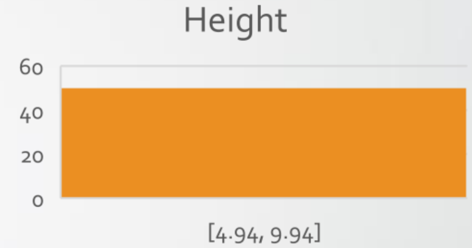
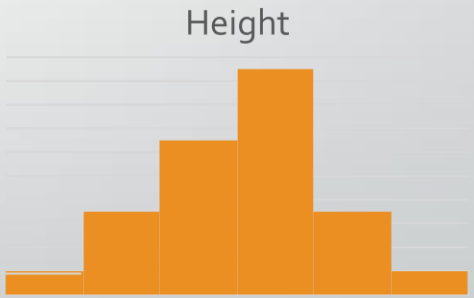
Normal Data

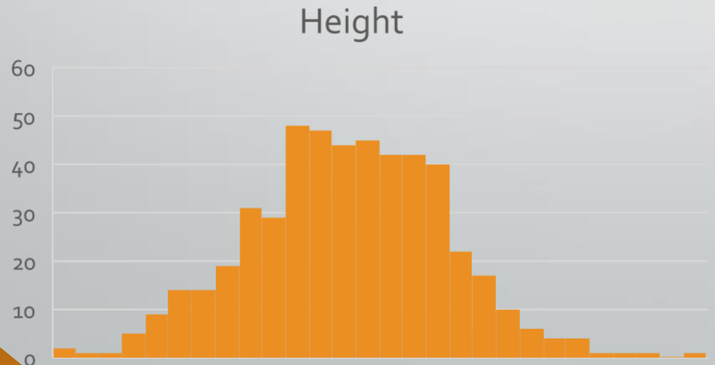
ID1050– Quantitative & Qualitative Reasoning

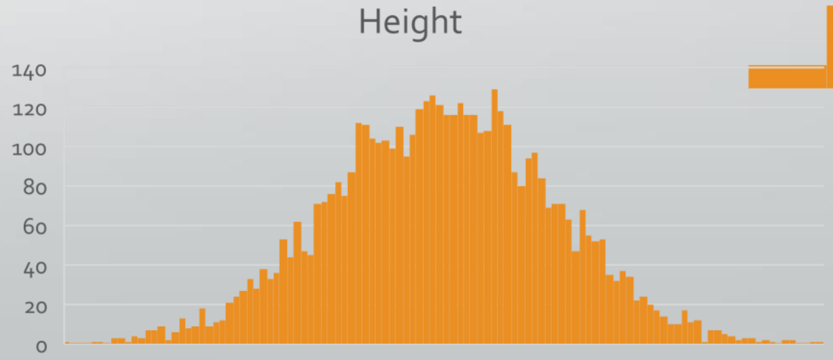
Histogram for Different Sample Sizes

- For a small sample, the choice of class (group) size dramatically affects how the histogram appears.
  - Say we're measuring heights of a group of 50 students.
    - If our classes are too wide, everyone fits into one bin.
    - If our classes are too narrow, each bin will have too few members.
    - If our classes are just right, we see a normal distribution
- As the sample gets bigger, we can have narrower classes and still see the normal distribution



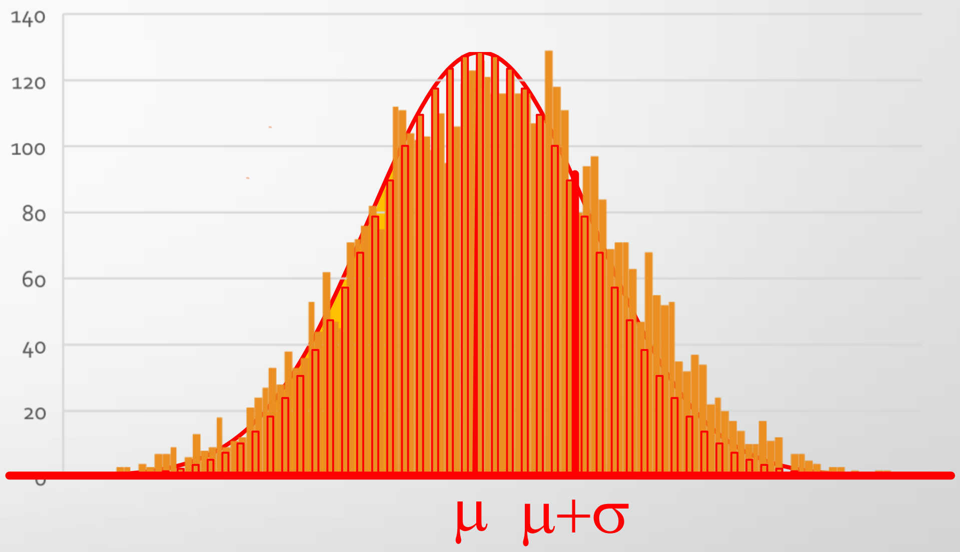






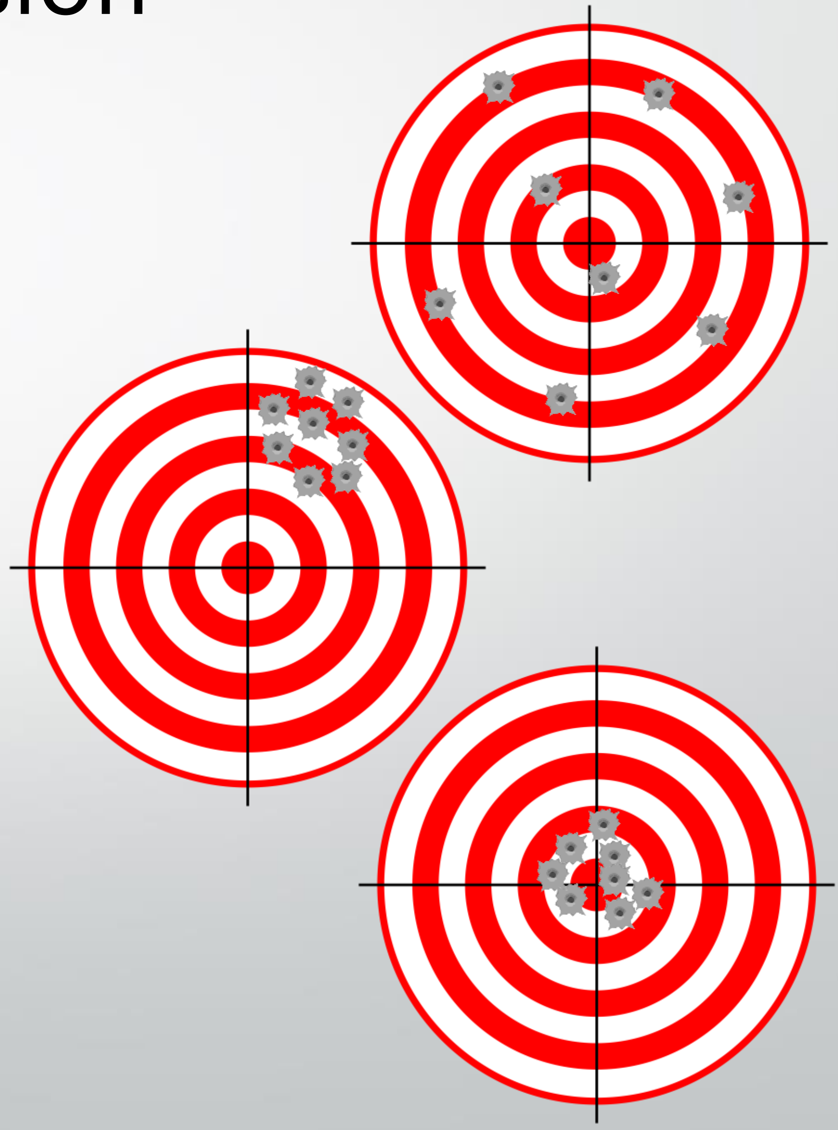
The Normal Curve

- We can replace the bars with just the curve across their tops
- In the ideal case, we get the Normal Curve (also called the 'bell curve' or the 'Gaussian curve'.
- Some properties of the ideal normal curve:
  - It has left-right symmetry about its middle.
  - 100% of population is under the curve.
  - The area under any part of the curve is directly related to the fraction of population in that region.
  - The left and right tails of the curve approach, but never cross, the x-axis.
  - The curve has a mathematically definition:
    - $y = \frac{1}{\sigma\sqrt{2\pi}} * e^{\frac{-(x-\mu)^2}{2\sigma^2}}$
  - There is a point where the curve changes from a downward curvature to an upward curvature. This is at 1 standard deviation (σ) above (and below) the middle, or average (μ).



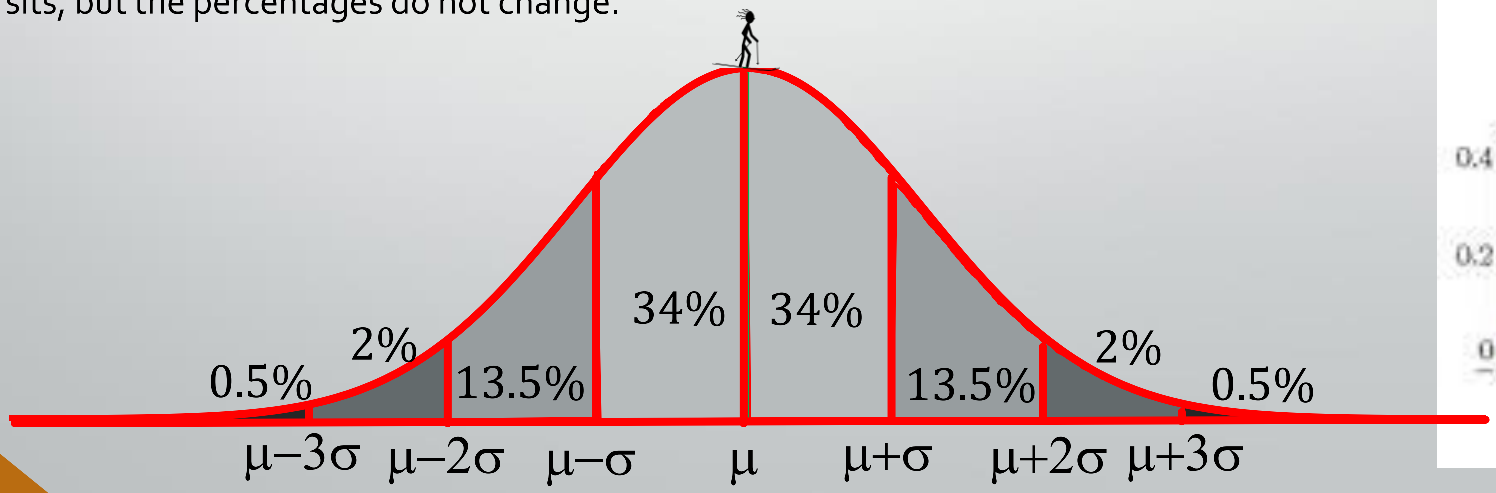
Accuracy vs. Precision

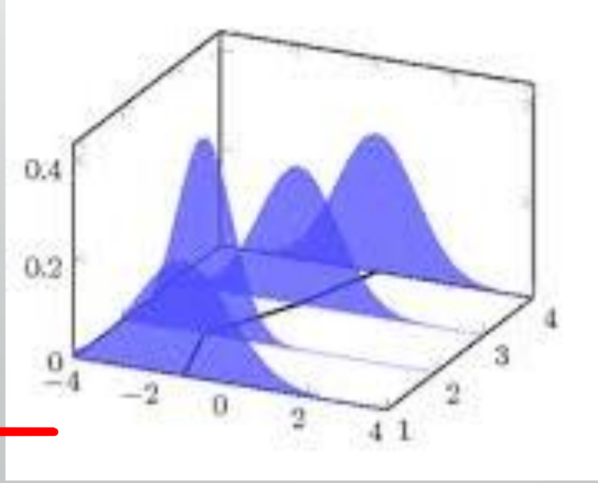
- Let's use the analogy of firing a gun at a target to illustrate the ideas of accuracy and precision.
- On one day, our target looks like this: we are hitting the target, but the holes are all over it. We have good accuracy, but low precision.
- On another day, our target looks like this: the holes are clustered close together, but they are not near the bulls-eye. We have good precision, but bad accuracy.
- On the last day, our target looks like this: we have both good accuracy and good precision.
- In statistics:
  - Sample bias leads to poor accuracy
  - Insufficient sample size leads to low precision



Normal Curve and Standard Deviation

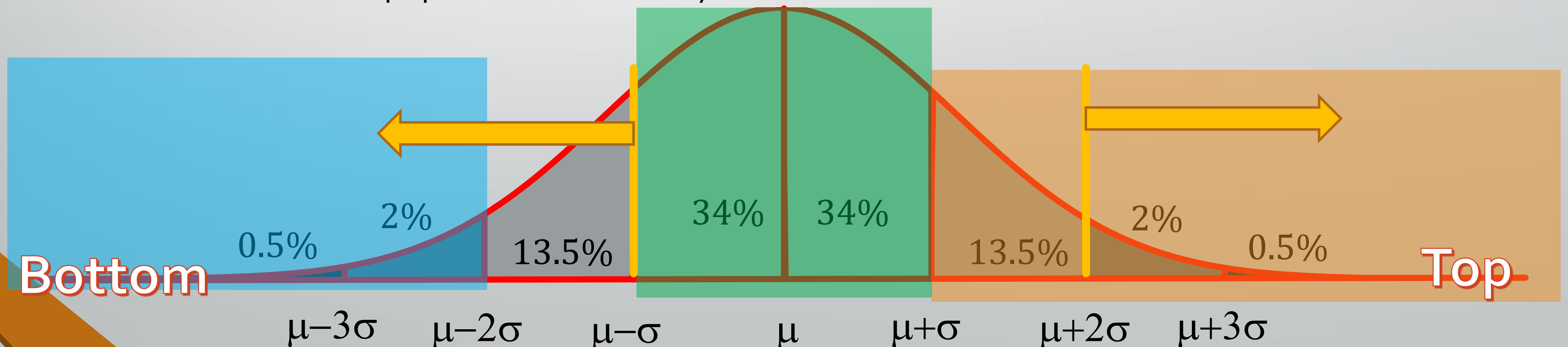
- Imagine the normal curve is a snowy hill
  - A skier at the top is standing where the hill has a downward curve. When the skier is near the bottom, the hill has begun to curve upward, toward the sky.
  - The point on the hill where the curvature changes from up to down (and where the slope is steepest) is at one standard deviation away from the mean.
  - Draw vertical lines at the mean, at one standard deviation left and right of the mean, and then at two and three times the standard deviation, both left and right.
- Using the equation for the normal curve, you could calculate the percentages (or fraction of the population) between these boundaries.
  - For every normal curve, these percentages will always the be same!
  - The standard deviation governs the general shape (thin, thick, etc.) and the mean determines where the center of the curve sits, but the percentages do not change.





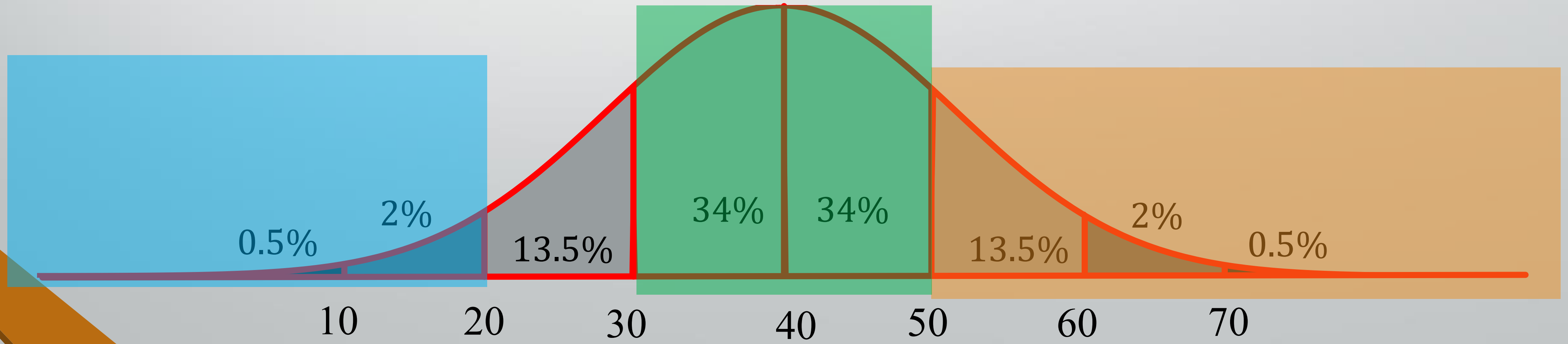
How can we use this information?

- Types of questions we can answer
  - “What fraction of the population is above (to the right of) below (to the left of), or between boundaries?”
  - “How many in the population is above, below, or between boundaries?”
  - “What is the least x-value (along the horizontal axis) required in order to be in some <u>top</u> fraction of the population?”
  - “What is the greatest x-value required in order to be in some <u>bottom</u> fraction?”
  - Percentile questions – “What percent of the population is below an x-value?”
- A question we can’t answer using this method:
  - “What fraction of the population had <u>exactly</u> some x-value?”



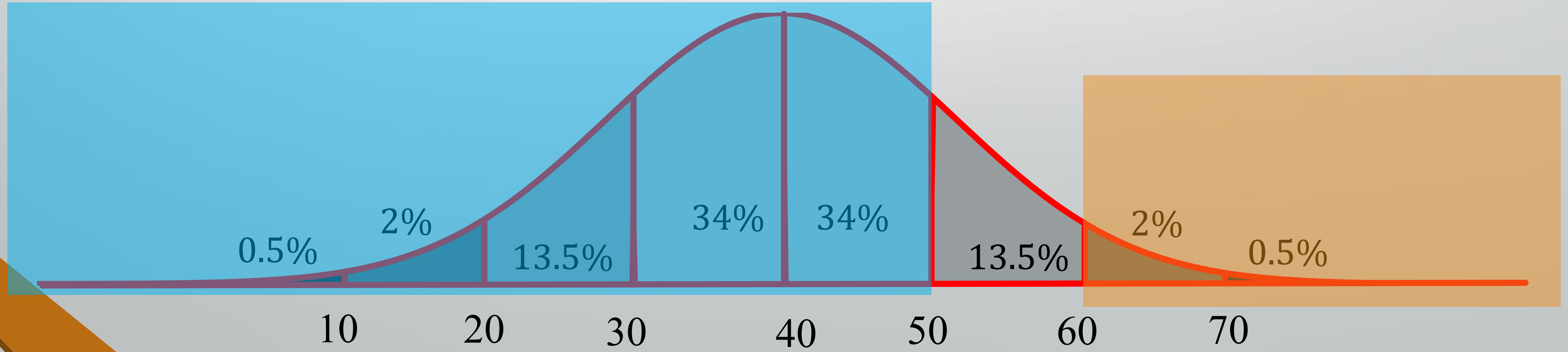
Given a Boundary, What Percentage?

- First, label the x-axis from the information given about the mean (μ) and standard deviation (σ).
  - For these examples, let's assume μ=40 and σ=10. We get the following labels along the x-axis.
- To answer the 'above, below, between' type of questions, we simply add up the percentages in the desired regions.
  - Example: "What percentage of the population is above 50?" Answer: 13.5%+2%+0.5%=16%
  - Example: "What percentage of the population is below 20?" Answer: 2%+0.5%=2.5%
  - Example: "What percentage of the population is between 30 and 50?" Answer: 34%+34%=68%
  - Note: 2/3 of the population is within 1σ of the mean, 95% is within 2σ of the mean, and 99% is within 3σ.



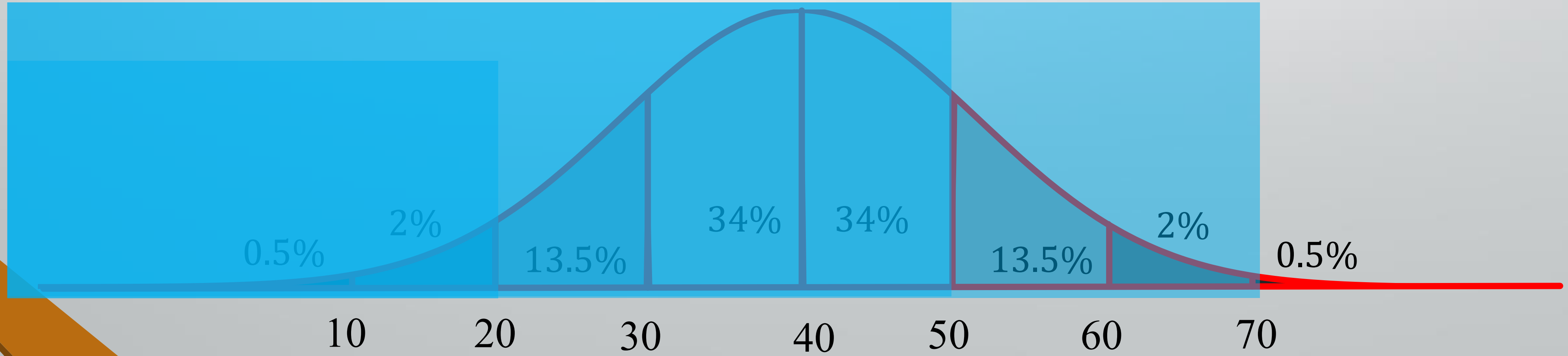
Given a Percentage, What Boundary?

- This is the converse of the previous types of questions. Here we are given a percentage, and we need to find the boundary(s) that give us that percentage.
  - Note: Only certain percentages can be given since our boundaries are limited and the percentages between them are fixed.
  - For these examples, let's again assume μ=40 and σ=10. Labels the x-axis using these values.
  - Example: "What x-value has 2.5% above it?"
    - Answer: 60 (sliding in from the right, we have 0.5% above 70, and 2%+0.5%=2.5% above 60)
  - Example: "What value has 84% below it?"
    - Answer: 50 (sliding in from the left, when we reach 50, we've added 0.5%+2%+13.5%+34%+34%=84%)



Percentile

- Percentile is a way of gauging where in the population a particular x-value appears.
  - The percentile is the percent of the population below the give x-value.
  - It is the percent of the population that that x-value beats.
  - If a value is at the 50th percentile, then that score is the average.
  - Lower percentiles lie below the average, higher percentiles lie above the average.
- For these examples, let's assume $\mu=40$ and $\sigma=10$.
  - An x-value of 20 is at the 2.5th percentile.
  - An x-value of 50 is at the 84th percentile.
  - An x-value of 70 is at the 99.5th percentile.

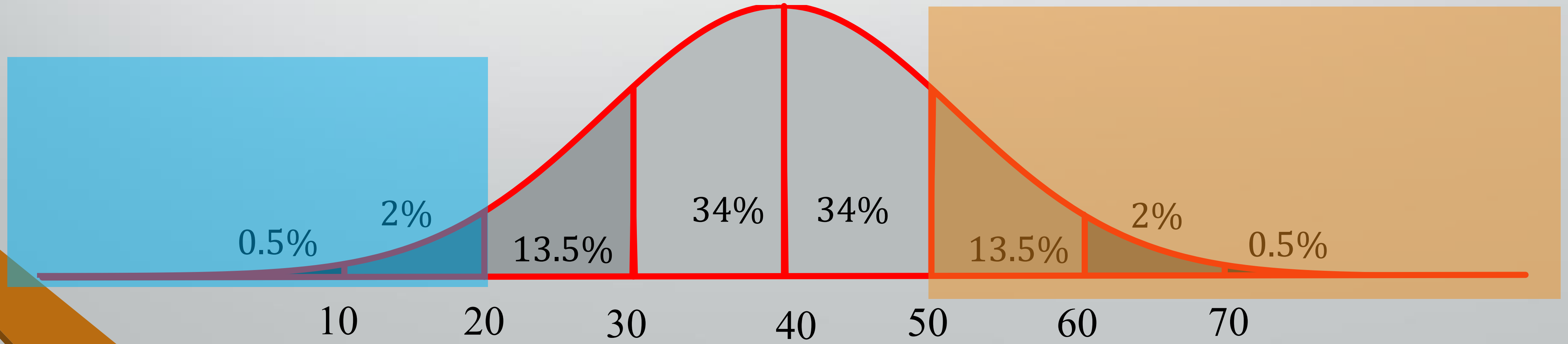


Converting Between Percentage and Fraction

- Some questions call for a *fraction* of the population instead of the *percentage*. This is an easy conversion:
  - Divide the percentage by 100 to get the fraction (or move the decimal point 2 positions to the left).
  - Example: $34\% = 0.34$ (out of 1, or 100% of, the whole population)
  - Example: $84\% = 0.84$
- Converting the other way is also easy:
  - Multiply the fraction by 100 to get the percentage (or move the decimal 2 positions to the right.)
  - Example: $0.025 = 2.5\%$
- The *chance* or *probability* of being in some portion of the population is the same as the *fraction* of that population.

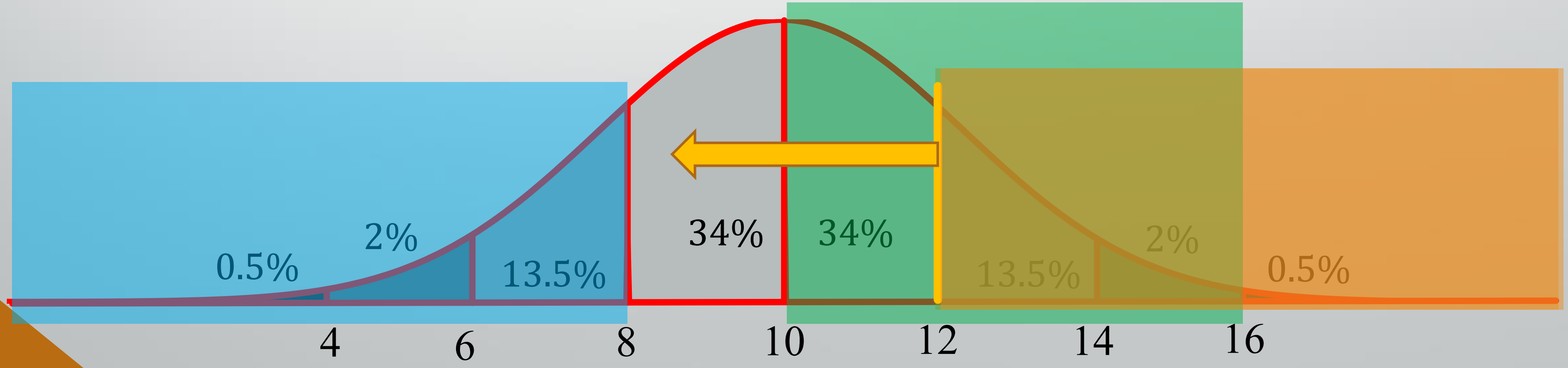
Converting Percentage or Fraction to 'How Many'

- If we are given a population size, N, or how many individuals there are in the population, we can also answer questions involving "How many of the population…?", not just percentages. Calculating this is simple:
  - 'How many' = (fraction) * (population size) or
  - 'How many' = (percentage/100) * (population size)
- For these examples, let's assume μ=40 and σ=10, and a population size of N=10000.
  - Example: "How many of the population is above 50?" Answer: (16% / 100) * 10000 = 1600
  - Example: "How many of the population is below 20?" Answer: (0.025) * 10000 = 250



Additional Example

- A quiz is given, and the resulting scores are normally distributed with μ=10, σ=2, and population N=2000.
  - What fraction of students have a score above 12? Answer: 0.16 (16%/100)
  - What maximum score would it take to be in the bottom 84%? Answer: 12
  - What is the percentile that a score of 8 gives a student? Answer: 16^{th} percentile
  - What are the chances a student scored between 10 and 16? Answer: 49.5% or 0.495
  - How many students scored above 12? Answer: (0.16)*(2000)=320



Conclusion

- Some populations of individuals have data that is *normally distributed*.
  - There are many more individuals near the center, and fewer near the extremes.
  - Idealized normal data is symmetric and always has the same general shape, which is determined by its standard deviation.
    - High precision data has a low standard deviation, meaning the spread of the data about the mean is narrow.
  - *Percentages* or *fractions* of the population between standard boundaries under the normal curve have been calculated.
  - We can use these percentages to answer questions about the data.
    - Like "How many" or "what fraction" is "above/below/between" some score?