Handout 3: Normal Distribution
Reading Assignment: Chapter 4

Towards Normality
Recall that the Empirical Rule states for bell-shaped data about

- 68% of the values fall within one standard deviation of the mean in either direction
- 95% of the values fall within two standard deviations of the mean in either direction
- 99.7% of the values fall within three standard deviations of the mean in either direction

Data that is distributed in a symmetric, bell shape around the mean; this is called the normal distribution. Numerous continuous variables common in business have distributions that closely resemble the normal distribution. For datasets that are reasonably symmetric and bell-shaped, the normal distribution is a good representation.

The standard deviation is useful for measuring how far an individual value falls from the mean. For instance, the ACT is a standardized test with scores that are bell-shaped and suppose that you were told your score was 2 standard deviations above the mean. Without even knowing your score or the class average score, using the Empirical Rule you would know that only _______% of other people that took the ACT exceeded your score. Coincidently, this puts you in the _______ percentile.

The standardized score or z-score is a useful measure of the relative value of any observation in a dataset and the formula of these scores is simple.

\[ Z = \frac{\text{Observed Value} - \text{Mean}}{\text{Standard Deviation}} \]

The resulting z-score is simply the distance between the observed value and the mean, measured in terms of number of standard deviations (or: the number of standard deviations above or below the mean an observed value falls). Negative z-scores are below the mean and positive z-scores are above.

Suppose you and a friend take the ACT and SAT, respectively. You both feel that you are a better Mathematician than the other (a very interesting quarrel) and wish to settle the argument. You obtained a score of 27 on your Math section while your friend scores a 680 on theirs. Further suppose that the Math scores on the ACT are normally distributed with a mean ($\mu$) of 18 points and a standard deviation ($\sigma$) of 6 points and that Math scores on the SAT are also normally distributed with $\mu = 500$ and $\sigma = 100$. Who is the better Mathematician?

Note: We use Greek symbols to denote the mean (mu) and the standard deviation (sigma) because when we talk about a distribution of data in general, we consider that to be the entire population. Mu and sigma do not denote sample statistics. What would the units be for these z-scores?

Let $X_1 = \text{ACT Math score}$
Let $X_2 = \text{SAT Math score}$

$X_1 = 27$
$X_2 = 680$
Since we were told the scores followed the normal distribution, these Z-scores are also normally distributed and we can compare them to each other using bell curves.

In the ACT/SAT example, we let $X_1$ and $X_2$ represent the ACT and SAT, respectively. We assign these numerical values to outcome of a random circumstance that we call **random variables**. A random variable assigns a number to each unit in a population. There are two broad classes of random variables: a discrete random variable or a continuous random variable. Recall from the second handout that we said a discrete variable is one that can take one of a countable list of values (e.g. number of roommates living with you), while a continuous variable can take any value in an interval (e.g. height, blood pressure). In this handout, we will be discussing continuous random variables.

If we know that our random variable, $X$, is normally distributed with mean $\mu$ and standard deviation $\sigma$, the standardized value (z-score) will have a standard normal distribution with mean 0 and standard deviation 1. Normal distribution can only differ from each other in terms of center ($\mu$) and spread ($\sigma$), but the standardized values will have the same center and spread.

From the z-scores we can compute probabilities using the normal distribution. Recall from Math 1160 that a probability is a number between 0 and 1 and is written as either a fraction or a decimal. A probability provides information about how likely a particular outcome occurs by random circumstances over the long run.

**Finding Probabilities for z-Scores**

When we convert values for any normal random variable to z-scores, we can use our calculator or Excel to calculate the probabilities for different areas on the standard normal curve. For example, say that we wanted to know the probability that someone scored as good as or better than you on the Math section of the ACT.

We can turn the previous statement into a probabilistic statement (recall how to do so from MATH 1160) to look as follows:

Let $X = $ ACT Math score

$$P(X \geq 27) = ?$$

If $X$ is a normally distributed random variable (which it is), we can convert the left-hand side of the above probabilistic statement into a z-score

$$P(X \geq 27) = P\left(\frac{X-18}{6} \geq \frac{27-18}{6}\right) = P(Z \geq 1.5)$$

Based on our knowledge of the standard normal distribution, we can draw this probability
And finally, using our TI-84 we obtain the probability (area of the shaded region) as _____ . Or we can say that \( P(Z \geq 1.5) = \) _______.

In your calculator you will be using normalcdf( LOWER, UPPER):

- Press [2nd], then [VARS]
- Select normalcdf( in the DISTR tab and then press [ENTER]
- Type in ‘LOWER’, ‘UPPER’ (the lower and upper values of the shaded region) on the main screen
- Press [ ) ] to close the parenthesis and then press [ENTER]

What does this probability mean?

Let’s try this again, but with a more ‘business-y’ application.

Many manufacturing problems involve the accurate matching of machine parts, such as shafts, that fit into a valve hole. This design requires a shaft with a diameter of 22 mm, but shafts with diameters between 21.9 mm and 22.01 mm are acceptable. Suppose that a new process delivers shafts with diameters normally distributed with an average of 22.002 mm and a standard deviation of 0.0005 mm. What is the probability of an acceptable shaft?
iClicker Question
General Hospital’s patient account division has compiled data on the age of accounts receivables. The data collected indicate that the age of the accounts follows a normal distribution with mean of 28 days and standard deviation of 8 days. What proportion of the accounts are less than 30 days old?

(a) $-0.4013$
(b) $-0.5987$
(c) $0.4013$
(d) $0.5987$
(e) Still trying to compute

Finding Percentiles
Recall that the $k$th percentile refers to the value of a variable that has $k\%$ of the data values at or below it and $(100-k)\%$ of the data values at or above it. For instance, the median is the $50^{th}$ percentile because for any distribution, exactly $50\%$ of the data is below it and $50\%$ of the data is above. For a symmetric bell-shaped distribution (the standard normal distribution), the mean and median are equal. The value that corresponds to the $50^{th}$ percentile on the standard normal curve would be _____ (how many standard deviations above or below the mean would it be?). Draw the bell curve if it helps.

Let’s now go back to the ACT Math scores example. Suppose a university has the following requirement for its’ applicants: In order to be admitted applicants must be in the top 25% of ACT Math scores. In other words, we wish to know the value for ACT scores that would result in the _____th percentile.

First, we draw a picture (see above), marking where the percentile should fall. We will find this value initially in terms of standard units and then convert back to an ACT score. In probabilistic notation, we are looking for

\[ P(X <= a) = _____ \]
Recall that the Math scores on the ACT are normally distributed with $\mu = 18$ points and $\sigma = 6$ points.

In your calculator you will be using invNorm(PERCENTILE):
press [2nd], then [VARS]
Select invNorm( in the DISTR tab and then press [ENTER]
Type in ‘PERCENTILE’ (the percentile, in decimal form, that we are looking for) on the main screen
Press [ ) ] to close the parenthesis and then press [ENTER]

Remember that this is a standardized value; we need to un-standardize it. Using the z-score formula we can plug in what we know

$$0.674 = \frac{X - 18}{6}$$

and solve for $X$.

$$X =$$

What score should students obtain on the Math section of the ACT to be considered for this university?

Again, lets try this with a more ‘business-y’ application.

iClicker Question
General Hospital’s patient account division has compiled data on the age of accounts receivables. The data collected indicate that the age of the accounts follows a normal distribution with mean of 28 days and standard deviation of 8 days. What is the number of days in which 75 percent of all accounts are above?

(a) Still trying to compute
(b) 22.6
(c) 33.4
(d) $-22.6$
(e) $-33.4$
Assessing Normality
As discussed so far in this handout, many continuous variables used in business closely follow a normal distribution, but how do you determine whether a set of data can be approximated by the normal distribution? We will look at two approaches:

1. Compare the characteristics of the data with the theoretical properties of the normal distribution.
2. Construct a normal probability plot.

Compare Data Characteristics to Theoretical Properties
Recall that the normal distribution has some important properties:

- It is symmetrical; this implies that the mean and median are equal.
- It is bell shaped; this implies that the empirical rule applies.

Based on these properties, we can construct plots of the data to observe their appearance, compute descriptive statistics to compare the mean and the median, and evaluate how the data are distributed.

Let's look back at the data from the Emissions Report on Michigan's CO2 emissions from other energy sources. Below is a histogram and descriptive statistics of the data - we could also look at the box-and-whisker plot.

<table>
<thead>
<tr>
<th>CO2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>121,806</td>
</tr>
<tr>
<td>Standard Error</td>
<td>13,072</td>
</tr>
<tr>
<td>Median</td>
<td>94,661</td>
</tr>
<tr>
<td>Mode</td>
<td>0</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>118,371</td>
</tr>
<tr>
<td>Coefficient of Variation</td>
<td>97%</td>
</tr>
<tr>
<td>Range</td>
<td>389,001</td>
</tr>
<tr>
<td>Minimum</td>
<td>0</td>
</tr>
<tr>
<td>Maximum</td>
<td>389,001</td>
</tr>
<tr>
<td>Count</td>
<td>82</td>
</tr>
</tbody>
</table>

Descriptive Statistics of Michigan's CO2 Emissions (Metric Tons) for Other Energy Sources

Histogram of Michigan’s CO2 Emissions (Metric Tons) for Other Energy Sources
Based on this information, what is your conclusion about the normality of Michigan’s CO2 emissions for other energy sources?

Construct a Normal Probability Plot
A normal probability plot is a graphical approach for evaluating whether data are normally distributed. In this method, you plot the data against a theoretical normal distribution in such a way the points should form an approximate straight line. Departures from this straight line indicate departures from normality. We will let Excel handle this plot, shown below, while we focus on the interpretation.

The red line is the theoretical normal distribution while the blue line dots are the data values. We can see from the normal probability plot that Michigan’s CO2 emissions do not appear approximately normal because of several departures from a normal distribution. This agrees with what we saw in the histogram and when comparing the mean and median.

iClicker Question
The time required to complete an audit of a small company’s financial report is normally distributed, with mean of 30 hours and a standard deviation of 4 hours. What is the probability an auditor will complete the audit in a time between 15 and 26 hours?

(a) 0.1586  
(b) Still trying to compute  
(c) 0.8414  
(d) −0.1586  
(e) −0.8414