COLLEGE OF CHARLESTON

Week 6

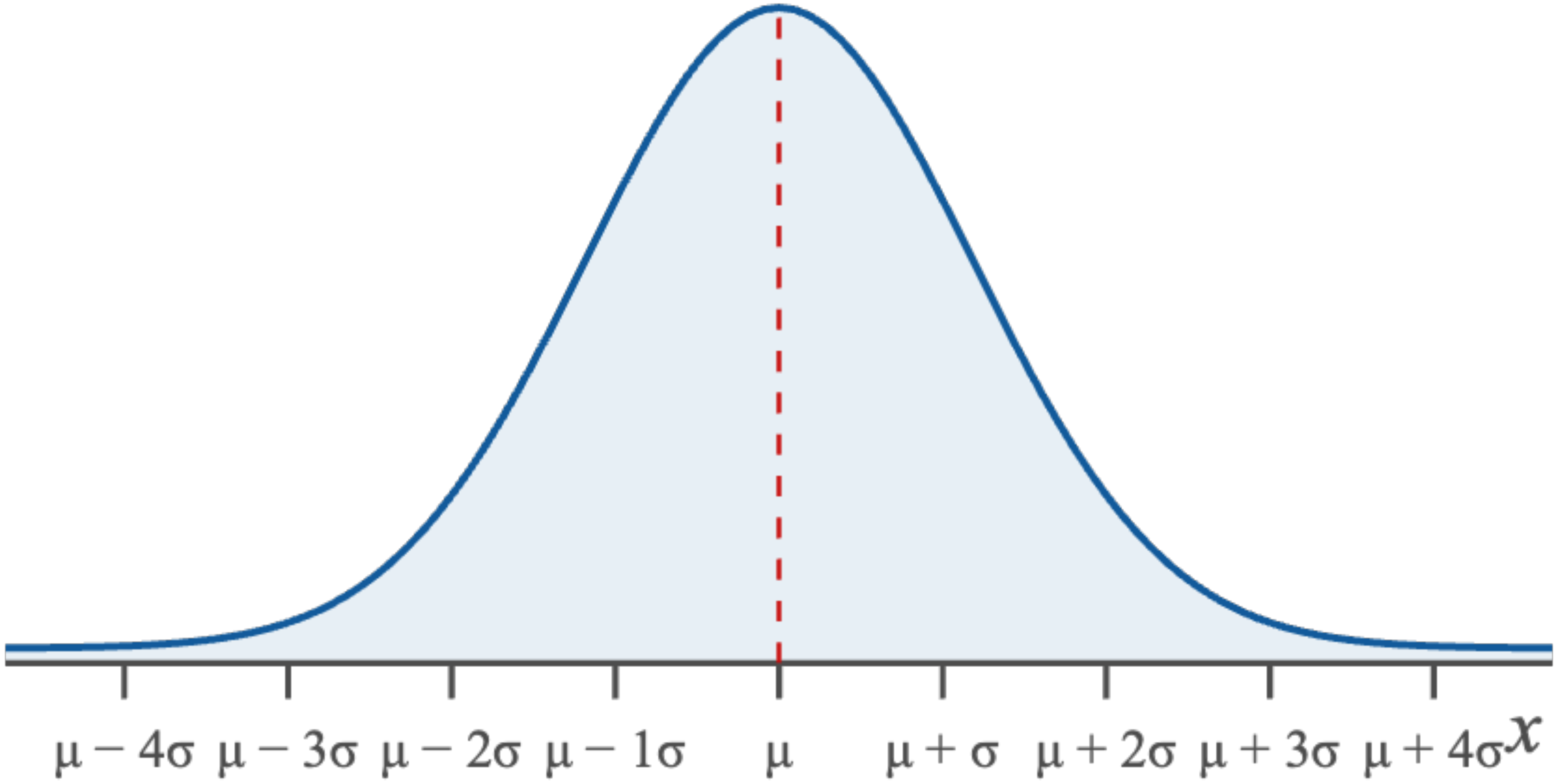
Math 104-03: Elementary Statistics

6.1 Introduction to the Normal Distribution

Normal Distributions

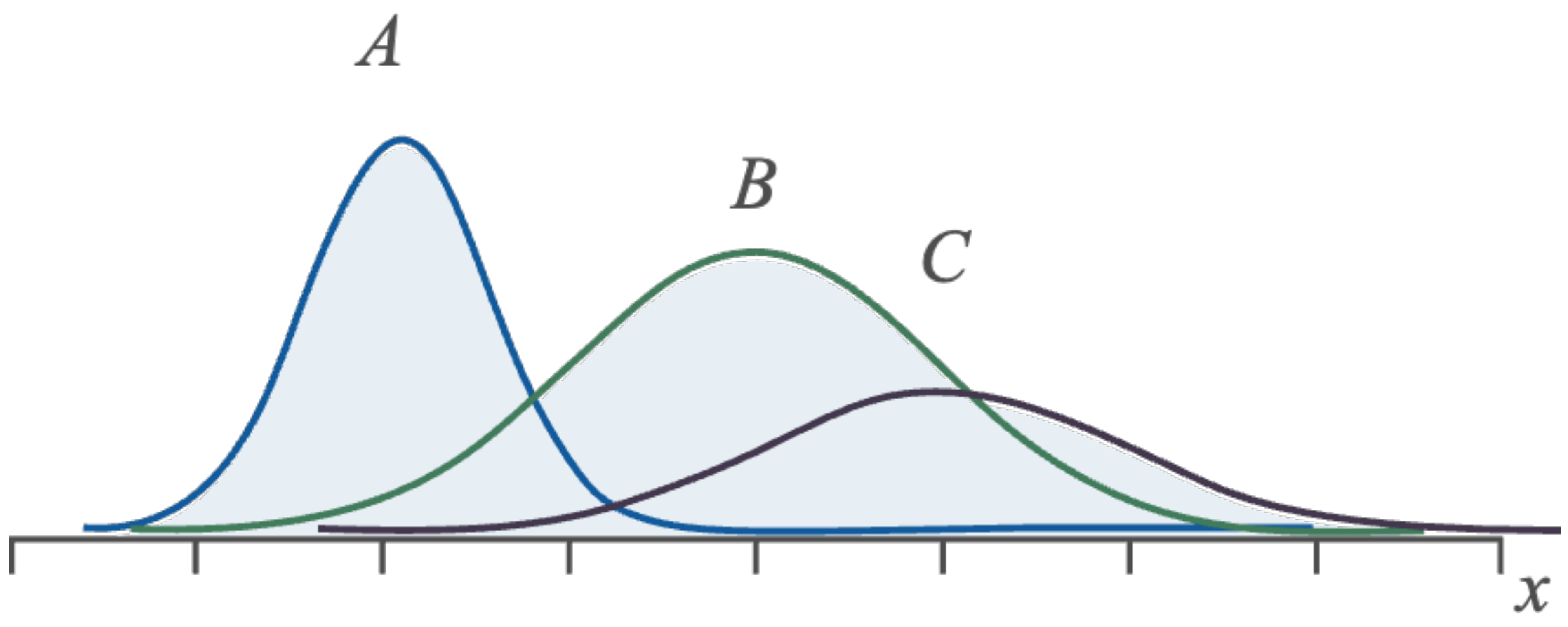
- A **continuous random variable** is a continuous variable whose numeric value is determined by the outcome of a probability experiment.
- A **normal distribution** is a probability distribution for a continuous random variable, X, defined completely by its mean and standard deviation, such that the following properties are true:
  - A normal distribution is bell-shaped and symmetric about its mean.
  - A normal distribution is completely defined by its
    - mean, μ, and
    - standard deviation, σ.
  - The total area under a normal distribution curve equals 1 (true for any probability distribution)
  - The x-axis is a horizontal asymptote for a normal distribution curve (thin tails)

Normal Distribution



Normal Distributions

- Example: use the following graphs of three normal distributions to answer the questions
  - Determine which distribution has the smallest standard deviation and which has the largest.
  - Determine which distribution has the smallest mean and which has the largest.
  - What does the position of curve B in relation to curve A tell you about the data of the distributions?



Normal Distributions

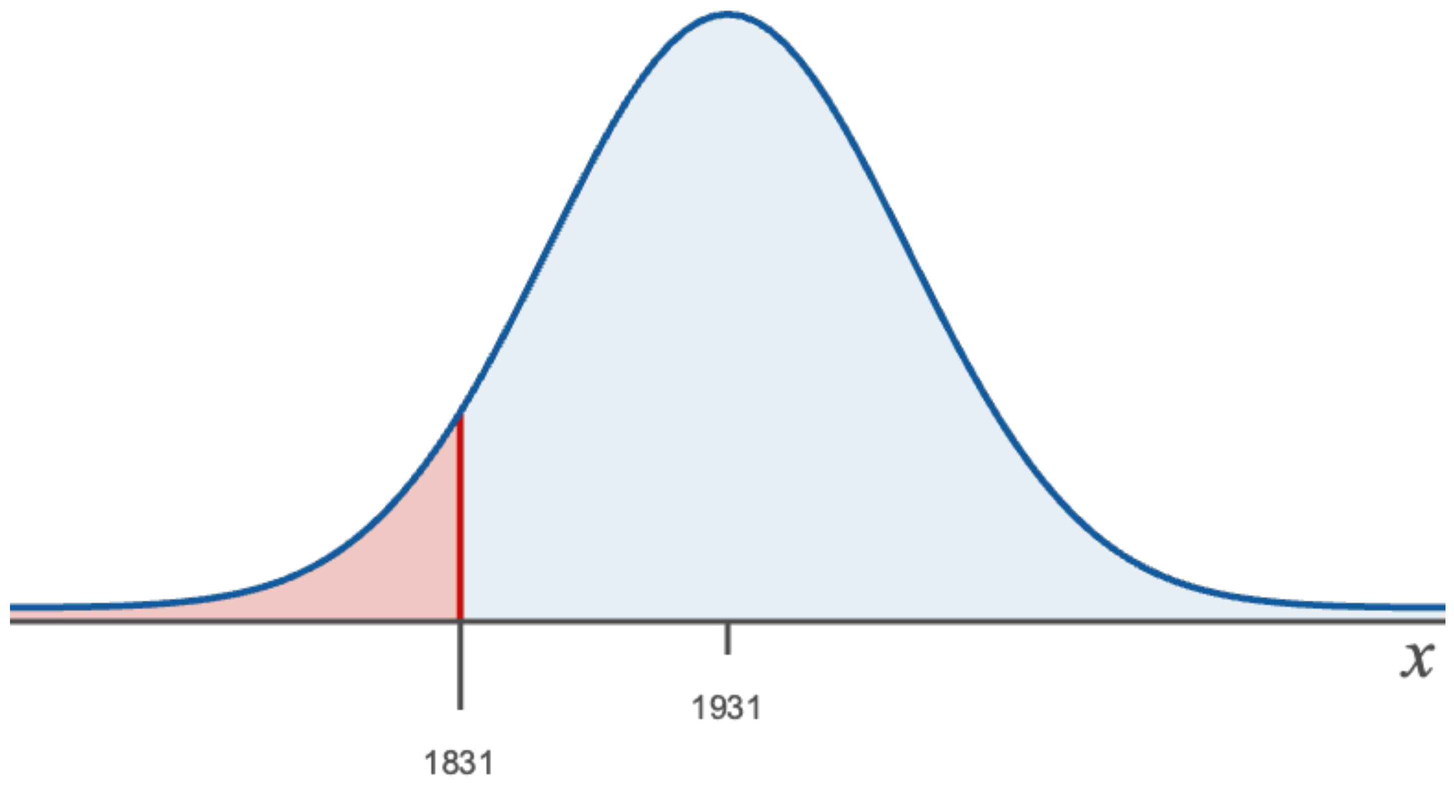
- Examples of normally distributed random variables:
  - Weight of apples in a large batch
  - Systolic/diastolic blood pressures of humans
  - Daily high temperature in the month of July in Charleston
  - Errors in a manufacturing process (e.g. diameter of chalk)
  - Household appliance lifespans
  - IQ scores (maybe?)

Normal Distributions

- More nonexamples:
  - Number of followers on a social media platform
  - Number infected in a pandemic
  - Housing prices in a city
  - Stock market returns
  - Wealth of individuals
  - Market capitalization of companies
  - Book sales
- Nassim Taleb: normality is the exception rather than the rule

Area Under a Normal Distribution

- Key: the area under the normal curve is equal to the probability of the random variable falling within that region.
  - The area to the *left* of a x-value is equal to $\mathbb{P}(X \leq x)$
  - The area to the *right* of a x-value is equal to $\mathbb{P}(X \geq x)$



X

6.2 Standard Normal Distribution

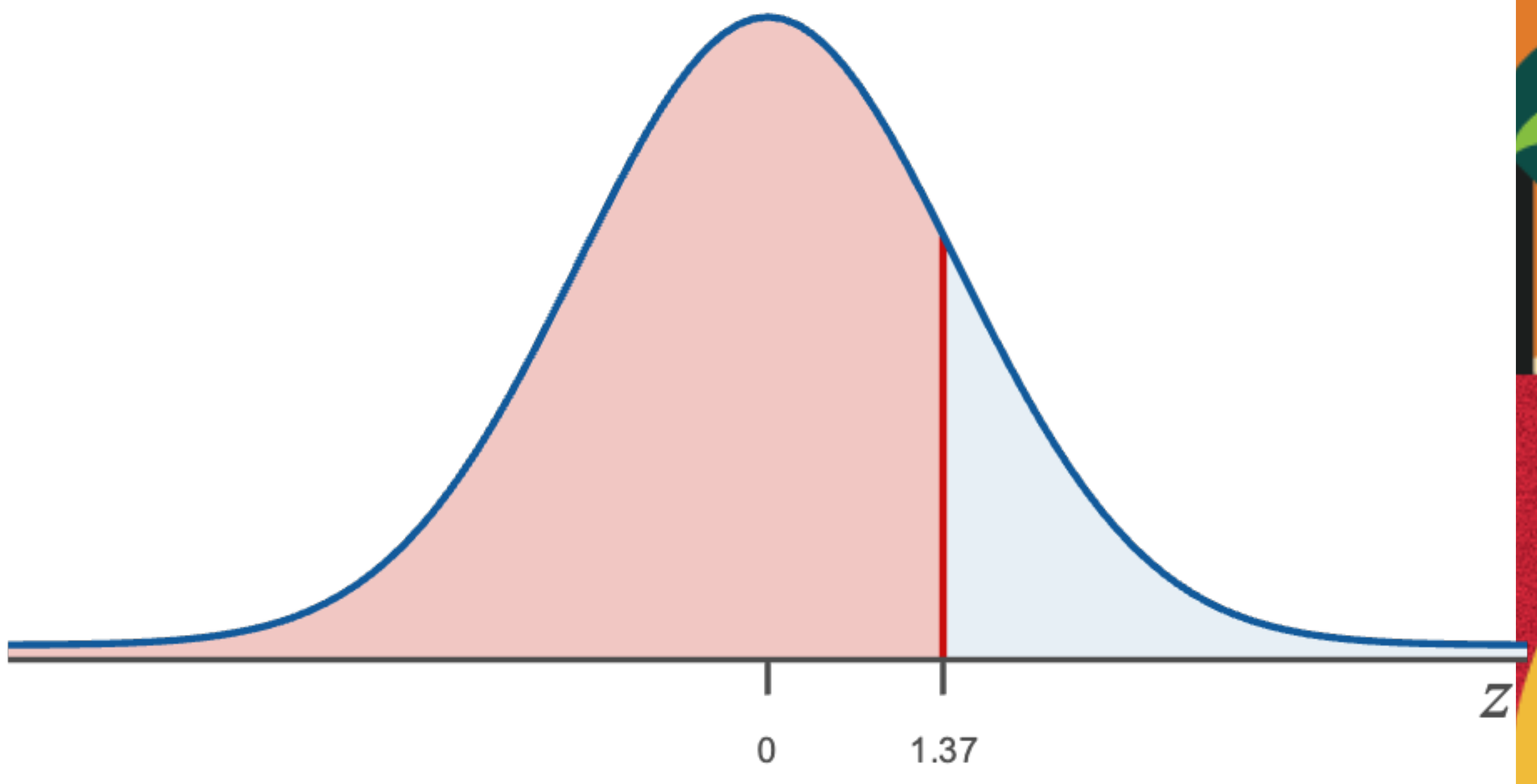
Standard Normal Distribution

- The standard normal distribution is a normal distribution with $\mu = 0$ and $\sigma = 1$
- Normal distribution function v.s. standard normal distribution function
  - $f(x) = \dfrac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$ v.s. $f(z) = \dfrac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}}$
- Standardizing a normal distribution
  - Compute the z-score for every x-value: $z = \dfrac{x-\mu}{\sigma}$
- Example: consider a normal distribution with $\mu = 48$ and $\sigma = 5$
  - Convert an x-value of 45 to a z-value
  - Draw two normal curves: one indicating the x-value in relation to the mean and standard deviation; the other indicating the placement of the z-value on the standard normal

Area left of z-score

- Find the area under the standard normal curve to the left of z=1.37.
  - $\mathbb{P}(Z \leq 1.37)$

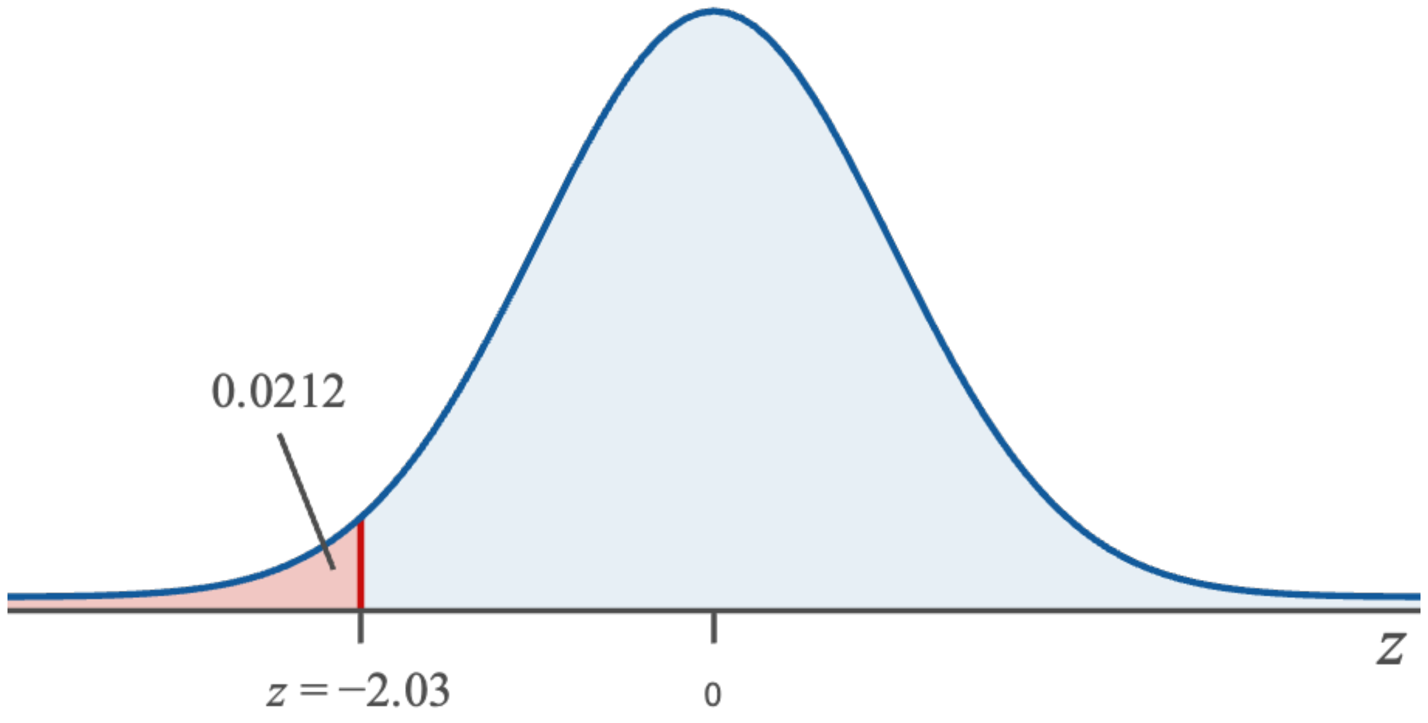
z	0.05	0.06	0.07	0.08	0.09
1.0	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9394	0.9406	0.9418	0.9429	0.9441



Area left of z-score

- Find the area under the standard normal curve to the left of z= -2.03
  - $\mathbb{P}(Z \leq -2.03)$

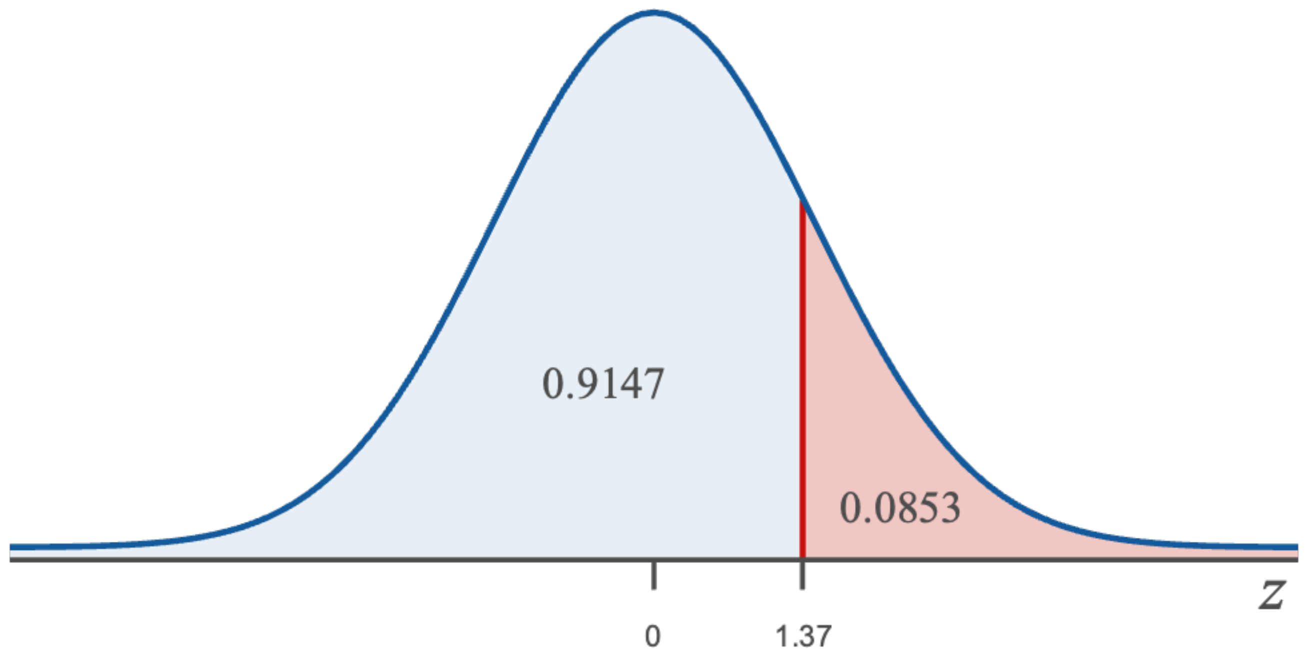
z	0.04	0.03	0.02	0.01	0.00
-2.2	0.0125	0.0129	0.0132	0.0136	0.0139
-2.1	0.0162	0.0166	0.0170	0.0174	0.0179
-2.0	0.0207	0.0212	0.0217	0.0222	0.0228
-1.9	0.0262	0.0268	0.0274	0.0281	0.0287
-1.8	0.0329	0.0336	0.0344	0.0351	0.0359
-1.7	0.0409	0.0418	0.0427	0.0436	0.0446



Area right of z-score

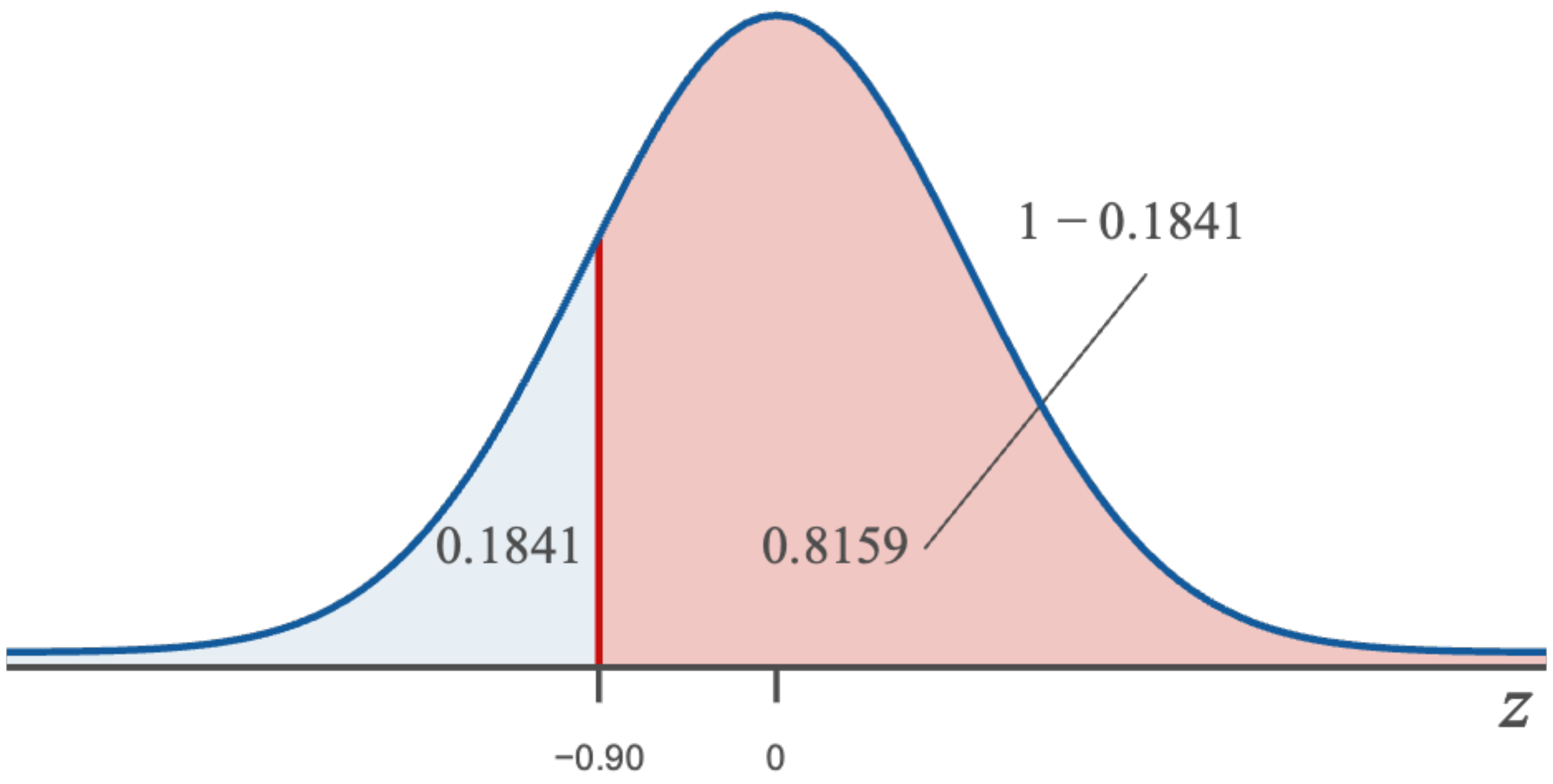
- Find the area under the standard normal curve to the right of z=1.37 using both methods described.
  - $\mathbb{P}(Z \geq 1.37)$

z	0.09	0.08	0.07	0.06	0.05
-1.6	0.0455	0.0465	0.0475	0.0485	0.0495
-1.5	0.0559	0.0571	0.0582	0.0594	0.0606
-1.4	0.0681	0.0694	0.0708	0.0721	0.0735
-1.3	0.0823	0.0838	0.0853	0.0869	0.0885
-1.2	0.0985	0.1003	0.1020	0.1038	0.1056
-1.1	0.1170	0.1190	0.1210	0.1230	0.1251



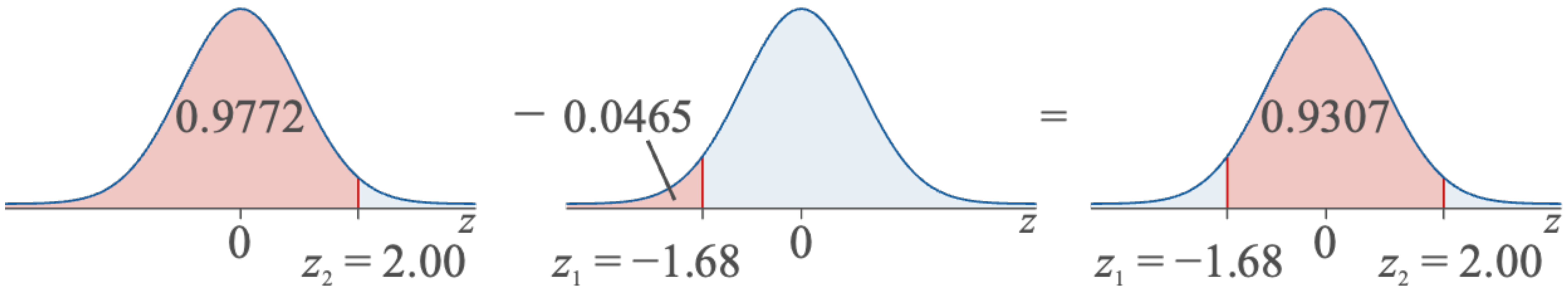
Area right of z-score

- Find the area under the standard normal curve to the right of z=-0.90.
  - $\mathbb{P}(Z \geq -0.90)$



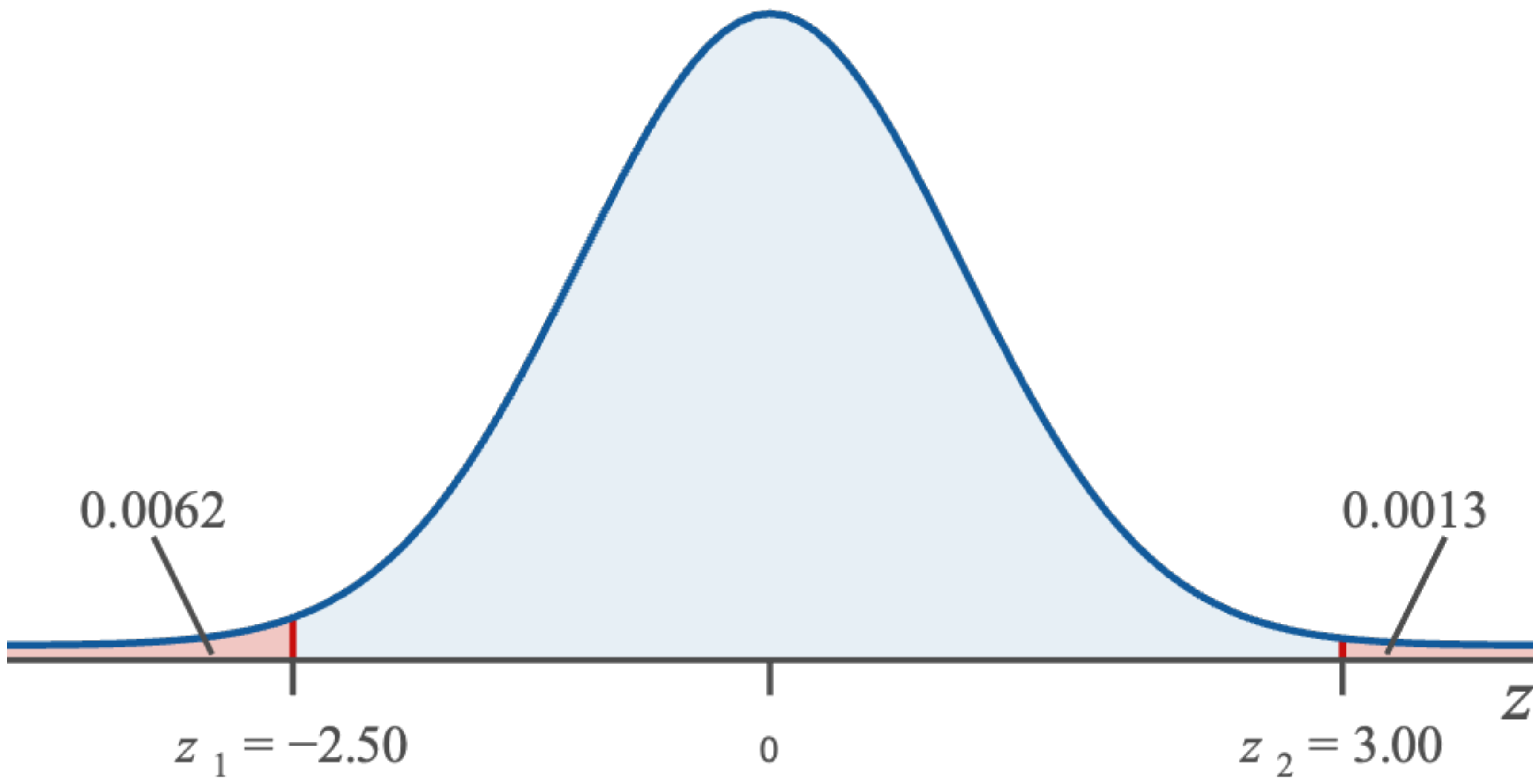
Area between two z-scores

- Find the area under the standard normal curve between $z_1 = -1.68$ and $z_2 = 2.00$
  - $\mathbb{P}(-1.68 \leq Z \leq 2.00)$



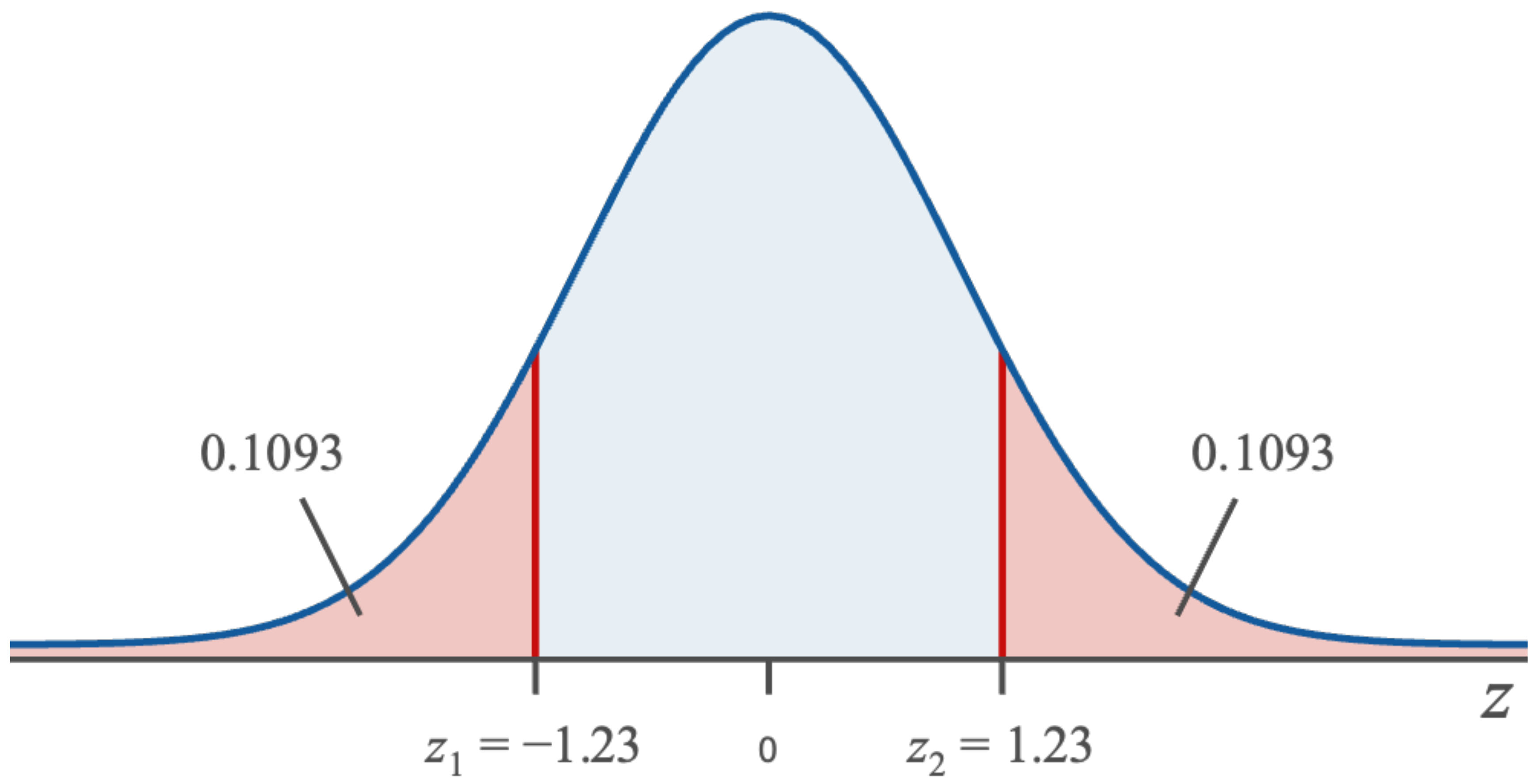
Area of tails

- Find the total of the areas under the standard normal curve to the left of $z_1 = -2.50$ and to the right of $x_2 = 3.09$
  - $\mathbb{P}(Z < -2.50 \text{ or } Z > 3.09)$



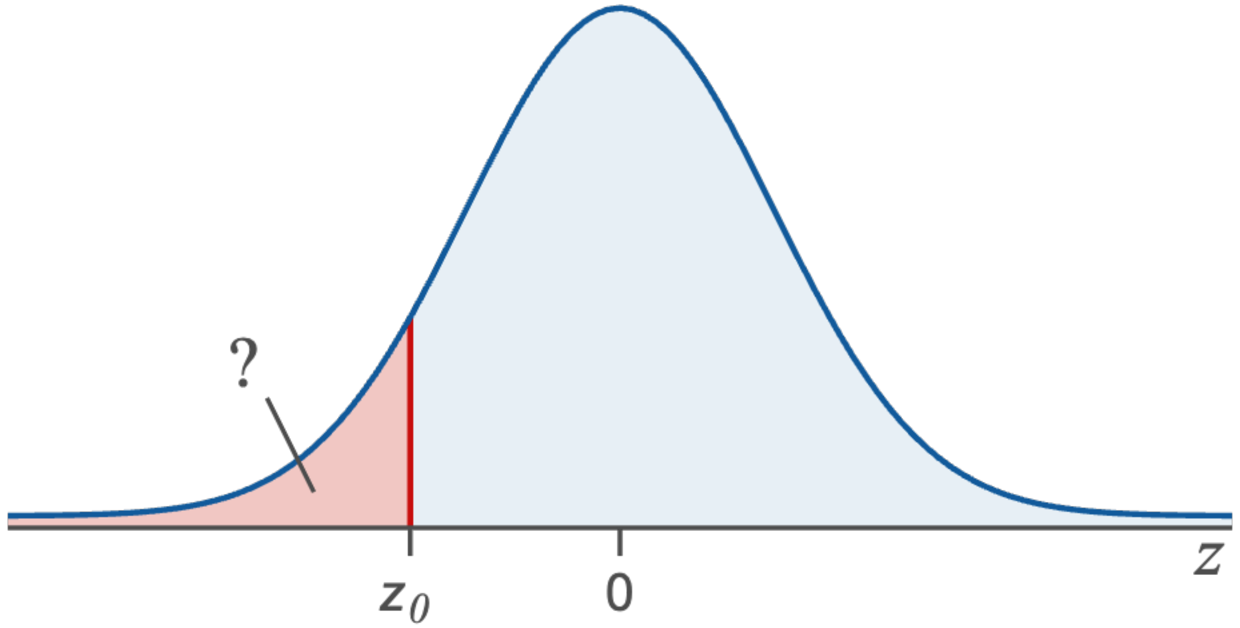
Area of tails

- Find the total of the areas under the standard normal curve to the left of $z_1 = -1.23$ and to the right of $x_2 = -1.23$
  - $\mathbb{P}(Z < -1.23$ or $Z > 1.23)$



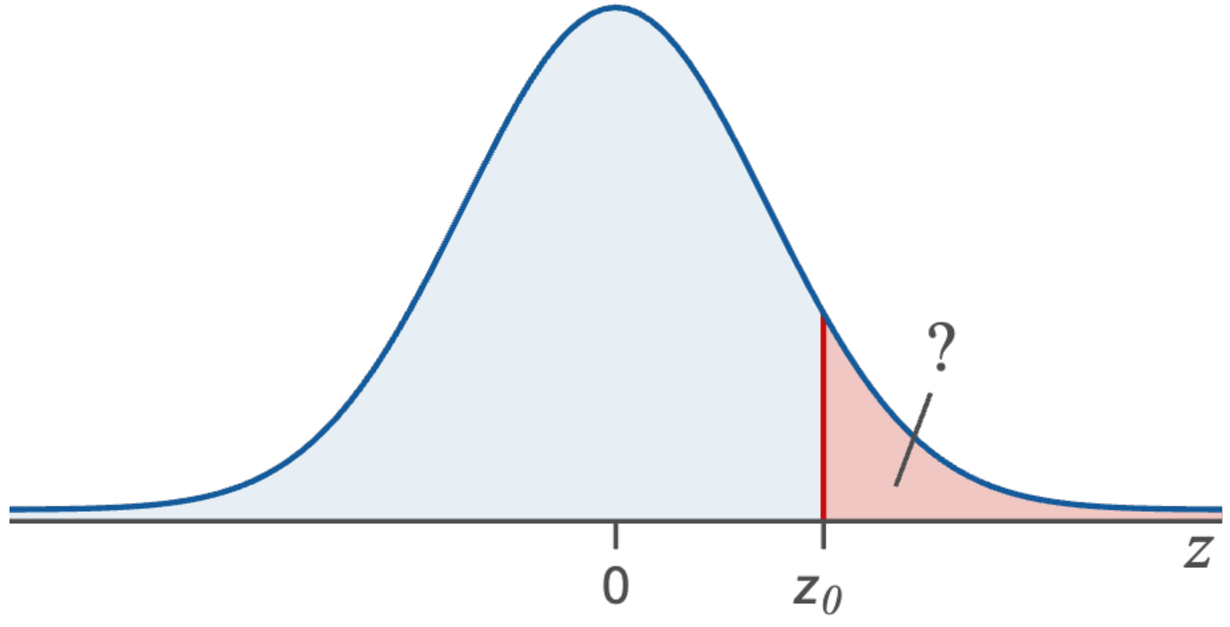
Computing areas

Area to the Left



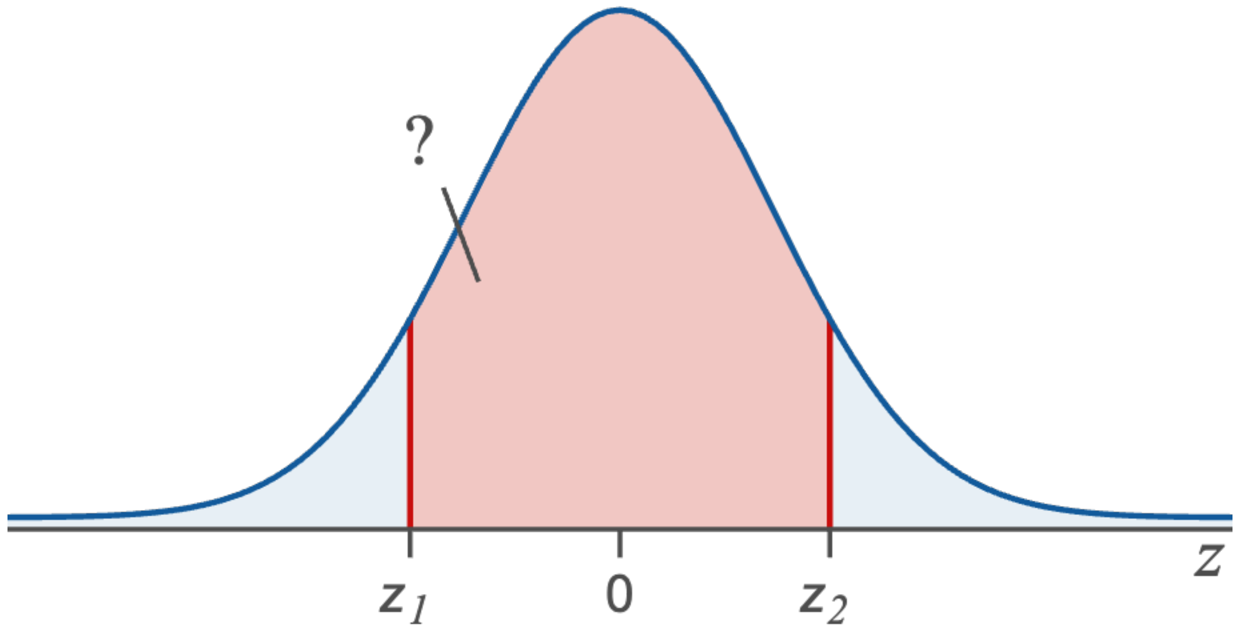
Look up z_0

Area to the Right



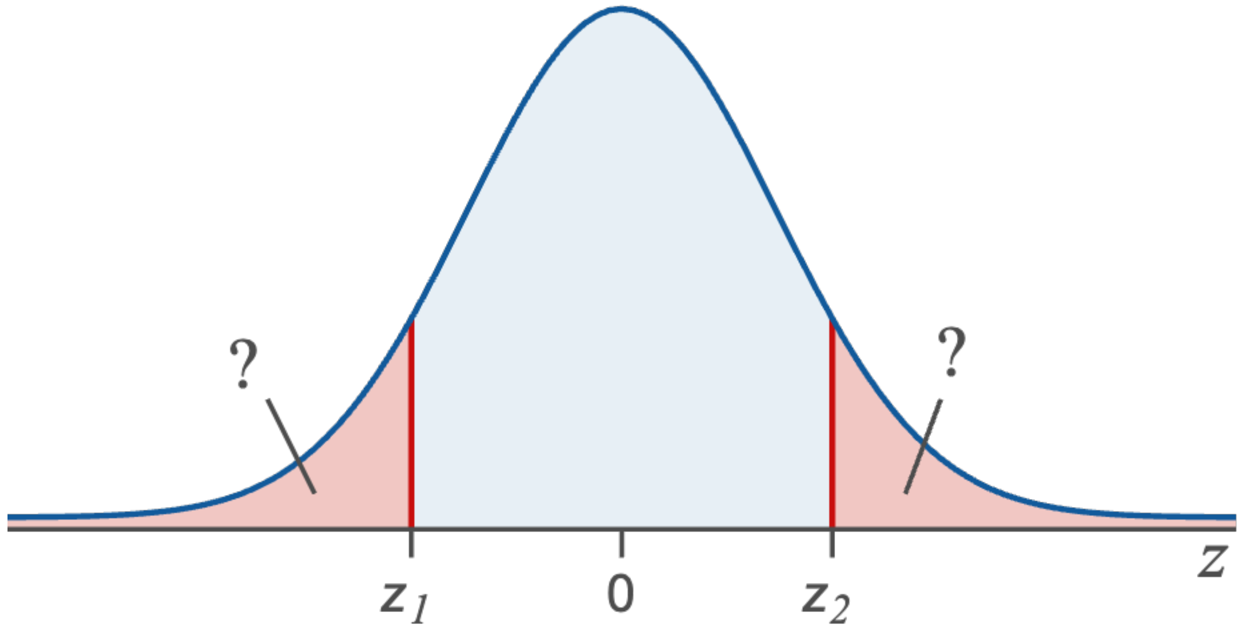
Look up $-z_0$

Area in Between



Look up z_1 and z_2 and subtract area to the left of z_1 from area to the left of z_2

Area in the Tails



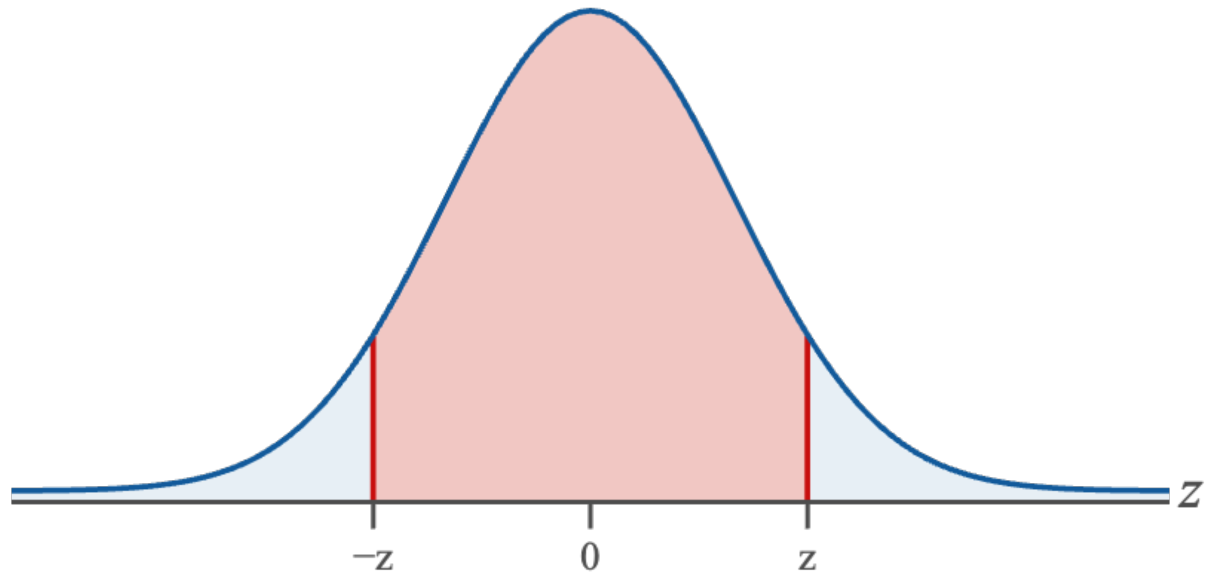
Look up z_1 and $-z_2$ and add areas

6.3 Finding Probability Using a Normal Distribution

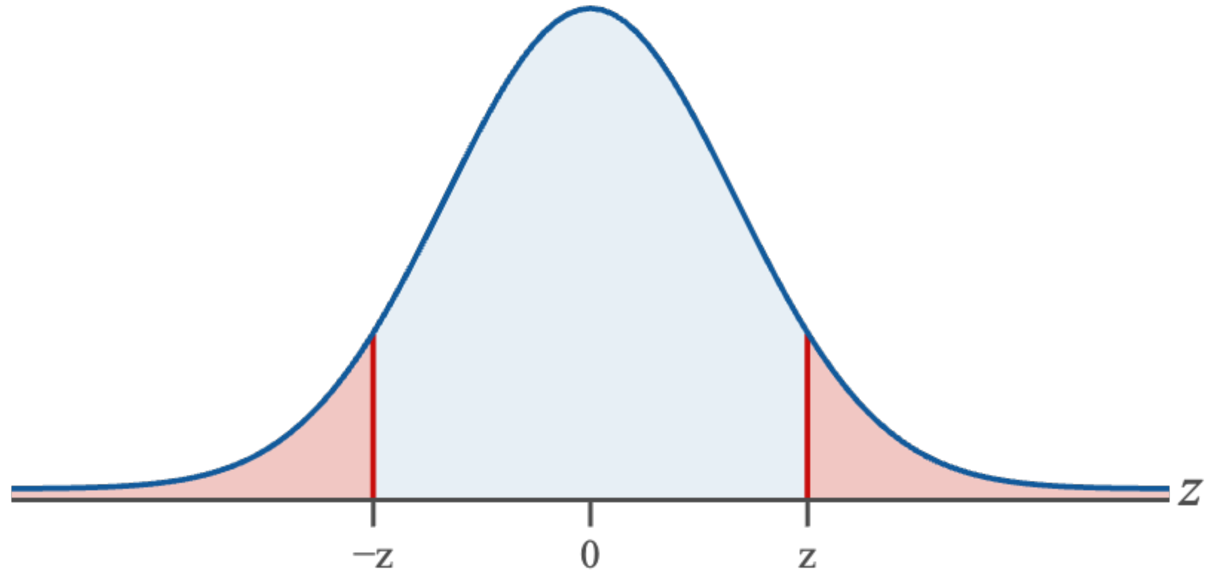
Finding probabilities

- How to find probabilities
  - Calculate z-scores
  - Compute areas using previous methods

- ‘Differs by less than’



- ‘Differs by more than’

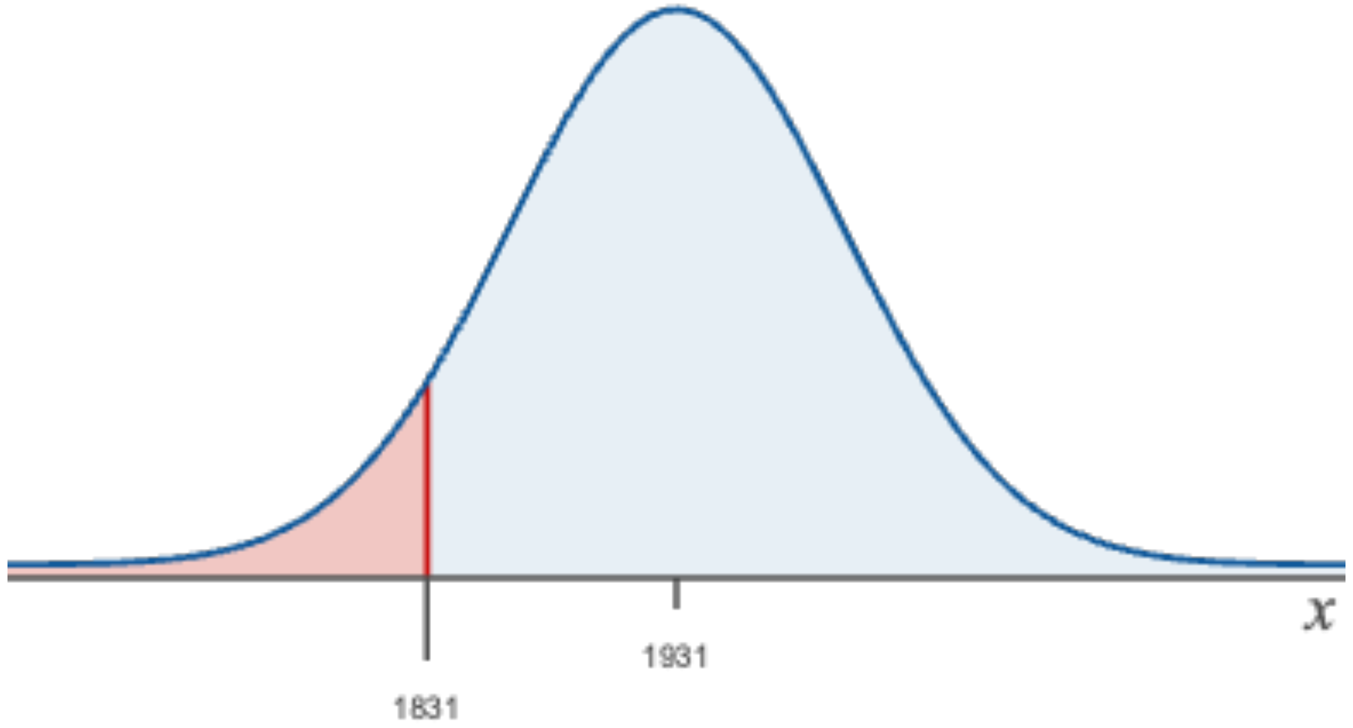


Example 1

- **Example 1:** Dairy Delight Farms has been milking cows for many generations. They have calculated that on average, the milk production per dairy cow over a month's time is 1931 lbs with a standard deviation of 67.1 lbs. The manager has noticed that the milk production from their cows seems to have been slowly decreasing in the past couple of years. However, their sales remain steady. Determine the probability that a random cow chosen from the farm produces less than 1831 pounds of milk per month. Assume that the milk production from the cows is normally distributed.

Example 1

- We are in a position to use the normal curve to answer this question since we are told to assume that milk production data are normally distributed, and we are given the mean and standard deviation of the distribution. Begin by sketching a normal curve to represent the distribution. Indicate the mean of the distribution, 1931 lbs, and the x-value we are interested in, 1831 pounds. Because this is an upper limit for the cow's milk production, we need to find the probability that X is less than 1831, written mathematically as $\mathbb{P}(X \leq 1831)$. Therefore, we shade the area to the left of 1831 on the curve.



Example 1

- In order to use cumulative probability tables, we need to convert 1831 to a standard score first.

$$\begin{aligned} z &= \frac{x-\mu}{\sigma} \\ &= \frac{1831-1931}{67.1} \\ &\approx -1.49 \end{aligned}$$

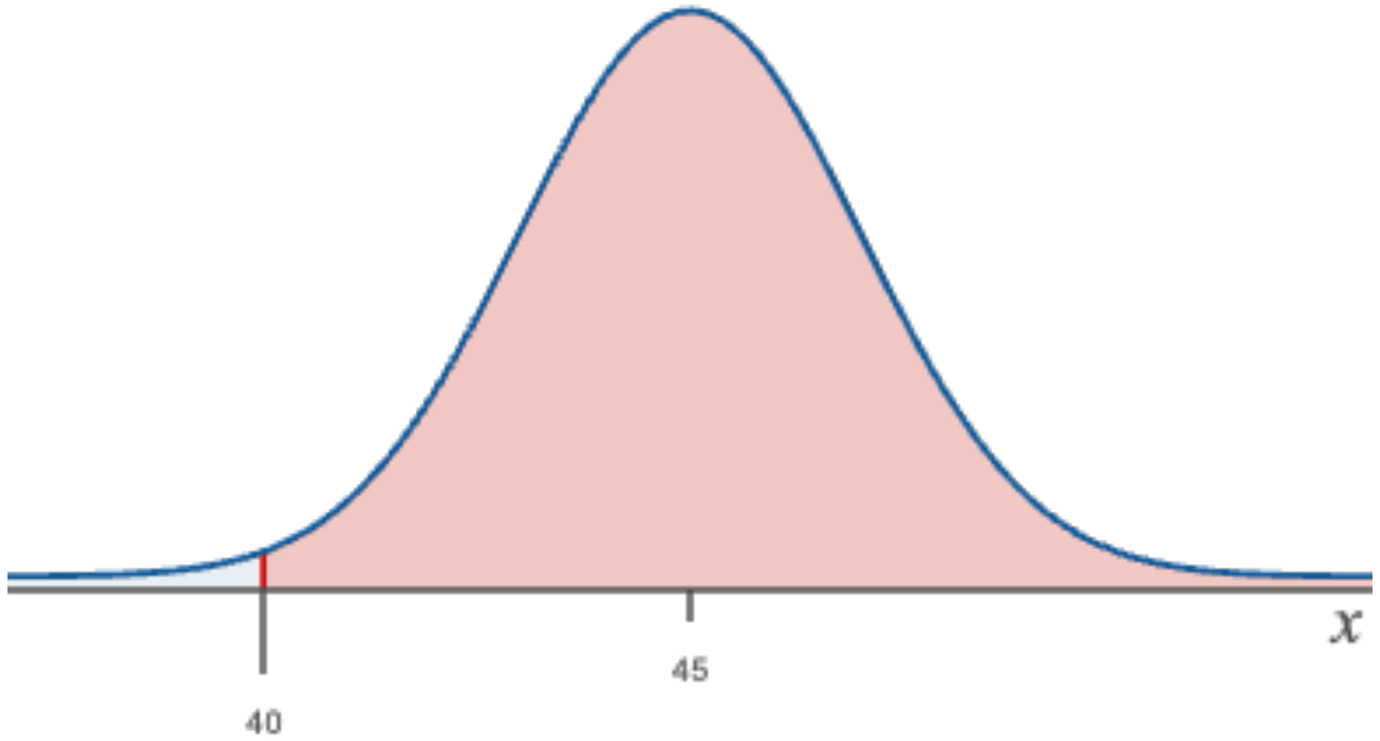
- Looking up -1.49 in the cumulative normal table we see that the shaded region has an area of approximately 0.0681.
- Thus, the probability of choosing a cow whose monthly milk production is less than 1831 pounds is approximately 0.0681 depending on the level of precision of the method used. This tells us that the farm manager could expect that around 6.8% of the dairy cows would be producing less than 1831 pounds of milk per month.

Example 2

- **Example 2**: Disney World is a dream for lots of children around the world. Full days of talking with live characters from the movies, wearing funny hats with ears, and riding rides of all types will wear out even the heartiest of souls at the end of the day. However, not everyone gets to ride every attraction. There are height restrictions on some of the rides. For instance, guests must be at least 40 inches tall to ride Splash Mountain. If heights for 6 year-olds are normally distributed with a mean of 3 ft 9 inches and a standard deviation of 2 inches, what percentage of 6-year-olds are allowed to ride Splash Mountain?

Example 2

- Note we are told that heights are normally distributed and can, therefore, use the normal distribution to answer the question. Since the height restriction is given in inches, we will first convert the mean into inches: $\mu =$ 3 ft 9 in = 45 in
- Draw a normal curve with a mean of 45 and standard deviation of 2 to represent the distribution of the heights of 6 year-olds. Denote 40 inches on the x-axis as well to indicate the height restriction for Splash Mountain. Because this is a lower limit for heights, we are interested in finding the probability that a 6 year-old's height X is more than 40 inches, written mathematically as $\mathbb{P}(X \geq 40)$. Therefore, we shade the area to the right of 40 on the curve.



Example 2

- In order to use the cumulative probability tables, we need to convert 40 to a standard score first.

$$z = \frac{x - \mu}{\sigma}$$

$$z = \frac{40 - 45}{2}$$

$$= -2.5$$

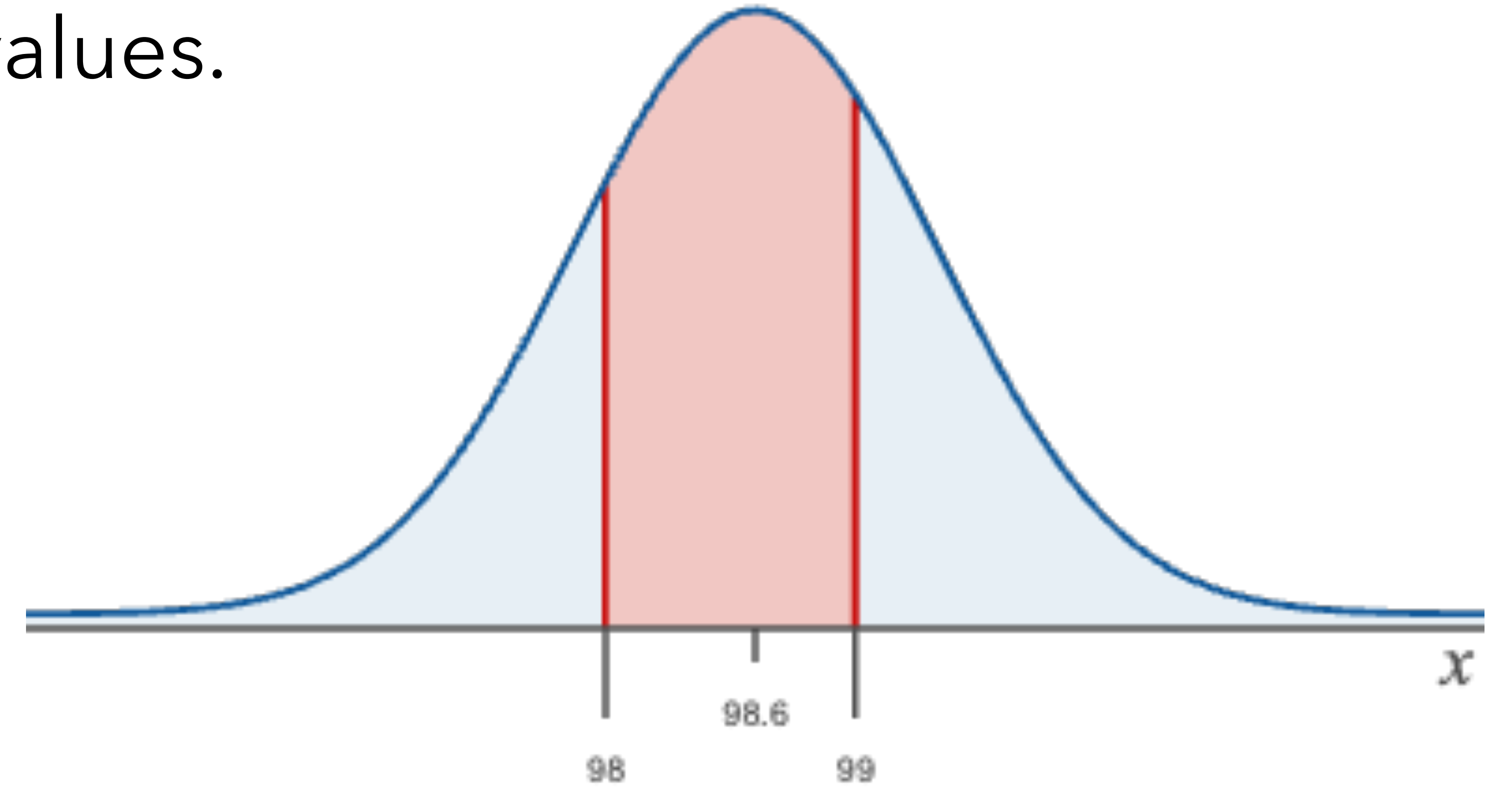
- Because we need the area to the right, we can look up the negative of the z-value we calculated. Looking up 2.5 in the cumulative normal table we see that the shaded region has an area of approximately 0.9938.
- Thus, the probability of a 6 year-old having a height of at least 40 inches is approximately 0.993790 using either method of calculation. We can also interpret this to mean that 99.38% of 6 year olds have a height of at least 40 inches. Therefore, only a small percentage of 6 year-olds (less than 1%) are excluded from riding Splash Mountain.

Example 3

- **Example 3**: As adults, our body temperatures are normally distributed with a mean of 98.6°F and a standard deviation of 0.73°F. Sometimes our bodies increase in temperature and we say we have a fever. Other times our body temperatures might drop to a place of hypothermia. Extreme fluctuations in either direction might be signals that we need to seek medical help. With this in mind, answer the following questions.
  - Part A: What percentage of healthy adults have a body temperature which is between 98°F and 99°F?
  - Part B: What percentage of healthy adults typically have either a consistently high body temperature of more than 100°F or a consistently low temperature of less than 97°F?
  - Part C: What is the probability of a healthy adult having a body temperature that differs from the population mean by more than 1°F?

Example 3A

- Begin by drawing a normal curve with a mean of 98.6 and standard deviation 0.73 to represent the distribution of body temperatures. Denote both 98 and 99 on the x-axis to indicate the two temperatures that form the interval we are interested in. This is written mathematically as P(98<X<99) and we shade the area under the curve which is in between these two values.



Example 3A

- Convert both temperatures to standard scores then use the cumulative probability tables:
  - For 98°F, $z = \dfrac{98 - 98.6}{0.73} \approx -0.82$
  - For 99°F, $z = \dfrac{99 - 98.6}{0.73} \approx 0.55$
- The area between these two z-values can be found by looking up both in the tables and then subtracting their values as shown below: $0.7088 - 0.2061 = 0.5027$
- This means that the percentage of healthy adults that have a body temperature which is between 98°F and 99°F is approximately 50%.

Example 3B

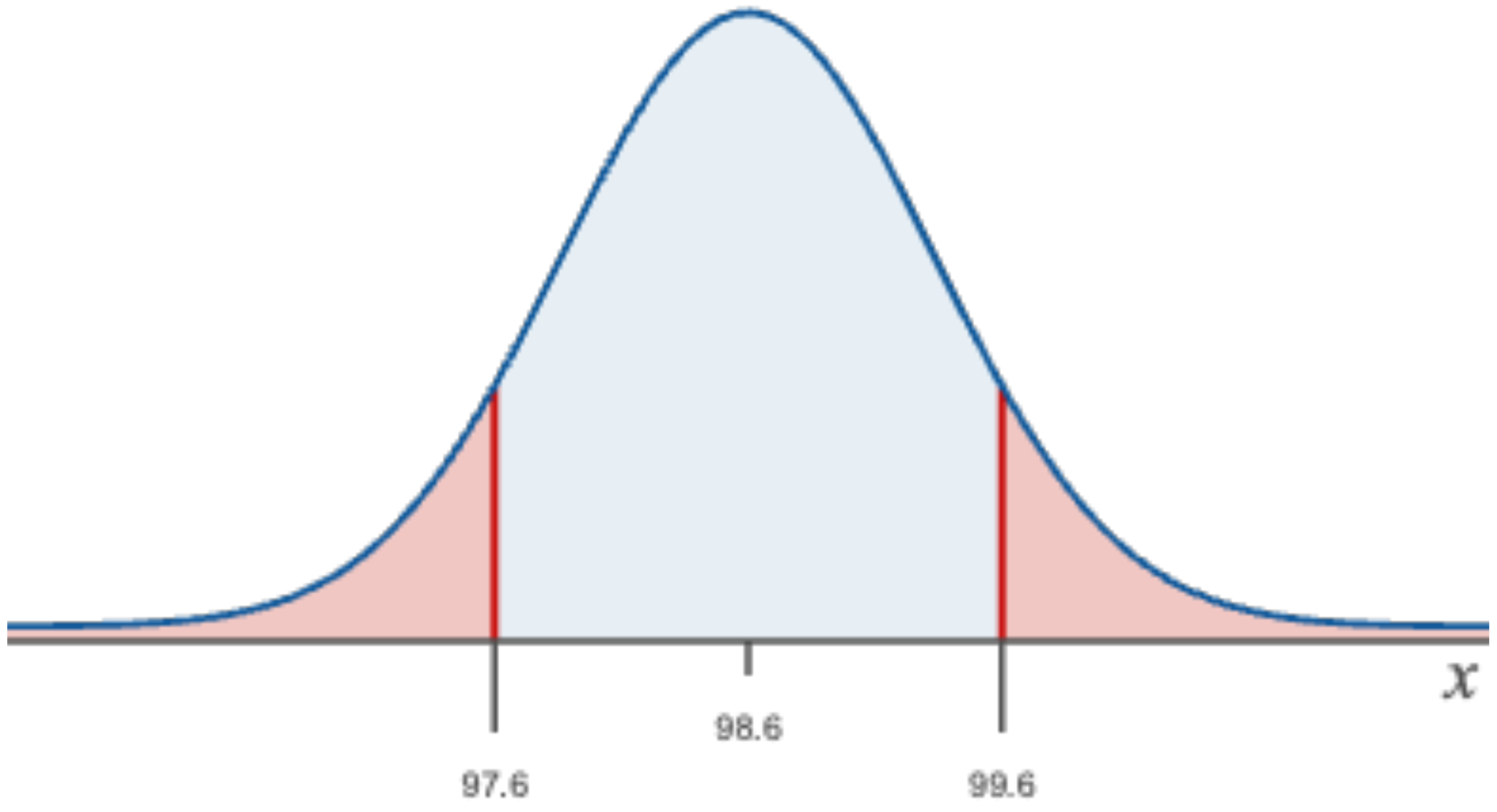
- b. Again, start by drawing a normal curve with a mean of 98.6 and standard deviation 0.73 to represent the distribution of body temperatures. This time denote both 97 and 100 on the x-axis to indicate the two temperatures that are the extremes we are interested in. Shade the area under the curve which is below 97 and the area which is above 100. This is written mathematically as $\mathbb{P}(X < 97 \text{ or } X > 100)$

Example 3B

- Convert both temperatures to standard scores and then use the cumulative probability tables:
  - For 97°F: $z = \dfrac{97 - 98.6}{0.73} \approx -2.19$
  - For 100°F: $z = \dfrac{100 - 98.6}{0.73} \approx 1.92$
- The area can be found by looking up the area to the left of z=-2.19 and the area to the right of z=1.92 and then adding their values: $0.0143 + 0.0274 = 0.0417$
- Thus, the percentage of healthy adults who have either a consistently high body temperature of more than 100°F, or a consistently low temperature of less than 97°F is approximately 4%.

Example 3C

- Because we want the temperatures to differ from the population mean by more than 1°F, we need to add 1°F to and subtract 1°F from the mean 98.6°F to get the temperatures in which we are interested. Doing this tells us we are interested in finding the probability that a person's body temperature X is less than 97.6°F or greater than 99.6°F. We can denote this mathematically as $P(X<97.6 \text{ or } X>99.6)$. Draw the normal curve as before, shading below 97.6 and above 99.6.



Example 3C

- Convert both temperatures to standard scores and then use the cumulative probability tables:
  - For 97.6°F: $z = \dfrac{x - \mu}{\sigma} = \dfrac{97.6 - 98.60}{0.73} \approx -1.37$
  - For 99.6°F: $z = \dfrac{99.6 - 98.60}{0.73} \approx 1.37$
- Notice that these z-scores are equal distances from the mean. Because of the symmetry of the curve, the shaded areas must then be equal. The easiest way to calculate the total area is to find the area to the left of $z_1 \approx -1.37$ and simply double it. We then have $(0.0853)(2) = 0.1706$.
- Thus, the probability of a healthy adult having a body temperature that differs from the population mean by more than 1°F is approximately 17%.

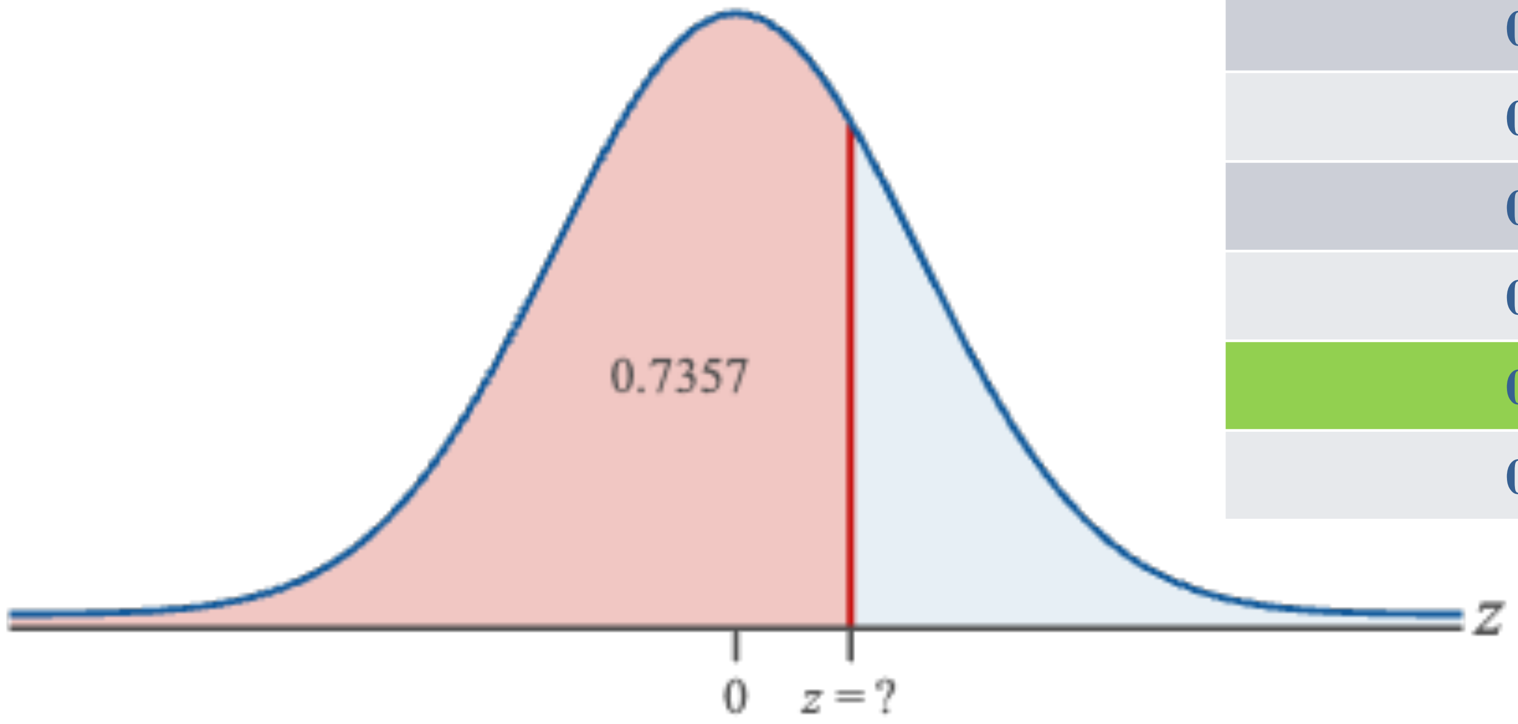
6.4 Finding Values of a Normally Distributed Random Variable

x- and z- values from probabilities

- We can do reverse lookups to calculate z-scores from probabilities
- Then, we can calculate x-values from z-scores given the mean and standard deviation
  - $z = \dfrac{x - \mu}{\sigma}$ implies that $x = \sigma z + \mu$

z-values from probabilities

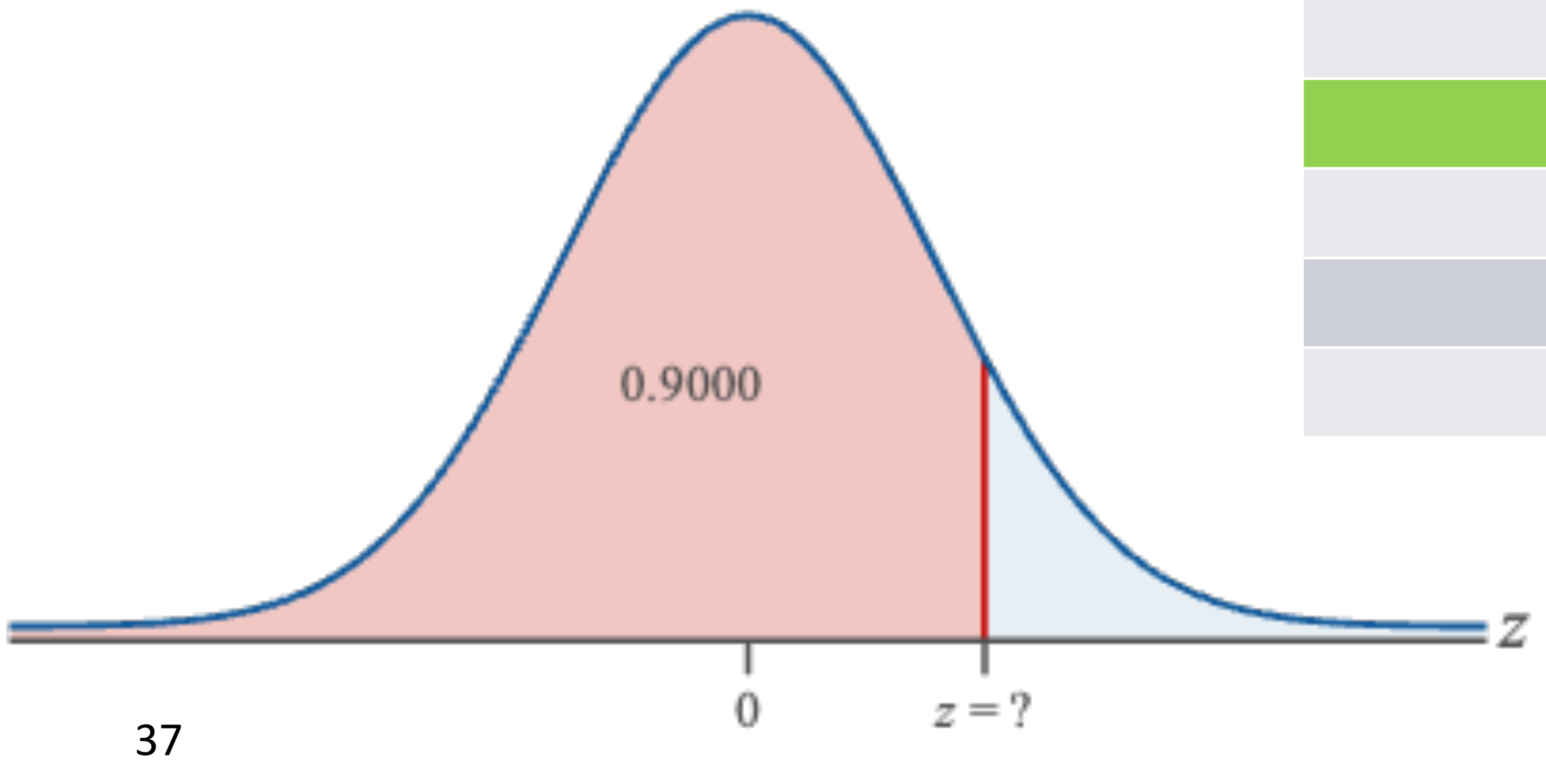
- Question: What z-value has an area of 0.7357 **to its left**?
- Answer:



z	0.00	0.01	0.02	0.03	0.04
0.0	0.5000	0.5040	0.5080	0.5120	0.5160
0.1	0.5398	0.5438	0.5478	0.5517	0.5557
0.2	0.5793	0.5832	0.5871	0.5910	0.5948
0.3	0.6179	0.6217	0.6255	0.6293	0.6331
0.4	0.6554	0.6591	0.6628	0.6664	0.6700
0.5	0.6915	0.6950	0.6985	0.7019	0.7054
0.6	0.7257	0.7291	0.7324	0.7357	0.7389
0.7	0.7580	0.7611	0.7642	0.7673	0.7704

z-values from probabilities

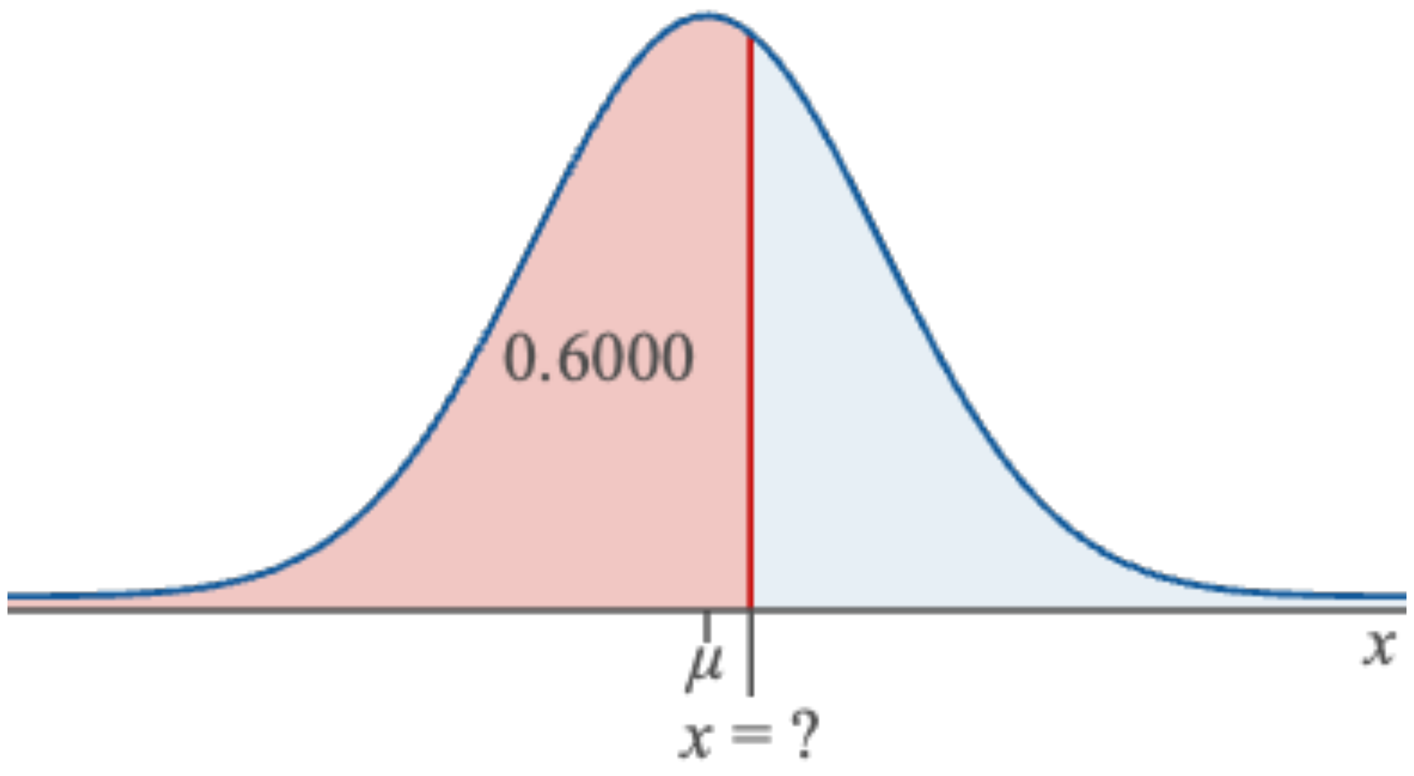
- Question: What z-value represents the 90th percentile?
- Answer: 90th percentile corresponds to z-value with 90% of the area under standard normal curve to the left of z. So we need to find a value of z that has an area of 0.9000 to its left.



z	0.05	0.06	0.07	0.08	0.09
1.0	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9394	0.9406	0.9418	0.9429	0.9441

x-values from probabilities

- At his most recent well-being checkup, Ethan's doctor told him that for his age, his cholesterol level was in a good range at the 60th percentile. Assume cholesterol levels are approximately normally distributed with a mean of 189.1 mg/dL and a standard deviation of 37.8 mg/dL for Ethan's age group. What was Ethan's cholesterol level at his last checkup?
- Answer: The 60th percentile means that 0.6000 of the area under the curve will be to the left as shown below. Then use either the cumulative normal tables or technology to find the value of x



x-values from probabilities

- To use the normal tables, we first need to find the z-value that represents the 60th percentile. Looking for 0.6000 (or the closest area to it) in the interior of the tables we find 0.5987, which corresponds to a z-value of 0.25.
- We can now substitute this value, along with the mean and standard deviation, in the formula for standard scores and use algebra to find the x value:

$$
\begin{aligned}
z &= \frac{x-\mu}{\sigma} \\
0.25 &= \frac{x-189.1}{37.8} \\
(0.25)(37.8) + 189.1 &= x \\
198.55 &= x
\end{aligned}
$$

6.5 Approximating a Binomial Distribution Using a Normal Distribution

Approximating a Binomial Distribution

- Recall, the probability of a binomial random variable: $\mathbb{P}(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$
- If the conditions that $np \geq 10$ and $n(1-p) \geq 10$ are met for a given binomial distribution, then a normal distribution can be used to approximate the binomial probability distribution with mean and standard deviation given by
  - $\mu = np$
  - $\sigma = \sqrt{np(1-p}$
  - n, number of trials
  - p, probably of success in a trail
- A **continuity correction** is a correction factor used to convert a value of a discrete random variable to an interval range of a continuous random variable when using a continuous distribution to approximate a discrete distribution.

Procedure

1. Determine the values of n and p.
2. Verify that the conditions $np \geq 10$ and $n(1-p) \geq 10$ are met.
3. Calculate the values of the mean and standard deviation of the binomial random variable using the formulas $\mu = np$ and $\sigma = \sqrt{np(1-p)}$.
4. Use a continuity correction to determine the interval corresponding to the given value of x.
5. Draw a normal curve using the mean and standard deviation calculated in Step 3, and label it with the information given in the problem.
6. Using either a z-value with normal distribution tables or available technology, find the appropriate area under the normal curve.

Example

- **Example 1**: Use the continuity correction factor to describe the area under the normal curve that approximates the probability that at least 2 people in a statistics class of 50 students regularly cheat on their math tests. Assume that the number of people in a statistics class of 50 students who consistently cheat on their math tests has a binomial distribution with a mean of 5.00 and a standard deviation of approximately 2.12.
- Begin by converting the discrete number 2 into an interval by adding 0.5 to and subtracting 0.5 from the number 2. The discrete number 2 is changed to the continuous interval from 1.5 to 2.5. Now, draw a normal curve with a mean of 5.00 and a standard deviation of 2.12, and indicate the interval from 1.5 to 2.5 to represent the number 2. Next, shade the area corresponding to the phrase at least 2. This would be the area under the curve for x-values greater than or equal to 2. Thus, the area corresponding to at least 2 would include the interval from 1.5 to 2.5 and all x-values to the right of 2.5.

Continuity correction

- Thus, the area under the normal curve with a mean of 5.00 and a standard deviation of 2.12 that approximates the probability that at least 2 people in the statistics class regularly cheat on their math tests is the area to the right of 1.5.