STA 4211 Lecture 3 Summary
Friday January 13, 2023

The topics were:

- The normal distribution; pp. 12 - 15 of the lecture notes,
- Homework 1: Problem 3 (application of z-scores), and a problem similar to Problem 5 (normal distribution in the linear regression model)

1 The normal distribution

The probability density function for the famous bell-shaped curve (or "normal distribution," or "Gaussian distribution") is given by:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

This is the function that is used to graph the bell-shaped curve, as illustrated for the standard normal curve (the one with $\mu = 0$ and $\sigma = 1$) in the notes p. 12. This is only one out of the infinite family of normal curves. You specify which curve by specifying μ and σ.

To be a pdf, the function $f(x)$ must satisfy two requirements:

- $f(x) \geq 0$, for all x, and
- The area under the whole curve must be equal to 1.

From the formula above for the pdf, we can tell quite a lot (but not everything) about the normal distribution. For instance, the parameter μ can take on any value on the real line; the parameter σ must be positive. From inspection of the formula, you see that μ can be either negative or positive (any value at all) without messing up the requirement that the pdf be non-negative, whereas if σ were less than zero, the pdf would become negative, which isn't allowed.

It is well-known that if μ changes value (while σ remains constant), the only change in the normal distribution is it shifts either left or right, so that the highest point of the curve is at the value of μ. That is, the whole curve is kept intact but moved left or right so its center is at μ. You can see this by inspection of the argument of the exponential function, the only place where μ occurs in the pdf

formula above. Also, the maximum value of the pdf occurs at $x = \mu$ and is given by $1/\sigma\sqrt{2\pi}$. So now let's sketch the pdf for $\mu = 0$ and $\sigma = .5$. (We did this in class on p. 12: The height of the pdf at 0 for $\sigma = .5$ is twice the height of the standard normal at 0, and the curve has to come down toward 0 more rapidly than the standard normal curve does, so it is more concentrated around 0, in order for the whole area under the curve to still be 1.)

Important facts (need calculus to derive these) are that the parameter μ is the mean of the distribution; the parameter σ is the standard deviation of the distribution. We will write $X \sim \mathcal{N}(\mu, \sigma)$ as shorthand for "the distribution of X is normal, with mean μ and SD σ."

1.1 Basic properties of the Normal Distribution

(p. 13)

1. Curve is symmetric, centered at the mean μ.
2. 50% of area lies to right of μ.
3. The SD σ measures the spread: The bigger the value of σ, the more spread out and flatter the curve.
4. Area under the curve is always 1.
5. (a) 68% of area is within σ of μ,
   (b) 95% of area is within 2σ of μ,
   (c) $\sim 99.7\%$ of area is within 3σ of μ.

See the graph depicting Property 5 on p. 13 of the notes. The first three properties can be seen from the pdf; to derive Properties 4 and 5 requires calculus. In this class, we will just state the calculus-based properties, and use them. To calculate probabilities involving the normal distribution, which involve areas under the pdf, we will use either normal tables or computer programs such as R.

1.2 Standard score

(p. 14)

Suppose X is an observation from a population with mean μ and SD σ. The standard score of X, usually denoted Z, is the number of SD's above $(+)$ or below $(-)$ the mean X is. Formula is

$$Z = \frac{X - \mu}{\sigma}.$$

The standard score indicates the relative standing of X in the population.

Calculating the standard score is a way of getting a common scale for different measurements which are approximately normally distributed.

Example. Suppose that a course has two midterms, for which the scores are approximately normally distributed, with means and SD's given below:

	Midterm 1	Midterm 2
Class Avg	55	50
Class SD	14	10
You get	76	67

On which test did you do better relative to the rest of the class? Answer: The standard (z) score for Midterm 1 is $\frac{76-55}{14} = 1.5$, and the z score for Midterm 2 is 1.7, so the answer is you did better relative to the whole class on Midterm 2.

If $X \sim N(\mu, \sigma)$, then the "probability density function" of X is given by the formula on p. 12. For most datasets in this course, we assume the normal distribution. We need the normal assumption to justify inference procedures based on the t and F distributions.

1.3 Useful facts

Suppose $X \sim N(\mu_X, \sigma_X)$, $Y \sim N(\mu_Y, \sigma_Y)$, and X and Y are independent random variables. Let a, b be any nonzero constants. Then:

- $aX \sim N(a\mu_X, |a|\sigma_X)$
- $aX + bY$ is normally distributed with mean: $E(aX + bY) = a\mu_X + b\mu_Y$,
  and variance: $a^2\sigma_X^2 + b^2\sigma_Y^2$

2 Some Homework 1 Problems

- Problem 3, on use of z-scores in diagnosing osteoporosis.

  Part A. A suggested way to start the write-up for this problem is the following:

  We need area to the left of -2.5 under the standard normal curve, which is found in the normal table to be (you fill in the blank here).

Note added after class: Then the student needs to interpret the numerical answer, for example, "so about XXX percent of healthy young adults have osteoporosis by the WHO criterion."

Part B. We discussed a way to approach the problem in class. In your write-up, you might want to use sketches of the two normal curves that are involved in the calculation, to supplement the written explanation you give.

- Problem like Problem 5

  Question:

  For data $(X_1, Y_1), \cdots, (X_n, Y_n)$, assume the simple linear regression model: $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, where the error terms $\epsilon_i, i = 1, \cdots, n$ are distributed independently of each other and have the normal distribution $\mathcal{N}(0, \sigma^2)$. Also assume we know $\beta_0 = 50$, $\beta_1 = 10$, and $\sigma = 8$.

  Consider the first two data pairs. Suppose we know $X_1 = 5$ and $X_2 = 6$, but we don't know the corresponding Y values. What is $P(Y_1 > Y_2)$?

  Solution *Reminder: You can get help from me, from the TA, and from other students on homework problems. You are responsible for writing up the problems yourself. The solution below gives a standard style for you to imitate.*

  We are asked to find $P(Y_1 > Y_2)$, which can be written $P(Y_1 - Y_2 > 0)$. So, we need to obtain the distribution of $Y_1 - Y_2$.

  First let's get the distributions of Y_1 and Y_2. Now $E(Y_1) = 50 + 10(5) = 100$, and $E(Y_2) = 50 + 10(6) = 110$. Also, $\sigma_{Y_1} = \sigma_{Y_2} = 8$. So we have $Y_1 \sim \mathcal{N}(100, \sigma = 8)$, and $Y_2 \sim \mathcal{N}(110, 8)$, from the linear regression model stated above.

  Next we get $E(Y_1 - Y_2) = 100 - 110 = -10$, and Var$(Y_1 - Y_2) = 2\sigma^2 = 128$. These follow from the "useful facts" we went over in class from the lecture notes p. 15.

  Now since a linear combination of normal random variables also has a normal distribution, we see that $Y_1 - Y_2 \sim \mathcal{N}(-10, 11.3137)$. So, $P(Y_1 - Y_2 > 0) = P(Z > (0 - (-10))/11.3137) = P(Z > .8839) = .1884$