AMS 572 Data Analysis I
Summarizing and Exploring Data

Pei-Fen Kuan

Applied Math and Stats, Stony Brook University

Content

- Types of variables
- Measures of location
- Measures of spread, shape
- Data displays

Types of variables

- A *variable* is a quantity that may vary from object to object
- A *sample* or *data set* is a collection of values of one or more variables.
- Quantitative variable intrinsically numerical
  e.g. age, height, counts
- Qualitative (categorical) - intrinsically nonnumerical
  e.g. gender, province, country

Types of variables

- Qualitative (categorical) - intrinsically nonnumerical
  - Binary, dichotomous
    e.g., alive/dead, female/male
  - Ordinal - natural ordering
    e.g., diagnosis (certain, probable, unlikely, ...)
    e.g., attitude (strongly agree, agree, neutral, ...)
  - Nominal - no natural ordering
    e.g., religion, race
- In recording qualitative data, numerical values may be assigned

Measures of Location

- (Arithmetic) Mean
- Percentiles
- Median
- Mode
- Geometric mean

Arithmetic mean

- Data:

$$x_1, x_2, \ldots, x_n$$

- Mean:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Example: Duration of hospital stay in days:

$$x_1 = 5, x_2 = 10, x_3 = 6, x_4 = 11$$

Mean:

$$\bar{x} = 8$$

Properties of Mean

- Let c be any constant
- If
  $$y_i = x_i + c \text{ for } i = 1, 2, 3, \ldots, n,$$
  then
  $$\bar{y} = \bar{x} + c$$
- If
  $$y_i = cx_i \text{ for } i = 1, 2, 3, \ldots, n,$$
  then
  $$\bar{y} = c\bar{x}$$

Properties of Mean - Example

- A sample of birth weights in a hospital found

$$\bar{y} = 3166.9 \text{ grams}$$

- 1 oz = 28.35 g
- Therefore the mean in ozs. is

$$\bar{x} = \frac{\bar{y}}{28.35} = 111.7$$

Order Statistics

- Data: $x_1, x_2, \ldots, x_n$
- Order data from smallest to largest

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$

- $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ are *order statisitics*
- Note

$$x_{(1)} = \min\{x_1, x_2, \ldots, x_n\}$$

$$x_{(n)} = \max\{x_1, x_2, \ldots, x_n\}$$

- Example: Duration of hospital stay in days:

$$x_1 = 5, x_2 = 10, x_3 = 6, x_4 = 11$$

Order statistics:

$$x_{(1)} = 5, x_{(2)} = 6, x_{(3)} = 10, x_{(4)} = 11$$

Theorem: Let $X_{(1)} \leq \ldots \leq X_{(n)}$ denote the order statistics of a random sample, $X_1, \ldots, X_n$ from a random population with c.d.f. $F(\cdot)$ and p.d.f. $f(\cdot)$. The p.d.f. of $X_{(k)}$ is

$$f_{X_{(k)}}(x) = \frac{n!}{(k-1)!(n-k)!}[F(x)]^{k-1}[1-F(x)]^{n-k}f(x)$$

Percentiles

- Intuitive definition: the x *percentile* is such that $x\%$ of the observations are less than that value
- Also known as sample *quantile*
- The $(p \times 100)^{th}$ percentile of a sample

$$\hat{\zeta}_p = \begin{cases} y_{(np+p)} & \text{if } np+p \text{ is an integer} \\ \{y_{(\lfloor np+p \rfloor)} + y_{(\lceil np+p \rceil)}\}/2 & \text{otherwise} \end{cases}$$

for $0 < p < 1$

Percentiles: General form

- General form (Hyndman and Fan, *Am Stat* 1996)

$$\hat{\zeta}_p = (1-\gamma)y_{(j)} + \gamma y_{(j+1)}$$

  where $j = \lfloor pn + k \rfloor$ for some $k \in \mathbb{R}$ and $0 \le \gamma \le 1$.

- Your textbook

$$\hat{\zeta}_p = \begin{cases} y_{(np+p)} & \text{if } np+p \text{ is an intege} \\ y_{(m)} + [np+p-m](y_{(m+1)} - y_{(m)}) & \text{otherwise} \end{cases}$$

  where $m = \lfloor np + p \rfloor$

Example

- In R, there are nine different quantile definitions (argument `type`)

```
> x <- 1:278
> quantile(x,.75,type=1)
75%
209
```

Median

- The sample median is the 50th percentile

$$\hat{\zeta}_{.5} = \begin{cases} y_{(\frac{n+1}{2})} & \text{if } n \text{ is odd} \\ \{y_{(n/2)} + y_{(n/2+1)}\}/2 & \text{if } n \text{ is even} \end{cases}$$

  for $0 < p < 1$

- Example: Duration of hospital stay in days:

$$x_1 = 5, x_2 = 10, x_3 = 6, x_4 = 11$$

  Median:

$$\hat{\zeta}_{.5} = 8$$

Mode

- The mode is the most frequently occurring value in the data set
- E.g., if
$$x_1 = 5, x_2 = 11, x_3 = 6, x_4 = 11$$
then mode is 11

Geometric Mean

- Data: $x_1, x_2, \ldots, x_n$
- The geometric mean of x is

$$\bar{x}_g = (x_1 x_2 \cdots x_n)^{1/n}$$

- Eg, suppose $x_1 = 10$ and $x_2 = 0.1$. Then $\bar{x}_g = 1$

Comments

- Mean is most often used measure
- Median is better if there are influential observations (more robust to extreme values)
- Mode rarely used (exception: nominal data)

Measures of Spread, Shape

- Range
- Variance and standard deviation
- Interquartile range
- Skewness, Kurtosis

Range

- Range:
$$r_a = x_{(n)} - x_{(1)}$$
- Easy to calculate
- Sensitive to unusual observations (outliers)
- Usually, the larger n is, the larger r_a
- A rough estimate of $\sigma = r_a/4$

Sample Variance and Standard Deviation

- Want to measure deviation from mean
- Sample variance

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

- Sample standard deviation

$$s = \sqrt{s^2}$$

Sample Variance and Standard Deviation

- An alternative form of the sample variance is

$$\hat{\sigma}_1^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- We have shown that s^2 is unbiased for population variance σ^2, however

$$E(\hat{\sigma}_1^2) = \sigma^2 - \frac{\sigma^2}{n}$$

Sample Standard Deviation

- The units of s are the same as the units of x_i
- If s is large, the data are spread over a wide range
- If c is a constant and

$$y_i = x_i + c,$$

  then

$$s_y = s_x$$

- If

$$y_i = cx_i$$

  then

$$s_y = cs_x$$

Some approximations

- The interval $\bar{x} \pm s$ will contain approx 68% of the observations
- The interval $\bar{x} \pm 2s$ will contain approx 95% of the observations
- Approx s by
$$s \approx \frac{\hat{\zeta}_{.75} - \hat{\zeta}_{.25}}{1.35}$$
- Note
$$\hat{\zeta}_{.75} - \hat{\zeta}_{.25}$$
is called *interquartile range*

Symmetry and Skewness

- Informally, define *symmetry* to indicate having a uniform or even distribution about the mean
- If a distribution is symmetric,

$$\text{mean}=\text{median}$$

- Data sets that are not symmetric are said to be *skewed*
- *Skewness* is a measurement of the degree to which a data set is skewed

Skewness

- Define rth sample moment about the mean

$$m_r = \frac{\sum_i (y_i - \bar{y})^r}{n} \text{ for } r = 1, 2, 3, \ldots$$

- Definition of sample skewness:

$$a_3 = \frac{\sum_i (y_i - \bar{y})^3 / n}{\{\sum_i (y_i - \bar{y})^2 / (n-1)\}^{3/2}}$$

- $a_3 > 0$ indicates skewness to the right

Kurtosis

- *Kurtosis* is a measure of the flatness or peakedness of a distribution; degree of archedness; thickness of tails
- Definition of *sample* kurtosis:

$$a_4 = \frac{\sum_i (y_i - \bar{y})^4 / n}{\{\sum_i (y_i - \bar{y})^2 / (n-1)\}^2}$$

- $a_4 > 3$ indicates the distribution has heavier tails than the normal distribution.
- In R, skewness and kurtosis can be computed using functions `skewness()` and `kurtosis()` from `library(e1071)`.

Data display

- Simplest form is a line listing
- A *frequency table* gives the frequency of observations within a set of ordered intervals
- Intervals should be mutually exclusive and exhaustive
- 8 to 10 intervals is usually sufficient
- With the exception of the end intervals, the length of the intervals should be constant

Frequency Table - Example

Blood Pressure	Pop1	Pop2	Pop3
< 106	218	4	23
106-114	272	23	132
116-124	337	49	290
126-134	362	33	347
136-144	302	41	346
146-154	261	38	202
156-164	166	23	109
> 164	314	52	112
Total	2232	263	1561

Frequency Tables

- Table on previous slide example of *empirical frequency distribution*
- Difficult to compare blood pressure distributions due to different sample sizes
- Divide by sample size to get *empirical relative frequency distribution*

ERFD - Example

Blood Pressure	Pop1	Pop2	Pop3
< 106	0.098	0.015	0.015
106-114	0.122	0.087	0.085
116-124	0.151	0.186	0.186
126-134	0.162	0.125	0.222
136-144	0.135	0.156	0.222
146-154	0.117	0.144	0.129
156-164	0.074	0.087	0.070
> 164	0.141	0.198	0.072
Total	2232	263	1561

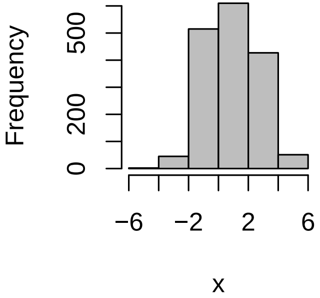
Graphs

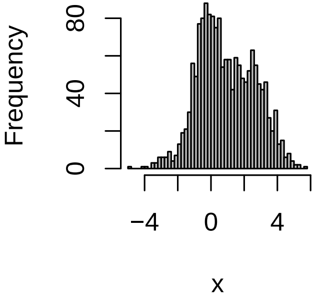
- Histogram
- Box plot
- Trellis/conditional plots

Histogram

- Data are divided into intervals as in a frequency table
- A histogram is a bar graph with the area of each bar equal to the relative frequency in the interval.
- Can compare histograms from samples of different size
- Intervals need not be the same width
- Beware effect of choice of interval width

```
> x <- c(rnorm(50,-2.5),rnorm(1000),rnorm(600,2.7))
> hist(x,breaks=5,col="gray",xlab="x",main="")
> hist(x,breaks=50,col="gray",xlab="x",main="")
```



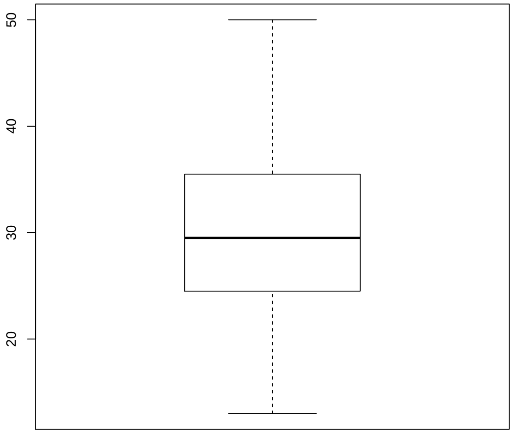


Box plot

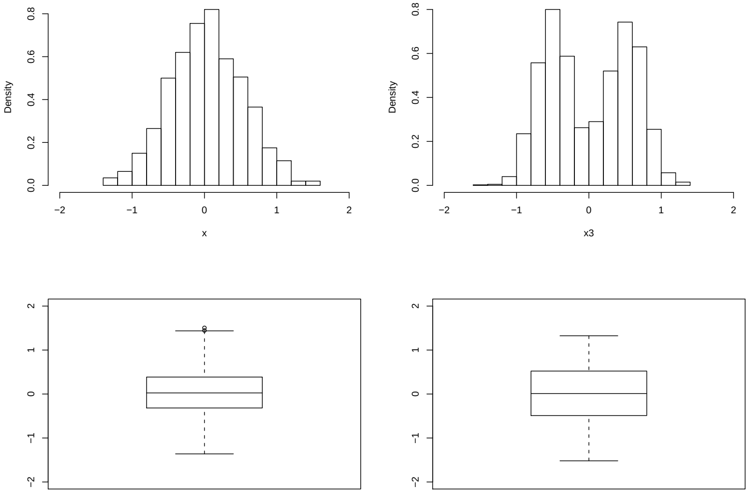
- The top of the box is the 75th percentile ($\hat{\zeta}_{.75}$); the bottom is the 25th percentile ($\hat{\zeta}_{.25}$)
- A line through the box is drawn at the median
- The lines extending out of the box (*whiskers*) may extend to
  - the 90th and 10th percentiles
  - the largest and smallest values
  - largest observation $\leq \hat{\zeta}_{.75} + 1.5$ x IQR;
    smallest observation $\geq \hat{\zeta}_{.25} - 1.5$ x IQR
- Data beyond whiskers may be plotted individually

Boxplot Example

```
> boxplot(mileage)
```

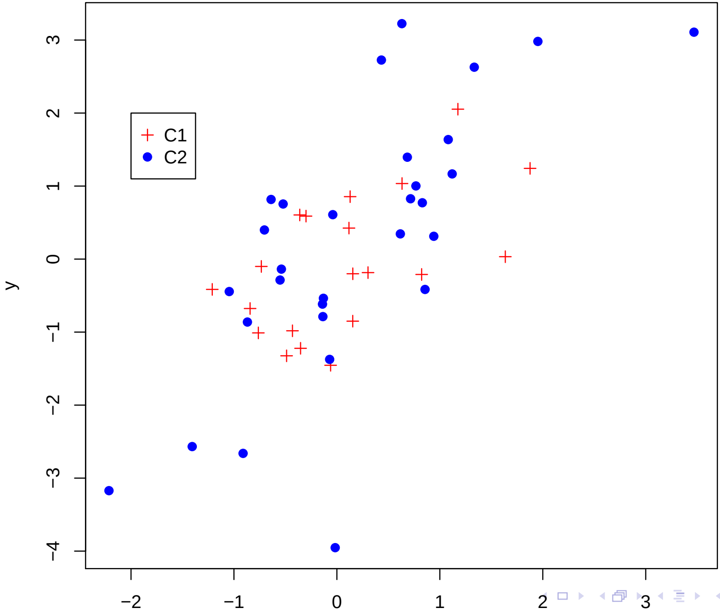


Box plot and Histogram Example



Multivariate plots

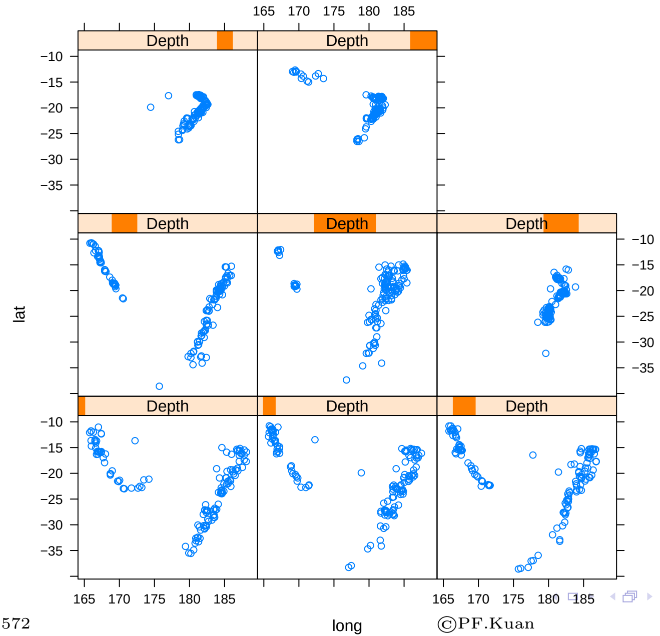
- Describe relationships/associations between more than one variable
- Scatterplots
  - Simple for two variables
  - Add color, symbols for > 2 variables



C1
C2
y
3
2
1
0
−1
−2
−3
−4
−2
−1
0
1
2
3

```
> x <- rnorm(50)
> y <- x+rnorm(50)
> id <- sample(1:50,size=20)
> plot(x,y,type="n")
> points(x[id],y[id],col="red",pch=3)
> points(x[-id],y[-id],col="blue",pch=19)
> legend(-2,2,c("C1","C2"),col=c("red","blue"),pch=c(3,19))
```

Trellis plots



Trellis plots

```
> library(lattice)
> require(stats)

## Tonga Trench Earthquakes

> Depth <- equal.count(quakes$depth, number=8, overlap=.1)
> xyplot(lat ~ long | Depth, data = quakes)
> update(trellis.last.object(),
      strip = strip.custom(strip.names = TRUE, strip.levels = TRUE),
      par.strip.text = list(cex = 0.75),
      aspect = "iso")
```

Tables or graphs?

- Tables best suited for looking up specific information
- Graphs better for perceiving trends, making comparisons and predictions

Read Chapter 7