6.7: Normal distribution and the normal deviate

Introduction

In Chapter 3.3 we introduced the normal distribution and the **Z score**, aka **normal deviate**, as part of a discussion about how some knowledge about characteristics of the dispersion of our data sampled from a population could be used to calculate how many samples we need (the empirical rule). We introduced Chebyshev's inequality as a general approach to this problem, where little is known about the distribution of the population, and contrasted it with the Z score, for cases where the distribution is known to be Gaussian or the normal distribution. The normal distribution is one of the most important distributions in classical statistics. All normal distributions are bell-shaped and symmetric about the mean. To describe a normal distribution only two parameters are needed: the population mean, μ, and the population standard deviation, σ. The normal distribution with mean equal to zero and standard deviation equal to one is called the **standard normal**, or **Z distribution**. With use of the Z score, any normal distribution can be quickly converted to the standard normal distribution.

Proportions of a Normal Distribution

This concept will become increasingly important for the many statistical tests we will learn over the next few weeks. What is the proportion of the populations that is greater than some specific value? Below, again, I have generated a large data set, now with population mean $\mu = 5$ and $\sigma = 2$. The red line corresponds to the equation of the normal curve using our values of $\mu = 5$ and $\sigma = 2$.

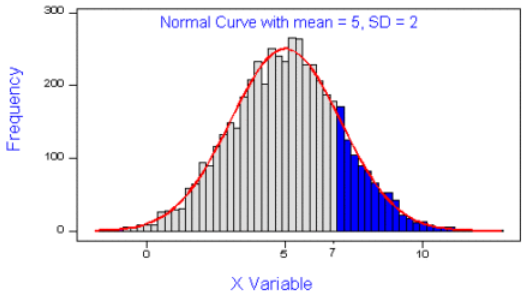

Figure 6.7.1: Frequency of observations expected to be greater than 7, from a large population with $\mu = 5$ and $\sigma = 2$.

Note that this is a crucial step! We assume that our sample distribution is really a sample from a population density (= "area under the curve") function (= "an equation") for a normal random (= "population") variable.

Once I (you) make this assumption, then we have powerful and easy to use tools at our command to answer questions like:

Question: What proportion of the population is greater than 7? (colored in blue).

This gets to the heart of the often-asked question, How many samples should I measure? If we know something about the mean and the variability, then we can predict how many samples will be of a particular kind. Let's solve the problem.

The Z score

We could use the formula for the normal curve (and a lot of repetitions), but fortunately, some folks have provided tables that short-cut this procedure. R and other programs also can find these numbers because the formulas are "built in" to the base packages. First, let's introduce a simple formula that lets us standardize our population numbers so that we can use established tables of probabilities for the normal distribution.

Below, we will see how to use `Rcmdr` for these kinds of problems.

However, it's one of the basic tasks in statistics that you should be able to do by hand. We'll use the Z score as a way to take advantage of known properties of the standard normal curve.

$$Z = \frac{(X_i - \mu)}{\sigma}$$

$Z = 1$ (with the mean = 0 and SD = 1). Z (say "Z-score") is called the normal deviate (aka "standard normal score"; it is also called the "Z-score"); it gives us a shortcut for finding the proportion of data greater than 7 in this case).

We use the normal deviate to do a couple of things; one use is to standardize a sample of observations so that they have a mean of zero and a standard deviation of one (the **Z distribution**). The data would then said to have been **normalized**.

The second use is to make predictions about how often a particular observation is likely to be encountered. As you can imagine, this last use is very helpful for designing an experiment — if we need to see a specified difference, we can conduct a pilot study (or refer to the literature) to determine a mean and level of variability for our observation of choice, plug these back into the normal equation and predict how likely we can expect to see a particular difference. In other words, this is one way to answer that question — how many observations need I make for my experiment to be valid?

Table of normal distribution

A portion of the table of the normal curve is provided at our web site and in your workbook. For our discussions, here's another copy to look at (Fig. 6.7.2).

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	
0.0	0.500000	0.496011	0.492022	0.488034	0.484047	0.480061	0.476078	0.472097	0.
0.1	0.460172	0.456205	0.452242	0.448283	0.444330	0.440382	0.436441	0.432505	0.
0.2	0.420740	0.416834	0.412936	0.409046	0.405165	0.401294	0.397432	0.393580	0.
0.3	0.382089	0.378281	0.374484	0.370700	0.366928	0.363169	0.359424	0.355691	0.
0.4	0.344578	0.340903	0.337243	0.333598	0.329969	0.326355	0.322758	0.319178	0.
0.5	0.308538	0.305026	0.301532	0.298056	0.294599	0.291160	0.287740	0.284339	0.
0.6	0.274253	0.270931	0.267629	0.264347	0.261086	0.257846	0.254627	0.251429	0.
0.7	0.241964	0.238852	0.235763	0.232695	0.229650	0.226627	0.223627	0.220650	0.
0.8	0.211855	0.208970	0.206108	0.203269	0.200454	0.197663	0.194895	0.192150	0.
0.9	0.184060	0.181411	0.178786	0.176186	0.173609	0.171056	0.168528	0.166023	0.
1.0	0.158655	0.156248	0.153864	0.151505	0.149170	0.146859	0.144572	0.142310	0.
1.1	0.135666	0.133500	0.131357	0.129238	0.127143	0.125072	0.123024	0.121000	0.
1.2	0.115070	0.113139	0.111232	0.109349	0.107488	0.105650	0.103835	0.102042	0.

Figure 6.7.2: Portion of the table of the normal distribution. Only values equal to or greater than $Z = 0$ are visible.

See Table 1 in the Appendix for a full version of the normal table.

We read values of Z from the first column and the first row. For $Z = 0.23$ we would scan the top row, scoot over to the fourth column, then trace to where the row and column intersect (Fig. 6.7.3); the frequency of occurrence of values at $Z = 0.23$ is 0.409046, or 40.9% (Fig. 6.7.3).

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	
0.0	0.500000	0.496011	0.492022	0.488034	0.484047	0.480061	0.476078	0.472097	0.
0.1	0.460172	0.456205	0.452242	0.448283	0.444330	0.440382	0.436441	0.432505	0.
0.2	0.420740	0.416834	0.412936	0.409046	0.405165	0.401294	0.397432	0.393580	0.
0.3	0.382089	0.378281	0.374484	0.370700	0.366928	0.363169	0.359424	0.355691	0.
0.4	0.344578	0.340903	0.337243	0.333598	0.329969	0.326355	0.322758	0.319178	0.
0.5	0.308538	0.305026	0.301532	0.298056	0.294599	0.291160	0.287740	0.284339	0.
0.6	0.274253	0.270931	0.267629	0.264347	0.261086	0.257846	0.254627	0.251429	0.
0.7	0.241964	0.238852	0.235763	0.232695	0.229650	0.226627	0.223627	0.220650	0.
0.8	0.211855	0.208970	0.206108	0.203269	0.200454	0.197663	0.194895	0.192150	0.
0.9	0.184060	0.181411	0.178786	0.176186	0.173609	0.171056	0.168528	0.166023	0.
1.0	0.158655	0.156248	0.153864	0.151505	0.149170	0.146859	0.144572	0.142310	0.
1.1	0.135666	0.133500	0.131357	0.129238	0.127143	0.125072	0.123024	0.121000	0.
1.2	0.115070	0.113139	0.111232	0.109349	0.107488	0.105650	0.103835	0.102042	0.

Figure 6.7.3: Highlighted $Z = 0.23$ in table, frequency is 0.409046.

Z on the standard normal table is going to range between -4 and $+4$, with $Z = 0$ corresponding to 0.500. The Normal table values are symmetrical about the mean of zero.

What to make of the values of Z, from $-4, -3, \ldots +2, +3$, up to $+4$ and beyond? These are the standard deviations! Recall that using the Z score you corrected to a mean of zero (got it!), and a standard deviation of one! $Z = 2$ is twice the standard deviation; a $Z = 3$ is therefore three times the standard deviation, and so forth. The distribution is symmetrical: you get the same frequency for negative as for positive values. So on the "X" axis on a standard normal distribution, we have units of standard deviation plus

(greater) or minus (less) than the mean. In Figure 6.7.4, the area under the curve representing less than -1 standard deviations is highlighted.

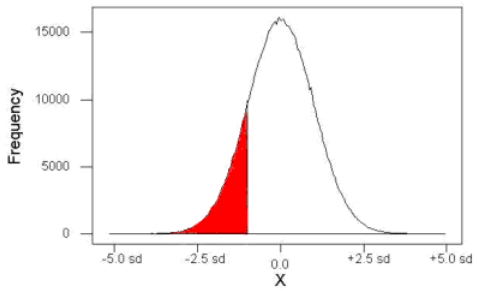

Figure 6.7.4: Plot of standard normal distribution; highlighted area under curve less than $X = -1\sigma$.

Question. How many multiples of standard deviations would you have for a Z score of $Z = 1.75$?

Answer = 1.75 times

Examples

See Table 1 in the Appendix for a full version of the normal table as you read this section.

What proportion of the data set will have values greater than 7? After applying our Z score equation, I get $Z = 1.0$, which translates to a frequency of 0.1587 or 15.87% that the observations are greater than 7.

What proportion of the data set will have values less than -7? After applying our Z score equation, I get $Z = -1.0$. Taking advantage of the symmetry argument, I just take my $Z = -1.0$ and make it positive — instead of values smaller than -1.0, we now have values greater than $+1.0$. And for $Z = 1.0$, 0.1587 or 15.87% of the observations are greater than 7, which means that 15.87% will be -7 or smaller.

What proportion of the data set will have values greater than 8? Again, apply the Z score equation. I get that for $Z = 1.5$, 0.0668 or 6.68% of the observations are greater than Z.

What proportion of the population is between 5 and 7? Draw the problem, as shown in Figure 6.7.5, where the subset of the population between 5 and 7 is colored red.

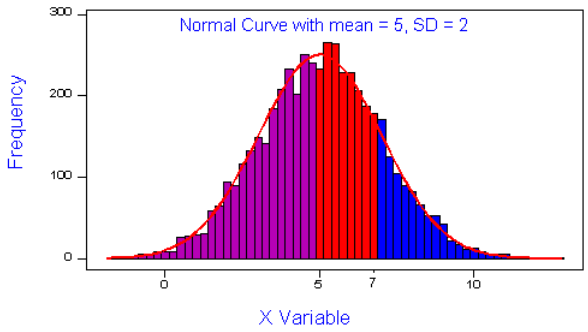

Figure 6.7.5: Proportion of the population between 5 and 7 is red (sorry about all of the colors — I kind of went crazy).

Worked problem

$1 - (\text{proportion beyond } 7) - (\text{proportion less than } 5)$

$1 - (0.1587) - (\text{proportion less than } 5)$

And the proportion less than 5?

Use the Z-score equation again. Now we find that $Z = 0$ and look up this Z-value in the table, which shows a 0.5 proportion or 50.0%.

Therefore, the proportion between 5 and 7 equals

$1 - 0.1587 - 0.50 = 0.3413$

Answer = 34.13% of the observations are between 5 and 7 when $\mu = 5$ and $\sigma = 2$.

Questions

1. Repeat the worked problem, but this time, find the proportion

- between 2 and 6.
- between 3 and 5.
- less than 5.
- greater than 7.