Descriptive Stats: categorical variables

Objectives

(1) Learn to graphically display categorical variables,

(2) Learn to analyze the relationship between 2 categorical variables.

Concept briefs:

* Frequency count = # of times each category appears in a cat. variable.

* Relative freq. OR Frequency % = Divide freq. count by total # of data values.

* Bar graphs, Pie charts = Common ways to display categorical variables.

* Association = If there is relationship between 2 cat. variables, such that values of one affect the other, we say there is an "association."

* Marginal / Conditional distributions = Two-way tables (contingency tables) used for studying relationship between 2 categorical variables.

* Segmented bar graphs, Pie charts = Used for displaying relationship between 2 categorical variables.

What is marginal distribution?

For any categorical variable, the marginal distribution shows the percentage breakdown of the categories. In other words, it is just the relative frequency.

Example

The number of students in this class who classify themselves as conservative, liberal, or moderate is shown below

conservative	liberal	moderate
6	10	12

Find the marginal distribution of political views of students in this class.

Solution:

Total = 6+10+12 = 28: conservative = 6/28, liberal = 10/28, mod = 12/28.

So, the marginal distribution of political views is:

conservative	liberal	moderate
21.4%	35.7%	42.9%

What is conditional distribution?

* This concept requires 2 categorical variables, say, V1 and V2.

* A conditional distribution shows the percentage breakdown of V1 for each categorical value of V2. [OR vice versa]

Example

The number of women and men in this class who classify themselves as conservative, liberal, or moderate is shown below

	conservative	liberal	moderate
Women	3	8	9
Men	3	2	3

(A) Find the conditional distribution of political views by sex.
(B) Find the conditional distribution of sex by political views.

Solution:

(A) Conditional distribution of political views by sex means
% breakdown of conservative, liberal, moderate among women,
and % breakdown of conservative, liberal, moderate among men.
Total women = 3+8+9 = 20: conservative = 3/20, liberal = 8/20, mod = 9/20.
Total men = 3+2+3 = 8: conservative = 3/8, liberal = 2/8, mod = 3/8.
Conditional distribution of political views by sex:

	conservative	liberal	moderate	Total
Women	15%	40%	45%	100%
Men	37.5%	25%	37.5%	100%

(B) Conditional distribution of sex by political views means % breakdown of women/men among conservatives, liberals, moderates. Notice that this adds to 100% vertically (not horizontally).

	conservative	liberal	moderate
Women	50%	80%	75%
Men	50%	20%	25%
Total	100%	100%	100%

Key point: There are 2 different ways to calculate conditional distributions (vertical, horizontal). It is important to know which way is right.

Association: How to determine whether it exists?

What is an association?

* There must be more than 1 categorical variable for an association

* If the values of one affect the other, there is an "association"

e.g., V1 = height of students in this class
V2 = political views of students in class [conservative, liberal, mod.]
V3 = major field of study (of students in this class)
V4 = political views of students' parents

There is likely no association between V1 and V2.

There is very likely an association between V2 and V4.

How to tell if there is association bet. 2 variables V1 and V2

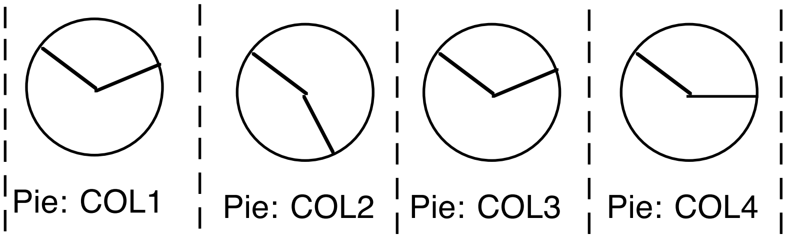
* Make a conditional distribution table that shows % distribution of V2 for each category of V1 (see illustration below).

* Check whether conditional distribution of V2 changes as V1 changes. If 'yes' there is an association; if 'no' there is no association. Best way to see this is via pie charts or segmented bar graphs.

V1 →

V2 ↓

	COL 1	COL 2	COL 3	COL 4
ROW 1	32%	54%	32%	40%
ROW 2	68%	46%	68%	60%



Think: Does it matter which way you calculate the %'s (by row or column)?

Ex. 42 (see copy of exercise on next page)

Strategy1: Let V1=Do you have hepatitis C? (2 categories: yes, no.)

V2=Who did your tattoo? (3 cats: comm. parlor, elsewhere, no tattoo.)

Find conditional distribution of V1 for each value of V2. Compare conditional distributions & see if they are the same or different.

V2=Commercial parlor

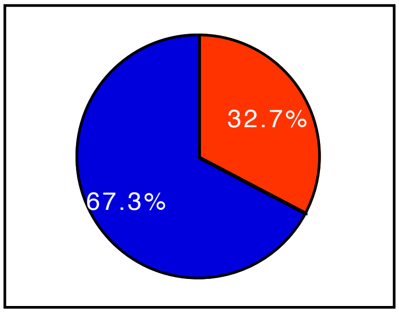
Hep. C	No Hep. C
17 32.7 %	35 67.3 %

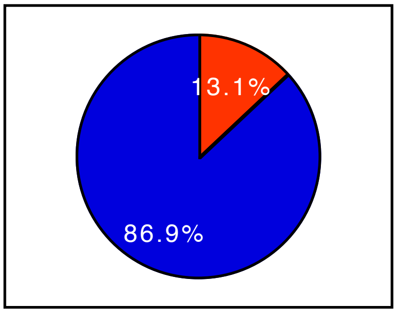
V2=Elsewhere

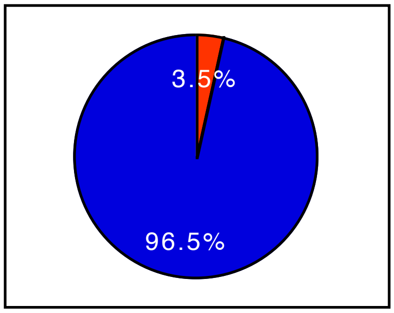
Hep. C	No Hep. C
8 13.1 %	53 86.9 %

V2=No tattoo

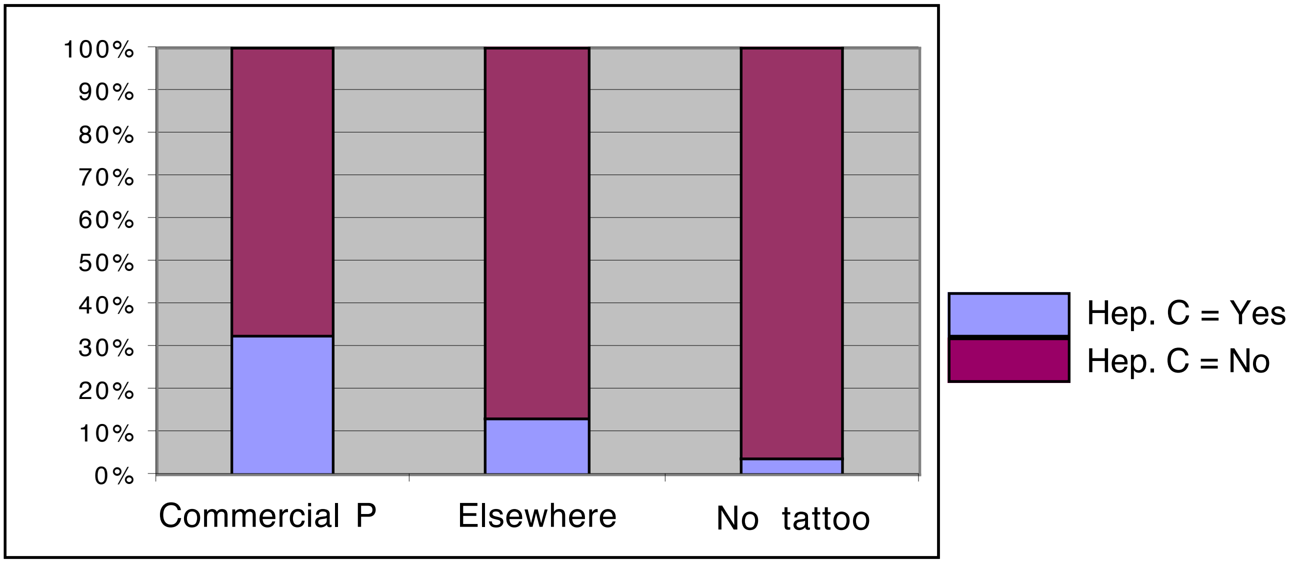
Hep. C	No Hep. C
18 3.5 %	495 96.5 %







An alternative way to visualize this is via segmented bar graphs:



42. Tattoos A study by the University of Texas Southwestern Medical Center examined 626 people to see if an increased risk of contracting hepatitis C was associated with having a tattoo. If the subject had a tattoo, researchers asked whether it had been done in a commercial tattoo parlor or elsewhere. Write a brief description of the association between tattooing and hepatitis C, including an appropriate graphical display.

	Tattoo Done in Commercial Parlor	Tattoo Done Elsewhere	No Tattoo
Has Hepatitis C	17	8	18
No Hepatitis C	35	53	495

Do the above results suggest any association between Hep. C and tattoos?

Strategy2: Find conditional distribution of V2 for each value of V1. Compare conditional distributions.

V1=Yes (has Hep. C)

Commercial parlor	Elsewhere	No tattoo
17 39.5 %	8 18.6 %	18 41.9 %

V1=No (No Hep. C)

Commercial parlor	Elsewhere	No tattoo
35 6.0 %	53 9.1 %	495 84.9 %

Make suitable graphical displays showing these conditional distributions.

Summarize the results of your analysis:

According to these data, there is a strong association between having a tattoo and the likelihood of contracting hepatitis C. Among the 3 groups of people studied, those whose tattoos were done in commercial parlors had the highest incidence of hepatitis C (33%). It was significantly lower (13%) for the group whose tattoos were done elsewhere, and very small for those with no tattoos.

Related questions

(a) Find the marginal distribution and conditional distribution of hepatitis C among the 3 groups of people studied.

Strategy: Marginal ⇒ combine all categories of V2 for each category of V1.

Yes Hep. C = 17+8+18 = 43 No Hep. C = 35+53+495 = 583

Marginal distribution: 6.9 % have Hep. C 93.1 % do not

Conditional distribution is given in the solution above (see 3 tables).

(b) Find the marginal distribution and conditional distribution of tattoo groups studied among the hepatitis C categories.

[solution is left as an exercise]

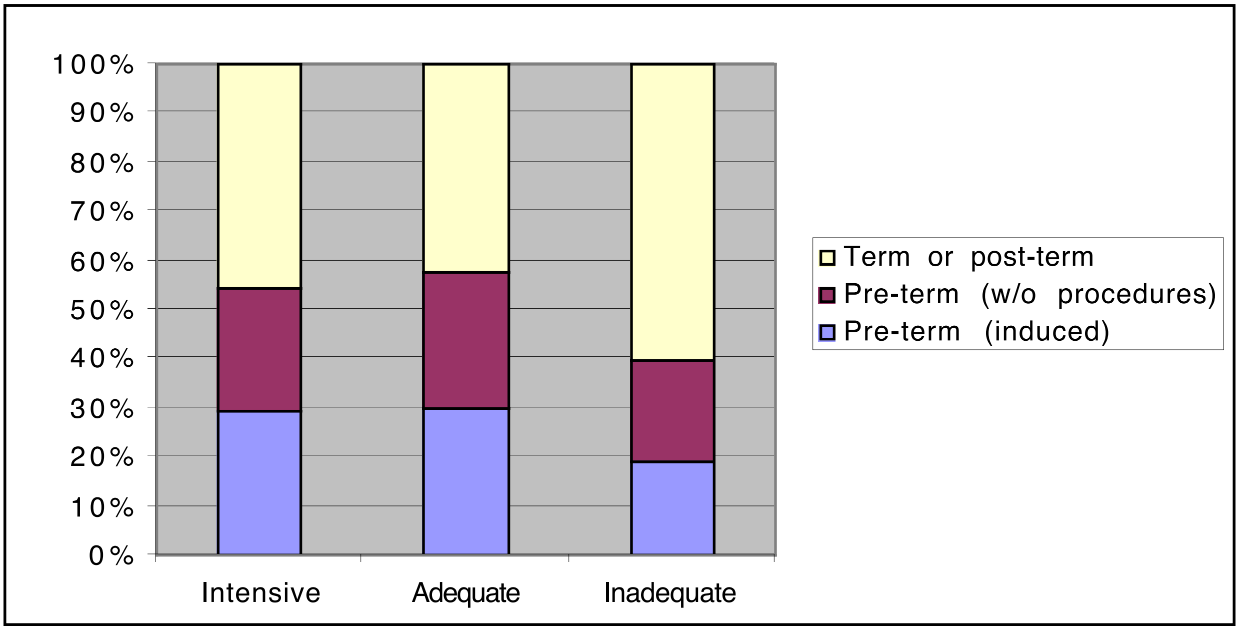
Ex. 36, (see copy of exercise on next pg.)

The 2 variables are:

V1=level of pre-natal care (w/ 3 categories: intensive, adequate, inadeq.)

V2=type of twin-birth (3 cat: preterm-induced, preterm, term or post-term)

To graphically explore the association, we can use segmented bar graphs or pie charts.



These data suggest there is some association between the level of pre-natal care a woman receives, and the type of twin-births she is likely to give. Those who receive inadequate care appear more likely to give birth to term or post-term twins than the other two groups. On the other hand, there doesn't appear to be much difference in types of births given by women who receive intensive or adequate care.

Extra exercise:

Find the marginal distribution and conditional distribution of birth types for different levels of pre-natal care.

36. Twin births In 2000, the *Journal of the American Medical Association (JAMA)* published a study that examined pregnancies that resulted in the birth of twins. Births were classified as preterm with intervention (induced labor or cesarean), preterm without procedures, or term/post-term. Researchers also classified the pregnancies by the level of prenatal medical care the mother received (inadequate, adequate, or intensive). The data, from the years 1995–1997, are summarized in the table below. Figures are in thousands of births. (Source: *JAMA* 284 [2000]:335–341)

Twin Births 1995–1997 (in thousands)

Level of Prenatal Care		Preterm (induced or cesarean)	Preterm (without procedures)	Term or Post-Term	Total
	Intensive	18	15	28	**61**
	Adequate	46	43	65	**154**
	Inadequate	12	13	38	**63**
	Total	**76**	**71**	**131**	**278**

Segmented bargraphs - some notes. [self-study]

* Only used for displaying or comparing conditional distributions.
* They are a great display when used for for this purpose.
* Some pointers on how to make a segmented bargraph:

[E.g., let V1=political views of students, V2=sex of student]

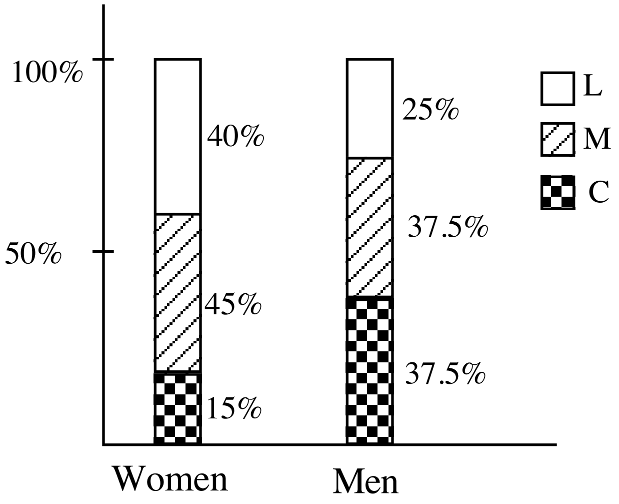
(1) The vertical axis is in %.

(2) All bars must have the same total height [i.e., 100%].

(3) Bars show conditional distribution of V1 for each category of V2. Thus, there must be as many bars as categories in V2.

(4) The # of colors (or shades) in each bar depends on the # of categories in V1.

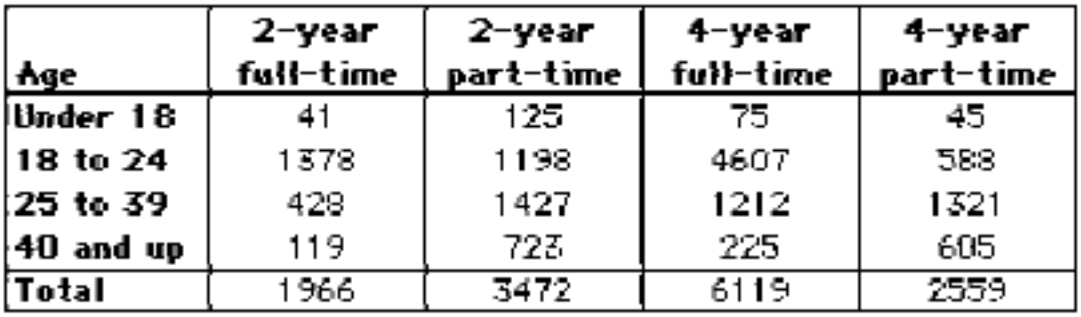
(5) Use consistent shades & position of categories within bars.



Exercise: Make a segmented bargraph showing conditional distribution of sex by political views. How many bars would it have? How many categories within each bar?

MRB-2. College Undergraduates

This two-way table reports data on all undergraduate students enrolled in U.S. colleges and universities in the fall of 1995 whose age was known.



Age	2-year full-time	2-year part-time	4-year full-time	4-year part-time
Under 18	41	125	75	45
18 to 24	1378	1198	4607	588
25 to 39	428	1427	1212	1321
40 and up	119	723	225	605
Total	1966	3472	6119	2559

1. How many undergraduate students were enrolled in colleges and universities?

2. What percent of all undergraduate students were 18 to 24 years old in the fall of the academic year?

3. Find the percent of the undergraduates enrolled in each of the four types of program who were 18 to 24 years old. Make a bar graph to compare these percents.

4. The 18 to 24 group is the traditional age group for college students. Briefly summarize what you have learned from the data about the extent to which this group predominates in different kinds of college programs.

Displaying & comparing quant. data

Objective

(1) Learn to graphically display and compare quantitative variables.

(2) Learn to summarize shape/center/spread of quantitative distributions using numbers.

(3) Learn how to choose good “summary statistics.”

Concept briefs:

* Numerical range of a quantitative variable.

* Distribution of quantitative variable = Frequency of occurrence within different numerical groups (or, bins).

* Basic display types = Histogram, dotplot, stem-and-leaf display.

* Key characteristics to observe = Shape, center, spread, outliers.

* Shape/spread descriptors = How many modes; what kind of skew; outliers.

* Mode = Refers to a "crest" in the overall profile of histogram (or any display).

* Left skew = Longish tail to the left. (Right skew = tail to the right).

* Median = Middle value of numerically ordered data.

* Mean = Arithmetic average of data values.

* Quartiles = Numerically ordered data; split into 4 equal groups;

find data values at interfaces between groups.

* IQR = Spans the central 50% of (ordered) data values.

Measure of spread that typically goes with median.

* Standard Deviation = Measure of spread that typically goes with mean.

* 5-number summary = Median/IQR based summary that gives: min, Q1,

med,Q3, max.

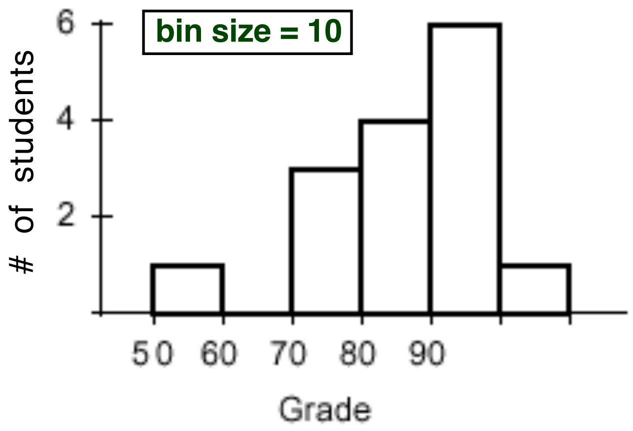
* Box plot = Graphical display of median/IQR based summary statistics.

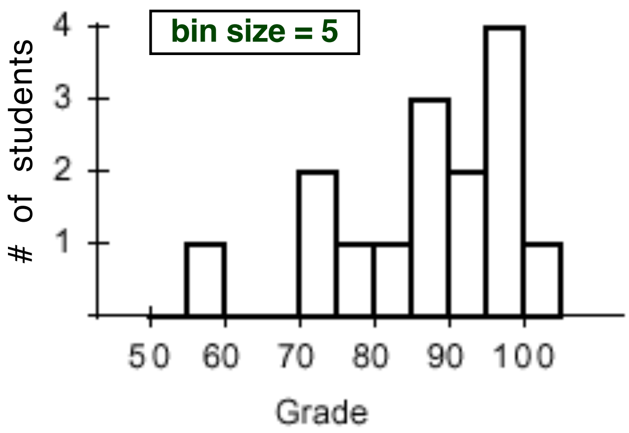
* Summary statistics choices = (1) Median & IQR (OR, 5-number summary);

(2) Mean & SD

Example: Histograms below show distribution of student grades in fictional "Statistics" class: 86, 91, 80, 74, 70, 100, 97, 92, 56, 87, 85, 95, 95, 98, 78.

[To learn how to make histograms step-by-step, see recipe at the end of this set of notes.]





Key things to observe in a histogram:

Shape/center/spread/outliers

- One mode
- Skewed to the left
- One low outlier in the 50's
- Center around 85-90

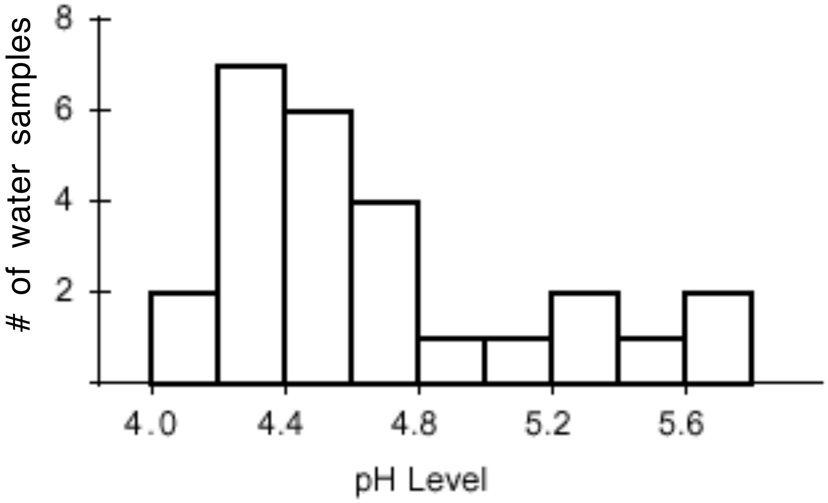
- One primary mode
- Skewed to the left
- One low outlier in 50's
- Center near 85-90

Ex. 51 (see copy of exercise on next pg.)

Solution:

* The histogram below displays their data.

* The distribution is skewed to the right, and pH levels are below 7, indicating that most samples of water the researchers collected were acidic. The distribution has one main mode; possibly a second smaller mode as well. The center is around pH levels of 4.4-4.6, which is well into the acidic range. The range is from pH levels of about 4.1 to 5.8.



51. Acid rain Two researchers measured the pH (a scale on which a value of 7 is neutral and values below 7 are acidic) of water collected from rain and snow over a 6-month period in Allegheny County, Pennsylvania. Describe their data with a graph and a few sentences:

4.57 5.62 4.12 5.29 4.64 4.31 4.30 4.39 4.45 5.67
4.39 4.52 4.26 4.26 4.40 5.78 4.73 4.56 5.08 4.41
4.12 5.51 4.82 4.63 4.29 4.60

How to make a histogram. [Recipe]

* Use the following dataset for illustration (student grades in Statistics class):
86, 91, 80, 74, 70, 100, 97, 92, 56, 87, 85, 95, 95, 98, 78.

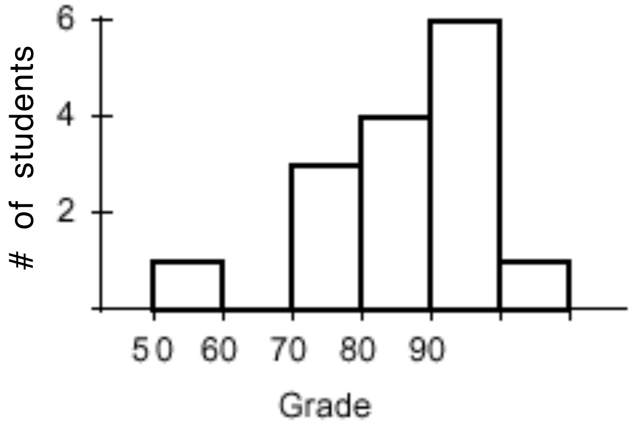
Step1: Find the range of your data. [E.g., 56 to 100 here --> 100-56 = 44.]

Step2: Select reasonable scale. [E.g., For range=44, either 10 or 5 is good.]

Step3: Make frequency table [E.g., see below for scale=10 and scale=5.]

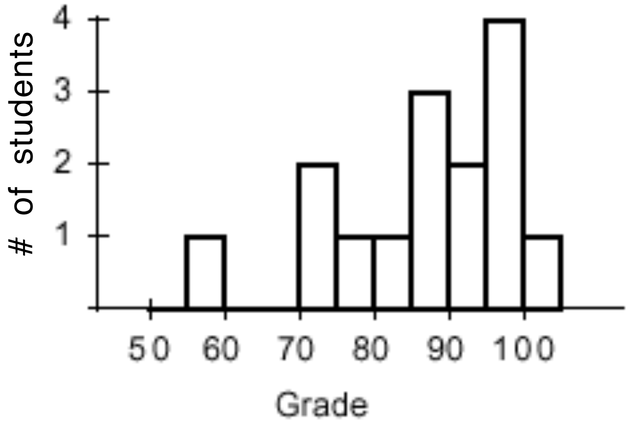
Step4: Sketch the histogram. Show clear labels on both axes.

bin	Frequency
50 to 60^-	1
60 to 70^-	0
70 to 80^-	3
80 to 90^-	4
90 to 100^-	6
100 to 110^-	1



__Note__: This frequency table has made the choice of putting "boundary" points on the right. E.g., The data value 80 is in the 80-90 bin, instead of the 70-80 bin. The 'minus' superscript on the bins says that the end-point stops short of the indicated value.

bin	Frequency
55 to 60^-	1
60 to 65^-	0
65 to 70^-	0
70 to 75^-	2
75 to 80^-	1
80 to 85^-	1
85 to 90^-	3
90 to 95^-	2
95 to 100^-	4
100 to 105^-	1



Some thinking & writing questions on basic concepts

(1) What are data? Is the following list of items data:
22.1, 14.6, 18.0, 8.7, 24.1, 4.5, 14.0 (units=inches, for all)
Why, or why not?

(2) What is the difference between categorical variables & quantitative variables? Can a categorical variable include numbers? Can a quantitative variable include letters or other symbols? Explain.

(3) What are the key differences between bar graphs & histograms? What are the key similarities?

(4) Suppose you have 2 different categorical variables. Can you display them on a common (single) bargraph? Explain why or why not.

(5) Suppose you have 2 different categorical variables. Can you display them on a common (single) pie chart? Explain.

(6) Suppose you have 2 different quantitative variables. Can you display them on a common (single) histogram? Explain.

(7) What are the key differences between bar graphs & segmented bar graphs? Explain with example.

Median & Mean

* Both measure the center of distributions, but they do it differently.

Median	Mean
(1) Arrange data items in numerically ascending or descending order.	(1) Add together all data items.
(2) Pick the data value at the exact center.	(2) Divide by total # of items.

NOTE: Both median and mean have units -- you must express them in the same units as the underlying dataset.

Example: For students in the hypothetical statistics class, find the mean and median age/grade of the first 5 records.

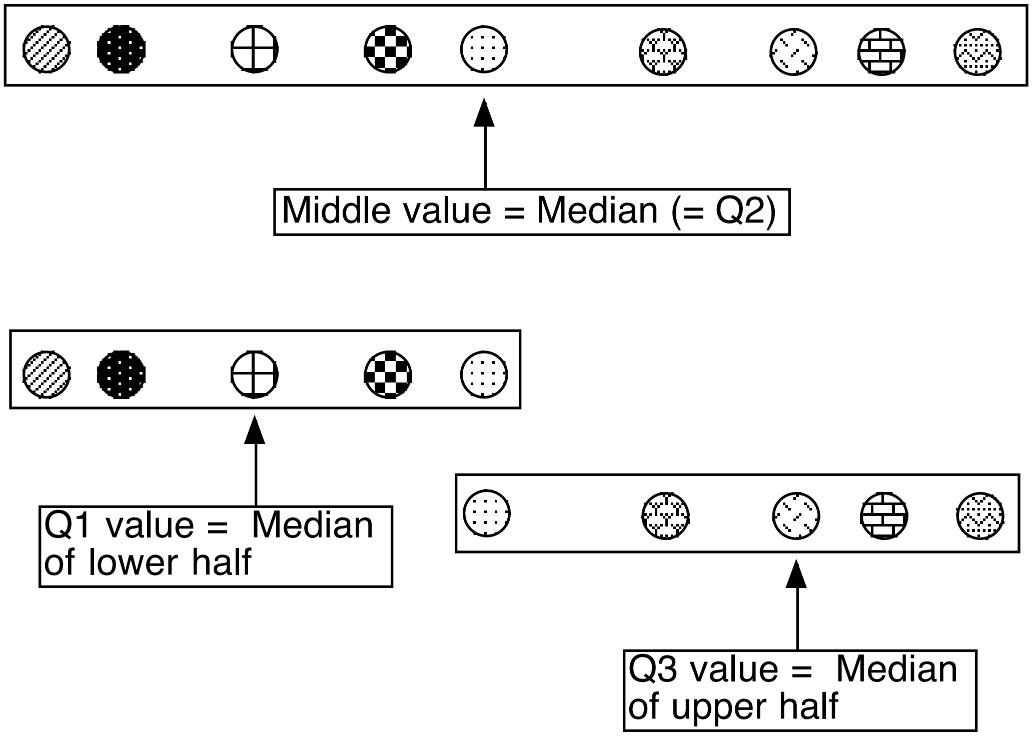
Student-ID	Age (Yrs)	Grade (%)
101	18	86
102	20	91
103	17	80
104	17	74
105	19	70

Answers: Age: median=18 years; mean=18.2 years

Grade: median=80 %; mean=80.2 %

Concept of Quartiles

Illustration: Let each “marble” represent a data value (in increasing order)



Middle value = Median (= Q2)

Q1 value = Median of lower half

Q3 value = Median of upper half

In a nutshell: Quartiles essentially split ordered data into 4 equal groups

Q1 = 1/4 data below this value [AKA: 25th percentile];

Q2 = 1/2 data below this value [AKA: 50th percentile]; etc.

Think about it: What subtle changes occur if we have even (vs. odd) number of data values? E.g., what is the median of 12, 14, 15, 15?

Inter Quartile Range (IQR) = Q3 - Q1
= 75th percentile - 25th percentile

(KEY CONCEPT)

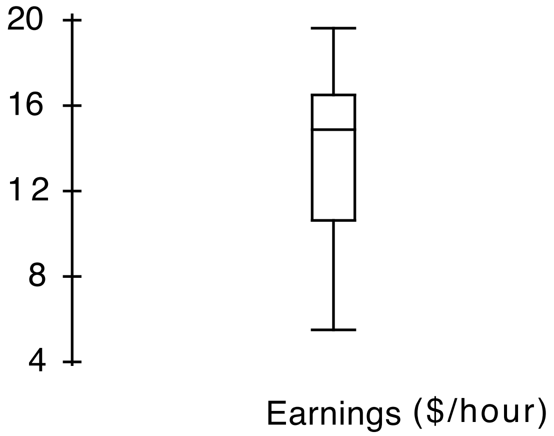
WEN-5. Labor Statistics

The Bureau of Labor Statistics collects data on employment and hourly earnings in private industry groups and publishes its findings in *Employment and Earnings*. Twenty people working in the manufacturing industry are selected at random: their hourly earnings, in dollars, are as follows.

16.70, 7.44, 13.78, 16.49, 7.49, 17.92, 17.21, 5.51, 10.40, 10.75, 15.27,
19.72, 10.68, 13.10, 14.70, 15.55, 16.67, 15.07, 14.02, 15.99

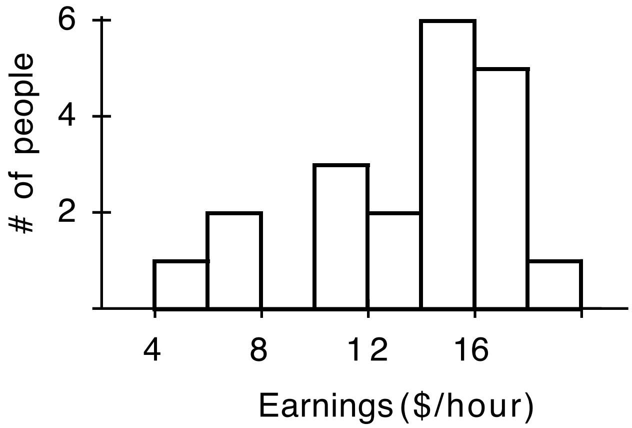
Obtain a boxplot and the five-number summary for this data.

Make a histogram of the data. What does the histogram tell you that the boxplot does not?



5-Number summary

minimum	= 5.51 $/hr
Q1	= 10.72
median	= 14.89
Q3	= 16.58
maximum	= 19.72



The histogram shows certain details about the data that the boxplot does not. For example, there is a gap between $8 and $10/hr that contains no data values. Also, there is a drop from $12 to $14, which could indicate the presence of a second mode.

Standard Deviation (SD)

SD measures the spread of a distribution **relative to** its **mean**

[In contrast to **IQR**, which measures spread **relative to** the **median**]

Key ideas [To compute the SD of a data set]

Step 1: Find the mean.
Step 2: Find "the distance" of each data value from the mean.
Step 3: Find the mean of these distances (after squaring each).
Step 4: Take the square root of Step 3.

Another key idea: The SD basically measures the average distance of the data values from the mean.

Example: For students in the hypothetical statistics class, find the standard deviation of grades for the first 5 records.

Solution:

Follow the basic steps above:

Student-ID	Grade (%)
101	86
102	91
103	80
104	74
105	70

(1) Mean = (86+91+80+74+70) / 5 = 80.2

(2) d1=(86-80.2)=5.8; d2=10.8; d3=-0.2; d4=-6.2; d5=-10.2

(3) $5.8^2 + 10.8^2 + (-0.2)^2 + (-6.2)^2 + (-10.2)^2$ = 292.8. So, the mean = 292.8 / 4=73.2

(4) $\sqrt{}$ 73.2 = 8.56

Answer: SD=8.56 % (**remember to use units!**)

What is variance?

* Another measure of spread, or variability, within a data set.

* It is almost the same as the standard deviation (SD). Numerically, it is simply the square of the SD. In practice, we actually compute the variance first, then square-root it to get the SD.

* Why is it useful?

Many theoretical concepts needed to understand & work with variability are based on the variance. This is because the variance has certain special mathematical properties that the standard deviation doesn't have. On the other hand, the standard deviation is much easier to understand & relate to in a practical sense. Thus, statisticians use both variance & SD, depending on context.

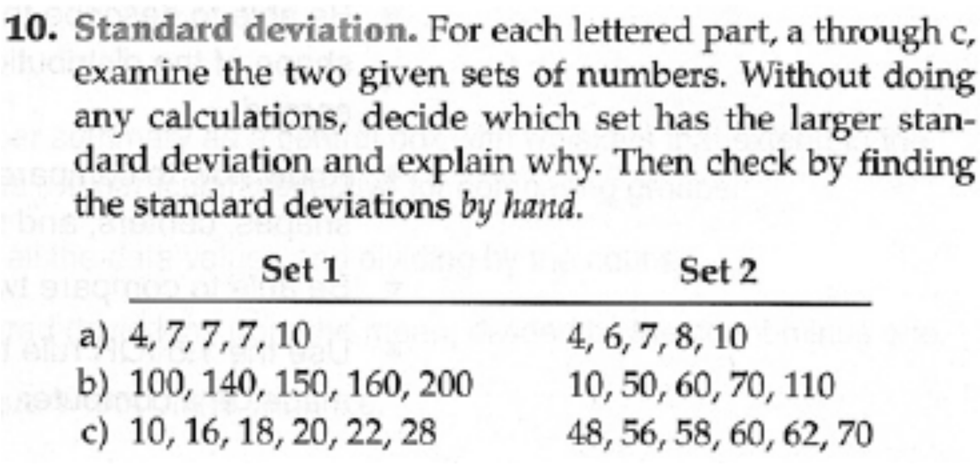
It is important to learn to get a "general feel" for the SD of small data sets. Here are some rules of thumb:

* If the range of a dataset gets larger, so does the SD.

* If most data values are "huddled" close together, the SD gets small.

* If many data values are out towards the extremes, the SD gets large.

Examples



10. Standard deviation. For each lettered part, a through c, examine the two given sets of numbers. Without doing any calculations, decide which set has the larger standard deviation and explain why. Then check by finding the standard deviations *by hand*.

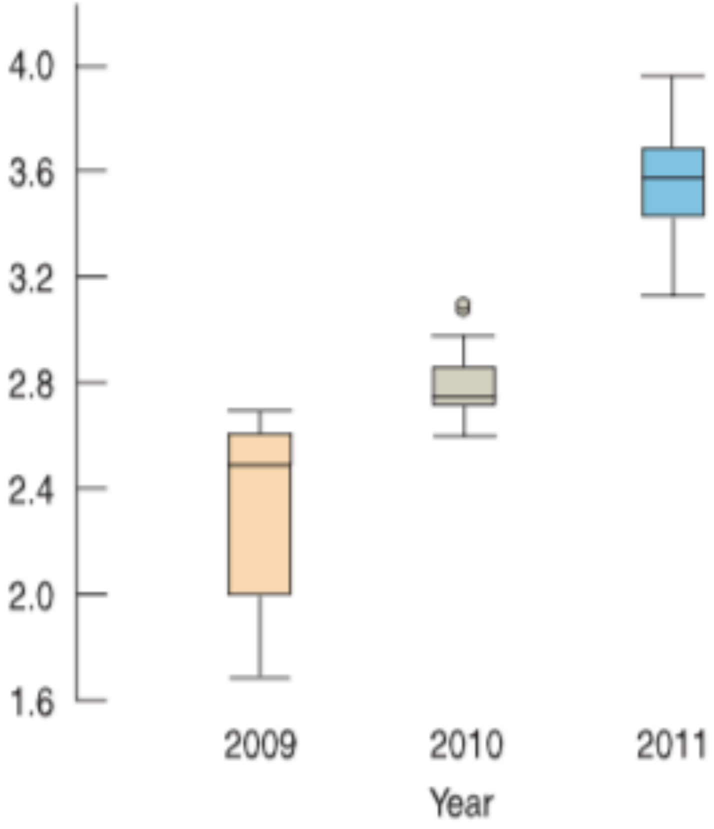
	Set 1	Set 2
a)	4, 7, 7, 7, 10	4, 6, 7, 8, 10
b)	100, 140, 150, 160, 200	10, 50, 60, 70, 110
c)	10, 16, 18, 20, 22, 28	48, 56, 58, 60, 62, 70

Ex. 26 (see copy of exercise on next pg.)

(a) Gas prices have been increasing steadily between the years 2009 and 2011. This is true not only in terms of average prices, which have gone from about $2.50/gallon in 2009 to $3.60/gallon in 2011, but also in terms of the yearly low & high price. Gas prices were much more consistent in 2010 than in the other two years. An interesting feature seen in the boxplots is that there was a significant right-skew in the prices in 2009, which means prices were much lower than the median on a few occasions. It is also interesting to note that each year's median price is as high (or higher) than the previous year's highest price.

(b) Gas prices were the least stable in 2009, because the spread of prices, measured in terms of either IQR or total range, is highest for that year.

26. Gas prices 2011 Here are boxplots of weekly gas prices for regular gas in the United States as reported by the U.S. Energy Information Administration for 2009, 2010, and 2011.



a) Compare the distribution of prices over the three years.

b) In which year were the prices least stable? Explain.

Warmup exercises

Find the median and mean for each of the following distributions, and sketch a rough histogram to show the distribution.

(1) −2, −1, 0, 1, 2.

(2) −20, −1, 0, 1, 2.

(3) −2, −1, 0, 1, 20.

(4) −2, 0, 0, 0, 1.

What is the moral of the story?

Another warmup exercise

(1) Create a dataset that satisfies the following conditions:

- It contains 10 items.
- The range is from 1 to 10.
- The mean is at least 1 unit smaller than the median.

(2) Repeat **(1)** with the following change:

- The mean is at least 1 unit larger than the median.

We now have 2 different ways of describing summary-statistics:

(1) Median (center), with IQR (spread) and five-number summary.

(2) Mean (center), with standard deviation (spread).

It is important to know which summary-statistics work best for different types of datasets and why.

Recap key points relevant to this question:

* The median & IQR are not sensitive to outliers, but the mean & SD are.

* The mean & SD are appropriate only for (roughly) symmetric distributions.

* For symmetric distributions, the median and mean are very close.

* It doesn't make sense to speak of center & spread if the distribution is not unimodal. In this case, it is best to separate the distribution into two (or more) sub-sets and look at the center/spread separately.

Rules of thumb when describing or comparing distributions

* Plot a (suitably-scaled) histogram to show the data.

* Recognize which type of summary-statistics work best for the dataset.

* Give the five-number summary and/or the mean+SD.

* Recognize any outliers and comment on them. If you have outliers, report the mean+SD with & without the outliers.

* Use boxplots when comparing two different datasets.

Ex. 34 (see copy of exercise on next pg.)

(a) Key things to notice:

> (i) mean is much larger than median --> suggests skew to the right;
>
> (ii) (median - Q1) = 15.0, but (Q3 - median) = 21.5. --> Right quartile much farther from median than the left. Confirms skew to the right.
>
> So, the distribution is not symmetric, and is skewed to the right.

(b) Think of points that might lie outside the ends of a boxplot.

Q1 - 1.5*IQR = - 36.25; Q3 + 1.5*IQR = 109.75
The maximum data value of 250 is larger than 109.75. So there is at least 1 outlier. There may be more outliers as well, since the SD + Mean = 95 --> not far from 109.75, so there are probably more values > 109.75.

(d) These 36 wineries range in size from 6 to 250 acres. The median size is 33.5 acres, so half the wineries are smaller and half are larger. The middle 50% of them range in size from 18.5 to 55 acres. The distribution of sizes is non-symmetric and skewed to the right, with at least the largest winery (at 250 acres) being an outlier.

34. Vineyards Here are summary statistics for the sizes (in acres) of Finger Lakes vineyards:

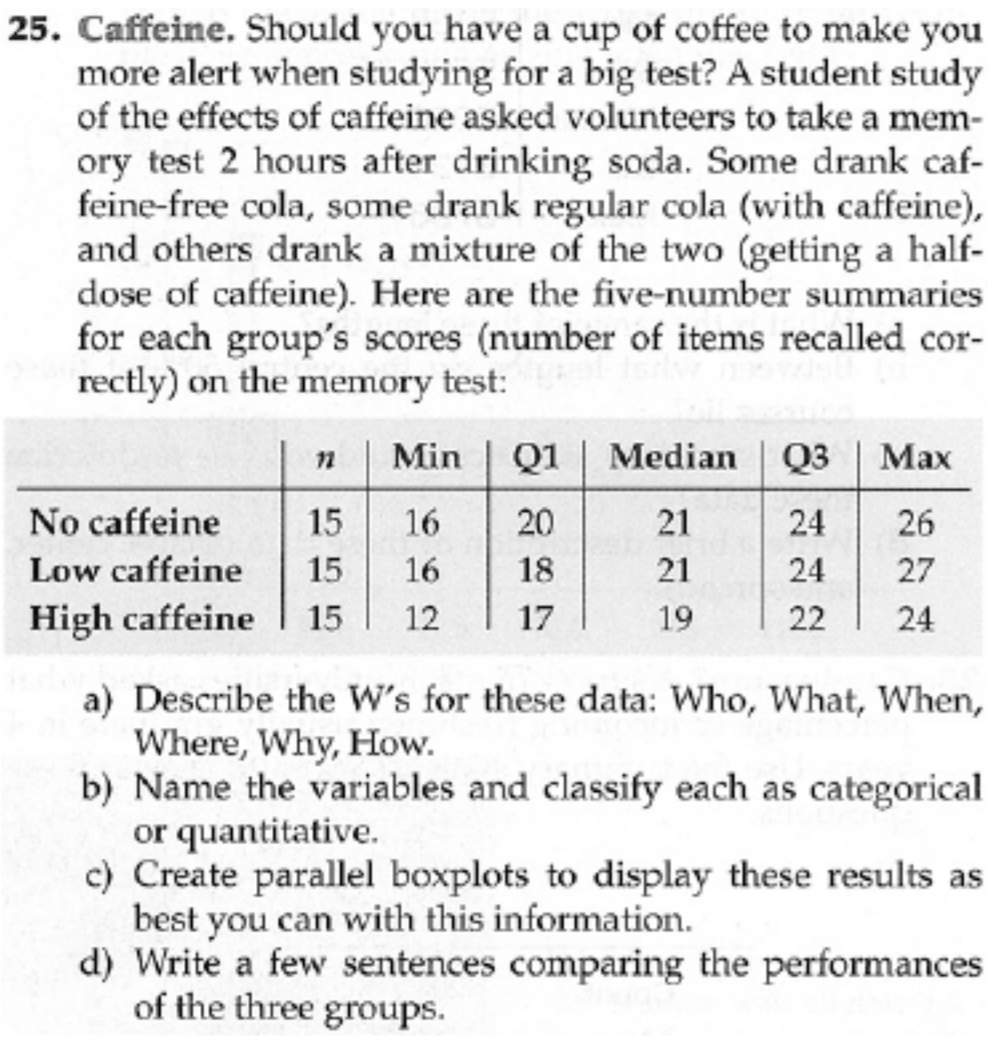
Count	36
Mean	46.50 acres
StdDev	47.76
Median	33.50
IQR	36.50
Min	6
Q1	18.50
Q3	55
Max	250

a) Would you describe this distribution as symmetric or skewed? Explain.

b) Are there any outliers? Explain.

c) Create a boxplot of these data.

d) Write a few sentences about the sizes of the vineyards.



25. Caffeine. Should you have a cup of coffee to make you more alert when studying for a big test? A student study of the effects of caffeine asked volunteers to take a memory test 2 hours after drinking soda. Some drank caffeine-free cola, some drank regular cola (with caffeine), and others drank a mixture of the two (getting a half-dose of caffeine). Here are the five-number summaries for each group's scores (number of items recalled correctly) on the memory test:

	n	Min	Q1	Median	Q3	Max
No caffeine	15	16	20	21	24	26
Low caffeine	15	16	18	21	24	27
High caffeine	15	12	17	19	22	24

a) Describe the W's for these data: Who, What, When, Where, Why, How.

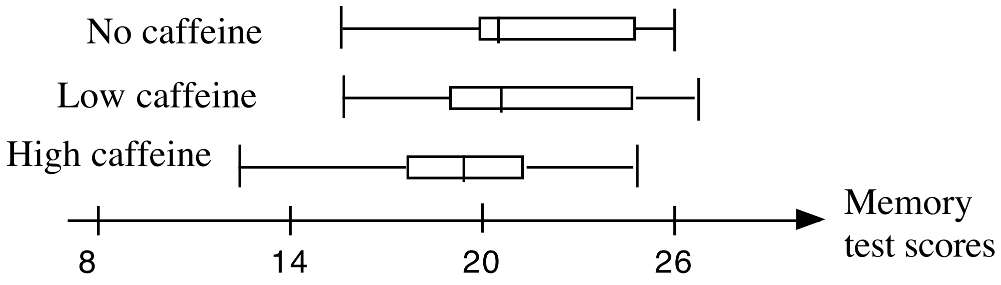
b) Name the variables and classify each as categorical or quantitative.

c) Create parallel boxplots to display these results as best you can with this information.

d) Write a few sentences comparing the performances of the three groups.

(c) To create boxplots you need the 5-number summaries, which are given.

Calculate fences of boxplot to check for outliers: Q1 - 1.5*IQR & Q3 + 1.5*IQR
No caff: IQR = 4; lower fence = 20-1.5*4 = 14; upper fence = 24+1.5*4 = 30
Low caff: IQR = 6; lower fence = 18-1.5*6 = 9; upper fence = 24+1.5*6 = 33
Hi caff: IQR = 5; lower fence = 17-1.5*5 = 9.5; upper fence = 22+1.5*5 = 29.5



Some conceptual questions on Ch.3-4

(1) Describe the Inter Quartile Range (IQR) without using (or even thinking about!) equations.

(2) Suppose the only thing you know about a distribution is the value of its mean and median. Is it possible to tell the direction of skew? Explain.

(3) Suppose the only thing you know about a distribution is the 5-number summary. Is it possible to tell the direction of skew? Explain.

(4) Is it possible for the “box” part of a boxplot to flatten down to a single value? Explain. [Hint: Think about what this would mean in terms of the 5-number summary.]

Some useful recipes from Ch.3-4

How to make a boxplot

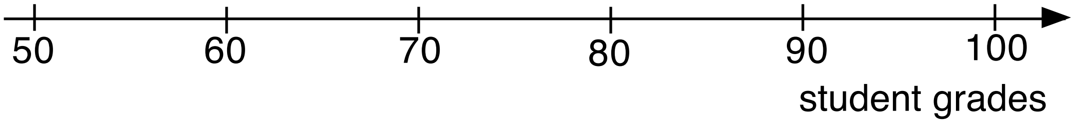
* Use the following dataset for illustration (student grades in Statistics class):

86, 91, 80, 74, 70, 100, 97, 92, 56, 87, 85, 95, 95, 98, 78.

Step1: Compute the 5-number summary (with calculator or by hand).

[E.g., min=56, Q1=78, med=87, Q3=95, max=100]

Step2: Sketch vertical (or horizontal) axis for boxplot. Pick reasonable scale, based on range, and label the axis [E.g., For range=44, let's use scale=10.]



Step3: Compute the invisible fences: lower=Q1−1.5*IQR, upper=Q3+1.5*IQR

[E.g., IQR=95−78=17. So, 1.5*IQR=25.5. lower=52.5, upper=120.5]

Step4: Sketch boxplot keeping in mind the following:

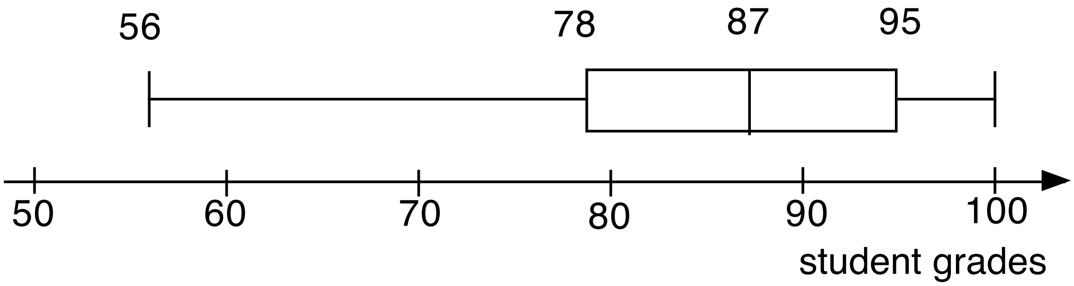
box_starts_at=Q1, box_middle_is_at=median, box_ends_at=Q3.

lower whisker goes to smallest data value inside the invisible fence.

upper whisker goes to largest data value inside the invisible fence .

When there are no outliers: lower whisker starts at min; upper ends at max.

[E.g., see below.]



How to describe a distribution

* Description is expected to be a short paragraph (or two), written in complete sentences, with correct grammar.

* The key objective of the description is to discuss shape, center, spread & unusual features (such as outliers). Be sure to address each of these items.

* For shape: mention how many modes, and what type of skew.
For center: give the value of the median and/or mean.
For spread: give the value of the IQR or the SD.
For outliers: mention how many, what type (low/high), approximate locations.

* Use units if they're given, and try to tie your discussion into the application context.