions:
  - data (observed, or empirical, distributions),
  - theoretical (that we use for modelling data).

Descriptive statistical analysis:

- use: plots, tables, simple statistics — by methods appropriate for data at hand...,
- purpose:
  - provide overview of the data, and focus attention on what is relevant,
  - detect errors / "different observations" (outliers[7]),
  - aid subsequent modelling of the data.

[7] Outlier (informal definition): observation that does not belong with the other values.

Graphs of Categorical Variables

Another data example: Migration to/from PEI July 2018 – June 2019:[8]

Province	Immigrants to PEI Count	Immigrants to PEI Proportion	Emigrants from PEI Count	Emigrants from PEI Proportion
NL	180	4.6%	112	3.0%
NS	606	15.5%	563	14.8%
NB	288	7.3%	264	7.0%
QC	98	2.5%	160	4.2%
ON	1646	42.0%	1817	47.9%
MB	73	1.9%	48	1.3%
SK	47	1.2%	44	1.2%
AB	485	12.4%	486	12.8%
BC	451	11.5%	292	7.7%
Territories	48	1.2%	7	0.2%
total	3922	100.1%	3793	101.1%

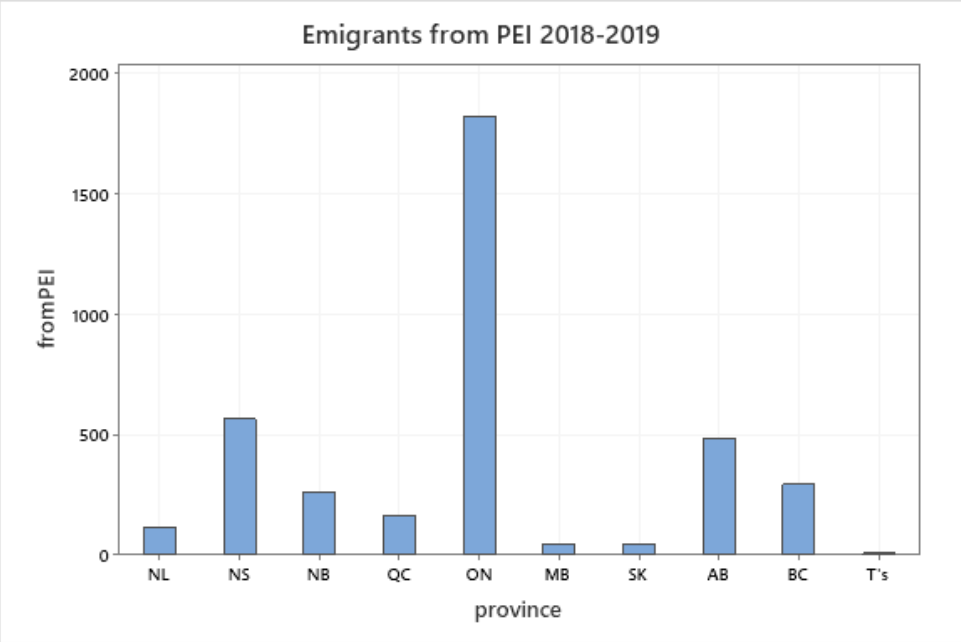
Graphs:

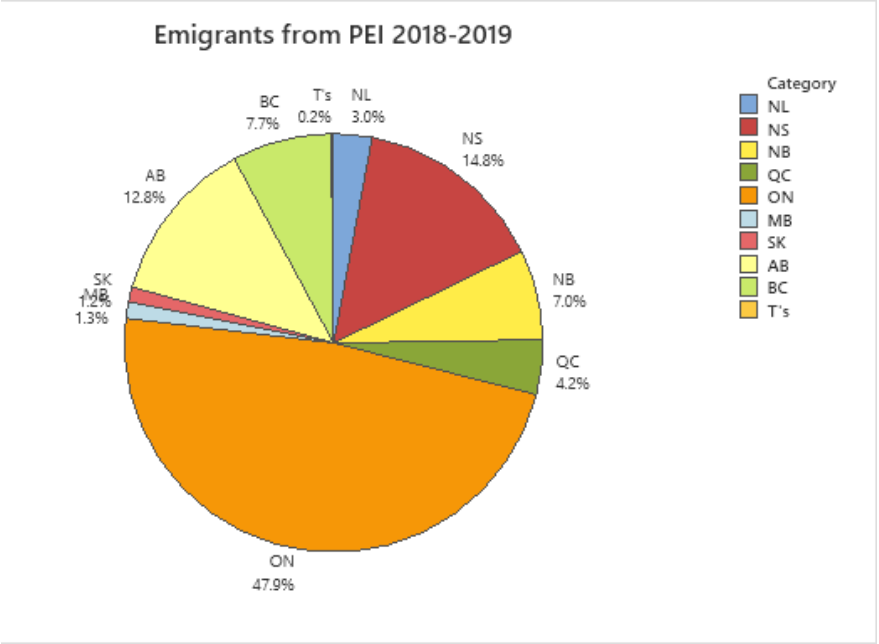
- bar graph: displays number in each group as a bar of corresponding height,
- pie chart: displays the proportions, or percentages (summing up to 100%), as corresponding parts of a pie,
- pie charts are only for proportions; in bar graphs, the numbers need not to be seen relative to their total.

[8] Table 5A from: www.princeedwardisland.ca/sites/default/files/publications/web/asr_2019.pdf

Bar Graph and Pie Chart for Migration Data

Graphs generated in Minitab
(using specifications below):





- (left) generated in Minitab using the menu Graph−Bar chart (using "Values from a table", and province as a categorical variable),
- (right) generated in Minitab using the menu Graph−Pie Chart (also using "Values from a table" etc.).

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,
- ideally, use back-to-back stem plots (with the same stems) for comparing two groups; however, this is rarely an option in statistical software.

Stemplots Produced by Software

Minitab plots: (cannot be done side-by-side in the software)

```
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 in the software)

```
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 and Dotplots

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

- divide the data range into classes of equal width,[9]
- 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.

Demonstration (next page) by another data set: weight in g of 162 crabs.

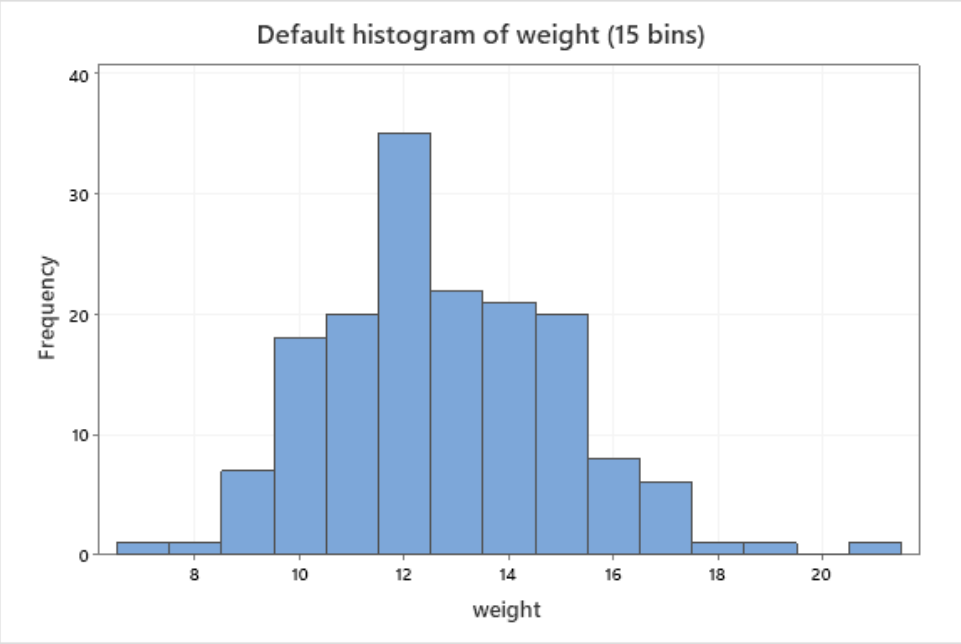
Histograms in practice:

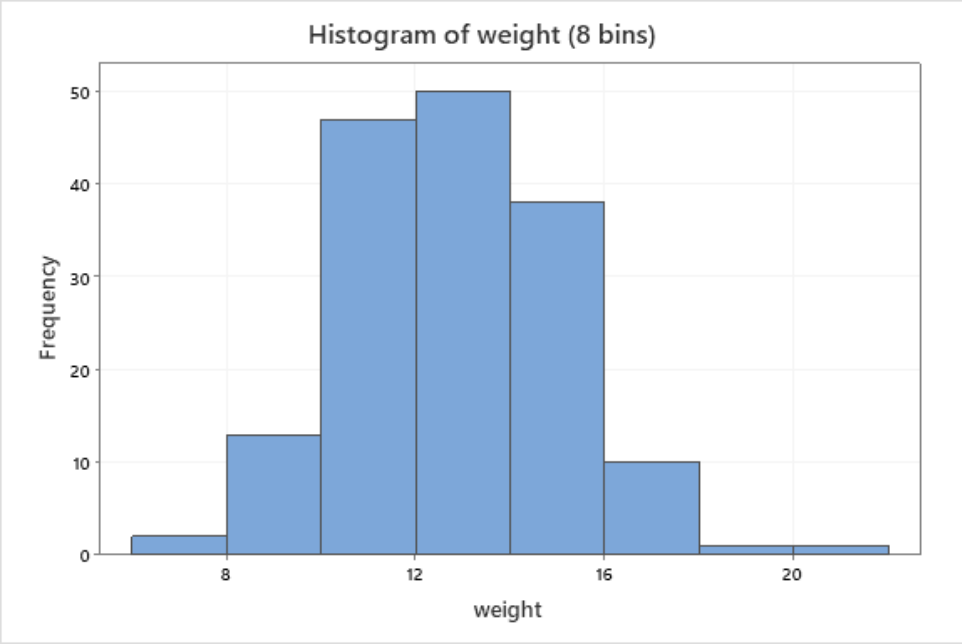
- generated by computer program; in Minitab, using command Graph−Histogram,
- displays the distribution's shape, and shows additional features such as mode(s)[10] and symmetry/skewness,
- the number and location of bars affect the shape of the histogram; rules of thumb exist.[11]

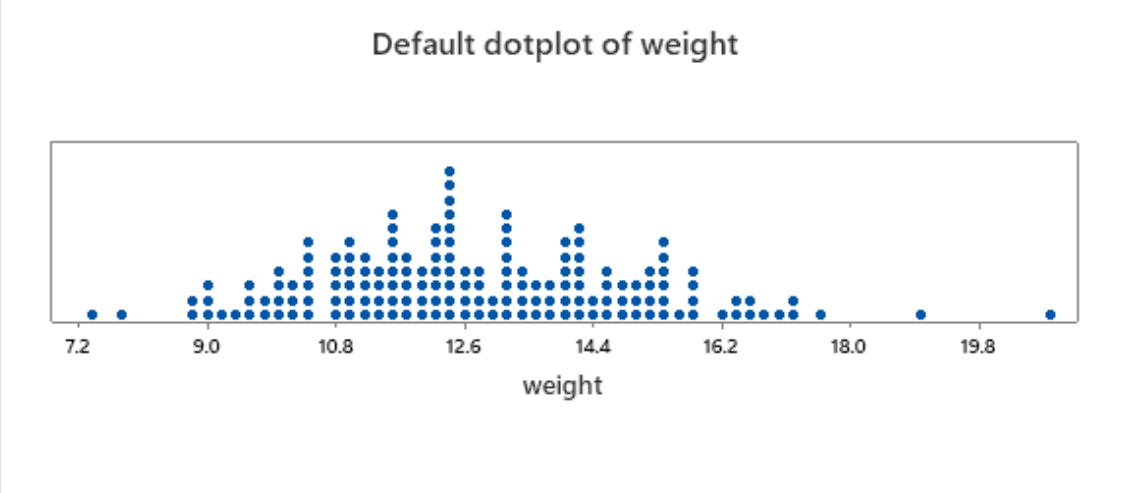
In a dotplot, data points are marked above a horizontal axis: most useful for small n.

[9] Histograms of unequal width exist as well, where the bar area and not the height reflects the numbers (Exercise 1.22).
[10] A mode shows as a (marked) peak in the histogram, and more than one mode may exist.
[11] 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: $x_1, \ldots, x_n$, — a total of n observations in arbitrary order.

Mean/average:

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

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

Median:

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

- has 50% of distribution to either side,
- is little affected by a few extreme observations ⇒ 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:

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

Simple Measures of Spread

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

- range, computed as: max − min,
  - simple and easy to compute, but not resistant,
  - no theoretical counterpart in unbounded distributions (⇒ tends to increase with sample size (not desirable)),
- interquartile range (IQR), computed as: $Q_3 - Q_1$,
  - more difficult to calculate, but resistant,
- 5-number summary of a distribution — (min, Q1, median, Q3, max),
  - gives a fair overview of the distribution's shape,
  - graphical representation = boxplot (next page).

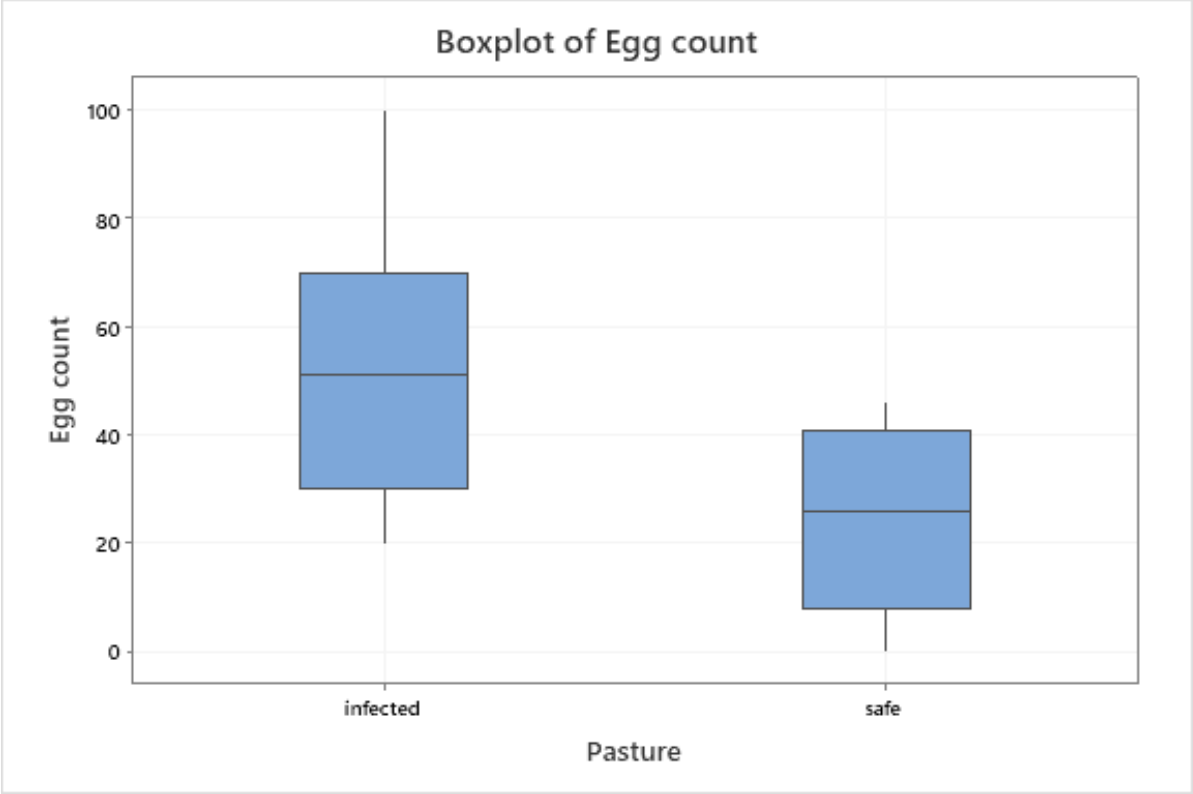
Interquartile criterion for suspected outliers:

- rule: "suspected outlier" if more than $1.5 \times$ IQR beyond Q_1 or Q_3
  ⇒ observations worth looking more closely at...,
  - for a normal distribution: 1.5 IQR beyond Q_3 (Q_1) $\sim$ 99.65% (0.35%) percentile,
- a rough guideline — based on a symmetrical distribution (but not all distributions are symmetrical), and often indicates too many outliers.

Boxplot for Parasite Data

Also called a box and whisker plot:

- the box is formed by Q_1 and Q_3, and is divided by the median (in some software, the mean is also indicated),[12]
- the whiskers extend at most 1.5 IQR beyond the box,
- observations beyond the whiskers $\sim$ asterisks,
- plot from Minitab's Graph−Boxplot menu.



[12] Some software also allows to adjust the width of the box (e.g. proportional to the square-root of the number of observations).

Standard Deviation

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

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

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

Properties of s and s^2:

- s most natural: same scale as the observations themselves,[13]
- s^2 mathematically simplest (but less important in this context),
- $s = 0$ means no dispersion (all observations equal), otherwise $s > 0$,
- s measures spread about the mean, and is the 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 distribution with a certain mean and spread looks like a nice (normal) distribution...,
- another definition: coefficient of variation (cv) $= s/\bar{x}$ ("relative spread").

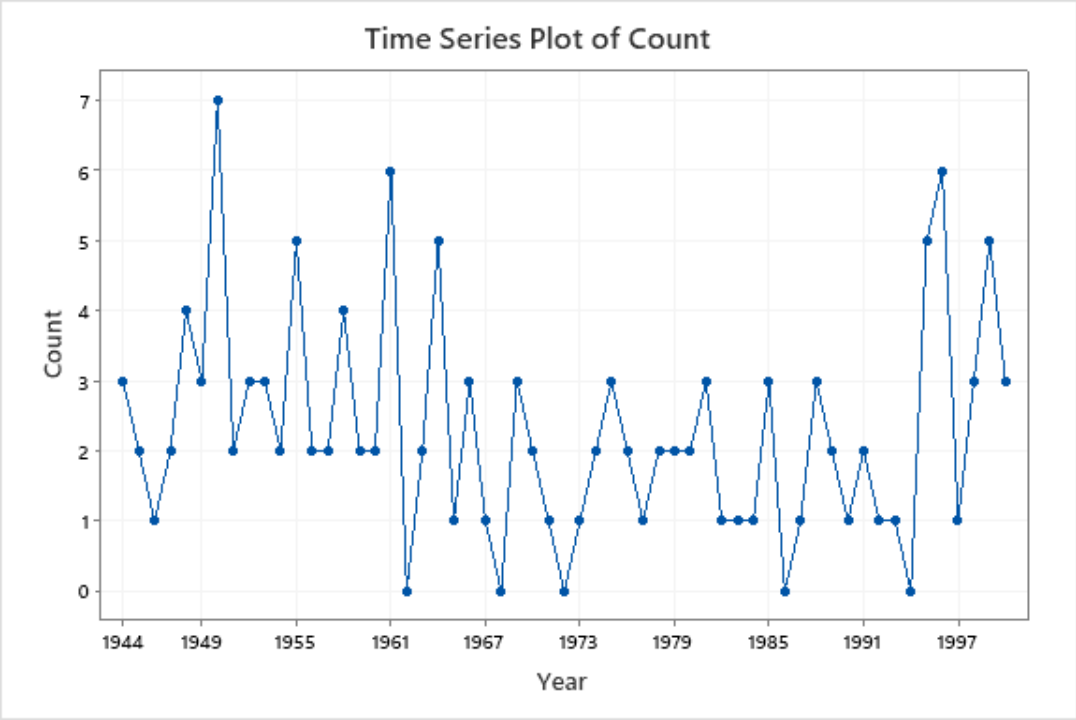
[13] Whereas s^2 is on scale the of squared observations.

Appendix: Time Plots

Time plots = plots against time (i.e., 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 yearly seasons,
- data errors or unstable conditions developing over time.

Example: yearly numbers of Atlantic hurricanes in the United States (Exercise 1.23):[14]



[14] Plot generated using Minitab command Graph−Time Series Plot (with Year as the "Stamp" variable).

Summary Notes

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,
* additional features of distributions are: symmetry/skewness, modality (uni- or multi-) and presence of extreme values.

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.