Math 316

Section 4.1 Introduction

First see **Online 4.1 Examples of non-normally distributed data**. Other distributions we've seen include **Figures 2.6.3**, **2.6.5** and **3.4.1**. Some distributions are "nicely shaped" (using the wording from page 65). Normal distribution: "Distribution" is "what are the outcomes and what portion/fraction of the time to they occur." "Normal" is synonymous with "common" or "typical."

Recall that histograms with many intervals (more possible with more data values) become "smooth," like a curve or function. See **Example 3.4.1** and **Figure 4.1.1**.

Section 4.2 The Normal Curves

Notation: $Y \sim N(\mu, \sigma)$ means random variable Y has a normal distribution with mean μ and standard deviation σ. There is actually a function with exactly these characteristics. See **Page 125. Figure 4.2.**2, then **Figure 4.2.1**. Recall the area under each of these curves is exactly 1.

Section 4.3 Areas under a Normal Curve

Recall that the area under a curve within a certain range is the probability of being within that range, and that we sometimes talk about what fraction of all of the data is within one (or two or …) standard deviation of the mean. We've seen this a few times already, including on **Page 64**. For normally distributed data, we can determine exactly what fraction of the data is less than a particular value or in a certain range of values.

So how do we compute this area and why do we care? An example to help answer both questions: we want to decide how good a 85 is in the following two sets of test scores. Both sets have a mean of $\mu = 80$.

Scores	How good is 85?	SD σ	How many SDs above 80 is 85?	That is:
$65, 70, 75, 80, 85, 90, 95$	Above average	10	$\frac{1}{2}$	$85 = 80 + \frac{1}{2}(10)$
$75, 78, 79, 80, 81, 82, 85$	High score!	2.93	1.71	$85 = 80 + 1.71(2.93)$

The number of standard deviations away from the mean a *particular* value of Y is the standardized value (or standard score or Z-value or Z-score) of Y:

$$Y = \mu + Z \cdot \sigma \quad \Rightarrow \quad Z = \frac{Y-\mu}{\sigma}$$

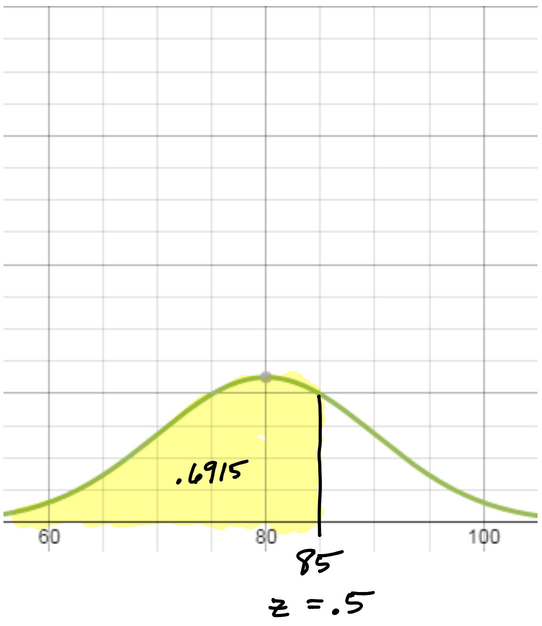
In the previous two scenarios we had standardized values for the score of 80 of:

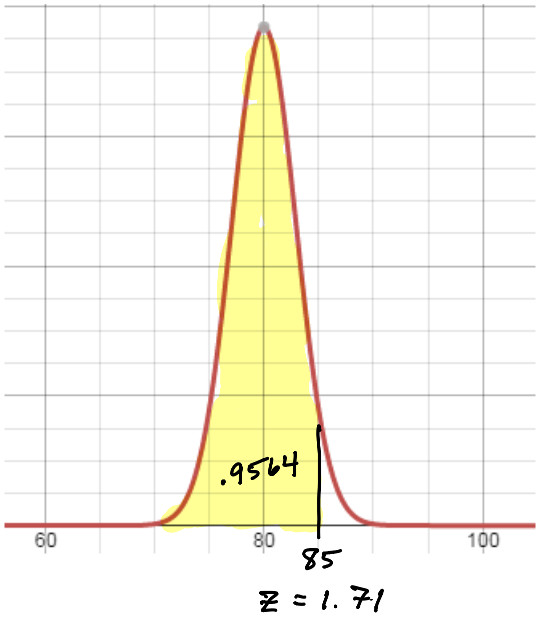
$$Z = \frac{Y-\mu}{\sigma} = \frac{85-80}{10} = \frac{1}{2} \qquad \text{That is, } 85 = 80 + \tfrac{1}{2}(10)$$

$$Z = \frac{Y-\mu}{\sigma} = \frac{85-10}{2.93} \approx 1.71 \qquad \text{That is, } 85 = 80 + 1.71(2.93)$$

In both cases the score of 85 is above the average of 80, but a score that is 1.71 standard deviations above average is more impressive than a score that is a mere 0.50 standard deviation above average. This will be a recurring theme: we will determine how extreme a particular value is by computing its standardized value, that is, *how many standard deviations above or below the mean* that particular value is.

Suppose we had two sets of many (say hundreds or even thousands, rather than just seven) normally distributed scores with mean 80 and standard deviations 10 and 2.93, as above. Suppose your own test score were 85. We can use **Table 3** (front of book) to find that the area under the standard normal curve below (to the left of) $z = 0.50$ is 0.6915 (or in Excel entering **= NORMSDIST(0.50)** returns the more precise value of 0.69146…). Similarly, the area below $z = 1.71$ is 0.9564. What do these numbers mean? They tell us that if there are a bunch of normally distributed scores with an average of 80 and standard deviation of 10, then a score of 85 (half a standard deviation above average) is better than about 69.15% of the other scores, and similarly if the average score is 80 and the standard deviation is 2.93, then a score of 85 (1.71 standard deviations above average) would be better than about 95.64% of the other scores. What does this look like?

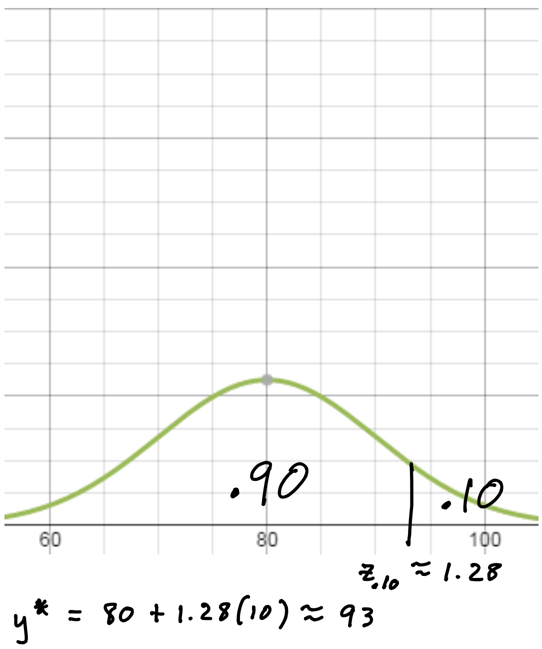


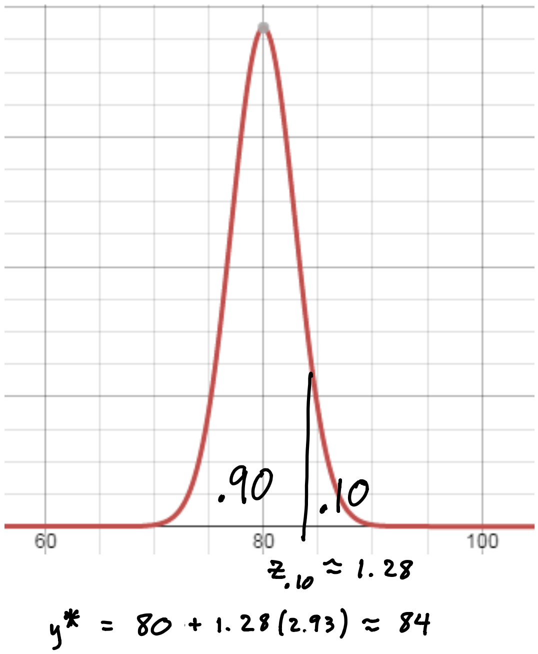


Here is a similar problem. Table 3 (or technology) again helps us.

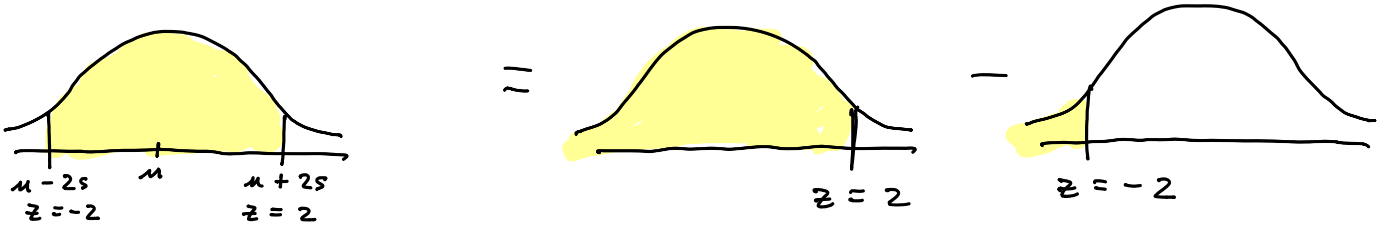
Problem: Suppose that there were a bunch of exam scores with an average of 80 and standard deviation of 10. What score would be at the 90th percentile? That is, what score would be higher than 90% of the other scores?

Solution: First, we need to find what Z-value has area of .90 below it. From **Table 3** we see that it is between $z = 1.28$ and $z = 1.29$. (In Excel entering **= NORMSINV(.9)** returns 1.281552….) I'll just use $z = 1.28$. So the score we are trying to find is one that is 1.28 standard deviations above average, that is, the score $80 + 1.28(10) \approx 93$. Similarly the 90th percentile score for the case with average 80 and standard deviation 2.93 would be $80 + 1.28(2.93) \approx 84$. What does this look like?





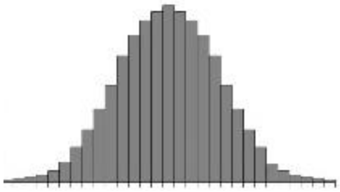
Figures 4.3.5, 4.3.6 and the text just below give us an idea of how much data is within 1 or 2 or 3 standard deviations of the mean. (Now you know where the **info on page 64** comes from.) For example, for 2 standard deviations:



Notation: $\boldsymbol{y^*}$ is the value we want to find (usually corresponding to some percentile) and $\boldsymbol{z_\alpha}$ is the Z-value which corresponds to an area α above that Z-value. See **Figure 4.3.11, Figure 4.3.12**. In the above example we found that $y^* = 93$, and since the area above a Z-value of 1.28 is .10, we could write $z_{.10} = 1.28$.

If time, let's work **HW 4.3.8b, HW 4.3.9af**.

Section 4.4 Assessing Normality

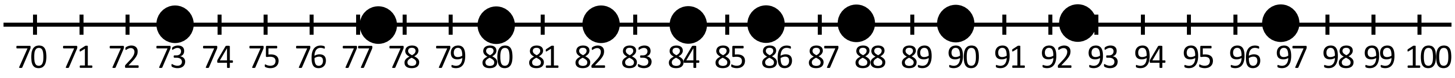


Recall the two main characteristics of a normal distribution: (1) more values/data "in the middle," i.e. near the mean; (2) symmetry/balance of values/data. Normally distrib-uted data are easier to work with than non-normally distributed data. So how can we check that data are normally distributed or not? Actually, it is not the case that data are either normally distributed or not—we actually want to determine *to what degree* the data are normally distributed. One idea is that we could create a histogram of the data and look at it, and then decide on normality. This is a bit of a pain and not very precise. See **Figure 4.4.2**. Is there a way to do this without a histogram?

Suppose that we had ten exam scores that were *perfectly* normally distributed. If we were to order the scores from lowest to highest, then we could find the standard scores (z scores) for those ten exam scores (fourth column). Suppose the mean were 85 and standard deviation were 7.3. Then we would have the scores listed in fifth column.

Which score	Range of scores	Percentile in middle	z-value	Corresponding exam score	Difference with 85
1	0 – 10%	5th %	-1.64	$85 - 1.64(7.3) \approx 73.0$	-12.0
2	10 – 20%	15th %	-1.04	$85 - 1.04(7.3) \approx 77.4$	-7.6
3	20 – 30%	25th %	$-.67$	$85 - 0.67(7.3) \approx 80.1$	-4.9
4	30 – 40%	35th %	$-.39$	$85 - 0.39(7.3) \approx 82.2$	-2.8
5	40 – 50%	45th %	$-.13$	$85 - 0.13(7.3) \approx 84.1$	-0.9
6	50 – 60%	55th %	$.13$	$85 + 0.13(7.3) \approx 85.9$	0.9
7	60 – 70%	65th %	$.39$	$85 + 0.39(7.3) \approx 87.8$	2.8
8	70 – 80%	75th %	$.67$	$85 + 0.67(7.3) \approx 89.9$	4.9
9	80 – 90%	85th %	1.04	$85 + 1.04(7.3) \approx 92.6$	7.6
10	90 - 100%	95th %	1.64	$85 + 1.64(7.3) \approx 97.0$	12.0

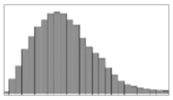
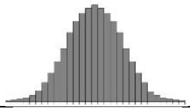
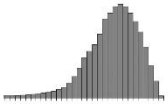
What these values look like on a number line:

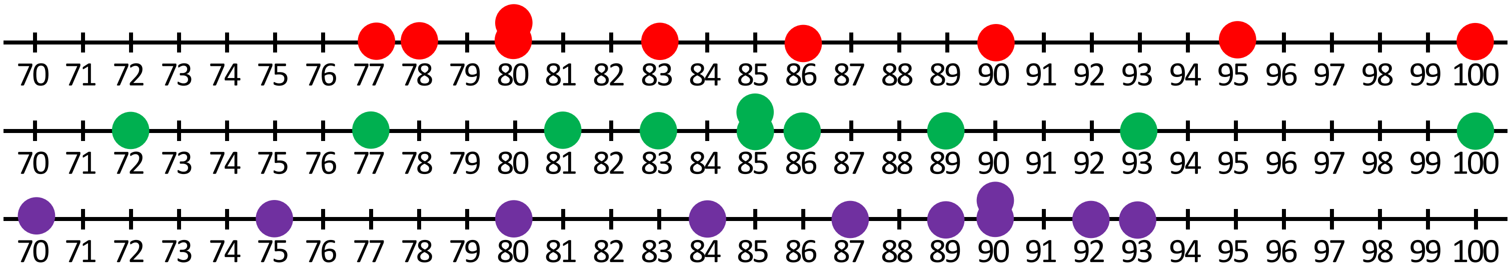


Compare this to **Figures 2.2.4/5**. *Let's go to the next page—what's there won't fit here. Then come back to Example 4.4.3 (below).*

Example 4.4.3 does a similar thing. In Table 4.4.1 they predict what the heights of eleven women should be if the data are perfectly normally distributed, then in Table 4.4.2 they compare those eleven expected heights with the actual heights in the data. They plot the actual heights vs. the expected heights (and Z scores). One could also plot Expected Z scores vs. Observed Z scores. If the data were perfectly normally distributed data, then the plotted points would be in a perfectly straight line. The more "straight-line-ish" the points, the more normally distributed the data are.

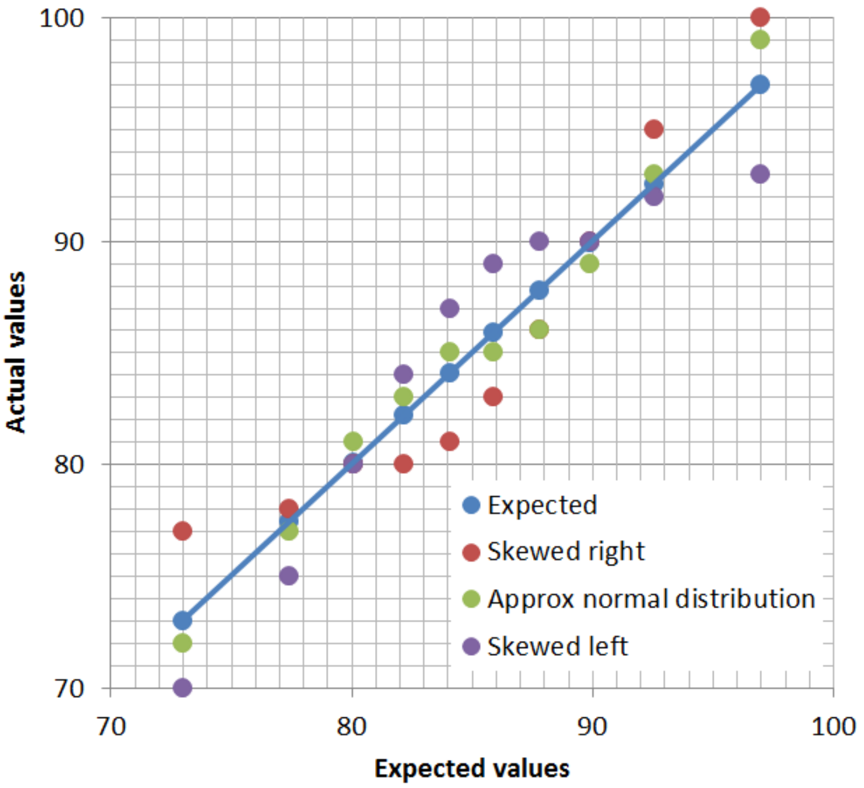
Let's look at the following three sets of values (for example, exam scores), *all of which have mean of 85 and standard deviation of approximately 7.3*. The **bolded values** are where the data seem more clustered, even though the mean is not exactly that value.

Type of data	Data	General shape of this type
Skewed right	77 78 **80** **80** 81 83 86 90 95 100	
Approximately normal	72 77 81 83 **85** **85** 86 89 93 99	
Skewed left	70 75 80 84 87 89 **90** **90** 92 93	



The normal probability plots. Recall: ten perfectly normal distributed values would be

73.0 77.4 80.1 82.2 **84.1** **85.9** 87.8 89.9 92.6 97.0



Also see **Figures 4.4.5 – 7**. In **Figures 5.2.6/7** (the next chapter) we see that even if the original data are normally distributed, a sample from that data is not necessarily perfectly normal. Now back to the bottom part of the front page of this handout.