0. Business from last time

Notation. Recall that we only need two parameters to describe a Binomial, we write $B(n, \pi)$. So for example, using the IRS auditor example from the last lecture, we would write “$B(4,.58)$”.

Here is a different example, with $\pi=0.50$

This is a lattice of walks for balls falling from the top to the bottom row. Each time a ball hits one of the pegs, it can choose to turn right or left with probability $\pi$ and $1-\pi$ respectively. This process therefore gives rise to a (discrete) binomial distribution.

If the number of balls dropped (number of trials) is sufficiently large, the diagram formed by the piles of balls will approximate a normal distribution.

1. The Normal Distribution

We get the name “Normal Distribution” from a Belgian mathematician named Quetelet who worked for Napoleon and was measuring all sorts of things about large numbers of people (e.g. 5000). He measured their height, weight, eyesight, etc. and found a recurring pattern. The distribution of human traits had "bell shape" and he thought it as an ideal shape or “normal”.

Definition (page 132 of the text)

The equation for this curve is rather formidable and it is given in definition 4-5.C, specifically a continuous random variable $X$ has a NORMAL DISTRIBUTION if its pdf is

$$
\frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \text{ for } -\infty \leq x \leq +\infty
$$

Fact: If $X$ has a normal $N(\mu, \sigma^2)$ distribution, then $E(X) = \mu$ and $\text{Var}(X) = \sigma^2$. The normal also uses two fundamental constants, $\pi$ and $e$. Generally, normal distributions can take any range of values for the mean and they always have positive standard deviations. In practice, we work with one special type of Normal Distribution: the Standard Normal.
2. The Standard Normal Distribution

Used to approximate or describe histograms of many (but not every) types of data. The standard normal is more frequently applied to problems because it is useful. Its properties are:

- Symmetric, bell-shaped, the "bell curve", see the inside cover of your textbook.
- Mean (expected value, \( \mu \)) = 0, Standard Deviation (\( \sigma \)) = 1
- The median is where 50% (half) of the observations are on either side. In this distribution, the mean is equal to the median. The values on the horizontal axis are called "Z SCORES" or "STANDARD (DEVIATION) UNITS". Values of Z above the average are positive, values of Z below the average are negative.
- Area under the curve is equal to 100% when expressed as a percentage. The shaded area under the curve represents the percentages of the observations in your data between given values of Z.
- 68%-95%-almost 100% rule (see p. 133) About 68% fall within plus or minus 1 SD of the mean. About 95% fall within plus or minus 2 SD of the mean. Nearly 100% (99.7%) fall within plus or minus 3 SD
- The curve never crosses the horizontal axis, it gets very close at the extremes though. It extends to negative and positive infinity.

3. More on Standard (Deviation) Units or Z Scores

A score \( z \) is in STANDARD UNITS if tells how many Standard Deviations the original score is above or below the average. For example, if \( z=1.3 \), then the original score was 1.3 SD's above average; if \( z = -0.55 \), then the original score was 0.55 SD's BELOW average. The formula for transforming data from original values to Z scores is:

\[
Z = \frac{X - \mu}{\sigma}
\]

you can call this a "normal calculation".

Typically, we want to calculate the probability (the area under the normal curve) above or below some given value of Z. In other words, what is of interest is to answer questions like “find \( P(Z \geq 1.3) \)” or “find \( P(Z \leq -0.55) \)” or “find \( P(-0.55 \leq Z \leq 1.3) \)”

4. Examples of the use of Standard Units (see Table IV handout page 774 of the text or inside front cover of the textbook)

First step: Draw a picture...

(i) Draw the x-axis.
(ii) Draw a normal curve centered above it. The area under the curve gives probabilities.
(iii) Shade in the area in question.

Then perform your calculations. Example: For securities in the S & P 500, the average percentage return on equity over the last 52 weeks was 16.94% with a standard deviation of 48.82%. The average return on assets was 4.56% with a standard deviation of 8.21%. Both variables are approximately normally distributed. Here are some securities which make up the index:
Company | Percent ROE | Z score | Probability P(Z ≥ z) | Percent ROA | Z score | Probability P(Z ≥ z)  
--- | --- | --- | --- | --- | --- | ---  
Office Depot | 13.25% | -.08 | 1-.468=.532 | 6.33% | +.22 | .413  
AutoZone | 90.56% | +1.51 | .066 | 14.62% | +1.23 | .109  
RJR Tobacco | -70.55% | -1.79 | 1-.037=.963 | -29.62% | -4.16 | nearly 1.0  

Note that the table in your textbook only gives POSITIVE Z SCORES and THE CUMULATIVE PROBABILITY TO THE RIGHT ASSOCIATED WITH THOSE Z SCORES. Other tables work differently.

The point of using a table is to avoid using integration to figure out the areas under the curve. That's your alternative. You will receive a copy of the table with your remaining exams, you do not need to memorize it, you just need to know how to work with it.

4. Converting Standard Units back to original values

Idea: suppose you are told that the minimum acceptable ROE percentage change for a security under consideration for investment is Z= -1.15 or -1.15 standard deviations below the mean, what does that translate into as an actual percentage change?

\[ Z = \frac{X - \mu}{\sigma} \]

\[ -1.15 = \frac{X - 16.94}{48.22} \]

solving for X we get X=-72.39

5. Why bother with Standard Units?

Standard Units allow quick comparisons across variables with different units of measure. For example, suppose all the test scores in a class are normally distributed. The first test was worth 45 points, the mean was 33, the standard deviation was 5. A student received a 40 on that test and was told it was an A-, her Z score was

\[ Z = \frac{40 - 33}{5} = 1.40 \]

The next test had a mean of 58 and standard deviation of 8. If she were to do as well on the second test as she did on the first (that is get an A-), what would her new score need to be?

\[ Z = 1.40 = \frac{\text{new score} - 58}{8} \]

or she would need to score a 69.2 or something around a 69 to 70. The thing to remember is this: the second test has more points, a different mean, and a different standard deviation – it’s different, but if we convert the “raw” scores to standard scores (Z), comparisons are easy. In a way, it lets us compare apples and oranges.

6. More about using the standard normal curve (page 129)

If the data are normally distributed, then raw scores can be converted into standard units to find probabilities; also, percentages can be converted into standard units and then converted back into raw scores (original numbers).

Examples

a. GMAT scores are normally distributed with a mean of 500 and an SD of 100. Total scores range from 200 to 800. UCLA's Anderson school reported an average admission score of 698. Question: what is P(X ≥ 698)?

(first step: draw a picture ... want an area under the curve or probability)

(next step: convert to standard units; \( z = 1.98 \) so P(Z ≥ 1.98)
(final step: look up the probability; it's .024 from the table)

b. What GMAT score defines the top 10% of all test takers?
   (first step: draw a picture ... want an original score from an area of 10% at the top)
   (next step: using the area, find the standardized score: \( z = 1.28 \))
   (final step: convert back from standard units using formula; score is 628)

c. Stanford reports a class average score of 727. What is the percentile rank among all test takers of the typical Stanford MBA? (this is just like the first question, but report it as a percentage instead of a probability, so \( Z = 2.27 \) and that's 1.2%)

d. The Quinnipiac University school of business (I didn't make this up) reports a class average score of 460. What percentage of all GMAT test takers have a score higher than 460? What is the probability of randomly selecting a GMAT test taker who has a score between Stanford's average (727) and Quinnipiac's (460)?
   (first step: draw a picture ... want the area between 460 and 727)
   (next step: using the scores, find the standardized scores: \( Z = -0.40 \) and \( Z = 2.27 \))
   (final step: look up both probabilities \( P(Z \leq 2.27) = 1 - 0.012 = 0.988 \) and \( P(Z \leq -0.40) = 0.345 \) take the difference to get the final answer or \( P(-0.40 \leq Z \leq 2.27) = 0.643 \). You have a 64.3% chance of selecting a student in that range)

You should not attempt normal calculations on non-normal variables

7. Assessing Normality

A. Common sense: if the normal curve implies nonsense results (for example, that people have negative incomes, or that some women have a negative number of children), the normal curve doesn't apply and using the normal curve will give the wrong answer.

Example: in 2000, the mean annual electric bill in West Los Angeles was $733 with a Standard Deviation of $622. Does the normal curve apply? Try calculating how much a family pays annually for electricity if they are two standard deviations below the mean...

B. Construct a histogram: if the data look like a normal curve, the normal curve probably applies; otherwise, it does not. If you are interested, in Stata, you can issue the command:

\[ \text{histogram variablename, normal} \]

and Stata will draw a normal curve over your histogram so you can compare your histogram to a normal.

C. Do the data fall in a 68-95-99.7% pattern? If yes, normality is probably being met. You can examine any numeric variable in Stata by issuing the command:

\[ \text{pnorm variablename} \]

and the variable will be graphed against a line that represents the 68-95-99.7 rule. Deviations from the line suggest deviations from normality.