Normal Distribution

Sravani Vadlamani

QMB 3200

Transformations

Transformations

- Often, we transform observed scores to make them more interpretable

- Example:
  - You received a score of 74 out of a possible 82 on the first exam
  - What’s the percentage?

$$\frac{74}{82} = \frac{x}{100} \qquad x = \frac{7400}{82} \qquad x = 90.2$$

Effects of Transformations

- When we perform transformations on the observed scores, it makes sense that the measures of central tendency and variability are also transformed

  - Let’s see how…

Transformations

	X_i	$Percent_i$ $(X_i/82)*100$
	74	90.2
	62	75.6
	81	98.8
	76	92.7
	80	97.6
	82	100
M	75.8	92.5
SD^2	55.4	82.6
SD	7.4	9.1

Effect of Linear Transformations on Mean, Variance, and Standard Deviation

- Adding a constant to all data
  - Increases the mean by that constant
  - Does not change the variance and standard deviation

- Multiplying all data by a constant
  - Multiplies the mean by that constant
  - Multiplies the standard deviation by that constant
  - Multiplies the variance by the constant^2

- Similar effects for subtracting and dividing by a constant

Z-Scores

Z-Score

- Transformation in terms of the mean and standard deviation
  - We always like to compare ourselves with others around us

- Example
  - Tom scored a 76 on the statistics test. How did he do?
    - If the Mean was 68 & Standard Deviation was 3, then we know Tom did well
    - If the Mean was 86 & Standard Deviation was 5, then we know Tom did poor

Uses of Z-Scores

- Describes where an individual score falls within a distribution
- Number of Standard Deviations above or below the mean an individual score is
  - Mean is the point of reference
    - Is individual above or below the mean?
  - Standard Deviation is the yardstick
    - How much is the individual above or below the mean?

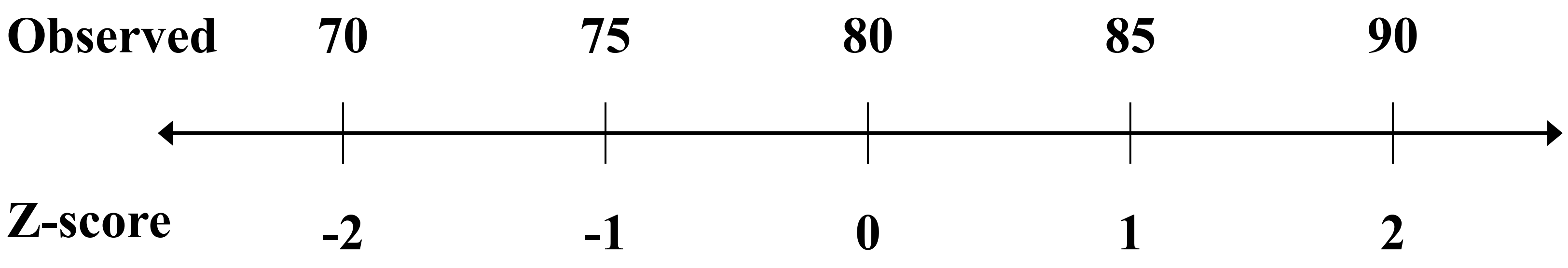
Formula to Calculate Z-Scores

$$z_i = \frac{x_i - \mu}{\sigma}$$

Note: This is a linear transformation. The distribution characteristics (e.g., symmetry, skew, modality) remain

Example

- Assume the distribution of the observed scores has a Mean of 80 and a Standard Deviation of 5
- Transformation to Z-scores will transform the Mean to be 0 and the Standard Deviation to be 1



Example

- Tom scores a 76 on the statistics test. How did he do?
  - Class Mean is 68.

- How much worse or better relative to other students in the class?
  - Standard Deviation is 8
  - Standard Deviation is 4

Tom scored 1 standard deviation above the class mean

Tom scored 2 standard deviations above the class mean

Z-scores

- Because we transform the observed (raw) scores to z-scores, it's easy to compare scores obtained from two different distributions

- Example
  - Tom scored 35 on his first exam (class mean = 30, class standard deviation = 5). Tom scored 25 on the second test (class mean = 21, class standard deviation = 2)

Interpretation of Z-Scores

- Positive or Negative Values (including 0)
  - Positive z-score – above the mean
  - Negative z-score – below the mean
  - 0 z-score – at the mean
- Magnitude (size) lets us know the distance from the mean (in terms of standard deviations)
  - Larger absolute values indicates the score is further away from the mean
  - Example:
    - Which is z-score is further from the mean?
      - Z-score of -.5 vs. Z-score of 1

Reverse Example

- The mean depression score was 12 with a variance of 16. Matthew had a z-score of -.75. What was his depression score?

$$z_i = \frac{x_i - \mu}{\sigma} \longrightarrow -.75 \cdot 4 + 12 = X_i$$

$$\downarrow$$

$$9 = X_i$$

Why Z-Scores (Standard Scores)

- Given certain assumptions (Normal Distribution), z-scores allow us to determine probabilities of events

- Statistics is inherently concerned with the probability of events

Statistics & Probability

- Statistics directly deal with the probability of an event
  - Goal is to determine the probability that the event occurred at random

  - If the probability that the event occurred at random is low (i.e., $< .05$), we claim that the event did not occur at random and suggest a link (e.g., causal mechanism)

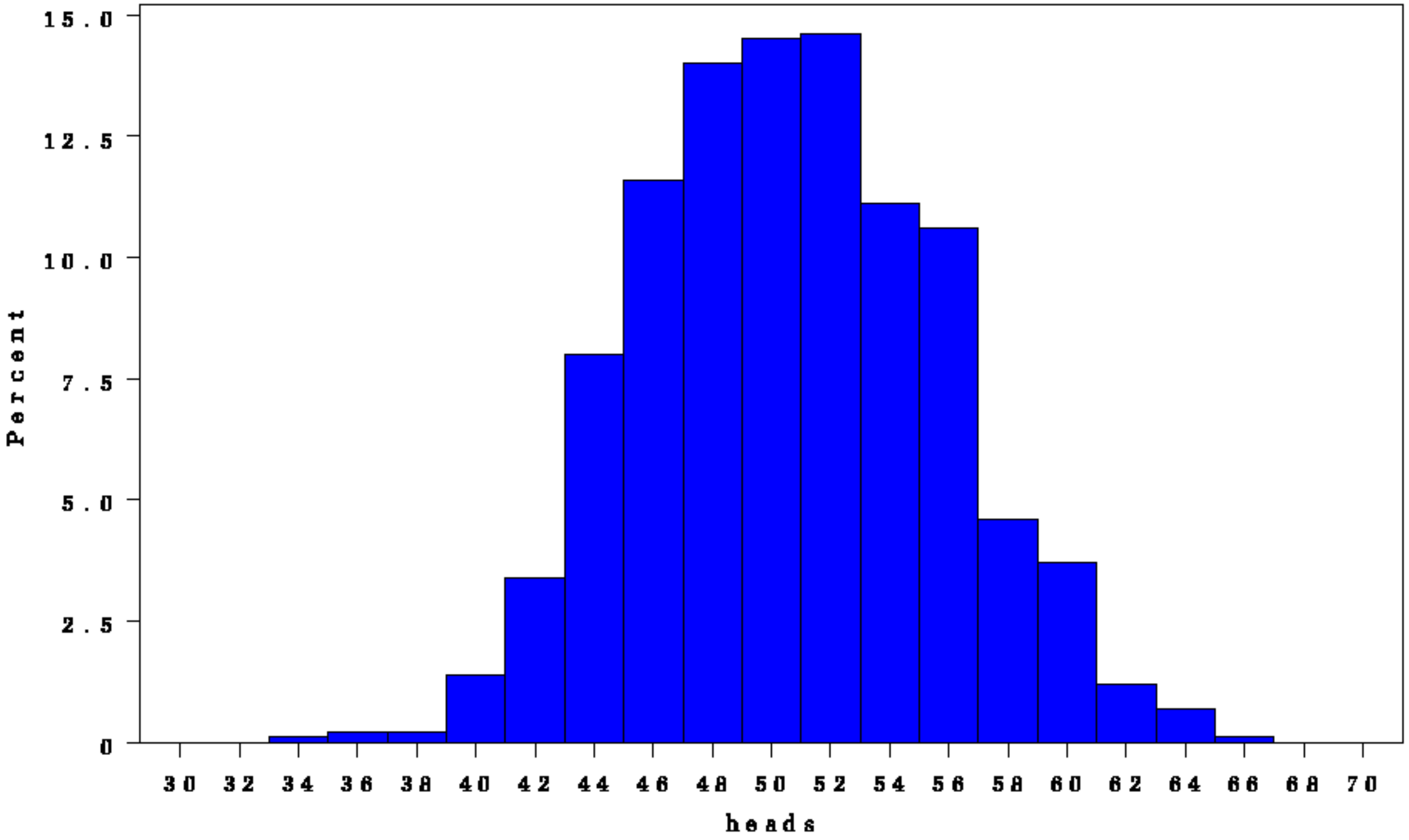
Statistics & Probability

- Statistics require several assumptions in order to easily calculate probabilities
- One of the biggest assumptions is that the distribution of the (outcome) variable in the *population* follows a *Normal Distribution*
- By making this assumption, we can easily calculate the probabilities of events and determine whether or not the event was unusual (i.e., unlikely to occur at random)

Statistics & Probability

- *Example*
  - A fair coin is tossed 100 times and the number of heads is counted. This process is repeated 1000 times
  - Outcome variable
    - The number of heads out of 100 tosses
  - What can we expect?

Histogram of the Coin Toss Data



N Mean Std Dev Minimum Maximum

ƒƒ

1000 50.15 5.05 34.00 65.00

ƒƒ

Statistical Question

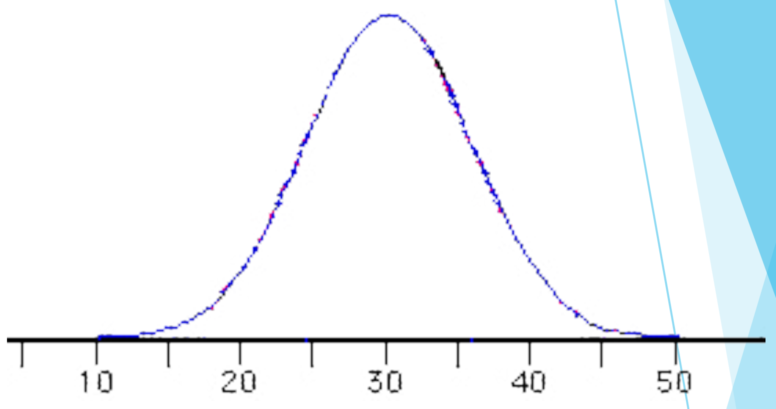
- How do I know if a *new* coin is fair?
  - Yes or no question
  - We can't answer this question definitively using statistics, but we can infer the answer probabilistically
    - Cannot prove something is true or false, but we can determine the probability the new coin is fair and make a determination based on that probability
- To answer this question
  - Toss the *new* coin 100 times
  - Determine the probability (likelihood) that the number of heads could have arose from a fair coin (based on our sampling distribution and assumption of a normal distribution in the population)

Normal Distribution

- The distribution of many quantitative variables are unimodal, roughly symmetric, and bell-shaped.
  - Approximate a normal distribution
- Most of statistics (and all statistics we consider in this class) assume the outcome is normally distributed

Normal Distribution

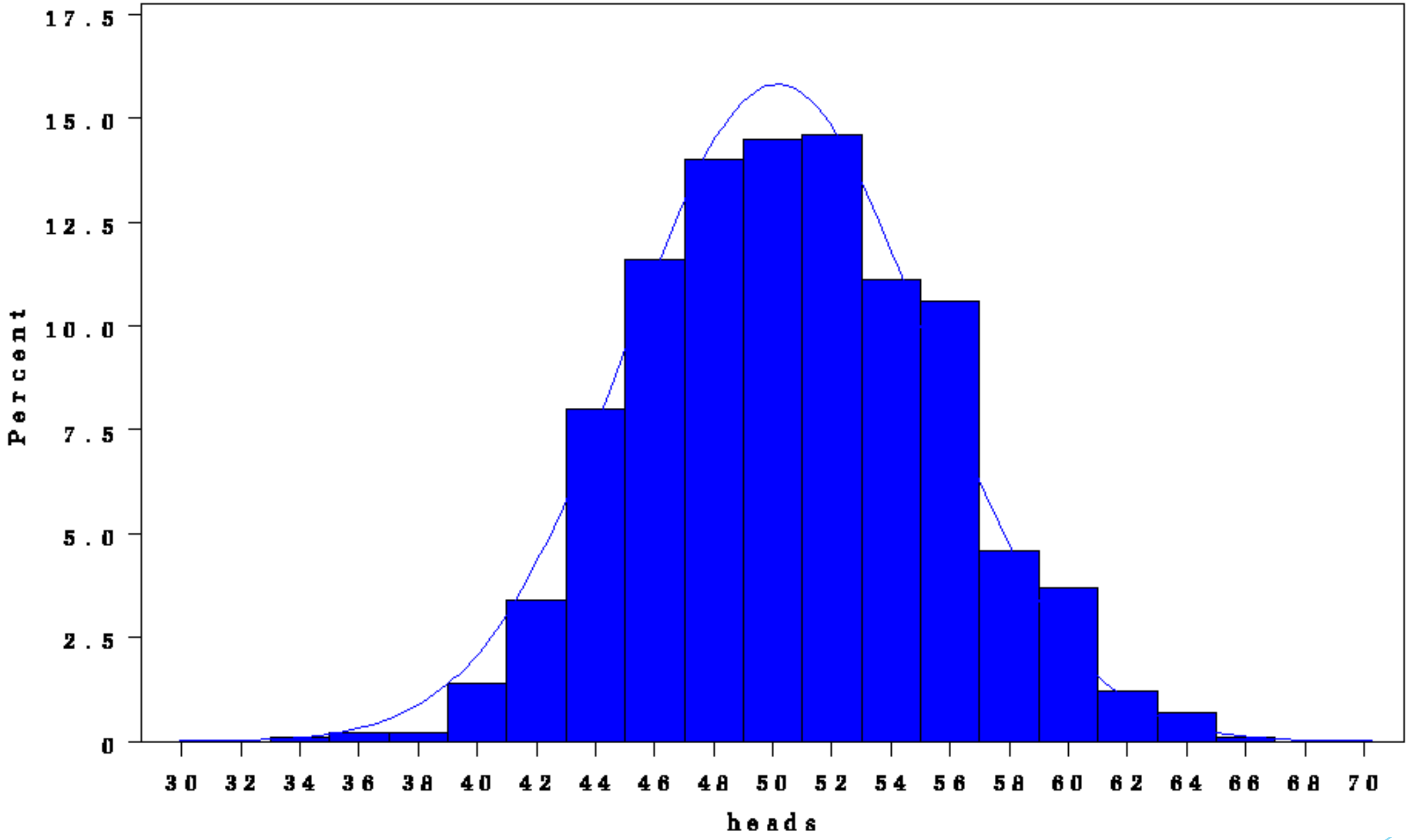
- Bell-shaped
- Symmetric about the Mean
- Mean = Mode = Median
- “Tails” Extend from ± Infinity
- Tails approach, never touch X-axis
- Can be Described by a Function (Equation)



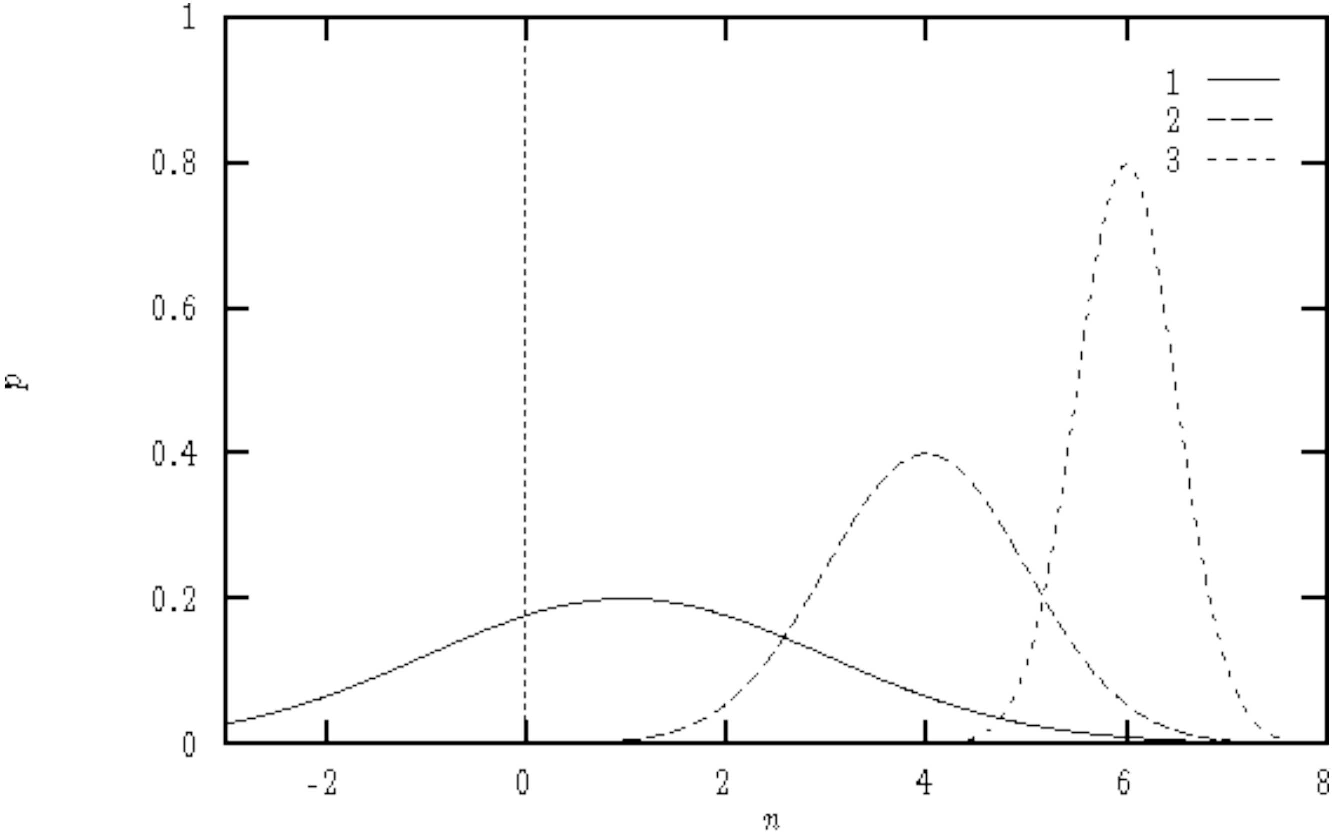
Normal Distribution

- Many variables are distributed normally in the population (Roughly speaking)
  - e.g. Intelligence & Introversion
- Statistical procedures often assume a normal distribution in the *population*, which allows us to make a number of inferences
- Normal Distributions can be defined by an arithmetic function
  - Can compute area under portions of the curve
- Sampling distributions of statistics are often assumed to be normally distributed (under certain conditions)
  - Enables hypothesis testing using normal distributions

Histogram of Coin Tossing Data - Normal Distribution Imposed



Normal Distributions with Different Means and Standard Deviations



1. $\mu = 1,\ \sigma = 2$
2. $\mu = 4,\ \sigma = 1$
3. $\mu = 6,\ \sigma = 0.5.$

Normal Distribution

Dependent only on the Mean and Standard Deviation

$$f(X) = \frac{1}{\sigma\sqrt{2\pi}}(e)^{-(X-\mu)^2/2\sigma^2}$$

X is the variable (e.g., extraversion)

e is a constant (~2.71828)

π is a constant (~3.1415)

μ is the mean of X

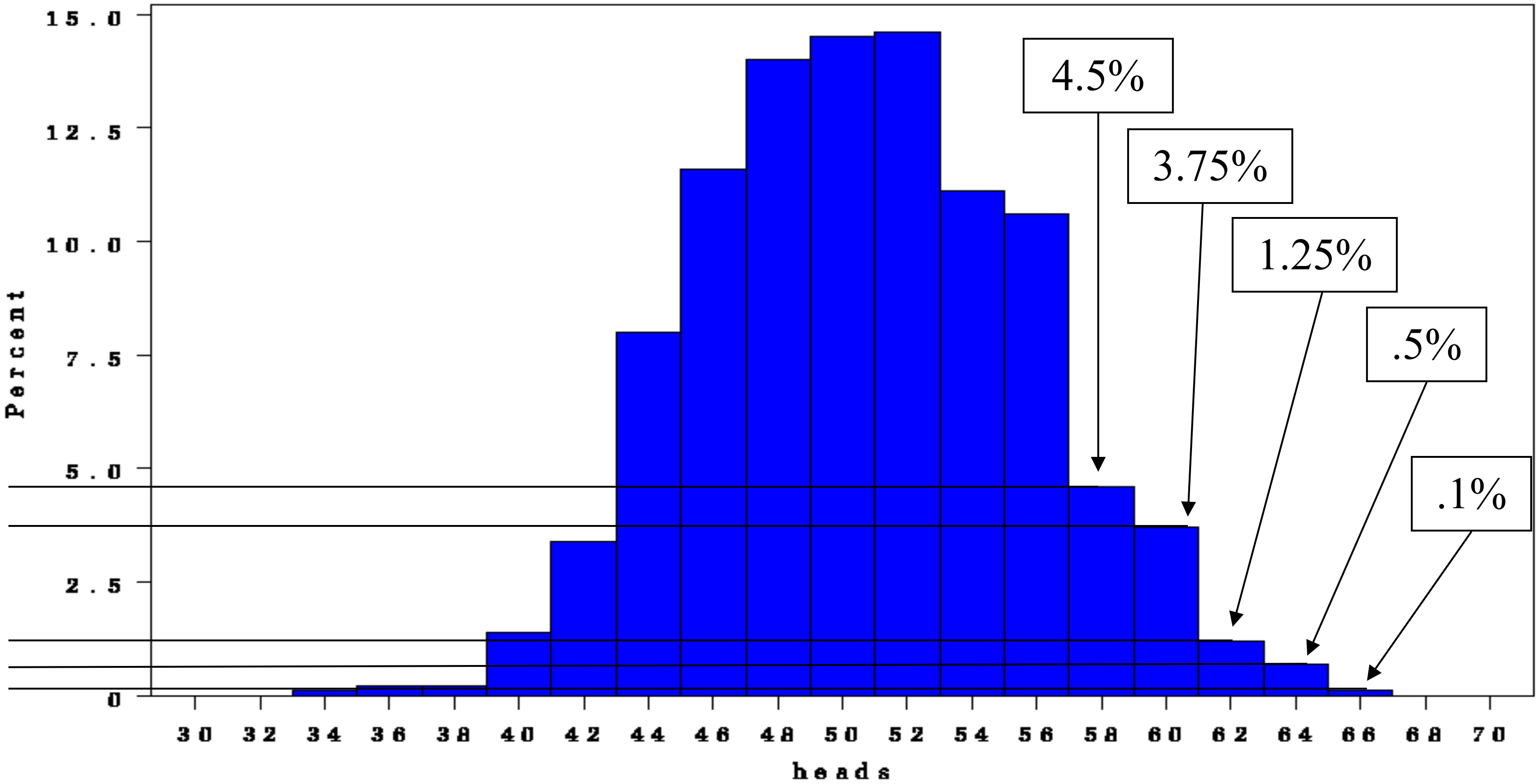
σ^2 is the variance of X

Importance of a Functional Form

- Since the normal distribution is defined by a function, we can determine the probability that a new value (# of heads in 100 tosses for the *new* coin) is part of the distribution

- Done by summing the area underneath the normal distribution, compared to the total area underneath the distribution (integral calculus)

Example



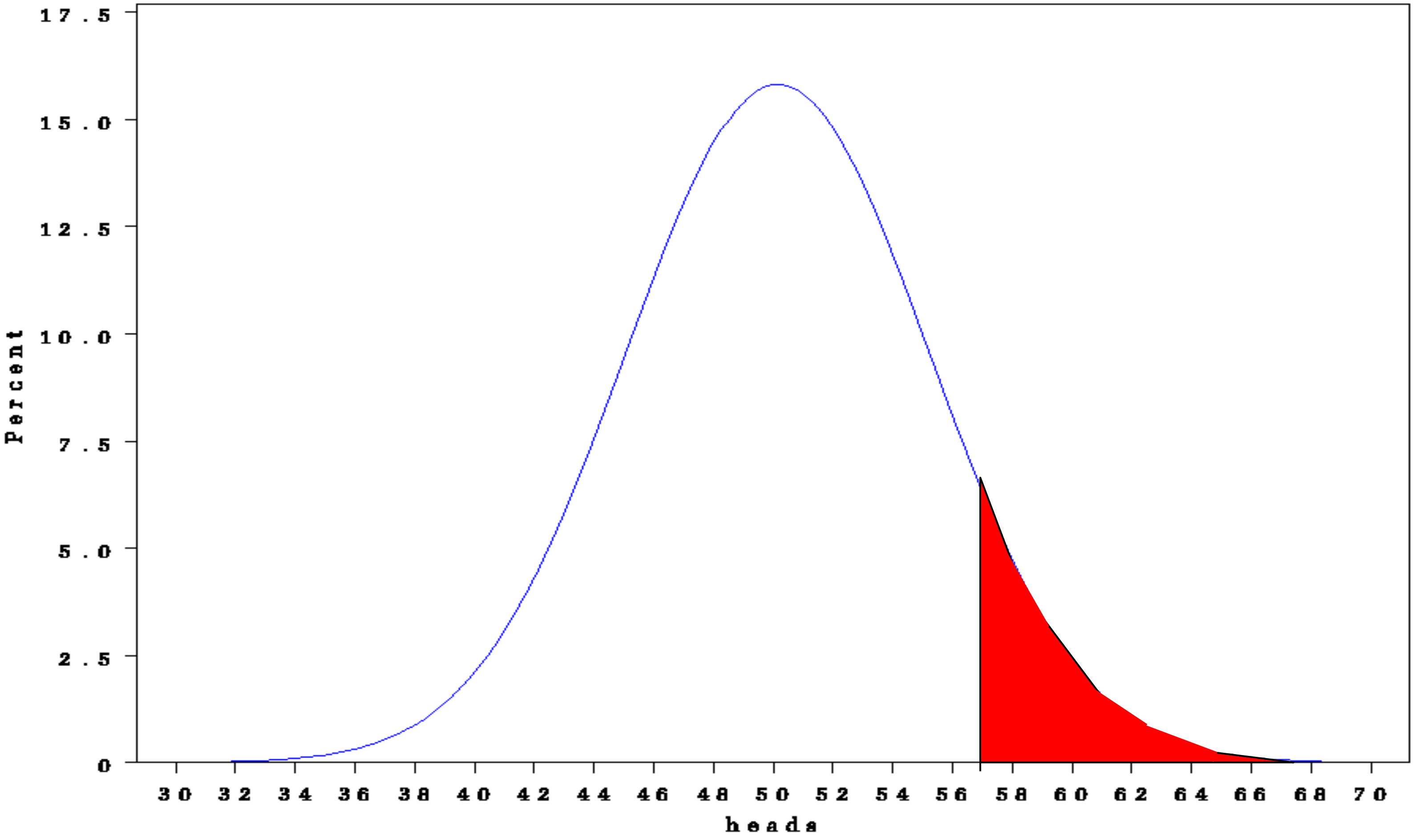
Based on our sampling distribution, what's the probability of 57 or 58 (bar labeled 58) heads in 100 coin tosses?

4.5% = .045

Based on our sampling distribution, what's the probability of 57 or more heads in 100 coin tosses?

4.5% + 3.75% + 1.25% + .5% + .1% = 10.1% = .101

Example



Based on a normal distribution with a mean of 50.15 and standard deviation of 5.05, what's the probability of 57 or more heads in 100 coin tosses?

Standard Normal Distribution

- Instead of integrating over each normal distribution, we transform each normal distribution to the **Standard Normal Distribution.**

- Properties
  - Mean equals 0, Standard Deviation equals 1 ($N(0,1)$)
  - Total area under standard normal curve is 1
  - Area under the curve represented in Z-Table (Table 3 at the back of the book)

- Transform observed scores to z-scores

Example

- Based on our sampling distribution (mean = 50.15, standard deviation = 5.05) what's the probability of having 57 or more heads?
  - Steps
    1. Transform the cut-score (e.g., 57) to a Z-score
    2. Use Table to determine the probability
    3. Interpret results

Step 1

- Calculate the Z-score

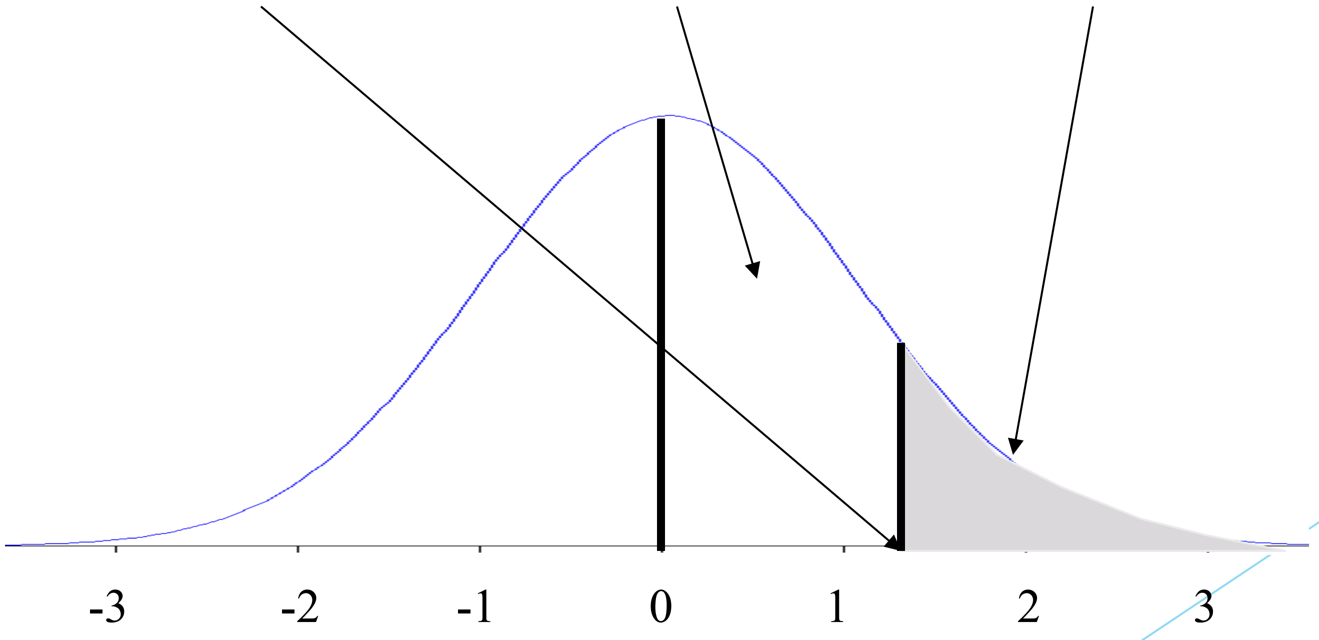
$$z_i = \frac{x_i - \mu}{\sigma} \qquad = \frac{57 - 50.15}{5.05} = 1.36$$

Interpretation:

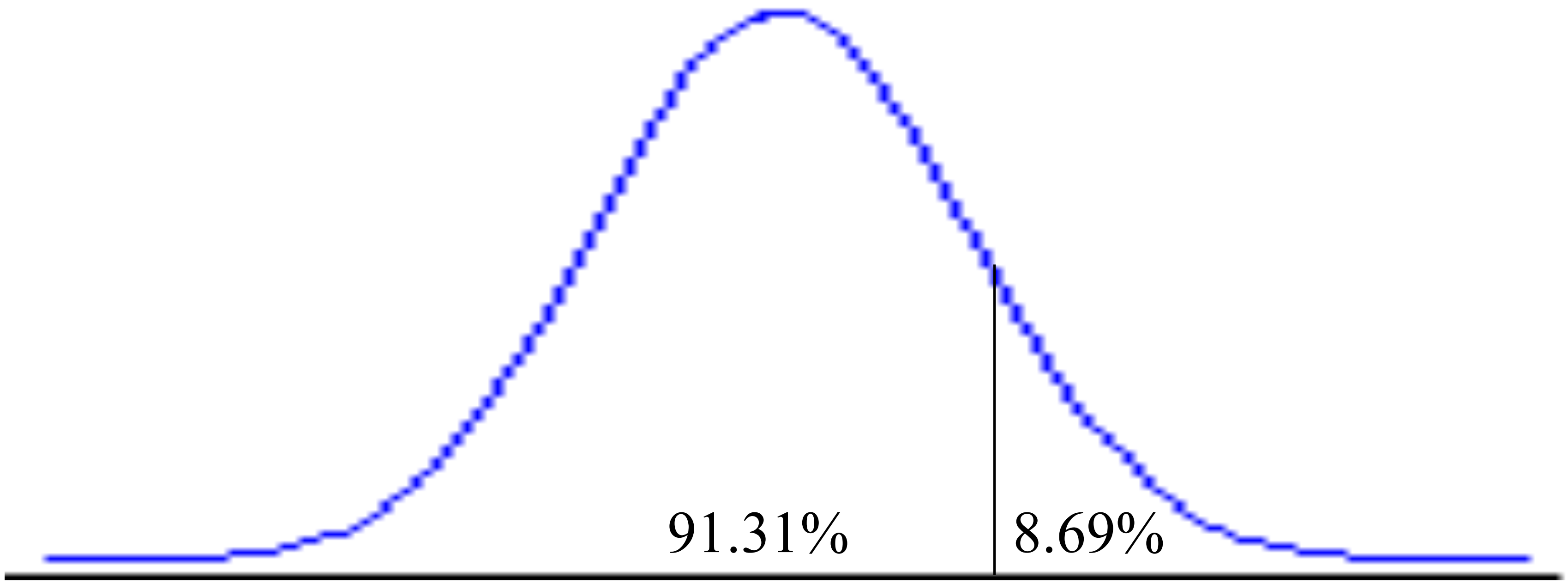
A score of 57 is 1.36 standard deviations above the mean.

Table 3 (Hand out and Back of Book)

z	% in Tail
1.36	8.69%



Step 3



X = 57
z = 1.36

Interpretation:

Only 8.69% of trials is expected to yield a score greater than or equal to 57 heads.

Additional Example (SAT Math Z-scores)

- The Math SAT has a mean of 600 and a standard deviation of 100.
  - Barbara scored a 610 and Billy scored a 540

1. What percent of the population did Barbara score greater than (percentile)?
2. What percent of the population did Billy score greater than (percentile)?
3. What percent of the population scored greater than Billy, but less than Barbara?

Step 1 – Convert to Z-Scores

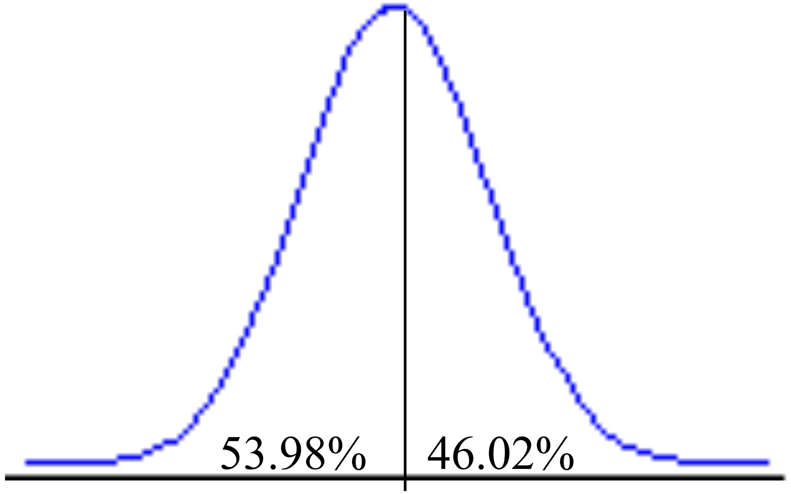
Barbara

$$z_i = \frac{x_i - \mu}{\sigma} = \frac{610 - 600}{100} = 0.10$$

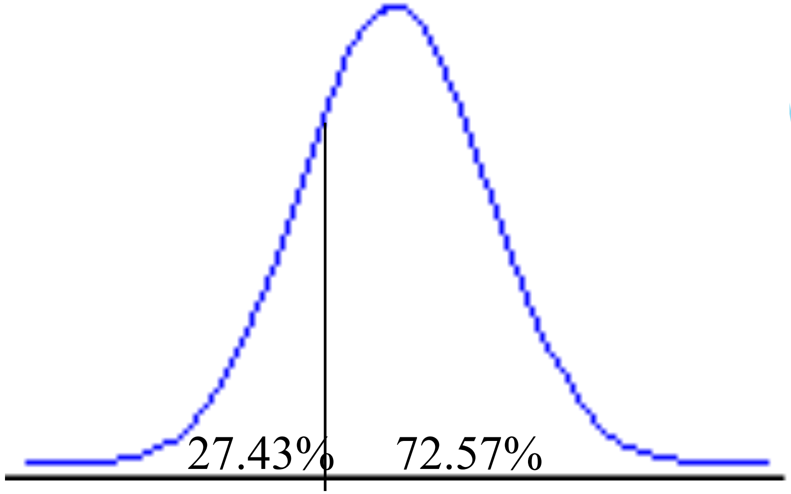
Billy

$$z_i = \frac{x_i - \mu}{\sigma} = \frac{540 - 600}{100} = -0.60$$

Step 3



X = 610
z = .10



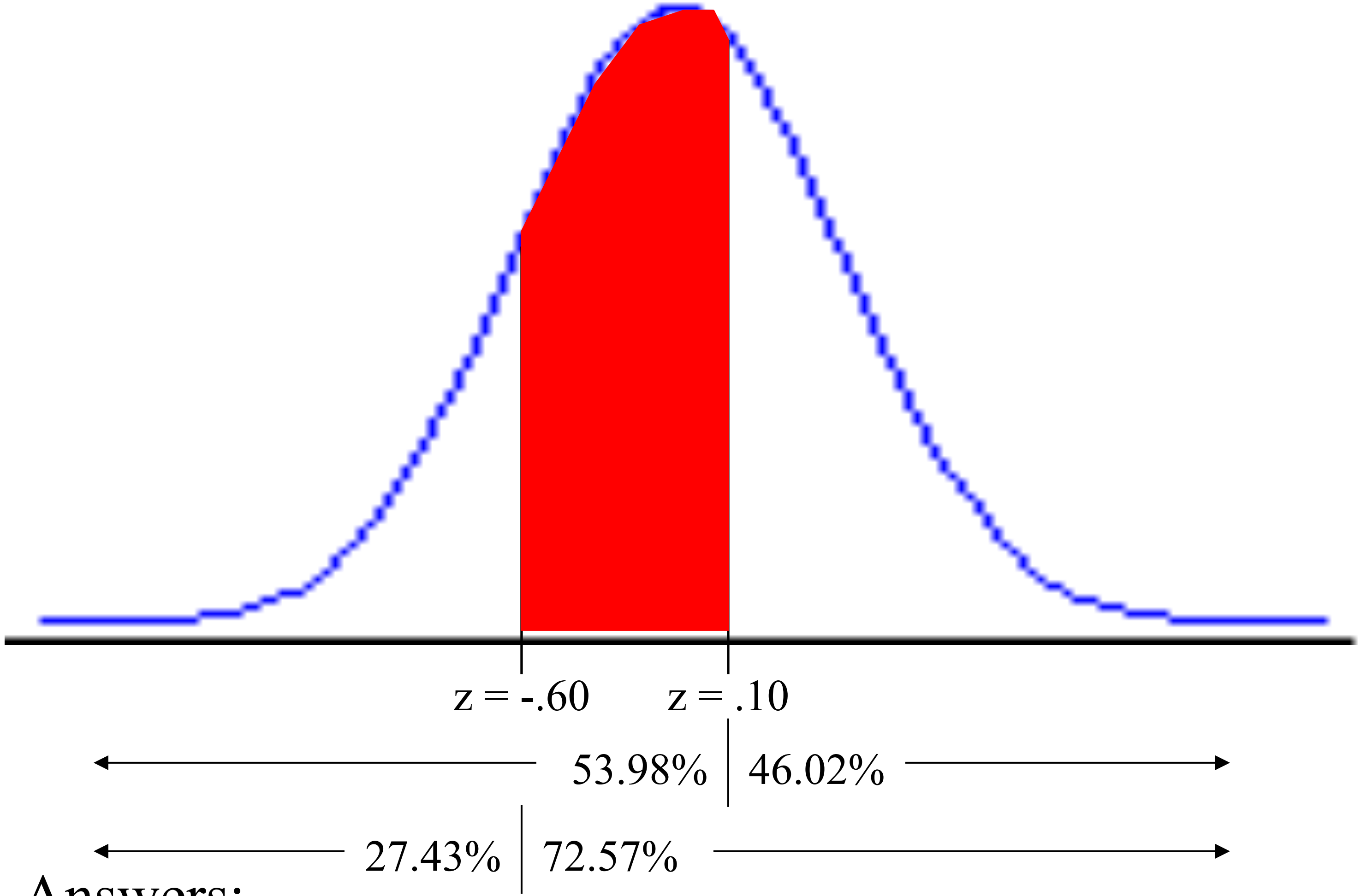
X = 540
z = -.60

Answers:

1. Barbara scored better than 53.98% of the population (~54 percentile)

2. Billy scored better than 27.43% of the population (~27th percentile)

Step 3



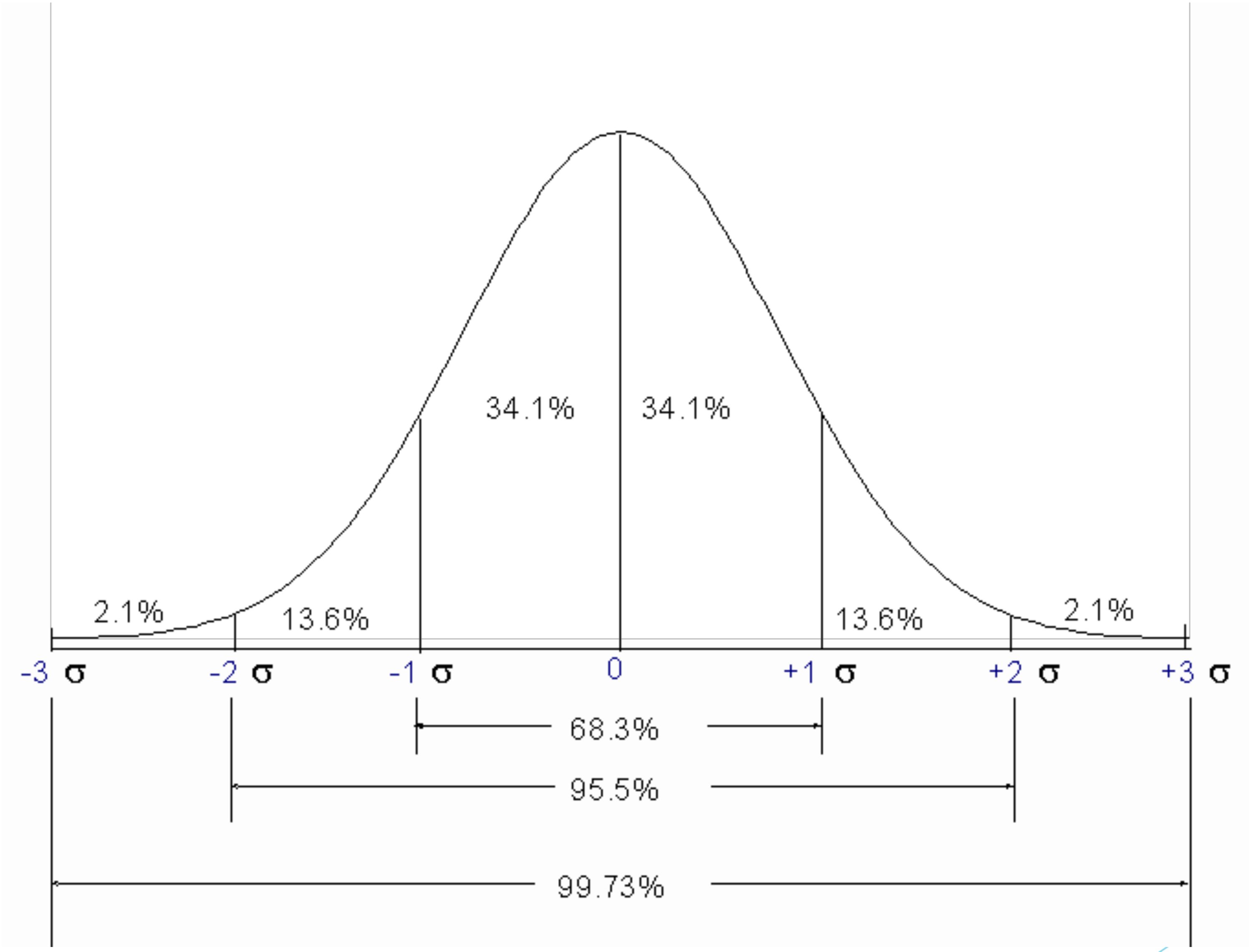
Answers:

3. (53.98-27.43) = 26.55% of the population scored better than Billy but worse than Barbara.

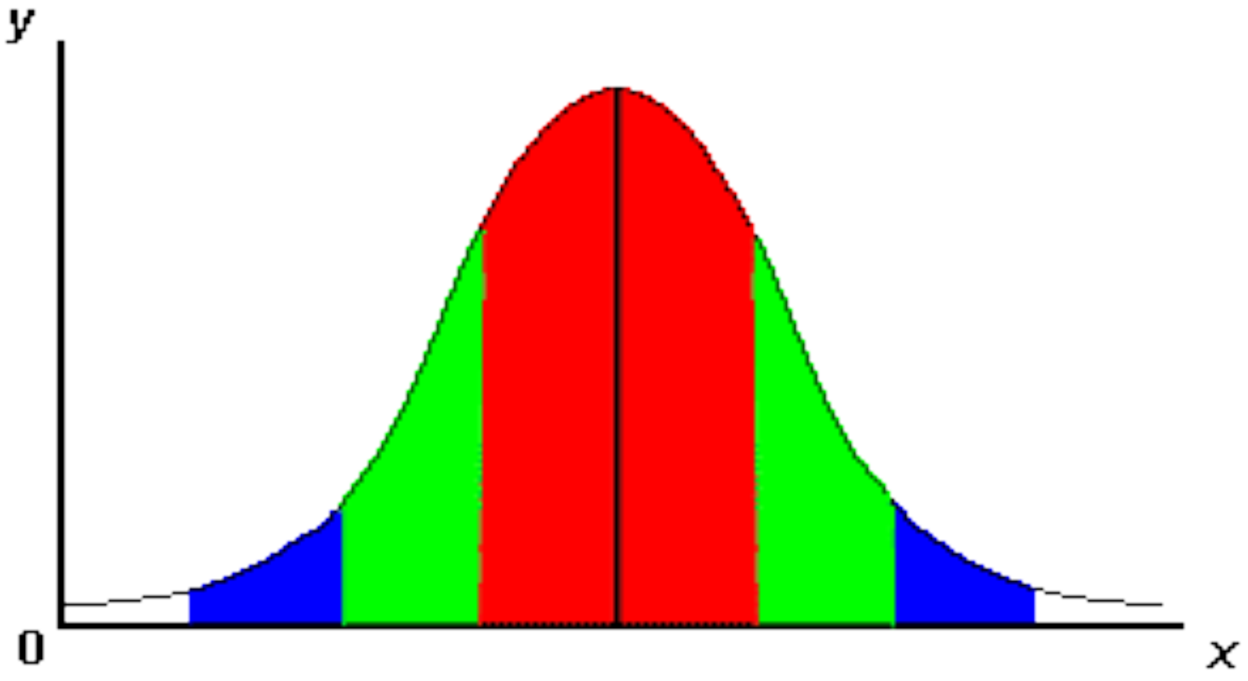
Characteristics of Z-scores

- Mean of 0, Standard Deviation of 1
- Scores are in standard deviation units from the mean
- Almost all of the data in a standard normal distribution will be within 3 SDs of the mean
- Extreme scores (outliers) will have large z-scores
  - Greater than +3 or less than -3

Properties of the Standard Normal Curve



"68-95-99+" Rule



Red = 68% of the area (± 1 standard deviation)

Red + Green = 95% of the area (± 2 σs)

Red + Green + Blue = 99+% of the area (± 3 σs)

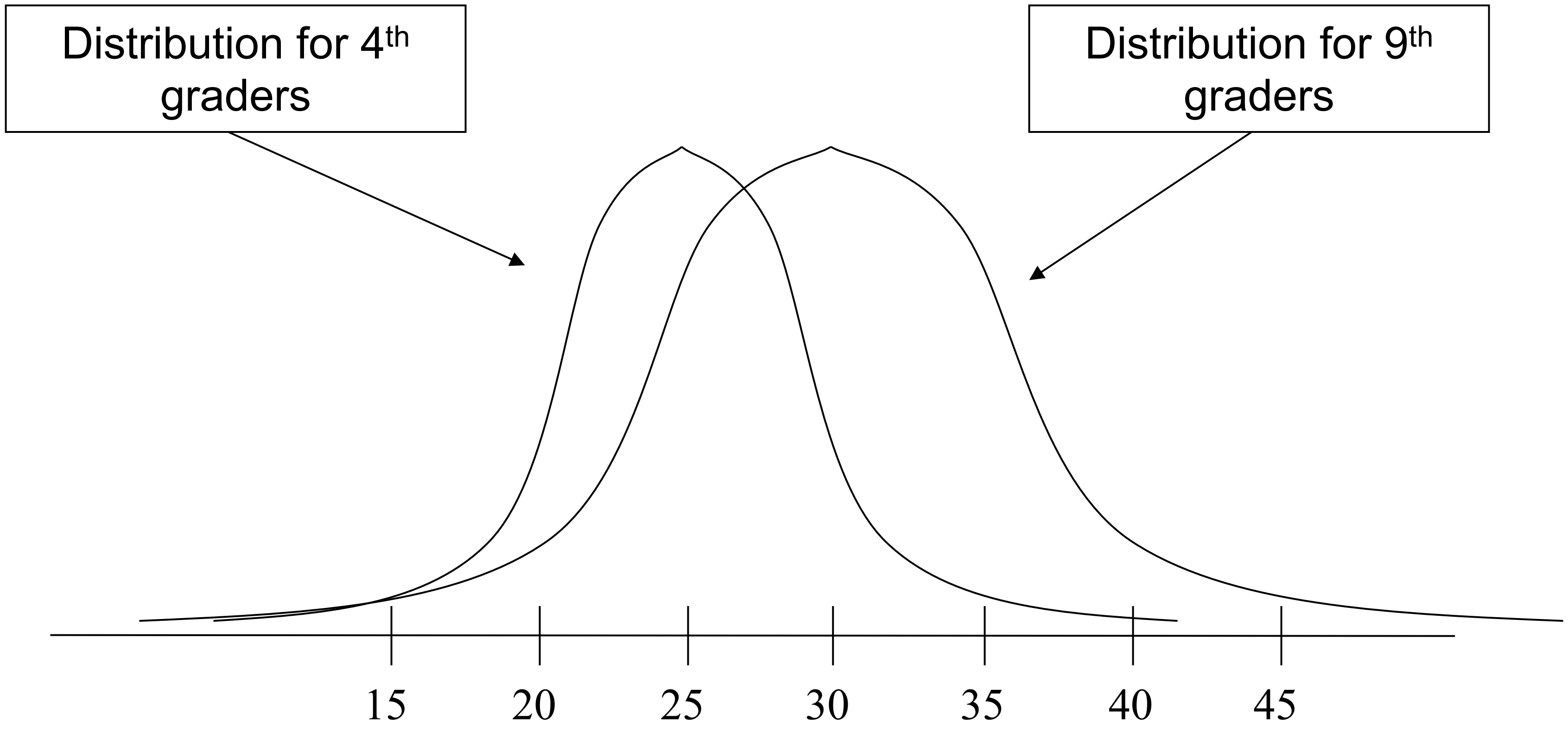
Summary

- Transformation are common
- Z-Score Transformation
  - Data transformation technique to make the mean 0 and the standard deviation 1
  - An individual z-score represents how many standard deviations the individual's score is above or below the mean
  - Determine probability that a score falls above or below a given value or within a range of values (using the tables in appendix)
- The Normal Distribution
  - Allows us to more easily calculate probabilities
- Standard Normal Distribution
  - Normal Distribution with a mean of 0 and a standard deviation of 1

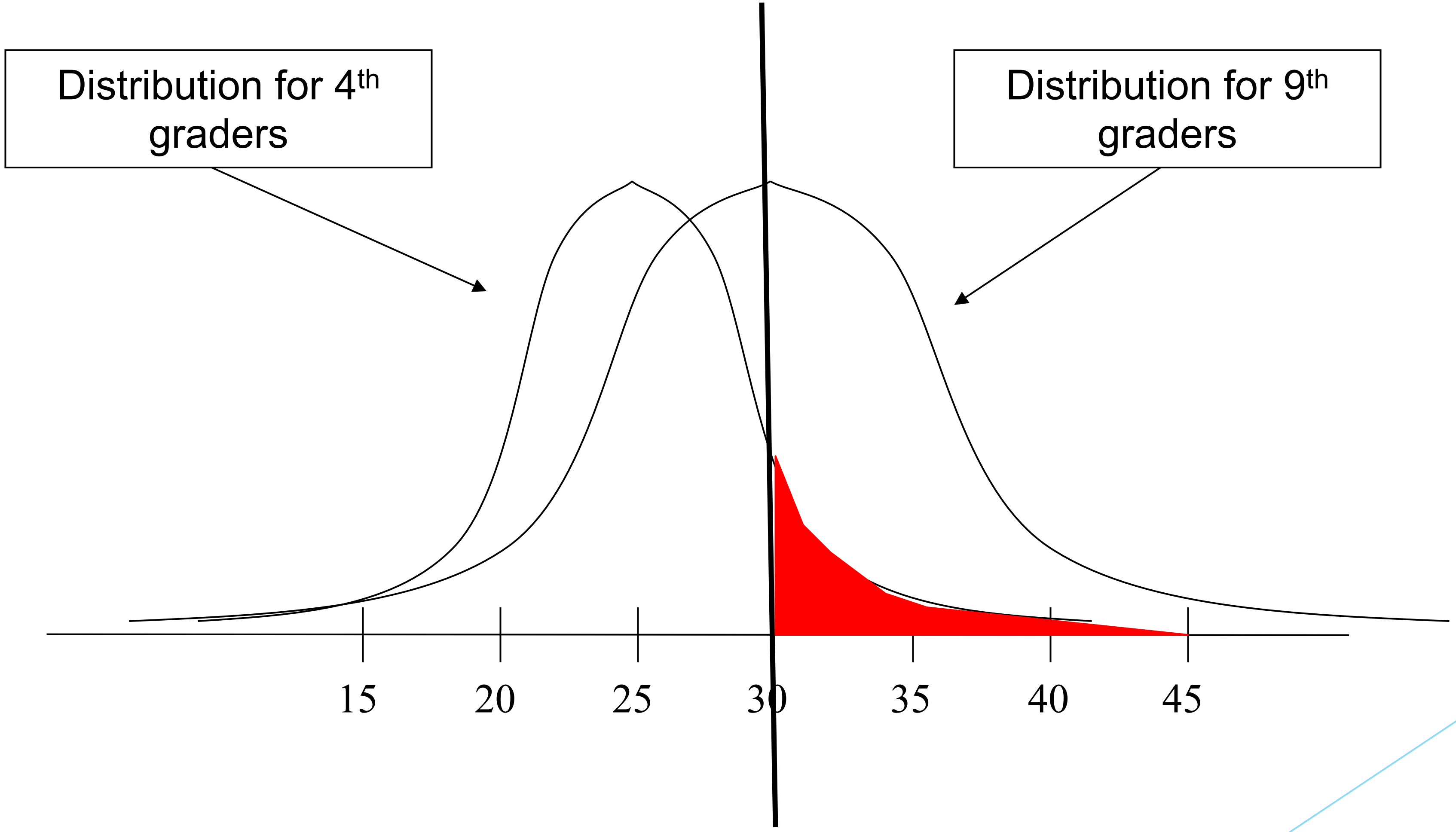
Example 2

- A set of reading scores for fourth-grade children has a mean of 25 and a standard deviation of 5. A set of scores for ninth-grade children has a mean of 30 and a standard deviation of 10. Assume that the distributions are normal.
  - Draw a rough sketch of these data, putting both groups in the same figure
  - What percentage of fourth graders score better than the average ninth grader?
  - What percentage of ninth graders score worse than the average fourth grader?

(A) Rough Sketch



(B) What percentage of 4th graders score better than the average 9th grader?



Rephrase Question

- What percent of 4th graders score better than a 30?

- Step 1: Convert 30 to Z-score based on 4th grade distribution

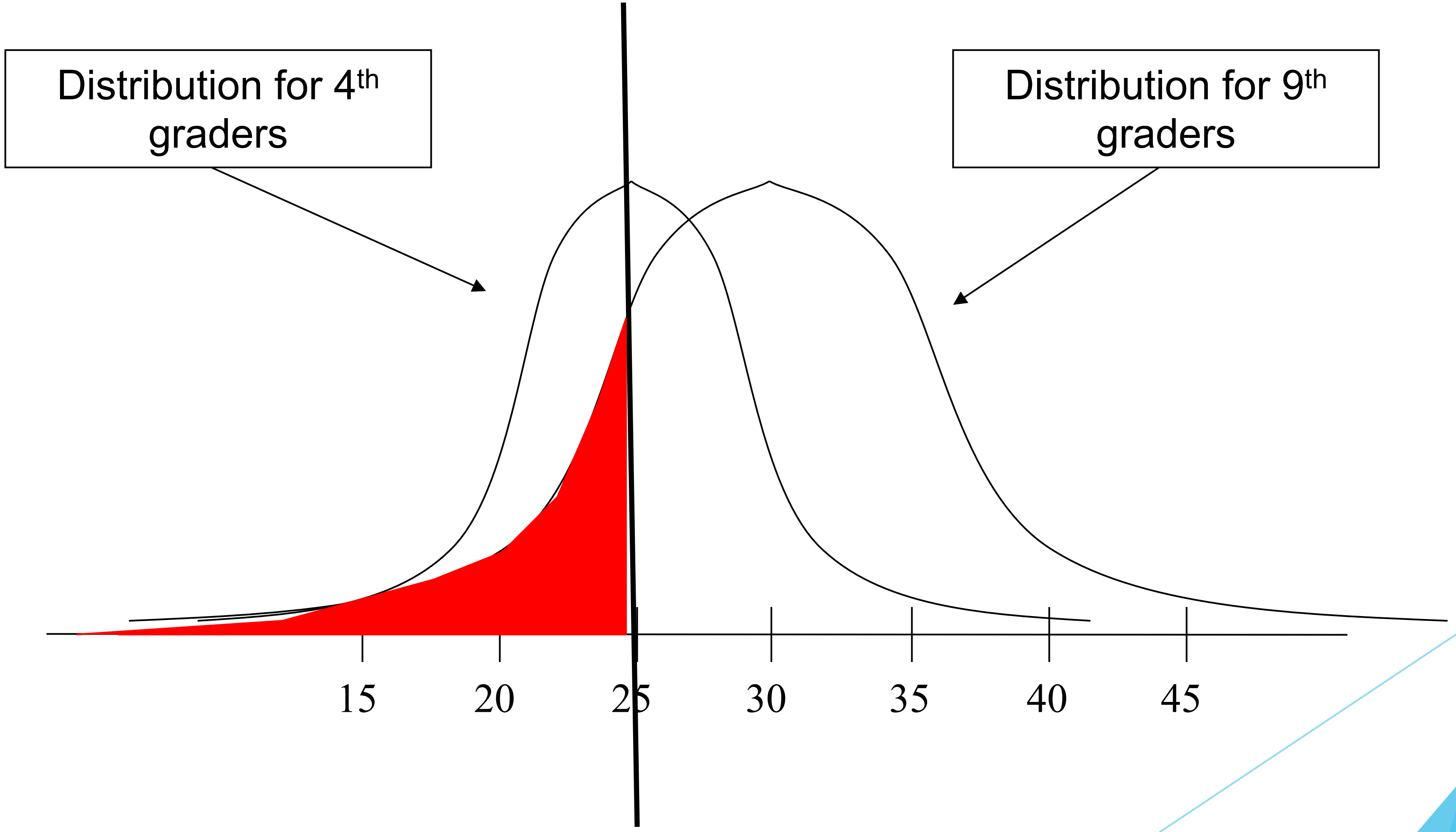
$$z_i = \frac{x_i - \mu}{\sigma} = \frac{30 - 25}{5} = 1.00$$

Step 2 & 3

z	% in Tail
1.00	15.87

- 15.87% of 4th graders score better than the average 9th grader.

(C) What percentage of 9th graders score worse than the average 4th grader?



Rephrase Question

- What percent of 9th graders score worse than a 25?

- Step 1: Convert 25 to Z-score based on 9th grade distribution

$$z_i = \frac{x_i - \mu}{\sigma} = \frac{25 - 30}{10} = -.50$$

Step 2 & 3

z	% in Tail
0.50	30.85

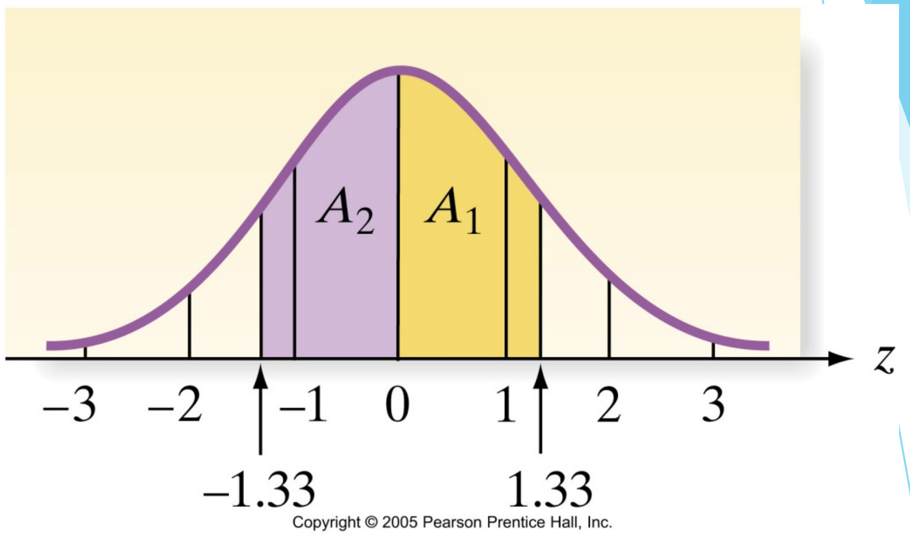
- 30.85% of 9th graders score worse than the average 4th grader.

Example 3

- What is $P(-1.33 < Z < 1.33)$?

Example 3

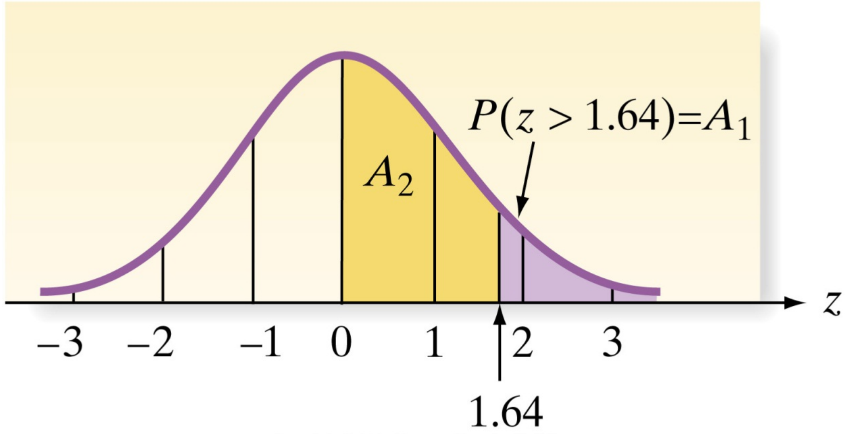
- What is P(-1.33 < Z < 1.33)?



P(-1.33 < Z < 1.33) = .9082 – 0.0918= .8164

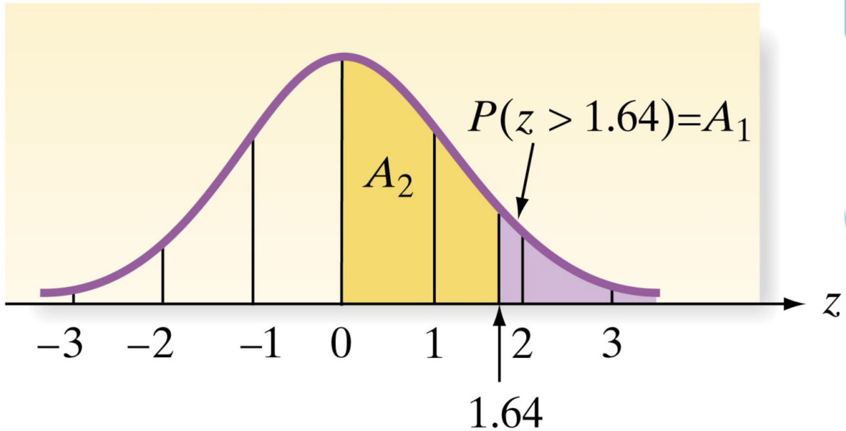
Example 4

- What is $P(Z > 1.64)$?



Example 4

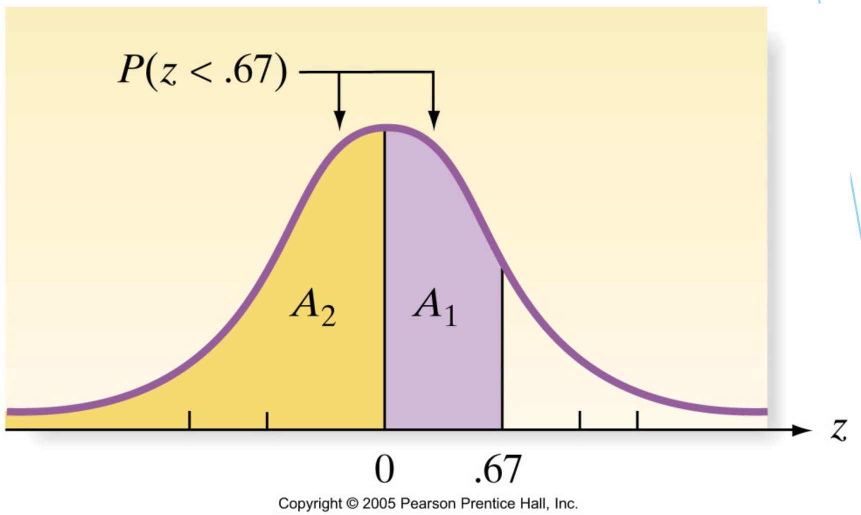
- What is P(Z > 1.64)?



$P(Z > 1.64) = 1-0.9495= .0505$

Example 5

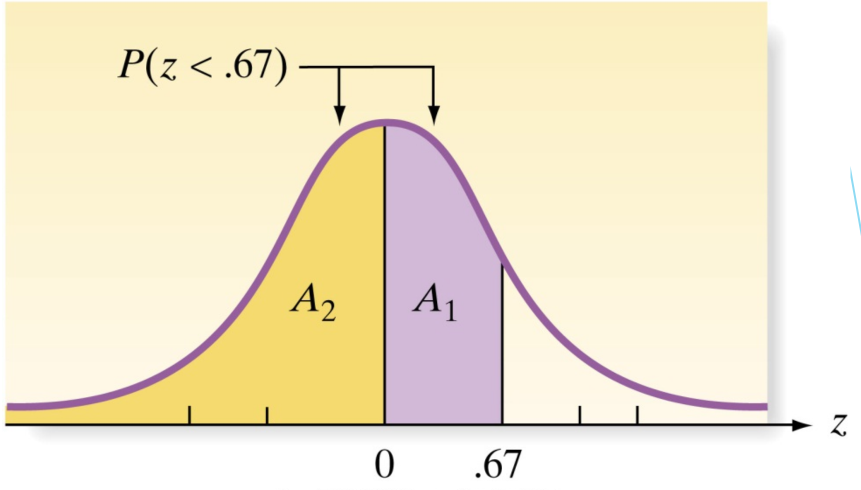
What is $P(Z < .67)$?



Example 5

What is P(Z < .67)?

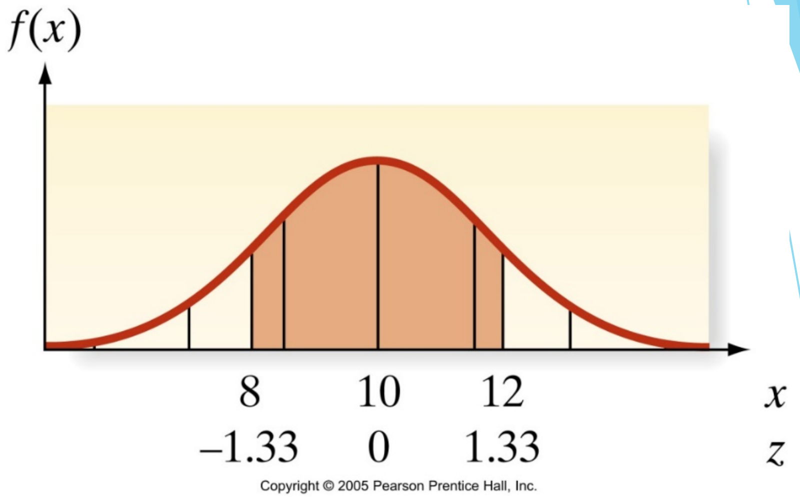
P(Z < .67) = .7486



Example 6

What if values are not normalized

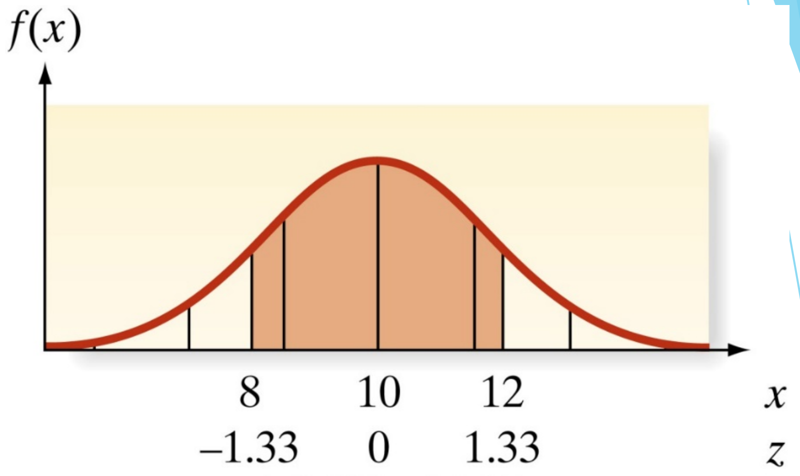
Calculate P (8<x<12), with μ=10 and σ=1.5



Example 6

What if values are not normalized

Calculate P (8<x<12), with μ=10 and σ=1.5

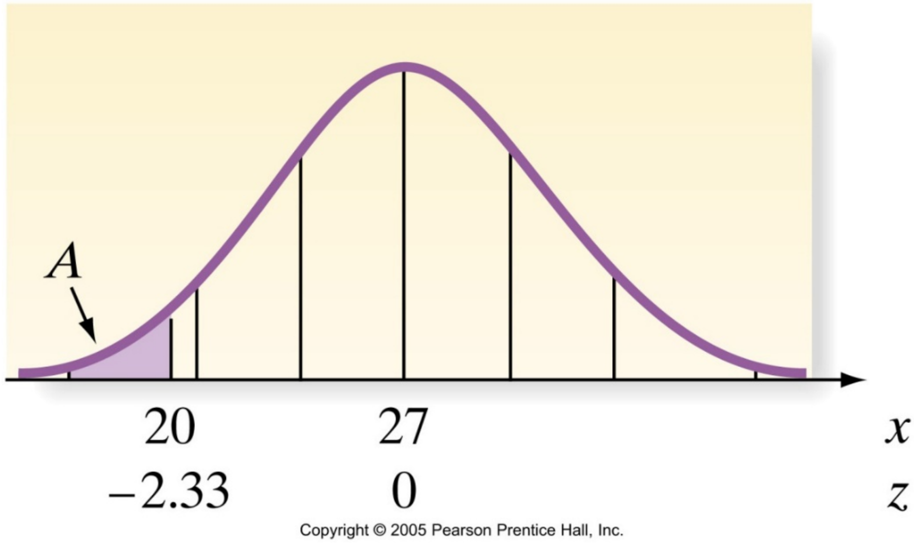


Convert to standard normal using

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

P(8<X<12) = P(-1.33<Z<1.33) = 0.9082 – 0.0918 = .8164

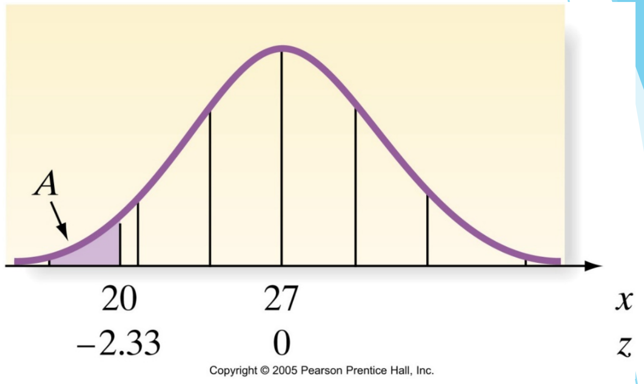
Example 7 – Making an Inference



How likely is an observation in area A, given an assumed normal distribution with mean of 27 and standard deviation of 3?

Example 7 – Making an Inference

How likely is an observation in area A, given an assumed normal distribution with mean of 27 and standard deviation of 3?



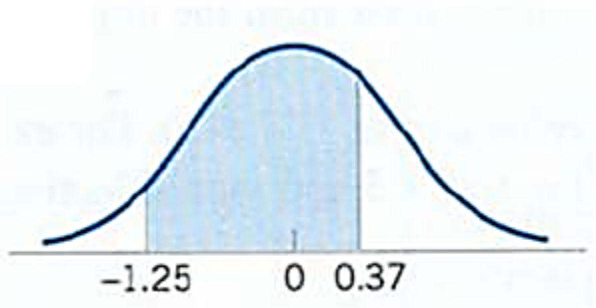
Z-value for x=20 is -2.33

P(x<20) = P(Z<-2.33) =.0099

You could reasonably conclude that this is a rare event.

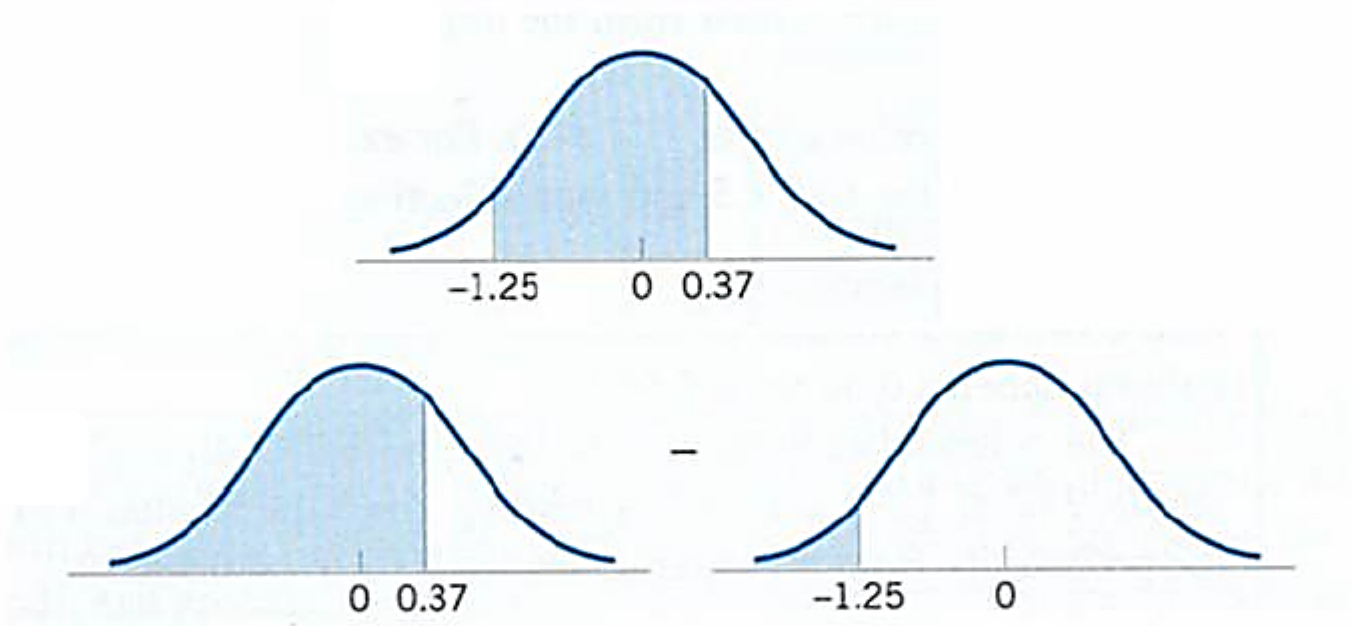
Example 8

Determine the probability that $P(-1.25 < Z < 0.37)$



Example 8

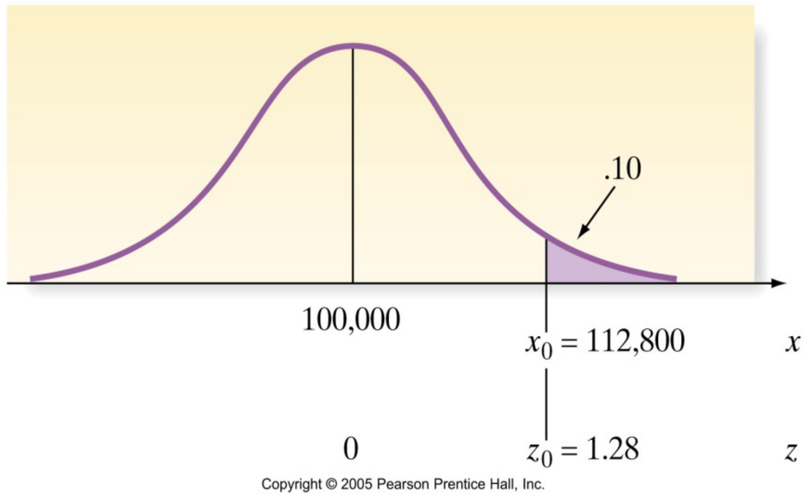
$$P(-1.25 < Z < 0.37) = p(Z < 0.37) - P(Z < -1.25)$$



P(Z<0.37)=0.644309

P(Z<-1.25)=0.105650

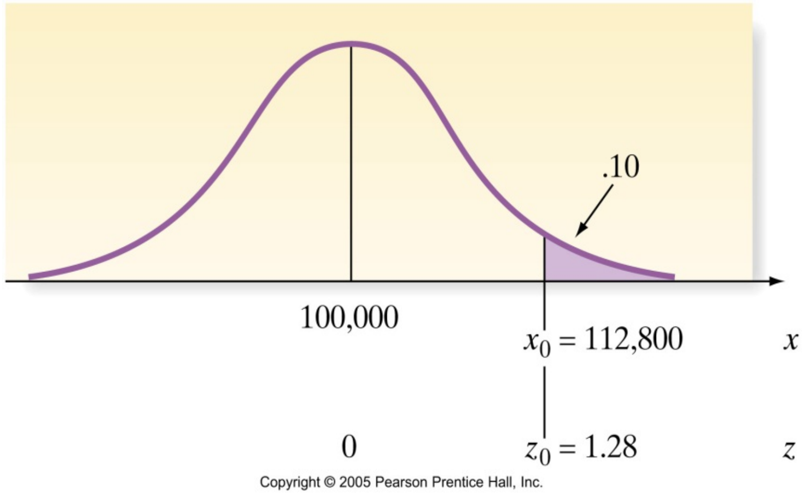
Example 9



Given a normally distributed variable x with mean 100000 and standard deviation of 10000, what value of x identifies the top 10% of the distribution?

Example 9

- Given a normally distributed variable x with mean 100000 and standard deviation of 10000, what value of x identifies the top 10% of the distribution?



$$P(X \le x_0) = P\left(Z \le \frac{x_0 - \mu}{\sigma}\right) = P\left(Z \le \frac{x_0 - 100,000}{10,000}\right) = .90$$

The z value corresponding with .40 is 1.28. Solving for x_0

x_0 = 100,000 +1.28(10,000) = 100,000 +12,800 = 112,800

Example 10

A very large group of students obtain test scores that are normally distributed with mean 60 and standard deviation 15. What proportion of the students obtained scores between 85 and 95?

Example 10

A very large group of students obtains test scores that are normally distributed with mean 60 and standard deviation 15. What proportion of the students obtained scores between 85 and 95?

$$P(85 < X < 95) = P\left(\frac{85-60}{15} < Z < \frac{95-60}{15}\right)$$

$$= P(1.67 < Z < 2.33)$$

$$= F(2.33) - F(1.67)$$

$$= 0.9901 - 0.9525 = 0.0376$$