Stat04-Normal.wxmx: Normal (μ, σ) Distributio

Table of Contents

1 Preface

In Stat04-Normal.wxmx we discuss the discrete Normal (μ, σ) probability distribution and its application, using Maxima tools and methods. μ is the mean (average) and σ is one standard deviation. The normal distribution is the most commonly-used probability distribution in all of statistics. This distribution is usually referred to as the Gaussian distribution by physicists; it is used so widely that it has become known as the "normal" distribution.
This is the fourth worksheet in my Statistics with Maxima section.

Edwin L. (Ted) Woollett
https://home.csulb.edu/~woollett/
April 25, 2024

2 References

In our series Statistics with Maxima we have used some examples and explanations (with much editing and additions) from:

Barlow, R. J.. Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences (Manchester Physics Series), 1993, Wiley

Gerhard Bohm, Günter Zech,
Introduction to Statistics and Data Analysis for Physicists, 3rd revised edition, 2017,
available from CERN webpage:
https://s3.cern.ch/inspire-prod-files-d/da9d786a06bf64d703e5c6665929ca01

Ch. 3 Reagle & Salvatore [RS], Statistics and Econometrics, 2nd ed, Schaum's Outlines, 2011, McGraw Hill,

Ch. 8 Fred Senese [FS], Symbolic Mathematics for Chemists: A Guide for Chemists, 2019, Wiley,

Louis Lyons, Statistics for Nuclear and Particle Physics, 1986, Cambridge Univ. Press,

Luca Lista, 'Statistical Methods for Data Analysis in Particle Physics', Lecture Notes in Physics 909, 2016, Springer-Verlag,

Frederick James, 'Statistical Methods in Experimental Physics', 2nd ed., 2006, World Scientific.

In this Stat04-Normal.wxmx worksheet we use examples from:
https://www.statology.org/example-of-normal-distribution/

https://www.scribbr.com/statistics/normal-distribution/

(%i4)
```
load (descriptive);
load (distrib);
fpprintprec : 6$
ratprint : false$
```
(%o1)

C:/maxima-5.43.2/share/maxima/5.43.2/share/descriptive/descriptive.mac

(%o2) *C:/maxima-5.43.2/share/maxima/5.43.2/share/distrib/distrib.mac*

Homemade functions fll, head, tail, Lsum are useful for looking at long lists.

(%i8)
```
fll ( aL) := [ first (aL), last (aL), length (aL) ]$
head(L) := if listp (L) then rest (L, - (length (L) - 3) ) else
      error("Input to 'head' must be a list of expressions ")$
tail (L) := if listp (L) then rest (L, length (L) - 3 )  else
       error("Input to 'tail' must be a list of expressions ")$
Lsum (aList) := apply ("+", aList)$
```

3 fracData (List, a, b)

fracData (alist, a, b) calculates the fraction of the list numbers which lie in the interval [a, b].

(%i9)
```
fracData (myData, xx1, xx2) :=
block ([ ccnt : 0 ],
   for j thru length (myData) do
     if myData[j] >= xx1 and myData[j] <= xx2 then ccnt : ccnt + 1,
   float (ccnt/ length (myData)) )$
```

4 confidence (q, m, s)

With q a number in the interval 0 < q < 1, and with m the mean and s the standard deviation of a Normal distribution, confidence (q, m, s) prints out the values dx, m - dx, m + dx, and outputs a list [m - dx, m + dx] which allows one to have 100*q % confidence a random value of x will lie within that interval, ie., within m +/- dx.

(%i10)
```
confidence (qq, mm, ss) :=
block ([ddx],
   ddx :
     float ( (quantile_normal (qq + (1 - qq)/2, mm, ss) -
                 quantile_normal ( (1 - qq)/2  , mm, ss))/2 ),
   print ( "delx = ", ddx ),
   print ( " x1 = ", mm - ddx, ",  x2 = ", mm + ddx ),
   [mm - ddx, mm + ddx])$
```

5 Mean and Variance from a Set of N Measurements

See Lyons, Sec. 1.4.2, pp. 9 - 11.

For a set of N separate measurements x_i (a sample of measurements - the size of the population is infinite) the sample mean <x> is defined as

$$E(x) = \langle x \rangle = \sum x_i / N = x_{av}$$

As N increases, both the numerator and denominator increase at roughly the same rate. As N --> ∞, <x> --> μ, the true mean of the population.

A measure of how wide the distribution of the N measurements is spread out about the mean is provided by the variance of the N measurements, which is defined to be the mean square deviation from the mean.

$$var(x) = s^2 = \sum (x_i - \mu)^2 / N$$

which is an estimate of the variance of the overall population (which is independent of the sample size N). As the size of N increases, both the numerator and denominator increase at roughly the same rate. As N --> ∞, s^2 --> σ^2, the variance of the population.

In general, the true mean μ is not known, so the above definition of s^2 needs to be replaced by

$$var(x) = s^2 = \sum (x_i - \langle x \rangle)^2 / (N - 1).$$

One measurement of a quantity (N = 1) does not allow us to estimate the spread of values, if the true value of the mean μ is not known. For large enough N, the factor (N-1) in the denominator can be safely replaced by N.

6 Continuous Probability Distributions

The Normal (μ, σ) distribution is an example of a continuous probability distribution, often usable to describe the distribution of a continuous random variable.

Quoting [RS] Sec. 3.5,
"A continuous random variable x is one that can assume an infinite number of values within any given interval.The probability that x falls within any interval is given by the area under the probability distribution (or 'density function') within that interval. The total area under the curve is 1."

If x is a continuous variable, and if P(x1) is the probability a measurement results in x lying in the interval [x1, x1 + dx], the normalization requirement for the probability function P(x) is that: ∫P(x) dx = 1, (integrating from xmin to xmax), or (in Maxima syntax):

integrate (P(x), x, xmin, xmax) = 1,

where the smallest and largest possible values of x are xmin and xmax, respectively. We then say P(x) is a 'normalized continuous probability density'.

The mean value of x (the expectation value of x, E(x)) is then predicted to be
E (x) = <x> = ∫ x P(x) dx = integrate (x*P(x), x, xmin, xmax).

The probability x willl be found in the interval [a,b] is
integrate (x*P(x), x, a, b).

7 The Normal (Gaussian) Distribution, pdf_normal (x, μ, σ)

The Maxima function pdf_normal (x, μ, σ) returns a symbolic value for the value of the Normal (μ, σ) probability distribution at position x. μ is the mean = <x> = E(x), σ is the value of one standard deviation about the mean.

Here we use an example of the Normal (5, 2) distribution (μ = mean = 5, σ = std = 2):

(%i11) pdf_normal (3, 5, 2);

(%o11) $$\frac{1}{2^{3/2}\sqrt{\%e}\sqrt{\pi}}$$

(%i12) pdf_normal (3, 5, 2), numer;

(%o12) 0.120985

(%i13) float (pdf_normal (3, 5, 2));

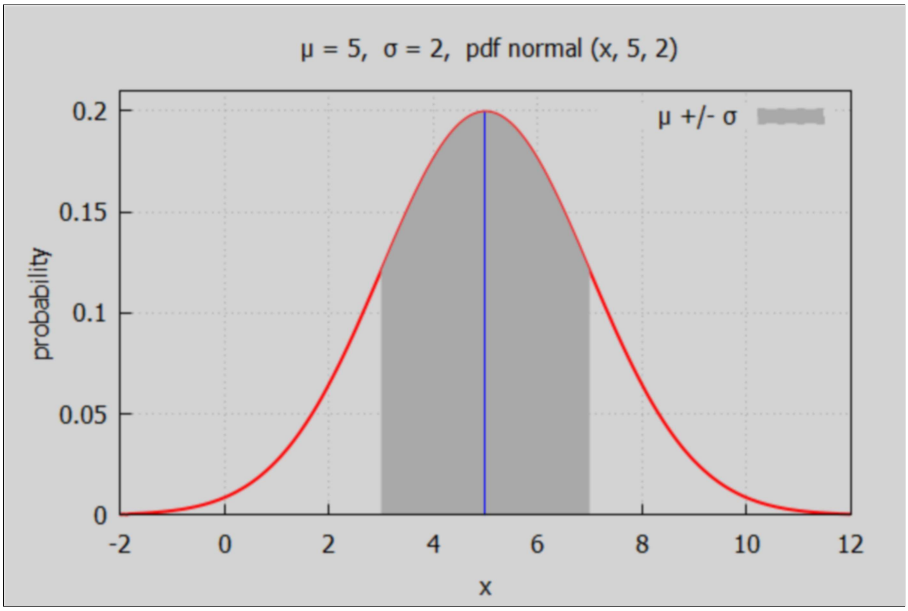
(%o13) 0.120985

(%i14) pmax : pdf_normal (5, 5, 2), numer;

(pmax) 0.199471

(%i15)
```
wxdraw2d (xrange = [-2, 12], yrange = [0, 0.21],
    xlabel = "x", ylabel = "probability", grid = true,
    title = " μ = 5,  σ = 2,  pdf normal (x, 5, 2)",
    background_color = light_gray, color = red, line_width = 2,
    explicit (pdf_normal (x, 5, 2), x, -2, 12),
    filled_func = true, fill_color = dark_gray,  key = " μ +/- σ ",
    explicit ( pdf_normal (x, 5, 2), x, 3, 7),
    line_width = 1, key = "",  color = blue, parametric (5, yy, yy, 0, pmax) )$
```

(%t15)



The most commonly used continuous probability distribution is the Normal distribution, called the Gaussian distribution by physicists.

Why do normal distributions matter?

All kinds of variables in the natural and social sciences are normally or approximately normally distributed. Height, birth weight, reading ability, job satisfaction, or SAT scores are just a few examples of such variables.

Because normally distributed variables are so common, many statistical tests are designed for normally distributed populations.

Understanding the properties of normal distributions means you can use inferential statistics to compare different groups and make estimates about populations using samples.

Quoting Fred Senese in: worksheet-continuous-distributions.pdf (see book preface),

"When continuous experimental data contains only small and purely random errors, the distribution for the measurements can often be approximated by the normal distribution."

quoting
https://www.varsitytutors.com/ap_statistics-help/
how-to-identify-characteristics-of-a-normal-distribution

"The two main parameters of the normal distribution are μ and σ. μ is a location parameter which determines the location of the peak of the normal distribution on the real number line. σ is a scale parameter which determines the concentration of the density around the mean. Larger σ's lead the normal [distribution] to spread out more than smaller σ's"

7.1 Empirical Rule

The fraction of the area under the **Standard** Normal (0, 1) curve can be calculated using Maxima's function pdf_normal (x, 0, 1) which describes a bell shaped curve with mean (average) at x = 0, and with one standard deviation σ = 1. In the following, the j = 1 case finds the fraction of the area within 1 standard deviation of the mean (0) ie., in the range [-1, 1]. The j = 2 case finds the fraction of the area within 2 standard deviations of the mean [-2, 2]. The j = 3 case corresponds to [-3, 3], three standard deviations.

7.1.1 dpc [j]

```
(%i16) for j thru 3 do (
         dpc[j] : float (integrate (pdf_normal (x, 0, 1), x, -j, j) ),
         print (" j  = ", j, "  dpc = ", dpc[j]) ) $
 j  =  1   dpc =  0.682689
 j  =  2   dpc =  0.9545
 j  =  3   dpc =  0.9973

(%i17) dpc[1];
(%o17) 0.682689
```

We show below that we can extrapolate these probability numbers to the general case Normal (μ, σ).

Roughly 68% of the area under the Normal (μ, σ) curve lies in the region [μ - σ, μ + σ], ie., μ - σ <= x <= μ + σ.

Roughly 95% of the area under the Normal (μ, σ) curve lies in the region [μ - 2 σ, μ + 2 σ], ie., (μ - 2 σ) <= x <= (μ + 2 σ).

Roughly 99.7% of the area under the Normal (μ, σ) curve lies in the region [μ - 3 σ, μ + 3 σ], ie., (μ - 3 σ) <= x <= (μ + 3 σ).

Here are some general comments from the webpage:

https://www.scribbr.com/statistics/normal-distribution/

"Empirical rule:

The empirical rule, or the 68-95-99.7 rule, tells you where most of your values lie in a normal distribution:

Around 68% of values are within 1 standard deviation from the mean.

Around 95% of values are within 2 standard deviations from the mean.

Around 99.7% of values are within 3 standard deviations from the mean"

"With μ the mean, σ the standard deviation, and σ^2 the variance, the Normal distribution has the mathematical form:

P(x, μ, σ) = (1/(σ*sqrt(2*π)) * exp (- (1/2) * ((x - μ)/σ)^2).

The normal curve is bell-shaped and symmetrical about its mean μ. It extends indefinitely in both directions, but most of the area is clustered around the mean μ.
68.27% of the area (probability) under the normal curve in included within one standard deviation the mean (i.e., within μ +/- 1 σ),
95.45% within two standard deviations of the mean (μ +/- 2 σ),
and 99.78% within three standard deviations of the mean (μ +/- 3 σ)."

7.2 Empirical Rule Problem 1

From the webpage:
https://www.scribbr.com/statistics/normal-distribution/

"You collect SAT scores from students in a new test preparation course. The data follows a normal distribution with a mean score (M) of 1150 and a standard deviation (SD) of 150.

Following the empirical rule:

Around 68% of scores are between 1,000 and 1,300, 1 standard deviation above and below the mean.
Around 95% of scores are between 850 and 1,450, 2 standard deviations above and below the mean.
Around 99.7% of scores are between 700 and 1,600, 3 standard deviations above and below the mean."

"The empirical rule is a quick way to get an overview of your data and check for any outliers or extreme values that don't follow this pattern.

If data from small samples do not closely follow this pattern, then other distributions like the t-distribution may be more appropriate. Once you identify the distribution of your variable, you can apply appropriate statistical tests."

7.3 Empirical Rule Problem 2

"Consider a normal distribution with a mean of 100 and a standard deviation of 5. Which of the following statements are true according to the Empirical Rule?

1. 84% of observations are at least 95.
2. 95% of observations are between 90 and 110.
3. 99.7% of observations are between 85 and 115.

Correct answer:
1, 2 and 3

Explanation:
2) and 3) are true by definition of the Empirical Rule - also known as the 68-95-99.7 Rule. Using our information with mean of 100 and a standard deviation of 5 we can create a bell curve with 100 in the middle. One standard deviation out from the mean would give us a range from 95 to 105 and would be in our 68% section. If we go two standard deviations out from 100 we would get the range 90 to 110 thus lying in the 95% section. Lastly, when we go out 3 standard deviations we get a range of 85 to 115 thus falling within the 99.7% section."

To show that 1) is true, since 68% of the values lie in the central region μ +/- σ within one standard deviation 95 <= x <= 105, 100% - 68% = 32% of the values lie in the tails, and 16% lie in the left-hand tail with x <= 95, and (because of symmetry) 16% lie in the right-hand tail x >= 105.

The percent of the values which lie in the region >= 95 is the sum of the 68% in the central 1σ region and the values which lie in the right-hand tail: (68 + 16)% = 84%.

Later we will discuss the Maxima function cdf_normal (x1, μ, σ) which is the integral of the probabilities pdf_normal (x, μ, σ) over the region [-∞, x1], which gives the probability of finding a value <= x1. Then you can use 1 - cdf_normal (x1, μ, σ) to find the probability of finding a value >= x1.

```
(%i19) cdf_normal (95, 100, 5), numer;
       1 - %;
(%o18) 0.158655
(%o19) 0.841345
```

This method is more exact, since the Rule: 68, 95, 99.7, is not exact:

```
(%i20) for j thru 3 do print (j, dpc[j])$
1  0.682689
2  0.9545
3  0.9973
```

7.4 **Standard** Normal Distribution: μ = 0, σ = 1

The 'standard normal distribution' is a normal distribution with μ = 0 and σ = 1. Let's plot the standard normal distribution over the interval [-4, 4].

```
(%i21) P(x) := (1/sqrt (2*%pi)) * exp (- x^2/2);
```

$$(\%o21)\quad P(x) := \frac{1}{\sqrt{2\,\pi}}\,\exp\left(\frac{-x^2}{2}\right)$$

```
(%i22) P(0), numer;
(%o22) 0.398942
```

From webpage:
https://www.scribbr.com/statistics/normal-distribution/

"The standard normal distribution, also called the z-distribution, is a special normal distribution where the mean is 0 and the standard deviation is 1."

"Every normal distribution is a version of the standard normal distribution that's been stretched or squeezed and moved horizontally right or left."

"While individual observations from normal distributions are referred to as x, they are referred to as z in the z-distribution. Every normal distribution can be converted to the standard normal distribution by turning the individual x- values into z-values."

"z-values tell you how many standard deviations away from the mean each x-value lies."

"You only need to know the mean and standard deviation of your distribution to find the z-value of a x-value."

$$z = (x - \mu)/\sigma$$

"We convert normal distributions into the standard normal distribution for several reasons:

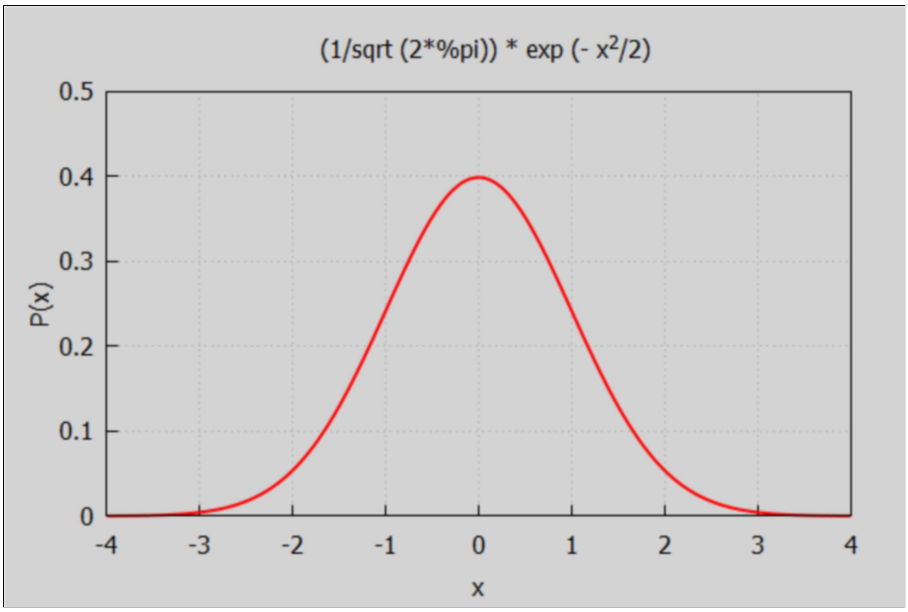
To find the probability of observations in a distribution falling above or below a given value.
To find the probability that a sample mean significantly differs from a known population mean.
To compare scores on different distributions with different means and standard deviations."

7.5 **Standard** Normal Distribution plot using P(x)

P(x) is an explicit formula for the standard normal distribution, defined above.

(%i23)
```
wxdraw2d (xrange = [-4, 4], yrange = [0, 0.5], xlabel = "x", ylabel = "P(x)", grid = true,
    title = "  (1/sqrt (2*%pi)) * exp (- x^2/2)", background_color = light_gray,
    color = red, line_width = 2,  explicit (P(x), x, -4, 4))$
```

(%t23)

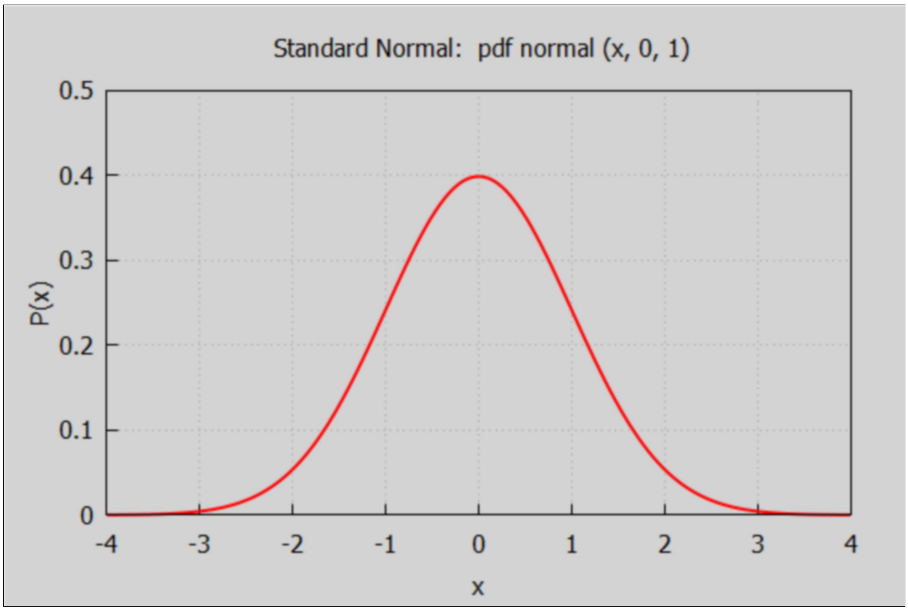


7.6 **Standard** Normal Distribution Plot using pdf_normal (x, 0, 1)

We could have used the Maxima function pdf_normal (x, 0, 1) instead of P(x) to get the same curve.

(%i24)
```
wxdraw2d (xrange = [-4, 4], yrange = [0, 0.5], xlabel = "x", ylabel = "P(x)", grid = true,
    title = " Standard Normal:  pdf normal (x, 0, 1)",
    background_color = light_gray, color = red, line_width = 2,
    explicit (pdf_normal (x, 0, 1), x, -4, 4))$
```

(%t24)



P(x)*dx is the probability of finding an x with value in the range [x, x + dx].
The sum of the numbers P(x)*dx from x = a to x = b is the probability of finding a value of x in the range [a, b].

Integrating (summing) from negative infinity to positive infinity, we get the total area under the curve.

(%i25) integrate (P(x), x, minf, inf);

(%o25) 1

Since the area under the whole curve is unity (1), the area under the curve from x = -1 to x = +1 gives the probability [p-value] of finding a value for x in the range [-1, 1]. We then replace 1 by 2 and then 3 (standard deviations from the mean (0)), etc.

(%i26) for j thru 5 do print (j, float (integrate (P(x), x, -j, j)))$

1 0.682689
2 0.9545
3 0.9973
4 0.999937
5 0.999999

The probability of finding a value of x in the region -5 σ <= x <= 5 σ is

(%i27) p5σ : integrate (P(x), x, -5, 5), numer;

(p5σ) 0.999999

To see all the digits used by Maxima here, set fpprintprec to either 16 or 0.

(%i29) fpprintprec : 0$
p5σ;

(%o29) 0.9999994266968537

The probability of finding a value is the regions x < - 5σ and x > 5σ (|x| >= 5 σ) is:

(%i30) 1 - p5σ;

(%o30) $5.73303146289561\ 10^{-7}$

The probability that x >= 5 σ is:

(%i31) %/2;

(%o31) $2.866515731447805\ 10^{-7}$

Thus the p-value that x >= 5 σ is about 3x10^(-7). To convert to per cent chance, multiply the p-value by 100%.

The chance that a value of x will be found such that x >= 5 σ:

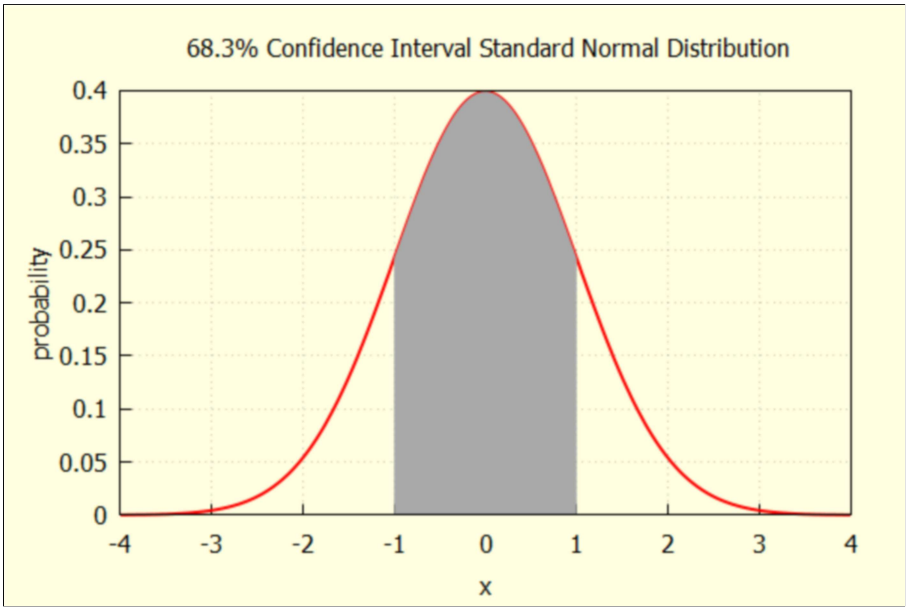
(%i32)
```
100*%;
```
(%o32) $2.866515731447805\ 10^{-5}$

Thus, the chance that a value of x will be found such that x >= 5 σ is roughly 3x10^(-5)% = 0.00003%.

Instead of using the explicit expression P(x), we can use the numerical function pdf_normal (x, 0, 1), and integrate over the same x intervals, getting the same numbers:

(%i34)
```
fpprintprec : 5$
for j thru 3 do print (j, float (integrate (pdf_normal (x, 0, 1), x, -j, j)) )$
```
1 0.68269
2 0.9545
3 0.9973

(%i35)
```
wxdraw2d (xrange = [-4, 4], yrange = [0, 0.4], xlabel = "x",
   ylabel = "probability", grid = true,
   title = " 68.3% Confidence Interval Standard Normal Distribution", line_width = 2,
    background_color = light_yellow,
   color = red,  explicit (pdf_normal (x, 0, 1), x, -4, 4),
    fill_color = dark_gray, filled_func = true, explicit (pdf_normal (x,0,1), x, -1, 1) )$
```
(%t35)



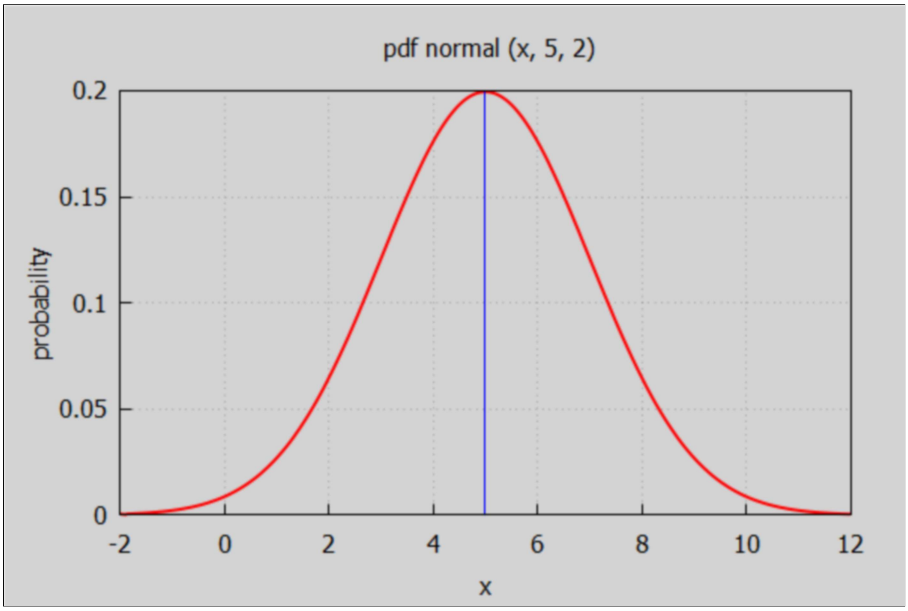
The size of one standard deviation gives us a '68% confidence interval', "that is, we can be 68.3% certain that a measurement of x will fall within one standard deviation of the mean", the shaded region in the above figure. Generalizing to a Normal distribution with mean μ and standard deviation σ, the probability that a trial measurement of x will fall in one of the tails x < (μ - σ), x > (μ + σ) is (100 - 68.3) = 31.7%.

Remembering that the Normal distribution is perfectly symmetric, we can then say the probability that a trial measurement of x will fall in the range x <= (μ - σ) is 15.85%, and the probability that a trial measurement of x will fall in the range x >= (μ + σ) is 15.85%

We can guess that for any normal distribution (having any mean and standard deviation) the same fractions of the total area under the curve will result as for the "standard normal distribution". Rather than using a change of integration variables to get a formal proof, we just use a numerical example in which μ = 5 and σ = 2. The size of σ is a measure of the spreading of the distribution about the mean of a series of measurements of the continuous random variable x.

(%i36)
```
wxdraw2d (xrange = [-2, 12], yrange = [0, 0.2], xlabel = "x",
    ylabel = "probability", grid = true,
    title = " pdf normal (x, 5, 2)", background_color = light_gray,
    color = red, line_width = 2,
    explicit (pdf_normal (x, 5, 2), x, -2, 12),
    color = blue, line_width = 1, parametric (5, y, y, 0, 0.2))$
```

(%t36)



The area under the whole curve from negative infinity to positive infinity is 1:

(%i37)
```
integrate (pdf_normal (x, 5, 2), x, minf, inf), numer;
```

(%o37) 1.0

The area under the curve in the range [μ - σ, μ + σ] = [5 - 2, 5 + 2] = [3, 7] is

(%i38) integrate (pdf_normal (x, 5, 2), x, 3, 7), numer;
(%o38) 0.68269

which is the same fraction of the area we found using the standard normal distribution.

The area under the curve in the range [μ - 2 σ, μ + 2 σ] = [5 - 4, 5 + 4] = [1, 9] is

(%i39) integrate (pdf_normal (x, 5, 2), x, 1, 9), numer;
(%o39) 0.9545

Again the same as when using the standard normal distribution.

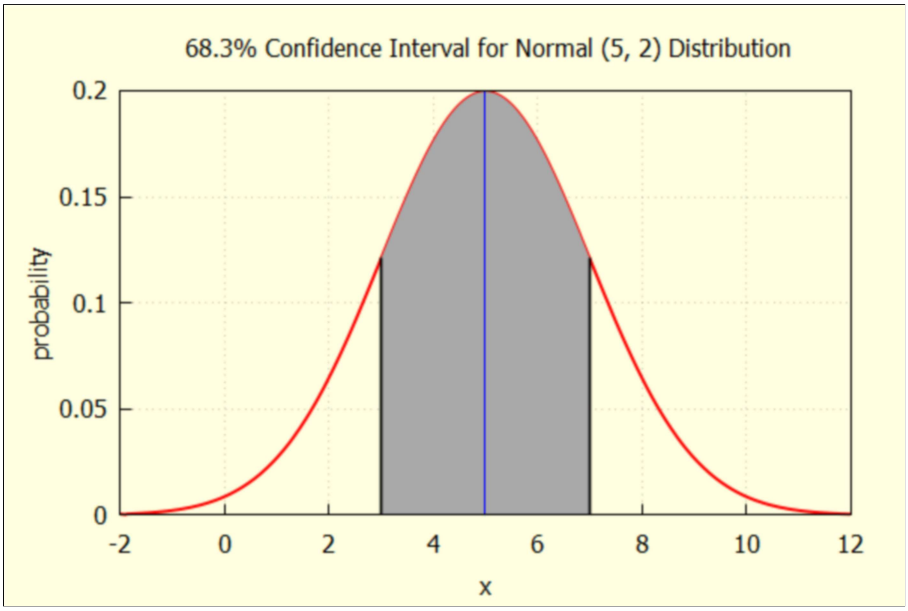
The area under the curve in the range [μ - 3 σ, μ + 3 σ] = [5 - 6, 5 + 6] = [-1, 11] is

(%i40) integrate (pdf_normal (x, 5, 2), x, -1, 11), numer;
(%o40) 0.9973

Again the same as when using the standard normal distribution.

(%i41)
```
wxdraw2d (xrange = [-2, 12], yrange = [0, 0.2], xlabel = "x",
    ylabel = "probability", grid = true,
    title = " 68.3% Confidence Interval for Normal (5, 2) Distribution",
    line_width = 2,  background_color = light_yellow,  color = red,
    explicit (pdf_normal (x, 5, 2), x, -2, 12), fill_color = dark_gray,
    filled_func = true, explicit (pdf_normal (x, 5, 2), x, 3, 7),
    color = blue, line_width = 1, parametric (5, y, y, 0, 0.2), line_width = 1.7,
    color = black, parametric (3, y, y, 0, pdf_normal (3, 5, 2)),
    parametric (7, y, y, 0, pdf_normal (7, 5, 2)) )$
```

(%t41)



The size of one standard deviation gives us a '68% confidence interval', "that is, we can be 68.3% certain that a measurement of x will fall within one standard deviation of the mean", the shaded region in the above figure. The chance that a trial measurement of x will fall in one of the tails x < (μ - σ), x > (μ + σ), is (100 - 68.3) = 31.7%.

Remembering that the Normal distribution is perfectly symmetric, we can then say the chance that a trial measurement of x will fall in the left tail x <= (μ - σ) is 15.85%, and the chance that a trial measurement of x will fall in the right tail x >= (μ + σ) is 15.85%

7.7 [RS] Example 13 and 14 (p. 41)

13) Suppose x is a normally distributed random variable with μ = 10 and σ^2 = 4. What is the chance of finding x in the interval [8, 12]?

Answer: σ = 2, so the interval [8,12] is [μ - σ, μ + σ], which is the 68.3% confidence interval; there is a 68.3% chance we will find x in this interval.

```
(%i42) integrate (pdf_normal (x, 10, 2), x, 8, 12), numer;
(%o42) 0.68269
```

14) What is the probability of finding x in the interval [7, 14]?

```
(%i43) integrate (pdf_normal (x, 10, 2), x, 7, 14), numer;
(%o43) 0.91044
```

So 91.04% chance x will be found in [7, 14].

7.7.1 Cumulative probability distribution cdf_normal (x1, μ, σ)

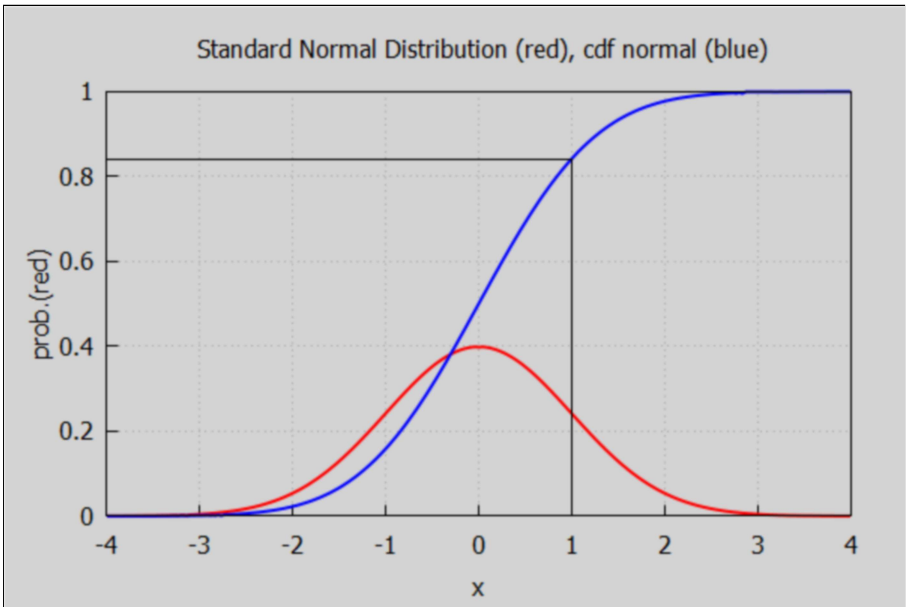
cdf_normal (x1, μ, σ) gives the probability that the value of x measured lies in the interval [- ∞, x1], given that the values measured are described by the normal distribution with mean μ and standard deviation σ.

We can put cdf_normal (x, 0, 1) on the plot (in blue) to show graphically that the area under the red standard normal distribution curve pdf_normal from x = - ∞ to x = 1 is 0.841. (0.16 from left-hand tail plus 0.68 from central 1σ region adds to 0.84.)

```
(%i44) cdf1 : cdf_normal (1, 0, 1), numer;
(cdf1) 0.84134
```

```
(%i45) wxdraw2d (xrange = [-4, 4], yrange = [0, 1], xlabel = "x",
    ylabel = "prob.(red)", grid = true,
    title = " Standard Normal Distribution (red), cdf normal (blue)",
    background_color = light_gray, color = red, line_width = 2,
    explicit (pdf_normal (x, 0, 1), x, -4, 4), color = blue,
    explicit (cdf_normal (x, 0, 1), x, -4, 4), color = black,
    line_width = 1, parametric (1,y,y,0,cdf1), explicit (cdf1, x, -4, 1))$
```

(%t45)



"... we could have computed the probability that [a measured value of] x will be within one standard deviation of the mean more simply (without explicit integration) as follows"

```
(%i46) cdf_normal (1, 0, 1) - cdf_normal (-1, 0, 1), numer;
(%o46) 0.68269
```

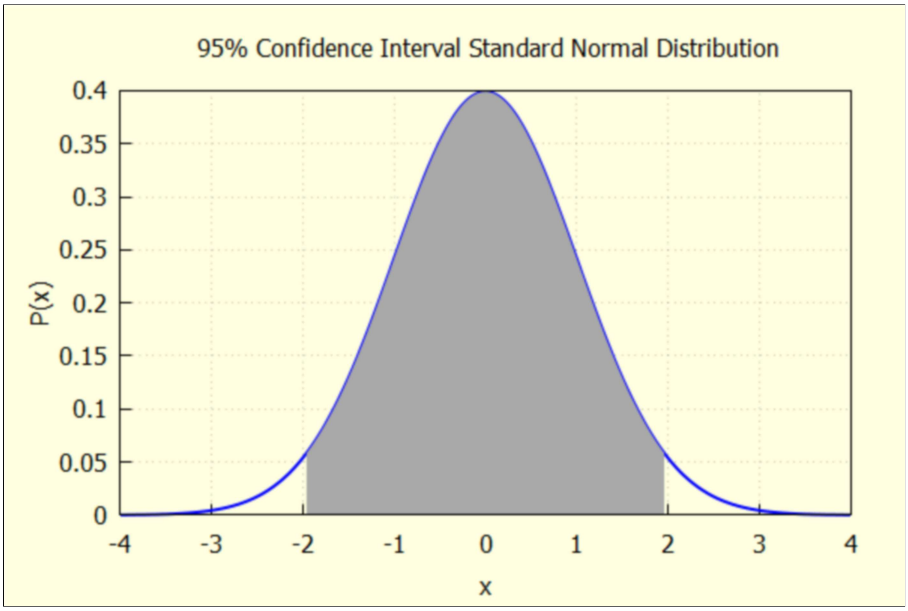
The size of one standard deviation gives us a 68% confidence interval, "that is, we can be 68% certain that a measurement of x will fall within one standard deviation of the mean."

"The 95% 'confidence interval' for the standard normal distribution with mean zero and standard deviation 1 is [- 1.96, 1.96]."

```
(%i47) cdf_normal (1.96, 0, 1) - cdf_normal (-1.96, 0, 1), numer;
(%o47) 0.95
```

```
(%i48) wxdraw2d (xrange = [-4, 4], yrange = [0, 0.4], xlabel = "x", ylabel = "P(x)",
    grid = true, title = " 95% Confidence Interval Standard Normal Distribution",
    line_width = 2,  background_color = light_yellow,
    explicit (pdf_normal (x, 0, 1), x, -4, 4), fill_color = dark_gray,
    filled_func = true, explicit (pdf_normal (x,0,1), x, -1.96, 1.96) )$
```

(%t48)



Generalizing to a Normal distribution with mean μ and standard deviation σ, and remembering that the Normal distribution is perfectly symmetric, we can then say that the probability that a trial measurement of x will fall in the range x <= (μ -1.96 σ) is 2.5%, and the probability that a trial measurement of x will fall in the range x >= (μ + 1.96 σ) is 2.5%.

7.8 quantile_normal (p, μ, σ)

The Maxima function quantile_normal (p, μ, σ) returns the value of c for which x <= c with probability p, assuming x are drawn from a random normal distribution with mean μ and standard deviation σ.

The value of c for which x <= c with chance 2.5%, assuming x are drawn from a STANDARD normal distribution, is then

```
(%i49) quantile_normal (0.025, 0, 1), numer;
(%o49) −1.96
```

We can therefore be 95% confident that x falls within the interval [-1.96, 1.96] when μ = 0 and σ = 1.

7.8.1 confidence (q, m, s)

The Maxima function confidence (q, m, s) has been defined near the top of the worksheet.

With q a number in the interval 0 < q < 1, and with m the mean and s the standard deviation of a Normal distribution, confidence (q, m, s) prints out the values dx, m - dx, m + dx, and outputs a list [m - dx, m + dx] which allows one to have 100*q % confidence a random value of x will lie within that interval, ie., within m +/- dx.

Consider, for example, the **standard** normal distribtion with μ = 0 and σ = 1; what is the 95% confidence interval dx such that we can be 95% confident that a random x value will be found in the range m +/- dx?

(%i50) confidence (0.95, 0, 1);

delx = 1.96

x1 = −1.96 , *x2* = 1.96

(%o50) [−1.96,1.96]

Thus we can be 95% confident that a random x value from a Normal (0,1) distribution will be found in the range x1 <= x <= x2, or -1.96 <= x <= 1.96.

Note that you can get results printed out on the screen without the final list, by ending with a $.

(%i51) confidence (0.95, 5, 2)$

delx = 3.9199

x1 = 1.0801 , *x2* = 8.9199

Thus we can be 95% confident that a random x value from a Normal (5, 2) distribution will be found in the range x1 <= x <= x2, or 1.08 <= x <= 8.92, or roughly 1 <= x <= 9; ie., within 2 standard deviations of the mean.

On the other hand, if you want to use the exact values of x1 and x2 in another calculation, you should end the command with a semicolon (;) rather than the dollar sign ($), and then refer to %[1] and %[2] to retrieve the computed values of x1 and x2 in the next calculation, or else give those values names, as in x1 : %[1], and x2 : %[2] for use in later calculations.

(%i52) confidence (0.95, 5, 2);

delx = 3.9199

x1 = 1.0801 , *x2* = 8.9199

(%o52) [1.0801,8.9199]

(%i53) [x1, x2] : [%[1], %[2]];

(%o53) [1.0801,8.9199]

An alternative way to use this function is to leave the dollar sign at the end of the command, but define the output list as [x1, x2] (for example):

```
(%i54) [x1, x2 ] : confidence (0.68, 0, 1)$
delx =  0.99446
 x1 =  −0.99446 ,  x2 =  0.99446

(%i55) [x1, x2];
(%o55) [−0.99446,0.99446]
```

Yet a third way to use this function is to again leave the dollar sign, and in the following command use [x1, x2] : %. Even though the final list is not evident, it can still be retrieved using %.

```
(%i56) confidence (0.68, 0, 1)$
delx =  0.99446
 x1 =  −0.99446 ,  x2 =  0.99446

(%i57) [x1, x2] : %;
(%o57) [−0.99446,0.99446]
```

7.9 Statology Ex. 1 Birthweight of Babies

"It's well-documented that the birthweight of newborn babies is normally distributed with a mean of about 7.5 pounds."

7.10 Statology Ex. 2 Height of Males

"The distribution of the height of males in the U.S. is roughly normally distributed with a mean of 70 inches and a standard deviation of 3 inches."

7.11 Statology Ex. 3 Shoe Sizes

"The distribution of shoe sizes for males in the U.S. is roughly normally distributed with a mean of size 10 and a standard deviation of 1."

7.12 Statology Ex. 4 ACT Scores

"The distribution of ACT scores for high school students in the U.S. is normally distributed with a mean of 21 and a standard deviation of about 5."

7.13 Statology Ex. 5 Average NFL Player Retirement Age

"The distribution of retirement age for NFL players is normally distributed with a mean of 33 years old and a standard deviation of about 2 years."

7.14 Statology Ex. 6 Blood Pressure

"The distribution of diastolic blood pressure for men is normally distributed with a mean of about 80 and a standard deviation of 20."

"The systolic blood pressures of adults, in the appropriate units, are normally distributed with a mean of 128.4 and a standard deviation of 19.6."

"Normal resting blood pressure in an adult is approximately 120 millimetres of mercury (16 kPa) systolic over 80 millimetres of mercury (11 kPa) diastolic."

See also:

https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/
BS704_BiostatisticsBasics/BS704_BiostatisticsBasics3.html

Basic Concepts for Biostatistics
Sample Statistics

Example of calculation of standard deviation from blood pressure data.

https://jhanley.biostat.mcgill.ca/bios601/AustinBradfordHillMedStatistics/BradfordHill-08.pdf

7.15 random_normal (m, s), random_normal (m, s, n)

The Maxima function random_normal (m, s) returns a Normal (m, s) random variate.
The parameter m is the requested average (mean) value of the distribution.
The parameter s is the requested standard deviation of the distribution.

```
(%i58) for j thru 5 do print (j, random_normal (5, 2))$
1  5.2619
2  7.0541
3  7.7588
4  4.7463
5  7.6203
```

The Maxima function random_normal (m, s, n) returns a list of n Normal distributed random integers, where m is the average (mean) of the distribution and s is one standard deviation.

7.15.1 Random Sample Size n = 10 Simulations, m = 5, s = 2

```
(%i59) rsample : random_normal (5, 2, 10);
(rsample) [5.1619,3.7268,6.0031,3.621,5.6681,3.7387,6.2303,5.5042,5.7958
,5.0287]
```

7.15.2 Random Sample Size n = 100 Simulations, m = 5, s = 2

```
(%i63) rsample : random_normal (5, 2, 100)$
fll (rsample);
head (rsample);
tail (rsample);
(%o61) [6.8148,6.1833,100]
(%o62) [6.8148,6.9837,3.4078]
(%o63) [5.7149,4.3547,6.1833]

(%i64) length (rsample);
(%o64) 100

(%i65) [lmin (rsample), lmax (rsample)];
(%o65) [-0.40486,10.934]
```

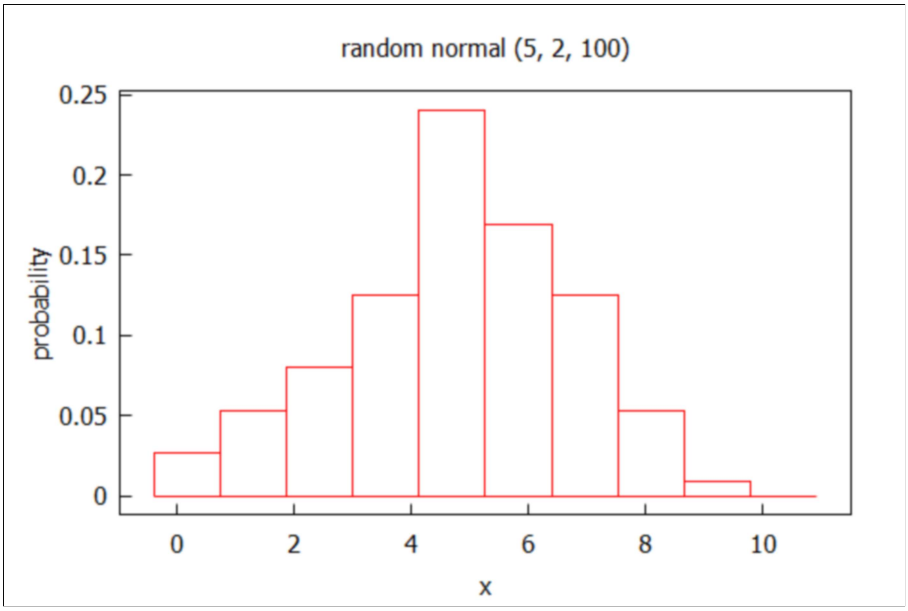
Find the properties of this set of 100 random floating point numbers drawn from the Normal (5,2) distribution.

```
(%i66) mean (rsample);
(%o66) 4.7997

(%i67) std (rsample);
(%o67) 2.0448

(%i68) fracData (rsample, 3, 7);
(%o68) 0.7

(%i69) fracData (rsample, 1, 9);
(%o69) 0.93

(%i70) fracData (rsample, -1, 11);
(%o70) 1.0
```

The default number of bins used by wxhistogram is 10.

(%i71)
```
wxhistogram (rsample, xlabel = "x", ylabel = "probability",
  title = " random normal (5, 2, 100) ",
  frequency = density)$
```

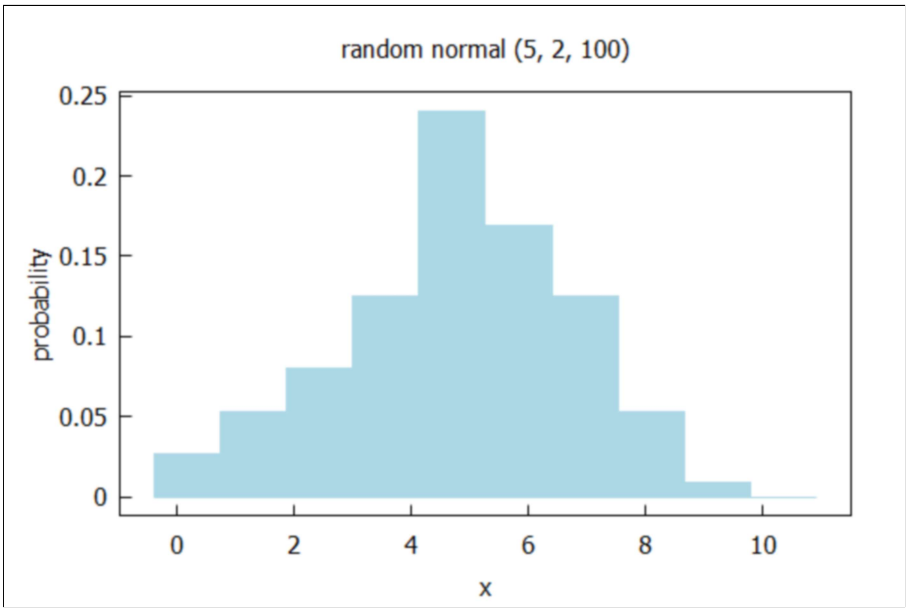
(%t71)



We can choose our own color style. (We can also change the number of bins to use with the nclasses = num option.)

(%i72)
```
wxhistogram (rsample, xlabel = "x", ylabel = "probability",
  title = " random normal (5, 2, 100) ", fill_color = light_blue,
  fill_density = 1, frequency = density)$
```

(%t72)



7.15.3 Random Sample Size m = 1000 Simulations, m = 5, s = 2

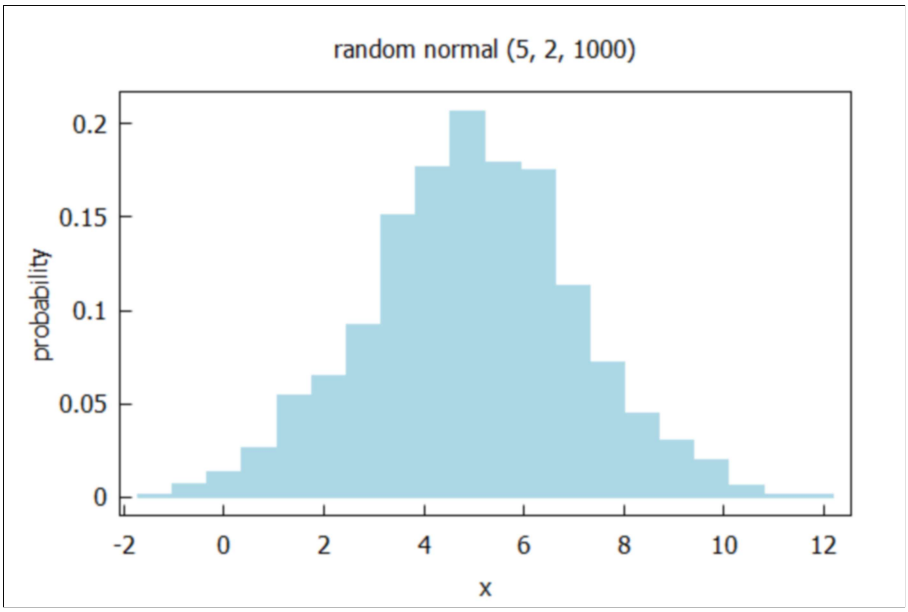
```
(%i76) rsample : random_normal (5, 2, 1000)$
       fll (rsample);
       head (rsample);
       tail (rsample);
(%o74) [5.2273,6.3679,1000]
(%o75) [5.2273,5.4091,6.9232]
(%o76) [7.0576,5.9009,6.3679]

(%i77) length (rsample);
(%o77) 1000

(%i78) [lmin (rsample), lmax (rsample)];
(%o78) [-1.7261,12.183]
```

Find the properties of this set of 1000 random floating point numbers drawn from the Normal (5,2) distribution.

```
(%i79) mean (rsample);
(%o79) 4.9506

(%i80) std (rsample);
(%o80) 2.0771

(%i81) fracData (rsample, 3, 7);
(%o81) 0.68

(%i82) fracData (rsample, 1, 9);
(%o82) 0.943

(%i83) fracData (rsample, -1, 11);
(%o83) 0.997
```

(%i84)
```
wxhistogram (rsample, xlabel = "x", ylabel = "probability",
    title = " random normal (5, 2, 1000) ", fill_color = light_blue,
    fill_density = 1, frequency = density, nclasses = 20)$
```

(%t84)



8 Statistical Significance of a Certain Sigma in Physics

Quoting The Perimeter Institute webpage "What is a Sigma?".

"Looking for a gravitational wave is like trying to hear a single bee in a hive. It is easy for the single buzz – the gravitational wave – to get lost in the overall hum of the hive – the noise of the universe. It's also easy to mistake another sound, such as someone starting a chainsaw nearby, for the signal."

"Statistical analysis is our most powerful tool for avoiding both of these unhappy outcomes and initial reports on new discoveries in physics are, therefore, often expressed in terms of their statistical significance. Saying something is statistically significant is the same as saying it's unlikely to have occurred by chance."

"In physics, statistical significance is usually expressed in units of the standard deviation, or σ (sigma), from the average value."

"For instance, consider an experiment where one flips a coin 100 times. The expected outcome is 50 heads and the standard deviation of such an experiment is five. If you get 55 heads, that's a one sigma effect. If you get 60 heads, that's a two sigma effect; 65 is three sigma; 70 is four sigma; 75 is five sigma, and so on. "

"Your intuition probably tells you that 55 is probably just chance, 60 is odd, 65 is startling, and more than that means there's something up with the coin. So it is in physics."

"Five-sigma corresponds to a p-value, or probability, of 3x10^(-7), or about one in 3.5 million. That is, there's less than one chance in 3.5 million that the effect being seen is due to random chance."

"The number of sigma we assign to an effect expresses how much confidence we have that the signal is not the result of random chance – that we have not mistaken a chainsaw for a bee. A five sigma effect is the gold standard of proof of a new discovery in physics."

In the following, we are quoting https://news.mit.edu/2012/explained-sigma-0209:

"Explained: Sigma

How do you know when a new finding is significant?
The sigma value can tell you — but watch out for dead fish."

David L. Chandler, MIT News Office
Publication Date:February 9, 2012

"It’s a question that arises with virtually every major new finding in science or medicine: What makes a result reliable enough to be taken seriously? The answer has to do with statistical significance — but also with judgments about what standards make sense in a given situation."

The unit of measurement usually given when talking about statistical significance is the standard deviation, expressed with the lowercase Greek letter sigma (σ). The term refers to the amount of variability in a given set of data: whether the data points are all clustered together, or very spread out."

"In many situations, the results of an experiment follow what is called a “normal distribution. For example, if you flip a coin 100 times and count how many times it comes up heads, the average result will be 50. But if you do this test 100 times, most of the results will be close to 50, but not exactly. You’ll get almost as many cases with 49, or 51. You’ll get quite a few 45s or 55s, but almost no 20s or 80s. If you plot your 100 tests on a graph, you’ll get a well-known shape called a bell curve that’s highest in the middle and tapers off on either side.
That is a normal distribution."

"The deviation is how far a given data point is from the average. In the coin example, a result of 47 has a deviation of three from the average (or “mean”) value of 50. The standard deviation is just the square root of the average of all the squared deviations. One standard deviation, or one sigma, plotted above or below the average value on that normal distribution curve, would define a region that includes 68 percent of all the data points. Two sigmas above or below would include about 95 percent of the data, and three sigmas would include 99.7 percent."

"So, when is a particular data point — or research result — considered significant? The standard deviation can provide a yardstick: If a data point is a few standard deviations away from the model being tested, this is strong evidence that the data point is not consistent with that model.

However, how to use this yardstick depends on the situation. John Tsitsiklis, the Clarence J. Lebel Professor of Electrical Engineering at MIT, who teaches the course Fundamentals of Probability, says, 'Statistics is an art, with a lot of room for creativity and mistakes.' Part of the art comes down to deciding what measures make sense for a given setting."

"For example, if you’re taking a poll on how people plan to vote in an election, the accepted convention is that two standard deviations above or below the average, which gives a 95 percent confidence level, is reasonable. That two-sigma interval is what pollsters mean when they state the “margin of sampling error,” such as 3 percent, in their findings."

"That means if you asked an entire population a survey question and got a certain answer, and then asked the same question to a random group of 1,000 people, there is a 95 percent chance that the second group's results would fall within two-sigma from the first result. If a poll found that 55 percent [plus or minus 3 percent] of the entire population favors candidate A, then 95 percent of the time, a second poll's result would be somewhere between 52 and 58 percent."

"Of course, that also means that 5 percent of the time, the result would be outside the two-sigma range. That much uncertainty is fine for an opinion poll, but maybe not for the result of a crucial experiment challenging scientists' understanding of an important phenomenon — such as last fall's announcement of a possible detection of neutrinos moving faster than the speed of light in an experiment at the European Center for Nuclear Research, known as CERN."

"Six sigmas can still be wrong"

"Technically, the results of that experiment had a very high level of confidence: six sigma. In most cases, a five-sigma result is considered the gold standard for significance, corresponding to about a one-in-a-million chance that the findings are just a result of random variations; six sigma translates to one chance in a half-billion that the result is a random fluke. (A popular business-management strategy called "Six Sigma" derives from this term, and is based on instituting rigorous quality-control procedures to reduce waste.)"

"But in that CERN experiment, which had the potential to overturn a century's worth of accepted physics that has been confirmed in thousands of different kinds of tests, that's still not nearly good enough. For one thing, it assumes that the researchers have done the analysis correctly and haven't overlooked some systematic source of error. And because the result was so unexpected and so revolutionary, that's exactly what most physicists think happened — some undetected source of error."

"Interestingly, a different set of results from the same CERN particle accelerator were interpreted quite differently."

"A possible detection of something called a Higgs boson — a theorized subatomic particle that would help to explain why particles weigh something rather than nothing — was also announced last year. That result had only a 2.3 sigma confidence level, corresponding to about one chance in 50 that the result was a random error (98 percent confidence level). Yet because it fits what is expected based on current physics, most physicists think the result is likely to be correct, despite its much lower statistical confidence level."