MSPP PPOL 501–02 & -06: Fall 2014
Course Notes #7: Continuous Random Variables: The Special Case of the Normal Distribution
Professor Carolyn Hill

I. THE NORMAL DISTRIBUTION

[picture here]

-- It is a kind of "frequency polygon" – think of it as a kind of super-histogram (with very very thin bins) – to describe an empirical distribution of observations (X_i) in a data set

-- It is smooth (think thin bins that are nicely distributed), symmetric (if you fold it in half, you can see that it's exactly the same on each side), and unimodal (i.e., "bell-shaped")

-- Because it is symmetric and unimodal, the mean, median, and mode are all the same of a set of observations that are normally distributed

-- In theory at least, it extends infinitely in both directions

-- There will always be a constant (and known) proportion of cases (or area) between particular values in this distribution.

--- These values can be standardized in standard deviation units: i.e., they are stated as a certain number of standard deviations above or below the mean.

-- The *total area* under the curve is 100% of the area of the empirical distribution of observations. Alternatively, the total area is equal to 1 if we are describing proportions (which range from 0 to 1) instead of percentages (which range from 0 to 100).

* $\mu \pm \sigma$: 68.26% of the area under the curve
* $\mu \pm 2\sigma$: 95.44% of the area under the curve
* $\mu \pm 3\sigma$: 99.74% of the area under the curve

- The mathematical formula for the pdf of a normally-distributed random variable is:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\left(\frac{(x-\mu)^2}{2\sigma^2}\right)\right], \text{ for } -\infty < x < \infty$$

- And, we can standardize any normal distribution so that it will have a mean of zero and a variance of 1:

$Z = \frac{X_i - \overline{X}}{s}$ if we have a sample that we know is normally distributed, or

$Z = \frac{X_i - \mu}{\sigma}$ if we have a population that we know is normally distributed.

What is Z?! : It is a score that

(a) expresses information about the distance from a point in the empirical distribution to the mean, and

(b) expresses this distance in a standardized way: in standard deviation units (i.e., it divides the difference calculated in (a) by the amount of the standard deviation for that distribution)

(c) Rescales empirical distributions to have a mean = 0; and a standard deviation = 1.

To understand what's happening here, it's helpful to think of an observation that is equal to the mean…..

II. WORKING WITH THE NORMAL DISTRIBUTION

- A key point to remember is that the *total area* under the curve is 100% of the area of the empirical distribution of observations. Alternatively, the total area is equal to 1 if we are describing proportions (which range from 0 to 1) instead of percentages (which range from 0 to 100).

- Recall from above that there is always a constant area in relation to a point on the normal curve. What does this mean in practice? Many different uses…..

Example A: A political party has local branches with membership sizes that are distributed normally with $\bar{X}$ =400 and s=36.

Q: What is the probability that a single local branch chosen at random will have more than 500 members? Steps to answer the question…:

(i) How far away is 500 from the mean value? = (500-400) = 100 members

(ii) Because it was stated that membership sizes are distributed normally, we can go further and say something about how likely it is that a branch chosen at random will have more than 500 members *if* we can figure out whether 500 members is "close to" or "far from" the mean of 400.

(iii) We can do this by comparing 100 (= 500 - 400) to a "typical deviation"[1] or distance from the mean. That is, we can compare it to the standard deviation by dividing the deviation by the standard deviation: 100 / 36 = 2.78

(iv) Now we know how far away from the mean a local branch of 500 members is (2.78 standard deviation units away from the mean of 400), and we have this information per standard deviation, or in standard deviation units.

(v) If we know that observations are normally distributed, we saw under "The Basics" above that we can find out the percentage of area under the curve from a given point on the curve if these points are expressed in standard deviation units

(vi) Check Normal Curve table…

Read through table to get bearings…..

(vii) So, for a Z-value of 2.78, 0.0027 is the proportion under the normal curve to the right of this value. In other words, 0.27 percent of the local party branches will have membership sizes greater than 500.

Q: What percentage of local party branches has membership sizes less than 500?

Q: What percentage of local party branches has membership sizes of 400 to 500?

[1] Recall that the standard deviation is actually the square root of the variance; and you can think of the variance as a kind of *average squared deviation*.

Example B: Suppose verbal reasoning scores of 2nd grade students are distributed normally, with $\overline{X}$ =25 and s=2.5.

Q: What is the probability that a 2nd grader selected at random has a verbal reasoning score **less than 22.5?**

(i) Calculate the number of standard deviations away from the mean that 27.5 is:

$$Z = \frac{X_i - \overline{X}}{s} = \frac{22.5 - 25}{2.5} = \frac{-2.5}{2.5} = \quad -1$$

(ii) Now we have a Z-score that is *lower* than the mean. But because the normal distribution is symmetric, we can treat the Z value as an absolute value:

look up a Z-score of |-1.00| = 1.00. The proportion of the area under the curve beyond this point = 0.1587. Or, 15.87 percent of second graders have scores less than 22.5 on the verbal reasoning test.

(iii) This example also shows the value of stepping back from cranking through the formula and just taking a look at the distance and comparing it informally to the standard deviation. Why?

Example C: Suppose that the amount of money earned by all GPPI grads from speaking engagements is normally distributed with a mean of $4,040 and a standard deviation of $510. What percentage of GPPI grads earn between $4,000 and $4,500 on speaking engagements?

(i) The first thing to notice here is that the range of interest has a lower bound that is *less than* the mean value, and an upper bound that is *greater than* the mean value. This should raise a flag for you that you'll need to calculate the percentage in 2 steps:

(ii) $Z = \frac{\$4{,}000 - \$4{,}040}{\$510} = \frac{-40}{510} = -0.078$ $\qquad$ $Z = \frac{\$4{,}500 - \$4{,}040}{\$510} = \frac{460}{510} = 0.902$

The first Z-value is barely less than the mean, while the second is about 1 standard deviation away from the mean.

(iii) Look up these Z-values in the table, and get your bearings about which proportions you are looking for….

(iv) _____ percent of GPPI grads make between $4,000 and $4,500 on speaking engagements.

_____

Example D: Back to the GPPI example of speaking fees (mean = $4,040 and standard deviation=$510).

Q: What speaking fee has only 10 percent of all GPPI grads' speaking fees above it? (i.e., what is the lower bound fee for the top 10 percent of all fees)?

(i) Similar kind of question – reverse the process: the "10 percent" tells you the area under the curve (i.e., 0.1000 in the table). Now you need to find the Z-score associated with that area… To the Normal Table!

(ii) the proportion 0.1000 in the right-hand tail is associated with a Z-score of 1.28.

(iii) Now we need to just back out a specific score:

$$Z = \frac{X_i - 4{,}040}{510} = 1.28 \quad \rightarrow \quad X_i = \$4{,}693$$

(iv) 10 percent of the speaking fees of GPPI students are greater than $4,693.

III. NEXT UP:

The Central Limit Theorem