Reading the Z tables

Values on the left side and top of the Z table give the value of Z, For example, to find Z=0.11,

read the integer portion and first decimal part (0.1) along the left side and,

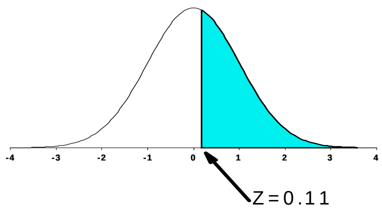
find the second decimal (0.01) along the top

The intersection of these gives the probability of a greater value of Z, in this case P(Z ≥ +0.11) = 0.4562.

Note that the value of Z=0.00 has a probability of 0.5, so half of the distribution is above this value (and half below)

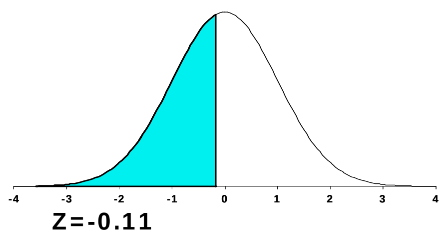
Working with Z tables

What did we just do?



We found the area under the curve above a value of Z=0.11. The values in the available tables will always be giving the r.c.f. of the upper area of the curve.

What if we want to work with the lower half of the curve?



Due to symmetry in the distribution, the probability of a randomly selected value falling in the negative area to the left is the same as the corresponding positive area, so P(Z ≥ +0.11) = P(Z ≤ – 0.11)

Some things we know from previous discussions of the empirical rule.

P(Z ≥ 0) = P(Z ≤ 0) = 0.5.

The probability that a randomly selected Z falls between the limits $\mu - 1\sigma$ and $\mu + 1\sigma$ is about 68%, and half of the remaining fall in each of the tails (about 16%). Since $\sigma = 1$ for the standard normal, we should have about 16% above +1, and 16% below –1. Looking this up in the table we see P(Z ≥ +1) = 0.1587. Due to symmetry P(Z ≤ –1) is also 0.1587.

The probability that a randomly selected Z falls between the limits $\mu - 1.96\sigma$ and $\mu + 1.96\sigma$ is 95%, and half of the remaining fall in each of the tails (about 2.5%). Since $\sigma = 1$ for the standard normal, we should have about 2.5% above 1.96, and 2.5% below –1.96. Looking this up in the table we see P(Z ≥ +1.96) = 0.0250, and P(Z ≤ –1.96) would be the same.

A memorable value, 1.96!

The probability that a randomly selected Z falls between the limits μ–2.576σ and μ+2.576σ is about 99%, and half of the remaining fall in each of the tails (about 0.5%). Since $\sigma = 1$ for the standard normal, we should have about 0.5% above 2.576, and 0.5% below –2.576. Attempting to look this up in the table we see that the value 2.576 does not occur exactly in the tables, but

P(Z ≥ +2.57) = P(Z ≤ –2.57) = 0.0051 and P(Z ≥ +2.58) = P(Z ≤ –2.58) = 0.0049

So the true value is somewhere between 2.57 and 2.58, it turns out to be exactly

P(Z ≥ +2.576) = P(Z ≤ –2.576) = 0.005

"In between" values would normally be determined by interpolation. Exact values can be obtained from various software packages, including SAS and EXCEL.

Note: On an exam, if a value does not occur exactly, I will accept either of the two limits on either side of the correct value, or anything in between.

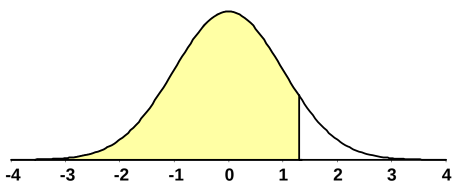
In the real world you can get "exact" values from EXCEL. In the even more real world, how much precision, or how many decimal places, do you really need to make this type of decision? All my tables were created in EXCEL

A few more examples of working with Z tables

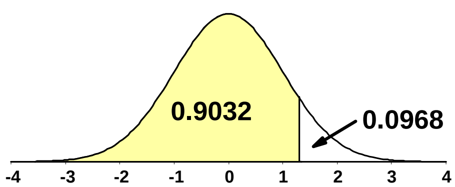
Find $P(Z \geq +1.35)$. This is an area in the upper half of the distribution (since Z is positive) so we can read it directly from the Z tables. $P(Z \geq +1.35) = 0.0885$

Find $P(Z \leq -2.22)$. This is an area from the lower half of the table, but due to symmetry $P(Z \leq -2.22) = P(Z \geq +2.22)$, so we can use the upper half of the table that we have available. $P(Z \leq -2.22) = 0.0132$

What about problems that do not ask for the area in the upper or lower tail? For example, $P(Z \leq 1.30)$. This value is in the upper half of the table, but the probability requested is for randomly chosen Z values **less than or equal**, this will go into the lower half of the distribution!



To solve this problem you must recall that the total area under the curve adds to 1. To find $P(Z \leq 1.30)$, we first find $P(Z \geq +1.30)$ and subtract from 1. $P(Z \leq 1.30) = 1 - P(Z \geq +1.30) = 1 - 0.0968 = 0.9032$.

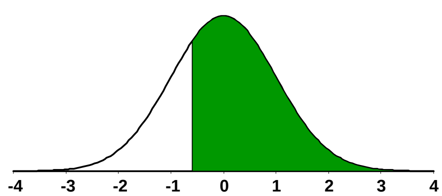


Even trickier Z distribution problems

Note that the value of Z = 0.00 has a probability of 0.5, so half of the distribution is above this value (and half below).

Find $P(Z \geq -0.65)$.

Now we are looking for a value greater than or equal to a value on the negative side of the distribution. From our tables we first find $P(Z \geq 0.65) = 0.2578 = P(Z \leq -0.65)$ due to symmetry, and so $1-P(Z \leq -0.65) = 1-0.2578 = 0.7422$

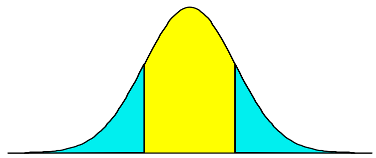


It is strongly advisable to sketch the problem, and to see if the answer makes sense. In this case we can see from the sketch that the desired area is over half of the total area, so the answer should be greater than 0.5, and of course it was $(P(Z) \geq -0.65) = 0.7422$).

A few extra examples

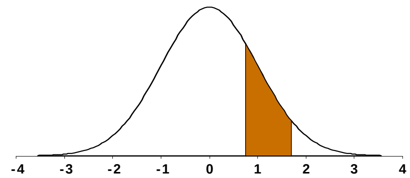
1) $P(Z \geq 3.50) = ?$	Read directly from the table
2) $P(Z \leq -2.00) = ?$	Read from the table, but for the upper (positive) end
3) $P(Z \geq 0.00) = ?$	Read directly from the table
4) $P(Z \leq 1.64) = P(Z \geq -1.64) = ?$	This is not in the table. Use $1-P(Z \geq 1.64)$
5) $P(Z \leq 1.96) = P(Z \geq -1.96) = ?$	This is not in the table. Use $1-P(Z \geq 1.96)$

Two-tailed problems and "area in the middle" problems



A common type of problem is to determine the area between two limits, or to determine the area in the tails outside some specific limits.

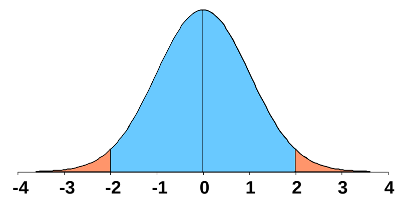
Probability expressions for these problems will take the form $P(Z_1 \le Z \le Z_2) = ?$



If the problem is symmetric, then $-Z_1 = Z_2$, call the limit Z_0 and we can rewrite $P(-Z_0 \le Z \le Z_0)$ as $P(|Z| \le Z_0)$. Probability expressions for areas in the tails will take the form $P(|Z| \ge Z_0)$ if the problem is symmetric. If not, we can write this as two expressions, $P(Z \le Z_1)$ OR $P(Z \ge Z_2)$.

It is also possible that problems involving sections may not be symmetric, and may occur entirely in the positive tail, or negative tail.

Some Examples

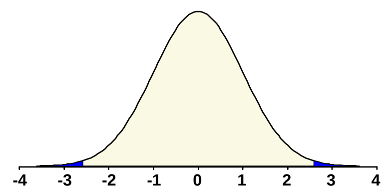


$P(|Z| \ge Z_0)$, where $Z_0 = 1.96$. Since we are taking the absolute value of a randomly chosen Z value in either tail, that value may be either positive or negative and its absolute value may be greater than or equal to 1.96.

$P(|Z| \ge Z_0)$, where $Z_0 = 1.96$.

$P(|Z| \ge Z_0) = P(Z \le -Z_0) + P(Z \ge Z_0)$, and since it is symmetric

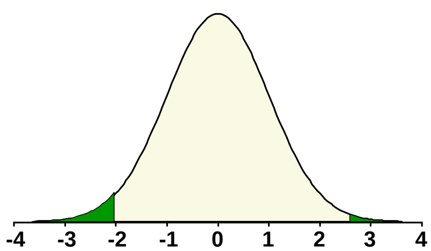
$P(|Z| \ge Z_0) = 2*P(Z \ge Z_0) = 2(0.0250) = 0.050$



$P(|Z| \le Z_0)$, where $Z_0 = 2.576$. This problem is similar to the previous, but it describes the area in the middle, between two limits.

$P(|Z| \le Z_0) = 1 - P(|Z| \ge Z_0) = 1-2*P(Z \ge Z_0) = 1 - 2(0.0050) = 0.99$

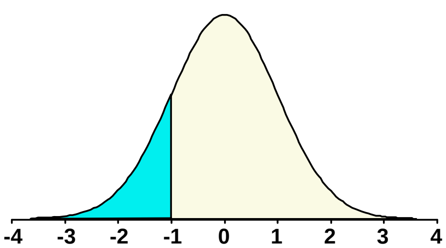
An asymmetric case



$P(-1.96 \le Z \le 2.576) = ?$ This is the area in the middle, the total minus the two tails. We already know these tails.

$P(-1.96 \le Z \le 2.576) = 1 - P(Z \ge 1.96) - P(Z \ge 2.576) = 1 - 0.0250 - 0.0050 = 0.97$

Working the Z tables, backward & forward

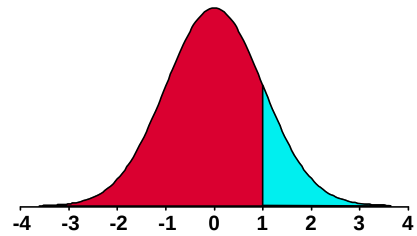


We have seen how to find a probability from a value of Z_0. Now we need to be able to find a value of Z_0 when a probability is known. Basically, we find the value of the probability in the body of our Z table, and determine the corresponding value of Z_0.

$P(Z \le Z_0) = 0.1587$, find the value of Z_0

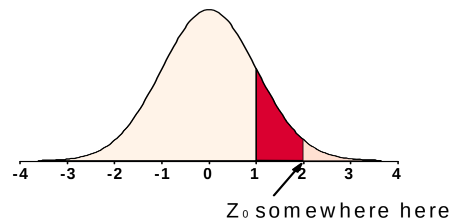
This probability is a value less than 0.5, so it is a tail and can be solved directly from our tables. We only need to find 0.1587 in the table and determine the corresponding value of Z_0. The value in the table occurs in the row corresponding to "1.0" and the column corresponding to "0.00". Finally note that the randomly chosen Z was to be **less than or equal** to Z_0, so we are in the lower tail. $Z_0 = -1.00$

$P(Z \leq Z_0) = 0.8413$, find the value of Z_0

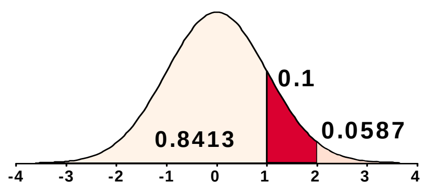


This probability is a value greater than 0.5. To read it from our tables we must determine the corresponding tail. The tail would be given by 1–0.8413 = 0.1587. So this is the same as the value we just looked up, it occurs in the row corresponding to "1.0" and the column corresponding to "0.00". Finally, note that the randomly chosen Z was to be **less than or equal** to Z_0, but since the probability was larger than 0.5 we are in the upper tail. $Z_0 = +1.00$

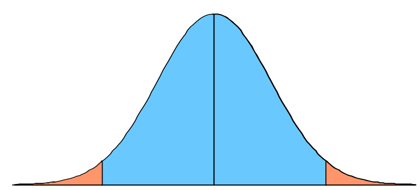
$P(1 \leq Z \leq Z_0) = 0.1$, find the value of Z_0



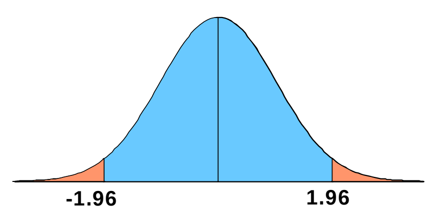
First note that the lower limit is 1.00, so we are working with the upper tail of the Z distribution. We can find the probability that Z is ≤ 1 as $1 - P(Z \geq 1) = 1 - 0.1587 = 0.8413$. The area of interest is stated to have an area of 0.1, so the upper tail would be what remains in the upper tail, 1 – 0.8413 – 0.1 = 0.0587. We find 0.0587 in our Z tables and note that it is between 1.56 and 1.57. I would accept either answer. The actual value, determined by the EXCEL "NORMDIST" function is Z_0 = 1.565781531.



$P(|Z| \geq Z_0) = 0.05$, find the value of Z_0

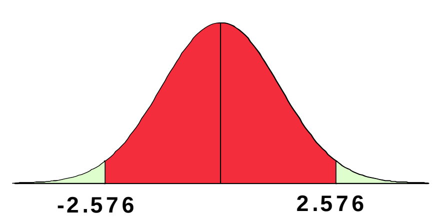


Note that the random value of Z is greater than Z_0, so we are examining an area in the tails, where the random Z may be either positive or negative and its absolute value will exceed the value of Z_0. Recall that our tables give only one tail, and this particular case has 0.05 in **two tails**, so we want to determine the value of Z_0 for only one of those tails to match our tables. Each tail would have half of the 0.05 since it is a symmetric problem. So the area in the tail is 0.05/2 = 0.025. We can examine the tables and determine that for a probability of 0.025 the Z value is 1.96, so $Z_0 = 1.96$. Some people like to write $Z_0 = \pm 1.96$, but this isn't really necessary for this case.



$P(|Z| \leq Z_0) = 0.99$, find the value of Z_0

Note that the random value of Z is **less** than Z_0, so we are examining an area in the middle of the distribution. Again, the randomly chosen Z may be either positive or negative. Since our tables give the tails of the distribution we need to determine the area in the tails. Since 0.99 occurs in the middle, the tails must have an area of 1 – 0.99 = 0.01. Half of this is in each tail, so the area in one tail would be 0.01/2 = 0.005. Given that the area in the tail is 0.005, the Z value is 2.576 (actually between 2.57 and 2.58; this is a value we have seen previously).



The Z transformation of any normal distribution

Most practical applications of the Z distribution will require that we take a real distribution $\{N(\mu,\sigma^2)\}$ and convert it to a Z distribution $\{N(0,1)\}$, calculate some probability statement, and then convert the results back to the original distribution.

So we work with two distributions

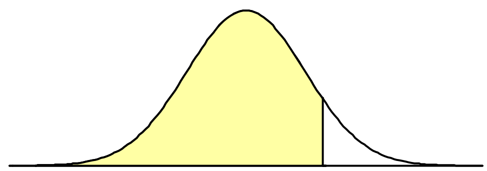
$Y \sim N(\mu,\sigma^2)$

$Z \sim N(0,1)$

In order to convert the observed Y distribution to the more workable Z distribution we need to transform the distribution. We use the Z transformation.

$Z_i = (Y_i - \mu) / \sigma$

For example, suppose we have a distribution where $Y \sim N(20,16)$ and we wish to determine $P(Y \le 24)$. We convert the probability statement for the original distribution into a Z distribution using our transformation.



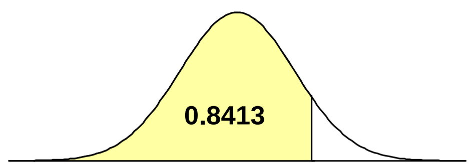
Transformation: $Z_i = (Y_i - \mu) / \sigma$ for N(20, 16)

$$P(Y \le 24) = P\left({(Y_i - \mu)}/{\sigma} \le {(24-20)}/{4} \right)$$

$= P(Z \le 1)$

$P(Z \le 1) = 1 - P(Z \ge 1) = 1 - 0.1587 = 0.8413$

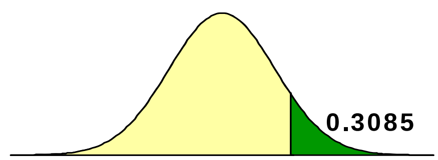
So, $P(Y \le 24) = 0.8413$



Another example

Using the same distribution $Y \sim N(20,16)$ where we wish to determine $P(Y \ge 22)$.

$$P(Y \ge 22) = P\left({(Y_i - \mu)}/{\sigma} \le {(22-20)}/{4} \right) = P(Z \ge 0.5)$$



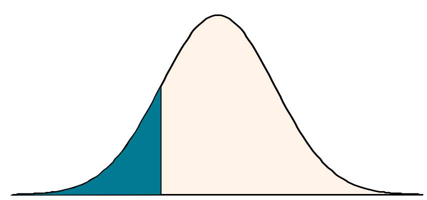
From the table we determine that $P(Z \ge 0.5) = 0.3085$

Another type of example

Find Y_0, where $P(Y \le Y_0) = 0.1515$ for the same distribution, $Y \sim N(20, 16)$. Again, using our transformation,

$$P(Y \le Y_0) = 0.1515 = P\left({(Y_i - \mu)}/{\sigma} \le {(Y_0 - \mu)}/{\sigma} \right) =$$

$P(Z \le Z_0) = 0.1515$

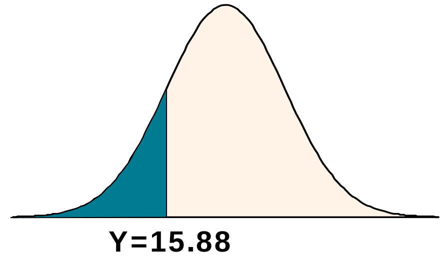


Notice that the sign is $\le$ and the probability small (less than 0.5), so we are in the lower half of the distribution. From the table the Z value is 1.03, so $Z_0 = -1.03$. So we know that $Z_0 = -1.03$, an area in the lower half of the distribution, but we don't know the value of Y_0 yet and our original problem was to find $P(Y \le Y_0) = 0.1515$. So we need to transform back to the Y scale. This back-transformation is a reversal of the Z transformation.

Transformation: $Z_i = (Y_i - \mu) / \sigma$

To transformation back calculate:

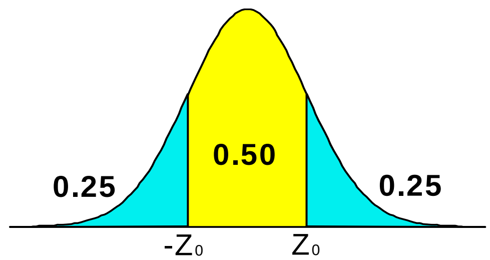
$Y_i = \mu + Z\sigma = 20 + (-1.03)4 = 20 + (-4.12) = 20 - 4.12 = 15.88$.



The interpretation of this problem is that there is a probability of 0.1515 of randomly drawing a value from this distribution and getting a value that is less than 15.88.

One last example

Determine within what limits 50% of the distribution lies. We will use the same hypothetical distribution Y~N(20,16). We will assume that the objective here is to form symmetric limits as shown below. Otherwise, the problem has no answer. After all, half of the distribution (50%) lies above $Z_0 = 0$, so the limits could be $(0, \infty)$. Half the distribution also lies below $Z_0 = 0$, so the limits could also be $(-\infty, 0)$. And there are an infinite number of other such limits in between. So, assuming symmetry, the problem becomes to determine the values of Y_0 such that half of the distribution lies between $-Z_1$ and $+Z_2$, where $Z_1 = Y_2$ due to symmetry and we will refer to this value as Y_0. Once we determine the values of Z that meet the probability limits we will transform the upper and lower value back to the Y distribution limits.



Determine:

$P(|Z| \le Z_0) = 0.5$

Since our tables give area in the tails, we need to change this from $P(|Z| \le Z_0) = 0.5$ to $1-2P(Z \ge Z_0) = 0.5$ and find $P(Z \ge Z_0) = (1-0.5)/2$ or $P(Z \ge Z_0) = 0.25$.

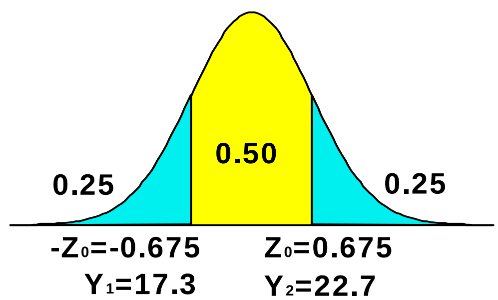
From our table: $Z_0 = 0.675$

Transforming $Z_0 = 0.675$ back to the Y scale gives,

$-Z_0$, $Y_1 = 20 - 0.675(4) = 20 - 2.7 = 17.3$

Z_0, $Y_2 = 20 + 0.675(4) = 20 + 2.7 = 22.7$

The final probability statement is best given in the form, $P(17.3 \le Y \le 22.7) = 0.50$



Summary on use of Z tables

Values of the relative cumulative frequency are given in the table.

the table is one – sided

the value given in the table is for the upper tail of the distribution

The total area under the curve is 1.0

The distribution is symmetrical.

e.g. for $Z = -1$ r.c.f. = 0.1587 (in the left tail)

for $Z = 1$ r.c.f. = 0.1587 (in the right tail)

We can transform from any normal distribution to the Z distribution by using the transformation $Z_i = (Y_i - \mu) / \sigma$

We can transform back from the Z distribution to any normal distribution by using the transformation $Y_i = \mu + Z\sigma$

Memorable values of Z (for 2 sided evaluations)

$\mu \pm 1\sigma = 0.68$

$\mu \pm 1.645\sigma = 0.90$

$\boldsymbol{\mu \pm 1.96\sigma = 0.95}$

$\mu \pm 2.576\sigma = 0.99$

A tip. Some tables give the one tailed probabilities at the top, some the two tailed probabilities. Some are cumulative from the lower end, some cumulative from the zero value (middle) up, and some give the cumulative area in the tails (like mine).

If you know that 1.96 leaves 2.5% in one tail and 5% in two tails you can look for this value and figure out what kind of tables you have.

Distribution of sample means

Sample means are the basis for testing hypotheses about μ, the most common types of hypothesis tests. Before discussing hypothesis tests we will need some additional information about the nature of sample means.

Imagine we are drawing samples from a population with the following characteristics.

Population size = N

Mean = μ

Variance = σ^2

The individual observations from this parent population are:

$Y_i = Y_1, Y_2, Y_3, \ldots, Y_N$

The set of samples of size n from the parent population form a derived population

There are N^n possible samples of size n that can be drawn from a population of size N (sampling with replacement, which simulates a very large population).

For each sample we calculate a mean $\bar{Y}_k = \sum_{i=1}^{n} Y_{ik} \Big/ n$, where k = 1, 2, 3, ... , N^n

The Derived Population of means of samples of size *n*

The sample size = n

Population size = N^n

Mean = $\mu_{\bar{Y}}$

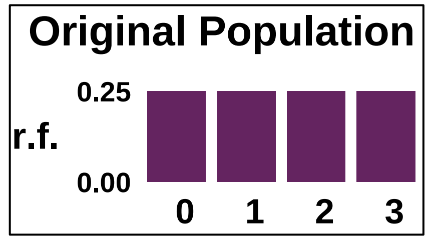
Variance = $\sigma_{\bar{Y}}^2$

Derived population values $\bar{Y}_k = \bar{Y}_1, \bar{Y}_2, \bar{Y}_3, \bar{Y}_4 \ldots, \bar{Y}_{N^n},$

Mean of the derived population $\mu_{\bar{Y}} = \sum_{k=1}^{N^n} \bar{Y}_k \Big/ N^n$ where k = 1, 2, 3, ... , N^n

Variance of the derived population $\sigma^2_{\bar{Y}} = \sum_{k=1}^{N^n} (\bar{Y}_k - \mu)^2 \Big/ N^n$ where k = 1, 2, 3, ... , N^n

Example of a Derived Population



Parent Population: Y_i = 0, 1, 2, 3

where $n = 2$ and $N^n = 4^2 = 16$

Draw all possible samples of size 2 from the Parent Population (sampling with replacement, so that values will occur more than once), and calculate a mean for each of the N^n samples.

For this discrete uniform distribution the Mean = (Max + Min)/2 = (3+0)/2 = 1.5 and the variance = ((Max–Min+1)^2–1)/12 = ((3–0+1)^2–1)/12 = 1.25, and the std. dev. = 1.1180.

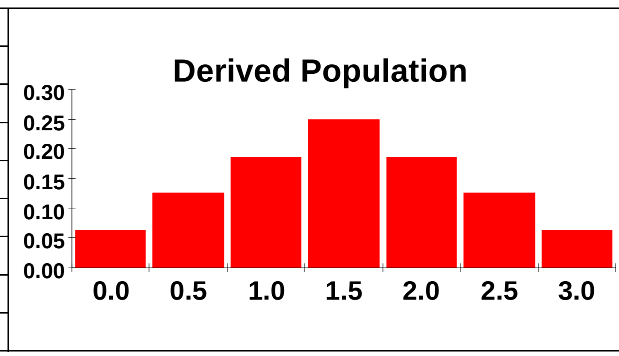
Sampling results for all possible means with replacement.

Sample	Mean
0, 0	0.0
0, 1	0.5
0, 2	1.0
0, 3	1.5
1, 0	0.5
1, 1	1.0
1, 2	1.5
1, 3	2.0
2, 0	1.0
2, 1	1.5
2, 2	2.0
2, 3	2.5
3, 0	1.5
3, 1	2.0
3, 2	2.5
3, 3	3.0

Deviation from 1.5	Squared Deviation
-1.5	2.25
-1	1
-0.5	0.25
0	0
-1	1
-0.5	0.25
0	0
0.5	0.25
-0.5	0.25
0	0
0.5	0.25
1	1
0	0
0.5	0.25
1	1
1.5	2.25

Distribution of the derived population of sample means

Means	Frequency	Relative Freq
0.0	1	0.0625
0.5	2	0.1250
1.0	3	0.1875
1.5	4	0.2500
2.0	3	0.1875
2.5	2	0.1250
3.0	1	0.0625
Sum =	16	1



$$\mu_{\bar{Y}} = \sum_{k=1}^{N^n} \bar{Y}_k \Big/ N^n = 24/16 = 1.5$$

$$\sigma_{\bar{Y}}^2 = \sum_{k=1}^{N^n}(\bar{Y}_k - \mu)^2 \Big/ N^n = \frac{\left(4(0.0)^2 + 6(\pm0.5)^2 + 4(\pm1.0)^2 + 2(\pm1.5)^2\right)}{16} =$$

$$\frac{\left(4(0.0) + 6(0.25) + 4(1.0) + 2(2.25)\right)}{16} = \frac{10}{16} = 0.625$$

$\sigma = 0.7906$

Note that the histogram of the derived population shows that the population is shaped more like the normal distribution than the original population.

Probability statement from the two distributions: Find P($1 \le Y \le 2$)

For the original, uniform population, P($1 \le Y \le 2$) = 0.5000

For the derived population, P($1 \le Y \le 2$) = 0.6250

THEOREM on the distribution of sample means

Given a population with mean μ and variance σ^2, if we draw all possible samples of size n (with replacement) from the population and calculate the mean, then the derived population of all possible sample means will have

Mean: $\mu_{\bar{Y}} = \mu$

Variance: $\sigma_{\bar{Y}}^2 = \sigma^2 / n$

Standard error: $\sigma_{\bar{Y}} = \sqrt{\sigma^2 / n} = \sigma / \sqrt{n}$

Notice that the variance and standard deviation of the mean have "n" in the denominator. As a result, the variance of the derived population becomes smaller as the sample size increases, regardless of the value of the population variance.

Central Limit Theorem

As the sample size (n) increases, the distribution of sample means of all possible samples, of a given size from a given population, approaches a normal distribution if the variance is finite. If the base distribution is normal, then the means are normal regardless of n.

Why is this important? (It is very important!)

If we are more interested in the MEANS (and therefore the distribution of the means) than the original distribution, then normality is a more reasonable assumption.

Often, perhaps even usually, we will be more interested in characteristics of the distribution, especially the mean, than in the distributions of the individuals. Since the mean is often the statistic of interest it is useful to know that it is possibly normally distributed regardless of the parent distribution.

NOTES on the distribution of sample means

Another property of sample means

as n increases, $\sigma_{\bar{Y}}^2$ and $\sigma_{\bar{Y}}$ decrease.

$\sigma_{\bar{Y}}^2 \le \sigma^2$ for any n