AP Statistics Name \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Chapter 2: Modeling Distributions of Data Date \_\_\_\_\_\_\_\_\_\_\_ Period \_\_\_\_\_\_\_\_\_\_

**2.1 Transforming Data and Density Curves**

**Definition:**

**Percentile**: the *pth percentile* of a distribution is the value with p percent of the observations less than it

**Ex: 1: Wins in Major League Baseball**

The stemplot below shows the number of wins for each of the 30 Major League Baseball teams in 2009.

5 9

6 2455

*Key*: 5|9 represents a team with 59 wins.

7 00455589

8 0345667778

9 123557

10 3

Calculate and interpret the percentiles for the Colorado Rockies who had 92 wins, the New York Yankees who had 103 wins, and the Cleveland Indians who had 65 wins.

**Cumulative Relative Frequency Graphs**

CreatingCumulativeRelativeFrequency- take the cumulative frequency, divide by the total. Use the resulting decimal or corresponding percent.

CreatingaCumulativeRelativeFrequencyGraph(**ogive)**- we plot a point corresponding to the cumulative relative frequency in each class at the smallest value of the *next* class

**Ex 2:** The frequency table below summarizes the ages of the first 44 U.S. Presidents when they were inaugurated.

a) Expand the table to show relative frequency, cumulative frequency, and cumulative relative frequency.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Age | Frequency | Relative Frequency | Cumulative Frequency | Cumulative Relative Frequency |
| 40-44 | 2 |  |  |  |
| 45-49 | 7 |  |  |  |
| 50-54 | 13 |  |  |  |
| 55-59 | 12 |  |  |  |
| 60-64 | 7 |  |  |  |
| 65-69 | 3 |  |  |  |

1. Plot a cumulative relative frequency graph (**ogive)** in the space below (use your separate sheet of paper if necessary)
2. Was Barack Obama, who was inaugurated at age 47, unusually young?
3. Estimate and interpret the 65th percentile of the distribution.

**Ex 3:**

1) What is the median age of school enrollment in 1996?



Age

Cumulative relative frequency

2) What is the interquartile range of school enrollment in 1996?

3) At or below what age is the bottom 10% of school enrollment in 1996?

4) At or above what age is the top 20% of school enrollment in 1996?

# Measuring Position: z-Scores

**Z-Score (Standardized Value):**  if x is an observation from a distribution that has known mean and standard deviation, then the z-score of x is

Z =

The z-score tells us how many standard deviations from the mean an observation falls, and in what direction.

Positive Z-Score= larger than the mean Negative Z-Score= smaller than the mean Zero Z-Score= equal to the mean

**Ex 4: Wins in Major League Baseball:** In 2009, the mean number of wins was 81 with a standard deviation of 11.4 wins.

Find and interpret the z-scores for the following teams.

1. The New York Yankees, with 103 wins.
2. The New York Mets, with 70 wins.

**Ex 5: Home run kings**

The single-season home run record for major league baseball has been set just three times since Babe Ruth hit 60 home runs in 1927. Roger Maris hit 61 in 1961, Mark McGwire hit 70 in 1998 and Barry Bonds hit 73 in 2001. In an absolute sense, Barry Bonds had the best performance of these four players, since he hit the most home runs in a single season. However, in a relative sense this may not be true. Baseball historians suggest that hitting a home run has been easier in some eras than others. This is due to many factors, including quality of batters, quality of pitchers, hardness of the baseball, dimensions of ballparks, and possible use of performance-enhancing drugs. To make a fair comparison, we should see how these performances rate relative to others hitters during the same year. Calculate the standardized score for each player and compare.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Year** | **Player** | **HR** | **Mean** | **SD** | **z-score** |
| 1927 | Babe Ruth | 60 | 7.2 | 9.7 |  |
| 1961 | Roger Maris | 61 | 18.8 | 13.4 |  |
| 1998 | Mark McGwire | 70 | 20.7 | 12.7 |  |
| 2001 | Barry Bonds | 73 | 21.4 | 13.2 |  |

In 2001, Arizona Diamondback Mark Grace’s home run total has a standardized score of *z* = –0.48. Interpret this value and calculate the number of home runs he hit.

A large z-score indicates a value that is\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

A small z-score indicates a value that is \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

A z – score near zero indicates a value that is \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

When is a z-score considered large? \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Transforming Data**

**Ex 6:**

1. Below are the test score’s from an AP Statistics Class first test.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 71 | 92 | 65 | 88 | 78 | 73 89 | 83 | 90 | 67 | 79 | 54 |
| 97 | 61 | 84 | 77 | 59 | 71 81 | 62 | 76 | 85 | 93 | 48 |

1. Plot the test scores on a dotplot and CUSS.
2. The teacher decides to scale the test scores by 10 points. Plot these new scores on the same dotplot (in a different color).
3. What effect did the 10 points have on the graph? (Be sure to CUSS)
4. What can we conclude about the effect of adding (or subtracting) a constant?
5. After much thought, the teacher decided to instead scale their test scores by a factor of 1.2.
6. Plot both the original scores and the newly scaled scores on a dotplot.
7. Describe the changes to the graph (be sure to CUSS).
8. What can we conclude about the effect of multiplying (or dividing) by a constant?

On the shape, center and spread of a distribution, what is the effect of adding or subtracting a constant from each observation?

On the shape, center and spread of a distribution, what is the effect of multiplying or dividing each observation by a constant?

**How are transformations connected to z-scores?**- now that we know how a set of data is effected by adding/subtracting a constant and multiplying/dividing a constant, notice that when finding a z-score I subtract and then divide.

**Density Curves**

After we have plotted our data, found an overall pattern, and calculated a numerical summary to describe center and spread, we find one more step to exploring our quantitative data- *sometimes the overall pattern of a large number of observations is so regular that we can describe it by a smooth curve*

A *Density Curve* is a curve that:

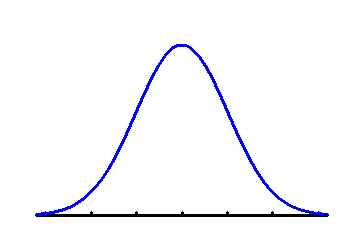
* + Is always on or above the horizontal axis
  + Has area exactly 1 underneath it
  + Describes the overall pattern of a distribution
  + The area under the curve and above any interval of values on the horizontal axis is the proportion of all observations that fall in that interval
  + Outliers are NOT described by this curve
  + The curve is an approximation, not exactly described

*Median* of a density curve- “equal-areas point”; half the area is on the left and half the area is on the right

*Mean* of a density curve- point at which the curve would balance if made of solid material

The median and mean are the **same** for a *symmetric density curve*. They both lie at the center of the curve.

The mean of a skewed curve is **pulled away from the median in the direction of the long tail.** The mode is the tallest value on the graph.

Identify the mean and median on the density curves below.

1. 2.

A B C A B C

3. Using the uniform distribution below, find the areas.

a. area below 0.25 = \_\_\_\_\_\_\_\_\_\_

b. area above 0.70 = \_\_\_\_\_\_\_\_\_\_

c. area between 0.45 and 1.25 \_\_\_\_\_\_\_\_\_\_

**0 0.5**

**2.2 Normal Distributions**

**Definition:**

A **Normal distribution** is described by a Normal density curve. Any particular Normal distribution is completely specified by two numbers: it’s mean μ and standard deviation σ . The mean of a Normal distribution is at the center of a symmetric **Normal curve.** The standard deviation is the distance from the center to the change-of-curvature points on either side.

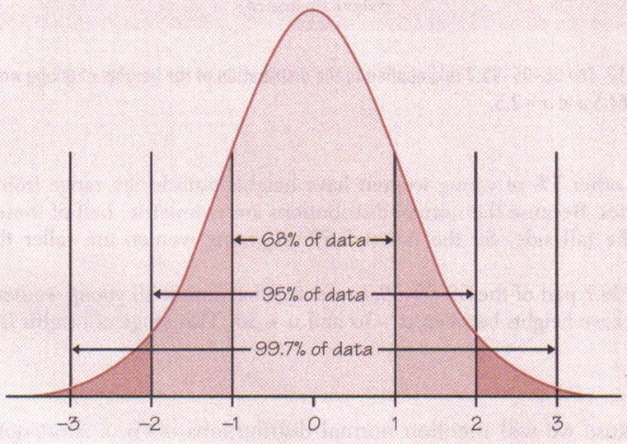
We abbreviate the Normal distribution with mean μ and standard deviation σ as *N* (μ, σ)

**3 Reasons why Normal distributions are important in statistics**

1. Normal distributions are good descriptions for some distributions of real data
2. Normal distributions are good approximations to the results of many kinds of chance outcomes
3. Most important, is that many statistical inference procedures based on normal distributions work well for other roughly symmetric distributions

**Empirical Rule/ 68-95-99.7 Rule**

* Approximately 68% of the observations fall within 1 σ of the mean μ
* Approximately 95% of the observations fall within 2 σ of the mean μ
* Approximately 99.7% of the observations fall within 3 σ of the mean μ



*Standard Normal Curve (has mean of 0 and standard deviation of 1)*

**Ex 1:** In 2010, the distribution of batting averages for Major League Baseball players was approximately Normal with a mean of 0.261 with a standard deviation of 0.034. Sketch this distribution, labeling the mean and the points one, two, and three standard deviations from the mean.



a. Approximately, what percentage of players had batting averages w/in 1 standard deviation of the mean? \_\_\_\_\_\_\_

W/in 2 standard deviation? \_\_\_\_\_\_\_\_\_\_\_ W/in 3 standard deviations? \_\_\_\_\_\_\_\_\_\_

b. What percentage of players had batting averages above .295? \_\_\_\_\_\_\_\_\_\_ Below .227? \_\_\_\_\_\_\_\_\_

c. What percentage of players had batting averages above .329? \_\_\_\_\_\_\_\_\_\_ Below .193? \_\_\_\_\_\_\_\_\_

d. What percentage of players had batting averages between .295 and .329? \_\_\_\_\_\_\_\_\_\_

e. 34% of players had batting averages between \_\_\_\_\_\_\_\_\_\_ and \_\_\_\_\_\_\_\_\_\_\_.

f. 50% of players had batting averages below \_\_\_\_\_\_\_\_\_.

**Ex. 2** Suppose that a distribution of test scores is approximately Normal and the middle 95% of scores are between 72 and 84. What are the mean and standard deviation of this distribution?

**Ex 3:** Can you calculate the percent of scores that are above 80? Explain.

**2.2 The Standard Normal Distribution**

**The standard Normal distribution** is the Normal distribution with mean 0 and standard deviation 1.

Z =

An area under a density curve is a proportion of the observations in a distribution. Any question about what proportion of observations lie in some range of values can be answered by finding an area under the curve. Because all Normal distributions are the same when we standardize, we can find the areas under any Normal curve from a single table; **Table A** is a table of areas under the standard Normal curve. The table entry for each value z is the area under the curve and to the *left* of z (Note: when using the table to find areas to the *right* of z, you will need to do 1 - table value. You will not need to do this if using the calculator)

**Table A in your book**

**Calculator function- 2nd, Vars, normalcdf (#2)**

1. Find the proportion of observations from the 2. Find the proportion of observations from the

standard Normal distribution that are less than 0.54. standard Normal distribution that are greater

than –1.12.

3. Find P(z ≥ 2.12) 4. Find P(z > 3.89)

5. Find P(0.49 < z < 1.82) 6. Find P()

7. On a Normal distribution, what proportion of observations from are within 1.5 standard deviations of the mean?

8. A distribution of test scores is approximately Normal and Joe scores in the 85th percentile. How many standard deviations above the mean did he score?

**How to Solve Problems Involving Normal Distributions:**

**Example: Serving Speed**

In the 2008 Wimbledon tennis tournament, Rafael Nadal averaged 115 miles per hour (mph) on his first serves. Assume that the distribution of his first serve speeds is Normal with a mean of 115 mph and a standard deviation of 6 mph.

(a) About what proportion of his first serves would you expect to exceed 120 mph?

(b) What percent of Rafael Nadal’s first serves are between 100 and 110 mph?

(c) The fastest 20% of Nadal’s first serves go at least what speed?

(d) What is the *IQR* for the distribution of Nadal’s first serve speeds?

**Example: Heights of 3-year old females**

The heights of three-year-old females are approximately Normally distributed with a mean of 94.5 cm and a standard deviation of 4 cm. What is the third quartile of this distribution?

**2.2: Assessing Normality**

Read 124-128

The measurements listed below describe the useable capacity (in cubic feet) of a sample of 36 side-by-side refrigerators. (Source: *Consumer Reports*, May 2010) Are the data close to Normal? Graph the data using a boxplot, histogram and normal probability plot.

12.9 13.7 14.1 14.2 14.5 14.5 14.6 14.7 15.1 15.2 15.3 15.3

15.3 15.3 15.5 15.6 15.6 15.8 16.0 16.0 16.2 16.2 16.3 16.4

16.5 16.6 16.6 16.6 16.8 17.0 17.0 17.2 17.4 17.4 17.9 18.4

When looking at a Normal probability plot, how can we determine if a distribution is approximately Normal?

Sketch a Normal probability plot for a distribution that is strongly skewed to the left.