ORIGIN ≡ 0

The Normal Distribution

The Normal Distribution, also known as the "Gaussian Distribution" or "bell-curve", is the most widely employed function relating observations X with probabilty P(X) in statistics. Many natural populations are approximately normally distributed, as are several important derived quantitities even when the original population is not normally distributed.

Properly speaking, the Normal Distribution is a continuous "probability density function" meaning that values of a random variable X may take on any numerical value, not just discrete values. In addition, because the values of X are infinite the "exact" probabiliy P(X) for any X is zero. Thus, in order to determine probabilities one typically looks at invervals of X such as X >2.3 or 1< X < 2 and so forth. It is interesting to note that because the probability P(X) = 0, we don't have to worry about correctly interpreting pesky boundaries, as seen in discrete distributions, since X > 2 means the same thing as $X \ge 2$ **and X < 2 is the same as** $X \le 2$.

As described previously, the Normal distribution N(μ,σ²) consists of a family of curves that are specified by supplying values for two parameters: μ **= the mean of the Normal population, and** σ^2 **= the variance of the same population.**

Prototyping the Normal Function using the Gaussian formula:

Making the plot of N(50,100):

$\mu := 50$ **< specifying mean (μ)**

$\sigma := \sqrt{100}$ $\sigma^2 = 100$ **< specifying variance (σ²)**

$i := 0..100$

$X_i := i$

< Defining a bunch of X's ranging in value from 0 to 100. Remember that the range of X is infinite, but we'll plot 101 point here. That should give us enough points to give us an idea of the Gaussian function shape!

$$Y1_i := \frac{1}{\sigma \cdot \sqrt{2 \cdot \pi}} \cdot e^{\left[\frac{-1}{2\cdot\sigma^2}\cdot(X_i-\mu)^2\right]}$$

< Formula for Normal distribution. Here we have computed P(X) for each of our X's. Zar 2010 Eq. 6.1, p. 66.

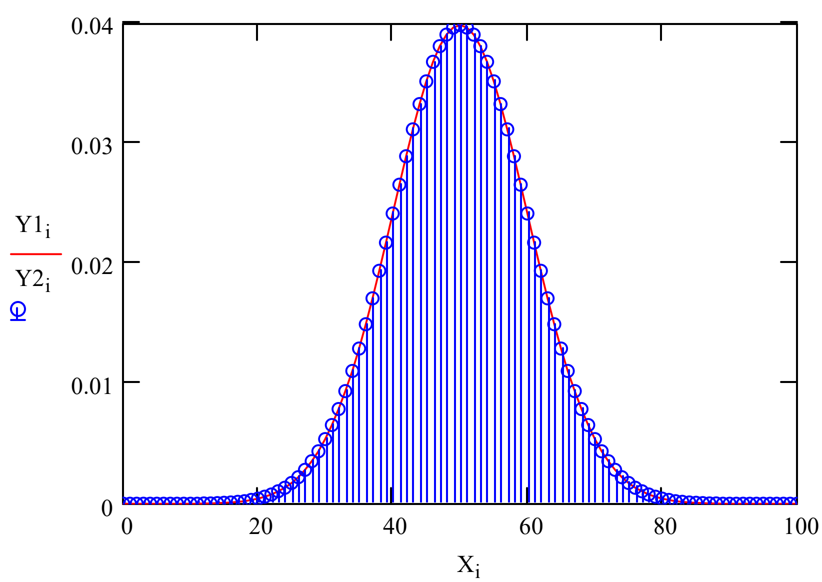
Now, let's compare with Mathcad's built-in function:

$Y2_i := \text{dnorm}(X_i, \mu, \sigma)$ $\sigma^2 = 100$

< MathCad's function asks us provide standard deviation rather than variance...

Plotting the two sets of Y's:

The two approaches give th same probability function P for X, so this prototype confirms the built-in functi



Prototype in R:

dnorm(x,mu,sigma)

^ R has a nearly identical function, see *Biostatistics* 070

What happens when μ or σ^2 is changed:

Location of mode changes (translation of μ) and width of hump changes showing greater or lesser variance - see *Biostatistics* 070.

Simulation of Normally Distributed Data:

$\mu := 65 \qquad \sigma := 25 \qquad \sigma^2 = 625$

$X := \text{rnorm}(1000, \mu, \sigma)$

Descriptive Statistics for X:

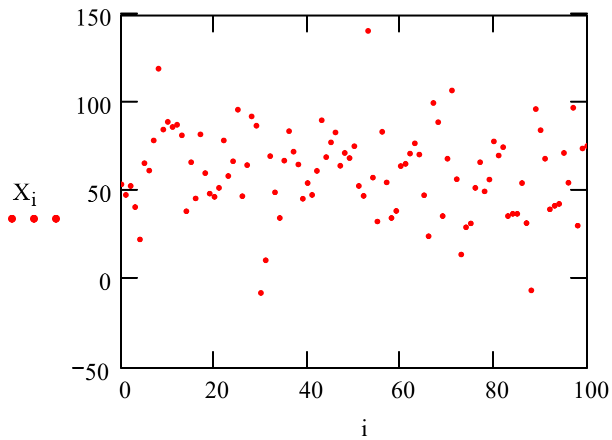
$n := \text{length}(X) \qquad n = 1000$

$\text{mean}(X) = 63.5061$

$\frac{n}{n-1} \cdot \text{var}(X) = 606.3107$

$\text{Var}(X) = 606.3107$

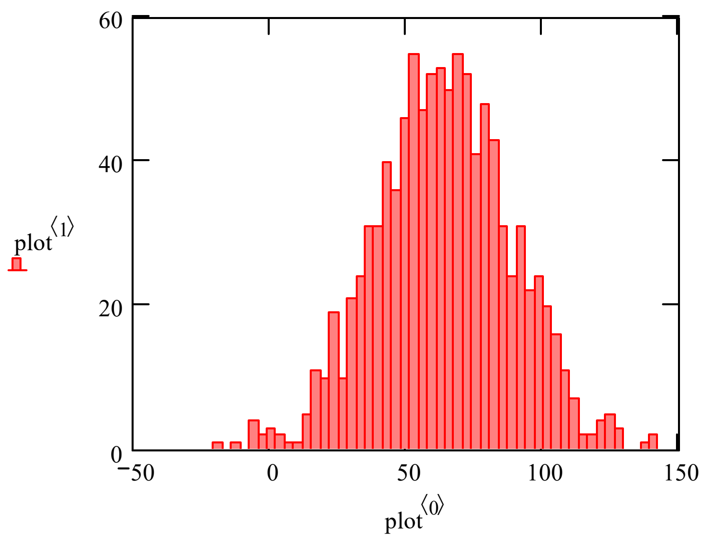
< Note: mathcad has two func
var(X) = population variance
Var(X) = sample variance



^ Mean and variance of this sample are close, but not exactly equal to N(65,625). This is to be expected of a sample as opposed to the entire population

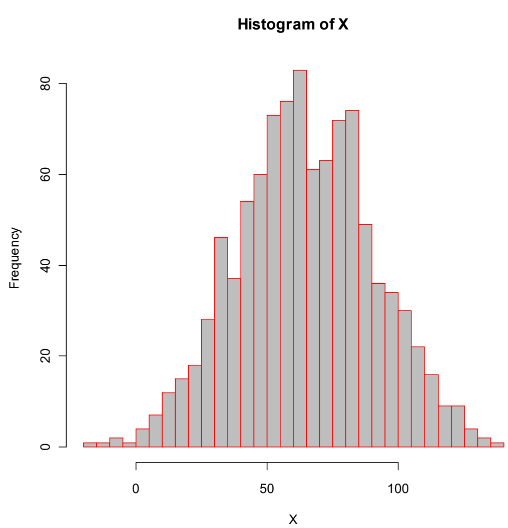
Histogram of X:

$\text{plot} := \text{histogram}(50, X)$



Prototype in R:

```
#CREATING A PSEUDORANDOM
 NORMAL DISTRIBUTION:
X=rnorm(1000,65,25)
hist(X,nclass=50,col="gray",border="red")
```



< R has a nearly identical function rnorm(n,mu,sigma) where n = number of points desired

Standardizing the Normal Distribution:

In many instances, we have a sample that we may wish to compare with a Normal Distribution. Using computer-based functions, as above, one has little difficulty calculating probabilities P(X) and simulating additional samples from a Normally Distributed population $N(\mu,\sigma^2)$. When using published tables, however, it is often useful to compare probabilities with the Standard Normal Distribution ~N(0,1). This is done by *Standardizing* the Data:

Given your X's ~$N(\mu,\sigma^2)$ you create a new variable Z ~N(0,1) by means of a Linear Transformation:

$i := 0..999$

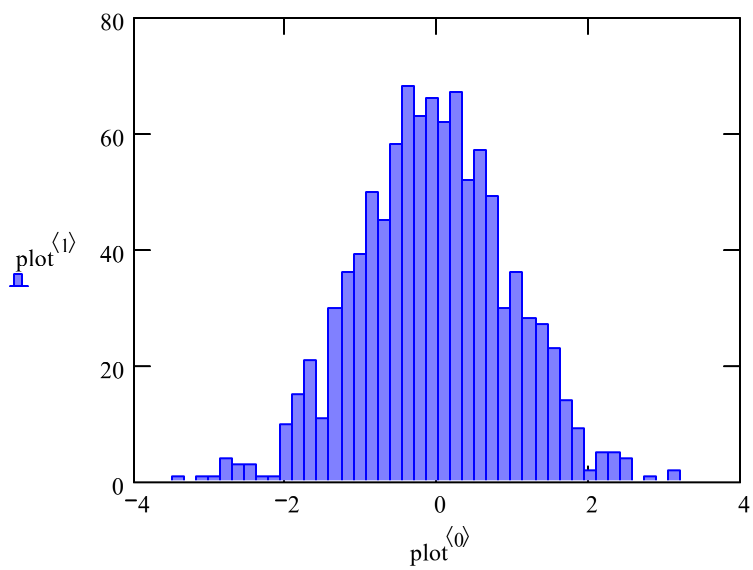
$$Z_i := \frac{(X_i - \mu)}{\sigma}$$ **< Z's are now Standardized ~N(0,1)**

$\text{mean}(Z) = -0.0598$

$\text{Var}(Z) = 0.9701$ **< sample estimates are close, but not exactly equal to N(0,1)**

Histogram of Z:

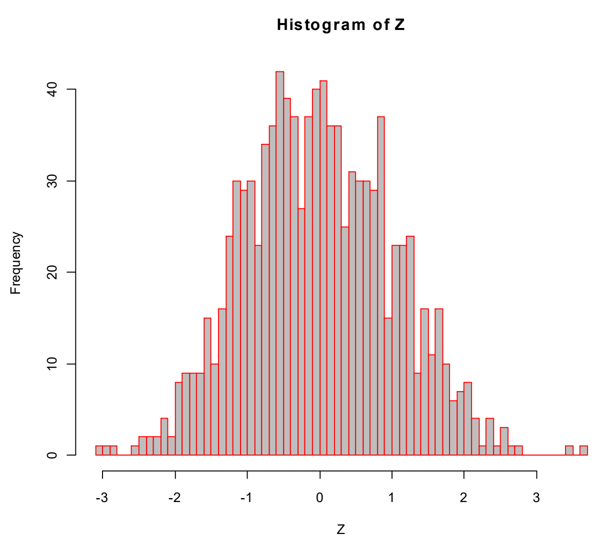
$plot := \text{histogram}(50, Z)$



Prototype in R:

```
#STANDARDIZING DATA:
mu=mean(X)
sigma=sd(X)
Z=(X-mu)/sigma
Z[1:10]
hist(Z,nclass=50,col="gray",border="red")

Z2=scale(X,center=TRUE,scale=TRUE)
Z2[1:10]
```



Note: in both cases here, we had prior knowledge of μ and σ^2.

With real-world data, we will have to estimate these values, usually with X_{bar} & s^2.

Calculating Probabilities & Quantiles:

The above graphs display the relationship between X values, or observations (also called quantiles), and the probability that a range (or bin) of X is expected to have given the assumption of Normal probability for X, indicated as P(X). Most statistical software packages have standard "p" and "q" functions allowing conversion from X to P(X) and vice versa. In the most useful form, the probability function is given as a *Cumulative Probability* Φ(X) starting from X values of minus infinity up to X. In each case a specific cumulative probability function reqires that one provides specific parameter values for the curve (μ,σ,), along with X or Φ(X).

Probabilities of the Normal Distribution and Cumulative Normal Distribution N(0,1):

$i := 0..100$

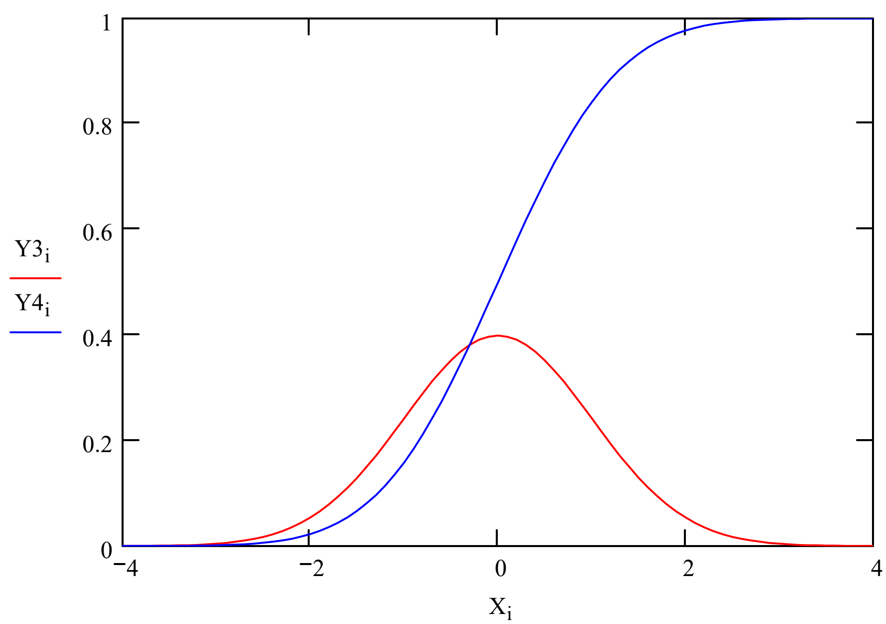
$X_i := \frac{i-50}{10}$ **< scaling 101 X's to a reasonable scale...**

$\mu := 0 \quad \sigma := 1 \quad \sigma^2 = 1$ **< parameters of the Normal N(0,1) distribution...**

$Y3_i := \mathrm{dnorm}\left(X_i, \mu, \sigma\right)$ **< *Interval Estimate* of probability P(X) for each X**

$Y4_i := \mathrm{pnorm}\left(X_i, \mu, \sigma\right)$ **< *Cumulative* probability Φ(X) for each X**

Plots of Normal Distribution and Cumulative Normal Distributions



Prototype in R:

```
#PQ FUNCTIONS FOR NORMAL DISTRIBUTION:
mu=0
sigma=1
X=1.6449
PHI=0.90
dnorm(x,mu,sigma)      # interval estimate P(X) given X
pnorm(x,mu,sigma)      # cumulative phi(X) given X
qnorm(PHI,mu,sigma)  # X given cumulative phi(X)
```

Calculating Probability Intervals of the Cumulative Normal Distribution:

$\mu = 0$ $\sigma = 1$ **< Normal distribution parameters (change these if desired)**

Probability that X ranges between -1 and 1:

$\mathrm{dnorm}(-1,\mu,\sigma) = 0.242$ $\mathrm{dnorm}(1,\mu,\sigma) = 0.242$ **< P(X)**

$\mathrm{pnorm}(-1,\mu,\sigma) = 0.1587$ $\mathrm{pnorm}(1,\mu,\sigma) = 0.8413$ **< Φ(X)**

$\mathrm{pnorm}(1,\mu,\sigma) - \mathrm{pnorm}(-1,\mu,\sigma) = 0.6827$ **< Calculating MAX cut-off - MIN cut-off**

^ cumulative value at MIN of interval

^ cumulative value at MAX of interval **68.27%**

Probability that X ranges between -2.576 and 2.576:

$\mathrm{dnorm}(-2.576,\mu,\sigma) = 0.0145$ $\mathrm{dnorm}(2.576,\mu,\sigma) = 0.0145$ **< P(X)**

$\mathrm{pnorm}(-2.576,\mu,\sigma) = 0.005$ $\mathrm{pnorm}(2.576,\mu,\sigma) = 0.995$ **< Φ(X)**

$\mathrm{pnorm}(2.576,\mu,\sigma) - \mathrm{pnorm}(-2.576,\mu,\sigma) = 0.99$ **< Calculating MAX cut-off - MIN cut-off**

^ cumulative value at MIN of interval

^ cumulative value at MAX of interval **99%**

Probability that X ranges between -1.96 and 1.96

$\mathrm{dnorm}(-1.96,\mu,\sigma) = 0.0584$ $\mathrm{dnorm}(1.96,\mu,\sigma) = 0.0584$ **< P(X)**

$\mathrm{pnorm}(-1.96,\mu,\sigma) = 0.025$ $\mathrm{pnorm}(1.96,\mu,\sigma) = 0.975$ **< Φ(X)**

$\mathrm{pnorm}(1.96,\mu,\sigma) - \mathrm{pnorm}(-1.96,\mu,\sigma) = 0.95$ **< Calculating MAX cut-off - MIN cut-off**

^ cumulative value at MIN of interval

^ cumulative value at MAX of interval **95%**

Prototype in R:

```
#EXAMPLE INTERVAL CALCULATIONS:
mu=0
sigma=1
MIN=pnorm(-1,mu,sigma)
MAX=pnorm(1,mu,sigma)
MAX-MIN
MIN=pnorm(-2.576,mu,sigma)
MAX=pnorm(2.576,mu,sigma)
MAX-MIN
MIN=pnorm(-1.96,mu,sigma)
MAX=pnorm(1.96,mu,sigma)
MAX-MIN
```

Calculating Quantiles of the Cumulative Normal Distribution:

Quantile Range for Probability 95%

$\mathrm{qnorm}(0.975, \mu, \sigma) = 1.96$

^ X values (quantiles) lying outside the 95% probability range can occur either above 1.96 or below value -1.96. This is the "two-tailed" situation where 5% lying outside the range is distributed 2.5% above 1.96 and 2.5% below -1.96.

Quantiles associated with a range of Cumulative Probabilities:

$\mathrm{ORIGIN} := 1$

$i := 1..39$

$\Phi_i := \frac{i}{40}$ **< Creating a range of Cumulative Probabilities Φ**

$Q_i := \mathrm{qnorm}(\Phi_i, \mu, \sigma)$ **< Finding the quantiles Q**

$$\Phi = \begin{pmatrix} 0.025 \\ 0.050 \\ 0.075 \\ 0.100 \\ 0.125 \\ 0.150 \\ 0.175 \\ 0.200 \\ 0.225 \\ 0.250 \\ 0.275 \\ 0.300 \\ 0.325 \\ 0.350 \\ 0.375 \\ 0.400 \\ 0.425 \\ 0.450 \\ 0.475 \\ 0.500 \\ 0.525 \\ 0.550 \\ 0.575 \\ 0.600 \\ 0.625 \\ 0.650 \\ 0.675 \\ 0.700 \\ 0.725 \\ 0.750 \\ 0.775 \\ 0.800 \\ 0.825 \\ 0.850 \\ 0.875 \\ 0.900 \\ 0.925 \\ 0.950 \\ 0.975 \end{pmatrix} \quad Q = \begin{pmatrix} -1.95996 \\ -1.64485 \\ -1.43953 \\ -1.28155 \\ -1.15035 \\ -1.03643 \\ -0.93459 \\ -0.84162 \\ -0.75542 \\ -0.67449 \\ -0.59776 \\ -0.5244 \\ -0.45376 \\ -0.38532 \\ -0.31864 \\ -0.25335 \\ -0.18912 \\ -0.12566 \\ -0.06271 \\ 0 \\ 0.06271 \\ 0.12566 \\ 0.18912 \\ 0.25335 \\ 0.31864 \\ 0.38532 \\ 0.45376 \\ 0.5244 \\ 0.59776 \\ 0.67449 \\ 0.75542 \\ 0.84162 \\ 0.93459 \\ 1.03643 \\ 1.15035 \\ 1.28155 \\ 1.43953 \\ 1.64485 \\ 1.95996 \end{pmatrix}$$

Prototype in R:

```
#QUANTILES
Q=qnorm(0.975,mu,sigma)
Q
PHI=seq(0.025,0.975,0.025)
PHI
Q=qnorm(PHI,mu,sigma)
Q
```