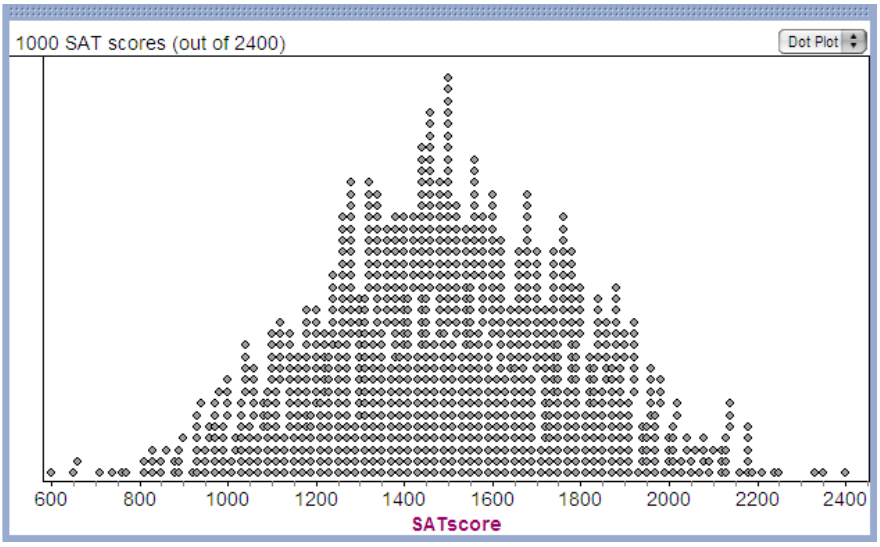
## 9 – Normal Distributions and other Non-Normal Distributions

Problem set 9-1

1. Open up "National SAT Scores.ftm" and make a dot plot of the data.
   a. What are the mean and standard deviation of the data?




   The data is only a sample of the populations, so supply the sample standard deviation (not the population standard deviation).
   Mean=1497.31, Sample standard deviation=298.12

   b. Give the SAT scores that are one standard deviation above and one standard deviation below the mean, What proportion of the data lies within one standard deviation of the mean? (Hint: put the scores in ascending order.)
   1497.31+298.12=1795.43 is one standard deviation above the mean.
   1497.31- 298.12=1199.19 is one standard deviation below the mean.
   682 out of 1000 or 68.2% of the SAT scores are between 1199.19 and 1795.43.

   c. Give the SAT scores that are two standard deviations above and two standard deviations below the mean. What proportion of the data lies within two standard deviations of the mean?
   1497.31+2*298.12=2093.55 is two standard deviations above the mean.
   1497.31- 2*298.12= 901.07 is two standard deviations below the mean.
   956 out of 1000 or 95.6% of the SAT scores are between 901.07 and 2093.55.

   d. Give the SAT scores that are three standard deviations above and three standard deviations below the mean. What proportion of the data lies within three standard deviations of the mean?
   1497.31+3*298.12=2391.67 is three standard deviations above the mean.
   1497.31- 3*298.12= 602.95 is three standard deviations below the mean.
   998 out of 1000 or 99.8% of the SAT scores are between 602.95 and 2391.67.

2. People often have the impression that if we are careful enough when measuring something we will get the exact measurement. But statisticians often say "error is inherent in measurement". This can be true even if we measure the same thing over and over, even if the same person is doing the measurement. In class we will use

an on-line stop watch to measure the length of time from when the starter's pistol goes off until the winner crosses the finish line. We will do this repeatedly, put all the times into a Fathom file, and make a dot plot of the data.

Answer the same questions a-d you did in question (1).

Problem set 9-2

The dot plots for both data sets in problem set 9-1 both had the same bell-shaped distribution.  In addition, the answer to "*What proportion of the data lies within one, two, or three standard deviations of the mean?*" was about the same for both data sets. This is not a coincidence. That is because both data sets were approximately normally distributed. A normally distributed dot plot looks somewhat like a bell and colloquially is called the bell-curve. "Normally distributed" data conform to what statisticians call the Empirical Rule or the 68%-95%-99.7% Rule. That is, about 68% of the data lie within one standard deviation of the mean, about 95% of the data lie within two standard deviations of the mean, and about 99.7% of the data lie within three standard deviations of the mean.

1. For normally distributed data, what proportion of the data:
   a. is above the mean? 50%
   b. is within one standard deviation of the mean? 68%
   c. is between the mean and one standard deviation above the mean? 34%
   d. is below two standards deviation above the mean? 97.5%
   e. is above three standard deviations below the mean? 99.85%

2. Open up again "National SAT Scores.ftm". Create a new attribute (column) called zscore. Each case (row) will be the associated z-score if you make the formula of this new attribute $\frac{\text{(SATscore-mean(SATscore)}}{\text{stdDev(SATscore)}}$. Create a summary table.
   a. What is the mean of "zscore"? 0
   b. What is the standard deviation of "zscore"? 1.0
   c. What proportion of the z-scores lie between -1 and 1? 682 out of 1000 or 68.2%
   d. What proportion of the z-scores lie between -2 and 2? 956 out of 1000 or 95.6%
   e. What proportion of the z-scores lie between -3 and 3? 998 out of 1000 or 99.8%
      Note: The answers to 2 c, d, and e are identical to 9-1-1 b, c and d.
   f. What proportion of the z-scores lie between -3 and 1? 837 out of 1000 or 83.7%
   g. What proportion of the z-scores lie between 2 and 3? 23 out of 1000 or 2.3%

3. You get back an exam and your teacher says that the grades for all the sections were approximately normally distributed.
   a. If you got a z-score of 2, you did better than what proportion of the students taking the test? 97.5%
   b. If you got a z-score of 2.1, you did better than what proportion of the students taking the test? You can't use the Empirical rule to this problem. Just estimate the answer. Any answer that is a little more than 97.5%, like 97.6%.

Problem set 9-3

So far we have learned only what proportion of normally distributed data lies within 1, 2, and 3 standard deviations of the mean. In this section we will learn how to answer questions like the one at the end of problem set 9-2. Traditionally statisticians used tables for these computations, but we will use the Normal Cumulative Density Function or Normalcdf on your calculator. The TI-84 function normalcdf gives us the proportion of normally distributed data we would expect to lie between a lower bound and upper bound, given the mean and standard deviation of the distribution. To access the function on the TI-84, press 2nd, DISTR, 2. The syntax of the function is Normalcdf(lowerbound,upperbound,μ,σ). Thinking back to the SAT data with mean of about 1500 and standard deviation of about 300, approximately 68% of the data was between 1200 and 1800. On the TI-84, if you enter Normalcdf(1200,1800,1500,300) the calculator returns 0.6826894809, indicating that about 68% of SAT scores lie between 1200 and 1800.

1. One of the most commonly used IQ scales is the Wechsler IQ scale, the scores of which are normally distributed with a mean of 100 and standard deviation of 15. Mensa is an organization for people with high IQs; you need an IQ of 130 or higher to become a member. What proportion of the population can be admitted to Mensa?

```
normalcdf(130,1E
99,100,15)
      .022750062
```

about 2.275%

Source: Rodrigo de la Jara, IQ Basics, http://www.iqcomparisonsite.com/IQBasics.aspx, 3/17/2009.

2. The Standard Normal Distribution is a normal distribution with a mean of 0 and standard deviation of 1. Use Normalcdf to find out what proportion of data of the Standard Normal Distribution are:
   a. within one standard deviation of the mean.

```
normalcdf(-1,1,0
,1)
      .6826894809
```

about 68.269%

   b. within two standard deviations of the mean. about 95.450%
   c. within three standard deviations of the mean. about 99.730%
   Show what you typed in your calculator as well as your answer as a percent.
   Hint: Your answers should verify the E……. R… .

3. A Nielsen study about cell phone usage by teenagers reported that the average number of text messages per month was 1,742. If the results were normally distributed and the standard deviation of number of text messages per month was 200, what portion:
   a. sent between 1,542 and 1,942 messages? about 68.269%
   b. sent between 1,000 and 2,000 message? about 90.137%
   c. sent fewer than 1,500 messages? about 11.314%

d. sent more than 2,000 messages? about 9.853%

Source: New York Post, This Kid's a Text Maniac, http://www.nypost.com/seven/01112009/news/nationalnews/this_kids_a_text_maniac_149614.htm, 3/17/2009.

4. You get back an exam and your teacher says that the grades for all the sections were approximately normally distributed. If you got a z-score of 2.1, you did better than what proportion of the students taking the test? about 98.214%

Problem set 9-4

In the same way we have used Normalcdf to find the *proportion* of values of a normal distribution that were between specified values, we can use Normalcdf to find the *probability* that a randomly chosen sample of a population that is normally distributed, is between specified values. For example we used the Empirical Rule to say that the *proportion* of normally distributed SAT scores between 1200 and 1800 (within one standard deviation of the mean) was about 68%. We can also say that if you randomly chose an SAT score, the *probability* that it falls between 1200 and 1800 is about 68%.

1. Assume that nationally SAT scores are normally distributed with a mean of 1500 and standard deviation of 300. (This cannot be true strictly speaking because the maximum SAT score is 2400; normally distributed data have no maximum or minimum.) If an SAT score is chosen at random:
   a. what is the probability that the score is below 2000? about 95.221%
   b. what is the probability that the score is below 1800? about 84.134%
   c. what is the probability that the score is between 1800 and 2000? about 11.086%
   d. what is the probability that the score is above 2400? about 0.135% Note: Because the theoretical answer is not zero, we have demonstrated that SAT scores are not exactly normally distributed. Nonetheless, the normal distribution is a fairly accurate and useful distribution to model the way SAT scores are distributed.

2. Blood alcohol content or BAC is typically measured in mg of alcohol /ml of blood. Adults over 21 in most states will be arrested for DUI if they have 80 mg/ 100 ml (0.08% BAC) or more. Many handheld BAC meters like the AlcoHAWK Elite Digital by Breathalyzer and the Alco-Sensor III by Intoximeters give a readout to three decimal places. According to a Clinical Chemistry article, the standard deviation for a new Model 1000 Breathalyzer is 0.0037. Let's say you are on a jury in which the defendant is charged for DUI and the BAC was measured to be exactly 0.080 which is the minimum amount necessary for a conviction.
   a. What is the probability that the defendant would have a measured BAC of 0.080 or greater if in fact the defendant's actual BAC was 0.079? about 39.3%
   b. What is the probability that the defendant would have a measured BAC of 0.080 or greater if in fact the defendant's actual BAC was 0.075? about 8.8%
   c. What is the probability that the defendant would have a measured BAC of 0.080 or greater if in fact the defendant's actual BAC was 0.070? about 0.3%
   d. Based on the evidence, describe whether you think it was safe for the defendant to be driving. Explain. (If you weigh 160 lbs. and drink three 12 oz. beers in succession, your BAC would be about 0.070)
   e. Based on the evidence, describe whether you think the defendant should be found guilty of DUI. Explain.

Sources:


- Breath-tester. Breath Alcohol Testers, http://www.breath-tester.com/, 3/17/2009
- Wikipedia, BAC, http://en.wikipedia.org/wiki/Blood_alcohol_content, 3/17/2009

- G. Simpson, Accuracy and ZPresicion of Breath-Alcohol Measurements for a Random Suject in the Postabsorptive State, Clinical Chemistry, 33/2, 261-268, 1987 (http://www.clinchem.org/cgi/reprint/33/2/261.pdf)

Note: In most states the BAC limit for persons under the age of 21 is 0.000 - 0.020. The majority of countries in the world convict for DUI at BAC of 0.05.

3. The TI-84 function invNorm gives us the value in the normal distribution associated with a specified percentile. For example, since Normalcdf(-E99,110,100,10) returns about .841, invNorm(0.841,100,10) returns about 110.

   Going back to scores that are normally distributed with a mean of 1500 and standard deviation of 300:

   a. What SAT score is at the 50th percentile? Verify your answer using invNorm.

   ```
   invNorm(.5,1500,
   300)
                    1500
   ```

   1500

   b. What SAT score is at the 25th percentile? about 1297.7
   c. What SAT score is at the 75th percentile? about 1702.3
   d. What is the interquartile range for the SAT scores? about 404.7
   e. How high would an SAT score need to be to be considered an outlier?
      Q3 + 1.5XIQR = 2309.4, so SAT scores of 2310 or above are outliers.

Problem set 9-5

1. The passage in italics below is taken verbatim from:
   The City of Saratoga, CA, 85th Percentile Speed, http://www.saratoga.ca.us/boards-commissions/traffic/documents/85thPercentileSpeed.pdf, 4/5/2009

   ***85th Percentile Speed***

   *On most local streets, the speed limit is posted as 25 miles per hour (mph). In all residential and business districts where a limit is not posted, 25 mph this is the implied limit. Speed limits on higher capacity streets including major collector and arterial streets are set based on engineering and traffic surveys that include a review of speed data, design parameters, and operational issues.*

   *Traffic engineers rely on the 85th percentile rule to help establish speed limits on nonlocal streets. Typically, the speed limit is set to the speed that separates the bottom 85% of vehicle speeds from the top 15%. For example, if speeds of 100 vehicles are measured and 85 vehicles are traveling at 37 mph or less, the speed limit for the subject street could be set at 35 mph. Statistically, the 85th percentile speed is slightly greater than the speed that is one standard deviation above the mean of a normal distribution.*

   *The theory behind this approach is that most drivers will travel at a speed that is reasonable and prudent for a given roadway segment. Most U.S. jurisdictions report using the 85th percentile speed as the basis for their limits. However, California law permits jurisdictions to implement and enforce lower speed limits where slower speeds are desired to better accommodate bicycle and pedestrian movement, especially in residential and business districts. Specific criteria regarding building density fronting on a street and a lack of bicycle facilities or sidewalks/paths can be used to justify reductions in speed limits initially set based on the 85 th percentile speed.*

   ***What is the 85th Percentile Speed?***

   *The 85th percentile speed is the speed at or below which 85 percent of the motorists drive on a given road unaffected by slower traffic or poor weather. This speed indicates the speed that most motorists on the road consider safe and reasonable under ideal conditions. It is a good guideline for the appropriate speed limit for that road.*

Two questions:

a. Use the Empirical Rule to explain where the 85 in the 85th Percentile Speed comes from. You answer will need to include both some simple arithmetic as well as prose that explains the relationship between the Empirical Rule and the 85th Percentile Speed.

b. Open the “Speed Limit” Fathom file. It gives the speed of 100 drivers on a non-residential road in which drivers were focusing only on a safe driving speed; they

were not worried about getting a speeding ticket. Establish a speed limit using the 100 speeds in the Fathom file and the 85th Percentile Speed.

2. Most plants cannot withstand a frost, so you need to wait until after the last frost to plant seedlings or seeds. The date of the last frost are normally distributed.
   a. Use the information below to determine mean and standard deviation of the date of the last spring frost in Amherst for 32 degrees Fahrenheit.
   b. What is the spring date associated with a 5% probability level?

## Freeze / Frost Occurrence Data

All probabilities in whole percent. See notes for probability level description.

- Indicates the probability of occurrence of threshold temperature is less than indicated probability.

| State And Station Name | Threshold (F) | Spring (Date) | | | Fall (Date) | | | Freeze Free Period (Days) | | | Probability Level (4) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Probability Level (1) | | | Probability Level (2) | | | Probability Level (3) | | | |
| Massachusetts | | 90 | 50 | 10 | 10 | 50 | 90 | 10 | 50 | 90 | |
| AMHERST | 36<br>32<br>28 | May06<br>Apr26<br>Apr17 | May23<br>May10<br>Apr28 | Jun08<br>May24<br>May09 | Sep01<br>Sep17<br>Sep27 | Sep16<br>Sep28<br>Oct13 | Oct01<br>Oct09<br>Oct29 | 140<br>155<br>186 | 116<br>140<br>167 | 91<br>125<br>148 | 50<br>42<br>34 |

Notes:

(1) Probability of later date in spring (thru Jul 31) than indicated.

(2) Probability of earlier date in fall (beginning Aug 1) than indicated.

(3) Probability of longer than indicated freeze free period.

(4) Probability of Freeze/Frost in the yearly period (percent of days with temperatures at or below the threshold temperature).

Source: National Climate Data Center, Freeze/Frost Occurrence Date,, http://cdo.ncdc.noaa.gov/climatenormals/clim20supp1/states/MA.pdf, 4/12/2009