UCLA STAT 10

Introduction to Statistical Reasoning

- **Instructor**: **Ivo Dinov**, **Asst. Prof. in Statistics and Neurology**
- **Teaching Assistants:** , **Yan Xiong, Will Anderson UCLA Statistics**

University of California, Los Angeles, Winter 2002

http://www.stat.ucla.edu/~dinov/

UCLA STAT 10

Introduction to Statistical Reasoning

Course Description,

Class homepage,

online supplements, VOH's etc.

http://www.stat.ucla.edu/~dinov/

What is Statistics? A practical example

Demography: Uncertain population forecasts

by Nico Keilman, Nature 412, 490 - 491 (2001)

Traditional population forecasts made by statistical agencies **do not quantify uncertainty**. But demographers and statisticians have developed methods to calculate probabilistic forecasts.

The demographic future of any human population is uncertain, but some of the many possible trajectories are more probable than others. So, forecast demographics of a population, e.g., size by 2100, should include two elements: a range of possible outcomes, and a probability attached to that range.

What is Statistics?

Together, ranges/probabilities constitute a *prediction interval* for the population. There are trade-offs between greater certainty (higher odds) and better precision (narrower intervals). Why?

For instance, the next table shows an estimate that the odds are 4 to 1 (an 80% chance) that the world's population, now at 6.1 billion, will be in the range [5.6 : 12.1] billion in the year 2100. Odds of 19 to 1 (a 95% chance) result in a **wider** interval: [4.3 : 14.4] billion.

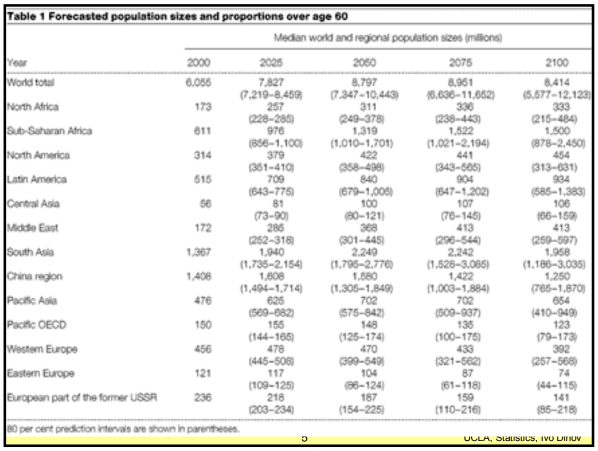


Table 1 Forecasted population sizes and proportions over age 60

	Median world and regional population sizes (millions)				
Year	2000	2025	2050	2075	2100
World total	6,055	7,827 (7,219–8,459)	8,797 (7,347–10,443)	8,951 (6,636–11,652)	8,414 (5,577–12,123)
North Africa	173	257 (228–285)	311 (249–378)	336 (238–443)	333 (215–484)
Sub-Saharan Africa	611	976 (856–1,100)	1,319 (1,010–1,701)	1,522 (1,021–2,194)	1,500 (878–2,450)
North America	314	379 (351–410)	422 (358–498)	441 (343–565)	454 (313–631)
Latin America	515	709 (643–775)	840 (679–1,005)	904 (647–1,202)	934 (585–1,383)
Central Asia	56	81 (73–90)	100 (80–121)	107 (76–145)	106 (66–159)
Middle East	172	285 (252–318)	368 (301–445)	413 (296–544)	413 (259–597)
South Asia	1,367	1,940 (1,735–2,154)	2,249 (1,795–2,776)	2,242 (1,528–3,085)	1,958 (1,186–3,035)
China region	1,408	1,608 (1,494–1,714)	1,580 (1,305–1,849)	1,422 (1,003–1,884)	1,250 (765–1,870)
Pacific Asia	476	625 (569–682)	702 (575–842)	702 (509–937)	654 (410–949)
Pacific OECD	150	155 (144–165)	148 (125–174)	135 (100–175)	123 (79–173)
Western Europe	456	478 (445–508)	470 (399–549)	433 (321–562)	392 (257–568)
Eastern Europe	121	117 (109–125)	104 (86–124)	87 (61–118)	74 (44–115)
European part of the former USSR	236	218 (203–234)	187 (154–225)	159 (110–216)	141 (85–218)

80 per cent prediction intervals are shown in parentheses.

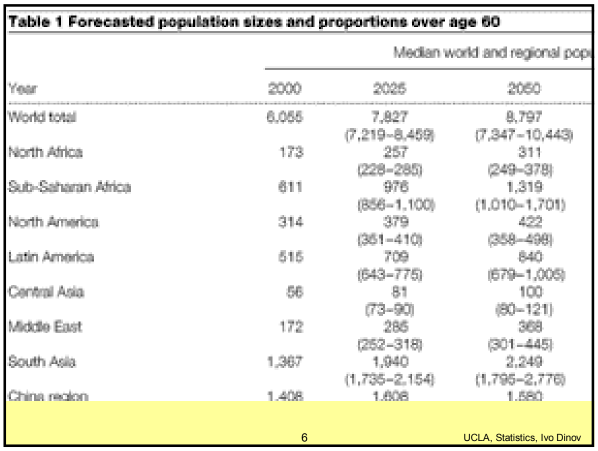


Table 1 Forecasted population sizes and proportions over age 60

		Median world and regional popu	
Year	2000	2025	2050
World total	6,055	7,827 (7,219–8,459)	8,797 (7,347–10,443)
North Africa	173	257 (228–285)	311 (249–378)
Sub-Saharan Africa	611	976 (856–1,100)	1,319 (1,010–1,701)
North America	314	379 (351–410)	422 (358–498)
Latin America	515	709 (643–775)	840 (679–1,005)
Central Asia	56	81 (73–90)	100 (80–121)
Middle East	172	285 (252–318)	368 (301–445)
South Asia	1,367	1,940 (1,735–2,154)	2,249 (1,795–2,776)
China region	1,408	1,608	1,580

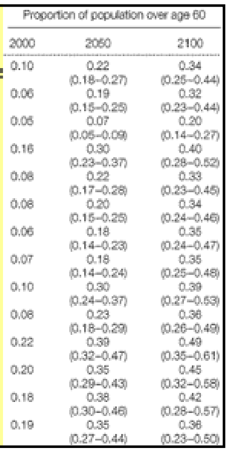
What is Statistics?

Demography: Uncertain population forecasts

by Nico Keilman, Nature 412, ,2001

Traditional population forecasts made by statistical agencies **do not quantify uncertainty**. But lately demographers and statisticians have developed methods to calculate probabilistic forecasts.

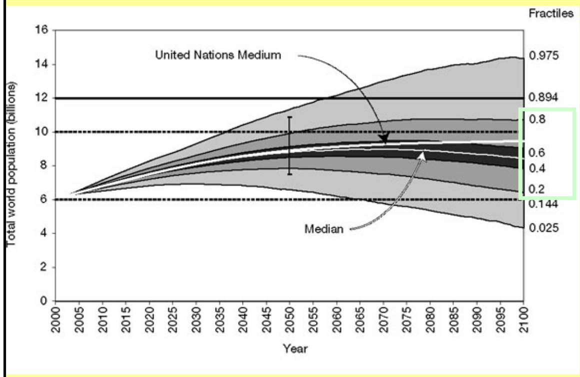
Proportion of population over 60yrs.



Proportion of population over age 60

2000	2050	2100
0.10	0.22 (0.18–0.27)	0.34 (0.25–0.44)
0.06	0.19 (0.15–0.25)	0.32 (0.23–0.44)
0.05	0.07 (0.05–0.09)	0.20 (0.14–0.27)
0.16	0.30 (0.23–0.37)	0.40 (0.28–0.52)
0.08	0.22 (0.17–0.28)	0.33 (0.23–0.45)
0.08	0.20 (0.15–0.25)	0.34 (0.24–0.46)
0.06	0.18 (0.14–0.23)	0.35 (0.24–0.47)
0.07	0.18 (0.14–0.24)	0.35 (0.25–0.48)
0.10	0.30 (0.24–0.37)	0.39 (0.27–0.53)
0.08	0.23 (0.18–0.29)	0.36 (0.26–0.49)
0.22	0.39 (0.32–0.47)	0.49 (0.35–0.61)
0.20	0.35 (0.29–0.43)	0.45 (0.32–0.58)
0.18	0.38 (0.30–0.46)	0.42 (0.28–0.57)
0.19	0.35 (0.27–0.44)	0.36 (0.23–0.50)

What is Statistics?



What is Statistics?

There is concern about the accuracy of population forecasts, in part because the rapid fall in fertility in Western countries in the 1970s came as a surprise. Forecasts made in those years predicted birth rates that were up to 80% too high.

The rapid reduction in mortality after the Second World War was also not foreseen; life-expectancy forecasts were too low by 1–2 years; and the predicted number of elderly, particularly the oldest people, was far too low.

What is Statistics?

So, during the 1990s, researchers developed methods for making probabilistic population forecasts, the **aim** of which is to calculate prediction intervals for every variable of interest.

Examples include population forecasts for the USA, AU, DE, FIN and the Netherlands; these forecasts comprised prediction intervals for variables such as age structure, average number of children per woman, immigration flow, disease epidemics.

We need accurate probabilistic population forecasts for the whole world, and its 13 large division regions (see Table). The conclusion is that there is an estimated 85% chance that the world's population will stop growing before 2100. Accurate?

What is Statistics?

There are three main methods of probabilistic forecasting:

- time-series extrapolation;
- expert judgment; and
- extrapolation of historical forecast errors.

Time-series methods rely on statistical models that are fitted to historical data. These methods, however, seldom give an accurate description of the past. If many of the historical facts remain unexplained, time-series methods result in excessively wide prediction intervals when used for long-term forecasting.

Expert judgment is subjective, and historic-extrapolation alone may be near-sighted.

Chapter 1

Preliminaries; Types of Measurements; Controlled Experiments

Types of variates

Qualitative Data

Quantitative Data

Hypothetical Data in a tabular form

Types of variates (variables)
(variate =data, variable = model)

We distinguish between two broad types of variables: qualitative and quantitative (or numeric). Each is broken down into two sub-types: qualitative data can be ordinal or nominal, and numeric data can be discrete (often, integer) or continuous.

Qualitative data always have a limited number of alternative values, such variables are also described as discrete. All qualitative data are discrete, while some numeric data are discrete and some are continuous.

For statistical analysis, quantitative data can be converted into discrete numeric data by simply counting the different values that appear.

Types of variables - Qualitative Data

Qualitative data arise when the observations fall into separate distinct categories.

Examples : Color of eyes : blue, green, brown etc

Exam result : pass or fail

Socio-economic status : low, middle or high.

Such data are inherently discrete, in that there are a finite number of possible categories into which each observation may fall.

Qualitative data are classified as:

nominal (Categorical) if there is no natural order between the categories (e.g., eye color), or

ordinal if an ordering exists (e.g., exam results, socio-economic status).

Types of variables - Quantitative Data

Quantitative data or numerical data arise when the observations are counts or measurements. The data are said to be discrete if the measurements are integers (e.g., number of people in a household, number of meals per day) and continuous if the measurements can take on any value, usually within some range (e.g., weight).

Quantities such as sex and weight are called variates, because the value of these quantities vary from one observation to another.

Numbers calculated to describe important features of the data are called statistics. For example, (i) the proportion of females, and (ii) the average age of unemployed persons, in a sample of residents of a town are statistics.

Types of variables - Quantitative Data

The following table shows a part of some (hypothetical) data on a group of 48 subjects.

'Age' and 'income' are continuous numeric variates, 'age group' is an ordinal qualitative variate, and 'sex' is a nominal (categorical) qualitative variate. The ordinal variate 'age group' is created from the continuous variate 'age' using five categories:

age group = 1	if age is	less than 20;
age group = 2	if age is	20 to 29;
age group = 3	if age is	30 to 39;
age group = 4	if age is	40 to 49;
age group = 5	if age is	50 or more

Types of variables

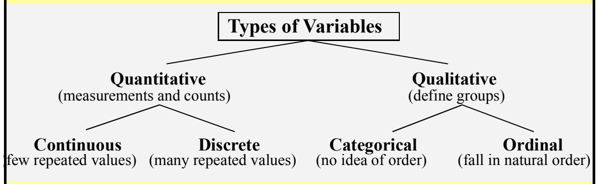


Table - Hypothetical Data

Subject No	Age (years)	Age Group	Annual Income (x $10,000)	Sex
1	32	3	4.1	F
2	20	2	1.5	M
3	45	4	2.4	F
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
47	19	1	0.5	F
48	32	3	1.9	F

Concept Grasping

1.A person's highest educational level is which type of variate?

- continuous
- discrete numeric
- ordinal
- nominal

2.The number of motor-vehicle accidents (in a section) of the Pacific Cost Highway in a week is which type of variate?

- continuous
- discrete numeric
- nominal
- ordinal

3. Nominal (categorical) data are often analyzed in the form of:

- counts
- averages
- ranks

Controlled Experiments

When a new drug is introduced its effectiveness needs to be evaluated. The basic method is comparison. Drug is administered to subjects in a treatment group and a second groups of subjects are used as controls (two groups should be randomly chosen).

Most of these experiments are carried as double-blind designs – neither the subjects taking the medicine nor the physicians who measure the response know which subject is in which group – to avoid biasing of the observed data.

Note: treatment and control groups need to be as similar (demographically) as possible, except for treatment.

Controlled Experiments

If the two groups differ in some factor, other than the treatment, we get this other factor possibly effecting the outcome of the study - this is called confounding effect. This should be avoided.

Example: Some disease fall more heavily on the poor. Hence, if a study tests the efficacy of a disease of hygiene (say) we'd need to match the subjects in the two groups for their socio-economic status (income), age, perhaps education level etc. to avoid the effects of these factors crippling into the results of the test.

Most often we use: randomized, controlled, double-blind studies - reduces the bias to a minimum.

Review

Data types: (quantitative, qualitative, etc.)

Population parameters and sample statistics

Controlled experiments

Confounding effects

Blindedness and placebo effects

Controlled Experiments

Randomized, controlled, double-blind studies are very hard to do, however. As a result sometimes we need to use designs that are not so perfect, but are more economical. Examples – using historical control groups.

Placebo groups: groups of subjects (patients) who receive fake treatment, sugar-pill, (not no-treatment, as in the treatment-control design). This design factors out the implicit psychological effects of been treated.

Randomization, Replication and Blocking

The use of chance to allocate experimental units into groups is called randomization. Randomization is the major principle of the statistical *design of experiments*.

Randomization produces groups of experimental units that are more likely to be similar in all respects before the treatments are applied than using non-random methods. At the end of the study if the differences in the outcome variable between the two groups is too large to attribute to chance, then the difference is called statistically significant. The decision about how large a difference is required to be ***significant*** depends on statistical inference using the laws of probability. This will be discussed in later sections.

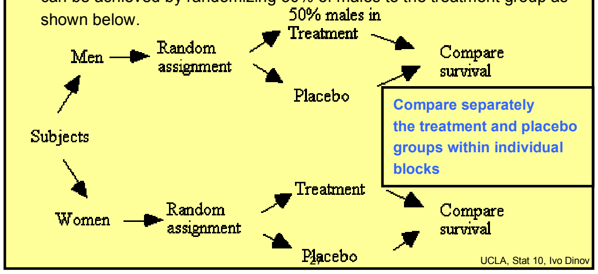
Randomization, Replication and Blocking

Another principle is that experiments with more subjects are more likely to detect differences than those with fewer subjects. Repeating an experiment on many subjects (or over many times) is called replication and increases the power of a statistical test.

If it is known, before the experiment is carried out, that other variables of no interest influence the outcome (e.g. age or sex of a patient), then randomization can be carried out within subsets of experimental units defined by these variables. This is called a block design.

Controlled Experiments - blocking

For example, the response to treatment for a given type of cancer is *expected* to depend on the sex of the patient. Ideally, equal numbers of males or females are required in the control or treatment groups and this can be achieved by randomizing 50% of males to the treatment group as shown below.



Compare separately the treatment and placebo groups within individual blocks

Chapter 2

Observational Studies

Observational Studies

Observational Studies are different from controlled experiments.

In controlled experiments the investigator decides who will be in the treatment and who will be in the control group.

In observational studies the subjects assign themselves to one of the groups – the researcher has no say, but just observes the outcome of the event. E.g., studying the effects of smoking – we can’t ask someone to smoke for 10 yrs just to satisfy the criteria of the study.

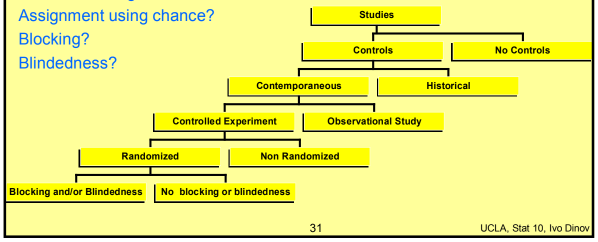
Observational Studies

Observational Studies can establish association between factors/predictors. Association may point to causation, but it can’t prove causality. The effects of treatment in observational studies, may be confounded with effects of factors that separated the units/subjects into control or treatment groups initially.

Examples?

Observational Studies

Identify subjects/unit .
Identify treatment.
Control group?
Treatment assignment?
Assignment using chance?
Blocking?
Blindedness?



Experimental vs. Observational studies

- A researcher wants to evaluate IQ levels are related to person's height. 100 people are are randomly selected and grouped into 5 bins: [0:50), [50;100), [100:150], [150:200), [200:250] *cm* in height. The subjects undertook an IQ exam and the results are analyzed.
- Another researcher wants to assess the bleaching effects of 10 laundry detergents on 3 different colors (R,G,B). The laundry detergents are randomly selected and applied to 10 pieces of cloth. The discoloration is finally evaluated.

Experimental vs. Observation study

- For each study, describe what *treatment* is being compared and what *response* is being measured to compare the treatments.
- Which of the studies would be described as *experiments* and which would be described as *observational* studies?
- For the studies that are observational, could an experiment have been carried out instead? If not, briefly explain why not.
- For the studies that are experiments, briefly discuss what *forms of blinding* would be possible to be used.
- In which of the studies has *blocking* been used? Briefly describe *what* was blocked and why it was blocked.

Experimental vs. Observation study

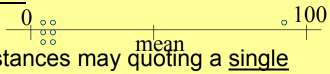
- What is the *treatment* and what is the *response?*
  1. Treatment is height (as a bin). Response is IQ score.
  2. Treatment is laundry detergent. Response is discoloration.
- *Experiment* or *observational* study?
  1. Observational – compare obs's (IQ) which happen to have the treatment (height).
  2. Experimental – experimenter controls which treatment is applied to which unit.
- For the observational studies, can we conduct an experiment?
  1. This could not be done as an experiment - it would require the experimenter to decide the (natural) height (treatment) of the subjects (units).
- For the experiments, is there *blinding?*
  2. The only form of blinding possible would be for the technicians measuring the cloth discoloration not to know which detergent was applied.
- Is there *blocking*?
  1. & 2. No blocking. Say, if there are two laundry machines with different cycles of operation and if we want to block we'll need to randomize which laundry does which cloth/detergent combinations, because differences in laundry cycles are a known source of variation.

Confounding Effects

Confounding means a difference between the treatment and control groups – other than the treatment – which effects the responses being studied. A confounder is a third variable which is associated with exposure and disease.

Mean, Median, Mode, Quartiles, 5# summary

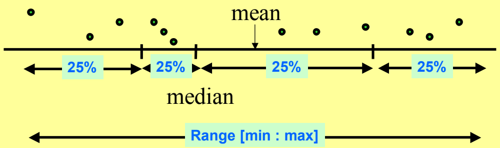
- The sample mean is the average of all numeric obs's.
- The sample median is the obs. at the index (n+1)/2 (note take avg of the 2 obs's in the middle for fractions like 23.5), of the observations ordered by size (small-to-large)?
- The sample median usually preferred to the sample mean for skewed data?



- Under what circumstances may quoting a single center (be it mean or median) not make sense?(multi-modal)
- What can we say about the sample mean of a qualitative variable? (meaningless)

Quartiles

The first quartile (Q_1) is the median of all the observations whose *position* is strictly below the *position* of the median, and the third quartile (Q_3) is the median of those above.



Five number summary

The five-number summery = (Min, Q_1, Med, Q_3, Max)

Chapter 3

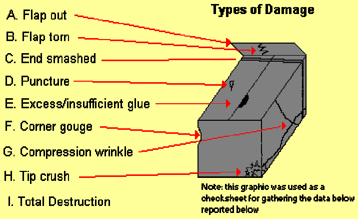
Frequency Distributions; Histograms

Frequency Distributions

A simple and effective way of summarizing discrete data is by counting the number of observations falling into each category. The number associated with each category is called the frequency and the collection of frequencies over all categories gives the frequency distribution of that variable.

The relative frequency is a number which describes the proportion of observations falling in a given category. This can be illustrated using the 'damaged boxes' example below. Observe which category a subject or object belongs to e.g. damaged box - corner gouge, tip crush, end smash. Count how many observations in each category - this gives 'frequency' or 'count' data. Tabulate results in frequency table showing frequencies or relative frequencies or percentages.

Frequency Distributions- **damaged boxes**



Note: this graphic was used as a checksheet for gathering the data below reported below

Frequency Distributions- **damaged boxes**

Type	Total Frequency	Relative Frequency	Percentage
A - Flap out	16	0.0096	1
B - Flap torn	17	0.0102	1
C - End smashed	132	0.0793	8
D – Puncture	95	0.0571	6
E - Glue problem	87	0.0523	5
→ F - Corner gouge	984	0.5913	59
G – Compr. wrinkle	15	0.0090	1
H - Tip crushed	303	0.1821	18
I - Tot. destruction	15	0.0090	1
Total	1664	0.9999*	100

(* the relative frequencies do not add to 1.0000 due to rounding)

Frequency Distributions- damaged boxes

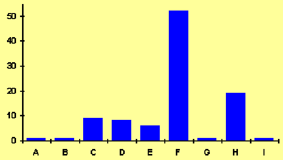
Relative frequency for type A is: $\frac{16}{1664} = 0.0096$

Percentage for type A is: $\frac{16}{1664} \times 100 = 0.96 \approx 1$ **percent.**

The usefulness of relative frequencies and percentages is clear: for example, it is easily seen that corner gouge accounts for 59% of the total number of damages.

Frequency Distributions- damaged boxes

The frequency distribution of a variable is often presented graphically as a bar-chart/bar-plot. For example, the data in the frequency table above can be shown as:



The vertical axis can be frequencies or relative frequencies or percentages. On the horizontal axis
all boxes should have the same width leave gaps between the boxes (because there is no connection between them) the boxes can be in any order.

Presenting continuous Data

The frequency distribution of a discrete quantitative variate may be summarized in a bar—graph.

The frequency distribution of a continuous quantitative variate can be constructed in the same way by first grouping the observations. That is, by choosing a set of contiguous, covering and non-overlapping intervals, called class intervals (or bins), the observations can be grouped to form a discrete variable from the continuous variable.

DEMO: SamplingDistributionApplet.html

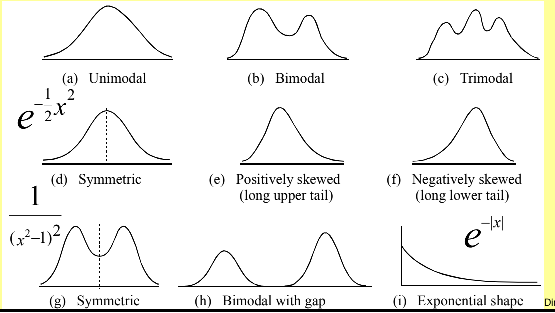
Presenting continuous Data

file:///C:/Ivo.dir/UCLA_Classes/Winter2002/AdditionalInstructorAids/NormalCurveInteractive.html

Histogram shapes

The natural number e is a constant: e ~ 2.7182…

.



(a) Unimodal (b) Bimodal (c) Trimodal

$e^{-\frac{1}{2}x^2}$

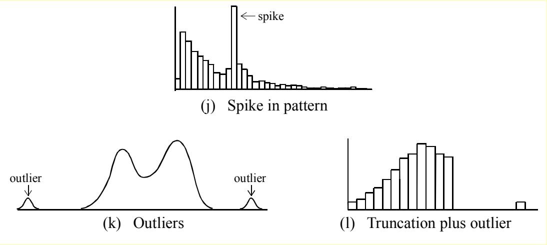
(d) Symmetric (e) Positively skewed (long upper tail) (f) Negatively skewed (long lower tail)

$\frac{1}{(x^2-1)^2}$

$e^{-|x|}$

(g) Symmetric (h) Bimodal with gap (i) Exponential shape

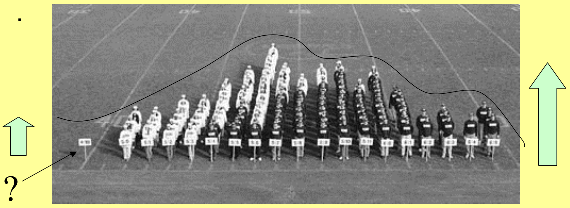
Histogram shapes – things to look for …



(j) Spike in pattern

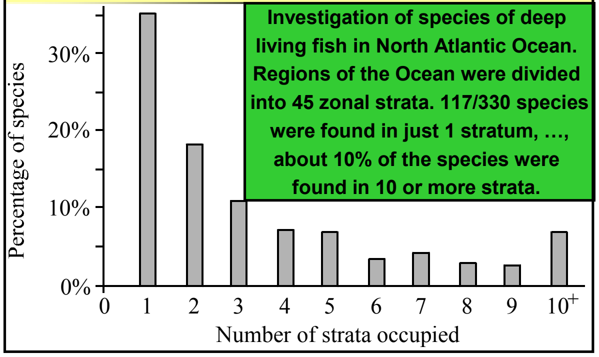
(k) Outliers (l) Truncation plus outlier

Histogram shapes – things to look for …



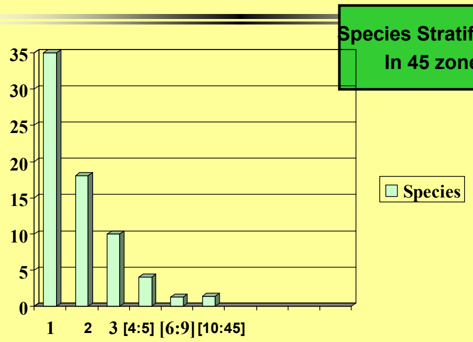
Subjects are 100 university genetics students, females in white and males in dark tops. Each student is in a bin corresponding to her/his height.

Histogram *density scale: height of each bar = (percentage of cases in strata)/(size of strata)*



Investigation of species of deep living fish in North Atlantic Ocean. Regions of the Ocean were divided into 45 zonal strata. 117/330 species were found in just 1 stratum, …, about 10% of the species were found in 10 or more strata.

Histogram *density scale: height of each bar = (percentage of cases in strata)/(size of strata)*



Species Stratification In 45 zones

TABLE 2.1.1 Data on Male Heart Attack Patients

A subset of the data collected at a Hospital is summarized in this table. Each patient has measurements recorded for a number of variables – ID, Ejection factor (ventricular output), blood systolic/diastolic pressure, etc.

- Reading the table
-Which of the measured variables (age, ejection etc.) are useful in predicting how long the patient may live.
-Are there relationships between these predictors?
-variability & noise in the observations hide the message of the data.

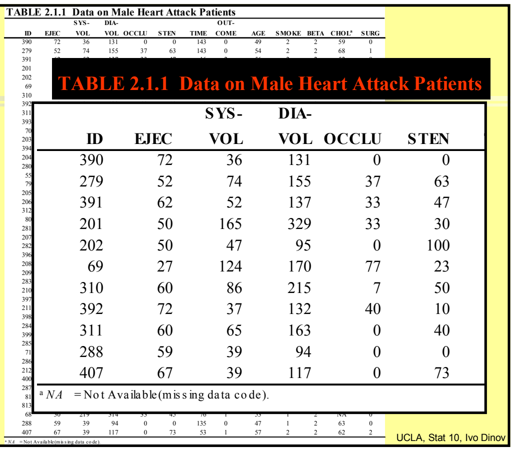


TABLE 2.1.1 Data on Male Heart Attack Patients

ID	EJEC	SYS- VOL	DIA- VOL	OCCLU	STEN	TIME	OUT- COME	AGE	SMOKE	BETA	CHOL[a]	SURG
390	72	36	131	0	0	143	0	49	2	2	59	0
279	52	74	155	37	63	143	0	54	2	2	68	1
288	59	39	94	0	0	135	0	47	1	2	63	0
407	67	39	117	0	73	53	1	57	2	2	62	2

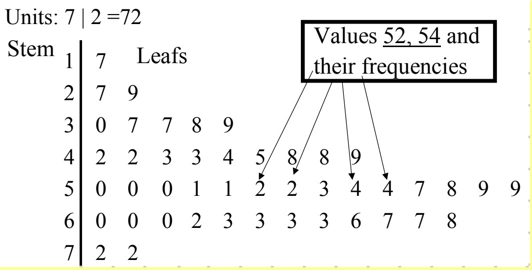
[a] NA = Not Available(missing data code).

TABLE 2.1.1 Data on Male Heart Attack Patients

ID	EJEC	SYS- VOL	DIA- VOL	OCCLU	STEN
390	72	36	131	0	0
279	52	74	155	37	63
391	62	52	137	33	47
201	50	165	329	33	30
202	50	47	95	0	100
69	27	124	170	77	23
310	60	86	215	7	50
392	72	37	132	40	10
311	60	65	163	0	40
288	59	39	94	0	0
407	67	39	117	0	73

[a] NA = Not Available(missing data code).

Example of a stem-and-leaf plot



```
Units: 7 | 2 =72
Stem
1 | 7     Leafs
2 | 7 9
3 | 0 7 7 8 9
4 | 2 2 3 3 4 5 8 8 9
5 | 0 0 0 1 1 2 2 3 4 4 7 8 9 9
6 | 0 0 0 2 3 3 3 3 6 7 7 8
7 | 2 2
```

Values 52, 54 and their frequencies

Stem-plot of the 45 obs's of the Ejection variable in the Heart Attack data table.

Traffic death-rates data

TABLE 2.3.1 Traffic Death-Rates (per 100,000 Population) for 30 Countries

17.4 Australia	20.1 Austria	19.9 Belgium	12.5 Bulgaria	15.8 Canada
10.1 Czechoslovakia	13.0 Denmark	11.6 Finland	20.0 France	12.0 E. Germany
13.1 W. Germany	21.1 Greece	5.4 Hong Kong	17.1 Hungary	15.3 Ireland
10.3 Israel	10.4 Japan	26.8 Kuwait	11.3 Netherlands	20.1 New Zealand
10.5 Norway	14.6 Poland	25.6 Portugal	12.6 Singapore	9.8 Sweden
15.7 Switzerland	18.6 United States	12.1 N. Ireland	12.0 Scotland	10.1England & Wales

Data for 1983, 1984 or 1985 depending on the country (prior to reunification of Germany)

Source: Hutchinson [1987, page 3].

```
Units: 17 | 4 = 17.4  deaths per  100,000
 5 | 4
 6 |
 7 |
 8 |
 9 | 8
10 | 1  1  3  4  5
11 | 3  6
12 | 0  0  1  5  6
13 | 0  1
14 | 6
15 | 3  7  8
16 |
17 | 1  4
18 | 6
19 | 9
20 | 0  1  1
21 | 1
22 |
23 |
24 |
25 | 6
26 | 8
      (a)
```

Round-off

Collapse to → 12 stems

```
Units: 1 | 7 = 17  deaths per  100,000
0 | 5
0 |
0 |
1 | 0  0  0  0  0  1  1
1 | 2  2  2  2  3  3  3
1 | 5  5
1 | 6  6  7  7
1 | 9
2 | 0  0  0  0  1
2 |
2 |
2 | 6  7
        (b)
```

Do we round off data for summaries/presentations, or for computations, prior to analyzing it?

FIGURE 2.3.7 Two stem-and-leaf plots for the traffic deaths dat

TABLE 2.3.2 Coyote Lengths Data (cm)

Females											
93.0	97.0	92.0	101.6	93.0	84.5	102.5	97.8	91.0	98.0	93.5	91.7
90.2	91.5	80.0	86.4	91.4	83.5	88.0	71.0	81.3	88.5	86.5	90.0
84.0	89.5	84.0	85.0	87.0	88.0	86.5	96.0	87.0	93.5	93.5	90.0
85.0	97.0	86.0	73.7								
Males											
97.0	95.0	96.0	91.0	95.0	84.5	88.0	96.0	96.0	87.0	95.0	100.0
101.0	96.0	93.0	92.5	95.0	98.5	88.0	81.3	91.4	88.9	86.4	101.6
83.8	104.1	88.9	92.0	91.0	90.0	85.0	93.5	78.0	100.5	103.0	91.0
105.0	86.0	95.5	86.5	90.5	80.0	80.0					

Coyotes captured in Nova Scotia, Canada. Data courtesy of Dr Vera Eastwood.

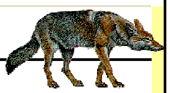


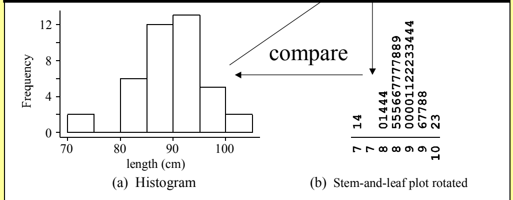
TABLE 2.3.3 Frequency Table for Female Coyote Lengths

Body length

Class Interval	Tally	Frequency	Stem-and-leaf plot
70-75⁻	||	2	7 | 1 4
75-80⁻		0	7 |
80-85⁻	𝍸 |	6	8 | 0 1 4 4 4
85-90⁻	𝍸 𝍸 ||	12	8 | 5 5 5 6 6 7 7 7 7 8 8 9
90-95⁻	𝍸 𝍸 |||	13	9 | 0 0 0 0 1 1 2 2 2 3 3 4 4 4
95-100⁻	𝍸	5	9 | 6 7 7 8 8
100-105⁻	||	2	10 | 2 3
Total		40	

TABLE 2.3.3 Frequency Table for Female Coyote Lengths

Class Interval	Tally	Frequency	Stem-and-leaf plot
70-75⁻	||	2	7 | 1 4
75-80⁻		0	7 |
80-85⁻	𝍸 |	6	8 | 0 1 4 4 4
85-90⁻	𝍸 𝍸 ||	12	8 | 5 5 5 6 6 7 7 7 7 8 8 9
90-95⁻	𝍸 𝍸 |||	13	9 | 0 0 0 0 1 1 2 2 2 3 3 4 4 4
95-100⁻	𝍸	5	9 | 6 7 7 8 8
100-105⁻	||	2	10 | 2 3
Total		40	



(a) Histogram (b) Stem-and-leaf plot rotated

Figure 2.3.8 Histogram of the female coyote-lengths data.

From *Chance Encounters* by C.J. Wild and G.A.F. Seber, © John Wiley & Sons, 2000.

Box plot compared to dot plot

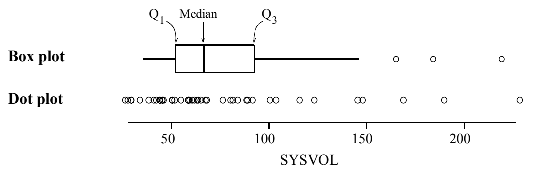
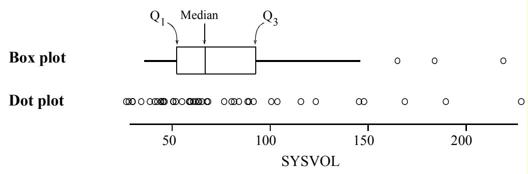


Figure 2.4.3 Box plot for SYSVOL.

From *Chance Encounters* by C.J. Wild and G.A.F. Seber, © John Wiley & Sons, 2000.

Box plot construction



1. Draw a box extending from Q_1 to Q_2, with a line across indicating the median
2. Calculate Q_1 –1.5 IQR and find the smallest observation not smaller than this number – this is the “left whisker”
3. Similarly find the “right whisker” – the Q_2 +1.5 IQR
4. Plot individually all obs’s outside the [L-whisker; R-whisker]

Construction of a box plot

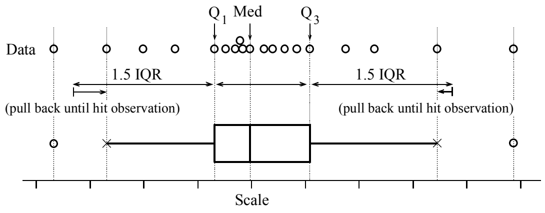

Q1
Med
Q3
Data
1.5 IQR
1.5 IQR
(pull back until hit observation)
(pull back until hit observation)
Scale

Figure 2.4.4 Construction of a box plot.

From *Chance Encounters* by C.J. Wild and G.A.F. Seber, © John Wiley & Sons, 2000.

Comparing 3 plots of the same data

```
Stem-and-leaf of strength  N  = 33
Leaf Unit = 10

  1   19 8
  5   20 0334
  5   20
 10   21 00233
 (8)  21 55668899
 15   22 000111112
  6   22 5
  5   23 014
  2   23
  2   24
  2   24
  2   25 2
  1   25 9
```

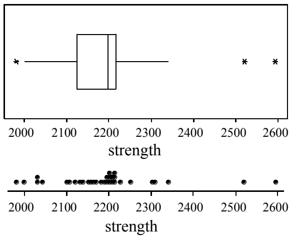


Figure 2.4.5 Three graphs of the breaking-strength data for} gear-teeth in positions 4 & 10 (Minitab output).

Frequency Table

TABLE 2.5.1 Word Lengths for the First 100 Words on a Randomly Chosen Page

3	2	2	4	4	4	3	9	9	3	6	2	3	2	3	4	6	5	3	4
2	3	4	5	2	9	5	8	3	2	4	5	2	4	1	4	2	5	2	5
3	6	9	6	3	2	3	4	4	4	2	2	4	2	3	7	4	2	6	4
2	5	9	2	3	7	11	2	3	6	4	4	7	6	6	10	4	3	5	7
7	7	5	10	3	2	3	9	4	5	5	4	4	3	5	2	5	2	4	2

Frequency Table

Value u	1	2	3	4	5	6	7	8	9	10	11
Frequency f	1	22	18	22	13	8	6	1	6	2	1

Mean from a frequency table

$$\bar{x} = \frac{1}{n}\text{Sum of (value} \times \text{frequency of occurrence)} =$$

$$\frac{1}{n}\text{(Sum of all observations)}$$

TABLE 2.5.2
Frequency Table for the Occurrence of Fish Species in Ocean Strata

No. of strata in which species occur (u_j)	**Frequency (No. of species)** (f_j)	**Percentage of species** $(\frac{f_j}{n} \times 100)$	**Cumulative Percentage**
1	117	35.5	35.5
2	61	18.5	53.9
3	37	11.2	65.2
4	24	7.3	72.4
5	23	7.0	79.4
6	12	3.6	83.0
7	14	4.2	87.3
8	10	3.0	90.3
9	9	2.7	93.0
10+	23	7.0	100.0
	n = 330	100	

Source: Haedrich and Merrett [1988]