Probability Distributions for Continuous Variables

Let X = lake depth at a randomly chosen point on lake surface

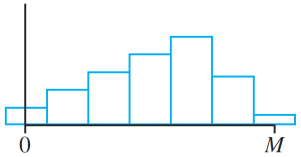
Let M = the maximum depth (in meters), so that any number in the interval [0, M] is a possible value of X.

If we "discretize" X by measuring depth to the nearest meter, then possible values are nonnegative integers less than or equal to M.

The resulting discrete distribution of depth can be pictured using a histogram.

Probability Distributions for Continuous Variables

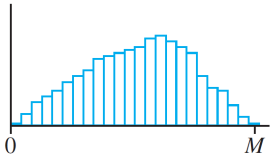
If we draw the histogram so that the area of the rectangle above any possible integer k is the proportion of the lake whose depth is (to the nearest meter) k, then the total area of all rectangles is 1:



Probability histogram of depth measured to the nearest meter

Probability Distributions for Continuous Variables

If depth is measured much more accurately, each rectangle in the resulting probability histogram is much narrower, though the total area of all rectangles is still 1.

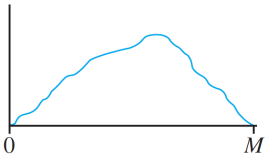


Probability histogram of depth measured to the nearest centimeter

Probability Distributions for Continuous Variables

If we continue in this way to measure depth more and more finely, the resulting sequence of histograms approaches a smooth curve.

Because for each histogram the total area of all rectangles equals 1, the total area under the smooth curve is also 1.



A limit of a sequence of discrete histograms

Probability Distributions for Continuous Variables

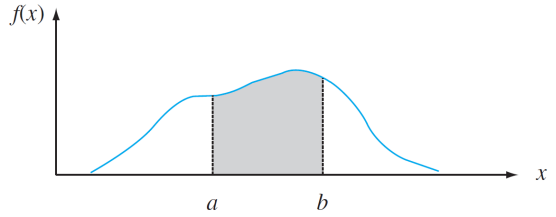
Definition

Let X be a continuous rv. Then a **probability distribution** or **probability density function** (pdf) of X is a function $f(x)$ such that for any two numbers a and b with $a \le b$,

$$P(a \le X \le b) = \int_a^b f(x)dx$$

Probability Distributions for Continuous Variables

The probability that X takes on a value in the interval $[a, b]$ is the area above this interval and under the graph of the density function:



$P(a \le X \le b)$ = the area under the density curve between a and b

Probability Distributions for Continuous Variables

For $f(x)$ to be a legitimate pdf, it must satisfy the following two conditions:

1. $f(x) \ge 0$ for all x
2. $\int_{-\infty}^{\infty} f(x)dx$ = area under the entire graph of $f(x)$ = 1

Example

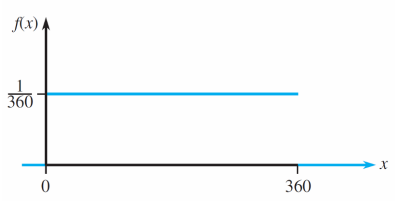
Consider the reference line connecting the valve stem on a tire to the center point.

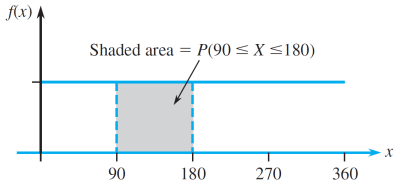
Let X be the angle measured clockwise to the location of an imperfection. One possible pdf for X is

$$f(x) = \begin{cases} \frac{1}{360} & 0 \le x < 360 \\ 0 & \text{otherwise} \end{cases}$$

Example, cont

cont'd





$$P(90 \le X \le 180) = \int_{90}^{180} \frac{1}{360}\, dx$$

$$= \frac{1}{4} = .25$$

Copyright Prof. Vanja Dukic, Applied Mathematics, CU-Boulder

Probability Distributions for Uniform Variables

Definition

A continuous rv X is said to have a **uniform distribution** on the interval $[A, B]$ if the pdf of X is

$$f(x; A, B) = \begin{cases} \dfrac{1}{B - A} & A \le x \le B \\ 0 & \text{otherwise} \end{cases}$$

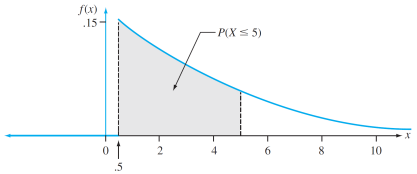
Copyright Prof. Vanja Dukic, Applied Mathematics, CU-Boulder

Exponential

"Time headway" in traffic flow is the elapsed time between the time that one car finishes passing a fixed point and the instant that the next car begins to pass that point.

Let X = the time headway for two randomly chosen consecutive cars on a freeway during a period of heavy flow

$$f(x) = \begin{cases} .15e^{-.15(x-.5)} & x \ge .5 \\ 0 & \text{otherwise} \end{cases}$$



Copyright Prof. Vanja Dukic, Applied Mathematics, CU-Boulder

Exponential example , cont

cont'd

Then

$$\int_{-\infty}^{\infty} f(x)\, dx = \int_{.5}^{\infty} .15e^{-.15(x-.5)}\, dx$$

$$= .15e^{.075} \int_{.5}^{\infty} e^{-.15x}\, dx$$

$$= .15e^{.075} \cdot \frac{1}{.15} e^{-(.15)(.5)}$$

$$= 1$$

Copyright Prof. Vanja Dukic, Applied Mathematics, CU-Boulder

Example , cont

cont'd

The probability that headway time is at most 5 sec is

$$P(X \le 5) = \int_{-\infty}^{5} f(x)\,dx$$

$$= \int_{.5}^{5} 15e^{-.15(x-.5)}\,dx$$

$$= .15e^{.075}\int_{.5}^{5} e^{-.15x}\,dx$$

$$= .15e^{.075}\cdot\left(-\frac{1}{.15}e^{-.15x}\Big|_{x=.5}^{x=5}\right)$$

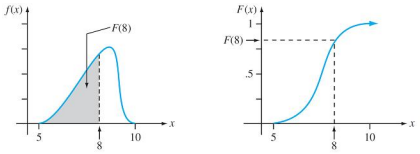
The Cumulative Distribution Function

The Cumulative Distribution Function

The **cumulative distribution function** $F(x)$ for a continuous rv X is defined for every number x by

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(y)\,dy$$

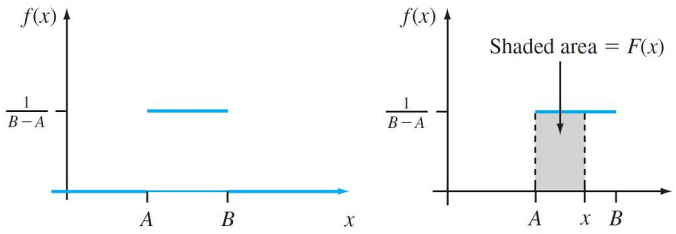
For each x, $F(x)$ is the area under the density curve to the left of x.



A pdf and associated cdf

Example

Let X, the thickness of a certain metal sheet, have a uniform distribution on $[A, B]$.



The pdf for a uniform distribution

Example , cont

cont'd

For $x < A$, $F(x) = 0$, since there is no area under the graph of the density function to the left of such an x.

For $x \geq B$, $F(x) = 1$, since all the area is accumulated to the left of such an x. Finally for $A \leq x \leq B$,

$$F(x) = \int_{-\infty}^{x} f(y)dy = \int_{A}^{x} \frac{1}{B-A}\,dy = \frac{1}{B-A} \cdot y \Big|_{y=A}^{y=x} = \frac{x-A}{B-A}$$

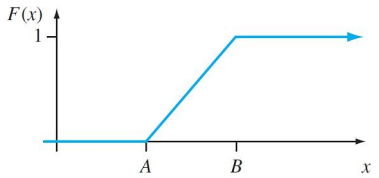
Example , cont

cont'd

The entire cdf is

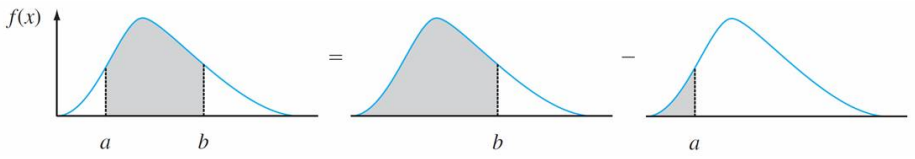
$$F(x) = \begin{cases} 0 & x < A \\ \dfrac{x-A}{B-A} & A \leq x < B \\ 1 & x \geq B \end{cases}$$

The graph of this cdf is



The cdf for a uniform distribution

Using $F(x)$ to Compute Probabilities



Percentiles of a Continuous Distribution

When we say that an individual's test score was at the 85th percentile of the population, we mean that 85% of all population scores were below that score and 15% were above.

Similarly, the 40th percentile is the score that exceeds 40% of all scores and is exceeded by 60% of all scores.

Percentiles of a Continuous Distribution

Proposition

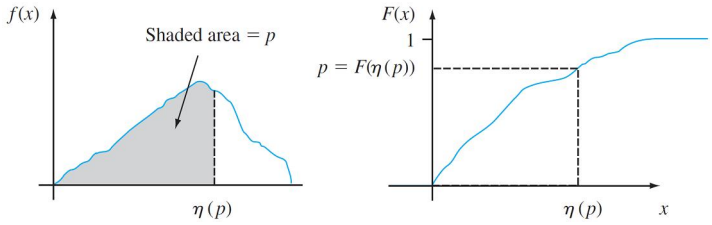
Let p be a number between 0 and 1. The **(100p)th percentile** of the distribution of a continuous rv X, denoted by $\eta(p)$, is defined by

$$p = F(\eta(p)) = \int_{-\infty}^{\eta(p)} f(y)\,dy$$

$\eta(p)$ is that value on the measurement axis such that 100p% of the area under the graph of $f(x)$ lies to the left of $\eta(p)$ and 100(1 – p)% lies to the right.

Percentiles of a Continuous Distribution

Thus $\eta(.75)$, the 75th percentile, is such that the area under the graph of $f(x)$ to the left of $\eta(.75)$ is .75.



The (100p)th percentile of a continuous distribution

Example 9

The distribution of the amount of gravel (in tons) sold by a particular construction supply company in a given week is a continuous rv X with pdf

$$f(x) = \begin{cases} \frac{3}{2}(1 - x^2) & 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

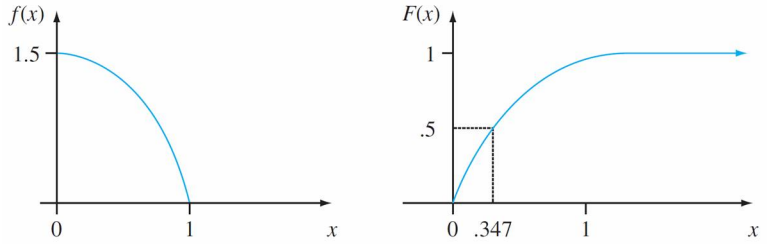
The cdf of sales for any x between 0 and 1 is

$$F(x) = \int_0^x \frac{3}{2}(1 - y^2)\,dy = \frac{3}{2}\left(y - \frac{y^3}{3}\right)\Big|_{y=0}^{y=x} = \frac{3}{2}\left(x - \frac{x^3}{3}\right)$$

Example 9

cont'd

The graphs of both $f(x)$ and $F(x)$ are



Example 9 cont'd

The (100p)th percentile of this distribution satisfies the equation

$$p = F(\eta(p)) = \frac{3}{2}\left[\eta(p) - \frac{(\eta(p))^3}{3}\right]$$

that is,

$$(\eta(p))^3 - 3\eta(p) + 2p = 0$$

For the 50th percentile, p = .5, and the equation to be solved is $\eta^3 - 3\eta + 1 = 0$; the solution is $\eta = \eta(.5) = .347$. If the distribution remains the same from week to week, then in the long run 50% of all weeks will result in sales of less than .347 ton and 50% in more than .347 ton.

Percentiles of a Continuous Distribution

Definition

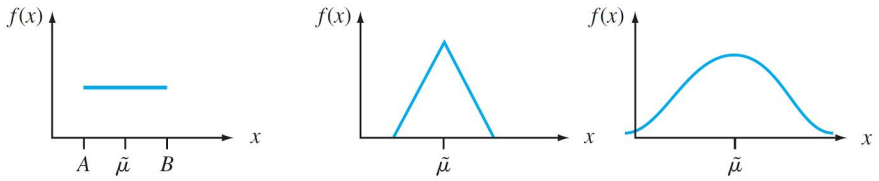
The **median** of a continuous distribution is the 50th percentile, so satisfies .5 = $F(\tilde{\mu})$

That is, half the area under the density curve is to the left of $\tilde{\mu}$ and half is to the right of $\tilde{\mu}$

A continuous distribution whose pdf is **symmetric**–the graph of the pdf to the left of some point is a mirror image of the graph to the right of that point–has median equal to the point of symmetry, since half the area under the curve lies to either side of this point.

Percentiles of a Continuous Distribution

Examples



Medians of symmetric distributions

Expected Values

Definition

The **expected** or **mean value** of a continuous rv X with the pdf $f(x)$ is:

$$\mu_X = E(X) = \int_{-\infty}^{\infty} x\ f(x)\,dx$$

Example, cont

The pdf of the amount of weekly gravel sales X is:

$$f(x) = \begin{cases} \frac{3}{2}(1 - x^2) & 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

So

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x)\, dx = \int_{0}^{1} x \cdot \frac{3}{2}(1 - x^2)\, dx$$

$$= \frac{3}{2}\int_{0}^{1}(x - x^3)\, dx = \frac{3}{2}\left(\frac{x^2}{2} - \frac{x^4}{4}\right)\Big|_{x=0}^{x=1} = \frac{3}{8}$$

Expected Values of functions of r.v.

If $h(X)$ is a function of X, then

$$E[h(X)] = \mu_{h(X)} = \int_{-\infty}^{\infty} h(x)\ f(x)\, dx$$

For $h(X)$, a linear function,

$$E[h(X)] = E(aX + b) = a\, E(X) + b$$

Variance

The **variance** of a continuous random variable X with pdf $f(x)$ and mean value μ is

$$\sigma_X^2 = V(X) = \int_{-\infty}^{\infty} (x - \mu)^2\ f(x)\, dx = E[(X - \mu)^2]$$
$$= E(X^2) - [E(X)]^2$$

The **standard deviation** (SD) of X is $\sigma_X = \sqrt{V(X)}$

When $h(X) = aX + b$, the expected value and variance of $h(X)$ satisfy the same properties as in the discrete case:

$$E[h(X)] = a\,\mu + b \quad \text{and} \quad V[h(X)] = a^2\ \sigma^2.$$

Example, cont.

For weekly gravel sales, we computed E(X) = $\frac{3}{8}$

$$E(X^2) = \int_{-\infty}^{\infty} x^2\ f(x)\, dx$$

$$= \int_{0}^{1} x^2\,(1 - x^2)\, dx$$

$$= \frac{3}{2}\int_{0}^{1} (x^2 - x^4)\, dx = \frac{1}{5}$$

$$V(X) = \frac{1}{5} - \left(\frac{3}{8}\right)^2 = .059$$

The Normal Distribution

The normal distribution is probably the most important distribution in all of probability and statistics.

Many populations have distributions that can be fit very closely by an appropriate normal (Gaussian, bell) curve.

Examples include

heights, weights, and other physical characteristics

scores on various tests,

etc.

The Normal Distribution

Definition

A continuous rv X is said to have a **normal distribution** with parameters μ and σ (or μ and σ^2), if the pdf of X is

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

e denotes the base of the natural logarithm system and equals approximately 2.71828

π is a mathematical constant with approximate value 3.14159.

The Normal Distribution

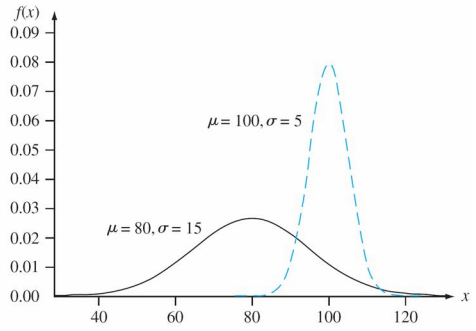
The statement that X is normally distributed with parameters μ and σ^2 is often abbreviated $X \sim N(\mu, \sigma^2)$.

Clearly $f(x; \mu, \sigma) \geq 0$, but a somewhat complicated calculus argument must be used to verify that $\int_{-\infty}^{\infty} f(x; \mu, \sigma)\, dx = 1$.

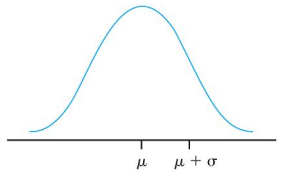
Similarly, it can be shown that $E(X) = \mu$ and $V(X) = \sigma^2$, so the parameters are the mean and the standard deviation of X.

The Normal Distribution

Graphs of $f(x; \mu, \sigma)$ for several different (μ, σ) pairs.



Two different normal density curves



Visualizing μ and σ for a normal distribution

The Standard Normal Distribution

The standard normal distribution almost never serves as a model for a naturally arising population.

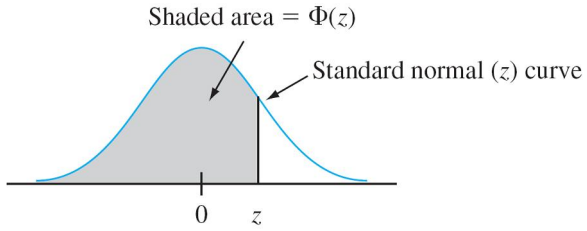
Instead, it is a **reference distribution** from which information about other normal distributions can be obtained via a simple formula.

$\Phi(z) = P(Z \le z)$, the area under the standard normal density curve to the left of z

This can also be computed with a single command in R, Matlab, Mathematica…

The Standard Normal Distribution

Figure below illustrates the probabilities



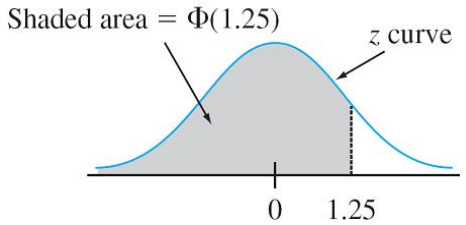
Example

cont'd

$P(Z \le 1.25) = \Phi(1.25)$,

The number is .8944, so $P(Z \le 1.25) = .8944$.

Figure below illustrates this probability:



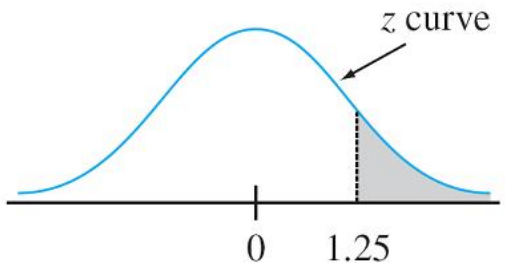
Example , cont.

cont'd

b) Since Z is a continuous rv,

$P(Z \ge 1.25) = 1 - P(Z < 1.25) =$

$= 1 - P(Z \le 1.25) = 1 - 0.8944 = 0.1056$



Example , cont.

cont'd

c. $P(Z \le -1.25) = \Phi(-1.25)$, a lower-tail area.

$$\Phi(-1.25) = .1056$$

By symmetry – the left tail is the same as the right tail, so this is the same answer as in part (b)

d. $P(-.38 \le Z \le 1.25)$ is the area under the standard normal curve above the interval whose left endpoint is -.38 and whose right endpoint is 1.25.

Recall, $P(a \le X \le b) = F(b) - F(a)$. Thus:

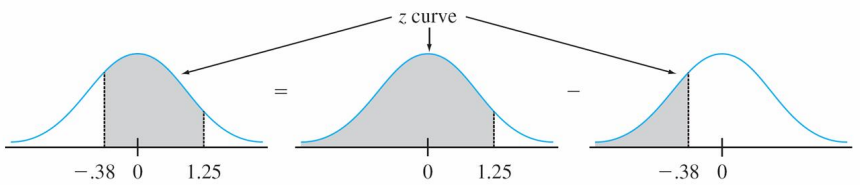
$$P(-.38 \le Z \le 1.25) = \Phi(1.25) - \Phi(-.38)$$

$$= .8944 - .3520$$

$$= .5424$$

Example , cont.

cont'd



$P(-.38 \le Z \le 1.25)$ as the difference between two cumulative areas

Percentiles of the Standard Normal Distribution

Example

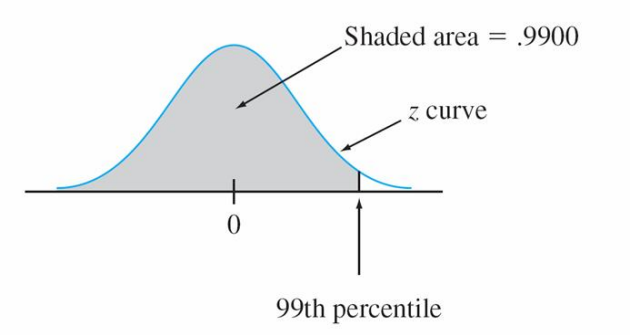
The 99th percentile of the standard normal distribution is that value of z such that the area under the *z* curve to the left of the value is .99

So far: for a fixed *z* the area under the standard normal curve to the left of *z*

Now: we have the area and want the value of *z*.

This is the "inverse" problem to $P(Z \le z) = ?$

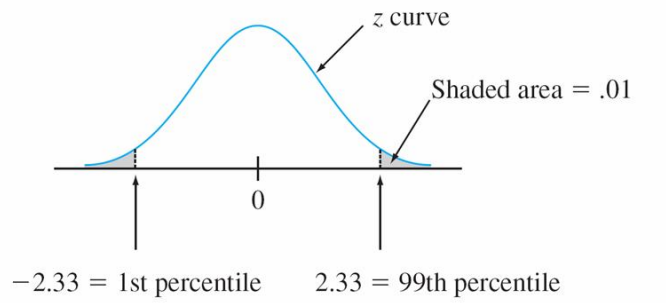
Example cont'd



The 99th percentile is (approximately) $z = 2.33$.

Example cont'd

By symmetry, the first percentile is as far below 0 as the 99th is above 0, so equals –2.33 (1% lies below the first and also above the 99th).

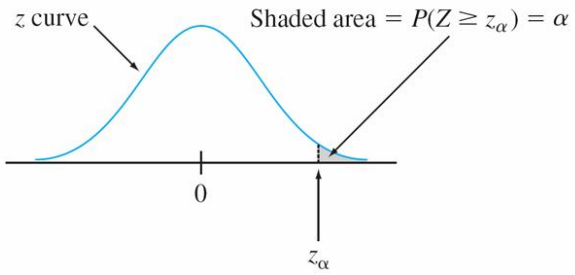


z_α Notation

In statistical inference, later, we will need the z values that give certain tail areas under the standard normal curve.

There, this notation will be standard:

z_α will denote the z value for which α of the area under the z curve lies to the right of z_α.



z_α Notation for z Critical Values

For example, $z_{.10}$ captures upper-tail area .10, and $z_{.01}$ captures upper-tail area .01.

Since α of the area under the z curve lies to the right of z_α, $1 - \alpha$ of the area lies to its left.

Thus z_α *is the* 100(1 – α)*th percentile of the standard normal distribution*.

By symmetry the area under the standard normal curve to the left of $-z_\alpha$ is also α. The z_α are usually referred to as **z critical values.**

z_α Notation for z Critical Values

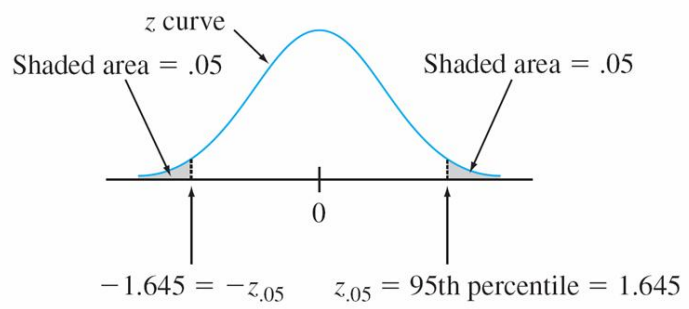
Table below lists the most useful z percentiles and z_α values.

Percentile	90	95	97.5	99	99.5	99.9	99.95
α (tail area)	.1	.05	.025	.01	.005	.001	.0005
$z_\alpha = 100(1-\alpha)$th percentile	1.28	1.645	1.96	2.33	2.58	3.08	3.27

Example – critical values

$z_{.05}$ is the 100(1 – .05)th = 95th percentile of the standard normal distribution, so $z_{.05}$ = 1.645.

The area under the standard normal curve to the left of $-z_{.05}$ is also .05



Nonstandard Normal Distributions

When $X \sim N(\mu, \sigma^2)$, probabilities involving X are computed by "standardizing." The **standardized variable** is $(X - \mu)/\sigma$.

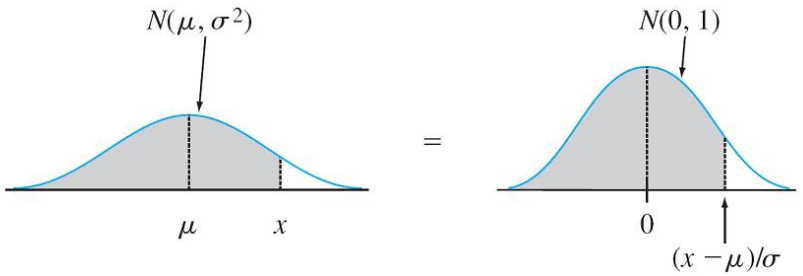
Subtracting μ shifts the mean from μ to zero, and then dividing by σ scales the variable so that the standard deviation is 1 rather than σ.

Proposition
If X has a normal distribution with mean μ and standard deviation σ, then

$$Z = \frac{X - \mu}{\sigma}$$

Nonstandard Normal Distributions

The key idea: by standardizing, any probability involving normally distributed X can be computed using standardized probabilities.



Equality of nonstandard and standard normal curve areas

Nonstandard Normal Distributions

$$P(a \le X \le b) = P\left(\frac{a-\mu}{\sigma} \le Z \le \frac{b-\mu}{\sigma}\right)$$

$$= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

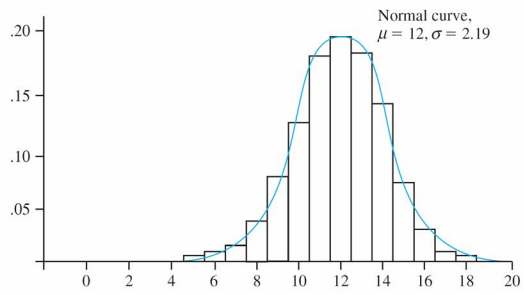
$$P(X \le a) = \Phi\left(\frac{a-\mu}{\sigma}\right) \qquad P(X \ge b) = 1 - \Phi\left(\frac{b-\mu}{\sigma}\right)$$

Using Normal to approximate the Binomial Distribution

Copyright Prof. Vanja Dukic, Applied Mathematics, CU-Boulder STAT 4000/5000

Approximating the Binomial Distribution

Figure below displays a binomial probability histogram for the binomial distribution with $n = 20$, $p = .6$,
for which $\mu = 20(.6) = 12$ and $\sigma = \sqrt{20(.6)(.4)} = 2.19.$



Binomial probability histogram for $n = 20$, $p = .6$ with normal approximation curve superimposed

Copyright Prof. Vanja Dukic, Applied Mathematics, CU-Boulder STAT 4000/5000

Approximating the Binomial Distribution

Let X be a binomial rv based on n trials with success probability p. Then if np is large (the binomial probability histogram is not too skewed), X has approximately a normal distribution with $\mu = np$ and $\sigma = \sqrt{npq}$.

In particular, for x = a possible value of X,

$$P(X \le x) = B(x, n, p) \approx \begin{pmatrix}\text{area under the normal curve} \\ \text{to the left of } x + .5\end{pmatrix}$$

$$= \Phi\left(\frac{x + .5 - np}{\sqrt{npq}}\right)$$

Copyright Prof. Vanja Dukic, Applied Mathematics, CU-Boulder STAT 4000/5000

Exponential Distribution

The Exponential Distributions

The family of exponential distributions provides probability models that are very widely used in engineering and science disciplines.

Definition

X is said to have an **exponential distribution** with the **rate parameter** λ ($\lambda > 0$) if the pdf of *X* is

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases} \qquad \textbf{(4.5)}$$

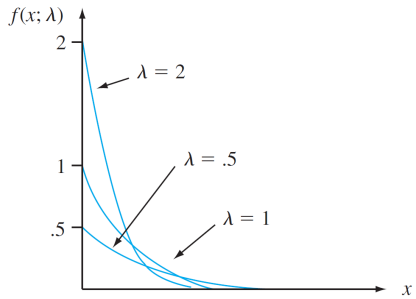
The Exponential Distributions

Integration by parts give the following results:

$$\mu = \frac{1}{\lambda} \qquad \sigma^2 = \frac{1}{\lambda^2}$$

Both the mean and standard deviation of the exponential distribution equal $1/\lambda$.

Several members of Exponential d'n →



CDF:

$$F(x;\lambda) = \begin{cases} 0 & x < 0 \\ 1 - e^{-\lambda x} & x \geq 0 \end{cases}$$

The Exponential Distributions

The exponential distribution is frequently used as a model for the distribution of times between the occurrence of successive events:

Suppose that the count of events follows a Poisson process with rate α (ie, mean αt for any time interval *t*).

Then the distribution of elapsed time between the occurrence of two successive events is exponential with parameter $\lambda = \alpha$.

The Exponential Distributions

Although a complete proof is beyond the scope of the course, the result is easily verified for the time X_1 until the first event occurs:

$P(X_1 \le t) = 1 - P(X_1 > t) = 1 - P$ [no events in $(0, t)$]

$$= 1 - \frac{e^{-\alpha t} \cdot (\alpha t)^0}{0!} = 1 - e^{-\alpha t}$$

which is exactly the cdf of the exponential distribution.

Example

Suppose that calls are received at an emergency room switchboard according to a Poisson process with rate $\alpha = .5$ call **per day**.

Then the **number of days** X between successive calls has an Exp distribution with parameter 0.5.

Ex: The probability that more than 2 days elapse between calls is then:

$P(X > 2) = 1 - P(X \le 2)$

$= 1 - F(2; .5)$

$= 1 - (1 - e^{-(.5)(2)}) = .368$

And the expected time between successive calls is 1/.5 = 2 days.

The Exponential Distributions

Another important application of the exponential distribution is to model the distribution of lifetimes.

A partial reason for the popularity of such applications is the **"memoryless" property** of the Exp distribution.

The Exponential Distributions

Suppose a light bulb's lifetime is exponentially distributed with parameter λ.

Say you turn the light on, and then we leave and come back after t_0 hours to find it still on. What is the probability that the light bulb will last for *at least additional t* hours?

In symbols, we are looking for $P(X \ge t + t_0 \mid X \ge t_0)$.

By the definition of conditional probability,

$$P(X \ge t + t_0 \mid X \ge t_0) = \frac{P[(X \ge t + t_0) \cap (X \ge t_0)]}{P(X \ge t_0)}$$

The Exponential Distributions

But the event $X \geq t_0$ in the numerator is redundant, since both events can only occur if $X \geq t + t_0$. Therefore,

$$P(X \geq t + t_0 | X \geq t_0) = \frac{P(X \geq t + t_0)}{P(X \geq t_0)} = \frac{1 - F(t + t_0; \lambda)}{1 - F(t_0; \lambda)} = e^{-\lambda t}$$

This conditional probability is identical to the original probability $P(X \geq t)$ that the component lasted t hours.

It's as if the light bulb "forgot" it was on.

The Gamma Distribution

Copyright Prof. Vanja Dukic, Applied Mathematics, CU-Boulder STAT 4000/5000

The Gamma Function

To define the family of gamma distributions, we first need to introduce a function that plays an important role in many branches of mathematics.

Definition

For $\alpha > 0$, the **gamma function** $\Gamma(\alpha)$ is defined by

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\, dx$$

Copyright Prof. Vanja Dukic, Applied Mathematics, CU-Boulder STAT 4000/5000

The Gamma Function

The most important properties of the gamma function are the following:

1. For any $\alpha > 1$, $\Gamma(\alpha) = (\alpha - 1)\,\Gamma(\alpha - 1)$ [via integration by parts]

2. For any positive integer, n, $\Gamma(n) = (n - 1)!$

3. $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$

Copyright Prof. Vanja Dukic, Applied Mathematics, CU-Boulder STAT 4000/5000

The Gamma Function

So if we let

$$f(x;\alpha) = \begin{cases} \dfrac{x^{\alpha-1}e^{-x}}{\Gamma(\alpha)} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

then f(x; α) ≥ 0 and $\int_0^\infty f(x;\alpha)\,dx = \Gamma(\alpha)/\Gamma(\alpha) = 1$

so *f*(*x*; a) satisfies the two basic properties of a pdf.

The Gamma Distribution

Definition

A continuous random variable *X* is said to have a **gamma distribution** if the pdf of *X* is

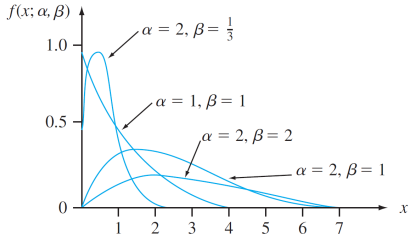
$$f(x;\alpha,\beta) = \begin{cases} \dfrac{1}{\beta^\alpha\Gamma(\alpha)}x^{\alpha-1}e^{-x/\beta} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

where the parameters α and β satisfy $\alpha > 0$, $\beta > 0$. The **standard gamma distribution** has $\beta = 1$.

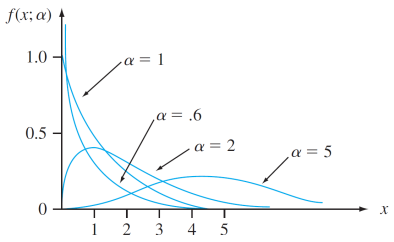
The Gamma Distribution

The Exp dist results from taking $\alpha = 1$ and $\beta = 1/\lambda$.

Figure on left illustrates the gamma pdf *f*(*x*; α, β) for several (α, β) pairs, and right the standard gamma pdf.



Gamma density curves



standard gamma density curves

The Gamma Distribution

The mean and variance of a random variable *X* having the gamma distribution *f*(*x*; α, β) are

$$E(X) = \mu = \alpha\beta \qquad V(X) = \sigma^2 = \alpha\beta^2$$

When *X* is a standard gamma rv, the cdf of *X*,

$$F(x;\alpha) = \int_0^x \frac{y^{\alpha-1}e^{-y}}{\Gamma(\alpha)}\,dy \qquad x > 0$$

is often called the **incomplete gamma function**

Routinely available from R (`pgamma`), Matlab, Mathematica…

Example

Suppose the survival time X (weeks) of a random mouse has a gamma distribution with $\alpha = 8$ and $\beta = 15$.

Then:

$E(X) = (8)(15) = 120$ weeks

$V(X) = (8)(15)^2 = 1800$

$\sigma_x = \sqrt{1800}$ 42.43 weeks.

Example 24 cont'd

The probability that a mouse survives between 60 and 120 weeks is

$$P(60 \le X \le 120) = P(X \le 120) - P(X \le 60)$$

$$= F(120/15; 8) - F(60/15; 8)$$

$$= F(8;8) - F(4;8)$$

$$= .547 - .051$$

$$= .496$$

Example 24 cont'd

The probability that a mouse survives at least 30 weeks is

$$P(X \ge 30) = 1 - P(X < 30)$$

$$= 1 - P(X \le 30)$$

$$= 1 - F(30/15; 8)$$

$$= .999$$

The Chi-Squared Distribution

The Chi-Squared Distribution

Definition

Let *v* be a positive integer. Then a random variable *X* is said to have a **chi-squared distribution** with parameter *v* if the pdf of *X* is the **gamma density with $\alpha = v/2$ and $\beta = 2$.** The pdf of a chi-squared rv is thus

$$f(x;\nu) = \begin{cases} \dfrac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{(\nu/2)-1} e^{-x/2} & x \geq 0 \\ 0 & x < 0 \end{cases} \qquad \textbf{(4.10)}$$

The parameter is called the **number of degrees of freedom** (df) of *X*. The symbol x2 is often used in place of "chi-squared."

The Weibull Distribution

The Weibull Distribution

The family of Weibull distributions was introduced by the Swedish physicist Waloddi Weibull in 1939; his 1951 article "A Statistical Distribution Function of Wide Applicability" (*J. of Applied Mechanics*, vol. 18: 293–297) discusses a number of applications.

Definition

A random variable *X* is said to have a **Weibull distribution** with parameters α and β ($\alpha > 0$, $\beta > 0$) if the pdf of *X* is

$$f(x;\alpha,\beta) = \begin{cases} \dfrac{\alpha}{\beta^{\alpha}} x^{\alpha-1} e^{-(x/\beta)^{\alpha}} & x \geq 0 \\ 0 & x < 0 \end{cases} \qquad (4.11)$$

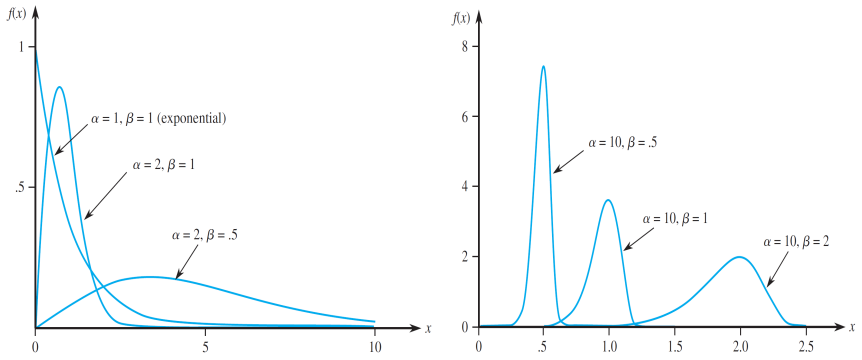
The Weibull Distribution

In some situations, there are theoretical justifications for the appropriateness of the Weibull distribution, but in many applications $f(x; \alpha, \beta)$ simply provides a good fit to observed data for particular values of α and β.

When $\alpha = 1$, the pdf reduces to the exponential distribution (with $\lambda = 1/\beta$), so the exponential distribution is a special case of both the gamma and Weibull distributions.

The Weibull Distribution

Both α and β can be varied to obtain a number of different-looking density curves, as illustrated in



Weibull density curves

The Weibull Distribution

β is called a scale parameter, since different values stretch or compress the graph in the *x* direction, and α is referred to as a shape parameter.

Integrating to obtain $E(X)$ and $E(X^2)$ yields

$$\mu = \beta\Gamma\left(1 + \frac{1}{\alpha}\right) \qquad \sigma^2 = \beta^2\left\{\Gamma\left(1 + \frac{2}{\alpha}\right) - \left[\Gamma\left(1 + \frac{1}{\alpha}\right)\right]^2\right\}$$

The computation of μ and σ^2 thus necessitates using the gamma function.

The Weibull Distribution

The integration $\int_0^x f(y; \alpha, \beta)\, dy$ is easily carried out to obtain the cdf of *X*.

The cdf of a Weibull rv having parameters α and β is

$$F(x; \alpha, \beta) = \begin{cases} 0 & x < 0 \\ 1 - e^{-(x/\beta)^\alpha} & x \geq 0 \end{cases} \qquad (4.12)$$

Example

In recent years the Weibull distribution has been used to model engine emissions of various pollutants.

Let *X* denote the amount of NO_x emission (g/gal) from a randomly selected four-stroke engine, and suppose that *X* has a Weibull distribution with α = 2 and β = 10

See the article "Quantification of Variability and Uncertainty in Lawn and Garden Equipment NO_x and Total Hydrocarbon Emission Factors," *J. of the Air and Waste Management Assoc.*, 2002: 435–448).

The Lognormal Distribution

The Lognormal Distribution

Definition

A nonnegative rv X is said to have a **lognormal distribution** if the rv $Y = \ln(X)$ has a normal distribution.

The resulting pdf of a lognormal rv when $\ln(X)$ is normally distributed with parameters μ and σ is

$$f(x;\mu,\sigma) = \begin{cases} \dfrac{1}{\sqrt{2\pi}\sigma x} e^{-[\ln(x)-\mu]^2/(2\sigma^2)} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

The Lognormal Distribution

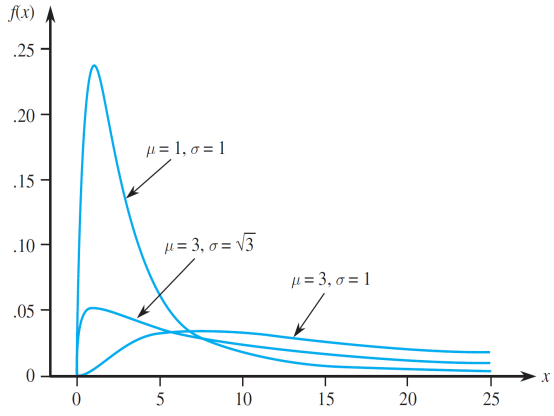
Be careful here; the parameters μ and σ are not the mean and standard deviation of X but of $\ln(X)$.

It is common to refer to μ and σ as the location and the scale parameters, respectively. The mean and variance of X can be shown to be

$$E(X) = e^{\mu+\sigma^2/2} \qquad V(X) = e^{2\mu+\sigma^2} \cdot (e^{\sigma^2} - 1)$$

The Lognormal Distribution

Figure below illustrates graphs of the lognormal pdf; although a normal curve is symmetric, a lognormal curve has a positive skew.



The Lognormal Distribution

Because ln(*X*) has a normal distribution, the cdf of *X* can be expressed in terms of the cdf $\Phi(z)$ of a standard normal rv *Z*.

$$F(x; \mu, \sigma) = P(X \le x) = P[\ln(X) \le \ln(x)]$$

$$= P\left(Z \le \frac{\ln(x) - \mu}{\sigma}\right)$$

$$= \Phi\left(\frac{\ln(x) - \mu}{\sigma}\right) \quad x \ge 0$$

The Beta Distribution

The Beta Distribution

All families of continuous distributions discussed so far except for the uniform distribution had positive density over an infinite interval (though typically the density function decreases rapidly to zero beyond a few standard deviations from the mean).

The beta distribution provides positive density only for *X* in an interval of finite length [A,B].

The standard beta distribution is commonly used to model variation in the proportion or percentage of a quantity occurring in different samples, such as the proportion of a 24-hour day that an individual is asleep or the proportion of a certain element in a chemical compound.

The Beta Distribution

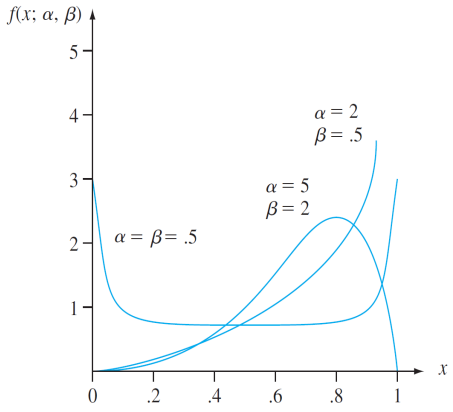
Definition
A random variable *X* is said to have a **beta distribution** with parameters α, β (both positive), *A*, and *B* if the pdf of *X* is

$$f(x; \alpha, \beta, A, B) = \begin{cases} \frac{1}{B-A} \cdot \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\cdot\Gamma(\beta)} \left(\frac{x-A}{B-A}\right)^{\alpha-1} \left(\frac{B-x}{B-A}\right)^{\beta-1} & A \le x \le B \\ 0 & \text{otherwise} \end{cases}$$

The case *A* = 0, *B* = 1 gives the **standard beta distribution.**

The Beta Distribution

Figure below illustrates several standard beta pdf's.



Standard beta density curves

The Beta Distribution

Graphs of the general pdf are similar, except they are shifted and then stretched or compressed to fit over [*A*, *B*].

Unless α and β are integers, integration of the pdf to calculate probabilities is difficult. Either a table of the incomplete beta function or appropriate software should be used.

The mean and variance of *X* are

$$\mu = A + (B - A) \cdot \frac{\alpha}{\alpha + \beta} \qquad \sigma^2 = \frac{(B - A)^2 \alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

Example

cont'd

Suppose that in constructing a single-family house, the time *X* (in days) necessary for laying the foundation has a beta distribution with $A = 2$, $B = 5$, $\alpha = 2$, and $\beta = 3$.

Then $\alpha/(\alpha + \beta) = .4$, so $E(X) = 2 + (3)(.4) = 3.2$.

The probability that it takes at most 3 days is:

$$P(X \le 3) = \int_2^3 \frac{1}{3} \cdot \frac{4!}{1!2!}\left(\frac{x - 2}{3}\right)\left(\frac{5 - x}{3}\right)^2 dx$$

$$= \frac{4}{27}\int_2^3 (x - 2)(5 - x)^2 dx = \frac{4}{27} \cdot \frac{11}{4} = \frac{11}{27} = .407$$

Examples...

Example 1

Suppose the pdf of the magnitude X of a dynamic load on a bridge (in newtons) is

$$f(x) = \begin{cases} \frac{1}{8} + \frac{3}{8}x & 0 \le x \le 2 \\ 0 & \text{otherwise} \end{cases}$$

For any number x between 0 and 2,

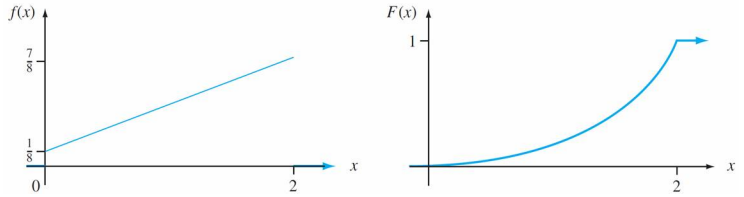
$$F(x) = \int_{-\infty}^{x} f(y)\,dy = \int_{0}^{x} \left(\frac{1}{8} + \frac{3}{8}y\right) dy = \frac{x}{8} + \frac{3}{16}x^2$$

Example 1 cont'd

Thus

$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{x}{8} + \frac{3}{16}x^2 & 0 \le x \le 2 \\ 1 & 2 < x \end{cases}$$

The graphs of $f(x)$ and $F(x)$ are shown in Figure 4.9.



Example 1 cont'd

The probability that the load is between 1 and 1.5 is

$P(1 \le X \le 1.5) = F(1.5) - F(1)$

$$= \left[\frac{1}{8}(1.5) + \frac{3}{16}(1.5)^2\right] - \left[\frac{1}{8}(1) + \frac{3}{16}(1)^2\right]$$

$$= \frac{19}{64}$$

$$= .297$$

The probability that the load exceeds 1 is

$P(X > 1) = 1 - P(X \le 1)$

$= 1 - F(1)$

Example 1 cont'd

$$= 1 - \left[\frac{1}{8}(1) + \frac{3}{16}(1)^2\right]$$

$$= \frac{11}{16}$$

$$= .688$$

Once the cdf has been obtained, any probability involving X can easily be calculated without any further integration.

Example 2

Two species are competing in a region for control of a limited amount of a certain resource.

Let X = the proportion of the resource controlled by species 1 and suppose X has pdf

$$f(x) = \begin{cases} 1 & 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

which is a uniform distribution on [0, 1]. (In her book *Ecological Diversity,* E. C. Pielou calls this the "broken- tick" model for resource allocation, since it is analogous to breaking a stick at a randomly chosen point.)

Example 2 cont'd

Then the species that controls the majority of this resource controls the amount

$$h(X) = \max(X, 1 - X) = \begin{cases} 1 - X & \text{if } 0 \le X < \frac{1}{2} \\ X & \text{if } \frac{1}{2} \le X \le 1 \end{cases}$$

The expected amount controlled by the species having majority control is then

$$E[h(X)] = \int_{-\infty}^{\infty} \max(x, 1 - x)\ f(x)dx$$

Example 2 cont'd

$$= \int_0^1 \max(x, 1 - x)\ 1\, dx$$

$$= \int_0^{1/2} (1 - x)\ 1\, dx + x \int_{1/2}^1 1\ \ dx$$

$$= \frac{3}{4}$$

Example 3

The time that it takes a driver to react to the brake lights on a decelerating vehicle is critical in helping to avoid rear-end collisions.

The article "Fast-Rise Brake Lamp as a Collision-Prevention Device" (*Ergonomics*, 1993: 391–395) suggests that reaction time for an in-traffic response to a brake signal from standard brake lights can be modeled with a normal distribution having mean value 1.25 sec and standard deviation of .46 sec.

Example 3 cont'd

What is the probability that reaction time is between 1.00 sec and 1.75 sec? If we let X denote reaction time, then standardizing gives

$$1.00 \le X \le 1.75$$

if and only if

$$\frac{1.00 - 1.25}{.46} \le \frac{X - 1.25}{.46} \le \frac{1.75 - 1.25}{.46}$$

Thus

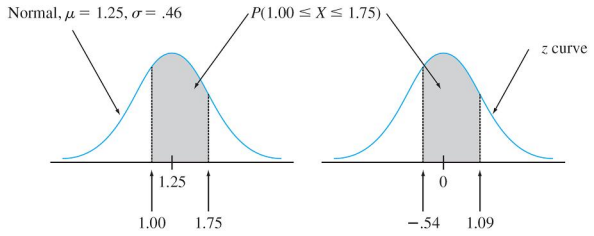
$$P(1.00 \le X \le 1.75) = P\left(\frac{1.00 - 1.25}{.46} \le Z \le \frac{1.75 - 1.25}{.46}\right)$$

Example 3 cont'd

$= P(-.54 \le Z \le 1.09) = (1.\Phi?) - (-.5\ \Phi$

$= .8621 - .2946 = .5675$

This is illustrated in Figure 4.22



Example 3 cont'd

Similarly, if we view 2 sec as a critically long reaction time, the probability that actual reaction time will exceed this value is

$$P(X > 2) = P\left(Z > \frac{2 - 1.25}{.46}\right) = P(Z > 1.63) = 1 - \Phi(1.63) = .0516$$

Example 4

According to the article "Predictive Model for Pitting Corrosion in Buried Oil and Gas Pipelines" (*Corrosion*, 2009: 332–342), the lognormal distribution has been reported as the best option for describing the distribution of maximum pit depth data from cast iron pipes in soil.

The authors suggest that a lognormal distribution with $\mu = .353$ and $\sigma = .754$ is appropriate for maximum pit depth (mm) of buried pipelines.

For this distribution, the mean value and variance of pit depth are

$$E(X) = e^{.353+(.754)^2/2} = e^{.6373} = 1.891$$

Example 4 cont'd

$$V(X) = e^{2(.353)+(.754)^2} \cdot (e^{(.754)^2} - 1) = (3.57697)(.765645) = 2.7387$$

The probability that maximum pit depth is between 1 and 2 mm is

$P(1 \le X \le 2) = P(\ln(1) \le \ln(X) \le \ln(2))$

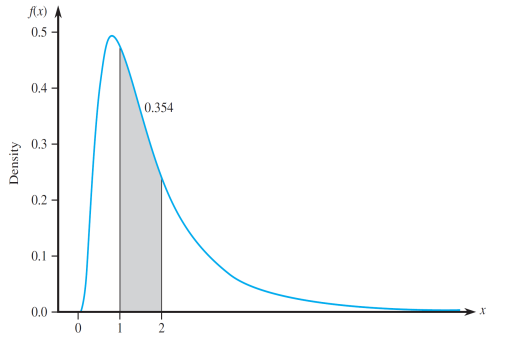
$= P(0 \le \ln(X) \le .693)$

$$= P\left(\frac{0 - .353}{.754} \le Z \le \frac{.693 - .353}{.754}\right)$$

$= \phi(.47) - \phi(-.45) = .354$

Example 4 cont'd

This probability is illustrated below



Lognormal density curve with location = .353 and scale = .754

Example 4 cont'd

What value c is such that only 1% of all specimens have a maximum pit depth exceeding c? The desired value satisfies

$$.99 = P(X \le c) = P\left(Z \le \frac{\ln(c) - .353}{.754}\right)$$

The z critical value 2.33 captures an upper-tail area of .01 ($z_{.01} = 2.33$), and thus a cumulative area of .99.

This implies that

$$\frac{\ln(c) - .353}{.754} = 2.33$$