Topic 3.Measurement REsults

3. Graphical Representation of Experimental Data

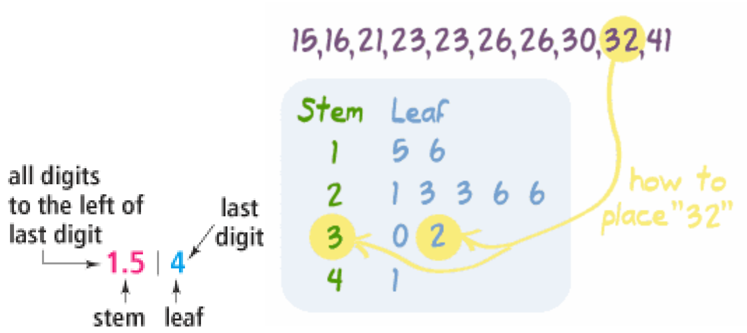
Using defined values for a sample or population, we only acquire some rough data about the nature of the measurement process. This situation is problematic for correct reasoning or decision making for example questions like “which one is the best value to represent sample mode, mean or median?” cannot be properly responded. Therefore, the graphs, which show how frequent the measured values occurred in sample, are used as a powerful tool for statistical analysis.

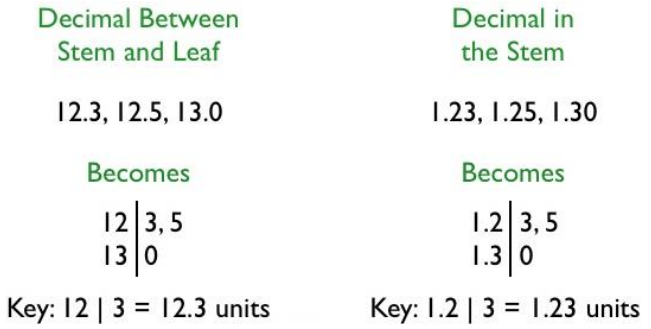
The distribution of a variable describes the values the variable takes and how often it takes each value
For qualitative results, there are two ways to describe data graphically

- Stem-and-leaf plots: a. Displays actual values of all observations and Good for small amounts of data
- Histograms: a. Displays only summary information and Used for large amounts of data

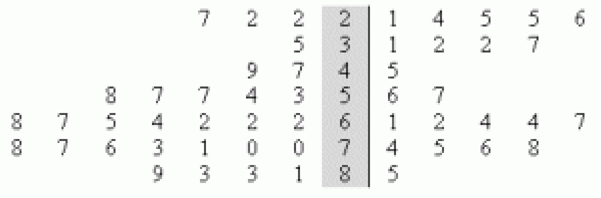
A. Stem-and-leaf plots

A stemplot is a tool to help you visualize a data set. Stemplots show a little more quantitative information than a histogram but they are typically used when there is a medium amount of quantitative variables to analyze; Stemplots of more than 50 observations are unusual. A stem and leaf plot is a way to plot data where the data is split into stems (the largest digit) and leaves (the smallest digits).





One feature of the stem plot is that by keeping stem constant, two distributions can be represented in one plot called back to back stemplot. This specific configuration allows a comparison of two distributions visually.



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

<u>How to plot Stem and Leaf plots</u>

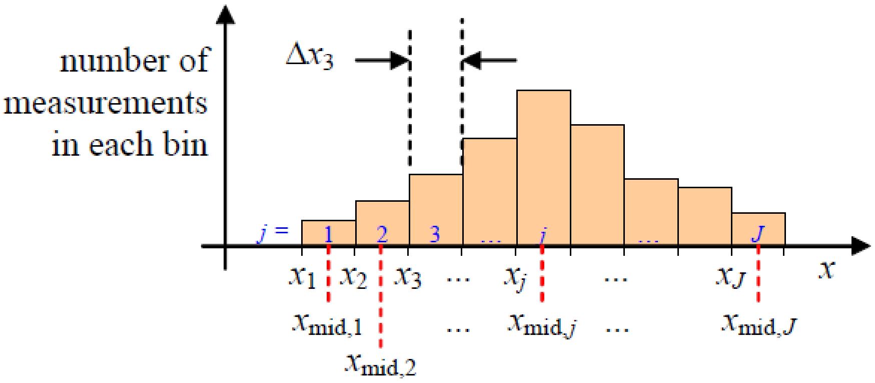
1. Put data in numerical order from smallest to largest.
2. Separate each observation into a “stem” and a “leaf” (Note:leaf = final digit, stem =remaining digits. In the case of a single digit number like 7, use 07 so that 0 is the stem.)
3. Write stem numbers in order in a vertical column with the smallest at the top. Do not skip numbers. Draw a vertical line next to the stem.
4. Write each leaf in the row to the right of its stem, in increasing order out from the stem.
5. Show each numerical value as many times as it appears in the dataset.
6. It is possible to trim any digits that you feel may be unnecessary. You can document that the numbers shown are “thousands”, for example, if you trim off the last three digits.

Topic 3.Measurement REsults

B.Histograms

Histograms are bar graphs and height of each bar shows number of the individuals that has a value within a particular class. This particular class is define with a sub-range in distribution and named as bin and showed on x axis.The y axis " the height of the bin" shows the frequency of the indivuals in bins.

A histogram is constructed by dividing up the n measurements of a sample into J bins or intervals (also called classes) such that forthe first bin (j = 1), $x_1 < x < x_2$, thesecond bin (j = 2), $x_2 < x < x_3$, etc.
$x_{\text{mid},j}$ is the middle value of x inbin j. For example,$x_{\text{mid},2} = (x_2 + x_3)/2$.Afterwards, a bar plot is made of the frequency(also called the class frequency) which is the number of measurements in each bin. If J is too small, your may not get the correct histogram shape and If J is too big, your histogram will look "choppy" – it will have bins with zero data points.


number of measurements in each bin
Δx_3
j = 1 2 3 … j … J
x_1 x_2 x_3 … x_j … x_J x
$x_{\text{mid},1}$ … $x_{\text{mid},j}$ … $x_{\text{mid},J}$
$x_{\text{mid},2}$

Although three is no definite formula for determining number of bins, there exist some methods or formulas for giving an idea of possible number of bins.
For samples with large data, the following two formulations may be used for initial guess

Sturgis rule: $J = 1 + 3.3\log_{10} n$

Rice rule: $J = 2n^{1/3}$

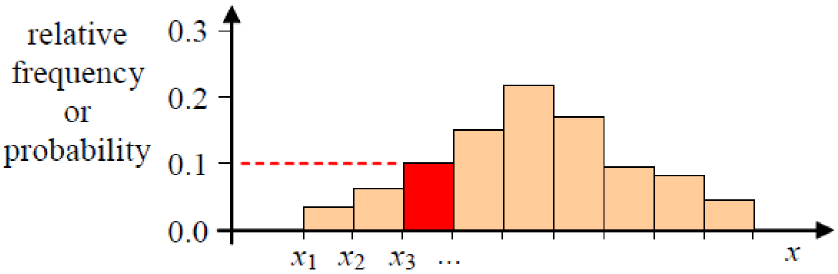
where*n* is the total number of measurements in the sample, and *J* is thenumber of bins or intervals in the histogram.In either case, it is unlikely that an integer number of bins will result. In that case, round off to thenearest integer, since we cannot have a non-integer number of bins.
There is also some basic rules that helps and improves the form of the histogram.

- Bins should be all the same size. For example, groups of ten or a hundred.
- Bins should include *all* of the data, even outliers. If your outliers fall way outside of your other data, consider lumping them in with your first or last bin. Boundaries for bins should land at whole numbers whenever possible (this makes the chart easier to read).
- Choose between 5 and 20 bins. The larger the data set, the more likely you'll want a large number of bins. For example, a set of 12 data pieces might warrant 5 bins but a set of 1000 numbers will probably be more useful with 20 bins.
- If at all possible, try to make your data set evenly divisible by the number of bins. For example, if you have 10 pieces of data, work with 5 bins instead of 6 or 7.

Topic 3.Measurement REsults

B.Probability histograms and normalized histograms:

$$probability_j = \frac{\text{number of measurements in range } x_j \le x < x_{j+1}}{n}$$

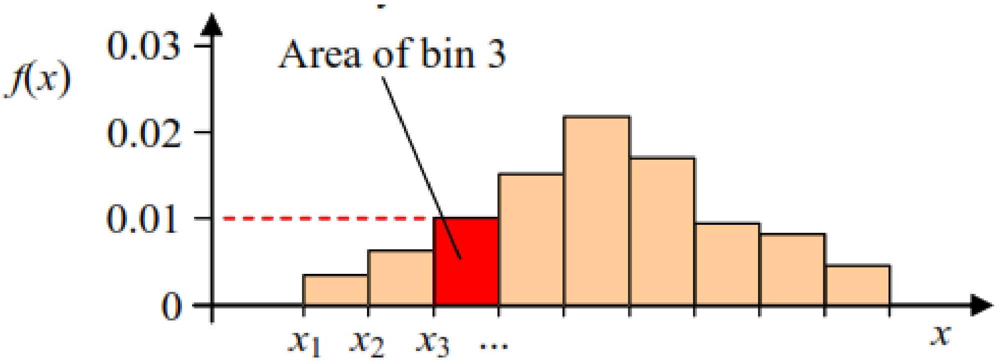


In this histogram, the probability that$x$ lies between x_3 and x_4 is about 0.1 or10%, as indicated by the red bar for j= 3, i.e., the third one.

We define a ***vertically normalizedhistogram*** by further dividing thevertical axis by the bin width orinterval width. The vertical axis of thevertically normalized histogram isdefined as

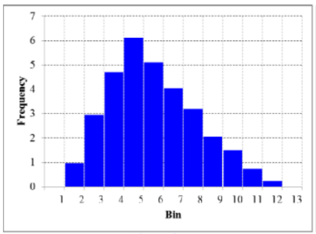
$$\frac{probability_j}{\Delta x} = \frac{\text{number of measurements in range } x_j \le x < x_{j+1}}{n\Delta x}$$

This guarantees mathematically that the area of the bin is equal to the probability that xlies in that bin, as sketched below. In this example,□x = 10 = constant; therefore, the values on thevertical axis decrease by a factor of 10.

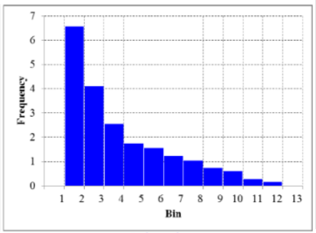


In many cases, the histogram is skewed to one side or may have more than one peak (e.g., exam grades often produce a strange-looking histogram. There are five standard histogram shapes that have been given standardized names:

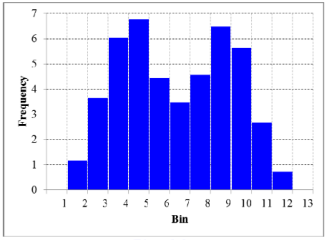
1. Symmetric: Nearly symmetric left to right, with a peak very close to the middle, like the ones above.
2. Skewed: Shifted to one side or the other, with a peak clearly located on one “preferred” side.
3. J-Shaped: Skewed very much to one side, with the peak at or near the lowest or highest bin.
4. Bimodal: Two distinct peaks, and usually fairly symmetric in the vicinity of either peak.
5. Uniform: Nearly the same frequency for each bin (no distinct peaks; fairly flat over whole range).



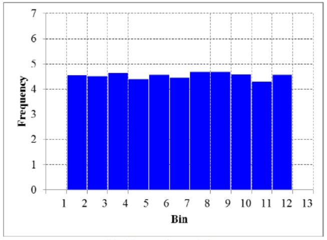
Skewed



J-shaped



Bimodal

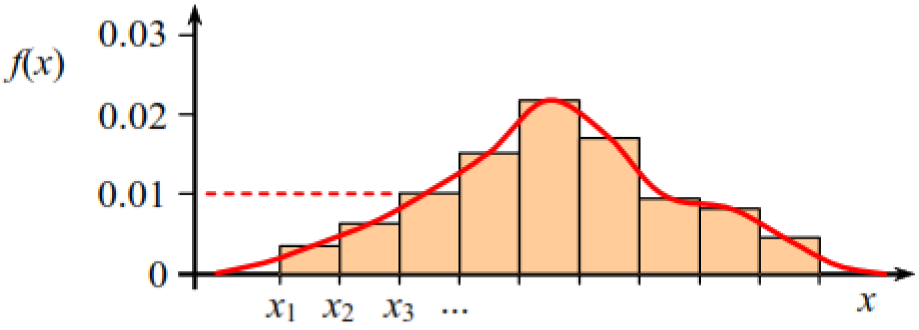


Uniform (nearly)

Topic 3.Measurement REsults

C. Probability Density Functions

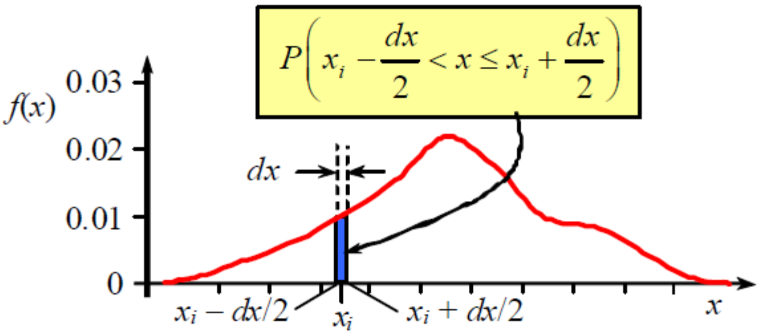
In simple terms, a probability density function (PDF) is constructed by drawing a smooth curve fit through the vertically normalized histogram assketched. You can think of a PDF as the smooth limit of a vertically normalized histogram if there were millions of measurements and a huge number of bins.



The main difference between ahistogram and a PDF is that ahistogram involves discrete data(individual bins or classes), whereas a PDF involves continuous data (a smooth curve), which is mathematically defined as follows:

$$f(x_i) = \frac{P(x_i - \frac{dx}{2} < x < x_i - \frac{dx}{2})}{dx}$$

Function P(.) represents the probability that variable x lies in the given range, and f(x) is the probability density function (PDF).
For the given infinitesimal range of width dx between xi – dx/2 and xi + dx/2, the integral under the PDF curve is the probability that a measurement lies within that range, as sketched.



As shown in the sketch, this probability is equal to the area (shaded blue region) under the f(x) curve – i.e., the integral under the PDF over the specified infinitesimal range of width dx.
The usefulness of the PDF is as follows: Suppose we choose a range of variable x, say between a and b. The probability that a measurement lies between a and b is simply the integral under the PDF curve between a and b, as sketched, where we define the probability as:

$$P(a < x < b) = \int_a^b F(x)dx$$

A quick inference from this formulation is that the probability of occurrences of all events in sample can be obtained by letting integral limits to go infinity. In this case, following equality is obtained:

$$P(\forall x) = \int_{-\infty}^{+\infty} F(x)dx$$

Expected Value (or mean of population)

The probability density function is continuous function. When the probability density function f(x) is defined, we leave the system of discrete random variables and enter the system of continuous random variables. It can also be thought to have infinite number of bins, so the infinite experiment.
For continuous random variables, some more formal definitions can be made:

Expected value is defined in terms of the probability density function as the mean of all possible x values in the continuous system.

$$\mu = E(x) = \int_{-\infty}^{+\infty} xF(x)dx$$

In an ideal situation in which f(x) exactly represents the population, μ is the mean of the entire population of x values, and that is why it is called the "expected" value. It is therefore also called the population mean.
In general, x ≠ μ, but x→ μ when n is large, i.e., the sample mean approaches the expected value when n is large. x and μ are often used interchangeably, but this should be done only if n is large.
Standard deviation: In an ideal situation in which f(x) exactly represents the population, σ is the standard deviation of the entire population. It is therefore also called the population standard deviation.

$$\sigma = \sqrt{\int_{-\infty}^{+\infty} (x-\mu)^2 F(x)dx}$$

If n is large, S → σ. Often, S and σ are used interchangeably, but this should be done only if n is large.

Normalized probability density function

A normalized probability density function is constructed by transforming both the abscissa (horizontal axis) and ordinate (vertical axis) of the PDF plot as follows:

$$z = \frac{x-\mu}{\sigma}$$

$$f(z) = \sigma f(x)$$

The above transformations accomplish two things:

The first transformation normalizes the abscissa such that the PDF is centered around z = 0.
The second transformation normalizes the ordinate such that the PDF is spread out in similar fashion regardless of the value of standard deviation. When normalized in this way, the normalized PDF can be directly compared to other PDFs,

Rules for Normalized PDF Calculation

1. Generate a frequency table. The table contains two columns, bin and frequency.
2. Transform table to histogram, Bin is the mid value of the range of each bin, and frequency is the number of data points in that bin range.
3. Calculate probability of each bin; in which you divide each frequency by the total number of data points.
4. Vertically normalize histogram; in which you divide each probability by the appropriate bin width to find f(x).
5. A smoothed plot of f(x) versus x is the PDF.
6. Apply z transform to find normalized histogram