## Record: 1

**Title:** STATISTICS: DIGGING INTO DATA.

**Authors:** Sullivan, Lisa M.
D'Agostino, Leilanie M.

**Source:** Odyssey; Dec2003, Vol. 12 Issue 9, p6-9, 4p, 2 charts, 3 graphs, 4 color

**Document Type:** Article

**Subject Terms:** STATISTICS
MATHEMATICS

**Abstract:** Focuses on a branch of mathematics called statistics that involves collecting, organizing, and interpreting information.

**Lexile:** 950

**Full Text Word Count:** 2146

**ISSN:** 01630946

**Accession Number:** 11546904

**Database:** Primary Search

# STATISTICS: DIGGING INTO DATA

A branch of math called statistics can help us better understand our world. But it can also be deceiving. How do you tell the good stats from the bad?

How does Nokia determine what cell phone design you're most likely to buy? Or what neat features on it you're likely to use most? And how do scientists determine if that cell phone is harmful to your health? The answer is statistics — an area of mathematics that focuses on collecting, organizing, and interpreting information or data.

In the terminology of statistics, you're a subject — someone (or something) to collect data on. In many statistical applications — from determining the types of jurors in a court case who might favor a conviction to determining which player on a sports team might perform best in the playoffs — subjects are people. The data to be collected are characteristics.

In the case of the cell phone, the characteristics to be measured are things like your age, how many calls you make, when and for how long you use a phone, and your economic status. That information is organized and summarized using specific techniques and procedures, and relationships are interpreted among the characteristics.

This process is called statistical analysis, and it can help marketers determine which Nokia model will be the hot phone of the season.

Cell phones aside, statistical analysis touches many areas of our lives. It tells investors what stocks achieve the best returns, points doctors to the medications that are most effective in treating disease, and can even help you to understand how exercising and eating right now will affect your health as an adult.

## The Big Picture

The goal of statistical analysis is to use a small group (a sample) to say something about a large group (a population). A population is all the subjects of interest — for example, ALL American males with heart disease or ALL middle school students who go to private schools.

 

In many situations, it's impossible or impractical (or both) to analyze an entire population. It may be too time-consuming or take so long that by the time the results are available they are no longer useful. And what about a characteristic that must be measured by a $2,000 laboratory test? It simply would be too expensive if every member of the population were tested.

So, in statistical analysis, a sample — or subset of subjects from the population — is analyzed. If the sample is representative of the population, then it is reasonable to assume that what is observed in the small group will be similar to what would be observed in the entire population.

The process of collecting and organizing information from a sample is called descriptive statistics (a sample is described). When predictions are made about the population based on that sample, the process is called statistical inference. (Statisticians infer what will happen — they can never be 100 percent certain about their predictions.)

Mathematical theorems and principles of probability are used to quantify how much error or imprecision exists. For example, the following example has a probability component: We are 95 percent confident that 30 minutes of exercise per day will reduce the chance of heart disease for an individual by 10 to 20 percent over the next 20 years.

**Example A** shows the two areas of statistical analysis — descriptive statistics and statistical inference.

## But Why Do We Need Statistics?

We use statistics to make important decisions in our lives. For example, most people believe that regular exercise and eating a low-fat diet promote good health. (We bet you've heard that bit of advice from your mom just a few times.) This belief sometimes motivates us to exercise when we don't feel like it or makes us feel guilty when we eat a four-cheese pizza or a burger with super-size fries. Should we really feel guilty?

Why do we believe that exercising and eating healthy are associated with better health in the first place? Well, there are a number of statistical research studies that have shown that these factors reduce the risk of heart disease.

An example is the Framingham Heart Study. It began in 1948 with over 5,000 men and women participants. Each had a complete physical examination at the study's start. The participants came back for repeat examinations every two years, and the study is still going on today. Along with monitoring such things as blood pressure, cholesterol, exercise, and nutrition, the study investigators also monitor whether participants experience heart attacks or other heart disease over time.

In the mid-1970s, the Framingham study was expanded to include a second generation, the children of the original subjects and their husbands and wives. In 2001, a third generation (the children of the second generation) was recruited, and these samples are now being analyzed for genetic factors associated with disease. Much of what we understand today about heart disease has been learned from this very important study.

The statistical results of the Framingham Heart Study often get reported in the newspaper or on television. But why should we believe that just because there was an association between exercise and heart disease among its participants (or those in any other study), the same would hold true for us?

The idea of generalizing or inferring associations from a study (a sample) to the population at large is the crux of statistical analysis. Let's look at an example.

## Comparing Weights

Because weight increases as we grow in our teens and then tends to rise with age, the

 

focus of this analysis is on boys and girls between the ages of 10 and 19. The sample includes a total of 578 boys and 640 girls, and the data is similar to that in the second generation of the Framingham Heart Study.

In statistics, "summary measures" are used to describe characteristics. You've probably all heard of the mean and the median. Both are measures of average value.

The mean represents a typical value on a characteristic and is calculated by summing all of the values and dividing by the total number of subjects. (Your teacher might report the mean test score for the class.)

The median also represents a typical value and is defined as the middle value when all of the values are placed in order from lowest to highest. The interpretation of the median is that half of the values are above the median and half are below.

**Describing the Sample**

In our sample, the mean weight is 128 pounds and the median weight is 129 pounds. They are very close, but this isn't always the case. For example, suppose we have a sample of five test scores: 25, 80, 85, 90, and 100. The mean test score is 76 and the median is 85 (the middle number when we arrange the scores from lowest to highest). In this example, the very low test score of 25 "pulls down" the mean, making it seem like a typical test score was 76, when in fact most of the scores are very high. For this example, the median (85) is a better measure of a typical test score.

The score of 25 is an "outlier" — it does not fit with the rest, because it is extremely low by comparison. When there are no outliers, the mean and median will usually be close in value and the mean is the preferred summary measure.

Since weights are substantially different for boys and girls, especially after age 14, in this study their weights were analyzed separately.

Example B shows one way to "describe" the weights. It is called a box and whisker plot, and shows the whole distribution of weights for boys (on the left) and girls (on the right). The lowest and highest weights are indicated by the lines at the bottom and top. The boys' weights range from a low of 60 to a high of 234, and the girls', from a low of 66 to a high of 255. The ranges are very similar.

The shaded box shows the middle 50 percent of the weights for boys and girls. For the boys, the middle 50 percent of weights fall between 100 and 153 pounds. The box includes the middle 50 percent, which leaves 25 percent above it (weighing more than 153 pounds) and 25 percent below (weighing less than 100 pounds). For the girls, the middle 50 percent of the weights are between 114 and 142, 25 percent are below 114, and 25 percent are above 142 pounds.

The bar across the middle of the shaded box is the median, and the dot is the mean. The mean weight for boys is 127 and the mean weight for girls is 130. Overall the weights look similar, but there is much less variability (spread) in the weights of the girls, at least in the middle 50 percent. Were you wondering why the girls' shaded box was narrower? That's why.

In this sample, the girls are, on average, three pounds heavier than the boys. Wow! Does that make sense? What's happening here?

To find out, let's look at Example C and another characteristic of our sample: the ages of the boys and girls.

Since there are a total of 578 boys and 640 girls in the sample, the best way to compare their ages is by using percentages. In general, are the boys younger or older than the

 

girls?

Right. . .the majority of the boys are younger, between ages 10 and 15 (85 percent), while the majority of the girls are older, between ages 16 and 19 (79 percent). A box and whisker plot of the ages readily shows the difference:

Notice that the middle 50 percent of the boys' ages are between 11 and 15, while the middle 50 percent of the girls' ages are between 16 and 19. There is more variability (or more spread) in the ages of the boys as compared to the girls (whose ages are mostly concentrated between 17 and 19). The mean age of the boys is 13, while the mean age of the girls is 18. How would this affect our interpretation of the weights of boys and girls?

## Making an Inference: So What Does It Tell Us?

As we look more closely, we realize that the girls are older than the boys in our sample. That's why they weigh more! To make a "fair" comparison of their weights, we should be looking at boys and girls of similar ages. Example D is a better comparison and shows the mean weights for boys and girls at each age.

Now how do the weights compare between boys and girls? When we look at each age, the mean weights for boys and girls are similar (remember that these are only samples) until about age 14. (In statistics, there are tests that can be run to determine when differences are large or when the differences are within what we would expect just by chance.)

For age 15 and older, the boys have higher mean weights than the girls. This is a more appropriate interpretation of the data, wouldn't you agree? Example E presents box and whisker plots of the weights for boys and girls at each age that show in detail the weight differences for this sample.

## Caution!

As our weight example shows, it's crucial to look carefully at how data is analyzed. Statistical analysis can be applied to almost anything, including agriculture, biology, sports, chemistry, psychology, sociology, political science, economics, engineering, and on and on. It can reveal much about our world if we view it with a critical and sometimes skeptical eye.

The next time you read stats from a newspaper, magazine, book, television newscast, or the Internet, do some analyzing on your own. Since just a few of us will ever participate in a research study as a subject, we have no choice but to rely on well-conducted research to keep us informed. Just remember to dig into the data!

EXamPLE C

| Age | Number (%) of Boys | Number (%) of Girls |
|---|---|---|
| 10 | 63 (11%) | 13 (2%) |
| 11 | 86 (15%) | 12 (2%) |
| 12 | 83 (14%) | 6 (1%) |
| 13 | 86 (15%) | 31 (5%) |
| 14 | 83 (14%) | 24 (4%) |
| 15 | 96 (16%) | 44 (7%) |
| 16 | 16 (3%) | 52 (8%) |
| 17 | 16 (3%) | 82 (13%) |
| 18 | 21 (4%) | 80 (12%) |
| 19 | 28 (5%) | 296 (46%) |
| TOTAL | 578 | 640 |

Ages of Boys and Girsl
EXamPLE D

 

| Age | Mean Weight for Boys | Mean Weight for Girls |
|---|---|---|
| 10 | 83 | 80 |
| 11 | 90 | 95 |
| 12 | 120 | 122 |
| 13 | 132 | 135 |
| 14 | 136 | 132 |
| 15 | 157 | 130 |
| 16 | 159 | 138 |
| 17 | 156 | 130 |
| 18 | 165 | 129 |
| 19 | 159 | 127 |
| ALL | 127 | 130 |

GRAPH: EXamPLE B

GRAPH: EXamPLE E Boys

GRAPH: EXamPLE E Girls

PHOTO (COLOR): EXamPLE a: STATISTICAL ANALYSIS

PHOTO (COLOR)

PHOTO (COLOR)

PHOTO (COLOR)

~ ~ ~ ~ ~ ~ ~ ~

By Lisa M. Sullivan, Ph.D. and Leilanie M. D'Agostino

She teaches biostatistics and statistics to undergraduate and graduate students and is involved in research work on the Framingham Heart Study.