STA 291
Lecture 10, Chap. 6

- **Describing Quantitative Data**
  - Measures of Central Location
  - Measures of Variability (spread)

- First Midterm Exam a week from today,
- Feb. 23 5-7pm
- Cover up to mean and median of a sample (begin of chapter 6). But not any measure of spread (i.e. standard deviation, inter-quartile range etc)

Summarizing Data Numerically

- Center of the data
  - Mean (average)
  - Median
  - Mode (…will not cover)
- Spread of the data
  - Variance, Standard deviation
  - Inter-quartile range
  - Range

Mathematical Notation: Sample Mean

- Sample size *n*
- Observations $x_1, x_2, \ldots, x_n$
- Sample Mean "x-bar" --- a statistic

$$\bar{x} = (x_1 + x_2 + \ldots + x_n)/n$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\sum = \text{SUM}$$

Mathematical Notation: Population Mean for a finite population of size *N*

- Population size (finite) N
- Observations $x_1 , x_2 ,\ldots, x_N$
- Population Mean "mu" --- a Parameter

$$m = (x_1 + x_2 + \ldots + x_N) / N$$

$$= \frac{1}{N} \sum_{i=1}^{N} x_i$$

$$\sum = \mathrm{SUM}$$

Infinite populations

- Imagine the population mean for an infinite population.
- Also denoted by mu or m
- Cannot compute it (since infinite population size) but such a number exist in the limit.
- Carry the same information.

Infinite population

- When the population consists of values that can be ordered
- Median for a population also make sense: it is the number in the middle….half of the population values will be below, half will be above.

Mean

- If the distribution is highly skewed, then the mean is not representative of a typical observation
- Example:

  Monthly income for five persons

  1,000   2,000   3,000   4,000   100,000
- Average monthly income: = 22,000
- Not representative of a typical observation.

- Median = 3000

Median

- The median is the measurement that falls in the middle of the *ordered* sample
- When the sample size *n* is odd, there is a middle value
- It has the ordered index $(n+1)/2$
- Example: 1.1, 2.3, 4.6, 7.9, 8.1

  $n=5$, $(n+1)/2=6/2=3$, *so index* $= 3$,

  Median = 3rd smallest observation = 4.6

Median

- When the sample size *n* is even, average the two middle values
- Example: 3, 7, 8, 9, *n=4,*

(n+1)/2=5/2=2.5, index = 2.5

Median = midpoint between

2nd and 3rd smallest observation

= (7+8)/2 =7.5

Summary: Measures of Location

Mean- Arithmetic Average

- Mean of a Sample - $\bar{x}$
- Mean of a Population - $\boldsymbol{\mu}$

Median – Midpoint of the observations when they are arranged in increasing order

Notation: Subscripted variables
- n = # of units in the sample
- N = # of units in the population
- x = Variable to be measured
- x_i = Measurement of the *ith* unit

Mode….

Mean vs. Median

Observations	Median	Mean
1, 2, 3, 4, 5	3	3
1, 2, 3, 4, 100	3	22
3, 3, 3, 3, 3	3	3
1, 2, 3, 100, 100	3	41.2

Mean vs. Median

- If the distribution is symmetric, then Mean=Median
- If the distribution is skewed, then the mean lies more toward the direction of skew
- Mean and Median Online Applet

Why not always Median?

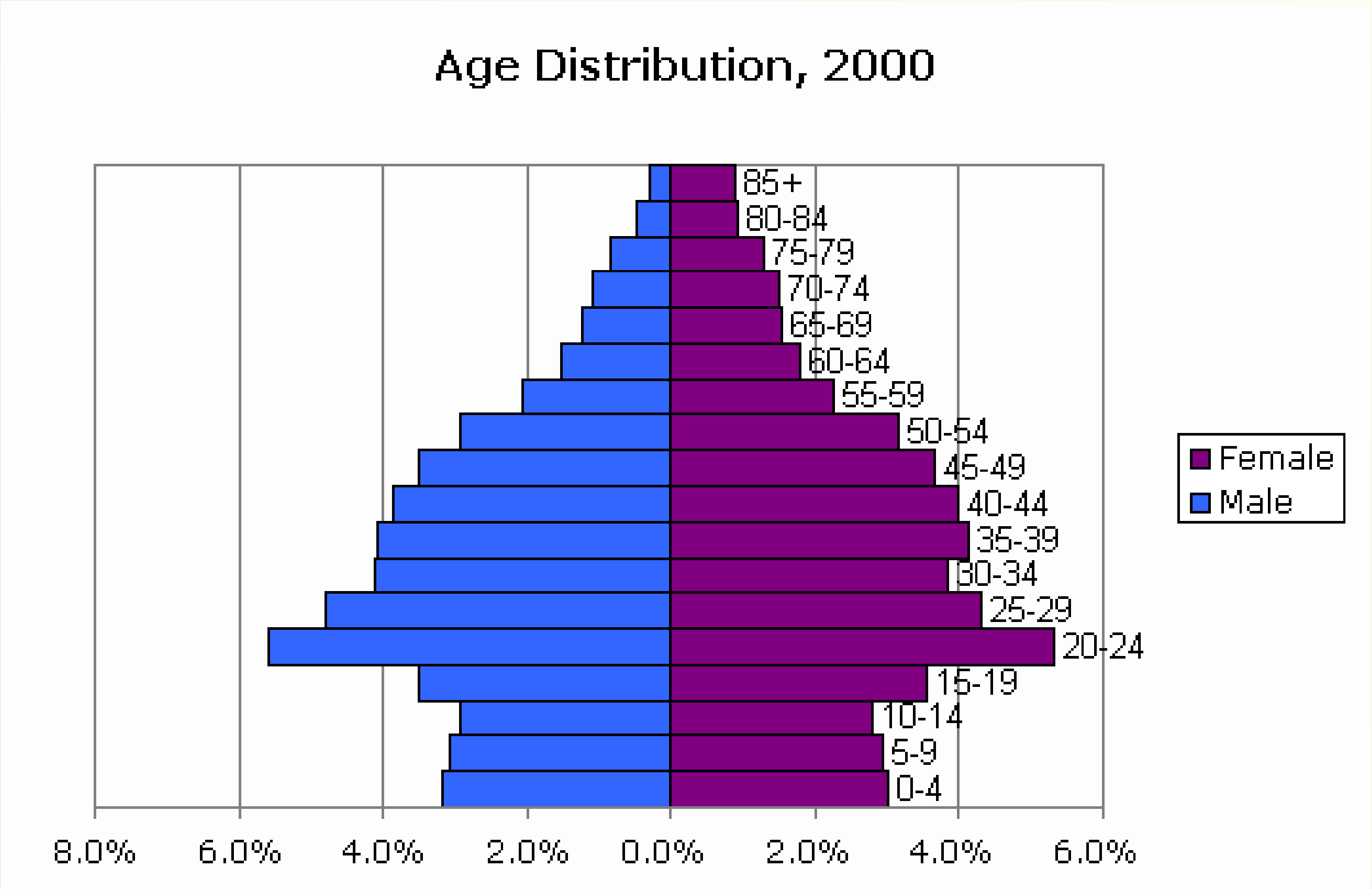
- Disadvantage: Insensitive to changes **within** the lower or upper half of the data
- Example: 1, 2, 3, 4, 5, 6, 7 vs.

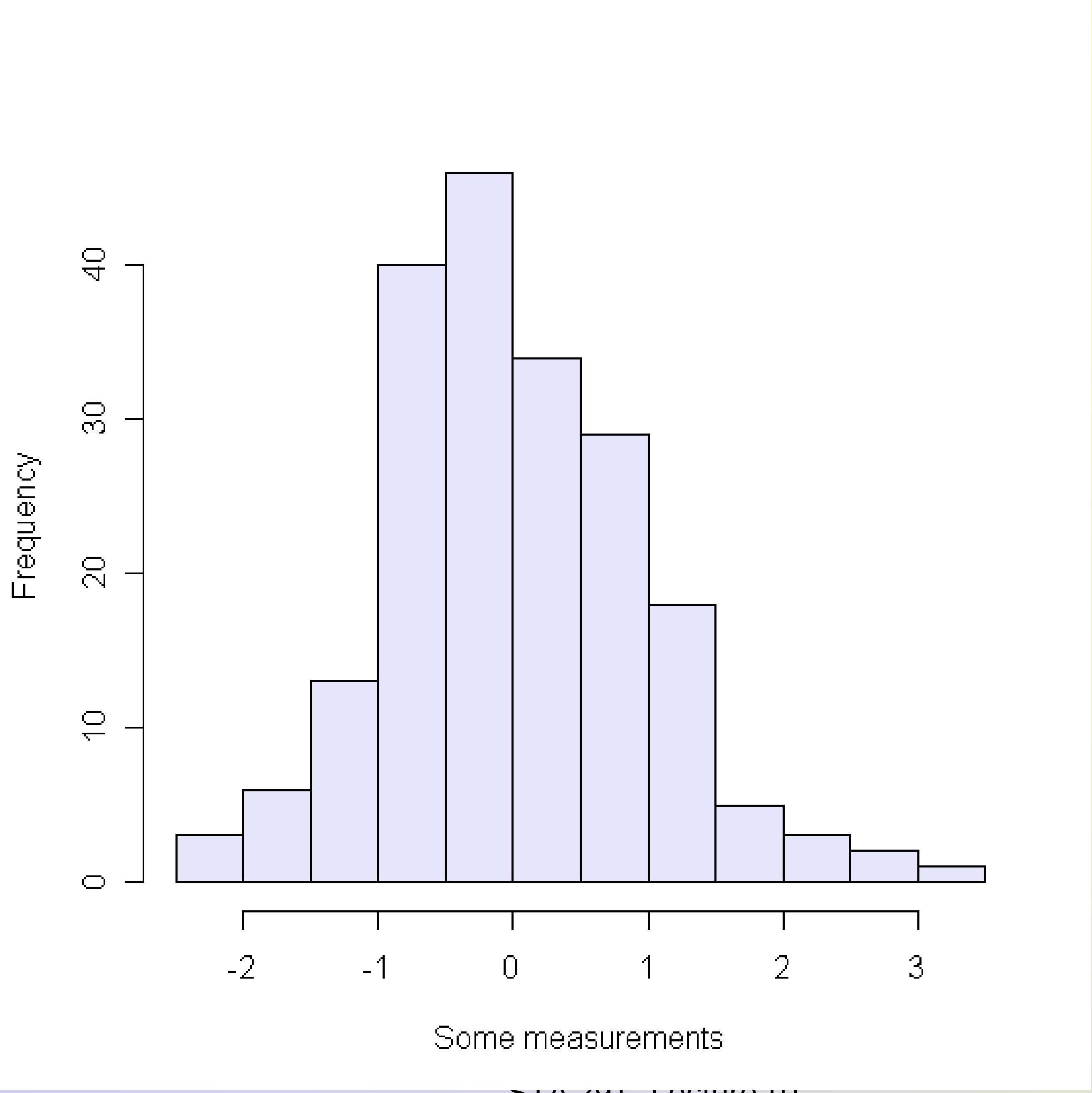
  1, 2, 3, 4, 100,100,100
- For symmetric, bell shaped distributions, mean is more informative.
- Mean is easy to work with. Ordering can take a long time
- *Sometimes*, the mean is more informative even when the distribution is slightly skewed

Census Data	Lexington	Fayette County	Kentucky	United States
Population	261,545	261,545	4,069,734	281,422,131
Area in square miles	306	306	40,131	3,554,141
People per sq. mi.	853	853	101	79
Median Age	35	34	36	36
Median Family Income	$42,500	$39,500	$32,101	$40,591

Real Estate Market Data	Lexington	Fayette County	Kentucky	United States
Total Housing Units	54,587	54,587	806,524	115,904,743
Average Home Price	$151,776	$151,776	$115,545	$173,585
Median Rental Price	$383	$383	$257	$471
Owner Occupied	52%	52%	64%	60%

Given a histogram, find approx mean and median




Frequency
0
10
20
30
40
-2
-1
0
1
2
3
Some measurements

Percentiles

- The *p*th percentile is a number such that *p*% of the observations take values below it, and (100-*p*)% take values above it
- 50^{th} percentile = median
- 25^{th} percentile = lower quartile
- 75^{th} percentile = upper quartile

Quartiles

- 25^{th} percentile = lower quartile
  = Q1

- 75^{th} percentile = upper quartile
  = Q3

Interquartile range = Q3 - Q1

(a measurement of variability in the data)

SAT Math scores

- Nationally (min = 210 max = 800)

  Q1 = 440

  Median = Q2 = 520

  Q3 = 610 (-- you are better than 75% of all test takers)

- Mean = 518 (SD = 115 what is that?)

SAT Percentile Ranks

Critical Reading, Mathematics, and Writing

Score	Critical Reading	Mathematics	Writing
800	99	99	99+
790	99	99	99+
780	99	99	99
770	99	99	99
760	99	98	99
750	98	98	99
740	98	97	98
730	97	97	98
720	96	96	97
710	96	95	97
700	95	93	96
690	94	92	95
680	93	91	94
670	92	89	93
660	90	88	92
650	89	86	90
640	87	83	89
630	85	81	87
620	83	79	85
610	82	76	83
600	79	74	81
590	77	71	79
580	74	68	76
570	71	66	73
560	68	63	71
550	65	60	68
540	62	56	64
530	58	53	62
520	55	50	58
510	51	47	54
500	48	43	51
490	44	40	47
480	41	36	44
470	37	33	40
460	34	30	37
450	31	27	33

Five-Number Summary

- Maximum, Upper Quartile, Median, Lower Quartile, Minimum
- Statistical Software SAS output (Murder Rate Data)

Quantile	Estimate
100% Max	20.30
75% Q3	10.30
50% Median	6.70
25% Q1	3.90
0% Min	1.60

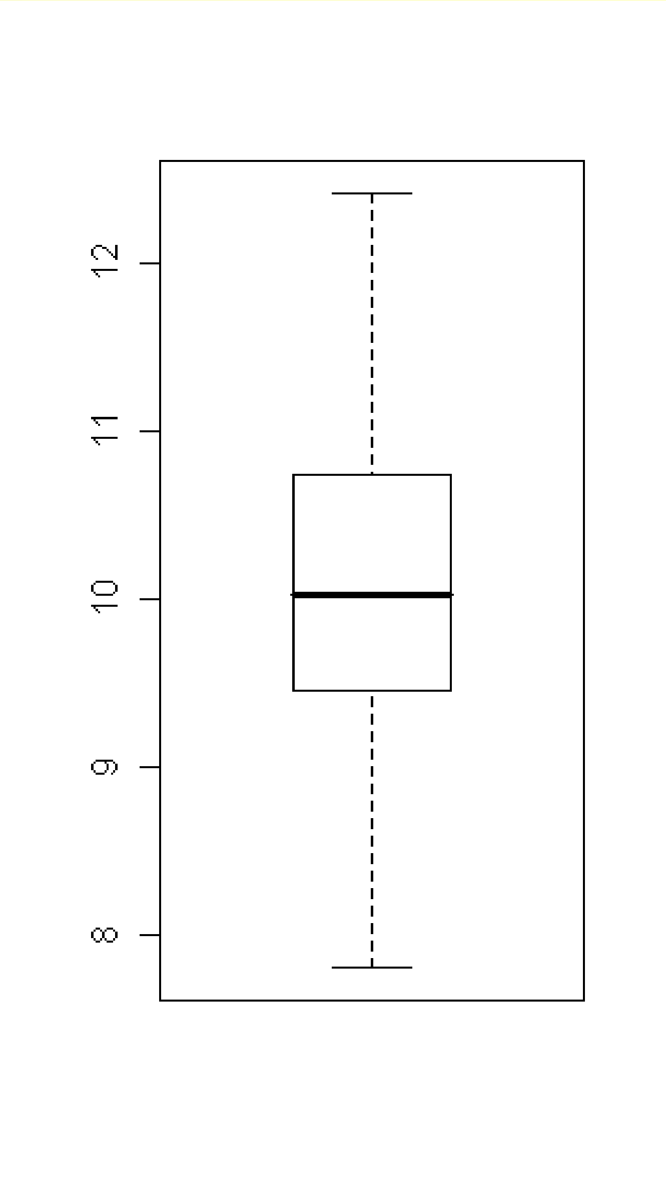
Five-Number Summary

- Maximum, Upper Quartile, Median,
  Lower Quartile, Minimum

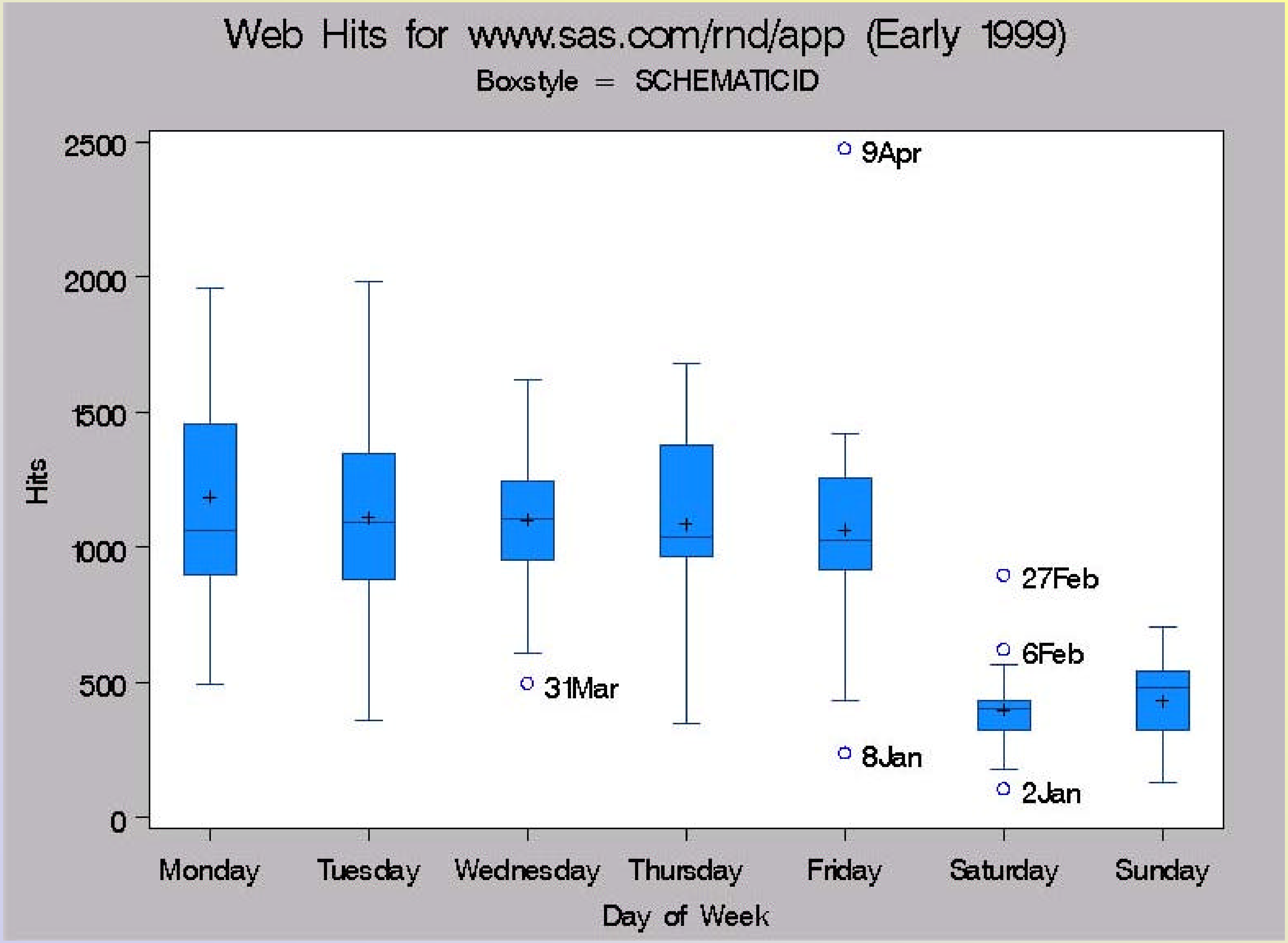
- Example: The five-number summary for a data set is min=4, Q1=256, median=530, Q3=1105, max=320,000.
- What does this suggest about the shape of the distribution?

Box plot

- A box plot is a graphic representation of the five number summary --- provided the max is within 1.5 IQR of Q3 (min is within 1.5 IQR of Q1)



- Otherwise the max (min) is suspected as an **outlier** and treated differently.


Web Hits for www.sas.com/rnd/app (Early 1999)
Boxstyle = SCHEMATICID
2500
2000
1500
1000
500
0
Hits
9Apr
31Mar
8Jan
27Feb
6Feb
2Jan
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Day of Week

- Box plot is most useful when compare several populations

Measures of Variation

- Mean and Median only describe the central location, but not the spread of the data
- Two distributions may have the same mean, but different variability
- Statistics that describe variability are called measures of spread/variation

Measures of Variation

- Range: = max - min

Difference between maximum and minimum value

- Variance: $s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$
-
- Standard Deviation: $s = \sqrt{s^2} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$

- Inter-quartile Range: = Q3 – Q1

Difference between upper and lower quartile of the data

Deviations: Example

- Data: 1, 7, 4, 3, 10
- Mean: (1+7+4+3+10)/5 =25/5=5

data	Deviation	Dev. square
1	(1 - 5)= -4	16
3	(3 - 5)= -2	4
4	(4 - 5) = -1	1
7	(7 - 5) = 2	4
10	(10 - 5) = 5	25
Sum=25	Sum = 0	sum = 50

Sample Variance

$$s^2 = \frac{\sum (x_i - \overline{x})^2}{n-1}$$

The variance of *n* observations is the sum of the squared deviations, divided by *n-1.*

Variance: Example

Observation	Mean	Deviation	Squared Deviation
1	5		16
3	5		4
4	5		1
7	5		4
10	5		25
Sum of the Squared Deviations			50
n-1			5-1=4
Sum of the Squared Deviations / *(n-1)*			50/4=12.5

- So, sample variance of the data is 12.5

- Sample standard deviation is 3.53

$$\sqrt{12.5} = 3.53$$

Attendance Survey Question

- On a 4”x6” index card
  - write down your name and section number
  - Question:
  - Lexington Average temperature in Feb.

Is about __________?

Example: Mean and Median

- Example: Weights of forty-year old men

  158, 154, 148, 160, 161, 182,

  166, 170, 236, 195, 162
- Mean =
- Ordered weights: (order a large dataset can take a long time)
- 148, 154, 158, 160, 161, 162,

  166, 170, 182, 195, 236
- Median =