THE NORMAL DISTRIBUTION

Documents prepared for use in course B01.1305,
New York University, Stern School of Business

Cover photo: Monarch butterfly caterpillar, Stony Brook, New York, 2006

Revision date 15 DEC 2006

※※※※※※※※ USE OF NORMAL TABLE ※※※※※※※※

The standard normal distribution refers to the case with mean $\mu = 0$ and standard deviation $\sigma = 1$. This is precisely the case covered by the tables of the normal distribution. It is common to use the symbol Z to represent any random variable which follows a normal distribution with $\mu = 0$ and $\sigma = 1$.

> The normal distribution is often described in terms of its variance σ^2. Clearly σ is found as the square root of σ^2.

If X is a normal random variable with general mean μ (not necessarily 0) and standard deviation σ (not necessarily 1), then it can be converted to standard normal by way of

$$Z = \frac{X - \mu}{\sigma} \text{ or equivalently } X = \mu + \sigma Z$$

The act of subtracting the mean and then dividing by a standard deviation is called "standardizing," and it enables you to use the normal table.

In the examples on this document, it is assumed that you are working from a table in which you have values of P[$Z \leq z$] for positive z only.

EXAMPLE 1: Suppose that Z is a standard normal random variable. What is the probability that Z is between -0.4 and +1.2?

SOLUTION: This is a routine table look-up exercise.

$$P[-0.4 \leq Z \leq +1.2] = P[-0.4 \leq Z \leq 0] + P[0 \leq Z \leq 1.2]$$

$$= P[0 \leq Z \leq 0.4] + P[0 \leq Z \leq 1.2]$$

$$= \quad 0.1554 \quad + \quad 0.3849 \quad = \quad 0.5403$$

EXAMPLE 2: Suppose that Z is a standard normal random variable. Find value w so that P[$-w \leq Z \leq +w$] = 0.60.

SOLUTION: This is again a routine use of the normal table, though here you have to use it in "reverse" order.

Now, P[$- w \leq Z \leq +w$] = 0.60 implies by symmetry that P[$0 \leq Z \leq +w$] = 0.30, and thus one should search the body of the table for the value closest to 0.3000.

The table reveals that

$P[\ 0 \le Z \le 0.84\ ] = 0.2995$
and
$P[\ 0 \le Z \le 0.85\ ] = 0.3023$

One could interpolate, but it's easier to just take the closer value. We'll use $P[\ 0 \le Z \le 0.84] = 0.2995 \approx 0.30$, so that we report our solution as $w = 0.84$. That is, we're claiming that $P[\ -0.84 \le Z \le +0.84\ ] \approx 0.60$.

Here's how you would interpolate to get this answer. Only in rare situations would you need to do this.

z-value	$P[0 \le Z \le z]$	Proportional distance from top
0.84	0.2995	0.0000
c (to be found)	0.3000 (desired)	0.2778
0.85	0.3023	1.0000

The proportional distance figure was found as $\frac{0.3000 - 0.2995}{0.3023 - 0.2995}$ = $\frac{3000 - 2995}{3023 - 2995} = \frac{5}{18} \approx 0.2778$. Now a straight ratio-proportion argument gives $\frac{c - 0.84}{0.85 - 0.84} = 0.2778$. This solves as $c = 0.84 + 0.2778 \times (0.01) = 0.84 + 0.002778 = 0.842778$. It seems reasonable to round this to 0.843.

EXAMPLE 3. Suppose that X is a normal random variable with mean $\mu = 200$ and standard deviation $\sigma = 40$. What is the probability that X will take a value greater than 228?

SOLUTION: If you remember to standardize, you can do problems like this almost without thinking.

$$P[X > 228\ ] = P\left[\frac{X - 200}{40} > \frac{228 - 200}{40}\right] = P[\ Z > 0.7\ ] = 0.5 - P[\ 0 \le Z \le 0.7\ ]$$
$$= 0.5 - 0.2580 = 0.2420$$

In this calculation, observe that $\frac{X - 200}{40}$ is renamed as Z.

EXAMPLE 4. Suppose that the latent load charge threshold for a population of investors is approximately normally distributed with mean 3.2 percentage points and standard deviation 0.8 of a percentage point. Find the lower 10% point. That is, find the load rate *A* so that, with probability 90%, an investor will happily tolerate rate *A*.

> Each person has an upper limit for the load charge; if Dave's limit is 3.8%, then he would not object to paying 3.1%. We can observe how any individual will behave at any specified rate, but we can't observe the individual's upper limit. For this reason, we call the limit *latent*.

SOLUTION: Note that we are given μ = 3.2 percentage points and σ = 0.8. Note also that the 10% point has probability 0.10 associated with lower values and a probability 0.90 associated with higher values.

Note also the use of "approximately normally distributed." We are not going to assert absolutely that this phenomenon, latent load charge threshold, follows a normal distribution perfectly.

You should standardize this problem just as you did the previous. The only difference is that some parts of the problem will be algebra expressions rather than numbers. Let *X* denote the limit for a randomly-chosen investor. Then

$$\text{P}[\, X \leq A \,] = \text{P}\left[\frac{X - 3.2}{0.8} < \frac{A - 3.2}{0.8} \right] = \text{P}\left[Z < \frac{A - 3.2}{0.8} \right]$$

and this is the quantity we want to be 0.10.

We seek a value v so that P[$Z \leq v$] = 0.10. Apparently such a value must be negative. We see that

0.1 of the probability is found between $-\infty$ and v
0.4 of the probability is found between v and 0
0.4 of the probability is found between 0 and $-v$ (note that $-v$ is positive)
0.1 of the probability is found between $-v$ and ∞

It appears that we must have P[$0 \leq Z \leq -v$] = 0.40.

From the normal table we find

$$\text{P}[\, 0 \leq Z \leq 1.28 \,] = 0.3997$$
$$\text{P}[\, 0 \leq Z \leq 1.29 \,] = 0.4015$$

We will simply use the closer of these; that is, we'll claim that $-v$ = 1.28, so that v = -1.28.

We can complete the solution by solving $\frac{A - 3.2}{0.8}$ = -1.28, giving A = 2.176. This means that, with probability 90%, an investor will have a limit above 2.176 percentage points (and presumably will not balk at having to pay 2.176 percentage points).

You can see the pointlessness of interpolating to get refined answers. Suppose you use the facts above to decide P[$0 \leq Z \leq 1.2817$] = 0.4000. This would cause you to replace "-1.28" above with "-1.2817" and this would change the answer to 2.17464 percentage points. The difference between the original 2.176 percentage points and the interpolated answer 2.17464 percentage points is 0.00236 percentage points, referring to the decimal 0.0000236. This is 23.6 cents on a $10,000 investment.

EXAMPLE 5: Suppose that an automobile muffler is designed so that its lifetime (in months) is approximately normally distributed with mean 26.4 months and standard deviation 3.8 months. The manufacturer has decided to use a marketing strategy in which the muffler is covered by warranty for 18 months. Approximately what proportion of the mufflers will fail the warranty?

SOLUTION: Observe the correspondence between

probability that a single muffler will die before 18 months

and

proportion of the whole population of mufflers that will die before 18 months.

We treat these two notions as equivalent.

Then, letting X denote the random lifetime of a muffler,

$$\text{P}[\ X < 18\ ] \ = \ \text{P}\left[\frac{X - 26.4}{3.8} < \frac{18 - 26.4}{3.8} \right] \approx \ \text{P}[\ Z < -2.21\ ] \ = \ \text{P}[\ Z > 2.21\ ]$$

$$= \ 0.5 - \text{P}[\ 0 \leq Z \leq 2.21\ ] \ = 0.5 - 0.4864 \ = \ 0.0136$$

From the manufacturer's point of view, there is not a lot of risk in this warranty.

EXAMPLE 6: Suppose that the manufacturer in the previous example would like to extend the warranty time to 24 months. Now the risk is considerable, since

$$P[\, X < 24 \,] \;=\; P\left[\frac{X - 26.4}{3.8} < \frac{24 - 26.4}{3.8} \right] \approx\; P[\, Z < -0.63 \,] \;=\; P[\, Z > 0.63 \,]$$

$$=\; 0.5 - P[\, 0 \leq Z \leq 0.63 \,] \;=\; 0.5 - \; 0.2357 = 0.2643$$

More than one-quarter of the mufflers would fail by this standard.

Suppose, though, that the warranty is “pro-rated” in that the customer recovers only the value of the time remaining to 24 months. For example, if a muffler fails at 22.5 months, then the time remaining to 24 months is only 1.5 months, and the customer receives $\frac{1.5}{24} = 0.0625$ of the value of a new muffler. If a muffler costs $64, this is worth $4. Moreover (and this is the point of the problem), of the mufflers that fail the warranty, most will fail by only a short amount of time.

Of all the mufflers that fail, what proportion of them have failures in the interval (20 months, 24 months)?

SOLUTION: We found previously that P[X < 24] = 0.2643.

The next task is to find P[20 < X < 24]

$$=\; P\left[\frac{20 - 26.4}{3.8} \;<\; \frac{X\; - 26.4}{3.8} \;<\; \frac{24 - 26.4}{3.8} \right] \;\approx\; P[\, -1.68 < Z\; -0.63 \,]$$

$$=\; P[\, 0.63 < Z < 1.68 \,] \;=\; P[0 \leq Z < 1.68\,] - P[0 \leq Z \leq 0.63\,] =\; 0.4535 - 0.2357$$

$$=\; 0.2178$$

Here’s how to make sense of these numbers. Suppose that 10,000 mufflers of this type were sold. Of these, you’d expect 2,643 (about) to fail the warranty. However, 2,178 (about) would fail during the period (20 months, 24 months), and the proportion is $\frac{2{,}178}{2{,}643} \approx 0.8241$. That is, about 82% of the warranty failures will be for very short periods of time....and thus very low cost to the manufacturer. In fact, this is probably an excellent strategy, because the customer who returns to collect petty cash on the warranty will probably also be a repeat purchaser!

⁂ ⁂ ⁂ ⁂ ⁂ ADDITIONAL NORMAL DISTRIBUTION EXAMPLE ⁂ ⁂ ⁂ ⁂ ⁂

EXAMPLE 1: The dressed weights of Excelsior Chickens are approximately normally distributed with mean 3.20 pounds and standard deviation 0.40 pound. About what proportion of the chickens have dressed weights greater than 3.60 pounds?

SOLUTION: Let *X* denote the dressed weight of a randomly-selected chicken. Then

$$P[X > 3.60] = P\left[\frac{X - 3.20}{0.40} > \frac{3.60 - 3.20}{0.40}\right] = P[Z > 1.0] = 0.1587$$

About 16% of the chickens will have dressed weights heavier than 3.60 pounds.

EXAMPLE 2: Suppose that the daily demand for change (meaning coins) in a particular store is approximately normally distributed with mean $800.00 and standard deviation $60.00. What is the probability that, on any particular day, the demand for change will be below $600?

SOLUTION: Let *X* be the random amount of change demanded. Then

$$P[X < 600] = P\left[\frac{X - 800}{60} < \frac{600 - 800}{60}\right] \approx P[Z < -3.33] = 0.0004$$

It is exceedingly unlikely that the demand will be below $600.

EXAMPLE 3: Consider the situation of the previous problem. Find the amount *M* of change to keep on hand if one wishes, with certainty 99%, to have enough change. That is, find *M* so that P[$X \leq M$] = 0.99.

SOLUTION:

$$P[X \leq M] = P\left[\frac{X - 800}{60} < \frac{M - 800}{60}\right] = P[Z < \frac{M - 800}{60}] \overset{want}{=} 0.99$$

We could equivalently say that $P[0 < Z < \frac{M - 800}{60}] \overset{want}{=} 0.49$

The normal table reveals that

P[0 < Z ≤ 2.32] = 0.4898 and
P[0 < Z ≤ 2.33] = 0.4901

We'll use the closer of these, namely 2.33.

Then we solve $2.33 = \frac{M - 800}{60}$, giving *M* = 800 + 60(2.33) = 939.80.

The store will need to have \$939.80 in change in order to have a 99% chance of having enough.

Of course, if they decide to keep a stock of \$950 in change, their chance of having enough is $\mathrm{P}[X < 950] = \mathrm{P}\left[\frac{X-800}{60} < \frac{950-800}{60}\right] = \mathrm{P}[Z < 2.5] = 0.9938$.

You can check that keeping a stock of \$1,000 in change will elevate this chance to 0.9996.

> NOTE: Some of the probabilities in examples 2 and 3 are very close to 0 or 1. Specific values encountered were 0.0004, 0.9938, and 0.9996. The phrase "approximately normally distributed" does not justify such refined answers.

EXAMPLE 4: A machine that dispenses corn flakes into packages provides amounts that are approximately normally distributed with mean weight 20 ounces and standard deviation 0.6 ounce. Suppose that the weights and measures law under which you must operate allows you to have only 5% of your packages under the weight stated on the package. What weight should you print on the package?

SOLUTION: Note first of all that the printed weighted weight, call it *w*, must be *below* 20 ounces. If you labeled the packages "20 ounces" you would have about 50% of your packages underweight. If you labeled the packages with something greater than 20 ounces, then more than half the packages would be underweight. Letting *X* denote the random amount dispensed, you want P[$X < w$] = 0.05. Then

$$\mathrm{P}[X \le w] = \mathrm{P}\left[\frac{X-20}{0.6} < \frac{w-20}{0.6}\right] = \mathrm{P}[Z < \frac{w-20}{0.6}] \overset{want}{=} 0.05$$

The normal table reveals that

P[$Z < -1.64$] = 0.0505 and
P[$Z < -1.65$] = 0.0495

This is such an obvious interpolation that we'll use the simple average -1.645.

> Actually, the facts from the table that we used were
> P[$0 < Z < 1.64$] = 0.4495
> P[$0 < Z < 1.65$] = 0.4505

Then we solve $\frac{w-20}{0.6}$ = -1.645 to get *w* = 20 - 0.6(1.645) = 19.013.

It would probably be adequate to label the packages "19 ounces."

⁂ ⁂ ⁂ ⁂ ⁂ ADDITIONAL NORMAL DISTRIBUTION EXAMPLE ⁂ ⁂ ⁂ ⁂ ⁂

EXAMPLE 5: A machine dispenses popcorn into cartons previously labeled "12 ounces." The machine has a setting to adjust the mean amount dispensed, but you have no idea about the standard deviation. Suppose that you set the dispenser at 12.5 ounces, and you find that 9% of the cartons are underweight (below 12 ounces). What is the standard deviation?

SOLUTION: Let X be the random amount dispensed. Let σ be the standard deviation. The facts are

$$0.09 = \text{P}[\,X < 12\,] = \text{P}\left[\frac{X - 12.5}{\sigma} < \frac{12 - 12.5}{\sigma}\right] = \text{P}\left[Z < \frac{12 - 12.5}{\sigma}\right]$$

The normal table reveals that

P[Z < -1.34] = 0.0901 and
P[Z < -1.35] = 0.0885

Let's use -1.34. We now match the facts $\text{P}\left[Z < \frac{12 - 12.5}{\sigma}\right] = 0.09$ and P[Z < -1.34] = 0.09. We decide that $\frac{12 - 12.5}{\sigma} = \frac{-0.5}{\sigma} = -1.34$ to decide that $\sigma = \frac{-0.5}{-1.34} \approx 0.37$.

Apparently the standard deviation of the amount dispensed is about 0.37 ounce.

EXAMPLE 6: An industrial process produces five-liter cans of paint thinner. The history of this process indicates a mean fill of 5.02 liters, with a standard deviation of 0.21 liter. The quality control experts watch this process and select a can for inspection every hour. This process runs for 12 hours every day. The exact contents of the selected can are then determined. The process is said to be "in control" if the volume is within the range 5.02 ± a(0.21). The question here involves the choice of a.

Suppose that a = 2.0. About how often *by chance alone*, will a can be declared out of control?

Repeat this for a = 2.5 and a = 3.0.

SOLUTION: The probability that a normally-distributed random quantity will be within two standard deviations of its mean is $2 \times P[0 < Z < 2] = 0.9544$. Thus, the probability that a single can will lie outside this range is 0.0456, about 4.5%. Since there are 12 inspections per working day, it follows that one will find an out-of-control can about every other day, based on chance alone.

If you change 2.0 to 2.5, the probability is 0.9876, about 99%. It follows that one will find an out-of-control can about every eight working days, based on chance alone.

If you use $a = 3.0$, the probability is 0.9974, about one in four hundred. It follows that one will find an out-of-control can about every thirty working days, based on chance alone.

The technique implied here is generally implemented by plotting the points on a *control chart*. The use of $a = 3.0$ is described as "3σ" limits, and is perhaps the most common choice.

These ideas about control charts should certainly be noted:

> Perhaps the most important benefit of control charts is that it causes people to watch the process.
>
> Control charts force people to confront the concept of statistical variability.
>
> The choice 3σ assures that very few false alarms will be issued. If the process suddenly shifts mean by one standard deviation, the shift will be detected immediately with probability about 2%. (In the paint thinner example, this could be a change from mean 5.02 to 5.02 + 0.21 = 5.23.) It could take a while to notice this shift. On the other hand, shift of two standard deviations will be noticed very quickly.
>
> The example used here dealt only with single cans of paint thinner. In other contexts, one takes small samples and watches the sample standard deviation as well as the sample mean.

It should be emphasized very strongly that the use of control charts puts you in a mindset to *control* the process, not to *improve* it. Among Deming's 14 points for management, here is point #3:

> Cease dependence on inspection to achieve quality. Eliminate the need for inspection on a mass basis by building quality into the product in the first place.

This document will deal with situations in which we talk about a sample of observations.

In statistical jargon, a "sample" denotes an independent set of random variables, each coming from the same distribution. (When the population size is finite, then the definition is modified slightly).

We will use the notation X_1, X_2, …, X_n to denote the sample when discussed as random variables. We would use x_1, x_2, …, x_n to denote the actual numeric values which occur when the sample is finally observed. (In some situations the distinction between random variables and their values is philosophically tortuous, and we will forsake notational rigidity.)

We will use X_i to denote the i^{th} observation in the sample. This is a generic use of the "*i*" symbol.

Since the X_i's all have the same distribution, they must all have the same mean and standard deviation.

Let $\mu = \text{E } X_i$ be used for the mean, and let $\sigma = \text{SD}(X_i)$ be used for the standard deviation.

Two statistically interesting quantities computed from the sample are

$T = \sum_{i=1}^{n} X_i$, the sample total, and $\overline{X} = \frac{T}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$, the sample average (or mean).

We will think of T and $\overline{X}$ as random variables. It happens that

$$\text{E } T = \text{n } \mu \quad \text{and} \quad \text{SD}(\, T\,) = \sigma\sqrt{n}$$

$$\text{E}\overline{X} = \mu \quad \text{and} \quad \text{SD}(\overline{X}) = \frac{\sigma}{\sqrt{n}}$$

There are several very important things to understand about these results:

(1) Here T and $\overline{X}$ are *derived* random variables. This means that they are computed from other random variables. They will have their own means and standard deviations, which exist at a deeper level of abstraction: they are the means and standard deviations of sampling distributions, not of any population of physically identifiable things.

(2) The means (expected values) and standard deviations of T and $\overline{X}$ depend on μ and σ but do not otherwise depend on the original distributions. (In particular, these facts do not require assumptions about normal distributions.)

(3) These results are sometimes described in terms of the variances (which are the squares of the standard deviations). That is, $\text{Var } T = n\sigma^2$ and $\text{Var}(\overline{X}) = \frac{\sigma^2}{n}$.

(4) The sample average $\overline{X}$ is generally used more commonly than the sample total T.

(5) The standard deviation of the sample total T grows with n; totals of many values are variable.

(6) The standard deviation of the sample average $\overline{X}$ decreases with n. As n gets very large, this standard deviation shrinks arbitrarily close to zero; this means that averages converge to the population mean. This fact is sometimes called the "Law of Averages" or "Law of Large Numbers."

(7) If the sample size n is not small, then T and $\overline{X}$ will be approximately normally distributed *even if the original population was not*. This property is called the Central Limit theorem. Generally, most people believe that n bigger than 30 allows you to invoke the Central Limit theorem, though you can often get away with n = 20 or even n = 10.

Specifically, the Central limit theorem allows you to claim that

> $\overline{X}$ is approximately normally distributed with mean μ and with standard deviation $\frac{\sigma}{\sqrt{n}}$.
>
> T is approximately normally distributed with mean $n\mu$ and with standard deviation $\sigma\sqrt{n}$.

The statements about $\overline{X}$ and T are equivalent; it is purely a matter of clerical convenience as to whether one works with totals or averages.

Some standard problems illustrate the use of these ideas.

EXAMPLE 1: Suppose that you have a sample of 100 values from a population with mean 500 and with standard deviation 80. What is the probability that the sample mean will be in the interval (490, 510)?

Each individual X_i has mean 500 and standard deviation 80. It follows that $\overline{X}$, the average of 100 such individuals, will have mean 500 and standard deviation $\frac{80}{\sqrt{100}} = 8$.

The Central Limit theorem allows us to assert that, to an excellent approximation, the average $\overline{X}$ will follow a normal distribution. The sample size, n = 100, is sufficiently large here that the use of the Central Limit theorem will not be questioned.

The mean of this distribution is 500, and the standard deviation is 8. Then

$$\text{P}[\ 490 < \overline{X} < 510\ ] = \ \text{P}\left[\frac{490-500}{8} < \frac{\overline{X}-500}{8} < \frac{510-500}{8}\right]$$

$$= \ \text{P}[\ -1.25 < Z < +1.25\ ] \ = 2\ \text{P}[\ 0 < Z < 1.25\ ] = 2(0.3944) = 0.7888.$$

Observe that in this problem $\overline{X}$ has its own mean and standard deviation.

EXAMPLE 2: Bluefish purchased at the Lime Beach Fishing Terminal produce a filet weight that has a mean of 4.5 pounds with a standard deviation of 0.8 pound. If you purchase five such fish, then what is the probability that you will have at least 21 pounds of filets?

It follows from the given facts that T, the total filet weight of five fish, will have mean $5\mu = 5 \times 4.5 = 22.5$ and standard deviation $0.8\sqrt{5} \approx 0.8 \times 2.236 = 1.7888$. We would like to assert that T is normally distributed, but the sample size $n = 5$ does not really permit use of the Central Limit theorem. We get around this difficulty with use of the caveat "Assuming that the population of filet weights is approximately normally distributed....." Stating the assumption does not make it true, but it is nonetheless important to make the statement.

We should also make the assumption that the five fish were, in some sense, randomly selected. This assumption may also be false, since the objective of clever shopping is to avoid random merchandise, but we state the assumption anyhow.

Answering the question is now very easy:

$$\text{P}[T \geq 21] = \text{P}\left[\frac{T-22.5}{1.7888} \geq \frac{21-22.5}{1.7888}\right] \approx \text{P}[Z \geq -0.84] = 0.5 + \text{P}[0 \leq Z \leq 0.84]$$

$$= 0.5 + 0.2995 = 0.7995 \approx 80\%.$$

It should be noted that this problem could also be done in terms of averages. Let $\overline{X} = \dfrac{T}{n}$ be the average. With $n = 5$, the condition $[T \geq 21]$ is equivalent to $\left[\ \overline{X} = \dfrac{T}{5} \ \geq \ \dfrac{21}{5} = 4.2\right]$.

We note that the expected value of $\overline{X}$ is 4.5, and the standard deviation of $\overline{X}$ is $\dfrac{\sigma}{\sqrt{n}}$ $= \dfrac{0.8}{\sqrt{5}} \approx 0.3578$. Thus we find

$$P[\ \overline{X} \geq 4.2\ ] \ = \ P\left[\frac{\overline{X} - 4.5}{0.3578} \ \geq \ \frac{4.2 - 4.5}{0.3578}\right] \ \approx \ P[\ Z \geq -0.84]$$

and this will necessarily produce exactly the same answer.

EXAMPLE 3: Sometimes the objective of sampling is to estimate the population mean μ, and the sample average $\overline{X}$ is the obvious estimate. The error that results in using $\overline{X}$ to estimate μ is $\overline{X}$ - μ . Since this can be positive or negative, we often unconsciously replace it by $\left|\overline{X} - \mu\right|$.

Suppose that you believe that a population has standard deviation at most 40 pounds. What sample size *n* is required if you want the error of estimation to be less than or equal to 10 pounds with probability at least 95% ?

You are asked to find the sample size *n* which makes certain things happen. There are a lot of loose ends to this problem.

(a) Here σ is assumed to be at most 40 pounds. What if really σ < 40 pounds ?

(b) The error limit has to be achieved with probability at least 95%. Why doesn't the problem ask for exactly 95%?

(c) We'd like to apply the Central Limit theorem to the sample average. Unfortunately, we have not yet determined *n*, so we don't know if the use of the Central Limit theorem will be legitimate.

The required error condition is $\left|\overline{X} - \mu\right| \leq 10$ pounds, and we will rewrite this as $-10 \leq \overline{X} - \mu \leq 10$.

We note that $\overline{X}$ has mean μ (unknown) and standard deviation $\frac{\sigma}{\sqrt{n}} = \frac{40}{\sqrt{n}}$.

Then P[$-10 \leq \overline{X} - \mu \leq 10$] =

$$P\left[\frac{-10}{\frac{40}{\sqrt{n}}} \ \leq \ \frac{\overline{X} - \mu}{\frac{40}{\sqrt{n}}} \ \leq \ \frac{-10}{\frac{40}{\sqrt{n}}}\right] \ = \ P\left[\frac{-\sqrt{n}}{4} \ \leq \ Z \ \leq \frac{\sqrt{n}}{4}\right] \ \overset{\text{want}}{=} \ 0.95$$

We can rephrase this as $P\left[0 \ \leq \ Z \ \leq \ \frac{\sqrt{n}}{4}\right] \ \overset{\text{want}}{=} \ 0.475.$

The normal table reveals that P[0 ≤ Z ≤ 1.96] = 0.4750, so we complete the problem by solving $1.96 = \frac{\sqrt{n}}{4}$, which leads to $n = 7.84^2 \approx 61.47$. We will elevate this answer to the next integer, and we will recommend the use of sample size $n = 62$.

If you trace through these steps algebraically, you can get this formula for n:

$$n \geq \left(\frac{z_{\alpha/2}\ \sigma}{E} \right)^2$$

If you use $z_{\alpha/2} = 1.96$, $\sigma = 40$, $E = 10$, you'll produce $n \geq 61.47$, exactly as above.

We can now address the loose ends noted before:

(a) Here σ is assumed to be at most 40 pounds, but we worked through the problem as though σ = 40 pounds. What if σ < 40 pounds ? In such a situation, our sample size will be overadequate; that is, we will achieve the desired error bound with a probability greater than 95%. (We could also say that the 95% error bound will be smaller than 10 pounds.)

(b) Why doesn't the problem ask for exactly 95%? The figure 95% will not be exactly achievable. In theory, and if σ = 40 pounds, then this requires a non-integer sample size. By going to the next larger integer, the probability is actually elevated a bit above 95%.

(c) The Central Limit theorem turns out to be legitimate here, since the required sample size is quite large. If this problem ended with a small value for *n*, then we simply would have added the assumption that the original population values follow a normal distribution.

EXAMPLE 4: You would like to make a bid on the stock of an out-of-business toy company. This stock consists of 2,860 sealed fiberboard cartons. Before making a bid, you would like to perform an audit to assess the value. How many cartons should you inspect and evaluate if you want to estimate the mean value per carton and if you want your estimate to be within 0.20 standard deviations of the correct value with probability at least 90%?

SOLUTION: This is very similar to the previous problem, except that your target error is expressed in standard deviation units — and you don't know the standard deviation!

In Example 3, the target error was 10 pounds. Here the target is 0.20 standard deviation.

Curiously, we can follow the approach of the previous problem, keeping σ as an unknown algebra symbol. Just note that the standard deviation of $\overline{X}$ is $\frac{\sigma}{\sqrt{n}}$.

Then P[-0.2 σ ≤ $\overline{X}$ - μ ≤ 0.2 σ] =

$$\text{P}\left[\frac{-0.2\sigma}{\frac{\sigma}{\sqrt{n}}} \leq \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq \frac{0.2\sigma}{\frac{\sigma}{\sqrt{n}}}\right] = \text{P}\left[-0.2\sqrt{n} \leq Z \leq 0.2\sqrt{n}\right] \overset{\text{want}}{=} 0.90$$

We can rephrase this as $\text{P}\left[0 \;\leq\; Z \;\leq\; 0.2\sqrt{n}\right] \overset{\text{want}}{=} 0.45$.

The normal table reveals that P[0 ≤ Z ≤ 1.645] = 0.45, so we complete the problem by solving $0.2\sqrt{n}$ = 1.645.

This gives $\sqrt{n} = \dfrac{1.645}{0.2} = 8.225$ and then $n \approx 67.65$.

We would need to sample 68 cartons to obtain an estimate with the desired precision. We note that a sample of 68 is sufficiently large to justify the use of the Central Limit theorem.

It should also be noted that the ultimate sample size of n = 68 is not a large fraction of the population size N = 2,860. Thus finite-population issues can be ignored.

By the way, if you decided to apply the finite-population correction, meaning SD($\overline{X}$) = $\dfrac{\sigma}{\sqrt{n}}\sqrt{\dfrac{N-n}{N-1}}$, you would end up working through the condition

$$\text{P}\left[\;\frac{-0.2\sigma}{\dfrac{\sigma}{\sqrt{n}}\sqrt{\dfrac{2{,}860-n}{2{,}860-1}}} \;\leq\; \frac{\overline{X}-\mu}{\dfrac{\sigma}{\sqrt{n}}\sqrt{\dfrac{2{,}860-n}{2{,}860-1}}} \;\leq\; \frac{0.2\sigma}{\dfrac{\sigma}{\sqrt{n}}\sqrt{\dfrac{2{,}860-n}{2{,}860-1}}}\;\right]$$

$$= \;\text{P}\left[\; -0.2\sqrt{n}\sqrt{\frac{2{,}859}{2{,}860-n}} \;\leq\; Z \;\leq\; 0.2\sqrt{n}\sqrt{\frac{2{,}859}{2{,}860-n}} \;\right]$$

$$= \;2\text{P}\left[\,0 \;\leq\; Z \;\leq\; 0.2\sqrt{n}\sqrt{\frac{2{,}859}{2{,}860-n}}\,\right] \overset{\text{want}}{\geq} 0.90.$$

The condition you'd deal with is $0.2\sqrt{n}\sqrt{\dfrac{2{,}859}{2{,}860-n}} = 1.645$. By squaring both sides, one gets the simple linear equation $0.04n\,\dfrac{2{,}859}{2{,}860-n} = 2.7060$, leading to n = 66.1094. This would be rounded up to n = 67. In this example the population size of 2,860 was fairly large, and dealing formally with the finite population issues has minimal impact on the sample size.

There are many methods to assess whether a sample of data might reasonably be assumed to come from a normal population. One very population graphical method carries the name “normal probability plot” and this is available in Minitab.

Suppose that the data are given by $x_1, x_2, x_3, \dots, x_n$. The procedure requires that these be sorted in increasing order. We’ll assume that this has already been done so that we may write $x_1 \le x_2 \le x_3 \le \dots \le x_n$. Now form points (x_i, y_i) with $y_i = \dfrac{i - A}{n + B}$.

The choices for A and B define the method. As a practical matter, the common choices for A and B are just not that important, and we recommend using the defaults.

The default method uses $A = \frac{3}{8}$ and $B = \frac{1}{4}$.

The Kaplan-Meier method uses $A = 0$ and $B = 0$. This is related to the famous Kaplan-Meier estimate used in survival analysis.

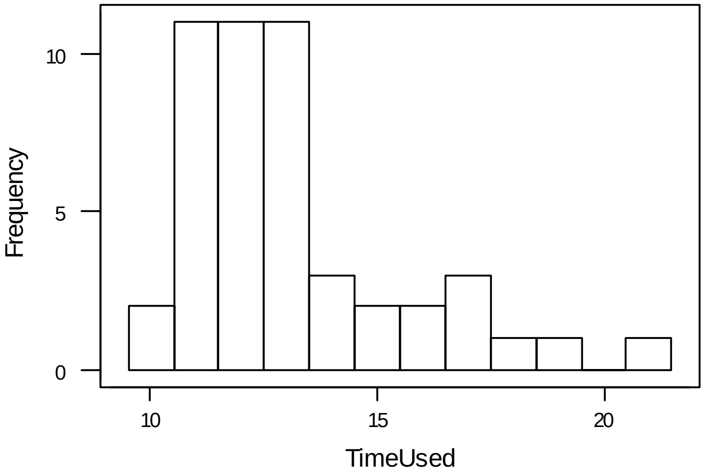
The Herd-Johnson method uses $A = 0$ and $B = 1$.

The points (x_i, y_i) are then plotted. The sorting of the x_i’s and the definition of the y_i’s guarantees that these points will increase from left to right on the graph paper. To check for normal distributions, the vertical scale is stretched and squeezed so that, if the data are perfectly normal, the plot will come out as a perfect straight line. Departures from straightness are used to assess possible non-normality. There are excellent pictures in Hildebrand and Ott, section 6.6.

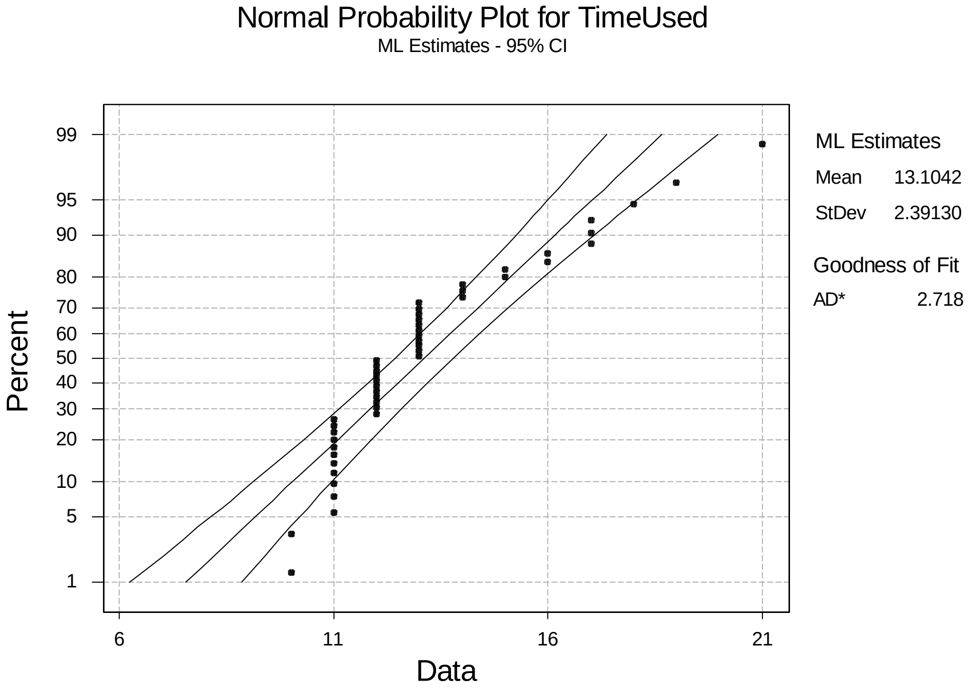
In Minitab, the actual y_i values, given as percents, appear on the vertical axis. In some other packages, the value y_i is replaced by z_i where $P[Z \le z_i] = y_i$. That is, z_i corresponds to normal standard scores, and these z_i’s are plotted on a linear equi-spaced scale.

It also happens that some computer packages reverse the axes, so that the x_i’s are vertical and the y_i’s (or z_i’s) are horizontal.

As an example, let’s consider the data given in Exercise 7.17, pages 211-212 of Hildebrand and Ott, which give the times in minutes for 48 oil change jobs at a “quick lube” shop. Here is a histogram of those data, obtained by **Graph ⇒ Histogram ⇒**.



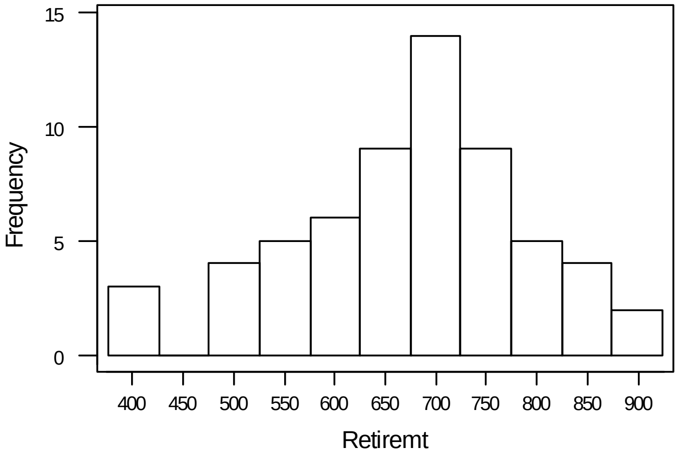
This picture certainly suggests that the sample comes from a non-normal population. We’ll now get the normal probability plot. In Minitab, do **Graph** ⇒ **Probability Plot** ⇒ and then choose **Distribution Normal**. You’ll get the following:



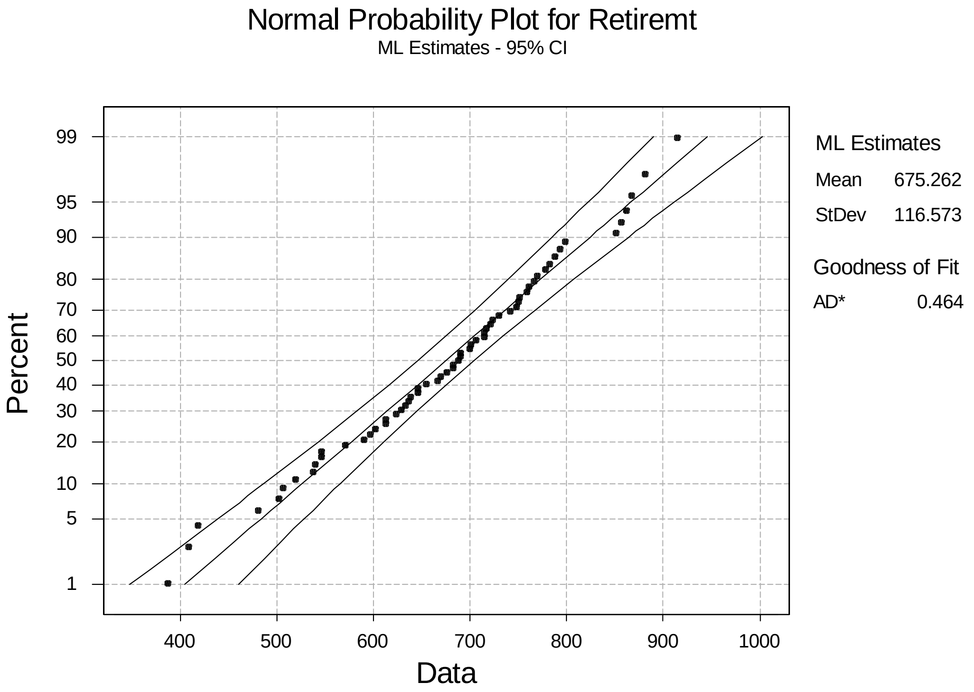
The data dots appear in vertical stacks since there were tied values among the set of 48.

The straight line is based on estimates of the mean and standard deviation, along with the assumption that the data actually come from a normal population. The curved bands represent “confidence limits” for the percentiles; for instance, the band opposite 20 on the vertical scale gives limits for the 20th percentile. The numbers relevant to these bands will be given in the Session window. The fact that data points fall outside these bands is strong evidence that the data really do not come from a normal population. The particular style of curvature shown here suggests positive skewness.

Let’s consider an example which would support the assumption of normal distributions. The file CASE07.MTP covers the story given on pages 246-247 of Hildebrand and Ott. Three variables are provided for each of 61 employees, and we’ll look at the column Retiremt, representing the yearly retirement costs for the employer. This is the histogram



This picture suggests a reasonably symmetric set of data, and normal distributions are certainly plausible. For these data, the normal probability plot is this:



Here the dots seem to stay generally within the 95% boundaries. There are some twists and wiggles, but we should be willing to say that the data are reasonably normal.

❽❽❽❽❽❽❽❽ CENTRAL LIMIT THEOREM ❽❽❽❽❽❽❽❽

In dealing with the Central Limit theorem, there are a number of important ideas.

(1) It’s important to distinguish between a sample of measured (continuous) data and a sample of yes/no data. The yes/no situation is simply that of binomial sampling, and we are usually interested only in the total number of “yes” responses.

(2) With continuous data, we need to distinguish the situation in which we are “sampling from a normally distributed population” from the situation in which we are “sampling from a population which is not normally distributed.” Obviously we cannot distinguish perfectly, and we may need to deal with the notion of approximately normal populations.

> (2a) With small sample sizes, we simply do not have enough data to make an informed judgment as to whether the population values do or do not follow a normal distribution.
>
> (2b) With moderate sample sizes (say n = 20 to n = 60) we can plot histograms and compute skewness coefficients, but these may not be decisive enough to help us decide.
>
> (2c) With large sample sizes, we can usually make a pretty good decision about normality. Because of the Central Limit theorem, however, it is not really very important whether the population values follow a normal distribution or not.

(3) It is vital to distinguish the sample X_1, X_2, …, X_n from the sample average $\overline{X}$.

> (3a) A large sample size does not do anything regarding the normality or non-normality of the sample values. If you are sampling from a non-normal population, then large n simply means that you have lots of values from a non-normal population.
>
> (3b) A large sample size will enable you to decide whether you are sampling from a normal population or a non-normal population.
>
> (3c) A large sample size allows you to invoke the Central Limit theorem, and this specifically lets you claim that $\overline{X}$ (or, equivalently, the total $T = n\overline{X}$) follows approximately a normal distribution.

(4) One makes assumptions about the population for the purpose of using some standard statistical procedures.

(4a) Stating an assumption about a population does not make that assumption true. Stating the assumption acts as a disclaimer.

(4b) With small sample sizes, generally meaning $n < 30$, from a continuous population, one assumes that the population values follow a normal distribution in order to use the conventional confidence interval $\bar{x} \pm t_{\alpha/2;n-1} \frac{s}{\sqrt{n}}$ for the population mean μ or to use the usual t test for the hypothesis H_0: $\mu = \mu_0$.

(4b1) You cannot state the assumption as "$\overline{X}$ is normal" because you need normality of the population in order to get the correct distribution for s (and legitimize the use of the cutoff point $t_{\alpha/2;n\text{-}1}$ from the t table).

(4b2) You cannot state the assumption as "the sample is normal" because this creates an inferential mystery. What could it mean to have a normal sample from a population which is not necessarily normal?

(4c) With large sample sizes, generally meaning $n \geq 30$, from a continuous population, the distribution of $\overline{X}$ is well approximated by a normal distribution. Thus we may use the conventional confidence interval or perform the usual t test.

(4c1) The confidence interval may be written as $\bar{x} \pm t_{\alpha/2;n-1} \frac{s}{\sqrt{n}}$.

(4c2) The use of the t value requires that s be based on a sample from a normal population, and the Central Limit theorem does not cover the distribution of s. However, with a large n, we can be sure that $s \approx \sigma$. Also with large n we have $t_{\alpha/2;n\text{-}1} \approx z_{\alpha/2}$. Thus, this interval is very close to the interval $\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$, which is the "known σ" case.

(4c3) This interval (with large n) is sometimes written as $\bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}}$ or even as $\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$.

(5) For binomial data, the Central Limit theorem simply says that the distribution of *X*, the total number of successes, may be approximated with a normal distribution. This works reasonably well if $n \geq 30$, $np \geq 5$, and $n(1 - p) \geq 5$.

(5a) If we do not meet the conditions on *n* and *p* noted above, there is no assumption that can be used as a prelude to a normal-based procedure. In particular, we cannot say "assuming that the population values follow a normal distribution."

(5b) Probability calculations should use a continuity correction. Thus, a question of the form P[$X \geq 20$] should be restated as P[$X > 19.5$]. Similarly, P[$X > 22$] should be restated as P[$X > 22.5$].

(5c) The usual estimate of *p* is $\hat{p} = \frac{X}{n}$, and the corresponding 1 - α confidence interval is $\hat{p} \pm \left[z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} + \frac{1}{2n} \right]$. This is usually given without the term $\frac{1}{2n}$; however the form given here comes closer to achieving the coverage probability 1 - α. A confidence interval procedure which seems to work even better is based on the calculation $\tilde{p} = \frac{X+2}{n+4}$; the interval is given as $\tilde{p} \pm \left[z_{\alpha/2} \sqrt{\frac{\tilde{p}\left(1-\tilde{p}\right)}{n+4}} \right]$. The $\tilde{p}$ form is especially useful if *n* is small or if $\hat{p}$ is very close to 0 or very close to 1.

(5d) The confidence interval should never be given with $t_{\alpha/2;n-1}$ as there is no logical connection to the *t* distribution.

(5e) The test of the hypothesis H_0: $p = p_0$ should be based on the test statistic $Z = \sqrt{n} \frac{\hat{p} - p_0}{\sqrt{p_0\left(1-p_0\right)}}$.

(5e1) The comparison point for *Z* comes from the normal table. There is no logical connection to the *t* distribution.

(5e2) Some users advocate a continuity correction and give the test statistic as $\sqrt{n} \frac{\left(\hat{p} - p_0\right) \pm \frac{1}{2n}}{\sqrt{p_0\left(1-p_0\right)}}$. The ± sign is used to bring the calculation closer to zero. Thus, if ($\hat{p}$ - p_0) > 0, use $-\frac{1}{2n}$; if ($\hat{p}$ - p_0) < 0, use $+\frac{1}{2n}$. There is considerable disagreement about the appropriateness of this continuity correction.

(5e3) You will sometimes see the test statistic in the form $\sqrt{n} \frac{\hat{p} - p_0}{\sqrt{\hat{p}(1-\hat{p})}}$. This is numerically very close to the form given in (5e).

❽❽❽❽❽❽❽❽ CENTRAL LIMIT THEOREM ❽❽❽❽❽❽❽❽

Finally, let’s see an illustration of the Central Limit theorem at work. Here is a sample of size 100 from a population about which we know very little:

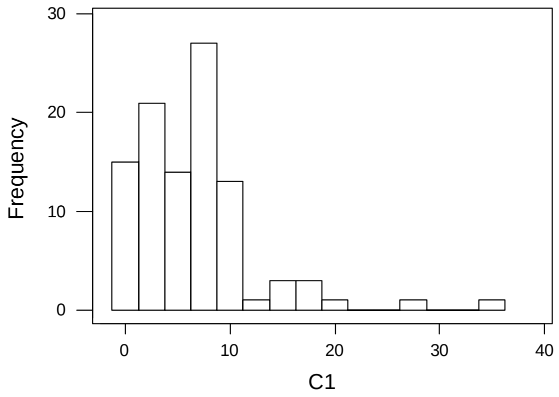


Figure 1

This is a fairly irregular shape, and we would certainly believe that the population values do not follow a normal distribution. For these data $\overline{x}$ = 6.481 and s = 5.773. We do not know the population mean and standard deviation, but with this sample of n = 100, we certainly believe that μ is near 6.481 and that σ is near 5.773.

> As this was a simulation, we can let you in on the secret. The population from which these were generated had μ = 5.9 and σ ≈ 4.5376.

If we took averages of samples of 5, we would at least begin to approximate normal distributions. The histogram below shows the results of taking 100 samples, each of size 5, and recording the averages.

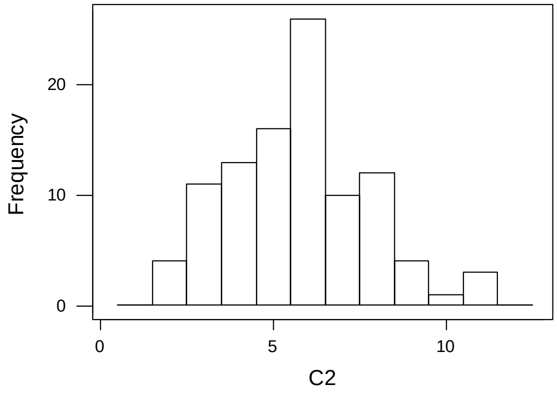


Figure 2

This distribution is much more symmetric, but it would be hard to say whether it is normal or not.

> By the way, the mean of the 100 versions of $\overline{x}$ (each of them an average of 5 values) here is 5.698, and this is rather close to the true mean $\mu = 5.9$. Also, the standard deviation of the 100 versions of $\overline{x}$ is 2.025; this should correspond to $\frac{\sigma}{\sqrt{n}} = \frac{4.5376}{\sqrt{5}} \approx 2.0293$.

Now let's consider what would happen if we took samples of size 30. The next histogram shows the results of taking 100 samples, each of size 30, and recording the averages.

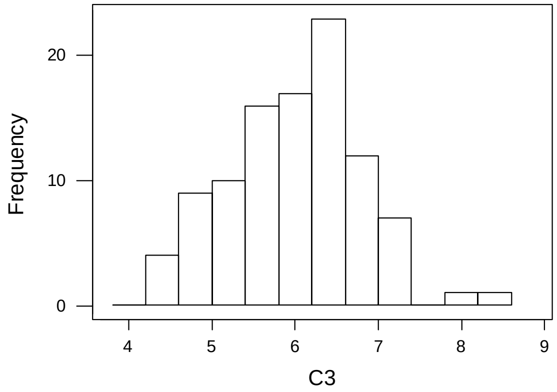


Figure 3

We now see a shape that is looking very close to normal.

The mean of the 100 versions of $\bar{x}$ (each of them an average of 30 values) here is 5.9662, and this is rather close to the true mean $\mu = 5.9$. Also, the standard deviation is 0.8031; this should correspond to $\frac{\sigma}{\sqrt{n}} = \frac{4.5376}{\sqrt{30}} \approx 0.8284$.

We can summarize the findings as follows:

Data collected	Unobserved population quantities		Calculated sample quantities		Histogram of 100
	Expected value	Standard deviation	Sample mean	Sample standard deviation	
Single value	5.9	4.5376	6.481	5.773	Figure 1
Average of sample of 5	5.9	2.0293	5.698	2.025	Figure 2
Average of sample of 30	5.9	0.8284	5.9662	0.8031	Figure 3

You might be curious as to how the data were actually generated.

With probability 0.7, a value was sampled from an exponential distribution with mean 5.

With probability 0.3, a data value was taken from an exponential distribution with mean 2, and then the value 6 was added.

This could be described as $0.7 \times \text{Expo}(\mu=5) + 0.3 \times [\, 6 + \text{Expo}(\mu=2)\,]$.

⌤⌤ NORMAL APPROXIMATION FOR BINOMIAL PROBABILITIES⌤⌤

Suppose that you sell 179 washing machines and with each sale you offer the buyer the opportunity to purchase an extended warranty. The probability that any individual will buy the extended warranty is 0.38. Find the probability that 70 or more will buy the extended warranty.

This is clearly a binomial situation. With n = 179 independent customers, we will let X be the (random) number of them who purchase the extended warranty. Thus X will be a binomial random variable with n = 179 and p = 0.38. We ask P[$X \geq 70$].

You can use a program like Minitab to get this probability. In fact, Minitab obtains this value as 1 - P[$X \leq 69$] = 1 - 0.5924 = 0.4076.

We will show here the workings of the normal approximation to the binomial. We note that E X = 179×0.38 = 68.02 and SD(X) = $\sqrt{179 \times 0.38 \times 0.62} \approx 6.4940$.

We will convert the request P[$X \geq 70$] into P[$X > 69.5$]. This half-integer adjustment is called the *continuity correction*. There are several explanations that could be made.

1. For the binomial, the event { $X \geq 70$ } is different from the event { $X > 70$ }. Since the normal distribution is continuous, it does not distinguish ≥ from >. The use of half-integer boundaries saves us from these confusions.
2. The probability histogram from the binomial distribution would consist of bars situated so that their centers align with the integers. That is, the bar representing P[$X = 65$] would be centered over 65. Said another way, that bar would extend from $64\frac{1}{2}$ to $65\frac{1}{2}$. The event { $X \geq 70$ } is really the event { $X = 70$ } ∪ { $X = 71$ } ∪ { $X = 72$ } ∪ … The corresponding probability bars run from 69.5 to 70.5, then 70.5 to 71.5, then 71.5 to 72.5, and so on. Thus, the probability accounting starts from 69.5.

The continuity correction greatly improves the answer (relative to the exact calculation) when n is small. For larger n, say 500 or more, the continuity correction offers only a small improvement.

Then

$$\text{P}[\ X > 69.5\ ] \ = \ \text{P}\left[\frac{X - 68.02}{6.4940} > \frac{69.5 - 68.02}{6.4940}\right] \approx \ \text{P}[\ Z > 0.228\ ]$$

$$= \ 0.5 - \text{P}[\ 0 \leq \ Z \leq 0.228\ ] \ = \ 0.5 \ - 0.0910 \ = \ 0.4090$$

This answer was found by grabbing the closer entry in the table; that is, we used P[$0 \leq \ Z \leq 0.23$] = 0.0910.

As an approximation to the exact answer 0.4076, this is reasonable, but not exquisite. The error of approximation is $\frac{0.4090 - 0.4076}{0.4076} \approx 0.0034 = 0.34\%$. The approximation was large by about $\frac{1}{3}$ of one percent.

> Comment 1: If a high-quality answer is critical, you can try to interpolate in using the normal table. Depending on how the rounding went, you do not necessarily get a better answer. Here we would do
>
> $$P[\ 0 \leq Z \leq 0.228\ ] = 0.0871 + 0.8 \times 0.0039 = 0.09022$$
>
> and an approximating probability of 0.40978. This is actually a little *farther* away from the exact answer.
>
> Comment 2: The continuity correction is important to getting a quality answer. If you had left this problem as P[$X \geq 70$], you'd get
>
> $$P\left[\frac{X - 68.02}{6.4940} > \frac{70 - 68.02}{6.4940} \right] \approx P[\ Z > 0.3049\ ]$$
>
> $$= 0.5 - P[\ 0 \leq Z \leq 0.3049\ ] \approx 0.5 - P[\ 0 \leq Z \leq 0.30\ ]$$
>
> $$= 0.5 - 0.1179 = 0.3821$$
>
> This is a *much* worse answer.

The normal approximation to the binomial works less well (in terms of proportional error) for events of very small probability, even with the continuity correction. Suppose that you had wanted P[$X \leq 55$]. Minitab gives the result as 0.0257. For the normal approximation we change P[$X \leq 55$] to P[$X \leq 55.5$], and then do this:

$$P[\ X \leq 55.5\ ] = P\left[\frac{X - 68.02}{6.4940} \leq \frac{55.5 - 68.02}{6.4940} \right] \approx P[\ Z \leq -1.928\ ]$$

$$= P[\ Z \geq 1.928\ ] = 0.5 - P[\ 0 \leq Z \leq 1.928\ ] = 0.5 - 0.4732 = 0.0268$$

The error of approximation is $\frac{0.0268 - 0.0257}{0.0257} \approx 0.0428 = 4.28\%$. This is really not very good.

ANOTHER LAYOUT FOR THE NORMAL TABLE

	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-4.0	.00003	.00003	.00003	.00003	.00003	.00003	.00002	.00002	.00002	.00002
-3.9	.00005	.00005	.00004	.00004	.00004	.00004	.00004	.00004	.00003	.00003
-3.8	.00007	.00007	.00007	.00006	.00006	.00006	.00006	.00005	.00005	.00005
-3.7	.00011	.00010	.00010	.00010	.00009	.00009	.00008	.00008	.00008	.00008
-3.6	.00016	.00015	.00015	.00014	.00014	.00013	.00013	.00012	.00012	.00011
-3.5	.00023	.00022	.00022	.00021	.00020	.00019	.00019	.00018	.00017	.00017
-3.4	.00034	.00032	.00031	.00030	.00029	.00028	.00027	.00026	.00025	.00024
-3.3	.00048	.00047	.00045	.00043	.00042	.00040	.00039	.00038	.00036	.00035
-3.2	.00069	.00066	.00064	.00062	.00060	.00058	.00056	.00054	.00052	.00050
-3.1	.00097	.00094	.00090	.00087	.00084	.00082	.00079	.00076	.00074	.00071
	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.0	.00135	.00131	.00126	.00122	.00118	.00114	.00111	.00107	.00104	.00100
-2.9	.00187	.00181	.00175	.00169	.00164	.00159	.00154	.00149	.00144	.00139
-2.8	.00256	.00248	.00240	.00233	.00226	.00219	.00212	.00205	.00199	.00193
-2.7	.00347	.00336	.00326	.00317	.00307	.00298	.00289	.00280	.00272	.00264
-2.6	.00466	.00453	.00440	.00427	.00415	.00402	.00391	.00379	.00368	.00357
-2.5	.00621	.00604	.00587	.00570	.00554	.00539	.00523	.00508	.00494	.00480
-2.4	.00820	.00798	.00776	.00755	.00734	.00714	.00695	.00676	.00657	.00639
-2.3	.01072	.01044	.01017	.00990	.00964	.00939	.00914	.00889	.00866	.00842
-2.2	.01390	.01355	.01321	.01287	.01255	.01222	.01191	.01160	.01130	.01101
-2.1	.01786	.01743	.01700	.01659	.01618	.01578	.01539	.01500	.01463	.01426
	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-2.0	.02275	.02222	.02169	.02118	.02068	.02018	.01970	.01923	.01876	.01831
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
-1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
-1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
-1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09

ANOTHER LAYOUT FOR THE NORMAL TABLE

	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
-0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
-0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
-0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
-0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
-0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
-0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
-0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
-0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
-0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
-0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7703	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389

	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09

ANOTHER LAYOUT FOR THE NORMAL TABLE

	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986

	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
3.0	.99865	.99869	.99874	.99878	.99882	.99886	.99889	.99893	.99897	.99900
3.1	.99903	.99906	.99910	.99913	.99916	.99918	.99921	.99924	.99926	.99929
3.2	.99931	.99934	.99936	.99938	.99940	.99942	.99944	.99946	.99948	.99950
3.3	.99952	.99953	.99955	.99957	.99958	.99960	.99961	.99962	.99964	.99965
3.4	.99966	.99968	.99969	.99970	.99971	.99972	.99973	.99974	.99975	.99976
3.5	.99977	.99978	.99978	.99979	.99980	.99981	.99981	.99982	.99983	.99983
3.6	.99984	.99985	.99985	.99986	.99986	.99987	.99987	.99988	.99988	.99989
3.7	.99989	.99990	.99990	.99990	.99991	.99991	.99992	.99992	.99992	.99992
3.8	.99993	.99993	.99993	.99994	.99994	.99994	.99994	.99995	.99995	.99995
3.9	.99995	.99995	.99996	.99996	.99996	.99996	.99996	.99996	.99997	.99997
4.0	.99997	.99997	.99997	.99997	.99997	.99997	.99998	.99998	.99998	.99998
	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09

THE RANDOM WALK MODEL FOR STOCK PRICES

Suppose that we have an infinite population represented by the generic random variable X. We can think of an unlimited random sampling process resulting in the unending string of random variables X_1, X_2, X_3, X_4, …

We will conceptualize these X_i's as random. We can use the lower case symbols x_1, x_2, x_3, x_4, ... for possible numeric values.

Let's suppose that the population has a mean μ and a standard deviation σ.

Let's also assume that X_0 is a known non-random starting value.

Let $T_n = X_0 + X_1 + X_2 + X_3 + X_4 + \ldots + X_n$ be the n^{th} total. Observe that $T_0 = X_0$, $T_1 = X_0 + X_1$, and in general $T_n = T_{n-1} + X_n$; that is, each total is the previous total plus one new X_n .

The sequence of running totals T_0, T_1, T_2, T_3, T_4, … is called a *random walk*. Implicit in this notion is the independence of the successive differences

$X_0 = T_0$

$X_1 = T_1 - T_0$

$X_2 = T_2 - T_1$

$X_3 = T_3 - T_2$

$X_4 = T_4 - T_3$

$X_5 = T_5 - T_4$

….

Since the T_n values are a form of sample totals, we have E $T_n = T_0 + n\mu$ and $\text{SD}(T_n) = \sigma\sqrt{n}$.

Many things can be conceptualized as random walks. For instance, if X_n is the number of papers that a news vendor sells on day n, then the sequence T_0, T_1, T_2, T_3, T_4, … gives the cumulative sales. (Here T_0 represents the carry-over from the previous accounting period; perhaps T_0 would be zero in this example.)

The role of random walks in stock prices has been viciously debated. We'll look at two such models. The first model is instructive; it is much more useful for modeling cumulative sales than it is for stock prices. The second model is more difficult, but it is used with great frequency in dealing with stock prices.

MODEL 1: Normal distributions for stock price changes.

Let P_0 be the price of a certain stock at the beginning of our observation period. The value of P_0 will be regarded as known and nonrandom.

> Of course P is also the symbol we use for probability. The context should make clear exactly which meanings are involved.

We will think of the daily *changes* X_1, X_2, X_3, X_4, … as independent random quantities, each with a normal distribution with mean μ and standard deviation σ. In this model, μ and σ are in money units (such as dollars).

> Daily prices need not be used. This description works for weekly prices or monthly prices. It also works for prices on 15-second intervals. In this discussion, we've ignored the discreteness of stock prices, which are traded in one-cent increments. (They used to be traded in eighths of dollars!) And yes, the use of normal distributions constitutes an assumption.

Observe that $P_n = P_0 + (X_1 + X_2 + \ldots + X_n)$. It's convenient to let $T_n = X_1 + X_2 + \ldots + X_n$, so that we can write $P_n = P_0 + T_n$.

Of course, $\mathrm{E}(T_n) = n\mu$ and $\mathrm{SD}(T_n) = \sigma\sqrt{n}$.

It follows that $\mathrm{E}(P_n) = P_0 + n\mu$ and $\mathrm{SD}(P_n) = \sigma\sqrt{n}$.

For stock-market applications, it is frequently assumed that $\mu = 0$, to be interpreted as no net drift for stock prices. This leads to $\mathrm{E}(P_n) = P_0$. (This makes the stock price sequence into a *martingale*, but that's another story.) Since P_n is the price of the stock n days from now, the result $\mathrm{SD}(P_n) = \sigma\sqrt{n}$ reflects our uncertainty about the future.

As an example, suppose that $P_0 =$ \$40, $\mu =$ \$0.01, and $\sigma =$ \$0.28. Let's find the probability that the price will exceed \$41 after 25 days. Let $T_n = X_1 + X_2 + \ldots + X_n$ be the cumulative sum of the daily changes. Note that $\mathrm{E}(T_n) = n\mu = 25 \times \$0.01 = \$0.25$ and $\mathrm{SD}(T_n) = \sigma\sqrt{n} = \$0.28\sqrt{25} = \$1.40$. Then

$$\mathrm{P}[\, P_{25} > \$41 \,] = \mathrm{P}[\, P_0 + T_n > \$41 \,] = \mathrm{P}[T_n > \$1 \,]$$

$$= \mathrm{P}\left[\frac{T_n - \$0.25}{\$1.40} > \frac{\$1 - \$0.25}{\$1.40} \right] \approx \mathrm{P}[\, Z > 0.54 \,]$$

$$= 0.50 - \mathrm{P}[\, 0 \le Z \le 0.54 \,] = 0.50 - 0.2054 = 0.2946$$

If the daily changes are assumed to follow a normal distribution, then the use of Z is exact. Even without this assumption, the sample size of 25 is probably enough to justify the use of the Central Limit theorem.

One can also give a 95% prediction interval for the price of the stock after 25 days. We note that $\text{E}(P_{25}) = P_0 + \text{E}(T_n) = \$40 + \$0.25 = \$40.25$. Next, $\text{SD}(P_{25}) = \text{SD}(Y) = \sigma\sqrt{25} = \$0.28 \times \sqrt{25} = \$1.40$. We predict with 95% probability that the stock price will be in the interval $\$40.25 \pm 1.96(\$1.40)$, which is ($37.51, $42.99).

> The "1.96" is an exact use of the normal table, since
>
> $P[\ -1.96 \le Z \le 1.96\ ] = 2\ P[\ 0 \le Z \le 1.96\ ] = 0.95.$

Many people are content to replace "1.96" with "2." This would give the interval as $\$40.25 \pm \$2.80$.

> Please note that this is a *prediction* interval, since we are making an inference about the future value of some random variable. (Confidence intervals, by way of contrast, are used to trap nonrandom parameters.)

One commonly expressed dissatisfaction with this model is that the normal distribution allows the possibility (albeit with very low probability) of negative values for the stock price. Certainly we would worry about this model for so-called penny stocks, which often trade at prices below $1.

MODEL 2: Log-normal distributions for stock price changes.

All logarithms in this discussion are base-*e*.

Let P_0 be the price of a certain stock at the beginning of our observation period. The value of P_0 will be regarded as known and nonrandom. We will think of the daily changes in terms of price ratios $\frac{P_i}{P_{i-1}}$. We will model the logarithms of these ratios, meaning things of the form $\log\left[\frac{P_i}{P_{i-1}}\right]$, as independent random quantities, each with a distribution with mean μ and standard deviation σ.

If the random quantities $\log\left[\frac{P_i}{P_{i-1}}\right]$ are assumed to follow a normal distribution, then the ratios $\frac{P_i}{P_{i-1}}$ are said to follow a *lognormal* distribution. As a result, this is often described as the lognormal model for stock prices.

In this model, μ and σ are parameters of a population of logarithms of ratios and thus are unit-free quantities. In particular, they are not in dollars or any other currency. The parameter μ represents the *drift* in the model and σ is a measure of *volatility*.

The notational scheme is this:

$P_0 = e^{X_0}$ $\qquad$ $X_0 = \log P_0$

$P_1 = P_0\, e^{X_1}$ $\qquad$ $X_1 = \log\left(\frac{P_1}{P_0}\right) = \log P_1 - \log P_0$

$P_2 = P_1\, e^{X_2}$ $\qquad$ $X_2 = \log\left(\frac{P_2}{P_1}\right) = \log P_2 - \log P_1$

$P_3 = P_2\, e^{X_3}$ $\qquad$ $X_3 = \log\left(\frac{P_3}{P_2}\right) = \log P_3 - \log P_2$

$P_4 = P_3\, e^{X_4}$ $\qquad$ $X_4 = \log\left(\frac{P_4}{P_3}\right) = \log P_4 - \log P_3$

or in general

$P_n = P_{n\text{-}1}\, e^{X_n}$ $\qquad$ $X_n = \log\frac{P_n}{P_{n-1}} = \log P_n - \log P_{n\text{-}1}$

By using back substitution, we can show that

$P_n = P_0\, e^{X_1+X_2+\ldots+X_n}$ $\qquad$ $X_1 + X_2 + \ldots + X_n = \log\frac{P_n}{P_0}$

Note that P_0 and X_0 are considered non-random. Let $T_n = X_1 + X_2 + X_3 + \ldots + X_n$; observe that $T_n = \log \frac{P_n}{P_0}$. Then T_n has an expected value of $n\mu$ and a standard deviation of $\sigma\sqrt{n}$.

The most elegant representation for the log-normal random walk is this:

$$P_n = P_0 \, e^{T_n}$$

There is an immediate parallel with the present-value formula $V_t = V_0 \; e^{r\,t}$.

As an example, suppose that P_0 = \$40, μ = 0, and σ = 0.02. Let's find the probability that P_{25}, the price of the stock after 25 days, will exceed \$45. Now T_{25} has mean value $25\mu = 0$ and standard deviation $\sigma\sqrt{25} = 0.02 \times 5 = 0.1$ and it is approximately normally distributed. (With a sample of $n = 25$, we can reasonably resort to the Central Limit theorem.) Then find

$$\text{P}[\ P_{25} > \$45\ ] = \text{P}\left[\ P_0\, e^{T_n} > \$45\ \right] = \text{P}\left[\ \$40\, e^{T_n} > \$45\ \right]$$

$$= \text{P}\left[\ e^{T_n} > 1.125\ \right] = \text{P}[\ T_n > \log(1.125)\ ] \approx \text{P}[\ T_n > 0.1178\ ]$$

It is important here that base-*e* logarithms are used (and not base-10).

$$= \text{P}\left[\ \frac{T_n - 0}{0.1} > \frac{0.1178 - 0}{0.1}\right] \approx \text{P}[\ Z > 1.18\ ] = 0.50 - \text{P}[\ 0 \le Z \le 1.18\ ]$$

$$= 0.50 \ - \ 0.3810 = 0.1190$$

A similar logic can be used for prediction intervals. We are 95% certain that $T_{25} = \log \frac{P_{25}}{P_0}$ is in the interval $0 \pm (1.96)(0.1)$, which is $(-0.196, 0.196)$. We are 95% confident that

$$-0.196 \le \log \frac{P_{25}}{P_0} \le 0.196$$

or equivalently

$$\log P_0 - 0.196 \le \log P_{25} \le \log P_0 + 0.196$$

This means that we are 95% confident that

$$e^{\log P_0}\ e^{-0.196} \quad \leq \quad e^{\log P_{25}} \quad \leq \quad e^{\log P_0}\ e^{0.196}$$

or, after a little clean-up,

$$P_0\ e^{-0.196} \quad \leq \quad P_{25} \quad \leq \quad P_0\ e^{0.196}$$

We can calculate (with the exponential function on a calculator) $e^{-0.196}$ = 0.8220 and $e^{0.196}$ = 1.2165, and this will lead us to

$$0.8220\ P_0 \ \leq\ P_{25} \ \leq\ 1.2165\ P_0$$

Since P_0 = \$40, the interval is \$32.88 to \$48.66.

A few comments about this....

While the quantity X_n can be positive or negative, the value of e^{X_n} is always positive. The relationship $P_n = P_{n\text{-}1}\ e^{X_n}$ will thus always produce positive prices. This gets around the objection to negative values which haunts model 1.

Model 2 puts a probability structure on the *proportional* changes $\dfrac{P_n}{P_{n-1}}$. This seems to be more reasonable than putting structure on the dollar-value changes, as in mode l.

There is an interesting side consequence to this model. You can see that the center of the confidence interval is at \$40.77, which is a little bit higher than \$40. This happens even though we put a mean of zero on the X_i 's. If you believe in the lognormal model, you must make money in the stock market because the gains tend to outweigh the losses, even when the market drift parameter μ is zero.

Here's a different numeric story for the second model, using a positive value of μ.

Suppose, as before, that P_0 = \$40 and σ = 0.02, but now let μ = 0.005. We want to give a 95% prediction interval for P_{25}, the price of the stock after 25 days. As before, $T_n = \log \dfrac{P_n}{P_0}$, but now T_{25} has mean value 25μ = 0.125 and standard deviation $\sigma\sqrt{25}$ = 0.1.

We are 95% certain that $T_{25} = \log \dfrac{P_{25}}{P_0}$ is in the interval 0.125 ± (1.96)(0.1), which is (-0.071, 0.321). We are 95% confident that

$$-0.071 \ \le\ \log \frac{P_{25}}{P_0} \ \le\ 0.321$$

or equivalently

$$\log P_0 - \ 0.071 \ \le\ \log P_{25} \ \le\ \log P_0 + 0.321$$

This means that we are 95% confident that

$$e^{\log P_0}\ e^{-0.071} \ \le\ e^{\log P_{25}} \ \le\ e^{\log P_0}\ e^{0.321}$$

or, after a little clean-up,

$$P_0\ e^{-0.071} \ \le\ P_{25} \ \le\ P_0\ e^{0.321}$$

We have $e^{-0.071}$ = 0.9315 and $e^{0.321}$ = 1.3785, and this will lead us to

$$0.9315\ P_0 \ \le\ P_{25} \ \le\ 1.3785\ P_0$$

Since P_0 = \$40, the interval is \$37.26 to \$55.14. The positive value of μ gives a nice kick to this interval.