Basic definition of variable:

A variable is any measurable attribute that can change from observation to observation.

For example, when you interview people about their favorite movie genre, you can also collect their personal information. You can ask them, for example, the following.

- Favorite genre
- Age
- City of residence
- Income
- Number of movies seen this year
- Marital status

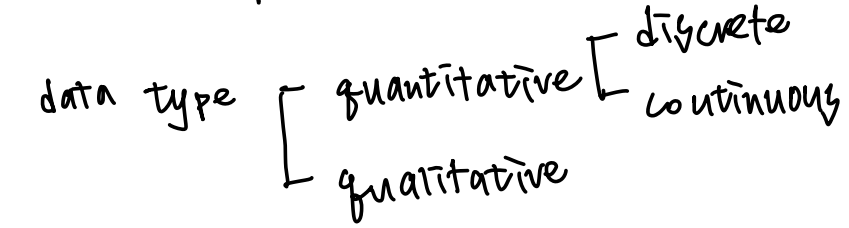
and the above would vary from person to person, thus they can be seen as variables.

Each of the variables can be classified as either qualitative (categorical) or quantitative (numerical), identifying what the data type is important when it comes to graphical summaries and statistical tests. Quantitative variables can also be classified into discrete or continuous.

- Discrete quantitative variables take only specific values. Usually going to be count data.

- Continuous quantitative variables take any value in a range.

- A mind map



<Hint: think about possible values of the variables>

- Favorite movie genre:
  possible values: thriller, romance, action, animation, ... → qualitative
- City of residence → qualitative
- Height → quantitative, continuous
- Number of movies seen last year → quantitative, discrete

▶ Different graphs you would learn:
Dot plot, stem leaf plot, histogram, bar chart, pie chart, Box plot.

How to choose suitable method to summarize your data?
→ Based on data type, sample size mainly.

- Data type
  - quantitative
    - small sample size
      - dot plot
      - stem leaf plot: at least two digits
    - Big sample size
      - Histogram
      - Box plot
  - qualitative
    - Bar chart: bar are not touching.
    - pie chart: can show the proportion/ percentage of each category, but are not good for multiple categories because you cannot see the differences of the sizes

* Dot plot

Steps:

1. Sort your data from smallest to largest
2. Create your horizontal axis and label, write the unit.
3. Plot each data, if repetitive value, stack them.

Example problem 1:

A federal government study of the oil reserves in Elk Hills, CA, included data on the amount of iron present in the oil. As follows:

20, 14, 22, 12, 34, 20, 22, 13, 12, 36, 25, 14, 17, 27, 17, 20, 29, 34, 36, 46.

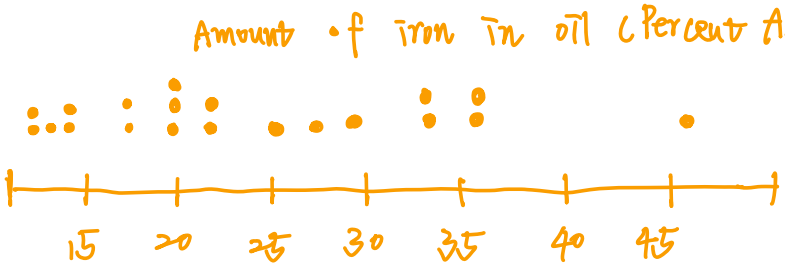
Create dot plot of this data.

Solution:

1. Sort the data:

12, 12, 13, 14, 14, 17, 17, 20, 20, 20, 22, 22, 25, 27, 29, 34, 34, 36, 36, 46

2. create the axis and label.



3. visually inspect the graph, and check if there is anything unusual. In this case, the data is slightly right skewed.

Right skewed: most data is on the left, outlier on the right.
Left skewed: most data is on the right, outlier on the left

- stem-leaf plot
  1. order the data from smallest to largest
  2. Divide numbers into longer stem portions for an ordered list
  3. Add ordered leaf portions
  4. create a key with units

Example problem 2:

A federal government study of the oil reserves in Elk Hills, CA, included data on the amount of iron present in the oil. As follows:

20, 14, 22, 12, 34, 20, 22, 13, 12, 36, 25, 14, 17, 27, 17, 20, 29, 34, 36, 46.

create stem leaf plot.

option1: you only have one stem per tens

1. order the data

12, 12, 13, 14, 14, 17, 17, 20, 20, 20, 22, 22,

12, 27, 29, 34, 34, 36, 36, 46

2. create the stem and create the leaf

1	2 2 3 4 4 7 7
2	0 0 2 2 4 5 7 9
3	4 6 6
4	6

key: 1|2 means 12

3. put the key

option 2: You have two stems per tens.
the first stem includes digits 0, 1, 2, 3, 4
the second stem includes digits 5, 6, 7, 8, 9

Key: 1|2 means 12

1	2 2 3 4 4
1	7 7
2.	0 0 0 2 2 4
2	5 7 9
3	4
3	6 6
4	
4	6

- Histogram

Histograms has horizontal and vertical axes. The area of a block represents a percentage of data locating into that interval.

Steps:

1. take a look at your data, and decide your intervals.
2. record how many data points in each interval, thus calculate the percentages.

3. Based on the percentage, determine the height of each block, which would be the percentage divided by the interval length. (Because you want the

Height of each block representing the density?>

4. make the axes, and draw the rectangles.

Example problem 3:

A concern of forest and conservation managers is the sale of forest lands that are then converted to non-forest uses. 775 quarter sections in a portion of northern Wisconsin were selected, and the number of acres of forest land converted to non-forest use during the period 1990-1999 was measured for each. Below is the distribution table for forest loss:

Forest loss in acres	0-20	20-40	40-80	80-100	100-130	130-150
# of quarter section	311	279	74	58	42	11

create the histogram of forest loss.

key observations

① histogram is for quantitative variable, what is the variable in this question?

→ forest loss in acres for each quarter. Each quarter section is seen as an observation / subject

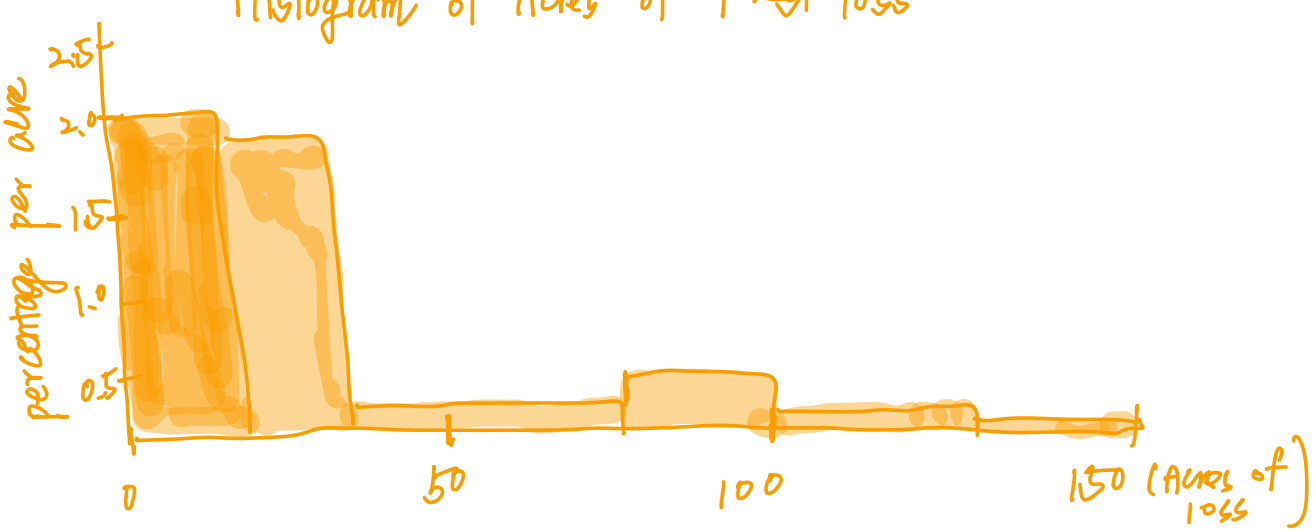
② Notice that for histogram, each interval can have different widths.

Create the histogram

1. first create the distribution table

Interval	Count	Percentage	Interval length	Density = %/meter
[0, 20)	311	$311/775 = 0.401$	20	$40.1/20 = 2.0$
[20, 40)	279	0.36	40	$36/20 = 1.8$
[40, 80)	74	0.095	40	$9.5/40 = 0.24$
[80, 100)	58	0.0748	20	0.374
[100, 130)	42	0.0542	30	0.18
[130, 150)	11	0.014	20	0.07
		area	Base	height

2. create the graph



Extended questions for Histogram

- for this data, approximate the percentage of quarter sections that experienced forest loss of less than 80 acres.

  :

  less than 80 acres is the intervals of $[0,20)+[20,40)+[40,80)$
  thus. $(311+279+74)/775 = 664/775 = 85.7\%$

- for this data, estimate the percentage of quarter sections with forest loss between 20 and 50 acres

  this is an important sample question. Notice that 20 is a boundary point but 50 is not. When it comes to histogram, we assume even distribution in an interval, which might not necessarily true.

  $[20,50] = [20,40) + (40,50)$.

  and we have 279 quarter sections in $[20,40)$
  $[40,50)$ is in the interval of $[40,80)$. thus we do not know for sure the amount of data in $[40,50)$
  but we assume the amount of data is proportional to the interval length. Thus, the length of $[40,50)$ is one quarter of $[40,80)$, so we estimate the amount of data from $[40,50)$ is $\frac{1}{4}\cdot 74$ (the amount in $[40,80)$)
  $= 18.5$.

Finally, then the percentage from $[20, 50)$ is
$(279 + 18.5)/775 = 38.4\%$

- Give an estimate for the median number of acres converted to non-forest use.

→ the median is around 25 to 26 acres, the logic is pretty similar to last question that you assume equal distribution in an interval / Bin.

Median is the value whose ranking is 50%.
since there are 775 values, the median happens between the 387 th and 388 th.
the $[0, 20)$ interval gives us 311 values already,
then we still need to find another $387 - 311 = 76$ values. since there are 279 values in $(20, 40]$,
then we're sure the median is in this interval.
$\frac{76}{279} = 0.27$, which is about $(40 - 20) \cdot 0.27 = 5.45$.
amount of data, this is how much we have to go beyond 20.
so $20 + 5.45 = 25.45$. we estimate our median is between 25 and 26.

It is very easy to get confused by bar chart and histogram.

They are obviously different that

→ Histogram is for quantitative data while bar chart is for qualitative.

→ Histogram has touching blocks, because the intervals do connect numerically while bar chart has no touching blocks, because it makes no sense to connect two categories together.

→ the height of histogram is density, while that for bar chart is percentage.

- Bar chart

  steps

  1. Create horizontal axis with equally spaced categories and label them
  2. Plot the bars so that the height represents the percentage of the category.
  3. Add the title.

Example problem 4

Births are not evenly distributed across the days of the week. Here are the average numbers of babies born on each day of the week at a hospital in the United States in a relent year.

Day	Births
Sunday	7374
Monday	11704
Tuesday	13169
Wednesday	13038
Thursday	13013
Friday	12664
Saturday	8459

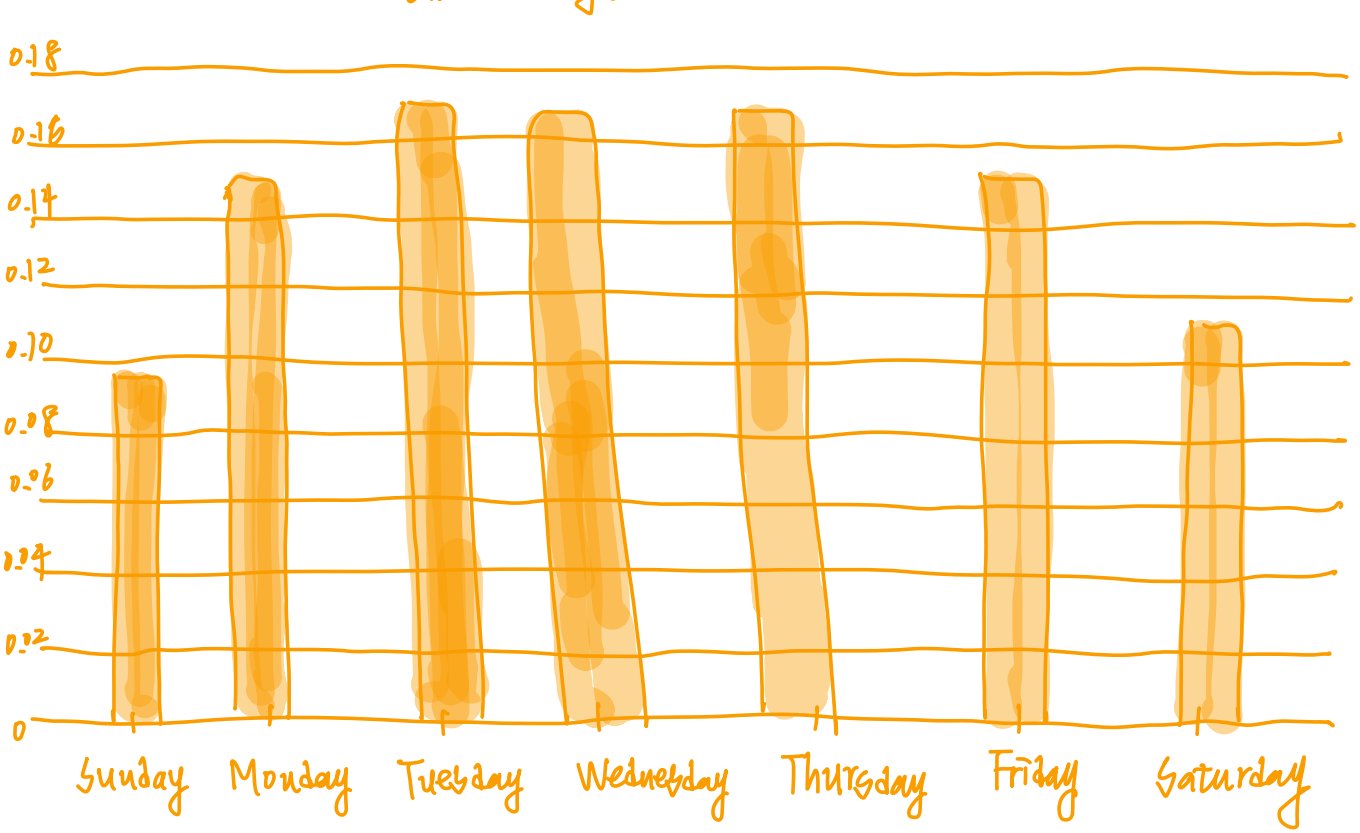
what type of graph is useful to display the data? Create the graph. What patterns do the data show?

Solution:

Bar chart would be useful since we have counts of categorical data.

To create the bar chart, we need to have the distribution table first.

Day	Births percentage
Sunday	7374/79421 = 0.093
Monday	11704/79421 = 0.147
Tuesday	13169/79421 = 0.166
Wednesday	0.164
Thursday	0.164
Friday	0.159
Saturday	0.107



→ possible reason to have lower percentage of births on weekends might be that C sections are not typically scheduled for weekends.

→ Pie chart is applicable since it is used to display categorical data, too. However, it would be difficult to compare values since they are similar.

→ It would be inappropriate to talk about the shape of these data as this is categorical data. Shape doesn't mean anything. You can adjust the order of the bars arbitrarily.

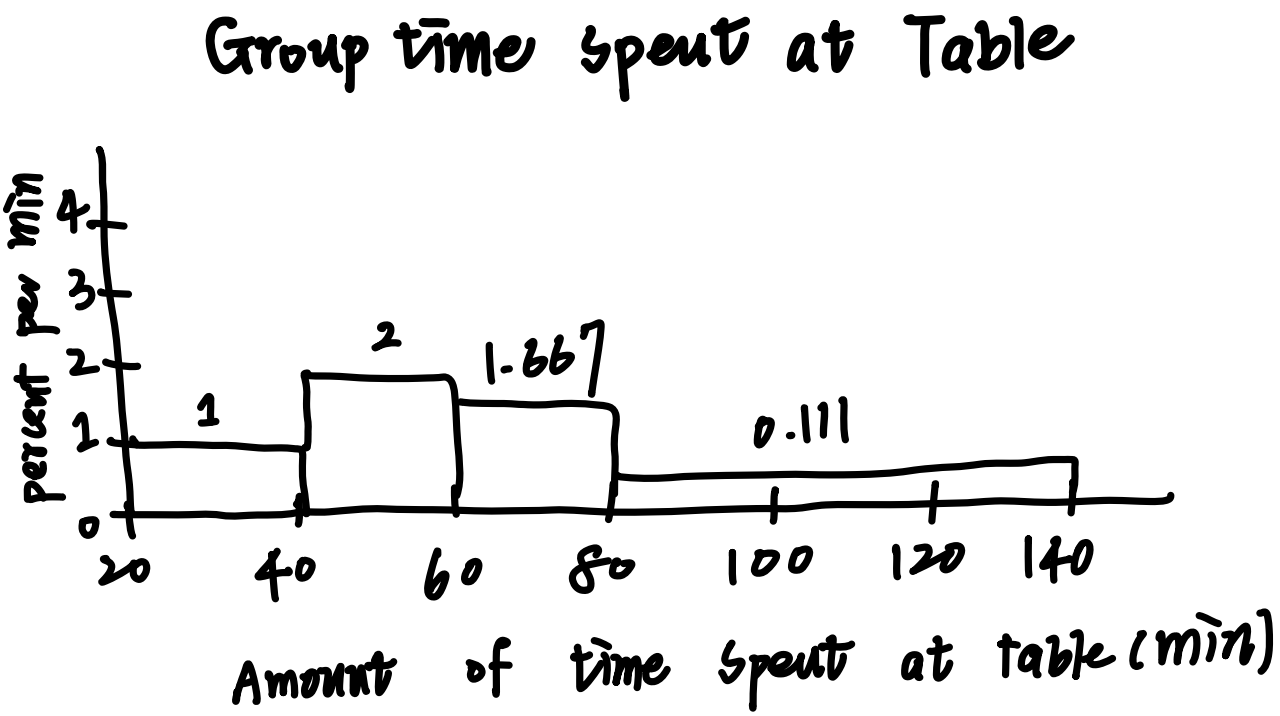
■ Start of a question • Start of a part of a big question

— solution in orange

Supplementary exercises.

■ You are being interviewed for managerial position at a restaurant. Your potential employer has asked you to discuss some aspects of the business as part of your interview.

He presents you with the density histogram of amount of time it takes from all customers in the group to be seated at the table to close out of the check. These are times he collected in the past month whenever a table was occupied



- Based on the above density histogram. what percent of customer groups spent more than 60 minutes at the table?

Solution:

this question mainly tests if you know the meaning of the area, height, and width of the bins.

- width of the bin means the width of the interval.
- height of the bin is the density of data in the interval
- area, which is the width multiplied by the height, of the bin, means the percentage of data contained in the interval.

Therefore.

From 60 - 80: $1.667 * (80-60) = 33.34\%$

From 80 - 120: $0.111 * (140-80) = 6.66\%$

$33.34\% + 6.66\% = 40\%$

the answer is 40%

- Based on the density histogram. what is the approximate amount of time below 50% of tables finished?

Solution:

this question asks us the median.

Firstly, from interval [20, 40), we have 20% already and the interval [40, 60) has 40% of data, so we know

median is somewhere between 40 and 60.

after 40, we still need $50\% - 20\% = 30\%$ of data. Thus, we have to go beyond 40 into $30\%/40\% = \frac{3}{4}$ into the interval of $[40, 60)$.

the length of $[40, 60)$ is 20, so the final answer is $40 + \frac{3}{4} \cdot 20 = 40 + 15 = 55$.

- If the restaurant wants to put information on their website that makes wait time appear shorter, should they report their mean or median?

Solution:

this question tests your understanding of mean, median, and skewness.

Concept 1:

mean is more sensitive to outliers compared to median.

For example:

A. $(1, 2, 3, 4, 5)$: median = 3, mean = 3

B. $(1, 2, 3, 4, 100)$: median = 3, mean = 22

A data and B data have the same value of median, but the means are quite different.

Concept 2:

Right skewed data has outlier on the right, thus the mean would be greater than the median. The B data above is this case.

Left skewed data has outlier on the left, for example, consider

$(-110, 1, 2, 3, 4)$: median $= 3$, mean $= -20$.

then mean is less than the median.

* Back to our density histogram, it is slightly right skewed. thus mean is greater than the median. Thus reporting median would be better.

- An astronomy researcher records several items related to individual stars in distant galaxies. The variables that she measures and records are the following:

  ✓ Distance from the earth (in light years)
  ✓ Number of other stars in that galaxy
  ✓ The type of each star (O, B, A, F, G, K, M)
  ✓ The star's luminosity (in joules per second)
  ✓ The name of the galaxy the star is in

- which of the above variables should be classified as quantitative, continuous?

  Solution:
  Distance from the earth & luminosity

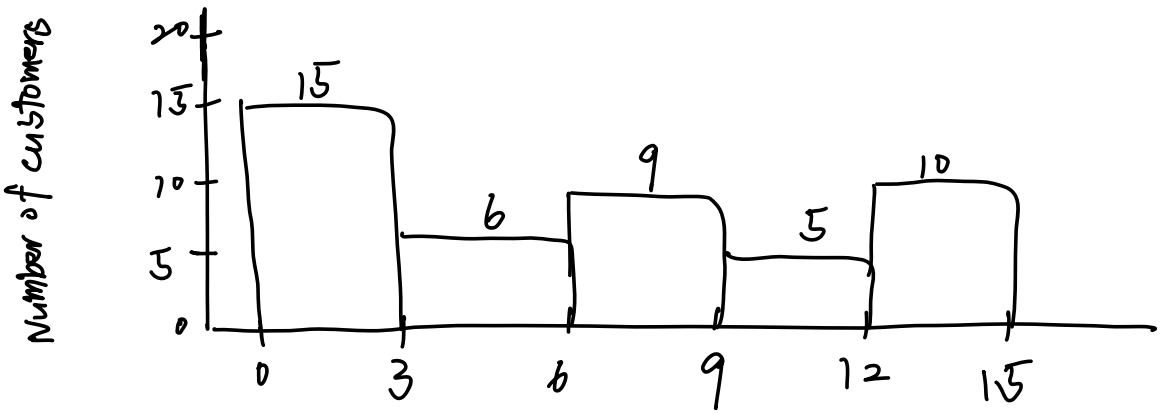
- which of the above variables should be classified as qualitative?

  Solution:
  type of stars & name of galaxy

- Which of these graphical summaries would be best to display data on the make of cars UW-Madison students drive? (eg. Ford, Toyota)
  a. Stem and leaf plot
  b. Bar chart
  c. Dot plot
  d. Car plot
  e. Histogram

Solution: b. Bar chart is the only one for categorical data

■ The histogram below displays the frequency of wait times, in minutes, for 45 customers of Ducky's car wash.



which of the following could be the median of the waiting times, in minutes?

a. 3.7 b. 5.1 c. 6.9 d. 9.1 e. 9.5

Solution: C. 6.9.

the median of 45 customers is the 23rd. which would be between 6 and 9.