Multiplicative models and multiplicative errors exist, but are not covered in basic statistical methods. Note that the error term in this model is additive.

Other models we will discuss this semester include:

$Y_{ij} = \mu_i + \varepsilon_{ij}$ ..... for the two sample t-tests:

$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$ ..... another form of the t-test also used for ANOVA

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ ..... Simple Linear Regression

$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$ ..... Multiple Linear Regression

Coding and Transformations

Objective – Hypothesis testing Background

Many applications in statistics require modifying an existing distribution to an alternative form of the distribution. Hypothesis testing, in particular, requires taking an observed distribution and transforming to a recognized statistical distribution with known properties. This modification involves a transformation.

Theorems

If a constant "a" is added to each observation then, the mean of the data set will increase by "a" units the variance and standard deviation will remain unchanged

Example: Population of size N = 4

$Y_i = 2, 4, 6, 8$

$$\mu = \frac{\sum_{i=1}^{N} Y_i}{N} = \frac{20}{4} = 5$$

$$\sigma_Y^2 = \frac{\sum_{i=1}^{N} (Y_i - \mu)^2}{N} = \sum_{i=1}^{N} Y_i^2 - \frac{\left(\sum_{i=1}^{N} Y_i\right)^2}{N} \Bigg/ N = \frac{(120 - 100)}{4} = 5$$

$$\sigma_Y = 2.24$$

Now add 10 to each observation, the population size is still N = 4

$Y_i = 12, 14, 16, 18$

$$\mu = \frac{\sum_{i=1}^{N} Y_i}{N} = \frac{60}{4} = 15$$

$$\sigma_Y^2 = \sum_{i=1}^{N} Y_i^2 - \frac{\left(\sum_{i=1}^{N} Y_i\right)^2}{N} \Bigg/ N = \frac{(920 - 900)}{4} = 5$$

$$\sigma_Y = 2.24$$

The mean increased by a factor of 10 while the variance and standard deviation did not change.

NOTE that "a" may be either negative or positive, so we and add or subtract a constant from all values of Y. If we took the values of Y_i = 12, 14, 16, 18 and subtracted 10 from each value we would reverse the previous example.

When subtracting the mean is REDUCED by the value subtracted and the variance and standard deviation remain unchanged. The mean would then ten less and the variance and standard deviation would be unchanged

Another theorem

If each observation Y_i is multiplied by a constant "a" then, the mean of the data set is "a" times the old mean, the new variance is "a^2" times the old variance and the standard deviation is "a" times the old standard deviation.

Example: using the same Population as before; N = 4

Y_i = 2, 4, 6, 8,

$\mu = 5; \quad \sigma^2 = 5; \quad \sigma = 2.24$

let "a" be 10; so we multiply each observation by 10.

Y_i = 20, 40, 60, 80

$$\mu = \frac{\sum_{i=1}^{N} Y_i}{N} = \frac{200}{4} = 50$$, which is equal to aμ = 10(5) = 50

$$\sigma_Y^2 = \frac{\sum_{i=1}^{N} Y_i^2 - \frac{\left(\sum_{i=1}^{N} Y_i\right)^2}{N}}{N} = \frac{(12000 - 10000)}{4} = 500$$, which is $a^2\sigma^2 = 10^2(5) = 500$

σ = 22.4, which is 10(2.24) = 22.4 or $\sqrt{500} = 22.4$

NOTE that "a" may also be an inverse (i.e. 1/a instead of a), so we can multiply or divide all values of Y_i by any constant

if we took the values of Y=20, 40, 60, 80 and divided each Y_i by 10, we would reverse the previous example.

For division, the mean is divided by the value "a" (1/10), the variance divided by "a^2" (1/100), and the standard deviation divided by "a" (1/10)

The transformation operations may be used in combination.

Example: Population of size N = 3

Y = 10, 20, 30: μ=20; σ^2 =66.67; σ = 8.16

The transformation is "divide by 10 (or multiply by 1/10) and subtract 2"

Y_i = –1, 0, 1 (much easier to work with)

$$\mu' = \frac{\sum_{i=1}^{N} Y_i}{N} = \frac{0}{3} = 0 \quad \text{and} \quad \sigma'^2_Y = \frac{\sum_{i=1}^{N} Y_i^2 - \frac{\left(\sum_{i=1}^{N} Y_i\right)^2}{N}}{N} = \frac{(2-0)}{3} = 0.66667$$

$$\sigma' = 0.816$$

Note that order is important. To get back the original values we must reverse the transformation.

Above we (1) divided and then (2) subtracted.

To reverse this we must (1) add and then (2) multiply; $\mu = 10(\mu' + 2) = 10(2) = 20$

Since addition and subtraction do not affect measures of dispersion, we need consider only the division; $\sigma_Y^2 = a^2\sigma'^2_Y = 100(0.66667) = 66.667$

$$\sigma_Y = a\sigma'_Y = 10(0.816) = 8.16$$

Note that there is no addition or subtraction for the measures of dispersion since they were unaffected by the original transformation.

Other transformations

The logarithmic transformation was mentioned previously.

$$Y'_i = \log(Y_i)$$

if we calculate statistics such as the mean using the log transformed values, and then back-transforming or detransform with the antilog,

$$\text{anti}\log\left(\sum \log(Y'_i) \Big/ n\right) = e^{\left(\sum \log(Y_i) / n\right)} = GM(Y_i)$$

This results in a "geometric mean"

HOWEVER, note that we cannot take the logarithm of 0 (zero), so if there are zeros in the data set we must combine two transformations. One common modification is to add 1 to all observations.

$$Y'_i = \log(Y_i + 1)$$

Be careful in back-transforming or detransforming to subtract 1 after taking the anti-log to detransform. Order is important.

The same is true for inverses used in calculating a harmonic mean with an inverse transformation

$$Y'_i = 1 \Big/ Y_i$$

If we calculate the mean of the inverse transformed values, then detransform with the inverse to get the harmonic mean.

The "Z" transformation

The Z transformation consists of a combination of several of the previously discussed transformations.

$$Z_i = \frac{Y_i - \mu}{\sigma} \qquad \text{or} \quad t_i = \frac{Y_i - \bar{Y}}{S} \text{ for a sample.}$$

This transformation standardizes any normal distribution to a different normal distribution with a mean of zero and a variance of one (i.e. $\mu = 0$; $\sigma^2 = 1$; $\sigma = 1$). This is called the standard normal distribution. This is necessary, because otherwise there are an infinite number of different normal distributions with different means and variances. By transforming to a standard normal distribution we can learn to work with a single distribution with known characteristics.

Example: transform the data for a population of N = 4.

$Y_i = 2, 4, 6, 8$

Initially, calculate the mean and variance

$\mu = 5$; $\sigma^2 = 5$; $\sigma = 2.24$

and the transformation is applied with the following result.

$Z_i = (2–5)/2.24, (4–5)/2.24, (6–5)/2.24, (8–5)/2.24 = –1.34, –0.45, 0.45, 1.34$

$\mu = 0 / 4 = 0$

$\sigma^2 = 4 / 4 = 1$

$\sigma = \sqrt{1} = 1$

NOTE: addition and subtraction do not affect calculations of the variance and could be ignored.

The Z distribution

This is the first statistical distribution that we will use and develop for hypothesis testing. These hypothesis testing techniques will require an understanding of the distribution, of how to work with tables of probabilities of the distribution, and of the Z transformation. Fortunately, other statistical distributions will be similar. Once these techniques are learned, they apply readily to other statistical tests and applications.

The Z transformation

Purpose – transforms values from any normal population to the corresponding values from the Standard Normal Distribution. The distribution is N($\mu = 0$, $\sigma^2 = 1$).

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

where;

μ = the mean of the original population

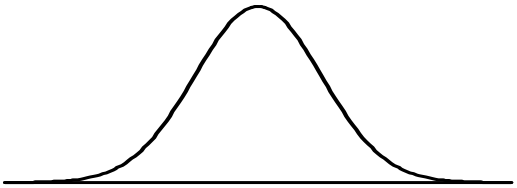
σ = the standard deviation of the original population

Y_i = the value of an observation from the original population

Z_i = the corresponding value from a Standard Normal Distribution

The purpose of this transformation is to deal with the infinite number of possible normal curves with different values of μ and σ by standardizing any normal curve so we can work with a single distribution. We will then work with these distributions from a table of Z values and probabilities related to the Z values. This will tie together much of what we have discussed (frequency and probability concepts, transformations, use of means, variances and standard deviations, and their calculations).

Probability statements



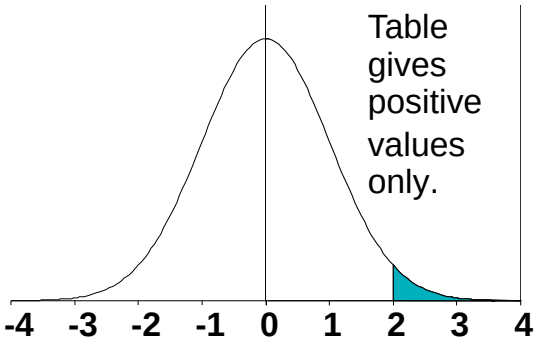
"Typical" probability statements are of the form.

$P[Z \leq Z_0]$ = r.c.f. at Z_0

where Z_0 is some hypothesized value

Tabulated Z distribution

We will need tables to work with the Z distribution. You should have those tables available. Your book has Z-tables, copies of my notes have the tables and I have a copy on the internet.



The Z table is exactly symmetric. As a result, the negative half (below zero) is a mirror image of the upper half. Therefore, our tables only need (and will only have) half of the distribution since it is exactly symmetric.

To work with these half tables, it is important to note that

$P[Z \leq 0] = P[Z \geq 0] = 0.5$ since half of the distribution is above 0 and half is below

$P[Z \leq -Z_0] = P[Z \geq +Z_0]$ since the table is symmetric

$P[Z \leq Z_0] = 1 - P[Z \geq Z_0]$ since the total area under the curve sums to one.

Z table

The table in the text is "one-sided" as only 1 side is required due to symmetry.

Values in the rows on the left side and top of the Z table give the value of Z, values in the body of the table are the probabilities of randomly choosing a larger Z value by random chance. For example, take Z = 0.11. What proportion of the distribution occurs above this value? Or, what is the probability of picking a Z value at random and it being larger than 0.11?

Only the first 6 rows and columns of the table are shown here, plus the row and column headings. Complete tables are available on the Internet linked to the departmental STATLAB web page.

	0.00	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**
0.00	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801
0.10	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404
0.20	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013
0.30	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632
0.40	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264
0.50	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912

Reading the Z tables

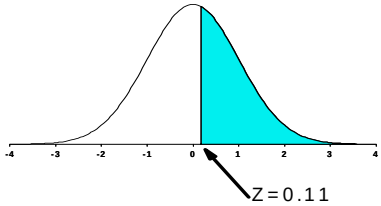
Values on the left side and top of the Z table give the value of Z, For example, to find Z=0.11,

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 $P(Z \geq +0.11) = 0.4562$.

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)



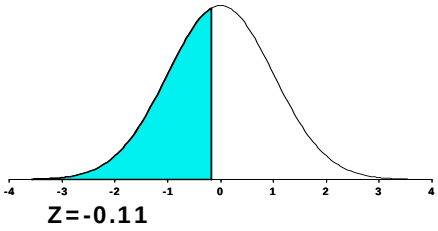
Working with Z tables

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 $P(Z \geq +0.11) = P(Z \leq -0.11)$



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

$P(Z \geq 0) = P(Z \leq 0) = 0.5$.

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 $P(Z \geq +1) = 0.1587$. Due to symmetry $P(Z \leq -1)$ 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 $P(Z \geq +1.96) = 0.0250$, and $P(Z \leq -1.96)$ would be the same.

A memorable value, 1.96!

The probability that a randomly selected Z falls between the limits $\mu - 2.576\sigma$ and $\mu + 2.576\sigma$ 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

$P(Z \geq +2.57) = P(Z \leq -2.57) = 0.0051$ and $P(Z \geq +2.58) = P(Z \leq -2.58) = 0.0049$

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

$P(Z \geq +2.576) = P(Z \leq -2.576) = 0.005$

"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.

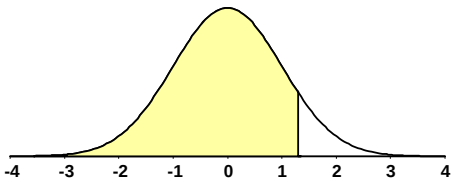
In the real world you can get "exact" values from EXCEL. In the even more real world, how much precision, or how many decimal places, do you really need to make this type of decision? All my tables were created in EXCEL

A few more examples of working with Z tables

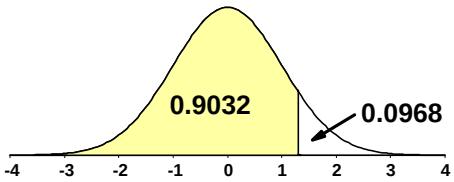
Find $P(Z \geq +1.35)$. This is an area in the upper half of the distribution (since Z is positive) so we can read it directly from the Z tables. $P(Z \geq +1.35) = 0.0885$

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. $P(Z \leq 1.30) = 1 - P(Z \geq +1.30) = 1 - 0.0968 = 0.9032$.

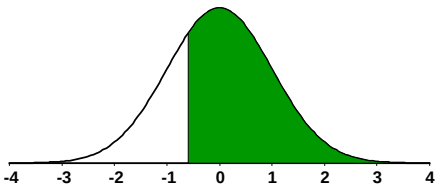


Even trickier Z distribution problems

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 $1 - P(Z \geq 1.96)$