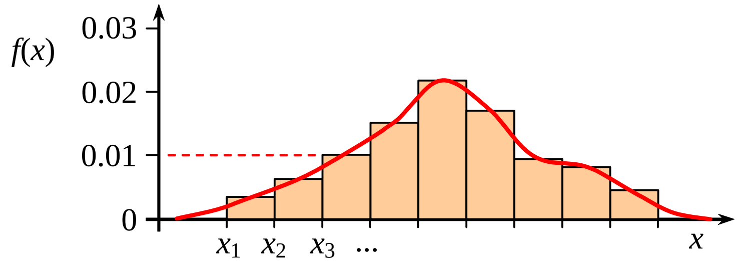
Probability Density Functions

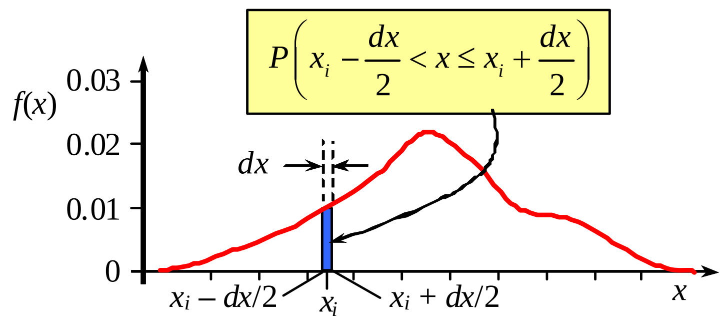
Author: John M. Cimbala, Penn State University
Latest revision: 20 January 2010

Probability Density Functions

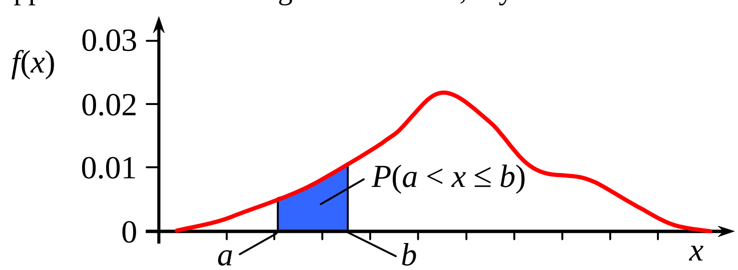
- ***Probability density function*** – In simple terms, a ***probability density function*** (***PDF***) is constructed by drawing a smooth curve fit through the vertically normalized histogram as sketched. You can think of a PDF as ***the smooth limit of a vertically normalized histogram*** if there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



  - The main difference between a histogram and a PDF is that a histogram involves ***discrete data*** (individual bins or classes), whereas a PDF involves ***continuous data*** (a smooth curve).
  - Mathematically, $f(x)$ is defined as $\boxed{f(x_i) = \dfrac{P\left(x_i - \dfrac{dx}{2} < x \le x_i + \dfrac{dx}{2}\right)}{dx}}$, where $\boxed{P\left(x_i - \dfrac{dx}{2} < x \le x_i + \dfrac{dx}{2}\right)}$ represents the probability that variable x lies in the given range, and $f(x)$ is the probability density function (PDF). In other words, for the given infinitesimal range of width dx between $x_i - dx/2$ and $x_i + dx/2$, 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 $\boxed{P(a < x \le b) = \int_{x=a}^{x=b} f(x)\,dx}$



  - If $a \to -\infty$ and $b \to +\infty$, the probability must equal 1 (100%), i.e., $\boxed{P(-\infty < x < \infty) = \int_{x=-\infty}^{x=\infty} f(x)\,dx = 1}$.
    In other words, the probability that x lies between $-\infty$ and $+\infty$ is 100% (a fact that should be obvious, since there are no other possibilities for real number x).
  - Once we have defined the probability density function $f(x)$, we leave the system of *discrete random variables* and enter the system of *continuous random variables*, on which we make some more formal definitions:
    - ***Expected value*** is defined in terms of the probability density function as ***the mean of all possible x values in the continuous system***. Namely, $\boxed{\text{expected value} = \mu = E(x) = \int_{-\infty}^{\infty} xf(x)\,dx}$. 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, $\bar{x} \ne \mu$, but $\bar{x} \to \mu$ ***when n is large***, i.e., ***the sample mean approaches the***

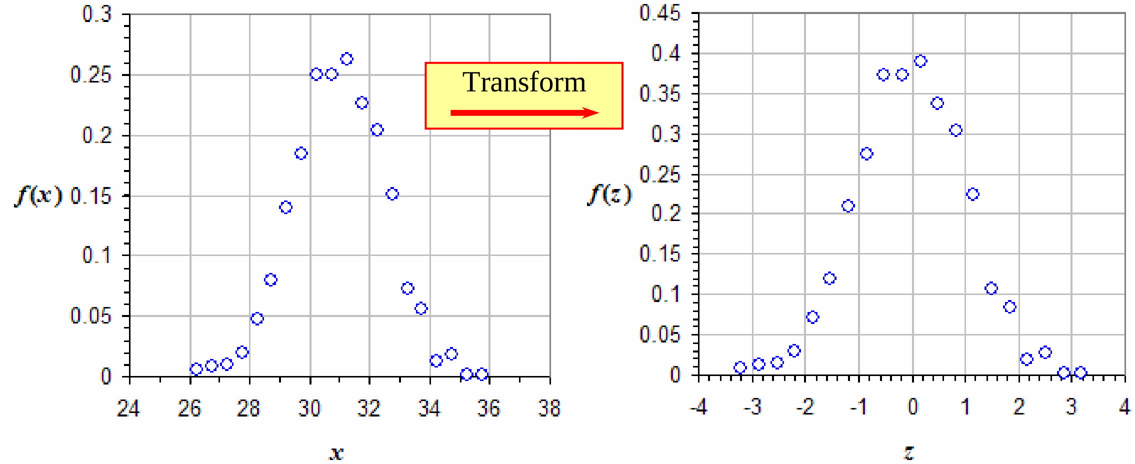
expected value when n is large. $\overline{x}$ and μ are often used interchangeably, but this should be done only if n is large.

- ***Standard deviation*** is defined in terms of the PDF as $\boxed{\text{standard deviation} = \sigma = \sqrt{\int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx}}$. 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***. ***If n is large, S → σ***. Often, S and σ are used interchangeably, but this should be done only if n is large.

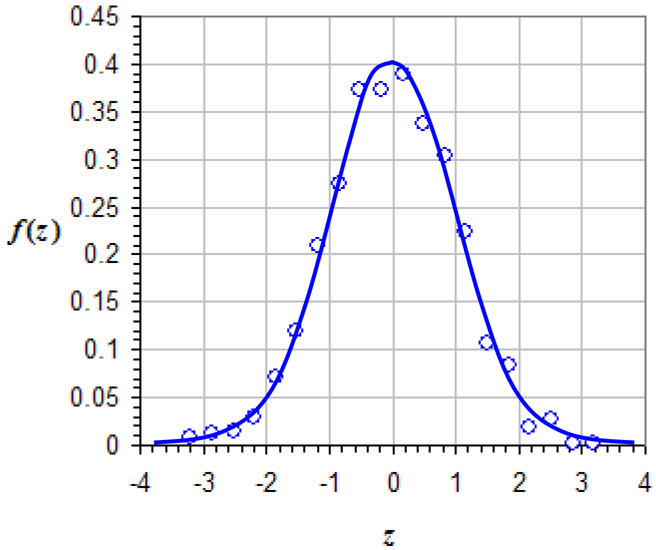
- ***Normalized probability density function*** – a ***normalized probability density function*** is constructed by transforming both the abscissa (horizontal axis) and ordinate (vertical axis) of the PDF plot as follows: $\boxed{z = \dfrac{x-\mu}{\sigma}}$ and $\boxed{f(z) = \sigma f(x)}$.
  - o The 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.
  - o When normalized in this way, the normalized PDF can be directly compared to standard PDFs, which we discuss in a later learning module.
  - o To summarize, here are several steps used in Excel to generate a normalized PDF of experimental data:
    1. Generate the histogram with Excel as discussed in the histogram learning module. Excel generates a table called a ***frequency table***. The table contains two columns, **bin** and **frequency**. *Bin* is the *maximum* value of the range of each bin, and *frequency* is the number of data points in that bin range. (For example, suppose there are 200 data points total, the mean value of x is 10.0, and the standard deviation of the data set is 3.0. Also suppose that 8 of those data points lie in the bin with x between 4 and 6 ($4 < x \le 6$). Thus, for this bin, Bin = 6 and Frequency = 8.)
    2. Create a new column called **probability** in which you divide each frequency by the total number of data points. This gives the probability that a data point lies in that bin, i.e. $\boxed{\text{probability} = \text{frequency}/n}$. (In the example here, probability = 8/200 = 0.040 or 4.0%.)
    3. Create a new column called $\boldsymbol{x_{mid}}$ in which you list the mid value of each bin: $\boxed{x_{mid} = (x_{min} + x_{max})/2}$. (In the example here, the mid value of the sample bin is (4 + 6)/2 = 5.0.)
    4. Create a new column called ***f(x)*** in which you divide each probability by the appropriate bin width, i.e., $\boxed{f(x) = \text{probability}/\Delta x}$. (In the example here, the bin width of the sample bin is $\Delta x = 6 - 4 = 2$, and $f(x) = 0.04/2 = 0.02$ at $x = x_{mid} = 5.0$.) A smoothed plot of $f(x)$ versus x is the PDF.
    5. Create a new column called ***z*** in which you normalize the x values into nondimensional z values. This is accomplished by converting each mid value of x into z: $\boxed{z = (x-\mu)/\sigma}$. (In the example here, z for the sample bin is $z = (5.0 - 10.0)/3.0 = -1.667$.)
    6. Create a new column called ***f(z)*** in which you normalize the PDF into the $f(z)$ values. This is accomplished by converting each $f(x)$ into $f(z)$: $\boxed{f(z) = \sigma \cdot f(x)}$. (In the example here, $f(z)$ of the sample bin is $f(z) = 0.02*3.0 = 0.060$ at $z = -1.667$.)
    7. Finally, a plot of $f(z)$ vs. z can be generated. A smooth curve through these data represents the ***normalized PDF***.

- **<u>Example</u>**:
  Given: The same 1000 temperature measurements used in a previous example for generating a histogram. The data are provided in an Excel spreadsheet (<u>Temperature_data_analysis.xls</u>) on the website.
  To do: Generate a PDF of these data. Normalize the PDF.
  Solution:
  - o In a previous example (see the Histogram learning module), we generated a histogram of the temperature data. We begin with the bin and frequency data generated in Excel.

- To generate the PDF, we follow the step-by-step instructions provided above. This will be shown in class in Excel. The vertically normalized PDF is shown below (left side).



- Finally, we transform to normalized variables – the fully normalized PDF is shown above (right side). Notice that the shape is the same, but the variable transformation to $f(z)$ is *nondimensional*, making it more useful for comparison with other probability density distributions.
- The final PDF should be continuous, not discrete. Because of scatter, it is difficult to get Excel to draw a smooth curve through these data. For lack of a better method at this point, we sketch the smooth curve "by eye" below:



Discussion:

- The peak in the vertically normalized PDF occurs at $x \approx 31$, which is very close to the sample mean. This peak transforms to $z \approx 0$ in the fully normalized PDF; this is a useful feature of the normalization.
- We can estimate the area under the $f(x)$ curve "by eye" by counting squares – the area is indeed approximately 1.0 or 100%, as it must be.
- We can also estimate the area under the $f(z)$ curve "by eye" – it is approximately 1.0 or 100%, as it also must be.

- There are several standard PDFs discussed in statistics literature. Of these, the ***normal PDF***, is the most common, and will be discussed next. We will also compare the above results with the normal PDF.