Random Variables:

Random variables are functions that describe the outcomes of an "experiment." They are usually denoted with a capital letter, such as *X* or *Y*. RVs can be either discrete of continuous. A discrete random variable has a "countable" number of values, meaning that they could be listed. A continuous random variable has an uncountable number of numeric values, meaning they cannot be listed. Here are some examples:

Discrete

- # spots on top face of die (1, 2, 3, 4, 5, 6)
- suit of drawn card (C, D, H, S)
- # aphids on leaf (0, 1, 2, 3, …)
- # defects in box of 1000 nails (0, 1, 2, …, 1000)
- # germinating seeds out of 50 (0, 1, 2, …, 50)

Continuous

- heights of people (0 -- ?)
- ph of soil (0 – 10)
- voltage in circuit (0 -- ?)

Discrete random variables have *probability (mass) function*s which we denote *p(x)*. The probability function gives probabilities of each individual value of the RV, e.g.

$$P(X = x) = p(x).$$

Continuous random variables have *probability density functions* which we denote *f(x)*. The probability density function can be used to give probabilities of ranges of values of the continuous RV. For example,

$$P(x_1 < X < x_2) = \int_{x_1}^{x_2} f(x)dx$$

The *Binomial* and *Normal* are very important random variables.

Binomial Random Variable

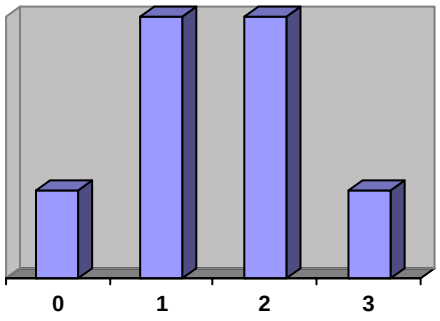
Let X = Number of successes out of n trials, in which the probability of success on each trial is a number π. Then X has a *binomial* distribution with parameters n and π. This is abbreviated $X \sim B(n, \pi)$.

Example: Consider flipping a coin, and declare a "success" if a head (H) appears. Then the probability of success is .5. That is, π=P(S)=.5. Suppose the coin is flipped n=3 times, and X = number of heads. Then $x \sim B(3, .5)$. The possible values of x are (0, 1, 2, 3). The probability of any event is $(1/2)^3 = 1/8$.

Events:	HHH	HHT	HTH	THH	HTT	THT	TTH	TTT
# Heads	3	2	2	2	1	1	1	0
P(Event)	1/8	1/8	1/8	1/8	1/8	1/8	1/8	1/8

P(3 H) = 1/8 P(2 H) = 3/8 P(1 H) = 3/8 P(0 H) = 1/8

In general: $P(k\ H) = \frac{3!}{k!(3-k)!} \cdot \left(\frac{1}{2}\right)^3 = \frac{3!}{k!(3-k)!} \cdot \left(\frac{1}{8}\right)$



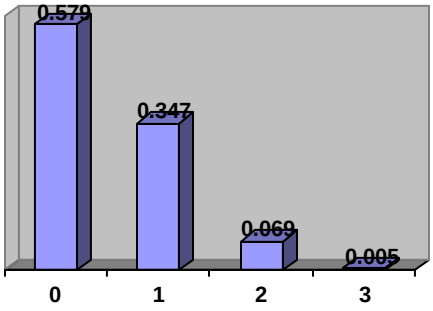
Binomial Random Variable (con't)

Example: X = number of 1's in 3 rolls of die (0, 1, 2, 3)

Events:	111	11X	1X1	X11	1XX	X1X	XX1	XXX
# 1's	3	2	2	2	1	1	1	0
P(Event)	$1^3/6^3$	$1^2 5^1/6^3$	$1^2 5^1/6^3$	$1^2 5^1/6^3$	$1^1 \cdot 5^2/6^3$	$1^1 \cdot 5^2/6^3$	$1^1 \cdot 5^2/6^3$	$5^3/6$

P(3 1's) =

$$\frac{1}{6}\cdot\frac{1}{6}\cdot\frac{1}{6}=\frac{1}{6^3}=\frac{1}{216}=.0046$$



P(k 1's) = $\dfrac{3!}{k!(3-k)!}\left(\dfrac{1}{6}\right)^k\left(\dfrac{5}{6}\right)^{3-k}$

The probability mass function is given by the so-called Binomial Formula:

π = probability of success on single trial

P(x successes in n trials) = $\dfrac{n!}{x!(n-x)!}\pi^x(1-\pi)^{n-x}$

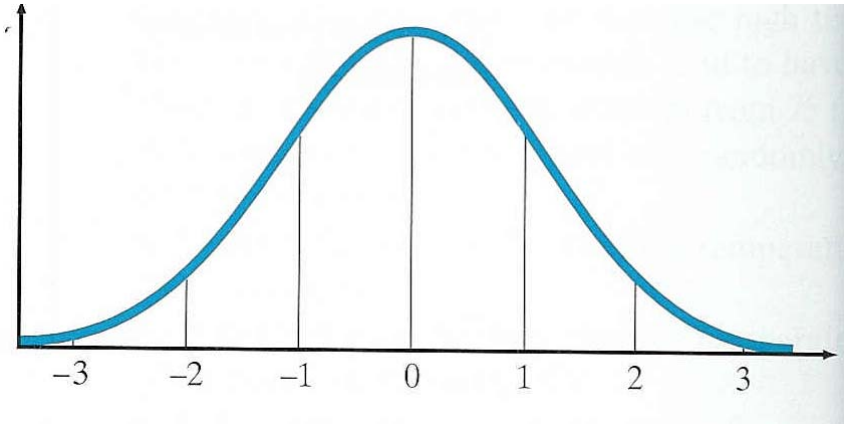
Mean of the binomial distribution: $n\pi$

Variance of the binomial distribution: $n\pi(1-\pi)$

Normal Distribution and normal random variable:

The notation $Y \sim N(\mu,\sigma^2)$ means "The random variable Y is distributed normally with mean μ and variance σ^2.

The *standard normal* distribution has mean $\mu=0$ and variance $\sigma^2=1$. The letter Z is reserved to represent the standard normal random variable.



Computer programs and tables are available to obtain probabilities from the normal distribution. For example, you can discover that

- $P(-1 < Z < 1) = .68$
- $P(Z > 1) = .16$
- $P(-1.96 < Z < 1.96) = .95$
- $P(Z > 1.96) = .025$
- $P(Z > 1.42) = .0778$.

The probability density function of the normal random variable with mean μ and variance σ^2 is given by the formula

$$f(y) = (2\pi\sigma^2)^{-1/2} \exp(-(y-\mu)^2 / 2\sigma^2)$$

Mean of Normal Distribution: μ
Variance of Normal Distribution: σ

Using the Normal Distribution

Standardizing a Normal Distribution:

If $Y \sim N(\mu,\sigma^2)$, then $Z = (Y - \mu)/\sigma \sim N(0,1)$.

This result allows us to compute probabilities from *any* normal distribution using tables or a computer program for the standard normal distribution.

If you wanted to calculate the probability that a random variable *y* is greater that 1.42 standard deviations above its mean, you would compute:

$$P(Y > \mu + 1.42\sigma) = P(\frac{Y - \mu}{\sigma} > 1.42) = P(Z > 1.42) = .0778$$

As a more specific application, suppose you believe the egg weights to be normally distributed with mean 65.4 and standard deviation 5.17. You would calculate the probability that a randomly drawn egg is greater than 72 as:

$$P(Y > 72) = P((Y - 65.4)/5.2) > (72 - 65.4)/5.2) = P(Z > 1.27) = .1$$

Using the normal distribution—an application

Egg weights are normally distributed with mean $\mu = 65$ (g) and standard deviation $\sigma = 5.0$.

1. What is the probability one randomly drawn egg will exceed:
   a. 65 b. 66 c. 70 d. 75

   Let Y = egg weight. Then

   a. $P(Y > 65) = P\left(\frac{Y-65}{5} > \frac{65-65}{5}\right) = P(Z > 0) = ½ = .5$
   b. $P(Y > 66) = P\left(\frac{Y-65}{5} > \frac{66-65}{5}\right) = P(Z > .2) = .4207$
   c. $P(Y > 70) = P\left(\frac{Y-65}{5} > \frac{70-65}{5}\right) = P(Z > 1) = .1587$
   d. $P(Y > 75) = P\left(\frac{Y-65}{5} > \frac{75-65}{5}\right) = P(Z > 2) = .0228$

2. What is the probability one egg is between 66 and 70 g?
   $P(66 < Y < 70) = P(Y > 66) - P(Y > 70)$
   $= .4207 - .1587 = .262$

Probabilities of this type can be expressed in terms of the *cumulative distribution function*

$$P(Z < z) = F(z) = \int_{-\infty}^{z} (2\pi)^{-1/2} \exp(-z^2/2)$$

The integral for the normal distribution is difficult to evaluate, so tables or computer programs are used to obtain actual values.

Normal Approximation to the Binomial

You can use the normal distribution to approximate binomial probabilities. This often simplifies a computation. For example, suppose you are shooting free-throws in basketball. You know that you make 75% of your shots; that is, the probability of making any one shot is .75. You have entered a contest that awards a prize if you make at least 18 out of 20 shots. What is the probability that you will win a prize?

You need to calculate P($Y \geq 18$), where y is the number of shots you make out of 20. The exact probability is given by the binomial formula with $\pi = .75$ and $n = 20$:

$$P(Y \geq 18) = P(Y = 18) + P(Y = 19) + P(Y = 20)$$

$$= 20!/(18!2!).75^{18}.25^{2}$$

$$+20!/(19!1!).75^{19}.25^{1}$$

$$+20!/(20!0!).75^{20}.25^{0}$$

$$= .0069 + .0211 + .0032$$

$$= .0912$$

Normal Approximation to the Binomial

The calculation on the previous page would be tedious by hand, but many computer programs are available that can readily do it. However, even good computer programs may fail for computations involving extremely large n.

The normal approximation sets $\mu = n\pi$ and $\sigma^2 = n\pi(1-\pi)$, and assumes $Y \sim N(\mu,\sigma^2)$ to evaluate the probability. The approximation is improved by using a *continuity correction*, which means you compute $P(Y \geq 18\text{-}.5) = P(Y \geq 17.5)$.

The normal approximation is then computed as:

$\mu = n\pi = 20(.75) = 15$
$\sigma^2 = n\pi(1-\pi) = 20(.75)(.25) = 3.75$
$\sigma = 3.75^{1/2} = 1.94$

$$P(Y \geq 17.5) = P((y - 15)/1.94 \geq (17.5 - 15)/1.94)$$

$$= P(Z \geq 1.29) = 1 - .901 = .099.$$

This is a reasonable approximation to the exact binomial probability of $P(Y \geq 18) = .091$.

Means and Variances of Random Variables

Means of random variables are called *expected values*, denoted

$$\mu = E(X)$$

If X is a continuous RV, then

$$\mu_X = E(X) = \int xf(x)dx$$

If X is a discrete RV, then

$$\mu_X = E(X) = \sum_i x_i p(x_i)$$

Variances of random variables are also expected values.

If X is a continuous RV, then

$$\sigma_X^2 = E((X-\mu)^2) = \int (x-\mu)^2 f(x)dx$$

If X is a discrete RV, then

$$\sigma_X^2 = E((X-\mu)^2) = \sum_i (x_i-\mu)^2 p(x_i)$$

Means and Variances of Linear Functions of Random Variables

If X is an RV with mean μ_X and variance σ_X^2, and $Y=a+bX$, where a and b are constants, then:

$$\mu_Y = E(a+bX) = a + bE(X) = a + b\mu_X$$

and

$$\sigma_Y^2 = E((Y-\mu_Y)^2) = E(((a+bX)-(a+b\mu_X))^2)$$
$$= E(b^2(X-\mu_X)^2) = b^2\sigma_X^2$$

If $X_1, \ldots, X_k$ are RVs and $Y= X_1 + \ldots + X_k$ then

$$\mu_Y = E(X_1 + \ldots + X_k) = E(X_1) + \ldots + E(X_k) = \mu_1 + \ldots + \mu_k$$

If $X_1, \ldots, X_k$ are independent RVs, each with mean μ and variance σ^2, and if $Y= X_1 + \ldots + X_k$ then

$$\sigma_Y^2 = E((X_1 + \ldots + X_k) - (\mu_1 + \ldots + \mu_k))^2$$
$$= E(X_1 - \mu_1)^2 + \ldots + E(X_k - \mu_k)^2 = \sigma_1^2 + \ldots + \sigma_k^2$$

It follows from the above results that the sample mean,

$$\overline{X} = (X_1 + \ldots + X_n)/n$$

is a random variable with mean $\mu_{\overline{X}} = \mu$ and variance $\sigma_{\overline{X}}^2 = \sigma^2/n$. This called the "sampling" distribution of $\overline{X}$.

Other Discrete Random Variables

The Binomial random variable is the most commonly used discrete random variable. It represents the number of "successes" of out on n independent Bernoulli trials. ("Bernoulli" means that a trial has only two possible outcomes; e.g. S-F, 0-1, Y-N, M-F, T-F, etc.) There are other discrete random variables that are used to represent outcomes from other situations.

Poisson Random Variable: The Poisson RV is used to represent the number of times that randomly occurring items are detected in a given interval of time or space. An example is given by the number of cars crossing a point on a road in a specified interval of time, assuming the cars come at random times. Another example is the number of particles suspended in a specified volume of air or liquid medium, assuming the particles are randomly distributed.

The distinction of the Poisson from the Binomial is that the Poisson counts are out of a given time or space, and the Binomial counts are out of a given number of trials. There is no upper bound on the Poisson counts; its possible values are 0, 1, 2, 3, … .

The probability mass function for the Poisson RV is

$$p(x) = \frac{\lambda^x}{x!} e^{-\lambda}, \text{ for } x = 0, 1, 2, 3, ...$$

The values of $p(x)$ form an infinite series that converges to 1:

$$\sum_{x=0}^{\infty} \frac{\lambda^x}{x!} e^{-\lambda} = 1$$

The mean and variance of the Poisson distribution are both equal to λ; that is, $\mu = \sigma^2 = \lambda$. Therefore the Poisson count, x, is an estimate of λ, and $\lambda^{1/2}$ is an estimate of the "uncertainty" expressed by the standard deviation.

The Poisson distribution provides a good approximation to the Binomial for large n and small π, with $\lambda = n\pi$.

Hypergeometric Random Variable:

This RV is used in a situation similar to that of the Binomial, except that the population of "S's and F's" is finite. Thus, the trials are not independent because the probability of S or F changes each time a value is drawn from the population.

Geometric and Negative Binomial Random Variable:

This RV represents the number of independent Bernoulli trials required to obtain a specified number r of successes. If $r=1$, then the Negative Binomial RV is called a Geometric RV.

Other Continuous Random Variables

If data are skewed, and thus do not fit the normal distribution, then other RV's can be used.

Lognormal Random Variable. In many types of applications the logarithms of data can be assumed normally distributed. It this assumption is true, then the data follow a lognormal distribution. Then, if X has a lognormal distribution, $Y=lnX$ will have a normal distribution. Usually, data are "transformed" and then analyzed in the log scale.

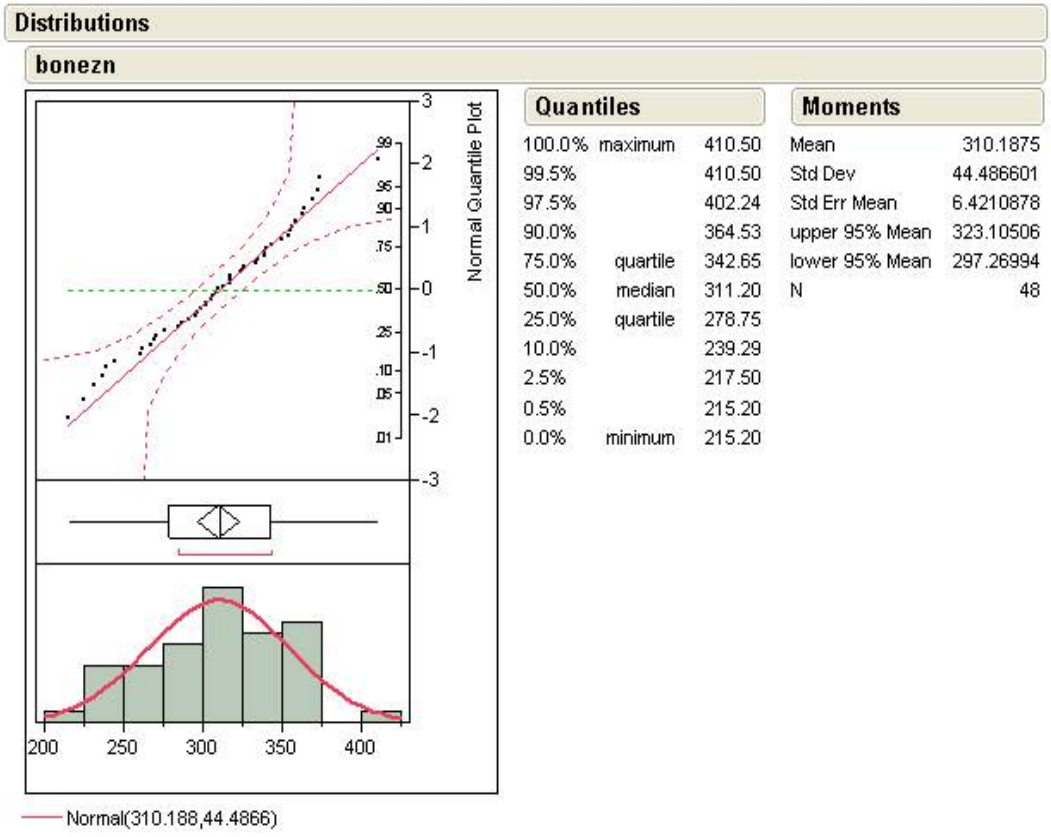
Exponential Random Variable. The exponential distribution is used as a "waiting time" distribution. If items are arriving at random (as with the Poisson distribution) the times between arrivals will follow the exponential distribution. The probability density function is

$$f(x) = \lambda e^{-\lambda x}, \text{ for } x > 0.$$

Probability Plots

Statistical graphs are useful to visually assess certain attributes of probability distributions. The figure below shows several such plots based on a data set containing measured values of bone zinc in 48 sheep. The quantiles and moments at the right of the plots give a computational summary of the data.

At the bottom is a histogram with the best fitting normal probability density curve plotted through it. The curve passes through the histogram bars fairly well, but there are a few departures. There are no data in the 375-400 interval, and a few too many values in the 350-375 interval. The 375-400 interval is empty, and there is only one value in the 400-425 interval. The familiar box-plot does not detect these departures.



Quantiles		
100.0%	maximum	410.50
99.5%		410.50
97.5%		402.24
90.0%		364.53
75.0%	quartile	342.65
50.0%	median	311.20
25.0%	quartile	278.75
10.0%		239.29
2.5%		217.50
0.5%		215.20
0.0%	minimum	215.20

Moments	
Mean	310.1875
Std Dev	44.486601
Std Err Mean	6.4210878
upper 95% Mean	323.10506
lower 95% Mean	297.26994
N	48

The departure from normality is depicted more clearly with a so-called "quantile-quantile" (Q-Q) plot, which is shown in the upper portion of the plot. The Q-Q plot is basically a plot of the quantiles of the fitted normal distribution versus the quantiles of the data distribution, i.e. the ordered observations $y_{1,\ldots,} y_{n}$. If the data represented a sample from a normal distribution, the points would be distributed about a straight line. Clusters of points above the line indicate more data that would be expected in that area, and points below the line indicate fewer. You can see the cluster from the 225-250 interval and the 350-375 interval. You can also see the gap due to the empty 375-400 interval.

Even though the Q-Q plot detects possible departures from normality, the results are not definitive because of the relative small sample of 48 values. Q-Q plots are more effective with larger data sets than n=48.

Central Limit Theorem

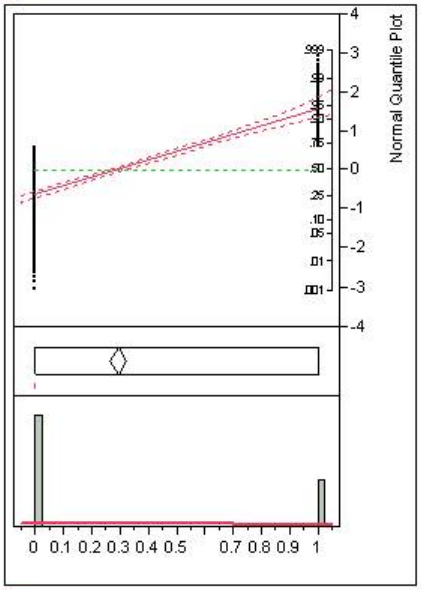
The sampling distribution is one of the most difficult concepts in all of statistics for students to comprehend. In order to grasp the concept of the sampling distribution, imagine obtaining a large number of samples, with each sample consisting of n observations. Then you compute the mean of each sample to generate a "population" of sample means. These means constitute the sampling distribution of $\overline{X}$.

Previously it was shown that the sampling distribution of the sample mean $\overline{X}$ has mean μ and variance σ^2/n. The *Central Limit Theorem* states that the sampling distribution of $\overline{X}$ also is approximate normally distributed, even though the distribution from which the samples were obtained is not normal. The closeness to the normal distribution increases as the sample size n increases.

To illustrate, consider a Bernoulli distribution that has only 0 and 1 as distinct values, and suppose there is a proportion π of 1's and a proportion 1-π of 0's. Imagine taking numerous samples from this population, each of size n, and computing the means for each of the samples. Now imagine doing this for several different values of n, say, n=1, 3, 5, 10, 30, and 100. Following are histograms of the means for each of the sample sizes, with π=.3.

Distributions

y1



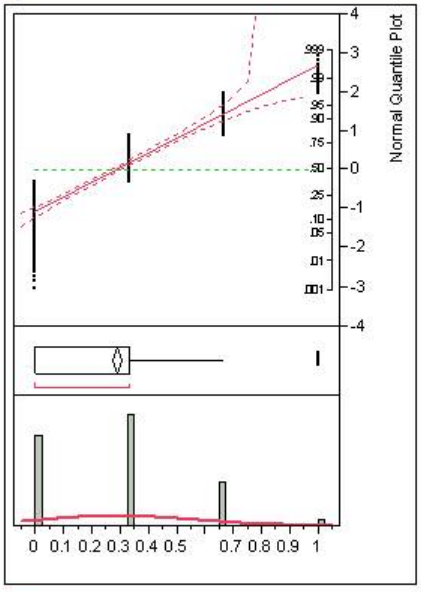
Normal(0.294,0.45582)

Quantiles		
100.0%	maximum	1.0000
99.5%		1.0000
97.5%		1.0000
90.0%		1.0000
75.0%	quartile	1.0000
50.0%	median	0.0000
25.0%	quartile	0.0000
10.0%		0.0000
2.5%		0.0000
0.5%		0.0000
0.0%	minimum	0.0000

Moments	
Mean	0.294
Std Dev	0.4558199
Std Err Mean	0.0144143
upper 95% Mean	0.3222858
lower 95% Mean	0.2657142
N	1000

Distributions

y3



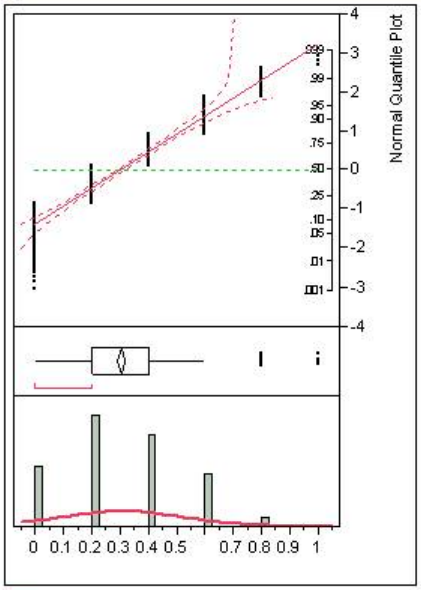
Normal(0.28967,0.26383)

Quantiles		
100.0%	maximum	1.0000
99.5%		1.0000
97.5%		1.0000
90.0%		0.6667
75.0%	quartile	0.3333
50.0%	median	0.3333
25.0%	quartile	0.0000
10.0%		0.0000
2.5%		0.0000
0.5%		0.0000
0.0%	minimum	0.0000

Moments	
Mean	0.2896667
Std Dev	0.2638319
Std Err Mean	0.0083431
upper 95% Mean	0.3060387
lower 95% Mean	0.2732947
N	1000

Distributions

y5



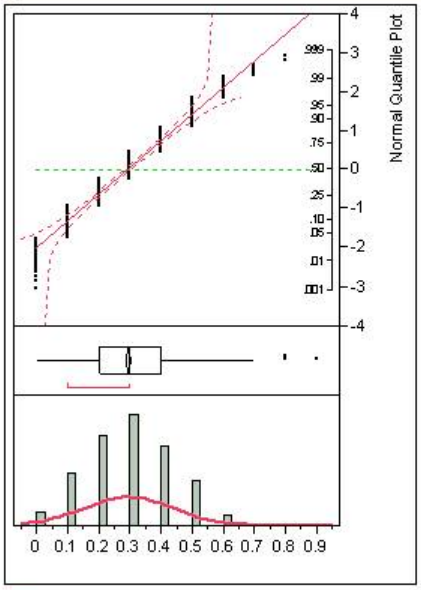
Normal(0.3048,0.2162)

Quantiles		
100.0%	maximum	1.0000
99.5%		0.8000
97.5%		0.8000
90.0%		0.6000
75.0%	quartile	0.4000
50.0%	median	0.2000
25.0%	quartile	0.2000
10.0%		0.0000
2.5%		0.0000
0.5%		0.0000
0.0%	minimum	0.0000

Moments	
Mean	0.3048
Std Dev	0.2162029
Std Err Mean	0.0068369
upper 95% Mean	0.3182164
lower 95% Mean	0.2913836
N	1000

Distributions

y10



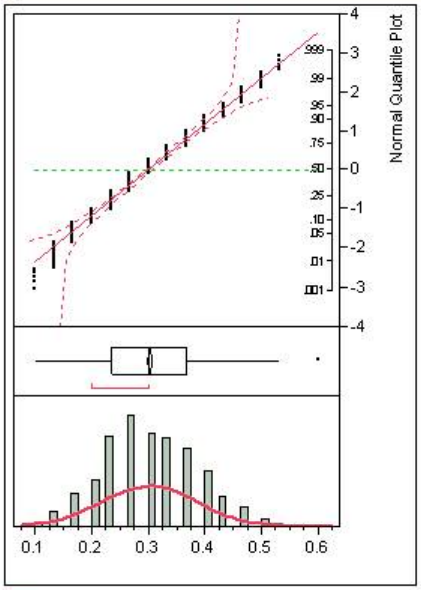
Normal(0.2953,0.14536)

Quantiles		
100.0%	maximum	0.90000
99.5%		0.70000
97.5%		0.60000
90.0%		0.50000
75.0%	quartile	0.40000
50.0%	median	0.30000
25.0%	quartile	0.20000
10.0%		0.10000
2.5%		0.00000
0.5%		0.00000
0.0%	minimum	0.00000

Moments	
Mean	0.2953
Std Dev	0.1453583
Std Err Mean	0.0045966
upper 95% Mean	0.3043202
lower 95% Mean	0.2862798
N	1000

Distributions

y30



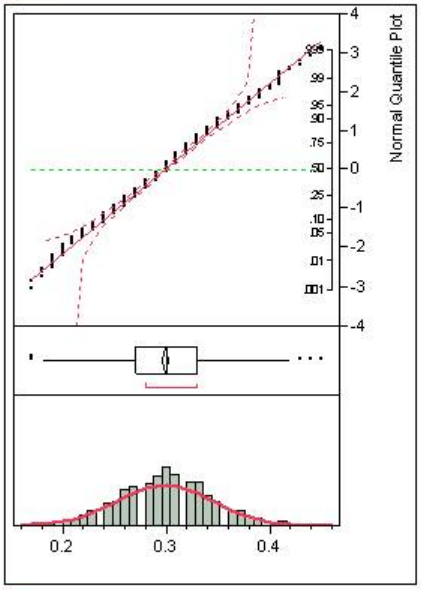
Normal(0.30223,0.08449)

Quantiles		
100.0%	maximum	0.60000
99.5%		0.53333
97.5%		0.46667
90.0%		0.40000
75.0%	quartile	0.36667
50.0%	median	0.30000
25.0%	quartile	0.23333
10.0%		0.20000
2.5%		0.13333
0.5%		0.10017
0.0%	minimum	0.10000

Moments	
Mean	0.3022333
Std Dev	0.0844915
Std Err Mean	0.0026719
upper 95% Mean	0.3074764
lower 95% Mean	0.2969902
N	1000

Distributions

y100



Normal(0.29912,0.04586)

Quantiles		
100.0%	maximum	0.45000
99.5%		0.42000
97.5%		0.39000
90.0%		0.36000
75.0%	quartile	0.33000
50.0%	median	0.30000
25.0%	quartile	0.27000
10.0%		0.24000
2.5%		0.21000
0.5%		0.18005
0.0%	minimum	0.17000

Moments	
Mean	0.29912
Std Dev	0.0458599
Std Err Mean	0.0014502
upper 95% Mean	0.3019658
lower 95% Mean	0.2962742
N	1000