EE3211
Modelling Techniques

Lecture 1
Descriptive Statistics

Topics to be covered:

- Descriptive statistics
- Probability distributions
- Estimation
- Hypothesis testing
- Regression and correlation methods
- Logistic regression
- Nonparametric methods
- Multisample Inference

Assessment criteria

- Continuous Assessment (60%)
  - In-class Exercise / Project (20%)
    - Skill test (5%)
    - Project (15%)
  - Test (20%)
  - Assignments (20%)
- Examination (40%)

Project

- Data source: NHANES publicly available datasets
- Choose one out of three available projects to work on
- Write a report which includes:
  - Background and objective (3%)
  - Methods (conduct statistical analyses with at least two methods taught) (5%)
  - Results (3%)
  - Conclusion (4%)

*within five A4 pages

(Deadline: End of Week 13, i.e. April 23rd 2021)

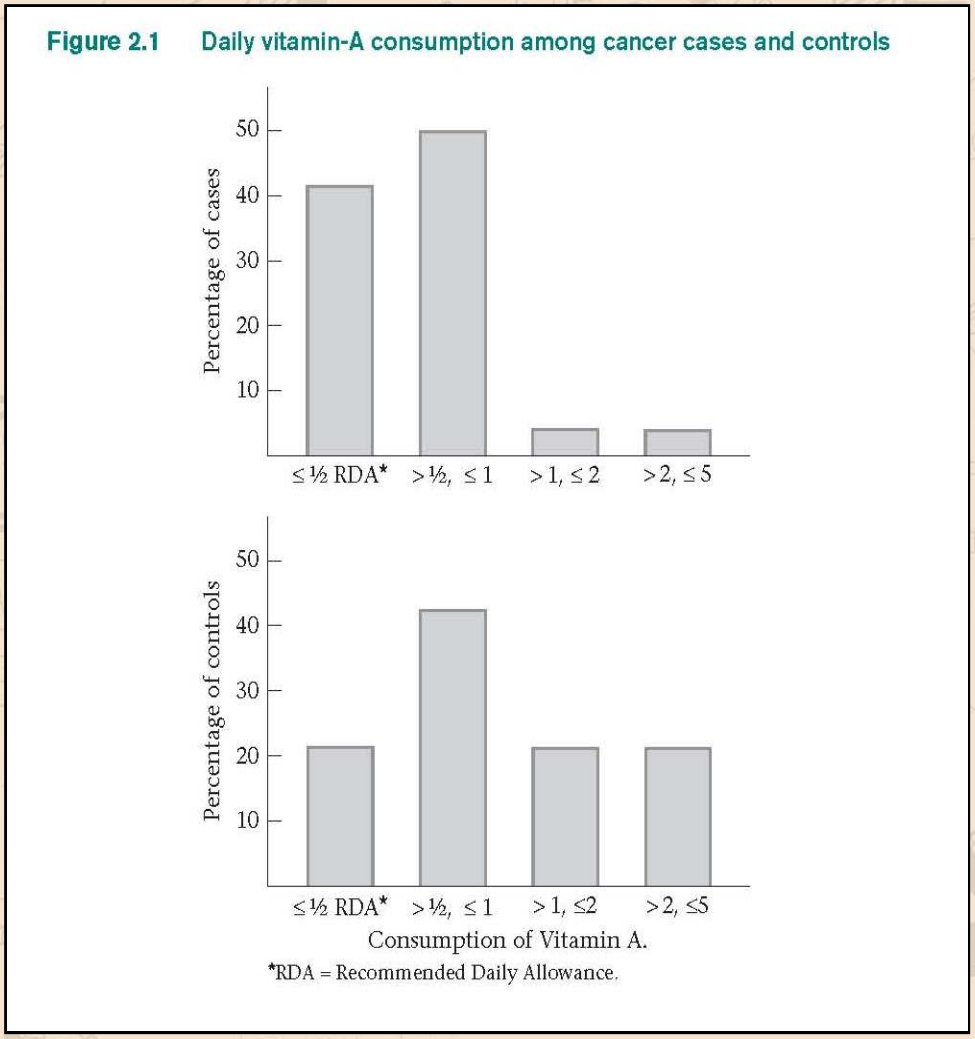
Descriptive Statistics

Why do we need descriptive statistics?

- Describe the data in some concise manner
- Numeric (table) or graphic : capturing and conveying the final results

- Good numeric (table) or graphic form of data summarization:
  - ➢Self-contained
  - ➢Understandable without reading the text
  - ➢Clearly labeled of attributes with well-defined terms
  - ➢Indicate principal trends in data

Bar graphs: Vitamin A and Cancer

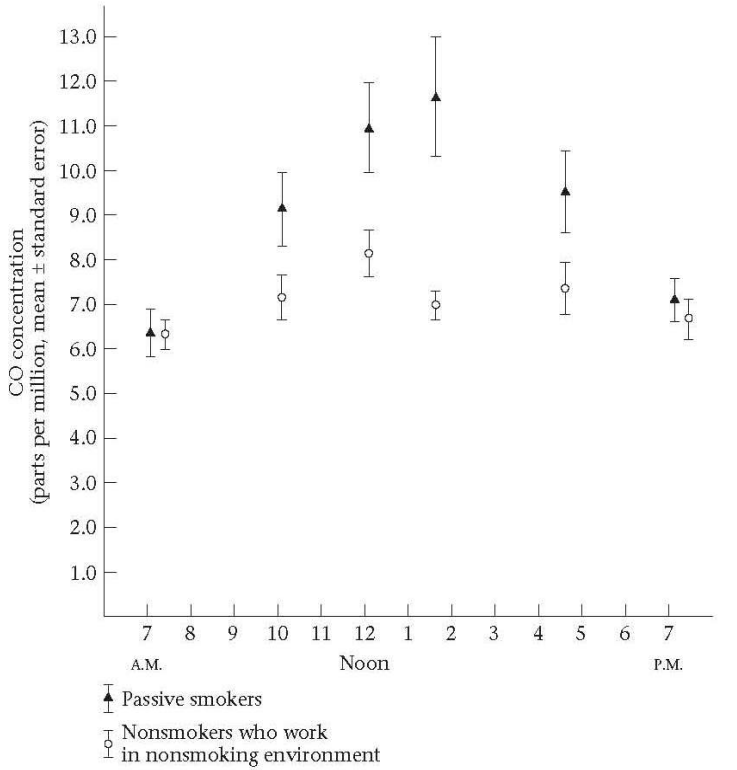


- Vitamin A consumption prevents cancer
- Dietary questionnaire: vitamin A
- Total cancer cases: 200
  Total matched controls (age and gender; unrelated disease): 200

<u>Vitamin A consumed by controls is more than that consumed by the patients with cancer.
(In some cases, the levels exceed the recommended daily allowance (RDA).)</u>

Scatter plot:
Passive smoking and pulmonary function

Figure 2.2 Mean carbon-monoxide concentration (± standard error) by time of day as measured in the working environment of passive smokers and in nonsmokers who work in a nonsmoking environment



Source: Reproduced with permission of *The New England Journal of Medicine*, 302, 720–723, 1980.

- Measure CO in the working environments of passive smokers and non-smoking workplace

<u>Early in the day: CO concentrations are about the same in the working environments of passive smokers and non-smokers</u>

- Supports the observation that passive smokers have lower pulmonary function than comparable nonsmokers

Descriptive statistics: graph

- Shows the key role of descriptive statistics:
  - Display data to give researcher a clue about main trends in the data
  - Suggest hints where to look for more details at the data (inferential statistics)

- Important to convey the findings in publications
- Influence the readers' impression of the work

Measure of Location

- No. of sample points: large
  -difficult to look at each sample points

- **Measure of location:** good for data summarization defining the center or middle of the sample

Measure of Location: The Arithmetic Mean

Table 2.1 Sample of birthweights (g) of live-born infants born at a private hospital in San Diego, California, during a 1-week period

i	x_i	i	x_i	i	x_i	i	x_i
1	3265	6	3323	11	2581	16	2759
2	3260	7	3649	12	2841	17	3248
3	3245	8	3200	13	3609	18	3314
4	3484	9	3031	14	2838	19	3101
5	4146	10	2069	15	3541	20	2834

- **Arithmetic mean**: sum of all the observations divided by the number of observations
- Statistically expressed as $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Limitation: Oversensitive to extreme values; not be representative of the location of the majority

Measure of Location: The Arithmetic Mean

Sigma $\sum$ is a summation sign

$\sum_{i=1}^{n} x_i$ implies $(x_1 + x_2 + \ldots + x_n)$

- If a and b are integers where $a \leq b$, then $\sum_{i=a}^{b} x_i = (x_a + x_{a+1} + \ldots + x_b)$
- If a = b, then $\sum_{i=a}^{b} x_i = x_a$
- If c is some constant, then $\sum_{i=1}^{n} cx_i = c\left(\sum_{i=1}^{n} x_i\right)$

Some Properties of Arithmetic Mean

- Original samples : $x_1, \ldots, x_n$
- Translated samples : $x_1 + c, \ldots, x_n + c$ (where c is some constant)
- Let $y_i = x_i + c$ $\quad\quad$ $i = 1, \ldots, n$

then $\bar{y} = \bar{x} + c$

* Change the "origin" of the sample data

TABLE 2.6 Translated sample for the duration between successive menstrual periods in college-age women

Value	Frequency	Value	Frequency	Value	Frequency
–4	5	1	96	6	7
–3	10	2	63	7	3
–2	28	3	24	8	2
–1	64	4	9	9	1
0	185	5	2	10	1

Note: $\bar{y} = [(-4)(5) + (-3)(10) + \ldots + (10)(1)]/500 = 0.54$

$\bar{x} = \bar{y} + 28 = 0.54 + 28 = 28.54$ days

- If the unit or scale changes, then using the **rescaled sample**

$$y_i = cx_i \qquad i = 1, \ldots, n$$

- Arithmetic mean : $\bar{y} = c\bar{x}$

- Let $x_1, \ldots, x_n$ be the original sample of data.
- Let $y_i = c_1 x_i + c_2 \qquad i = 1, \ldots, n$ represent a transformed

- sample obtained by multiplying each original sample point by a factor c_1 and then shifting over by a constant c_2

- If $y_i = c_1 x_i + c_2 \qquad i = 1, \ldots, n$

then $y = c_1 x + c_2$

*change both the origin and the scale of the data

Measure of Location: Median

- N observations in a sample (smallest to largest)
- Sample median is

➢ $\left(\frac{n+1}{2}\right)$ th the largest observation if n is odd

➢ Average of the $\left(\frac{n}{2}\right)$ th and the $\left(\frac{n}{2}+1\right)$ th observation if n is even

---

White-blood counts of admitted patients

Table 2.2 Sample of admission white-blood counts (× 1000) for all patients entering a hospital in Allentown, PA, on a given day

i	x_i	i	x_i
1	7	6	3
2	35	7	10
3	5	8	12
4	9	9	8
5	8		

- Order the sample:
  3, 5, 7, 8, 8, 9, 10, 12, 35
- n is odd, sample median of white-blood cell count:
  8 (5^{th} largest point)

Measure of Location: Median

- Strength:
  - insensitive to very large or very small values
  - E.g. white blood cell counts:

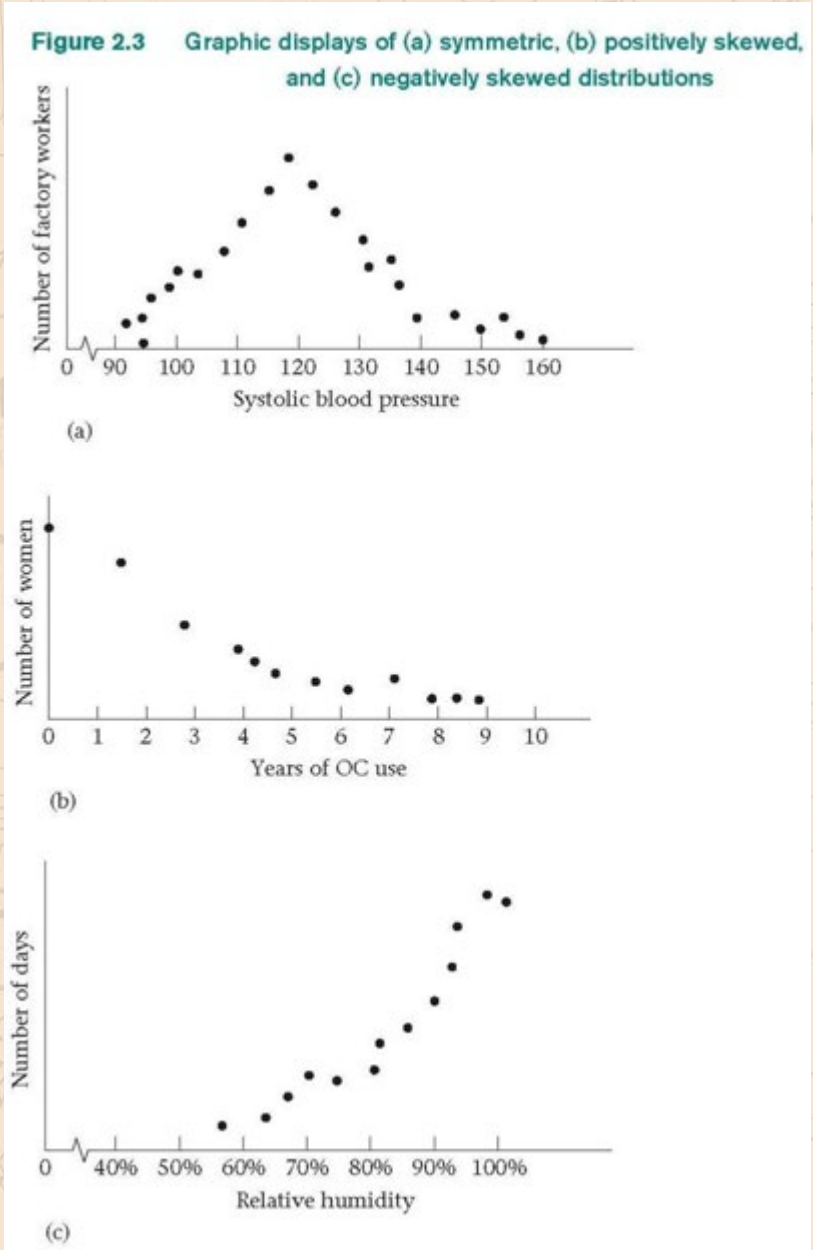
    3, 5, 7, 8, 8, 9, 10, 12, 35

    Vs. 3, 5, 7, 8, 8, 9, 10, 12, 65

    →Same median

- Arithmetic mean: increase a lot (10778 → 14111)
- Weakness:
  - determined mainly by the middle point(s)
  - less sensitive to the actual values of the other data points

Comparing Mean and Median



- **Symmetric** distributions:
arithmetic mean is approximately the same as the median
E.g. systolic blood-pressure measurements for 30-39 aged factory workers

- **Positively skewed (skewed to the right)** distributions:
arithmetic mean tends to be larger than the median
E.g. years of oral contraceptive use among women aged 20-29

- **Negatively skewed (skewed to the left** distributions:
the arithmetic mean tends to be smaller than the median
E.g. relative humidities in a humid climate at the same time of day over a no of days

Measure of Location: Mode

- Mode: the most frequent value among all the observations in a sample
- Data distributions may have one or more modes

  -One mode = unimodal

  -Two modes = bimodal

  -Three modes = trimodal

Time intervals between successive menstrual periods of 500 college women (aged 18-21)

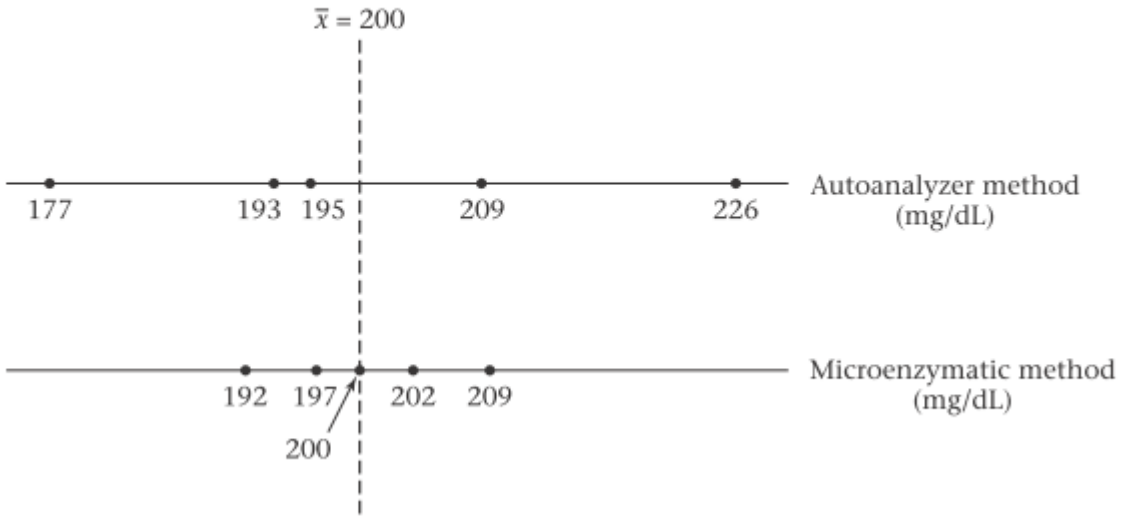
TABLE 2.4 Sample of time intervals between successive menstrual periods (days) in college-age women

Value	Frequency	Value	Frequency	Value	Frequency
24	5	29	96	34	7
25	10	30	63	35	3
26	28	31	24	36	2
27	64	32	9	37	1
28	185	33	2	38	1

Mode = 28

Measures of Spread

Figure 2.4 Two samples of cholesterol measurements on a given person using the Autoanalyzer and Microenzymatic measurement methods



- The mean obtained by the two methods is the same (same centre)
- **Variability** or **spread** of the Autoanalyzer method appears to be greater
- samples can be described by: measure of location

+

measure of spread

Measure of Spread
Range or variability

- Simplest measure about variability of a sample: **range**
- Range: difference between the largest and smallest observations in a sample
- Pros: easy to compute with ordered samples
- Cons:

  -very sensitive to **extreme observations** or **outliers**

  -depends on the sample size (n) (larger n→ larger range)

  → difficult to compare ranges of data with different n

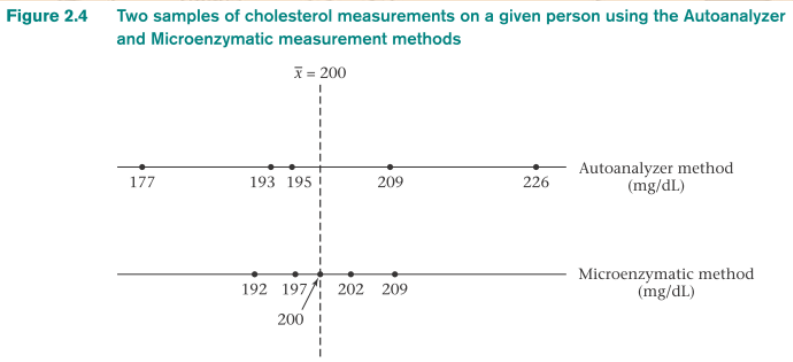
- A better measure of spread: **percentiles or quantiles**

  - less sensitive to outliers and are not greatly affected by the sample size

Measure of Spread
Quantiles or percentiles

- The *p*th percentile: value V_p such that *p* percent of the sample points are ≤ V_p.
- **Median**: 50th percentile *special case of a quantile*
- The ***p*th percentile**
  - ➢ If *np*/100 is not an integer (where *k* = largest integer less than *np*/100) : The (*k*+1)th largest sample point
  - ➢ If *np*/100 is an integer: average of the (*np*/100)th and (*np*/100 +1)th largest observations
- Percentiles computation: sample points must be ordered
- If *n* is large, a stem-and-leaf plot or a computer program may be used
- Frequently used percentiles are
  - ➢quartiles (25^{th}, 50^{th}, and 75^{th} percentiles)
  - ➢quintiles (20^{th}, 40^{th}, 60^{th}, and 80^{th} percentiles)
  - ➢deciles (10^{th}, 20^{th},…, 90^{th} percentiles)

Measure of Spread
Variance and Standard Deviation



- If center of the sample = arithmetic mean:

  - a measure summarizing the difference between the individual sample points and the arithmetic mean: $x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x}$

- **Sample variance or variance: average of the squares of the deviations from the sample mean:**

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

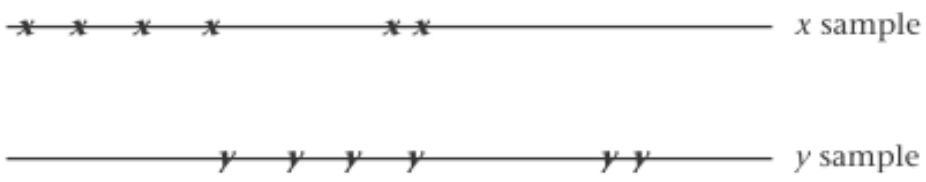
- **Sample standard deviation:** (Commonly used)

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{\text{sample variance}}$$

Measure of Spread
Properties of Variance and Standard Deviation

- Samples $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$
  where $y_i = x_i + c$ $\quad i = 1, \ldots, n$

Figure 2.5 Comparison of the variances of two samples, where one sample has an origin shifted relative to the other



- If respective sample variances: s_x^2 and s_y^2 then $s_y^2 = s_x^2$
- Samples $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$
  where $y_i = cx_i$ $\quad i = 1, \ldots, n$ *and* $c>0$
  then $s_y^2 = c^2 s_x^2$ which is $s_y = cs_x$

$$s_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} = \frac{\sum_{i=1}^{n}(cx_i - c\bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n}[c(x_i - \bar{x})]^2}{n-1} = \frac{\sum_{i=1}^{n}c^2(x_i - \bar{x})^2}{n-1}$$

$$= \frac{c^2\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = c^2 s_x^2$$

$$s_y = \sqrt{c^2 s_x^2} = cs_x$$

Coefficient of Variation (CV)

- Useful to link arithmetic mean and standard deviation
- CV = 100% × ($s/\bar{x}$)

Table 2.6 Reproducibility of cardiovascular risk factors in children, Bogalusa Heart Study, 1978–1979

	n	Mean	*sd*	*CV* (%)
Height (cm)	364	142.6	0.31	0.2
Weight (kg)	365	39.5	0.77	1.9
Triceps skin fold (mm)	362	15.2	0.51	3.4
Systolic blood pressure (mm Hg)	337	104.0	4.97	4.8
Diastolic blood pressure (mm Hg)	337	64.0	4.57	7.1
Total cholesterol (mg/dL)	395	160.4	3.44	2.1
HDL cholesterol (mg/dL)	349	56.9	5.89	10.4

- Remains the same regardless of units used
- CV is useful:
  - comparing variability of different samples with different arithmetic means
  - comparing the reproducibility of different variables

Grouped Data

- Sample size is large: data collected in grouped form
  -less accurate to measure certain quantities (measurement error and imprecise patient recall)

Table 2.7 Sample of birthweights (oz) from 100 consecutive deliveries at a Boston hospital

58	118	92	108	132	32	140	138	96	161
120	86	115	118	95	83	112	128	127	124
123	134	94	67	124	155	105	100	112	141
104	132	98	146	132	93	85	94	116	113
121	68	107	122	126	88	89	108	115	85
111	121	124	104	125	102	122	137	110	101
91	122	138	99	115	104	98	89	119	109
104	115	138	105	144	87	88	103	108	109
128	106	125	108	98	133	104	122	124	110
133	115	127	135	89	121	112	135	115	64

- The simplest way to display the data: frequency distribution (usually using a statistical package)

- **Frequency distribution:** ordered display of each value in a data set and its **frequency** (number of times that value occurs)

TABLE 2.10 Frequency distribution of the birthweight data on Table 2.9 using the FREQ procedure of SAS

Birthweight	Frequency	Percent	Cumulative Frequency	Cumulative Percent
32	1	1.00	1	1.00
58	1	1.00	2	2.00
64	1	1.00	3	3.00
67	1	1.00	4	4.00
68	1	1.00	5	5.00
83	1	1.00	6	6.00
85	2	2.00	8	8.00
86	1	1.00	9	9.00
87	1	1.00	10	10.00
88	2	2.00	12	12.00
89	3	3.00	15	15.00
91	1	1.00	16	16.00
92	1	1.00	17	17.00
93	1	1.00	18	18.00
94	2	2.00	20	20.00
95	1	1.00	21	21.00
96	1	1.00	22	22.00
98	3	3.00	25	25.00
99	1	1.00	26	26.00
100	1	1.00	27	27.00
101	1	1.00	28	28.00
102	1	1.00	29	29.00
103	1	1.00	30	30.00
104	5	5.00	35	35.00
105	2	2.00	37	37.00
106	1	1.00	38	38.00
107	1	1.00	39	39.00
108	4	4.00	43	43.00
109	2	2.00	45	45.00
110	2	2.00	47	47.00
111	1	1.00	48	48.00
112	3	3.00	51	51.00
113	1	1.00	52	52.00
115	6	6.00	58	58.00
116	1	1.00	59	59.00
118	2	2.00	61	61.00
119	1	1.00	62	62.00
120	1	1.00	63	63.00
121	3	3.00	66	66.00
122	4	4.00	70	70.00
123	1	1.00	71	71.00
124	4	4.00	75	75.00
125	2	2.00	77	77.00
126	1	1.00	78	78.00
127	2	2.00	80	80.00
128	2	2.00	82	82.00
132	3	3.00	85	85.00
133	2	2.00	87	87.00
134	1	1.00	88	88.00
135	2	2.00	90	90.00
137	1	1.00	91	91.00
138	3	3.00	94	94.00
140	1	1.00	95	95.00
141	1	1.00	96	96.00
144	1	1.00	97	97.00
146	1	1.00	98	98.00
155	1	1.00	99	99.00
161	1	1.00	100	100.00

- Unique sample values is large:
a frequency distribution may still be too detailed

TABLE 2.11 General layout of grouped data

Group interval	Frequency
$y_1 \le x < y_2$	f_1
$y_2 \le x < y_3$	f_2
.	.
.	.
.	.
$y_i \le x < y_{i+1}$	f_i
.	.
.	.
.	.
$y_k \le x < y_{k+1}$	f_k

- If the data is too large: data can be **categorized** into broader groups

TABLE 2.12 Grouped frequency distribution of the birthweight (oz) from 100 consecutive deliveries

The FREQ Procedure

Group_interval	Frequency	Percent	Cumulative Frequency	Cumulative Percent
$29.5 \le x < 69.5$	5	5.00	5	5.00
$69.5 \le x < 89.5$	10	10.00	15	15.00
$89.5 \le x < 99.5$	11	11.00	26	26.00
$99.5 \le x < 109.5$	19	19.00	45	45.00
$109.5 \le x < 119.5$	17	17.00	62	62.00
$119.5 \le x < 129.5$	20	20.00	82	82.00
$129.5 \le x < 139.5$	12	12.00	94	94.00
$139.5 \le x < 169.5$	6	6.00	100	100.00

Graphic Methods

- Graphic methods: quick overall impression of data

1) **Bar graphs**:
   ➢used to display grouped data
   ➢Cons:
   -groups are defined in arbitrary way
   -Identity of the sample points within the respective groups is lost

2) Stem-and-Leaf plots:
   ➢Leaves: general shape of the distribution of data points
   preserve actual data points + display grouped data (pros)
   ➢easy to compute the median and other quantiles
   ➢Each data point is converted into stem and leaf, e.g., 438 (stem: 43; leaf: 8)

3) **Box plots**:
   ➢median, upper quantile + lower quantile: skewness or symmetry of a distribution
   ➢Skewness: compare arithmetic mean vs. median

Stem-and-leaf plots

Figure 2.6 Stem-and-leaf plot for the birthweight data (oz) in Table 2.7

Stem-and-leaf of birthwgt N = 100
Leaf Unit = 1.0

1	3	2
1	4	
2	5	8
5	6	478
5	7	
15	8	3556788999
26	9	12344568889
45	10	0123444445567888899
(17)	11	00122235555556889
38	12	01112222344445567788
18	13	222334557888
6	14	0146
2	15	5
1	16	1

Stem |Leaf

- 5|8: 58
- 11|8: 118

- Overall data distribution (not sacrificing the individual values)

- First column: lowest / highest value in cumulative frequency

- (17): leaves including the median

Stem-and-leaf vs. Box Plot

Figure 2.8 Stem-and-leaf and box plots for the birthweight data (oz) in Table 2.7 as generated by the SAS UNIVARIATE procedure

Stem	Leaf	#	Boxplot
16	1	1	
15	5	1	
15			
14	6	1	
14	014	3	
13	557888	6	
13	222334	6	
12	5567788	7	
12	0111222234444	13	+------+ ← Upper quartile
11	5555556889	10	| |
11	0012223	7	*--+--* ← median
10	5567888899	10	| |
10	012344444	9	| |
9	568889	6	+------+ ← lower quartile
9	12344	5	
8	556788999	9	
8	3	1	
7			
7			
6	78	2	
6	4	1	
5	8	1	0
5			
4			
4			
3			
3	2	1	0

----+----+----+----+

Multiply Stem.Leaf by 10**+1

negatively skewed

Box plot

- Visually describe the distribution of sample points + point out possible outliers

➢ **Symmetric distribution**: upper and lower quartiles approximately equally spaced from median

➢**Positively skewed distribution:** upper quartile is farther from median than lower quartile

➢**Negatively skewed distribution:** lower quartile is farther from median than upper quartile

Box plot

- **outlying value:**

  $x >$ upper quartile + 1.5 × (upper quartile – lower quartile)
  or $x <$ lower quartile – 1.5 × (upper quartile – lower quartile)

- **extreme outlying value:**

  $x >$ upper quartile + 3.0 × (upper quartile – lower quartile)
  or $x <$ lower quartile – 3.0 × (upper quartile – lower quartile)

➢A vertical bar:

- connects upper quartile to largest nonoutlying value among the data points
- connects lower quartile to smallest nonoutlying value among the data points

Summary

- **Numeric** or **graphic** methods for displaying data help in
  - ➢quickly summarizing a data set
  - ➢presenting results (publications)

- A data set can be described **numerically** in terms of **measure of location** and a **measure of spread**

Measure of location	**Measure of spread**
Arithmetic mean	Standard deviation
Median	Quantiles
Mode	Range

- **Graphic methods** include bar graphs and more exploratory methods such as **bar graphs**, **stem-and-leaf plots** and **box plots**.