Ch 5: The Normal Approximation to Data

The _____ or _____ is the most important curve in statistics. This curve is that important that it is even honored on a German bank note!

Many data sets approximate the normal curve closely, and many statistical procedures use the normal curve.

The equation of the normal curve is

$$y = \frac{100\%}{\sqrt{2\pi}} e^{-x^2/2},$$

where $e = 2.71828...$, but we won't use the equation.

The graph is symmetric, has a total area of 100% between the curve and the horizontal (x-) axis, and the curve is always above the x-axis (though it gets very close).

Many histograms fit the normal curve closely, if they are drawn in _____.

Standard units say how many SD's above (+) or below (-) the average a value is.

Ex: In the HANES sample, women 18-74 had an average height of 63.5", and an SD of 2.5".

If a woman is 66" tall, she is 66"-63.5" = 2.5" above average, which is 1 SD above average. Therefore, 66" = +1 standard unit.

- What is a height of 56" in standard units?

- 67.5"?

- 63.5"?

- What height is -.6 standard units?

If a histogram follows the normal curve, the area under the histogram is approximately the same as the area under the normal curve.

Recall from Chapter 4: about 68% of values are often within 1 SD of the average, about 95% are within 2 SD's, and about 99.7% are within 3 SD's. This is based on the normal curve.

The table on page A-105 gives percentages within certain numbers of standard units ("z") of the average. The modified table at the end of this set of handouts does not contain the "Height" column which we don't use.

We can calculate the area for any range of values by using this table.

Sketches often help in using this table.

Note: different books contain different tables.

Ex: What is the area under the normal curve?

- Between -0.5 and 0.5?
- Between 0 and 1.5?
- Between -2 and 0?
- Between -2 and 1.5?
- Between -1.5 and 2?
- More than 1.5?
- Less than 0.5?
- Between 0.5 and 1.5?

The Normal Approximation

We use the _____ to estimate areas for histograms which closely follow the normal curve.

Ex: For women in HANES, avg = 63.5", SD = 2.5".

- What percentage were between 62" and 68.5"?

- What percentage were less than 67.5"?

How to proceed:

1. Rewrite the question, putting all values in standard units.
2. Answer the question, as though it were asked about the normal curve.

Percentiles

If data is not normal, then the average and SD are not as good as summary statistics.

Ex: For the 1987 income data, avg = \$44,500, and SD = \$32,000. By the normal approximation, \$0 is at -1.4 standard units. How did we get this value of -1.4?

The area to the left of -1.4 under the normal curve is about 8%. Thus, by the normal approximation, about 8% of families should have a negative income!

With non-normal data, we can summarize more accurately than with the average and SD by using _____.

The _____ is the value such that p% of the values are below, and (100-p)% of the values are above.

We commonly use percentiles including 1, 10, 25, 50, 75, 90, and 99.

The 50th percentile is called the _____.
The 25th percentile is called the _____.
The 75th percentile is called the _____.

Percentiles of 1992 US family income

Percentile	Income
1	$1,300
10	$10,200
25	$20,100
50	$36,800
75	$58,100
90	$85,000
99	$151,800

Interquartile Range

The _____ (IQR) is

$$IQR = \text{75th percentile} - \text{25th percentile.}$$

The IQR is used as a measure of spread when the SD is too heavily influenced by one or two extreme tails.

Ex: What is the IQR of the 1992 US family income data?

Calculating Percentages

When a histogram follows the normal curve, we can use a normal table to estimate the percentiles of the data.

1. Work backwards in the table to go from area to z (if a percentile greater than 50) or -z (if less than 50).

2. Convert z or -z to the original units.

Ex: What is the 99th percentile of women's heights in the HANES study?

Ex: What is the first quartile of the women's heights??

Change of Scale

Suppose we wish to work with our data in new units, such as changing meters to feet, or degrees Fahrenheit to degrees Celsius. Generally, this will involve multiplying every value by the same constant, and perhaps adding another constant to every value. How will this change the average and SD?

- Adding the same constant to every value on a list adds the same constant to the average. The SD doesn't change.

- Multiplying every value on a list by the same constant will multiply the average by the same constant and the SD by the absolute value of the constant.

- Neither will affect the standard units.

Ex: To convert degrees Fahrenheit to degrees Celsius, we use the formula

$$C = \frac{5}{9}(F - 32) = \frac{5}{9}F - \frac{160}{9}.$$

Given Fahrenheit temperatures 32, 50, 59, 68, and 86, with average 59 and SD 18:

1. Convert the temperatures to Celsius, and find the average and the SD.

2. How could we calculate the Celsius average and SD without converting all of the values?