Important Probability Distributions

OPRE 6301

Important Distributions...

Certain probability distributions occur with such regularity in real-life applications that they have been given their own names. Here, we survey and study basic properties of some of them.

We will discuss the following distributions:

- Binomial
- Poisson
- Uniform
- Normal
- Exponential

The first two are discrete and the last three continuous.

Binomial Distribution...

Consider the following scenarios:

— The number of heads/tails in a sequence of coin flips

— Vote counts for two different candidates in an election

— The number of male/female employees in a company

— The number of accounts that are in compliance or not in compliance with an accounting procedure

— The number of successful sales calls

— The number of defective products in a production run

— The number of days in a month your company's computer network experiences a problem

All of these are situations where the binomial distribution may be applicable.

Canonical Framework...

There is a set of assumptions which, if valid, would lead to a binomial distribution. These are:

- A set of n experiments or **trials** are conducted.
- Each trial could result in either a **success** or a **failure**.
- The probability p of success is the *same* for all trials.
- The outcomes of different trials are *independent*.
- We are interested in the total number of successes in these n trials.

Under the above assumptions, let X be the total number of successes. Then, X is called a **binomial random variable**, and the probability distribution of X is called the **binomial distribution**.

Binomial Probability-Mass Function. . .

Let X be a binomial random variable. Then, its probability-mass function is:

$$P(X = x) = \frac{n!}{x!(n-x)!}\, p^x (1-p)^{n-x} \qquad (1)$$

for $x = 0, 1, 2, \ldots, n$.

The values of n and p are called the *parameters* of the distribution.

To understand (1), note that:

- The probability for observing *any* sequence of n independent trials that contains x successes and $n - x$ failures is $p^n(1-p)^{n-x}$.
- The total number of such sequences is equal to

$$\binom{n}{x} \equiv \frac{n!}{x!(n-x)!}$$

  (i.e., the total number of possible combinations when we randomly select x objects out of n objects).

Example: Multiple-Choice Exam

Consider an exam that contains 10 multiple-choice questions with 4 possible choices for each question, only one of which is correct.

Suppose a student is to select the answer for every question randomly. Let X be the number of questions the student answers correctly. Then, X has a binomial distribution with parameters $n = 10$ and $p = 0.25$. (Convince yourself that all assumptions for a binomial distribution are reasonable in this setting.)

What is the probability for the student to get no answer correct? Answer:

$$\begin{aligned} P(X = 0) &= \frac{10!}{0!(10-0)!}(0.25)^0(1-0.25)^{10-0} \\ &= (0.75)^{10} \\ &= 0.0563 \end{aligned}$$

What is the probability for the student to get two answers correct? Answer:

$$\begin{aligned} P(X=2) &= \frac{10!}{2!8!}(0.25)^2(1-0.25)^8 \\ &= 45 \cdot (0.25)^2 \cdot (0.75)^8 \\ &= 0.2816 \end{aligned}$$

What is the probability for the student to fail the test (i.e., to have less than 6 correct answers)? Answer:

$$\begin{aligned} P(X \leq 5) &= \sum_{i=0}^{5} P(X=i) \\ &= 0.0563 + 0.1877 + 0.2816 + 0.2503 \\ &\qquad +0.1460 + 0.0584 \\ &= 0.9803 \end{aligned}$$

Binomial probabilities can be computed using the Excel function BINOMDIST(). Two other examples are given in a separate Excel file.

Binomial Mean and Variance...

It can be shown that

$$\mu = E(X) = np$$

and

$$\sigma^2 = V(X) = np(1-p)\,.$$

For the previous example, we have

- $E(X) = 10 \cdot 0.25 = 2.5$.
- $V(X) = 10 \cdot (0.25) \cdot (1 - 0.25) = 1.875$.

Poisson Distribution...

The Poisson distribution is another family of distributions that arises in a great number of business situations. It usually is applicable in situations where random "events" occur at a certain *rate* over a period of *time*.

Consider the following scenarios:

— The hourly number of customers arriving at a bank

— The daily number of accidents on a particular stretch of highway

— The hourly number of accesses to a particular web server

— The daily number of emergency calls in Dallas

— The number of typos in a book

— The monthly number of employees who had an absence in a large company

— Monthly demands for a particular product

All of these are situations where the Poisson distribution may be applicable.

Canonical Framework...

Like the Binomial distribution, the Poisson distribution arises when a set of canonical assumptions are reasonably valid. These are:

- The number of events that occur in any time interval is independent of the number of events in any other disjoint interval. Here, "time interval" is the standard example of an "exposure variable" and other interpretations are possible. Example: Error rate per *page* in a book.
- The distribution of number of events in an interval is the same for all intervals of the same size.
- For a "small" time interval, the probability of observing an event is proportional to the length of the interval. The proportionality constant corresponds to the "rate" at which events occur.
- The probability of observing two or more events in an interval approaches zero as the interval becomes smaller.

Under the above assumptions, let λ be the rate at which events occur, t be the length of a time interval, and X be the total number of events in that time interval. Then, X is called a **Poisson random variable** and the probability distribution of X is called the **Poisson distribution**.

Let $\mu \equiv \lambda t$; then, μ can be interpreted as the average, or mean, number of events in an interval of length t.

Poisson Probability-Mass Function...

Let X be a Poisson random variable. Then, its probability-mass function is:

$$P(X = x) = e^{-\mu} \frac{\mu^x}{x!} \tag{2}$$

for $x = 0, 1, 2, \ldots$.

The value of μ is the *parameter* of the distribution. For a given time interval of interest, in an application, μ can be specified as λ times the length of that interval.

Example: Typos

The number of typographical errors in a "big" textbook is Poisson distributed with a mean of 1.5 per 100 pages.

Suppose 100 pages of the book are randomly selected. What is the probability that there are no typos? Answer:

$$P(X = 0) = e^{-\mu} \frac{\mu^x}{x!} = e^{-1.5} \frac{1.5^0}{0!} = 0.2231$$

Suppose 400 pages of the book are randomly selected. What are the probabilities for having no typos and for having five or fewer typos? Answers:

$$P(X = 0) = e^{-1.5 \cdot 4} \frac{(1.5 \cdot 4)^0}{0!}$$

$$= 0.002479$$

and

$$\begin{aligned} P(X \leq 5) &= \sum_{i=0}^{5} P(X = i) \\ &= 0.0025 + 0.0149 + 0.0446 + 0.0892 \\ &\quad +0.1339 + 0.1606 \\ &= 0.4457 \end{aligned}$$

Poisson probabilities can be computed using the Excel function POISSON(). Further numerical examples of the Poisson distribution are given in a separate Excel file.

Mean and Variance

It can be shown that

$$E(X) = \mu$$

and

$$V(X) = \mu\,.$$

Interpretation of (2)

The form of (2) seems mysterious. The best way to understand it is via the binomial distribution.

Consider a time interval and divide it into n equally-sized subintervals. Suppose n is very large so that either one or zero event can occur in a subinterval. Suppose further that the probability for an event to occur in a subinterval is μ/n, independent of what occurs in other subintervals.

Under these assumptions, the total number of events, X, in that interval has a binomial distribution with parameters n and μ/n. That is,

$$P(X = x) = \frac{n!}{x!(n-x)!} \left(\frac{\mu}{n}\right)^x \left(1 - \frac{\mu}{n}\right)^{n-x} \qquad (3)$$

for $x = 0, 1, 2, \ldots, n$.

Note that $E(X) = n \cdot (\mu/n) = \mu$, suggesting that (3) and (1) are "consistent." Indeed, it can be shown that as n approaches ∞, (3) becomes (2). This useful fact is called **Poisson approximation** to the binomial distribution.

We will see several other examples of such limiting approximations in future chapters. They provide simple and accurate approximations to otherwise unmanageable expressions.

General Continuous Distributions...

Recall that a continuous random variable or distribution is defined via a probability *density* function. Let $f(x)$ (nonnegative) be the density function of variable X. Then, $f(x)$ is the rate at which probability accumulates in the neighborhood of x. In other words,

$$f(x)\,h \approx P(x < X \leq x + h)$$

when h (a positive number) is sufficiently small. It follows from this rate interpretation that for any interval $(x_1,\, x_2]$, we have

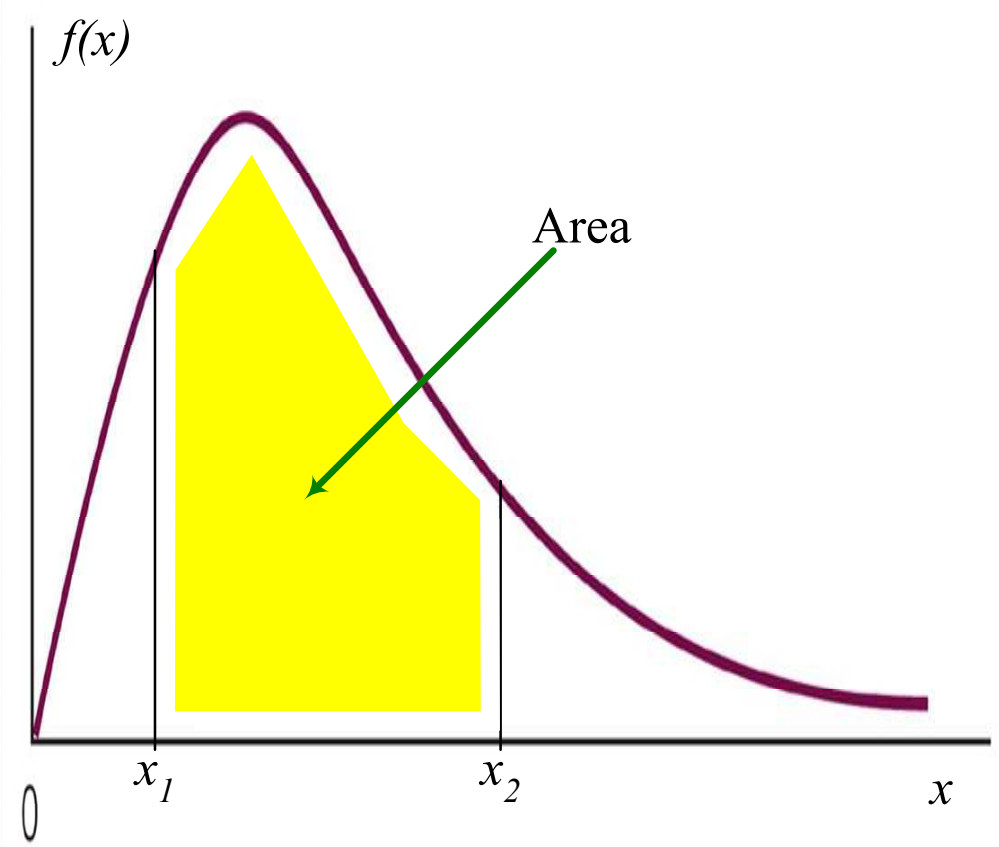
$$P(x_1 < X \leq x_2) = \int_{x_1}^{x_2} f(x)\,dx\,; \tag{4}$$

moreover, we must have

$$\int_{-\infty}^{\infty} f(x)\,dx = 1\,.$$

Note that the probability for a continuous random variable to assume any particular value is 0; this can be seen by setting $x_1 = x_2$ in (4).

Recall further that the integral of a function over an interval is the area under that function over the given interval. We can therefore visualize $P(x_1 < X \leq x_2)$ as the area of the yellow region below:



For $-\infty < x < \infty$, the function

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

(i.e., let $x_1 = -\infty$ and $x_2 = x$ in (4)) is called the **cumulative** distribution function of X. $F(x)$ can also be used to describe a random variable, since $f(x)$ is the derivative of $F(x)$.

Various probabilities of interest regarding a variable X can all be computed via either $f(x)$ or $F(x)$.

We next discuss three important continuous distributions: uniform, normal, and exponential.

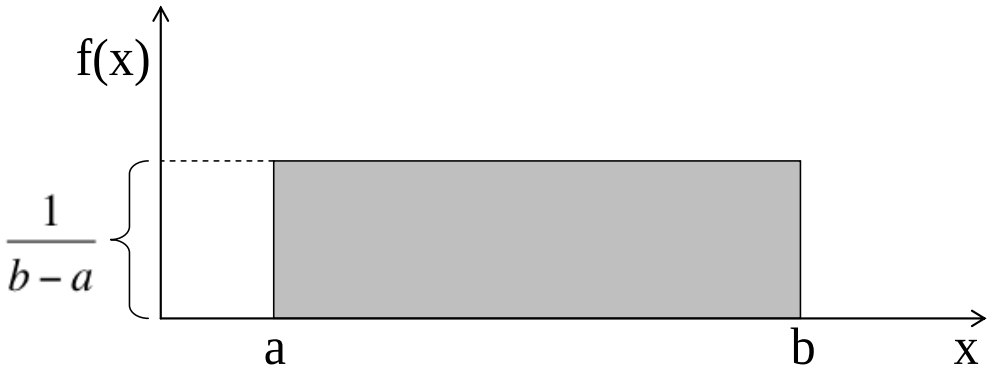
Uniform Distribution...

The uniform distribution is the simplest example of a continuous probability distribution. A random variable X is said to be uniformly distributed if its density function is given by:

$$f(x) = \frac{1}{b-a} \tag{5}$$

for $-\infty < a \leq x \leq b < \infty$.

Visually, we have



where the shaded region has area $(b-a)[1/(b-a)] = 1$ (width times height).

The values a and b are the parameters of the uniform distribution. It can be shown that

$$E(X) = \frac{a+b}{2} \qquad \text{and} \qquad V(X) = \frac{(b-a)^2}{12}\,.$$

The *standard* uniform density has parameters $a = 0$ and $b = 1$; and hence $f(x) = 1$ for $0 \leq x \leq 1$ and 0 otherwise. The Excel function RAND() "pretends" to generate independent samples from this density function.

Example: Gasoline Sales

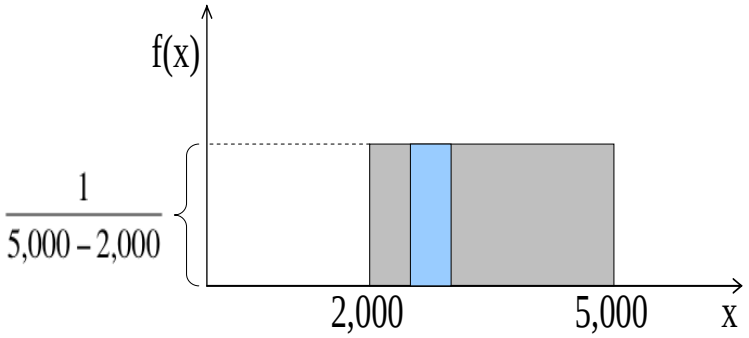
Suppose the amount of gasoline sold daily at a service station is uniformly distributed with a minimum of 2,000 gallons and a maximum of 5,000 gallons.

What is the probability that daily sales will fall between 2,500 gallons and 3,000 gallons? Answer:

$$P(2500 < X \leq 3000) = \frac{1}{5000 - 2000}(3000 - 2500)$$

$$= 0.1667\,.$$

Visually, we have



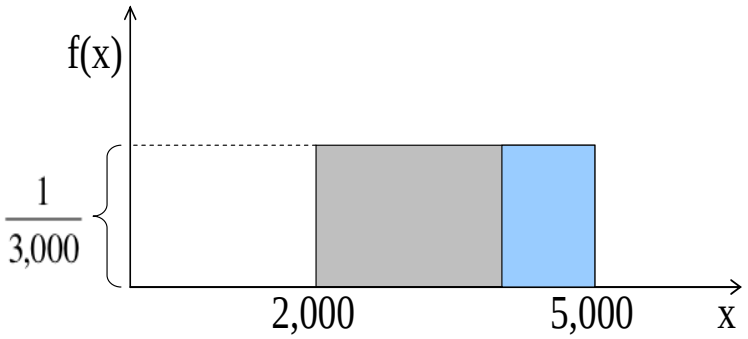
and the answer corresponds to the area in blue.

What is the probability that the service station will sell *at least* 4,000 gallons? Answer:

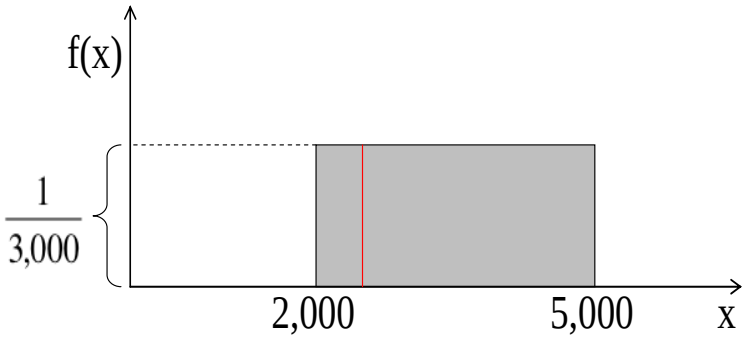
$$P(X > 4000) = \frac{1}{5000 - 2000}(5000 - 4000)$$

$$= 0.3333\,.$$

Visually, we have



What is the probability that the service station will sell *exactly* 2,500 gallons? Answer: $P(X = 2500) = 0$, since the area of a "vertical line" at 2,500 is 0.



Normal Distribution...

The normal distribution is the most important distribution in statistics, since it arises naturally in numerous applications. The **key reason** is that **large** sums of (small) random variables often turn out to be normally distributed; a more-complete discussion of this will be given in Chapter 9.

A random variable X is said to have the normal distribution with parameters μ and σ if its density function is given by:

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{ -\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2 \right\} \qquad (6)$$

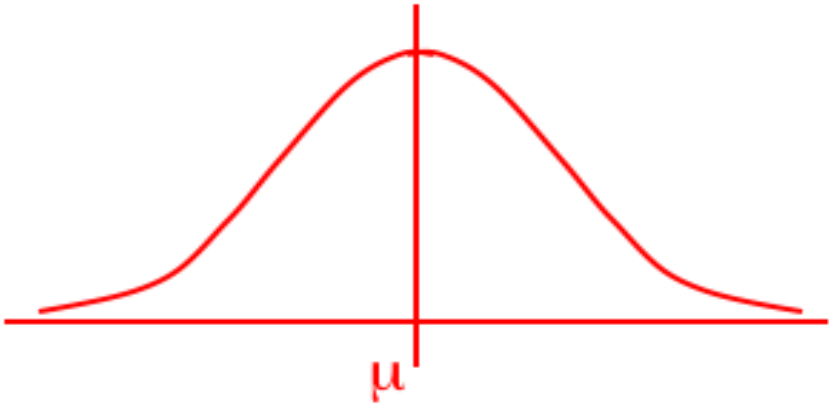
for $-\infty < x < \infty$.

It can be shown that

$$E(X) = \mu \qquad \text{and} \qquad V(X) = \sigma^2 \,.$$

Thus, the normal distribution is characterized by a mean μ and a standard deviation σ.

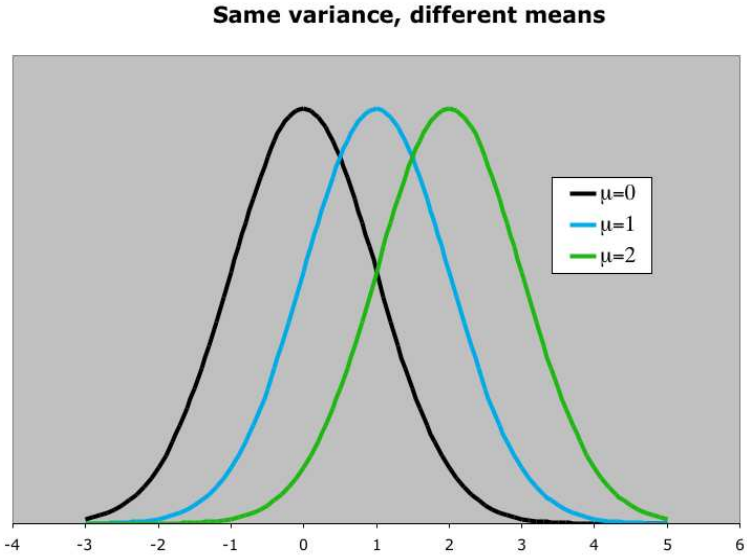
A typical normal density curve looks like this:



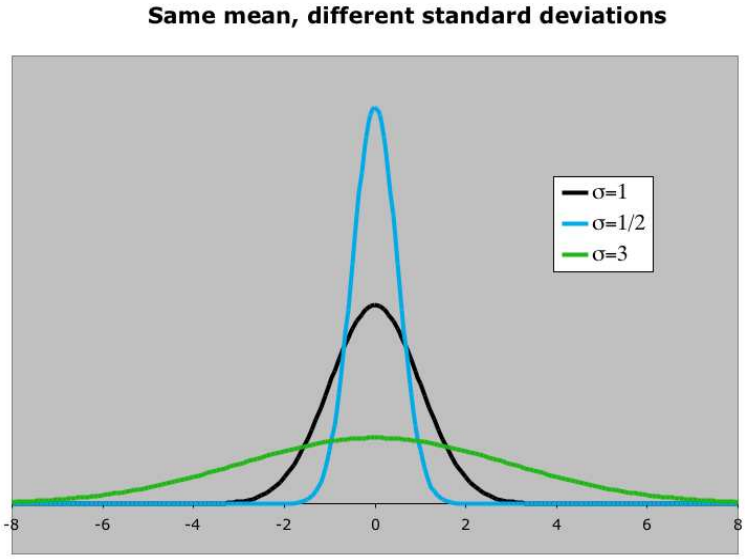
Thus, the curve is bell shaped and is symmetric around the mean μ. The standard deviation σ controls the "flatness" of the curve.

Details ...

Increasing the mean *shifts* the density curve to the right . . .



Increasing the standard deivation *flattens* the density curve . . .



Calculating Normal Probabilities...

A normal distribution whose mean is 0 and standard deviation is 1 is called the **standard** normal distribution. In this case, the density function assumes the simpler form:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \tag{7}$$

for $-\infty < x < \infty$.

Table 3 in Appendix B of the text can be used to calculate probabilities associated with the standard normal distribution. The Excel function NORMSDIST() (where "S" is for "standard") can also be used.

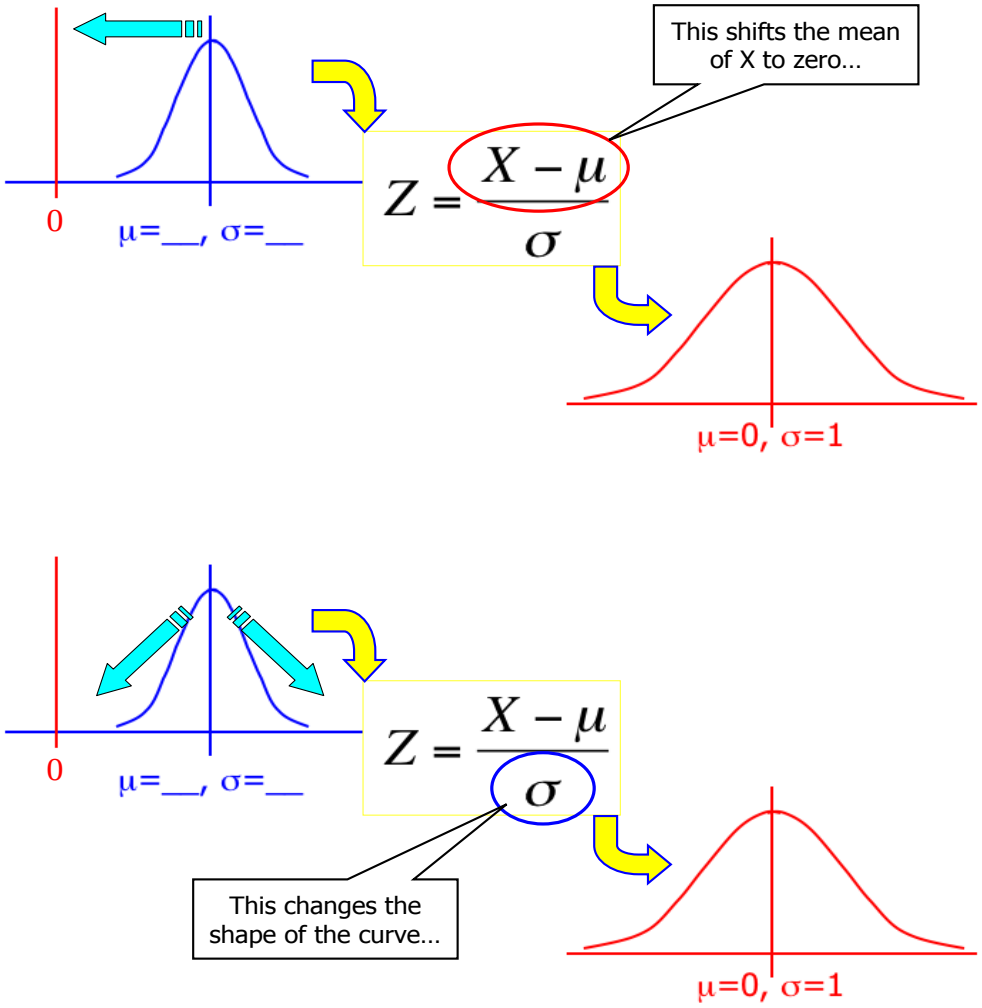
Denote by Z a random variable that follows the standard normal distribution. Then, Table 3 gives the probability $P(0 < Z \leq z)$ for any nonnegative value z; whereas NORMSDIST() returns $P(Z \leq z)$ for any z from $-\infty$ to ∞, i.e., values of the *cumulative* distribution function.

For general parameter values, the Excel function NORMDIST() (without "S" in the middle) can be used directly. However, ...

A standard practice is to convert a normal random variable X with arbitrary parameters μ and σ into a **standardized** normal random variable Z with parameters 0 and 1 via the transformation:

$$Z = \frac{X - \mu}{\sigma}; \tag{8}$$

this is illustrated in:



Example 1: Build Time of Computers

Suppose the time required to build a computer is normally distributed with a mean of 50 minutes and a standard deviation of 10 minutes.

What is the probability for the assembly time of a computer to be between 45 and 60 minutes? Answer:

We wish to compute $P(45 < X \leq 60)$. To do this, we first rewrite the event of interest into a form that is in terms of a standardized variable $Z = (X - 50)/10$, as follows.

$$P\left(\frac{45-50}{10} < \frac{X-50}{10} \leq \frac{60-50}{10}\right)$$

$$= P(-0.5 < Z \leq 1)\,.$$

Next, observe that

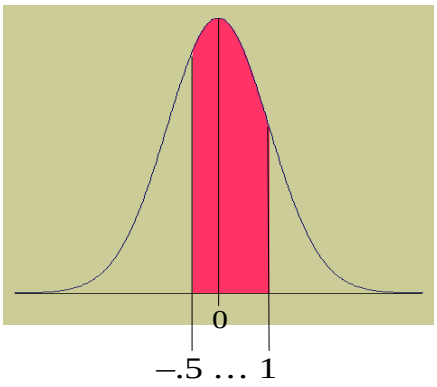
$$P(-0.5 < Z \leq 1) = P(Z \leq 1) - P(Z \leq -0.5)\,.$$

Using the Excel function NORMSDIST(), we find that $P(Z \leq 1) = 0.8413$ and $P(Z \leq -0.5) = 0.3085$. Hence, the answer is $0.8413 - 0.3085 = 0.5328$.

Table 3 can also be used for this calculation:

$$\begin{aligned} & P(-0.5 < Z \leq 1) \\ & \quad = P(-0.5 < Z \leq 0) + P(0 < Z \leq 1) \\ & \quad = P(0 < Z \leq 0.5) + P(0 < Z \leq 1) \\ & \quad = 0.1915 + 0.3414 \\ & \quad = \boxed{0.5328}\,, \end{aligned}$$

where the first equality follows from



the second equality is due to the fact that the normal density curve is *symmetric*, and the third equality is from Table 3.

Is it reasonable to assume that the build time is normally distributed? Reasoning: The build time can be thought of as the sum of times needed to build many individual components.

Example 2: Stock Returns

Suppose the return of an investment in a stock over a given time period is normally distributed with a mean of 10% and a standard deviation of 5%. Reasoning: Price movement of a stock over the given period can be thought of as the sum of a "long" sequence of small movements.

What is the probability of losing money over the given period? Answer: We wish to determine $P(X \leq 0)$. Following the steps in the previous example, we obtain

$$\begin{aligned} & P(X \leq 0) \\ & \quad = P\left(\frac{X-10}{5} \leq \frac{0-10}{5}\right) \\ & \quad = P(Z \leq -2) \\ & \quad = 0.02275 \,. \end{aligned}$$

What is the effect of doubling the standard deviation to 10? Answer: A similar calculation yields

$$\begin{aligned} P(X \leq 0) &= P\left(\frac{X-10}{10} \leq \frac{0-10}{10}\right) \\ &= P(Z \leq -1) \\ &= 0.1587\,, \end{aligned}$$

which is almost 7 times larger than the previous answer. Thus, increasing the standard deviation increases the probability of losing money. This reiterates the fact that the standard deviation is a measure of risk.

Example 3: Midterm Scores

Why did the frequency distribution of the Midterm scores resemble a normal density curve? Reasoning: The total score of an exam is the sum of scores for many individual problems/parts.

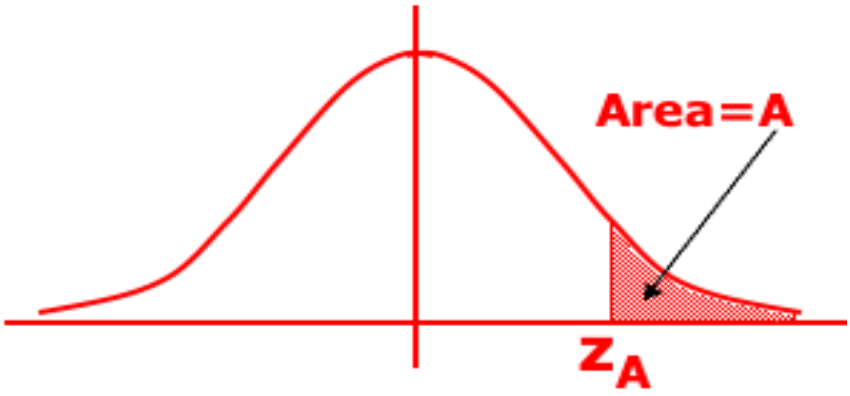
Finding "z" for Given Probability...

Most of the calculations above are of the form: Find the probability $P(Z \leq z)$ for a given value of z. Often times, we are also interested in an inverse problem: Find the value of z_A such that the probability for Z to be greater than z_A equals a specified value A.

Formally, our question is: For what value of z_A do we have

$$P(Z > z_A) = A\,? \tag{9}$$

This can be visualized as:



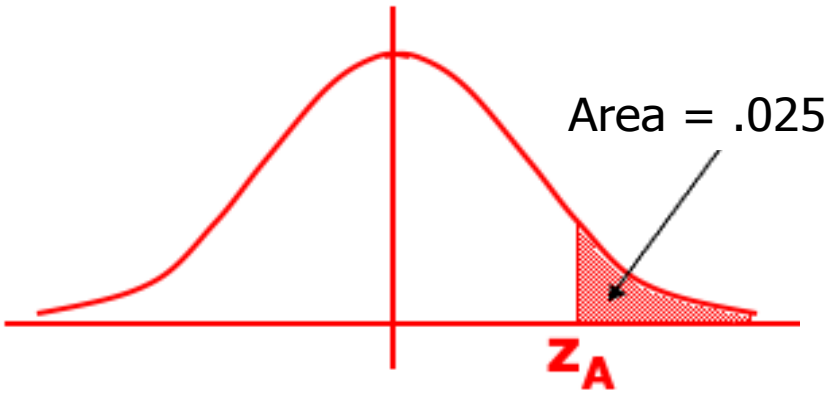
Questions like these will be relevant in statistical inference.

Examples:

Find z_A for $A = 0.025$ (or 2.5%). That is, what is $z_{0.025}$?

Answer: Observe that

$$P(Z > z_{0.025}) = 1 - P(Z \leq z_{0.025})\,.$$



Observer further that

$$\begin{aligned} P(Z \leq z_{0.025}) &= 1 - P(Z > z_{0.025}) \\ &= 1 - 0.025 \\ &= 0.975\,, \end{aligned}$$

where the second equality follows from the definition of $z_{0.025}$.

Hence, our problem is equivalent to that of finding $z_{0.025}$ such that $P(Z \leq z_{0.025}) = 0.975$. That is, we are interested in the inverse of a cumulative distribution function; this is similar to finding *percentiles* using an ogive. The Excel function NORMSDIST() (which is a cumulative distribution function) has an **inverse**: NORMSINV(). Using this inverse function with argument 0.975, we find that $z_{0.025} = 1.96$.

For $A = 0.05$, we have $z_{0.05} = 1.645$.

For $A = 0.01$, we have $z_{0.01} = 2.33$.

Exponential Distribution...

Another useful continuous distribution is the **exponential distribution**, which has the following probability density function:

$$f(x) = \lambda e^{-\lambda x} \tag{10}$$

for $x \geq 0$.

This family of distributions is characterized by a single parameter λ, which is called the *rate*. Intuitively, λ can be thought of as the instantaneous "failure rate" of a "device" at any time t, given that the device has survived up to t.

The exponential distribution is typically used to model time intervals *between* "random events"...

Examples:

— The length of time between telephone calls

— The length of time between arrivals at a service station

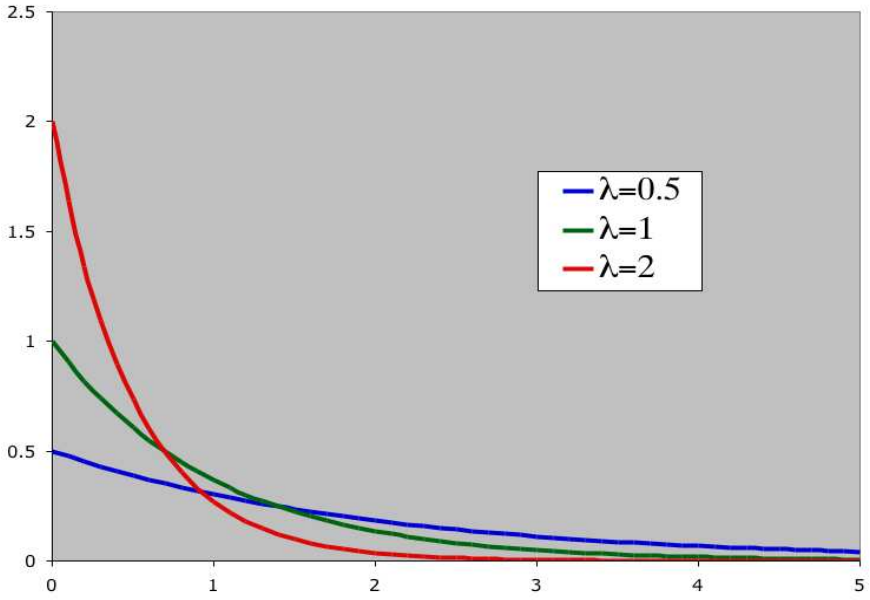
— The life time of electronic components, i.e., an inter-failure time

An important fact is that when times between random "events" follow the exponential distribution with rate λ, then the total number of events in a time period of length t follows the Poisson distribution with parameter λt.

If a random variable X is exponentially distributed with rate λ, then it can be shown that

$$E(X) = \frac{1}{\lambda} \qquad \text{and} \qquad V(X) = \left(\frac{1}{\lambda}\right)^2 .$$

For $\lambda = 0.5$, 1, and 2, the shapes of the expenential density curve are:



Observe that the greater the rate, the faster the curve drops. Or, the lower the rate, the flatter the curve.

Several useful formulas are:

$$P\{X \leq x\} = 1 - e^{-\lambda x}$$
$$P\{X > x\} = e^{-\lambda x}$$
$$P\{x_1 < X \leq x_2\} = e^{-\lambda x_1} - e^{-\lambda x_2}$$

These correspond to the areas under the density curve to the left of x, to the right of x, and between x_1 and x_2, respectively.

Example 1: Lifetime of a Battery

The lifetime X of an alkaline battery is exponentially distributed with $\lambda = 0.05$ per hour.

What are the mean and standard deviation of the battery's lifetime? Answer:

$$E(X) = SD(X) = \frac{1}{0.05} = 20 \text{ hours.}$$

What are the probabilities for the battery to last between 10 and 15 hours and to last more than 20 hours? Answer:

$$P(10 < X \leq 15) = e^{-0.05 \cdot 10} - e^{-0.05 \cdot 15} = 0.1341$$
$$P(X > 20) = e^{-0.05 \cdot 20} = 0.3679$$

(The Excel function EXP() can be used for these calculations.)

Example 2: Arrivals at a Gas Station

The arrival rate of cars at a gas station is $\lambda = 40$ customers per hour. (This is equivalent to saying that the interarrival times are exponentially distributed with rate 40 per hour.)

What is the probability of having no arrivals in a 5-minute interval? Answer:

$$P(X > \frac{5}{60}) = e^{-40 \cdot (5/60)} = 0.03567$$

What are the mean and variance of the number, N, of arrivals in 5 minutes? Answer:

The variable N has a Poisson distribution with parameter $\mu = \lambda t = 40 \cdot (5/60) = 3.333$. Hence,

$$E(N) = 3.333 \qquad \text{and} \qquad V(N) = 3.333\,.$$

What is the probability for having 3 arrivals in a 5-minute interval? Answer:

$$P(N = 3) = e^{-3.333}\,\frac{3.333^3}{3!} = 0.2202\,.$$