Normal Probability Distribution

Animal Health Risk Analysis
Canadian Food Inspection Agency
3851 Fallowfield Road
Ottawa, Ontario, Canada K2H 8P9

Analyse des risques de la santé des animaux
Agence canadienne d'inspection des aliments
3851, chemin Fallowfield
Ottawa (Ontario) Canada K2H 8P9



Canadian Food Inspection Agency (CFIA)

Agence canadienne d'inspection des aliments (ACIA)

Properties of the Normal Distribution

Probability density function:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

Cumulative distribution function:

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-(v-\mu)^2/2\sigma^2} dv$$

Parameters: $-\infty < \mu < \infty$, $\sigma > 0$. The mean (μ) represents the location parameter and the standard deviation (σ) the scale parameter.

Domain: $-\infty < x < \infty$

Mean (μ): μ

Variance (σ^2): σ^2

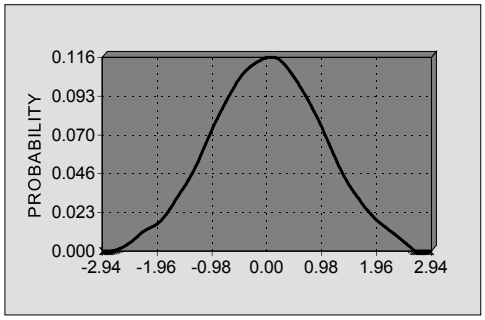
Mode: μ

Coefficient of skewness (α_3): 0

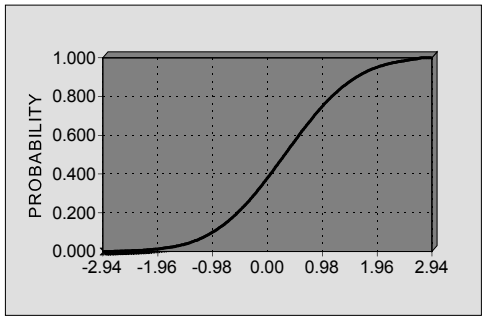
Coefficient of kurtosis (α_4): 3

Properties of the Normal Distribution

Probability density function for N(μ = 0, σ = 1) as an output of @RISK simulation



Cumulative distribution function for N(μ = 0, σ = 1) as an output of @RISK simulation

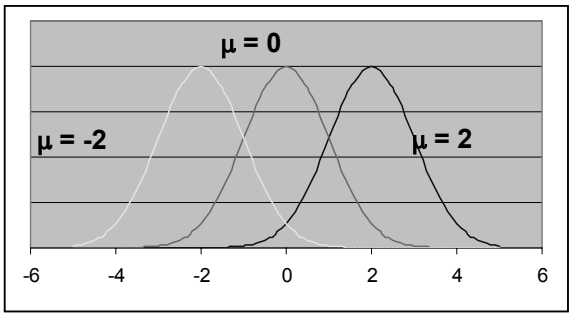


Properties of the Normal Distribution

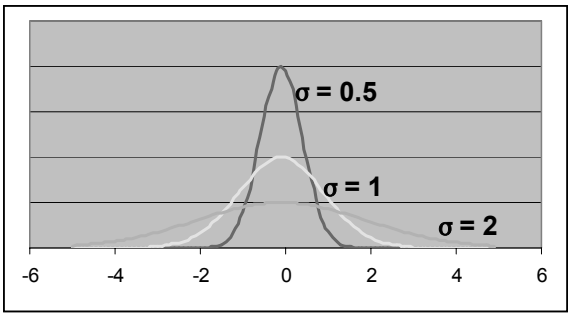
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

where e = 2.718 and π = 3.142; μ and σ (σ > 0) are the parameters that represent the population mean and standard deviation. The population mean μ locates the center of the distribution and the distribution is symmetric about μ . The areas to the right and left of μ are both .5. The shape of the distribution is determined by σ, the population standard deviation.

The bell-shaped normal curves change as μ varies and σ remains fixed (σ = 1).



In the graph below, the curves change as σ varies and μ remains fixed (μ = 0).



Properties of the Normal Distribution

Large values of σ reduce the height of the curve and increase the spread. Small values of σ increase the height of the curve and reduce the spread. Biological phenomena result in data distributed in a manner sufficiently normal that the distribution is the basis of much of the statistical theory used by biologists. Many positive random variables, such as heights, weights of animals and muscle groups, viral titres in tissues, viral titres in vaccines, physiological characteristics, pH of tissues and body fluids, milk and egg production, and time, have distributions that might be well approximated by a normal distribution.

Since the normal distribution has infinite tails, for physical quantities which are non-negative, the coefficient of variation should be less than 0.3. The coefficient of variation is defined as the sample standard deviation expressed as a percentate of the sample mean as shown in:

$$CV = \frac{100s}{\bar{Y}}$$

The CV is a relative measure of variation, in contrast to the standard deviation, which is in the same units as the observations. If the coefficient of variation is much higher than this, there is a significant probability of predicting negative values. With a high coefficient of variation, then a normality assumption may be inappropriate.

Properties of the Normal Distribution

Pertaining to the graphs on the previous slide, some important properties of the normal distribution need emphasizing:

- **The mode, which is the point on the horizontal axis where the curve is a maximum, occurs at $x = \mu$.**
- **The curve is symmetric about a vertical axis through the mean μ.**
- **The curve has its points of inflection at $x = \mu \pm \sigma$, is concave downward if $\mu - \sigma < X < \mu + \sigma$, and is concave upward otherwise.**
- **The normal curve approaches the horizontal axis asymptotically as we proceed in either direction away from the mean.**
- **The total area under the curve and above the horizontal axis is equal to 1.**

Properties of the Normal Distribution

Standardized Normal Distribution or Standard Normal Random Variable

The probability that a continuous random variable x assumes a value in the interval a to b is the area under the probability density function between the points a and b. A standardization procedure permits the use of the same tables of probabilities for all normal distributions.

A normal random variable is standardized by expressing its value as the number of standard deviations (σ) it lies to the left or right of its mean μ. The standardized normal random variable z, is defined as $z = (x - \mu)/\sigma$, or equivalently, $x = \mu + z\sigma$.

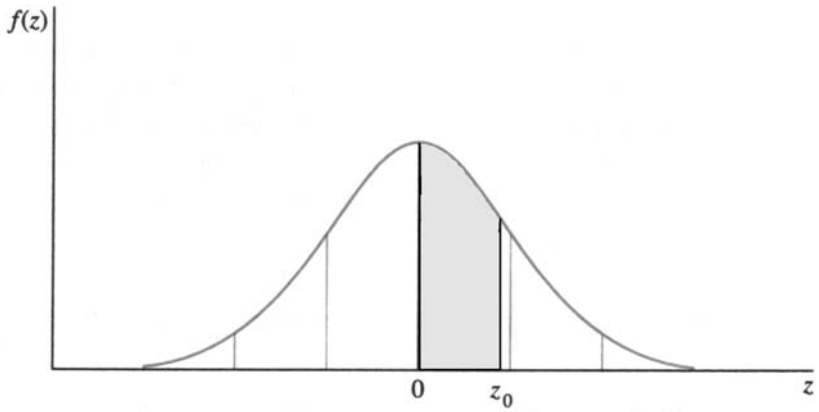
If x is less that the mean μ, the value of z is negative.
If x is greater that the mean μ, the value of z is positive.
If $x = \mu$, the value of z = 0.

The standard probability distribution has a mean of zero and a standard deviation of 1.

Properties of the Normal Distribution

Standardized Normal Distribution or Standard Normal Random Variable

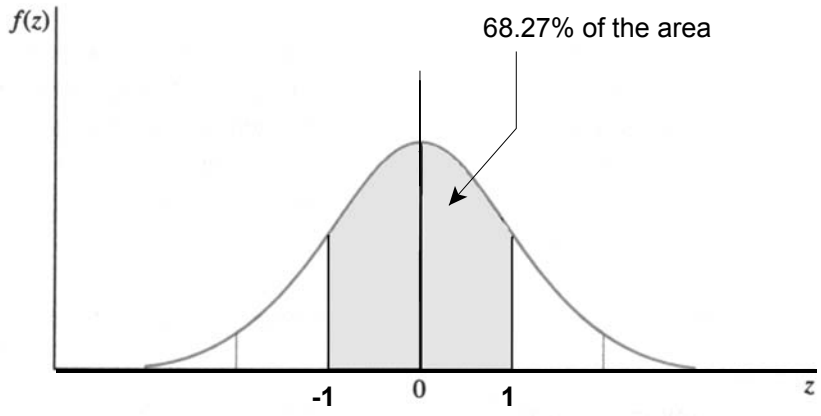
The area under the standard normal curve between mean z = 0 and a specified positive value of z, z_0, is the probability $P(0 \le z \le z_0)$.

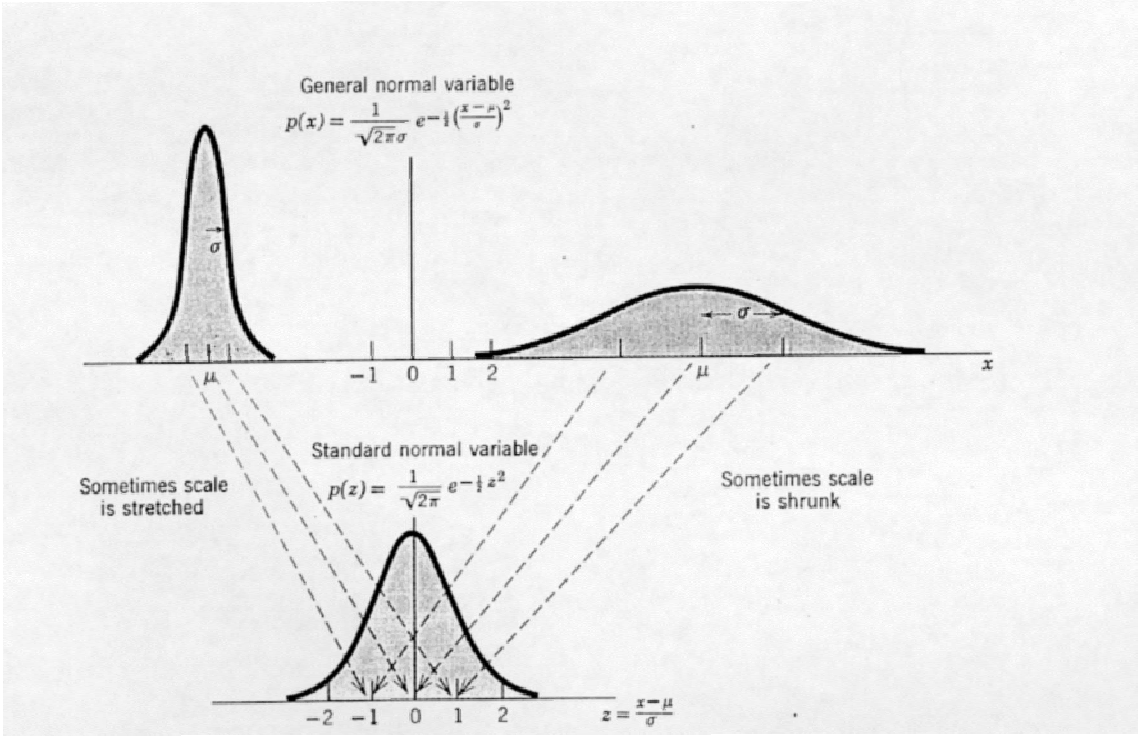


Properties of the Normal Distribution

Standardized Normal Distribution or Standard Normal Random Variable

The area under the standard normal curve between z = -1 and z = 1 is the probability P (-1 ≤ z ≤ 1) = 0.8413 - 0.1587 = 0.6826.





Properties of the Normal Distribution

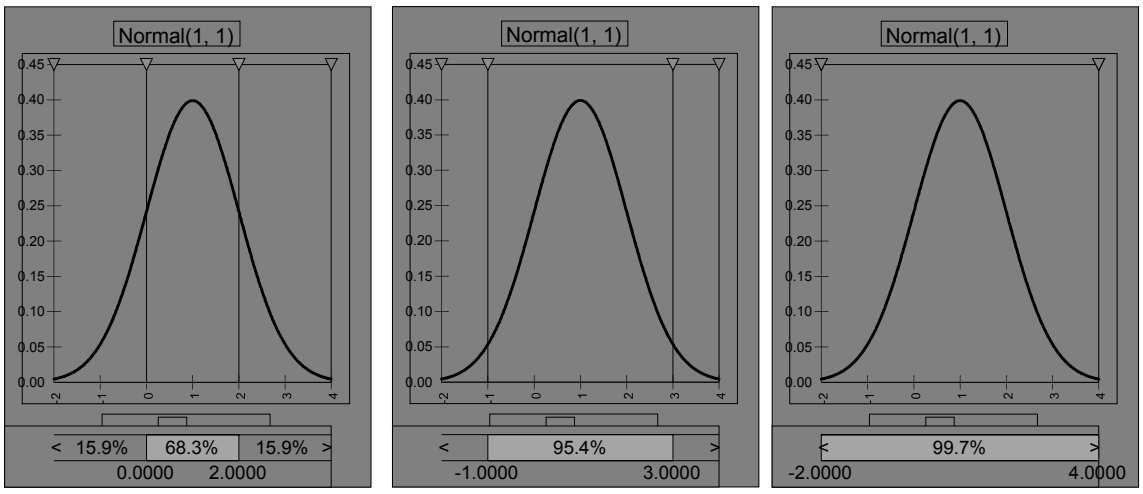
If $X \approx N(\mu, \sigma^2)$, then $Y = cX + d \approx N(c\mu + d, c^2\sigma^2)$ for $c, d \in \mathbb{R}$
If $X_1 \approx N(\mu_1, \sigma^2_1)$ and $X_2 \approx N(\mu_2, \sigma^2_2)$ are independent random variables, then $X_1 \pm X_2 \approx N(\mu_1 \pm \mu_2, \sigma^2_1 \pm \sigma^2_2)$.

The first and third quartiles are given as $Q_1 = \mu - \frac{2}{3}\sigma$ and $Q_3 = \mu + \frac{2}{3}\sigma$ so that the interquartile range becomes $Q_3 - Q_1 = 1\frac{1}{3}\sigma$.

Under the normal curve, 68.26%. 95.44%, 99.73% and 99.994% of the area lie respectively within 1, 2, 3 and 4 standard deviations from the mean.

Properties of the Normal Distribution

Normal distribution (1, 1) indicating the area under the curve within 1, 2 and 3 standard deviations from the mean.



Central Limit Theorem

The Central Limit Theorem states that, under rather general conditions, sums and means of random samples of measurements drawn from a population tend to possess an approximately normal distribution. In other words, for random samples from any underlying distribution with mean μ and variance σ^2, however nonnormal it may be, the sampling distribution of $\bar{x}$, the sample mean fro random samples of size *n*, is approximately normal, the approximation improving as *n* increases. The approximating normal distribution will, of course, have mean μ and variance σ^2/n. The only assumptions are that random sampling has been achieved and that a mean and variance exists. All distributions from a very wide class, including continuous, discrete, skewed, symmetric, unimodal and multi-modal distributions produce samples means that tend to be normally distributed.

Central Limit Theorem
If random samples of *n* observations are drawn from a nonnormal population with finite mean μ and standard deviation σ, then, when *n* is large, the sampling distribution of the sample mean is approximately normally distributed, with mean and standard deviation:

$$\mu_{\bar{x}} = \mu, \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

The approximation becomes more accurate as *n* becomes large.
Regardless of its shape, the sampling distribution of $\bar{x}$ always has a mean identical to the mean of the sampled population and a standard deviation equal to the population standard deviation σ divided by $\sqrt{n}$. The spread of the distribution of sample means is considerably less than the spread of the sampled population.

Central Limit Theorem

The Central Limit Theorem can be restated to apply to the sum of the sample measurements. The sum of the sample measurements $\sum x_i$ also has an approximately normal distribution with mean $n\mu$ and standard deviation $\sigma\sqrt{n}$ as *n* becomes large.

The important contribution of the Central Limit Theorem is in statistical inference. Many estimators that are used to make inferences about population parameters are sums or averages of the sample measurements. When the sample size is sufficiently large, you may expect these estimators to have sampling distributions that are approximately normal. One can then describe the behaviour of these estimators in repeated sampling and evaluate the probability of observing certain sample results using the normal distribution. These probabilities are calculated in terms of the standard normal random variable:

$$z = \frac{Estimator - mean}{S\tan dard\ deviation}$$

The appropriate value of *n* depends on the shape of the population from which you sample. If the sampled population is normal, then the sampling distribution of $\bar{x}$ will also be normal, no matter what sample size. When the sampled population is approximately symmetric, the sampling distribution of $\bar{x}$ becomes approximately normal for relatively small values of *n*. When the sampled population is skewed, the sample size *n* must be larger, with n at least 30 before the sampling distribution of $\bar{x}$ becomes approximately normal.

Out-to-Sea Example

Dear Abby;
You wrote in your column that a woman is pregnant for 266 days. Who said so? I carried my baby for ten months and five days, and there is no doubt about it because I know the exact date my baby was conceived. My husband is in the Navy and it couldn't have possibly been conceived any other time because I saw him only once for an hour, and I didn't see him again until the day before the baby was born.

I don't drink or run around, and there is no way this baby isn't his, so please print a retraction about the 266-day carrying time because otherwise I am in a lot of trouble. Signed Worried Writer.

To help answer the question, one can estimate the statistical likelihood of a pregnancy being 310 days long. One needs to compute the probability that, by chance alone, a pregnancy will be 310 days long or longer. If Y represents an arbitrary pregnancy duration, one wants to compute P(Y≥310), the smaller this probability is, the less credibility Worried Writer has.

According to evidence-based norms, the mean and standard deviation for Y are 266 days and 16 days, respectively. Assuming that the distribution of Y is normal,

$$P(Y \geq 310) = P\left(\frac{Y - 266}{16} \geq \frac{310 - 266}{16}\right) = P(Z \geq 2.75).$$

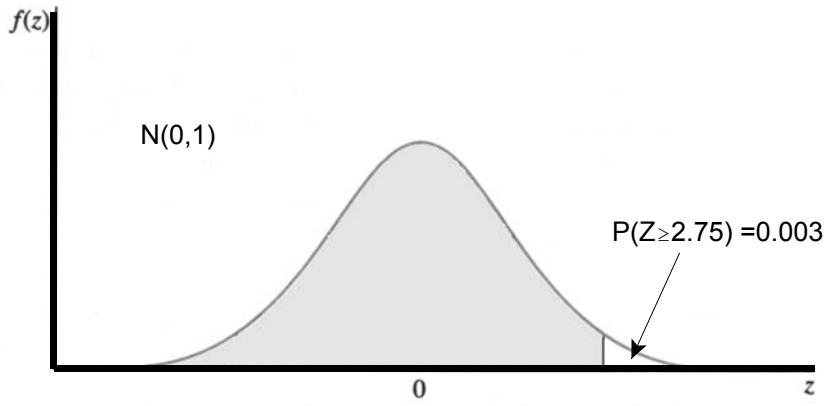
Out-to-Sea Example

From the table below (computed in Microsoft Excel using the NORMSDIST function) P(Z≥2.75) = 1 - 0.9970 = 0.003.

Cumulative Area under the Standard Normal Distribution										
Z	0	1	2	3	4	5	6	7	8	9
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Out-to-Sea Example

The yellow area of the graph comprises the cumulative probability of 0.9970 whereas the area of interest is to the right of 2.75 on the standard normal curve.



Cumulative area under the standard normal curve.

CWD Incubation Example

The incubation period of CWD in farmed elk varies from 3 to 20 years, a mean value of 8 years and a standard deviation of 4 years. The incubation periods of 30 elk that developed CWD were randomly selected from a database of cases in North America. The question of interest is the probability that the sample mean incubation period is less than 7 years.

First, the shape of the sampled population is not symmetric because the mean $\mu = 8$ does not lie midway between the maximum and minimum values. Since the mean is closer to the minimum, the distribution is skewed to the right. Despite the shape of the population distribution, the sampling distribution has a mean $\mu = 8$ and a standard error of:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{4}{\sqrt{30}} = 0.73$$

The probability that $\bar{x}$ is less than 7 is estimated by calculating the z value corresponding to $\bar{x} = 7$.

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{7 - 8}{0.73} = -1.37$$

From tables of cumulative areas under the standard normal distribution or using the Microsoft Excel funtion NORMSDIST(-1.37), the area to the left of -1.37, $P(\bar{x} < 7) = P(z < -1.37) = 0.0853$.

CWD Incubation Example

The probability that $\bar{x}$ is less than 7 is given by the yellow area in the graph below. This area represents $P(\bar{x} < 7) = P(z < -1.37)$.

