Stat 213: Intro to Statistics 1

H. Kim

Fall 2007

introduction

- Statistics deals with the collection, description, and interpretation of data
- data may arise naturally as observations of "things as they are" (observational study), or as a result of a controlled or designed experiment (experimental study)

observational study

- a study based on data in which no manipulation of factors has been employed: the studies focus on discovering trends and possible relationships between groups or associations between variables
- even though study may find an important variables related to outcome, there is no guarantee that we found the right or the most important related variables
- it is not possible to demonstrate a causal relationship

experimental study

- to establish cause and effect
- manipulates factor (explanatory veraiable) level to create treatments, randomly assigns subjects to these treatment levels, and then compares the responses of the subject groups across treatment levels
- experimental units are not always drawn randomly from a population: the randomization is in assigning subjects to treatments; the focus is on difference in the effects of the treatments, not on estimating the population parameters

example

- to study a presumed inverse relationship between black tea consumption and cardiovascular disease, investigators randomly selected a sample of 13500 subjects, 50-64 years of age, and queried them about tea consumption and current treatment for cardiovascular disease; no significant association was found between drinking black tea and having cardiovascular disease

example

- to see if hypnosis could reduce chronic pain, investigators randomly assigned 120 volunteers with osteoarthritis to either hypnosis therapy or a control group; pain levels were checked among the subjects every three months during the 12-month study using validated pain measures; those assigned to the hypnosis group reported reduced pain compared to those in the control group after three and six months of follow-up

population and sample

- a population is the set of all measurements or objects of interest in a particular study
- if the entire population were available for analysis we would know everything about it
- however, in practice one cannot know the entire population because it is either too expensive, or simply impossible or impractical to examine each member
- thus, a sample from the population is used to obtain information about the population; a sample is subset of the population

sample surveys

- randomness is an essential tool of Statistics, but it is not easy being random
- selecting a sample to represent the population fairly is difficult: bias
- randomizing protects us from bias and makes it possible for us to draw inferences about the population when we see only a sample

sample surveys

- how large a random sample do we need for the sample to be reasonably representative of the population ? do we need a large fraction of the population ?
  - it's the **sample size**: the number of subject in the sample (a size of spoonful is enough to make a decision about the entire pot, no matter how large the pot)
  - how big a sample do you need? that depends on what you are estimating

parameter and statistic

- building a model for questions: models use mathematics to represent reality
- a parameter used in a model for a population: population parameter
  - mean (μ), variance (σ^2), proportion (p)
- we use the data to try to estimate the population parameters: any summary from the data is a statistic
- sample statistic : when we match the statistics with the parameters they estimate
  - mean ($\bar{x}$), variance $\left(\hat{\sigma}^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}\right)$, proportion ($\hat{p}$)
- we need to be sure that the statistics we compute from the sample reflect the corresponding parameter accurately

sampling methods

- sampling frame: a list of subjects from which the sample is drawn
  - in defining the sampling frame, we must deal with the details of defining the population: subjects who may be in the population of interest, but who are not in the sampling frame cannot be included in any sample
- sampling methods: we need to be sure that the statistics we compute from the sample reflect the corresponding parameter accurately
  - simple random sample (SRS), stratified random sampling, cluster and multistage sampling, systematic samples

simple random sample (SRS)

- every possible sample of the size we plan to draw has an equal chance to be selected
- SRS is the standard against which we measure other sampling methods, and the sampling method on which the theory of working with sampled data is based

systematic sampling

- selecting individuals systematically from a randomly selected individual
- justify the systematic method is not associated with any of the measured variables
- less expensive

stratified sampling

- the population is sliced into homogenous groups, called strata before the sample is selected
- simple random sampling is used within each stratum before the results are combined
- reduce the sampling variability (eg. stratifying by race, income, age, sex, and other characteristics, depending on the questions in the survey)

cluster sampling

- splitting population into similar parts called cluster
- select one or a few cluster at random and perform a census within each of them
- make sampling more practical
- each cluster fairly represents the full population (each stratum doesn't represent the population; it different from the others)

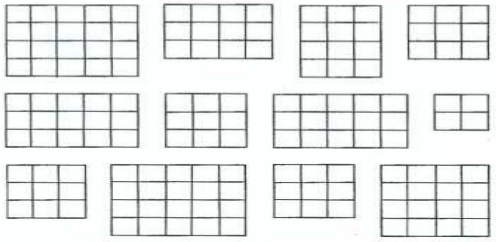
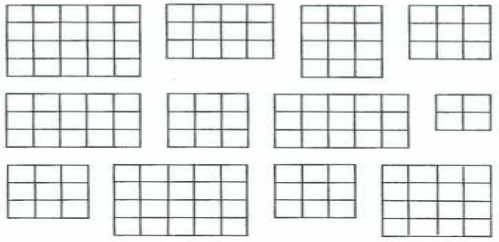
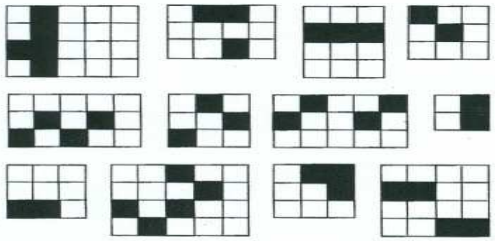
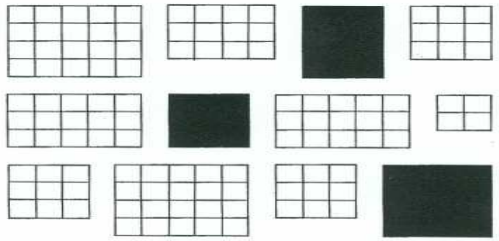
multistage sampling

- sampling schemes that combine several methods
  - eg: selecting sentences from a book:

    stratify by the part of book, and randomly choose a chapter to represent each stratum, within each selected chapter, choose pages as clusters, finally, SRS of sentence within each cluster

stratified and cluster samplings

FIGURE 5.1
Similarities and differences between cluster sampling and stratified sampling

Stratified Sampling	Cluster Sampling
Each element of the population is in exactly one stratum.	Each element of the population is in exactly one cluster.
Population of H strata; stratum h has N_h elements:	One-stage cluster sampling; population of N clusters:
	
Take an SRS from *every* stratum:	Take an SRS of clusters; observe all elements within the clusters in the sample:
	
Variance of the estimate of $\bar{y}_U$ depends on the variability of values *within* strata.	The cluster is the sampling unit; the more clusters we sample, the smaller the variance. The variance of the estimate of $\bar{y}_U$ depends primarily on the variability *between* cluster means.
For greatest precision, individual elements within each stratum should have similar values, but stratum means should differ from each other as much as possible.	For greatest precision, individual elements within each cluster should be heterogeneous, and cluster means should be similar to one another.

Sampling: Design and Analysis, Sharon L. Lohr, 1999

example

- Wilk et al. (1977) report data on the number and types of fish and environmental data for the area of the Atlantic continental shelf between eastern Long Island, New York and Cape May, New Jersey; sampling was done at a higher rate close to shore than farther away from shore: in-shore and off-shore, and calculating average numbers of fish caught and numbers of species

example

- a student wants to estimate the average grade point average (GPA) in his dormitory; he notices that the dorm consists of 100 suits, each with four students; he chooses 5 of those suits at random and asks every person in the 5 suits what her or his GPA is.

know where you are headed and why..

who were measured: subjects, experimental units, observations

what was measured: variables

why the study was performed

when the study was performed

where the data were collected

how the data were collected: design of methods

variable and data

- a variable is a characteristic that changes or varies for different individuals or objects under consideration
- an experimental unit is the individuals or object on which a variable is measured
- when a variable is measured on a set of experimental units, a set of data result

types of variables

- qualitative or categorical variables measure a quality or characteristic on each experimental unit; tell us what group and categories each subject belong to (counts how many subject belong in each categories)
- quantitative or numerical variables measure a numerical quantity on each subject and have measurement units: the units tells us how much of something we have or how far apart two variables are (average, range)
  - there are two types of quantitative variables: quantitative and continuous, quantitative and discrete
- some variables have ordered values: the amount of perceived value (order, quantitative), but no units
  - education level (1: worthless 2: slightly 3: middling 4: reasonably 5: invaluable): average or count the number in each category

example

- age: teen, adult, senior ( )
- age: average age ( )
- amount of time it takes to assemble a simple puzzle ( )
- province in which a person lives ( )
- number of students in a first grade classroom ( )
- rating of a newly elected politician (excellent, good, fair, poor) ( )
- mark out of 100 obtained on a statistic test ( )
- letter grade obtained on a statistic test ( )
- number of red cars in a parking lot ( )

descriptive statistics

- once collected, it is often useful to describe the data through displays or summaries that bring attention to certain characteristics which may be of interest
- the techniques used for such descriptions encompass the area of "descriptive statistics"

displaying quantitative data

- stem -and- leaf plot; frequency histogram
  - does the histogram have humps ? (shape, center, spread)
  - modes (uni- modal, bimodal, multi-modal, no mode (uniform);
  - symmetric or asymmetric ? (tails: thinner ends of distribution - skewed to the side of the long tail)
  - do any unusual features stick out ? (outliers: stand off away from the body of distribution; are they informative ? or mistake ? - treat them specially and discuss it when you tell about your data)

stem-and-leaf plot by hand

- 31 13 8 21 16 40 42 20 28 29 33 45 41 40 20 22 28 34 39 12 19 27 22 30

1. separate each observation into two parts: stem and leaf
2. list the stem in increasing order from top to bottom in a column; draw a vertical line to separate stem from leaf; add the leaf to the right of the line
3. arrange the leaf in increasing order from left to right in each stem
4. provide a key to your stem and leaf plot: leaf unit (LU=1); stem unit (SU=10)

- if the number of leaf on each stem are too large, increase the number of stem; divide each stem into 2 points; leaf 0-4 goes to upper stem; leaf 5-9 goes to lower stem

stem-and-leaf plot by hand

-

data

- fish: dover sole are a species of flatfish: 50 Dover sole taken from shallow water (1200 to 1800 feet) and 50 Dover sole taken from deep water (2700 to 3300 feet) were examined in terms of their body length, depth, width, weight and gender
- length in millimeters of 58 male dover sole:

  386 358 370 381 398 341 339 347 397 387 340 427 381 394 384
  331 415 420 395 366 371 364 408 398 389 405 339 390 414 410
  436 439 419 423 421 388 403 411 392 419 398 409 371 371 373
  420 387 410 417 451 346 393 439 374 412 404 425 413

- length in millimeters of 42 female dover sole:

  363 345 361 337 442 351 341 357 292 373 360 380 375 357 388 384 349 350 386 394 334
  379 355 376 382 327 393 366 316 389 390 405 356 342 400 366 385 362 369 412 364 457

stem-and-leaf plot by Minitab

```
Stem-and-Leaf Display: male

Stem-and-leaf of male  N  = 58
Leaf Unit = 1.0


 3   33  199
 7   34  0167
 8   35  8
 10  36  46
 16  37  011134
 24  38  11467789
(9)  39  023457888
 25  40  34589
 20  41  0012345799
 10  42  001357
 4   43  699
 1   44
 1   45  1
```

Minitab: Graph >Stem-and-Leaf ...> enter a variable

(relative) frequency histogram by hand

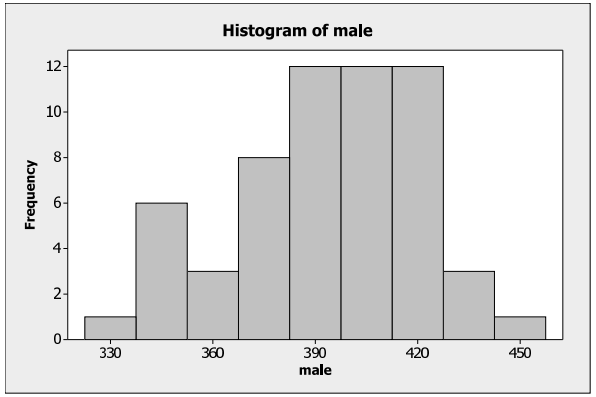
- a bar graph in which the height of the bar represents the proportion or relative frequency of occurrence for a particular class of the variable being measured
- 82 85 73 69 77 72 68 68 73 65 51 97 76 77 76 41 75 72 74 68 74 66 69 55 68 79 72 65 86 89 68 54 94 58 50 88
  1. choose the number o classes, say 6 classes
  2. calculate the approximate class width by dividing the difference between smallest and largest value by the number of classes
  3. locate the class boundaries (frequency distribution table)
  4. construct the histogram plotting classes intervals on the horizontal axis and relative frequencies as the height of the bars

(relative) frequency histogram by hand

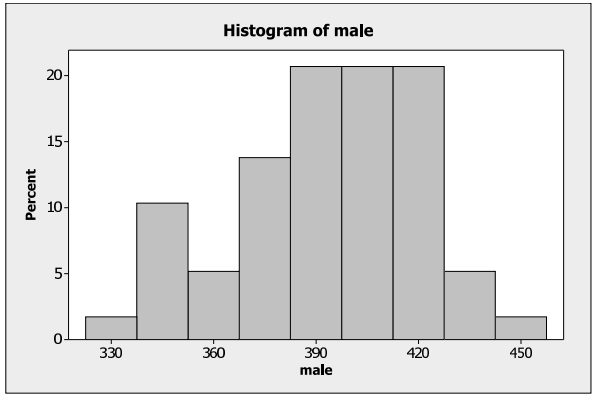
-

histogram (fish data) by Minitab

frequency historgram of male



percent histogram of male



Minitab: Graph > Histogram > Simple (enter variables)

displaying quantitative data numerically

- data: management in a medical-device manufacturing facility wishes to compare two water -purificaiton systems with respect to pH value. There are two systems: ultra-pure water system and deionizing system

  ultra-pure:

  $5.04, 5.18, 5.32, 5.33, 5.39, 5.09, 5.41, 5.20, 5.31, 5.46, 4.31, 5.40$

  deionizing:

  $5.18, 5.49, 5.34, 5.34, 5.19, 5.29, 5.33, 5.45, 5.42, 5.63, 5.55, 4.39$

mean, mode and median

- if you had to pick a single number to describe all the data, what would you pick ?
  - mean: the average of the values observed
    ($\bar{x} = \frac{x_1+x_2+\cdots+x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$)
  - median: the middle value that divides the data into two equal groups: 50% of the observation in the sample are less than or equal its value
    - order the values first- counting in the ends
    - find the value in the $\frac{n+1}{2}$ position
    - if n is even, median is the average of the two value in the $\frac{n}{2}$ and $\frac{n}{2}+1$ position
  - mode: the most frequently occurring value

example

- data with pH values of ultra-pure system

  $5.04, 5.18, 5.32, 5.33, 5.39, 5.09, 5.41, 5.20, 5.31, 5.46, 4.31, 5.40$

  – mean $= \frac{5.04+\cdots+5.40}{12} = 5.2033$

  – median: $\frac{12+1}{2} = 6.5$ position $\Longrightarrow$ median=average of the values in the 6 and 7 position

  $4.31, 5.04, 5.09, 5.18, 5.20, 5.31, 5.32, 5.33, 5.39, 5.40, 5.41, 5.46$

  that is, $\frac{5.31+5.32}{2} = 5.315$

- data with pH values of deionizing system

  $4.39, 5.18, 5.19, 5.29, 5.33,$ `5.34, 5.34`$, 5.42, 5.45, 5.49, 5.55, 5.63$

  – mode: 5.34

new data

- add 10.7 to the ultra-pure system data
  - $5.04, 5.18, 5.32, 5.33, 5.39, 5.09, 5.41, 5.20, 5.31, 5.46, 4.31, 5.40, 10.7$
  - mean $= \frac{5.04+\cdots+5.40+10.7}{13} = 5.6261$
  - median: $\frac{13+1}{2} = 7$ position $\Longrightarrow$ median

    $4.31, 5.04, 5.09, 5.18, 5.20, 5.31, 5.32, 5.33, 5.39, 5.40, 5.41, 5.46, 10.7$

    median $= 5.32$

sensitive ?

- add 10.7 to the ultra-pure system data
  - observation: 10.7 is outlier, i.e. it is not consistent with all the other points
  - the mean of the sample increased substantially (5.2033 to 5.6261) , but the increase in median is marginal (5.315 to 5.32); the sample median is less sensitive to irregular data points than the sample mean
  - such points can occur when the random sample is taken from a population, the variable of which possesses a skewed distribution - the sample median should be used as the appropriate measure of center

variability

- money (economics); life (biology); why (philosophy); what, where and when (history); how (engineering); variation (statistics)
  - range=max value - min value
  - variance: horizontal distances between measurements (x_i) and mean ($\bar{x}$) will help to measure the variability
  - one way to combine the n deviations into one numerical measure is to average them, but it will not work; why ?

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

variability

- the variance of a population of N measurements: the average of the squares of the deviations of the measurements about their mean μ

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{N}$$

- the variance of a sample of n measurements: the average of the squares of the deviations of the measurements about their mean $\bar{x}$ divided by $n - 1$

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

- sample standard deviation $\left(\hat{\sigma} = \sqrt{\hat{\sigma}^2}\right)$; quantifies the spread in the original data, in the same units as the original data

calculation

- sample variance:

sample	pH value	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
1	5.04	$5.04 - 5.2033 = -0.1633$	0.0267
2	5.18	$5.18 - 5.2033 = -0.0233$	0.0005
$\vdots$	$\vdots$	$\vdots$	$\vdots$
12	5.40	$5.40 - 5.2033 = \ 0.1967$	0.0387
sum			1.0613

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{1.0613}{11} = 0.0965$$

- standard deviation : $\hat{\sigma} = \sqrt{0.0965} = 0.3106$

calculation

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

Tchebysheff's theorem

- given a number k greater than or equal to 1 and a set of n measurements, at least $\left[1 - \frac{1}{k^2}\right]$ of measurements will lie within k standard deviation of their mean
  - at least $\frac{3}{4}$ of the measurements lie in the interval $(\mu - 2\sigma, \mu + 2\sigma)$
  - at least $\frac{8}{9}$ of the measurements lie in the interval $(\mu - 3\sigma, \mu + 3\sigma)$
  - at least none of the measurements lie in the interval $(\mu - \sigma, \mu + \sigma)$

eg: $n = 20$, $\bar{x} = 10$ and $\hat{\sigma}^2 = 4$

- at least $\frac{3}{4}$ of the measurements lie in the interval ( )
- at least $\frac{8}{9}$ of the measurements lie in the interval ( )

empirical rule

- given a distribution of measurements that is approximately mound-shaped:
  - the interval $(\mu - \sigma, \mu + \sigma)$ contains approximately 68% of the measurements
  - the interval $(\mu - 2\sigma, \mu + 2\sigma)$ contains approximately 95% of the measurements
  - the interval $(\mu - 3\sigma, \mu + 3\sigma)$ contains approximately almost all of the measurements

comments

- for a bell-shaped distribution, the empirical rule is more accurate than Tchebysheff's theorem
- according to both rules, most of the observations lie within the interval $(\bar{x} - 2\hat{\sigma}, \bar{x} + 2\hat{\sigma})$
- the range of data is approximately $4\hat{\sigma}$

  i.e. the range $\approx (\bar{x} + 2\hat{\sigma}) - (\bar{x} - 2\hat{\sigma}) = 4\hat{\sigma}$
- use the range approximation to check the calculation of the sample standard deviation; the range for a sample of n measurements will depend on the sample size n
  - $n \geq 50$, range $\approx 6\hat{\sigma}$
  - $5 \geq n < 50$, range $\approx 4\hat{\sigma}$
  - $n < 5$, range $\approx 2.5\hat{\sigma}$
- the range approximation is not included to provide an accurate value for $\hat{\sigma}$, rather its purpose is to detect gross error in calculating $\hat{\sigma}$

measure of relative standing

- need to know the position of one observation relative to others in a set of data
  - the sample Z-score$=\frac{x_i-\bar{x}}{\hat{\sigma}}$
  - it measures the distance between an observation and the mean, measured in units of standard deviation

eg: 46 54 42 46 32 $\bar{x} = 44$, $\hat{\sigma}^2 = 64$

- 54 lies ( ) standard deviations (above or below) the sample mean
- if Z-score exceeding 3 in absolute value are very unlikely, such an unusually large or small observation is called ( )

percentile and quartiles

- p^{th} percentile: the value of observations that exceeds $p\%$ of the measurements and is less than the remaining $(100 - p)\%$
- the lower quartile, Q_1: the value corresponding to the $\frac{n+1}{4}$ ordered observation
- the second quartile is the median
- the upper quartile, Q_3: the value corresponding to the $\frac{3(n+1)}{4}$ ordered observation
- inter-quartile range (IQR) $= Q_3 - Q_1$

  cutting data into four quarters: lower quartile (Q_1), median, upper quartile (Q_3)

example

$4.31, 5.04, 5.09, 5.18, 5.20, 5.31, 5.32, 5.33, 5.39, 5.40, 5.41, 5.46$

- range $= 5.46 - 4.31 = 1.15$
- $Q_1 = 5.09$ since $\frac{13}{4} = 3.25 \Longrightarrow 3^{th}$ position
- $Q_3 = 5.40$ since $\frac{3 \times 13}{4} = 9.75 \Longrightarrow 10^{th}$ position
- inter-quartile range (IQR) $= Q_3 - Q_1 = 5.40 - 5.09 = 0.31$

$7, 17, 9, 28, 11, 82, 21, 18, 19 \Longrightarrow 7, 9, 11, 17, 18, 19, 21, 28, 82$

- range $= 82 - 7$
- the position of $Q_1 = 0.25(n + 1) = 2.5$
- the position of $Q_3 = 0.75(n + 1) = 7.5$
- $Q_1 = \frac{9+11}{2} = 10$
- $Q_3 = \frac{21+28}{2} = 24.5$
- inter-quartile range (IQR) $= Q_3 - Q_1$

descriptive (fish data) by Minitab

Descriptive Statistics: male, female

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
male	58	0	393.26	3.72	28.35	331.00	372.50	396.00	414.25	451.00
female	42	0	369.29	4.68	30.32	292.00	350.75	366.00	386.50	457.00

Minitab: Stat > Basic Statistics > Display Descriptive Statistics (or Graphical Summary)

box plot

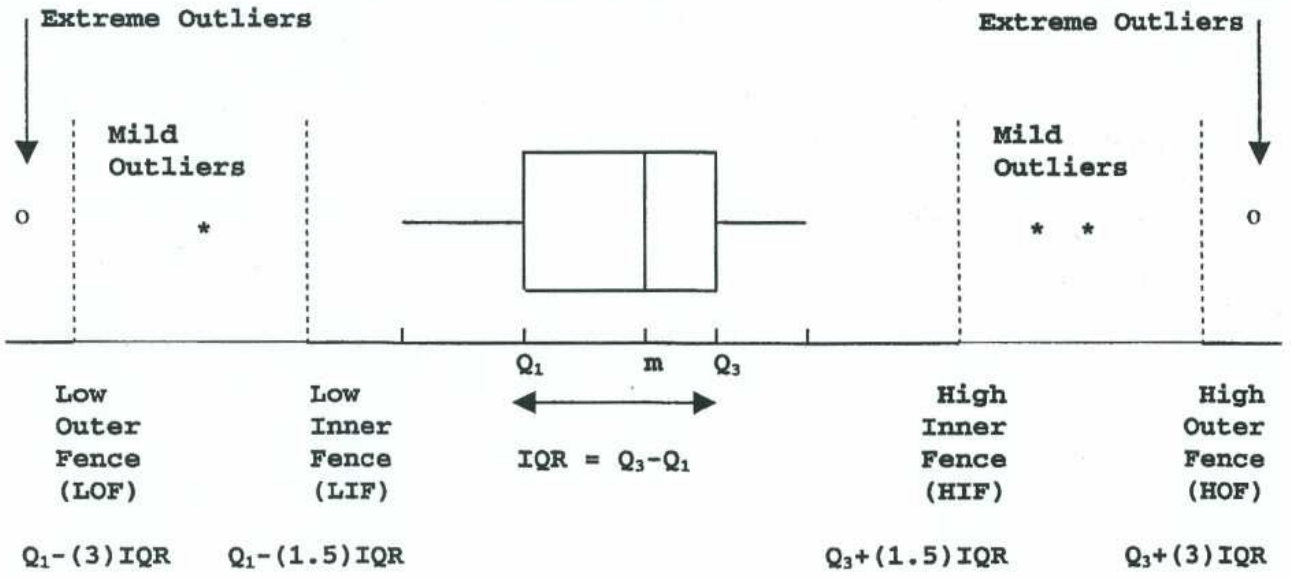
- from box plot, detect any skewness in the shape of the distribution and see whether there are outliers in the data set
  1. min, max, Q_1, median, Q_3
  2. calculate inner fences
     - lower fence (LF) $= Q_1 - 1.5(IQR)$
     - upper fence (UF) $= Q_3 + 1.5(IQR)$
  3. calculate outer fences
     - lower fence (LF) $= Q_1 - 3(IQR)$
     - upper fence (UF) $= Q_3 + 3(IQR)$
  - draw a horizontal line representing the scale of measurements; form a box; draw a vertical line through the box at the location of the median

box plot

- look for Z-score grater than 2 or 3 in absolute value
- measurements between inner and outer fences: mild outliers
- measurements beyond the outer fences: extreme outliers
- whiskers are marked by adjacent values, the last observations in the ordered set before the inner fences
- skewed distribution usually have a long whiskers in the direction of the skewness; the median line is drawn away from the direction of the skewness

Boxplots

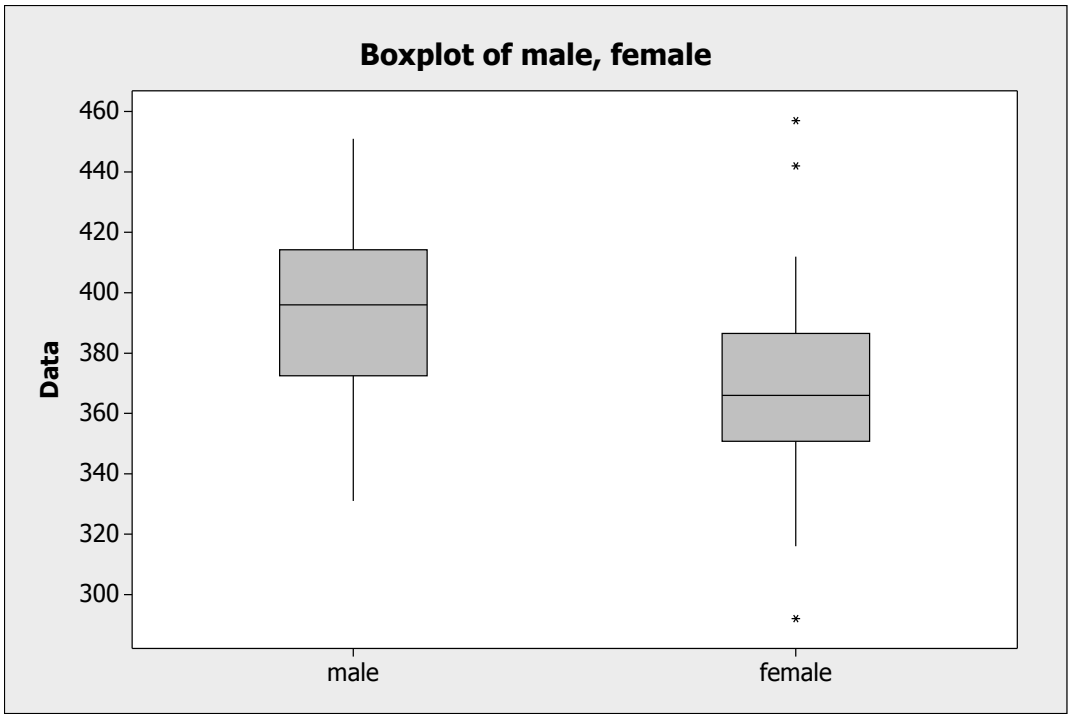
Boxplots are graphical data displays which highlight several key features of a data set. The layout for a boxplot is shown here:



box plot by hand

- 25 26 25 28 26 28 28 27 26 27 24 25 26 26

box plot (fish data) by Minitab



Minitab: Graph > BoxPlot> One Y (simple) or Multiple Y's (simple)

symmetric and skewed

- symmetric: the upper half is a mirror image of the lower half of the distribution
  - distance (Q_1 to median) = distance (Q_3 to median)
  - distance (min to Q_1) = distance (max to Q_3)
  - median= mean= $\frac{min+max}{2} = \frac{Q_1+Q_3}{2}$
- skewed: not symmetrical
  - mean $>$ median: skewed to the right (a greater proportion of the measurements lies to the right of the peak value)
  - mean $<$ median: skewed to the left (a greater proportion of the measurements lies to the left of the peak value)

interpreting graphs

- check the horizontal and vertical data to see what is being measured (scales)
- examine the horizontal axis to find the center of the distribution (location)
- examine the shape of the distribution symmetric or skewness
- check for outliers (determine if any obtains unusually larger than or smaller than all the other data values)

inferential statistics

- the description of data is not usually an end in itself: we may be interested in what the data can tell us about the group from which it was drawn, but which remains mostly un-sampled and thus unknown
- the methods employed in dealing with such generalization are known under the label of *inferential statistics*. i.e., generalizing the information in a sample to the population from which it was drawn