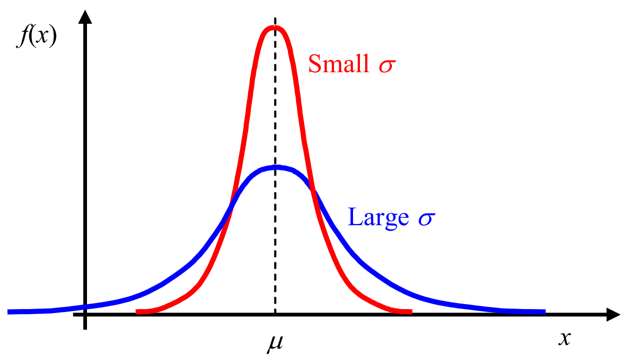
The Gaussian or Normal Probability Density Function

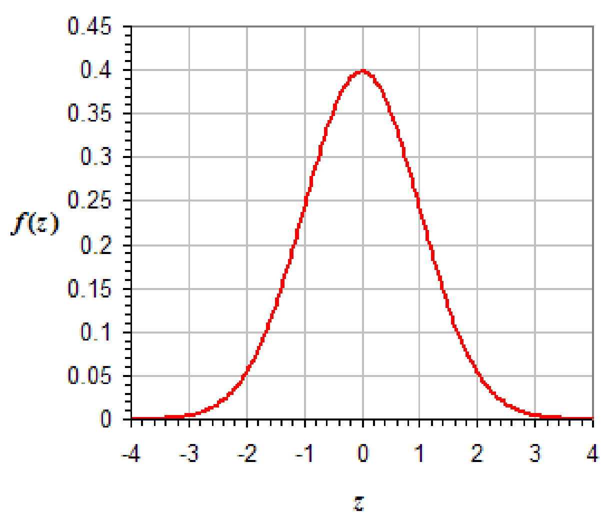
Author: John M. Cimbala, Penn State University
Latest revision: 11 September 2013

The Gaussian or Normal Probability Density Function

- ***Gaussian or normal PDF*** – The ***Gaussian probability density function*** (also called the ***normal probability density function*** or simply the ***normal PDF***) is ***the vertically normalized PDF that is produced from a signal or measurement that has purely random errors***.
  - The normal probability density function is $\boxed{f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{-(x-\mu)^2}{2\sigma^2}}=\frac{1}{\sigma\sqrt{2\pi}}\exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)}$.
  - Here are some of the properties of this special distribution:
    - It is symmetric about the mean.
    - The mean and median are both equal to μ, the *expected value* (at the peak of the distribution). [The mode is undefined for a smooth, continuous distribution.]
    - Its plot is commonly called a "***bell curve***" because of its shape.
    - The actual shape depends on the magnitude of the standard deviation. Namely, if σ is small, the bell will be tall and skinny, while if σ is large, the bell will be short and fat, as sketched.



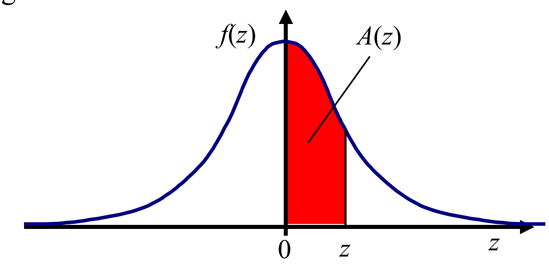
- ***Standard normal density function*** – All of the Gaussian PDF cases, for *any* mean value and for *any* standard deviation, can be collapsed into *one normalized curve* called the ***standard normal density function***.
  - This normalization is accomplished through the variable transformations introduced previously, i.e., $\boxed{z=\frac{x-\mu}{\sigma}}$ and $\boxed{f(z)=\sigma f(x)}$, which yields $\boxed{f(z)=\sigma f(x)=\frac{1}{\sqrt{2\pi}}e^{-z^2/2}=\frac{1}{\sqrt{2\pi}}\exp\left(-z^2/2\right)}$.
    This standard normal density function is valid for *any* signal measurement, with *any* mean, and with *any* standard deviation, provided that the errors (deviations) are *purely random*.
  - A plot of the standard normal (Gaussian) density function was generated in Excel, using the above equation for $f(z)$. It is shown to the right.



  - It turns out that the probability that variable x lies between some range x_1 and x_2 is the *same* as the probability that the transformed variable z lies between the corresponding range z_1 and z_2, where z is the transformed variable defined above. In other words, $\boxed{P(x_1<x\le x_2)=P(z_1<z\le z_2)}$ where $\boxed{z_1=\frac{x_1-\mu}{\sigma}}$ and $\boxed{z_2=\frac{x_2-\mu}{\sigma}}$.
  - Note that z is *dimensionless*, so there are no units to worry about, so long as the mean and the standard deviation are expressed in the *same units*.
  - Furthermore, since $\boxed{P(x_1<x\le x_2)=\int_{x_1}^{x_2}f(x)dx}$, it follows that $\boxed{P(x_1<x\le x_2)=\int_{z_1}^{z_2}f(z)dz}$.
  - We define ***$A(z)$*** as ***the area under the curve between 0 and z***, i.e., the special case where $z_1 = 0$ in the above integral, and z_2 is simply z. In other words, $A(z)$ is ***the probability that a measurement lies between 0 and z***, or $\boxed{A(z)=\int_0^z f(z)dz}$, as illustrated on the graph below.

- For convenience, integral $A(z)$ is tabulated in statistics books, but it can be easily calculated to avoid the round-off error associated with looking up and interpolating values in a table.
- Mathematically, it can be shown that $\boxed{A(z) = \frac{1}{2}\text{erf}\left(\frac{z}{\sqrt{2}}\right)}$

  where erf(η) is the ***error function***, defined as $\boxed{\text{erf}(\eta) = \frac{2}{\sqrt{\pi}}\int_{\xi=0}^{\xi=\eta}\exp(-\xi^2)d\xi}$.



- Below is a table of $A(z)$, produced using Excel, which has a built-in error function, ERF(*value*). Excel has another function that can be used to calculate $A(z)$, namely $\boxed{A(z) = \text{NORMSDIST}(\text{ABS}(z)) - 0.5}$.

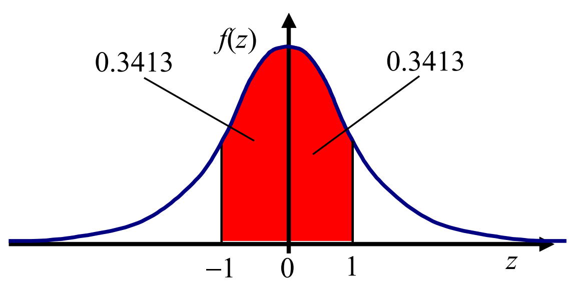
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.00000	0.00399	0.00798	0.01197	0.01595	0.01994	0.02392	0.02790	0.03188	0.03586
0.1	0.03983	0.04380	0.04776	0.05172	0.05567	0.05962	0.06356	0.06749	0.07142	0.07535
0.2	0.07926	0.08317	0.08706	0.09095	0.09483	0.09871	0.10257	0.10642	0.11026	0.11409
0.3	0.11791	0.12172	0.12552	0.12930	0.13307	0.13683	0.14058	0.14431	0.14803	0.15173
0.4	0.15542	0.15910	0.16276	0.16640	0.17003	0.17364	0.17724	0.18082	0.18439	0.18793
0.5	0.19146	0.19497	0.19847	0.20194	0.20540	0.20884	0.21226	0.21566	0.21904	0.22240
0.6	0.22575	0.22907	0.23237	0.23565	0.23891	0.24215	0.24537	0.24857	0.25175	0.25490
0.7	0.25804	0.26115	0.26424	0.26730	0.27035	0.27337	0.27637	0.27935	0.28230	0.28524
0.8	0.28814	0.29103	0.29389	0.29673	0.29955	0.30234	0.30511	0.30785	0.31057	0.31327
0.9	0.31594	0.31859	0.32121	0.32381	0.32639	0.32894	0.33147	0.33398	0.33646	0.33891
1.0	0.34134	0.34375	0.34614	0.34849	0.35083	0.35314	0.35543	0.35769	0.35993	0.36214
1.1	0.36433	0.36650	0.36864	0.37076	0.37286	0.37493	0.37698	0.37900	0.38100	0.38298
1.2	0.38493	0.38686	0.38877	0.39065	0.39251	0.39435	0.39617	0.39796	0.39973	0.40147
1.3	0.40320	0.40490	0.40658	0.40824	0.40988	0.41149	0.41309	0.41466	0.41621	0.41774
1.4	0.41924	0.42073	0.42220	0.42364	0.42507	0.42647	0.42785	0.42922	0.43056	0.43189
1.5	0.43319	0.43448	0.43574	0.43699	0.43822	0.43943	0.44062	0.44179	0.44295	0.44408
1.6	0.44520	0.44630	0.44738	0.44845	0.44950	0.45053	0.45154	0.45254	0.45352	0.45449
1.7	0.45543	0.45637	0.45728	0.45818	0.45907	0.45994	0.46080	0.46164	0.46246	0.46327
1.8	0.46407	0.46485	0.46562	0.46638	0.46712	0.46784	0.46856	0.46926	0.46995	0.47062
1.9	0.47128	0.47193	0.47257	0.47320	0.47381	0.47441	0.47500	0.47558	0.47615	0.47670
2.0	0.47725	0.47778	0.47831	0.47882	0.47932	0.47982	0.48030	0.48077	0.48124	0.48169
2.1	0.48214	0.48257	0.48300	0.48341	0.48382	0.48422	0.48461	0.48500	0.48537	0.48574
2.2	0.48610	0.48645	0.48679	0.48713	0.48745	0.48778	0.48809	0.48840	0.48870	0.48899
2.3	0.48928	0.48956	0.48983	0.49010	0.49036	0.49061	0.49086	0.49111	0.49134	0.49158
2.4	0.49180	0.49202	0.49224	0.49245	0.49266	0.49286	0.49305	0.49324	0.49343	0.49361
2.5	0.49379	0.49396	0.49413	0.49430	0.49446	0.49461	0.49477	0.49492	0.49506	0.49520
2.6	0.49534	0.49547	0.49560	0.49573	0.49585	0.49598	0.49609	0.49621	0.49632	0.49643
2.7	0.49653	0.49664	0.49674	0.49683	0.49693	0.49702	0.49711	0.49720	0.49728	0.49736
2.8	0.49744	0.49752	0.49760	0.49767	0.49774	0.49781	0.49788	0.49795	0.49801	0.49807
2.9	0.49813	0.49819	0.49825	0.49831	0.49836	0.49841	0.49846	0.49851	0.49856	0.49861
3.0	0.49865	0.49869	0.49874	0.49878	0.49882	0.49886	0.49889	0.49893	0.49896	0.49900
3.1	0.49903	0.49906	0.49910	0.49913	0.49916	0.49918	0.49921	0.49924	0.49926	0.49929
3.2	0.49931	0.49934	0.49936	0.49938	0.49940	0.49942	0.49944	0.49946	0.49948	0.49950
3.3	0.49952	0.49953	0.49955	0.49957	0.49958	0.49960	0.49961	0.49962	0.49964	0.49965
3.4	0.49966	0.49968	0.49969	0.49970	0.49971	0.49972	0.49973	0.49974	0.49975	0.49976
3.5	0.49977	0.49978	0.49978	0.49979	0.49980	0.49981	0.49981	0.49982	0.49983	0.49983
3.6	0.49984	0.49985	0.49985	0.49986	0.49986	0.49987	0.49987	0.49988	0.49988	0.49989
3.7	0.49989	0.49990	0.49990	0.49990	0.49991	0.49991	0.49992	0.49992	0.49992	0.49992
3.8	0.49993	0.49993	0.49993	0.49994	0.49994	0.49994	0.49994	0.49995	0.49995	0.49995
3.9	0.49995	0.49995	0.49996	0.49996	0.49996	0.49996	0.49996	0.49996	0.49997	0.49997
4.0	0.49997	0.49997	0.49997	0.49997	0.49997	0.49997	0.49998	0.49998	0.49998	0.49998

- To read the value of $A(z)$ at a particular value of z,
  - Go down to the *row* representing the *first two digits* of z.
  - Go across to the *column* representing the *third digit* of z.
  - Read the value of $A(z)$ from the table.
  - Example: At $z = 2.54$, $A(z) = A(2.5 + 0.04) = 0.49446$. These values are highlighted in the above table as an example.
  - Since the normal PDF is symmetric, $A(-z) = A(z)$, so there is no need to tabulate negative values of z.

- ***Linear interpolation***:
  - By now in your academic career, you should be able to linearly interpolate from tables like the above.
  - As a quick example, let's estimate $A(z)$ at $z = 2.546$.
  - The simplest way to interpolate, which works for both increasing and decreasing values, is to ***always work from top to bottom***, equating the fractional values of the known and desired variables.

$\underline{z}$	$\underline{A(z)}$
2.54	0.49446
2.546	$A(z)$ = ?
2.55	0.49461

  - We zoom in on the appropriate region of the table, straddling the z value of interest, and set up for interpolation – see sketch. The ratio of the red difference to the blue difference is the same for either column. Thus, keeping the color code, we set up our equation as $\dfrac{2.546-2.55}{2.54-2.55}=\dfrac{A(z)-0.49461}{0.49446-0.49461}$.
  - Solving for $A(z)$ at $z = 2.546$ yields $A(z)=\dfrac{2.546-2.55}{2.54-2.55}(0.49446-0.49461)+0.49461=0.49455$.
- ***Special cases***:
  - If $z = 0$, obviously the integral $A(z) = 0$. This means physically that there is zero probability that x will exactly equal the mean! (To be exactly equal would require equality out to an infinite number of decimal places, which will never happen.)
  - If $z = \infty$, $A(z) = 1/2$ since $f(z)$ is symmetric. This means that there is a 50% probability that x is greater than the mean value. In other words, $z = 0$ represents the median value of x.
  - Likewise, if $z = -\infty$, $A(z) = 1/2$. There is a 50% probability that x is *less* than the mean value.
  - If $z = 1$, it turns out that $\boxed{A(1)=\int_0^1 f(z)dz=0.3413}$ to four significant digits. This is a special case, since by definition $z=(x-\mu)/\sigma$. Therefore, $z = 1$ represents a value of x exactly *one* standard deviation *greater* than the mean.
  - A similar situation occurs for $z = -1$ since $f(z)$ is symmetric, and $\boxed{A(-1)=\int_0^{-1} f(z)dz=0.3413}$ to four significant digits. Thus, $z = -1$ represents a value of x exactly one standard deviation *less* than the mean.



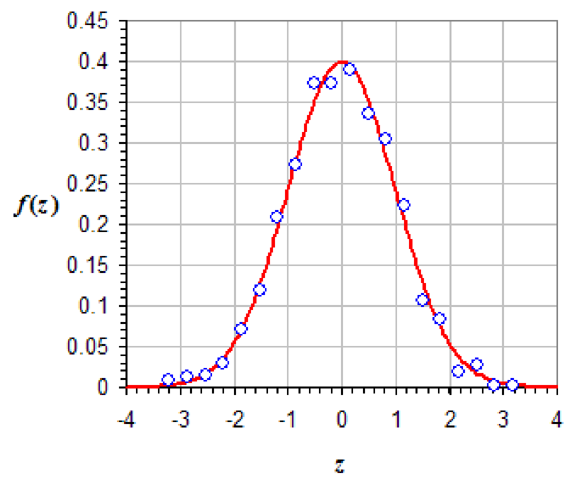
  - Because of this symmetry, we conclude that the probability that z lies between –1 and 1 is 2(0.3413) = 0.6826 or 68.26%. In other words, ***there is a 68.26% probability that for some measurement, the transformed variable z lies within ± one standard deviation from the mean*** (which is zero for this pdf).
  - Translated back to the original measured variable x, $\boxed{P(\mu-\sigma<x\le\mu+\sigma)=68.26\%}$. In other words, ***the probability that a measurement lies within ± one standard deviation from the mean is 68.26%***.
- ***Confidence level*** – The above illustration leads to an important concept called ***confidence level***. For the above case, we are 68.26% confident that ***any random measurement of x will lie within ± one standard deviation from the mean value***.
  - I would not bet my life savings on something with a 68% confidence level. A higher confidence level is obtained by choosing a larger z value. For example, for $z = 2$ (two standard deviations away from the mean), it turns out that $\boxed{A(2)=\int_0^2 f(z)dz=0.4772}$ to four significant digits.
  - Again, due to symmetry, multiplication by two yields the probability that x lies within *two* standard deviations from the mean value, either to the right or to the left. Since 2(0.4772) = 0.9544, we are ***95.44% confident that x lies within ± two standard deviations of the mean***.
  - Since 95.44 is close to 95, most engineers and statisticians ignore the last two digits and state simply that there is about a ***95% confidence level that x lies within ± two standard deviations from the mean***. This is in fact the ***engineering standard***, called the "***two sigma confidence level***" or the "***95% confidence level***."
  - For example, when a manufacturer reports the value of a property, like resistance, the report may state "R = 100 ± 9 Ω (ohms) with 95% confidence." This means that the mean value of resistance is 100 Ω, and that 9 ohms represents two standard deviations from the mean.

- In fact, the words "with 95% confidence" are often not even written explicitly, but are *implied*. In this example, by the way, you can easily calculate the standard deviation. Namely, since 95% confidence level is about the same as 2 sigma confidence, $2\sigma \approx 9\ \Omega$, or $\sigma \approx 4.5\ \Omega$.
- For more stringent standards, the confidence level is sometimes raised to *three* sigma. For $z = 3$ (three standard deviations away from the mean), it turns out that $\boxed{A(3) = \int_0^3 f(z)\,dz = 0.4987}$ to four significant digits. Multiplication by two (because of symmetry) yields the probability that x lies within ± *three* standard deviations from the mean value. Since 2(0.4987) = 0.9974, we are ***99.74% confident that x lies within ± three standard deviations from the mean***.
- Most engineers and statisticians round down and state simply that there is about a ***99.7% confidence level that x lies within ± three standard deviations from the mean***. This is in fact a stricter engineering standard, called the "***three sigma confidence level***" or the "***99.7% confidence level***."
- Summary of confidence levels: The ***empirical rule*** states that for any normal or Gaussian PDF,
  - Approximately 68% of the values fall within 1 standard deviation from the mean in either direction.
  - Approximately 95% of the values fall within 2 standard deviations from the mean in either direction. [This one is the standard "two sigma" engineering confidence level for most measurements.]
  - Approximately 99.7% of the values fall within 3 standard deviations from the mean in either direction. [This one is the stricter "three sigma" engineering confidence level for more precise measurements.]
- More recently, many manufacturers are striving for "***six sigma***" confidence levels.

- **Example**:

Given: The same 1000 temperature measurements used in a previous example for generating a histogram and a PDF. The data are provided in an Excel spreadsheet (Temperature_data_analysis.xls).

To do: (***a***) Compare the normalized PDF of these data to the normal (Gaussian) PDF. Are the measurement errors in this sample purely random? (***b***) Predict how many of the temperature measurements are greater than 33.0°C, and compare with the actual number.

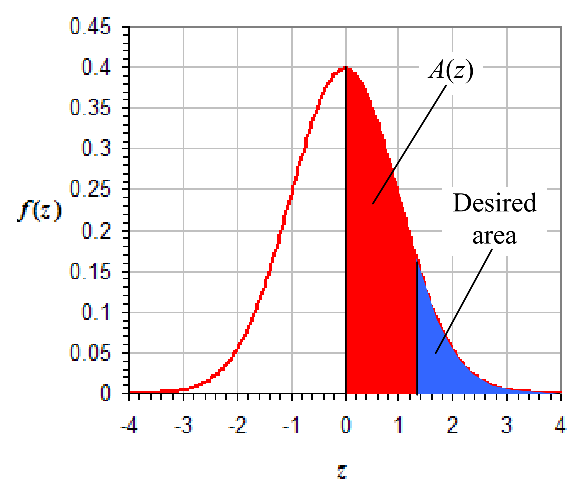


Solution:

(***a***) We plot the experimentally generated PDF (blue circles) and the theoretical normal PDF (red curve) on the same plot. The agreement is excellent, indicating that **the errors are very nearly random**. Of course, the agreement is not *perfect* – this is because n is finite. If n were to increase, we would expect the agreement to get better (less scatter and difference between the experimental and theoretical PDFs).

(***b***) For this data set, we had calculated the sample mean to be $\bar{x} = 31.009$ and sample standard deviation to be $S = 1.488$. Since $n = 1000$, the sample size is large enough to assume that expected value μ is nearly equal to $\bar{x}$, and standard deviation σ is nearly equal to S. At the given value of temperature (set $x = 33.0$°C), we normalize to obtain z, namely,

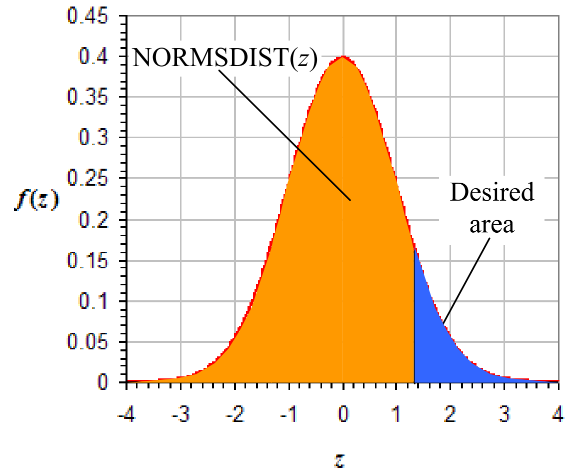
$$z = \frac{x-\mu}{\sigma} \approx \frac{x-\bar{x}}{S} = \frac{(33.0-31.009)\,^\circ\mathrm{C}}{1.488\,^\circ\mathrm{C}} = 1.338$$



(notice that z is nondimensional). We calculate area $A(z)$, either by interpolation from the above table or by direct calculation. The table yields $A(z) = 0.40955$, and the equation yields $A(z) = \frac{1}{2}\mathrm{erf}\left(\frac{z}{\sqrt{2}}\right) =$ $\frac{1}{2}\mathrm{erf}\left(\frac{1.338}{\sqrt{2}}\right) = 0.409552$. This means that 40.9552% of the measurements are predicted to lie between the mean (31.009°C) and the given value of 33.0°C (red area on the plot). The percentage of measurements *greater than* 33.0°C is 50% – 40.9552% = 9.0448% (blue area on the plot). Since n = 1000, we predict that 0.090448·1000 = 90.448 of the measurements exceed 33.0°C. Rounding to the nearest integer, we predict that $\boxed{90}$ **measurements are greater than 33.0°C**. Looking at the actual data, we count $\boxed{81}$ **temperature readings greater than 33.0°C**.

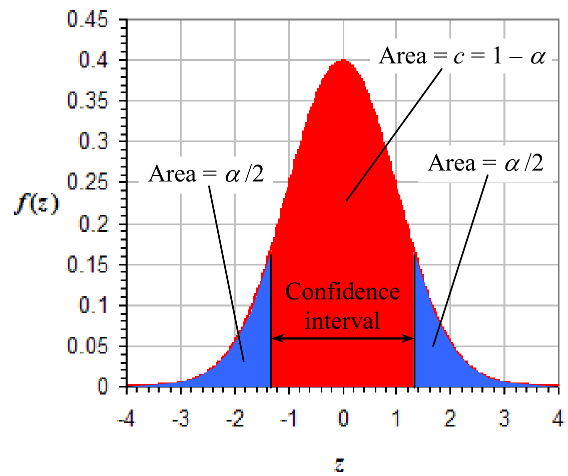
Discussion:

- The percentage error between actual and predicted number of measurements is around –10%. This error would be expected to decrease if n were larger.
- If we had asked for the probability that T lies between the mean value and 33.0°C, the result would have been 0.4096 (to four digits), as indicated by the red area in the above plot. However, we are concerned here with the probability that T is *greater than* 33.0°C, which is represented by the blue area on the plot. This is why we had to subtract from 50% in the above calculation (50% of the measurements are greater than the mean), i.e., the probability that T is greater than 33.0°C is 0.5000 – 0.4096 = 0.0904.



- Excel's built-in NORMSDIST function returns the *cumulative* area from -∞ to z, the orange-colored area in the plot to the right. Thus, at z = 1.338, NORMSDIST(z) = 0.909552. This is the entire area on the left half of the Gaussian PDF (0.5) plus the area labeled $A(z)$ in the above plot. The desired blue area is therefore equal to 1 - NORMSDIST(z).

- ***Confidence level*** and ***level of significance***
  - ***Confidence level***, c, is defined as ***the probability that a random variable lies <u>within</u> a specified range of values***. The range of values itself is called the ***confidence interval***. For example, as discussed above, we are 95.44% confident that a purely random variable lies within ± two standard deviations from the mean. We state this as a confidence level of c = 95.44%, which we usually round off to 95% for practical engineering statistical analysis.



  - ***Level of significance***, α, is defined as ***the probability that a random variable lies <u>outside of</u> a specified range of values***. In the above example, we are 100 – 95.44 = 4.56% confident that a purely random variable lies either *below* or *above* two standard deviations from the mean. (We usually round this off to 5% for practical engineering statistical analysis.)
  - Mathematically, confidence level and level of significance must add to 1 (or in terms of percentage, to 100%) since they are complementary, i.e., $\boxed{\alpha + c = 1}$ or $\boxed{c = 1 - \alpha}$.
  - Confidence level is sometimes given the symbol c% when it is expressed as a percentage; e.g., at 95% confidence level, c = 0.95, c% = 95%, and $\alpha = 1 - c = 0.05$.
  - Both α and confidence level c represent ***probabilities***, or areas under the PDF, as sketched above for the normal or Gaussian PDF.
  - The blue areas in the above plot are called the ***tails***. There are *two* tails, one on the far left and one on the far right. The two tails *together* represent all the data outside of the confidence interval, as sketched.
  - ***Caution***: The area of *one* of the tails is only $\alpha/2$, not α. This factor of two has led to much grief, so be careful that you do not forget this!