Section 3.3: The Empirical Rule and Measures of Relative Standing

The mean and standard deviation tell us a lot about the spread of data from the center. Chebyshev's inequality indicates an approximate percentage of data that falls within a certain number of standard deviations of the mean. The interval $\bar{x} \pm ks$ for samples or $\mu \pm k\sigma$ for populations captures values that fall within k standard deviations of the mean.

1 Chebyshev's inequality

Theorem 1 ***Chebyshev's inequality***: *For every distribution of data, at least* $1 - \frac{1}{k^2}$ *percent of the data falls within* k *standard deviations of the mean for all* $k > 1$.

The most common values used in Chebyshev's inequality are $k = 2$ or $k = 3$.

At least $1 - \frac{1}{2^2} = 1 - \frac{1}{4} = .75 = 75\%$ of the data falls within 2 standard deviations of the mean.

At least $1 - \frac{1}{3^2} = 1 - \frac{1}{9} = .889 = 88.9\%$ of the data falls within 3 standard deviations of the mean.

Problem 2 *Consider a class whose test results have an average of 80 and a standard deviation of 6. What percentage of students earned a grade between 70 and 90?*
The difference from the mean to either endpoint is 10. So $k = \frac{10}{6} \approx 1.67$. *So at least* $1 - \frac{1}{1.67^2} \approx 0.641 = 64.1\%$ *of the scores fall between 70 and 90.*

Problem 3 *Consider a class whose test results have an average of 78 and a standard deviation of 10. What percentage of students earned a grade between 58 and 98?*

Problem 4 *Consider a class whose test results have an average of 78 and a standard deviation of 10. What percentage of students earned a grade between below 58?*

Problem 5 *Par on the Jurassic Mini Golf Course (http://myrtlebeachfamilygolf.com/jurassic-golf/) in Myrtle Beach, SC is 44 strokes with a standard deviation of 8.*
At least 88.9% of the scores fall into what interval?

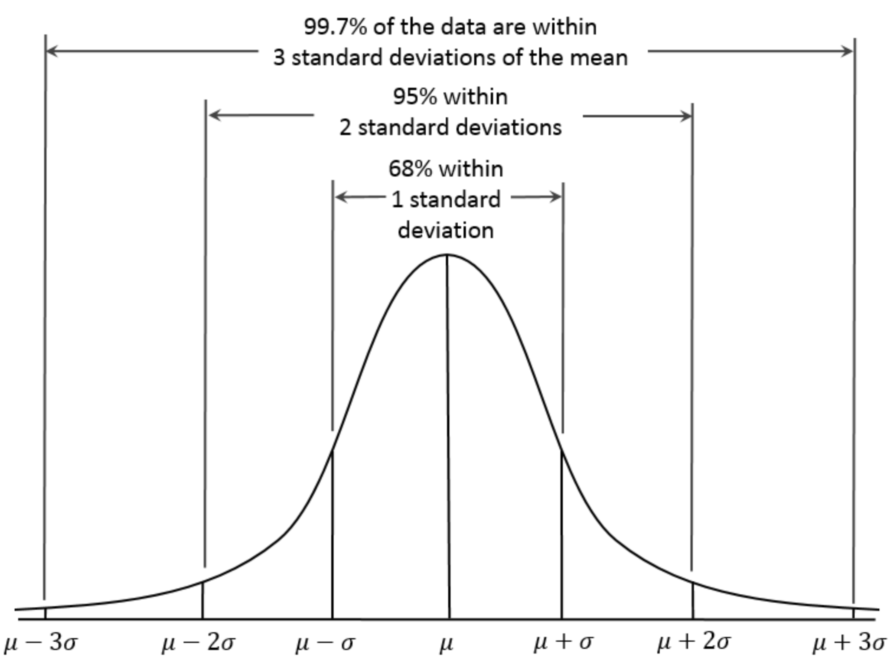
What percentage of scores fall between 20 and 68?

The strength of Chebyshev's inequality is also its weakness. It always works. That's why its conclusion is a lower bound on the amount of data that must be contained within k standard deviations of the mean. If we know more about a specific distribution then we can greatly improve on the result of Chebyshev.

2 The Empirical Rule (68-95-99.7 Rule)

For data distributions that have a **bell-shape distribution** (**normal curve**), the mean and standard deviation tell us a lot about the spread of data from the center.

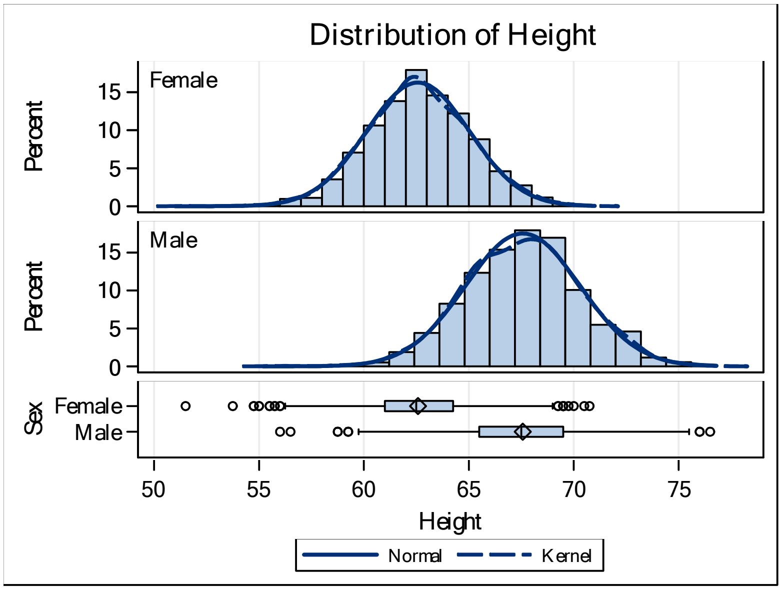
Theorem 6 *The Empirical Rule (68-95-99.7 Rule) states that every normal distribution 68% of the data falls within one standard deviation of the mean, 95% of the data falls within two standard deviations of the mean and 99.7% of the data (almost all) falls within three standard deviations of the mean.*

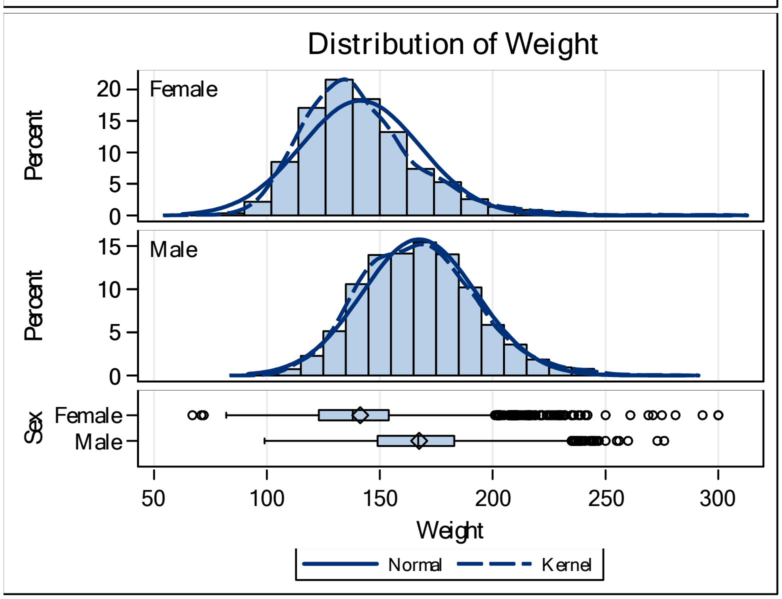


https://en.wikipedia.org/wiki/Normal_distribution

Remark 7 *We frequently denote a normal distribution by* $N(\mu, \sigma)$.

Example 8 *Heights and weights of men and women follow a normal distribution.*


Distribution of Height
Female
Male
Percent
15
10
5
0
Sex
Female
Male
50
55
60
65
70
75
Height
Normal
Kernel


Distribution of Weight
Female
Male
Percent
20
15
10
5
0
Sex
Female
Male
50
100
150
200
250
300
Weight
Normal
Kernel

Problem 9 *Heights of men follow a normal distribution with an average of 69" and a standard deviation of 2.8" (I'm rounding to make the arithmetic easier).*

This indicates that the central 68% of men are from to tall.

What percentage of men are between 60.6" and 77.4" tall?

What percentage of men are between 69" and 71.8" tall?

Problem 10 *Consider a class whose test results have a distribution of N(75,6).*

What grades comprise the central 68% of the students?

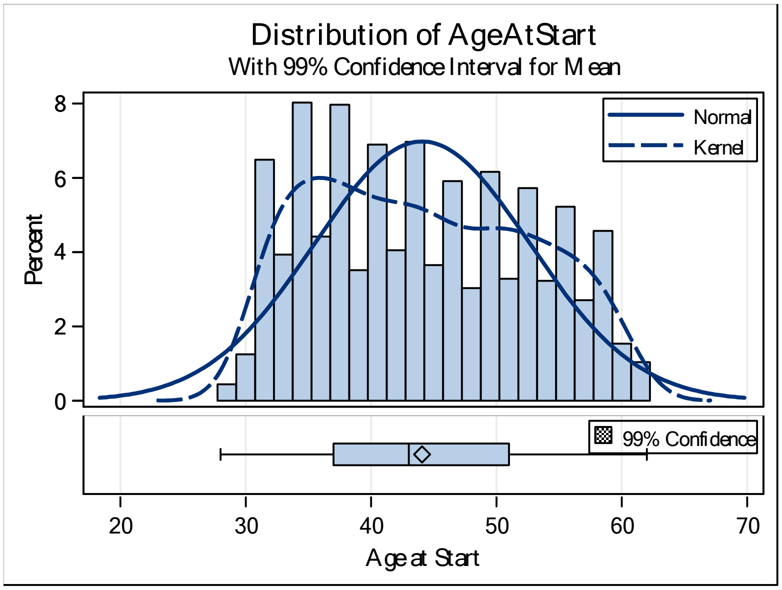
What percentage of grades are between 75 and 81?

What percentage of grades are between 81 and 87?

What percentage of grades are below 57?

If 2000 students took this test, how many students earned a grade less than 57?

Problem 11 *Do we use Chebyshev's inequality or the empirical rule to estimate the amount of data that falls within k standard deviations of the mean for the variable of age at the beginning of the study from our health data set?*



3 z-scores

We've used the empirical rule and Chebyshev's inequality to examine distributions of data. We use a z-score to measure the distance a value is from the mean relative to standard deviation.

$$z = \frac{x-\mu}{\sigma} \text{ for populations}$$

$$z = \frac{x-\bar{x}}{s} \text{ for samples}$$

Example 12 *Consider Math 1107/01, with a test 1 average of 10 and standard deviation of 2. Also, consider Math 1107/02 with a test 1 average of 150 and standard deviation of 15. Use z-scores to determine which score is better, Chris who scored a 13 in Math 1107/01 or Debbie who scored a 180 in Math 1107/02? The z-score for Chris is* $z = \frac{x-\mu}{\sigma} = \frac{13-10}{2} = 1.5$. *The z-score for Debbie is* $z = \frac{x-\mu}{\sigma} = \frac{180-150}{15} = 2.0$. *Since* $2 > 1.5$, *Debbie has the better score.*

Problem 13 *Use z-scores to determine which score is better, Evan who scored a 12 in Math 1107/01 or Francine who scored a 160 in Math 1107/02?*

Problem 14 *Determine which score is better, a 14.5 in Math 1107/01 or a 133 in Math 1107/02? Do we really need to use a z-score for this problem?*

Remark 15 *The sign on a z-score is important! A negative z-score tells us that the data value is below the mean, while a positive z-score tells us that the data value is above the mean.*

Example 16 *What is the original test score for a z-score of* -1.5 *in Math 1107/01? We solve* $-1.5 = \frac{x-10}{2}$ *for* x *and find the test score* $x = 7$

Problem 17 *What is the original test score for a z-score of* -2 *in Math 1107/02?*

Problem 18 *Par on the Jurassic Mini Golf Course (http://myrtlebeachfamilygolf.com/jurassic-golf/) in Myrtle Beach, SC is 44 strokes with a standard deviation of 8. Par on the Captain Kidd's Challenge (http://www.piratesislandgolf.com/) in Hilton Head, SC is 56 strokes with a standard deviation of 12.*

1. Which score is better, a 42 at Jurassic Golf or a 60 at Captain Kidd's Challenge?

2. Which score is better, a 50 at Jurassic Golf or a 60 at Captain Kidd's Challenge?

4 Exercises

1. Kokoska 3rd edition Section 3.3: 3.72-3.76, 3.78-3.81, 3.83, 3.84, 3.87, 3.93, 3.94, 3.103