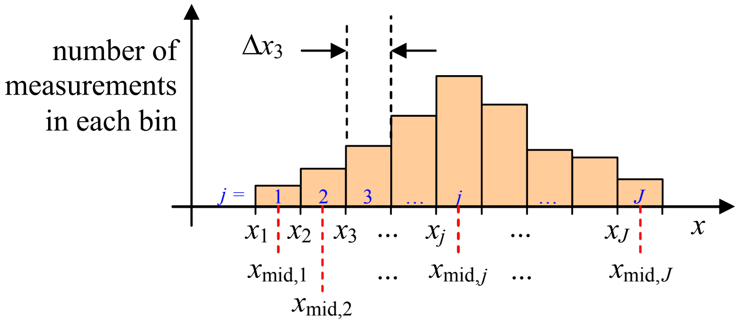
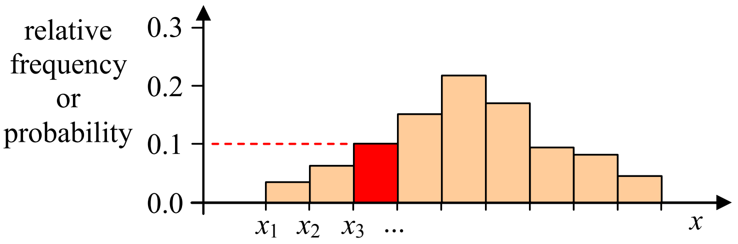
Histograms

Author: John M. Cimbala, Penn State University
Latest revision: 28 August 2014

Histograms

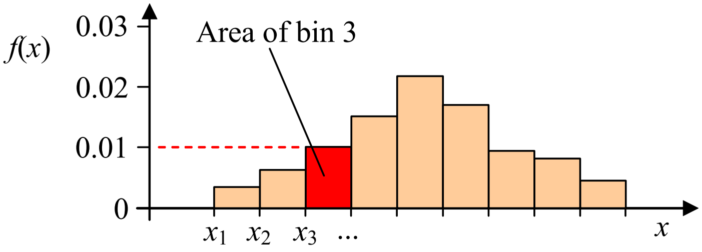

number of measurements in each bin
Δx_3
j = 1 2 3 … j … J
x_1 x_2 x_3 … x_j … x_J x
$x_{\text{mid},1}$ … $x_{\text{mid},j}$ … $x_{\text{mid},J}$
$x_{\text{mid},2}$

- ***Histogram*** – a ***histogram*** is constructed by divvying up the n measurements of a sample into J ***bins*** or intervals (also called ***classes***) such that for the first bin (j = 1), $x_1 < x \le x_2$, for the second bin (j = 2), $x_2 < x \le x_3$, etc.
- We define $x_{\text{mid},j}$ as the middle value of x in bin j. For example, $x_{\text{mid},2} = (x_2 + x_3)/2$. Then a bar plot is made of the ***frequency*** (also called the ***class frequency***) – the number of measurements in each bin versus the value of x, as sketched.
- The ***bin width*** (also called the ***interval width*** or ***class width***) is usually constant, although it does not have to be. In the sketch above, the bin width Δx_3 for the third bin (j = 3) is shown.
- There are some rules of thumb for choosing how many bins to use in a histogram:
  - ***Sturgis rule***: $\boxed{J = 1 + 3.3\log_{10} n}$, where n is the total number of measurements in the sample, and J is the number of bins or intervals in the histogram.
  - ***Rice rule***: $\boxed{J = 2n^{1/3}}$.
  - In either case, it is unlikely that an integer number of bins will result. In that case, round off to the nearest integer, since we cannot have a non-integer number of bins.
  - For example, if 20 measurements are taken (n = 20), The Sturgis rule yields J = 5.3 bins. The Rice rule yields J = 5.4 bins. We see that for small n, Sturgis and Rice give similar results. Since there can only be an integer number of bins, we would use 6 bins here, since 6 is the generally accepted minimum number of bins in a histogram.
  - As another example, suppose there is a larger number of measurements in the sample, e.g., n = 1000. The Sturgis rule yields J = 10.9 bins, which we round to 11 bins. The Rice rule yields J = 20 bins. So, we see that with a large n, Sturgis and Rice yield quite different results. Which one to use? We suggest somewhere between the two – perhaps the average of the two rules, around 14 to 15 bins.
  - There is no generally agreed-upon *maximum* number of bins to use in a histogram, although some authors say the maximum number of bins is 15; others say 20. Bottom line: depends on your data.
  - As a general rule of thumb in this course, use between 6 and 20 bins in a histogram, i.e., $\boxed{6 \le J \le 20}$.
  - *Note*: In the end, it is best to experiment with a few different bin numbers, and pick the one that yields the most meaningful, useful histogram.
    - If J is too small, your may not get the correct histogram shape.
    - If J is too big, your histogram will look "choppy" – it will have bins with zero data points.
- Probability histograms and normalized histograms:
  - The histogram can be modified by dividing the vertical axis by the total number of measurements, n. The resulting ***probability histogram*** has the same shape, but the vertical axis represents a ***relative frequency*** or ***probability***, i.e., $\boxed{\text{probability}_j = \dfrac{\text{number of measurements in bin } j}{n}}$.



    In this histogram, the probability that x lies between x_3 and x_4 is about 0.1 or 10%, as indicated by the red bar for j = 3, i.e., the third one.
  - We define a ***vertically normalized histogram*** by further dividing the vertical axis by the bin width or interval width. The vertical axis of the vertically normalized histogram is defined as $\boxed{f\left(x_{\text{mid},j}\right) = \dfrac{\text{probability}_j}{\Delta x_j} = \dfrac{\text{number of measurements in bin } j}{n \cdot \Delta x_j}}$ where n is the total number of

measurements. This guarantees mathematically that the *area* of the bin is equal to the probability that x lies in that bin, as sketched below. In this example, $\Delta x = 10$ = constant; therefore, the values on the vertical axis decrease by a factor of 10.



- **Creating histograms with Microsoft Excel**
  - A histogram macro is built into Excel.
    - **Office 2003**: Tools-Data Analysis-Histogram.
    - **Office 2007**: *Data Tab*. Then, in the *Analysis* area, Data Analysis-Histogram.
  - Highlight the actual data for the *Input Range*, highlight the bin data (optional – if you want your own bins) for the *Bin Range*, specify a cell as the top left cell of the *Output Range*, and check the Chart Output option to generate the actual histogram. OK.

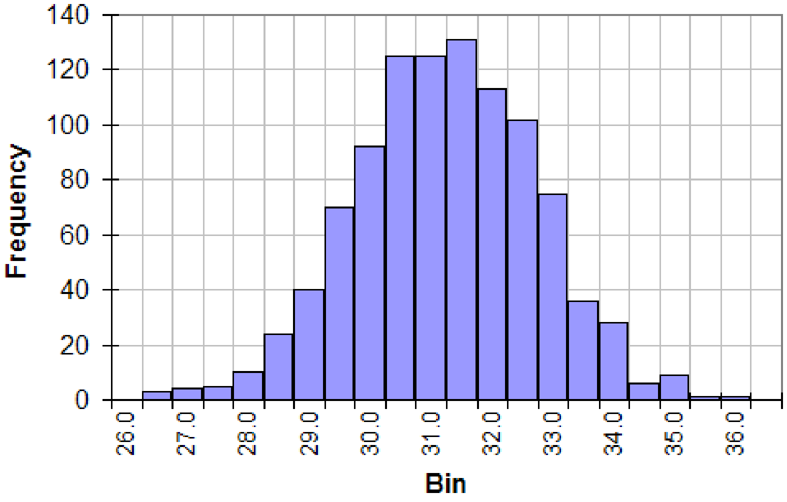
- Example:

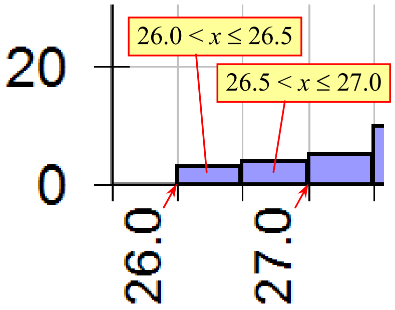
Given: One thousand temperature measurements are taken with a thermocouple of a steady-state process. The data are provided in an Excel spreadsheet (Temperature_data_analysis.xls).

To do: Calculate the basic statistics of these data (sample mean, sample standard deviation, etc.), and generate a nice-looking histogram.

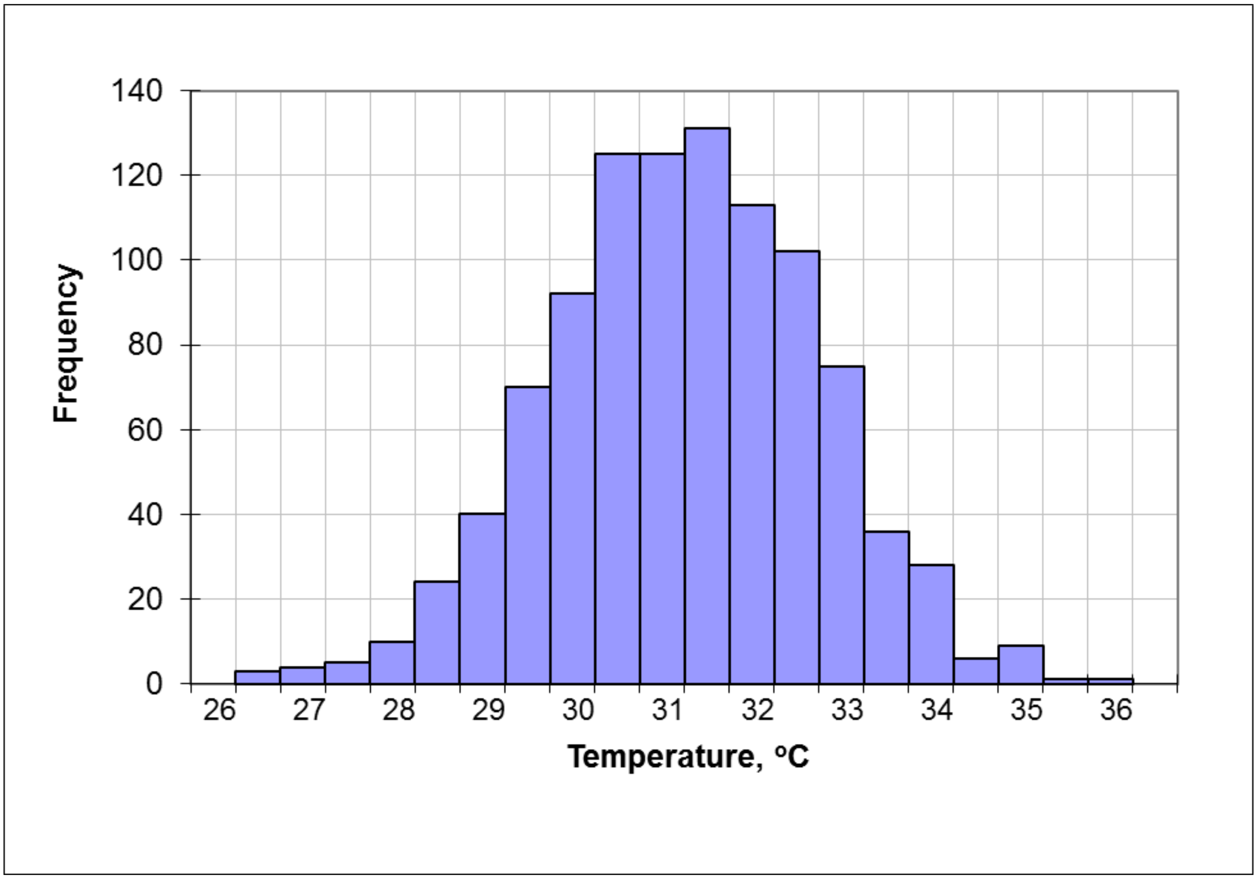
Solution:

- We use Excel to calculate the basic statistics: **sample mean = 31.009°C** (to 5 significant digits – it is acceptable to add one extra digit of precision when taking an average, especially when n is large), **sample standard deviation = 1.488°C**, **sample median = 31.015°C**, and **sample mode = 30.27°C**. Notice that the sample mean, median, and mode are close, as expected since $n = 1000$.
- To generate a histogram, we define our own bins so that the histogram is "cleaner" than the default histogram created by Excel. The histogram is quite symmetric, so we are confident that the errors in the readings are nearly random. Details will be shown in class using Excel.

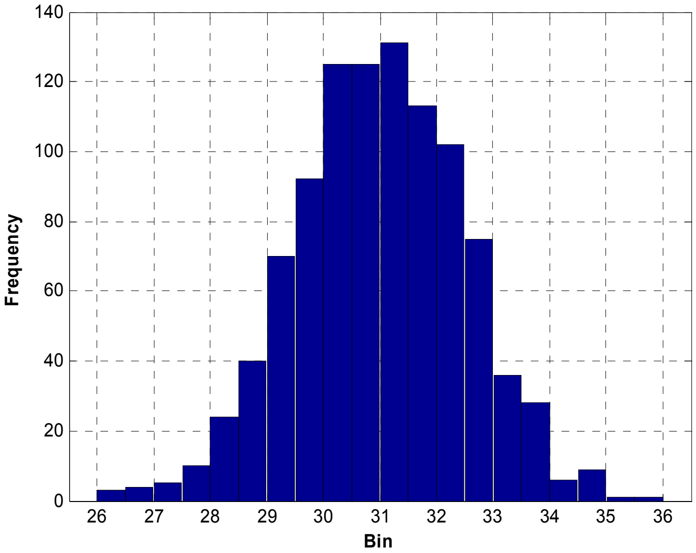




Discussion: Be careful with the interpretation of the horizontal axis! Excel always plots the value to the *left* of its actual location. For example, the label "26.0" points to the tick to its immediate right, as shown in the annotated magnified view of the bottom left portion of the histogram. The left-most blue frequency value represents data for which $26.0 < x \leq 26.5$, the second frequency represents $26.5 < x \leq 27.0$, etc. Further manipulation of the plot can sort of correct this. Click on the horizontal axis, and then click on the "Align Text Right" icon (right justify) in the top tool bar. This moves the axis labels to the right, but still not centered on the tick mark. I also used the Format Axis option to make the numbers horizontal instead of vertical, and changed the label from "Bin" to "Temperature, ^{o}C". Below is what would be considered a nice-looking, proper histogram.

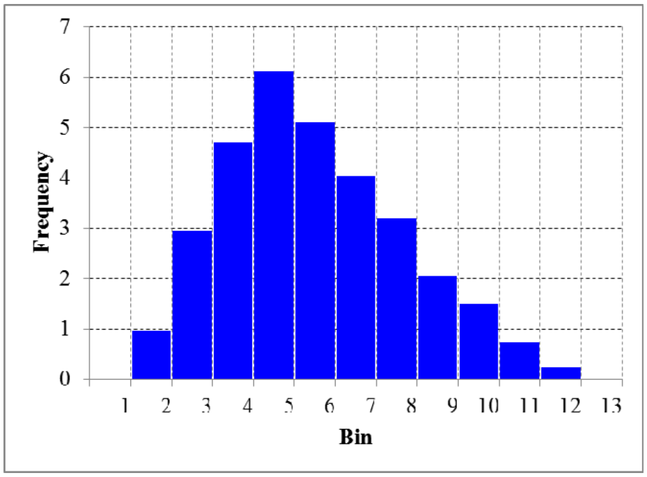


- **Creating histograms with Matlab**
  - Matlab also has a built-in histogram generator: **[n,xout] = hist(x);** where **x** is the raw data array, **n** is the frequency, and **xout** is the center value of each bin.
  - The plot is created using the bar function: **bar(xout, n, width);** where **width** is typically set to 1 (no gaps between bars).
  - Matlab draws the horizontal axis properly – it does not suffer from the shifted axis problem as in Excel. A Matlab version of the above histogram is shown here.

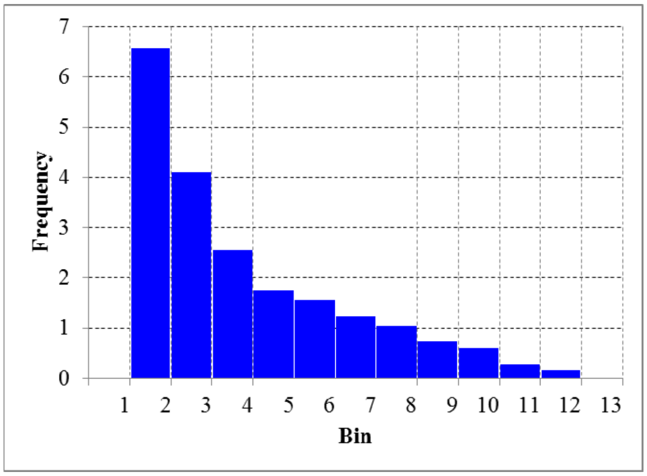


- **Standard types of histograms**
  - Not all data generate a nice symmetric histogram like the ones shown above. In many cases, the histogram is skewed to one side or may have more than one peak (e.g., exam grades often produce a strange-looking histogram)!
  - There are five standard histogram shapes that have been given standardized names:
    - **Symmetric**: Nearly symmetric left to right, with a peak very close to the middle, like the ones above.
    - **Skewed**: Shifted to one side or the other, with a peak clearly located on one "preferred" side.
    - **J-Shaped**: Skewed very much to one side, with the peak at or near the lowest or highest bin.
    - **Bimodal**: Two distinct peaks, and usually fairly symmetric in the vicinity of either peak.
    - **Uniform**: Nearly the same frequency for each bin (no distinct peaks; fairly flat over whole range).

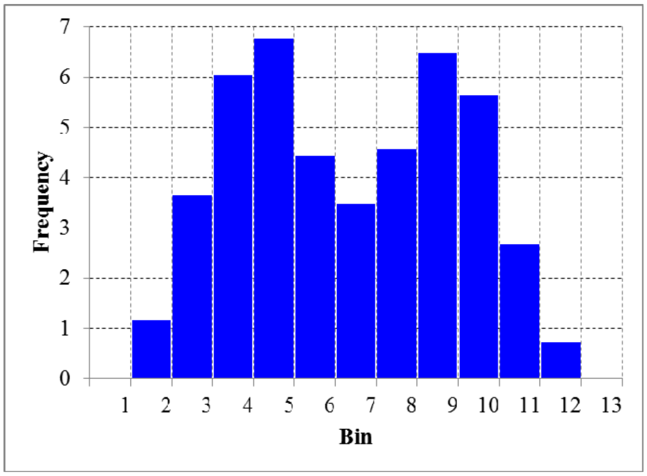
- The last four of these are sketched below:



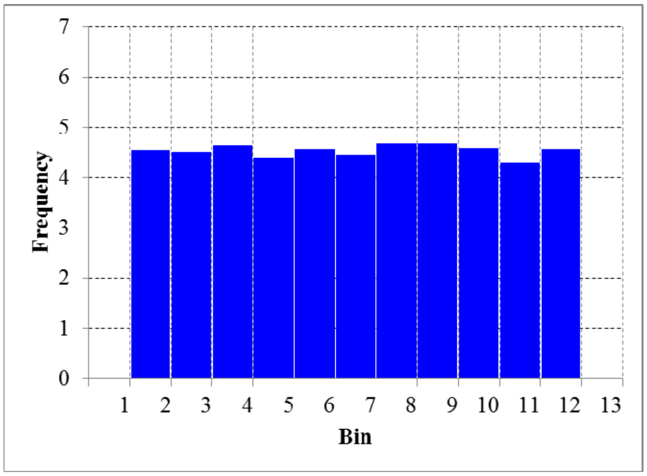
Skewed



J-shaped



Bimodal



Uniform (nearly)