MAT 105 - Summary of Lecture 3
Tuesday, January 12, 2016

> *Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.* (H.G. Wells)

1 Measurement levels

Most of the time we keep track and analyze numerical data, but that is not always the case. Here are different data types and their measurement levels:

1. *Nominal level* (in name only, qualitative in nature, describes the data). Examples:
   (a) the colors for Thunderbird cars on my dealer's lot are: red, yellow, blue, white, black.
   (b) the soups in my favorite restaurant are: tomato, lentil, French onion, potato-leek
   (c) Ski resorts in WA: Snoqualmie Pass, Stevens Pass, Crystal Mountain
   (d) Degree Programs at DigiPen: BAGD, BSGD, RTIS, BSESD, BAMSD, BSCE, BSCS, BFA
2. *Ordinal level* (includes **relative** comparisons, there is an order, but differences do not have meaning). Examples:
   (a) Qualification for performance: good, average, bad
   (b) In the class, Joe is 3rd, Nick is 7th and Steve is 11th. (Note that we cannot say that the difference between Nick an Joe is the same as that between Nick and Steve, since the ranking does not measure the ability of a student, only his **relative** position in reference to all others)
   (c) The temperature of the food is: cold, lukewarm, hot
   (d) Grades in the class: A, B, C, D, E, F
3. *Interval level* (still ordered, but difference between data is **meaningful** and ratios are **not** meaningful). Examples:
   (a) Years in which Democrats won presidential elections (with no reference to year 0)
   (b) Temperature on January 1st for the past 50 years (eg: today is 10 degrees warmer than last year, ie, the difference in data has meaning)
   (c) The scores in the MAT 105 course (if John has 93.2 and Jason has 89.1, then John has 4.1 more points than Jason)
   (d) Time of a students' first class (my MAT 580 is 3 hours before my MAT 105)
4. *Ratio level* (interval level + meaningful ratios, there is a zero used as a starting point). Examples:
   (a) Time from beginning to finish of a race
   (b) Length of salmon in Sammamish River
   (c) Time from deposit of check until it clears
   (d) Temperature of an object, measured in Kelvins, where zero Kelvins means NO heat.

Class Example 1. Measurement/ Data Types: We collect the following data from a robotics company. For each type, decide its level of measurement.

- Salesperson performance: below average, average or above average. → Ordinal
- Price of company's stock. → Ratio
- Names of new products. → Nominal
- Room temperature in CEO's office, in °F. → Interval
- Gross income for each of the past 5 years. → Ratio
- Color of packaging. → Nominal
- Room temperature in CEO's office: cold, average, hot. → Ordinal

2 Organizing Data

One of the first people to use graphical representations of statistical data to prove her point was Florence Nightingale (1820-1910), a nurse and a statistician, well known for improving military hospitals during the Crimean War. She used a special form of pie charts (Nightingale rose diagram) to illustrate how improved sanitation decreases mortality. In fact, due to her recommendation for improved sanitation practices, mortality rates in military hospitals saw a decrease from 42.7% to about 3%.
Here are a few of the most common data representation types and some examples:

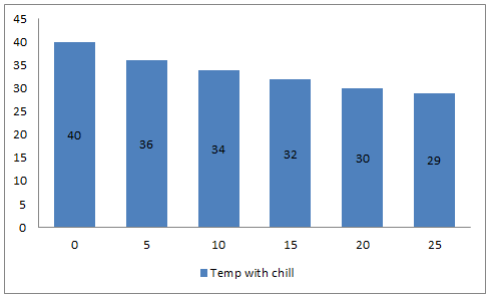
(a) bar graphs (c) pie charts (e) graphs (g) stem-leaf displays

(b) Pareto charts (d) histograms (f) time charts (h) tree diagrams

Class Example 2: Data from National Weather Service on wind chill at 40°F:

Temp with chill (°F)	40	36	34	32	30	29
Wind (mi/h)	0	5	10	15	20	25

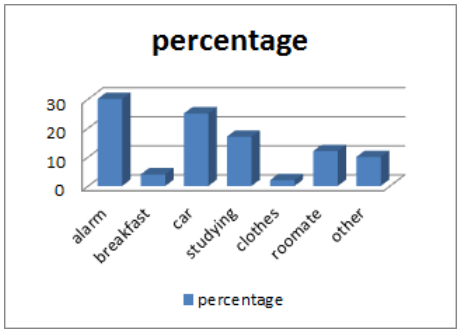
A **bar graph** displays the data, using bars with uniform width and uniform spacing between bars:



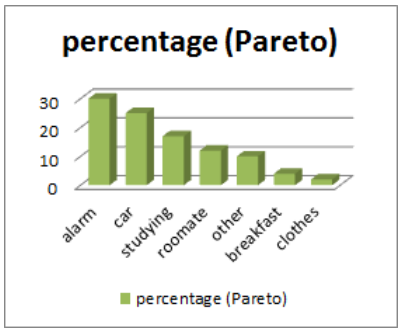
Class Example 3: Students plan to arrive to class 15 minutes early, to account for unforseen delays. When they arrive late, they cite the following as reason for the delay, given in percentages of those surveyed:

Reason	Alarm	Car trouble	Late breakfast	Studying	Clothes	The roommate	Other
%	30	25	4	17	2	12	10

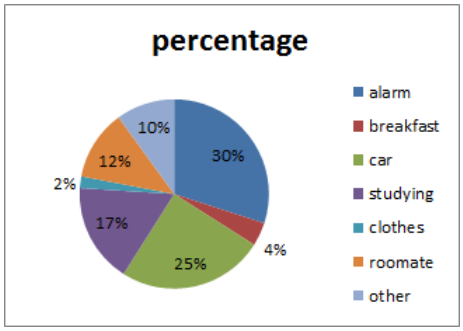
Again, here is a **bar chart**:



A **Pareto chart** is a bar graph in which heights represent frequency (or percentage), with bars arranged according to height, from tallest to smallest.



A **pie chart** is a circular pie divided into parts according to given percentages. For example, the slice corresponding to students who miss the alarm has a central angle of 30% of 360°, which equals 108°. The slice corresponding to students who blame it on the car has a central angle of 25% of 360°, which equals 90° and we continue this way until we fill out the pie. Note that the angles we obtain should add up to 360°.



Class Example 4: Weight of carry on luggage in pounds in a sample of 40:

30	27	12	40	35	17	38	36	27	35
22	17	29	3	21	0	39	15	40	33
26	36	18	0	18	32	31	32	19	21
33	31	28	29	26	12	32	18	21	26

A **histogram** is a representation of frequencies. It has the following components:

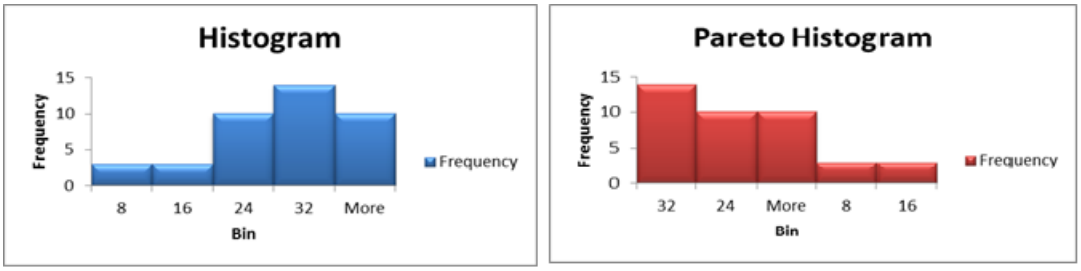
- a number of classes, also called bins or ranges, typically between 5 and 15 of them, represented by bars
- bar width = class width = $\dfrac{\text{largest element} - \text{smalles element}}{\text{number of classes}}$, rounded up to the nearest integer.
- usually, the bars touch each other, there is no spacing between them

We organize the data using a histogram with 5 classes first, followed by a histogram with 8 classes.

(a) Using 5 classes (bins). The class width is $\dfrac{40-0}{5} = 8$. The first class is between 0 and 8, with 8 included – there are 3 pieces of luggage in this range. The second class is between 8 and 16 (8 excluded, 16 included) which has 3 elements. The third bin is between 16 and 24, with 10 elements. The fourth bin is between 24 and 32, with 14 elements and the last bin is for elements larger than 32, with 10 elements. We collect this data:

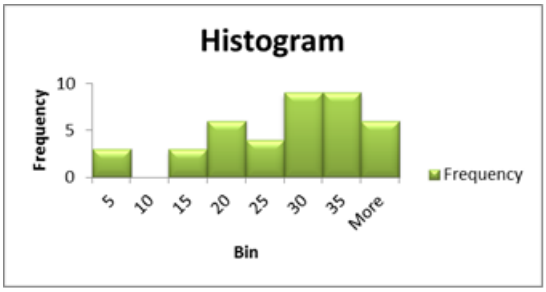
Bin	Frequency
8	3
16	3
24	10
32	14
more	10

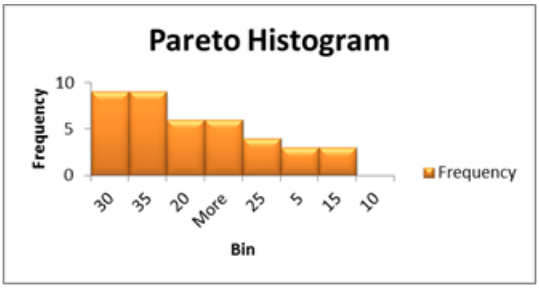
Then the histogram for 5 classes for this data is displayed below, along with a Pareto histogram (ordered bins from largest to smallest.)



(b) Using 8 classes (bins). We take a similar approach to complete a histogram with 8 classes, as well as a Pareto histogram.

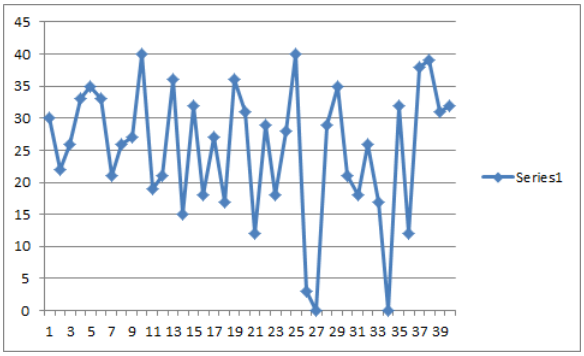
Bin	5	10	15	20	25	30	35	more
Frequency	3	0	3	6	4	9	9	6



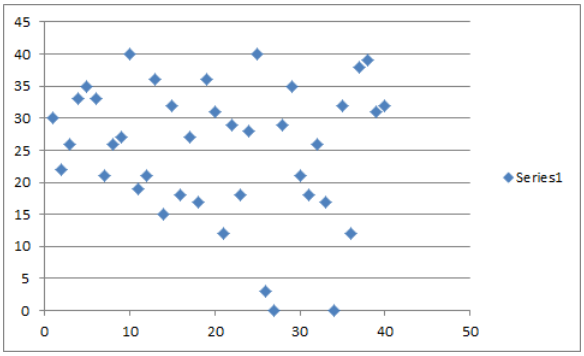


Note that once the histogram is done, the Pareto histogram is very easy to find, and it does not add much additional value to our analysis of the data.

Suppose that carry-on luggage arrives at equally spaced intervals. Then one can talk about a time plot, where the x-axis describes the time and the y-axis is the weight of the luggage.



Similarly, we can leave the data points on the plot without connecting them, in which case we get a scatter plot:



Note that the last two representation charts do not lose information, but they do not encode the frequency. A representation that both addresses the frequency, and it does not lose any data (like a histogram might) is the stem and leaf display. To make a stem and leaf display, we break the digits of each data value into two parts: a stem and a leaf. We are free to choose the number of digits to be included in the stem! In this kind of display, we list each stem ONCE on the left and all its leaves in the same row, to the right of the stem.

The weights in our carry-on example consist of two-digit numbers, so it is natural to choose the stem to be digit for the tens and the leaf to be the digit for the ones. We represent a carry on of weight ab lbs. by $a|b$. For example, $1|2 = 12$ lbs, $0|3 = 3$ lbs. etc. Sometimes, the leaves will be ordered from smallest to largest in the display, but that will not be necessary. Here are both an unordered and the ordered display for our example.

Stem and leaf (unordered) **key: 1|7=17**

0	3	0	0											
1	2	7	7	5	8	8	9	2	8					
2	7	7	2	9	1	6	1	8	9	6	1	6		
3	0	5	8	6	5	9	3	6	2	1	2	3	1	2
4	0	0												

Stem and leaf (ordered) **key: 1|7=17**

0	0	0	3											
1	2	2	5	7	7	8	8	8	9					
2	1	1	1	2	6	6	6	7	7	8	9	9		
3	0	1	1	2	2	2	3	3	5	5	6	6	8	9
4	0	0												

Note that the display looks like a rotated histogram. In particular, one can read off the display which range occurs with most and least frequency.

3 Averages and variation

The next step is to analyze data. Now that we collected the data, represented it, what can you infer from it? There are two quantities we would like to consider: "average" and "spread". The average will reflect the central tendency of data. However, average can mean many things: most common (mode), the midpoint mark (median), the average of all data points (mean)? In terms of spread, we will consider the range of data as well as its deviation from the mean (variance and standard deviation.)

3.1 Averages

Def: The mode is the value or property that occurs most often (frequently) in the data. It is the easiest to compute.

Example: To illustrate this idea, we find the mode of the data represented by the lengths of all words in the definition above. We make a tally of the lengths:

Length	2	3	4	5	6	7	8	10
Frequency	6	4	4	2	1	2	1	1

The mode is 2, as words of length 2 occur most often in this text.

While the mode is the most common data point in the set, it is not very stable, as small variations in the data can shift the mode by large amounts. The next average type is more stable.

Def: The median is the middle value in the ordered list of data points (from smallest to largest).

If there is an odd number of data points, the median is easy to find. If there is an even number of data points, the median is found by taking the arithmetic mean (average) of the data points immediately to the left and the right of the midpoint in the list.

Example: Consider the following sets of data.

(1). 12 14 18 22 30 63 75

(2). 12 14 18 22 30 63 75 80

(3). 0 0 2 22 90 172 390

For examples (1) and (3), the median is 22, for (2) the median is $\frac{22+30}{2} = 26$. Note that even though the medians for examples (1) and (3) are the same, the data sets are quite different, in particular the range in (3) is much larger than in (1). The next measure of central tendency is what we refer to as average in everyday life.

Def: The mean is the arithmetic mean of all the values in the data set. That is, if the data points are $n_1, n_2, \cdots, n_k$, then

$$\mu = \text{mean} = \frac{n_1 + n_2 + \cdots + n_k}{k}.$$

When the mean is derived from a *sample*, we denote it by $\overline{x}$, if it is the mean of *all population*, we denote it by μ.

Example: To find the means in examples (1)–(3) above, we have

(1). $\mu = \frac{234}{7} = 33.42$

(2). $\mu = \frac{314}{8} = 39.25$

(3). $\mu = \frac{676}{7} = 96.57$

Another example: Suppose there are 4 tests in this class, equally valued and on the first 3 tests you scored 84, 90, 88. What score do you need on the fourth exam in order to get an A- in the class (to *average* at least a 90)? Since the average must be at east a 90, the sum of the 4 scores must be at least 360, so the score on the fourth exam should be at least $360 - (84 + 90 + 88) = 98$.

While the mean takes into account the range, it is less *resistant* than the median. Here a *resistant measure* is one which is not influenced by extremely high or low data (outliers). To account for the outliers, one may consider the following modification of the mean.

Def: The trimmed mean is the mean of the data set obtained by removing the smallest 5% and the largest 5% of the data. If 5% of the number of data points is not an integer, we round it to the *nearest* integer.

Example H: Here is the time spent on homework on the first week by 20 students in MAT 105:

0	.5	.8	1.2	1.5	1.5	1.5	1.5	1.5	2
2	2	2	2.1	2.2	2.2	2.4	3	4.2	9

The mode of this data is 1.5, the median is 2, the mean is 2.155 and to find the trimmed mean, we remove $5\% * 20 = 1$ data point from the beginning and the end of the list, to get a trimmed mean of 1.895.

Remarks:

(a) You can find these averages at the following measurement levels:

- mode – at any measurement level
- median – in ordinal level data or above
- mean – in interval level data and above.

(b) Here are some examples as to when these averages are used

- mode may be used in deciding which sizes of clothing to re-stock
- median may be used to measure the average salary in a company
- mean may be used for the average amount of rainfall per day in a given year.

Most of the time, knowing averages is not enough to describe the data. One needs a measure of the spread of data points too.

3.2 Spread

Def: The range of a data set is the difference between the largest data point and the smallest data point.
In **example H** above, the range is $9 - 0 = 9$.
To measure the spread between the data points and the mean, basically measuring the frequency of data points within given distance from the mean, we use the variance and the standard deviation.
Def: The variance of a population is

$$\sigma^2 = \frac{\sum_{k=1}^{n}(x_k - \mu)^2}{n} = \frac{(x_1 - \mu)^2 + \cdots + (x_n - \mu)^2}{n}$$

where μ is the mean of the population, n is the number of data points, with data points $x_1, x_2, \ldots, x_n$.
If the data refers to a sample set, not the entire population, we use the notation $\overline{x}$ instead of μ and s^2 for variation instead on σ^2. In fact, as we will see later in the course, there is a correction in the denominator term as well, but for now you can use n for both sample data set and population data set.
In **example H**, the variation of the data is

$$\sigma^2 = \frac{(0 - 2.045)^2 + (0.5 - 2.045)^2 + \cdots + (4.2 - 2.045)^2 + (9 - 2.045)^2}{20} \approx 3.2.$$

We define the standard deviation of this data set to be the square root of variance, that is, σ (or s when working with samples). The following theorem gives a good estimate on the spread of the data for ANY data set.

3.3 Chebyshev's Theorem

Theorem. For any set of data, the proportion of data that must lie within k standard deviations on either side of the mean is *at least* $1 - \frac{1}{k^2}$, for any $k > 1$. In particular,

- at least 75% of data lies within two standard deviations of the mean, that is, in the interval $[\mu - 2\sigma, \mu + 2\sigma]$
- at least 88.9% of data lies within three standard deviations of the mean, that is, in the interval $[\mu - 3\sigma, \mu + 3\sigma]$
- at least 93.8% of data lies within four standard deviations of the mean, that is, in the interval $[\mu - 4\sigma, \mu + 4\sigma]$

See Problem 10 on Homework 1 for an application of this theorem.