

Predicting the sales and airplay of popular music

Topics covered: Data summary. Examination of univariate data. Prediction.

Key words: Scatter plot.

Data File: rock.dat

Each week the music industry magazine *Billboard* publishes lists of the most popular songs from the last week. The songs are ranked in various categories of interest to radio station disk-jockeys, music recording industry executives and, of course, the general public. The “Hot 100 Airplay” list ranks the songs in terms of the amount of airplay that they receive at radio stations over the last week. The data are compiled based on a national sample of 194 stations electronically monitored 24 hours per day. The “Hot 100 Singles Sales” list ranks the songs in terms of the number of sales of singles (records and compact disks) in the last week. The data are compiled from a national sample of point-of-sales equipped retail stores. Both lists provide the ranking of the song on the list from the previous week and the number of weeks that the song has been on the list. The music industry (i.e., recording, radio and television) uses this information to track trends and guide airplay.

Despite the names of these lists, each contains only the highest ranked 75 songs. As we wish to compare the songs in terms of airplay and sales, we will only consider the songs that are on both lists. The data are the 44 songs on both lists for the week ending September 10, 1994.

What is the relationship between the amount of airplay a song receives and the retail sales of that song in a given week? Do some songs appear to be different?

Do songs that receive more airplay in a previous week sell more this week?

One useful measure of the change in sales of a song is the change in ranking of the song on the “Hot 100 Singles Sales” list this week from last week (positive values mean that the song is ranked higher this week than last week). Similarly we can measure the change in airplay of a song by the change in ranking of the song on the “Hot 100 Airplay” list this week from last week.

Is there a relationship between the change in sales of a song and the change in airplay?

How does the change in sales of a song depend on the number of weeks that it has been on the chart? How does the change in airplay of a song depend on the number of weeks that it has been on the chart?

Another look at the “Old Faithful” geyser and adoption visas

Topics covered: Examination of univariate data.

Data Files: `geyser1.dat`, `adopt.dat`

As we saw in the two cases “Eruptions of the ‘Old Faithful’ geyser” and “International adoption rates,” the histogram can be a very useful way to investigate the characteristics of a variable for a set of data. The histogram is (the simplest) example of what is called a **density estimator**. As occurs for all density estimators, the appearance of the histogram is dependent on the degree of smoothing chosen by the data analyst. For the histogram, smoothness is determined by the width of the histogram bins (or, equivalently, the number of bins); large bins imply more smoothness. In addition, the form of the histogram is dependent on the “anchor” of the histogram — that is, the beginning value of the first bin in the histogram.

In this case, the effect of these choices on a histogram estimator is investigated.

For both the “Old Faithful” intereruption times and the (logged) adoption visa counts, construct histogram estimates based on a wide range of bin widths. How stable are the resultant estimates when the bin width is varied? If the appearance of the histogram does not change very much when varying the bin width over a reasonably wide range, then the data analyst can feel confident that any observed patterns are genuine. On the other hand, if the appearance changes in a fundamental way depending on the bin width, any observed patterns when using a particular bin width might just be an accidental result of that choice, and cannot be trusted as much. Do the histograms for these data sets vary enough so as to cast doubt on the impressions discussed in the two earlier cases? Is the appearance of the histogram strongly dependent on the choice of the anchor position? Which choice seems to have a stronger effect on the appearance of the histogram — the bin width or the anchor position?

Random drug and disease testing

Topics covered: Conditional probability.

Key words: False negative. False positive. Probability. Sensitivity. Specificity.

A controversial political issue in recent years has been the possible implementation of random drug and/or disease testing (e.g., drug testing of transportation workers, or testing medical workers for HIV [human immunodeficiency virus, which causes AIDS]). Such testing inevitably involves the tradeoff of the benefit to society of detecting drug abuse or disease versus the cost to the individual of potential invasion of privacy.

Consider, as an example, HIV testing. The standard test is the Wellcome Elisa test. For any diagnostic test, the two key attributes are summarized by **conditional probabilities**:

- (1) the *sensitivity* of the test:

$$\text{sensitivity} = P(\text{Positive test result} \mid \text{Person is actually HIV - positive})$$

- (2) the *specificity* of the test:

$$\text{specificity} = P(\text{Negative test result} \mid \text{Person is actually not HIV - positive})$$

According to the Food and Drug Administration, the sensitivity of the Elisa test is approximately .993 (so only .7% of the people who are truly HIV-positive would have a negative test result), while the specificity is approximately .9999 (so only .01% of the people who are truly HIV-negative would have a positive test result).

That sounds pretty good. However, these are not the only numbers to consider when evaluating the appropriateness of random testing. A person who tests positive is interested in a different conditional probability: $P(\text{person actually is HIV-positive} \mid \text{a positive test result})$. That is, what proportion of people who test positive (which would result in great mental distress, possible employment ramifications, etc.) actually are HIV-positive? If the incidence of the disease is low, most positive results could be *false positives*.

W.O. Johnson and J.L. Gastwirth, in a 1991 paper in *Journal of the Royal Statistical Society, Series B*, estimate the incidence of HIV-positive in the general population of people without known risk factors to be .000025. Consider a group of 10,000,000 people without known risk factors. We would expect that about 250 of these people are actually HIV-positive. Based on the sensitivity figure of .993, $(250)(.993) \approx 248$ of these people would have positive Elisa tests. Now think about the 9,999,750 true HIV-negative people. Based on the specificity figure of .9999, $(9,999,750)(.0001) \approx 1000$ false positive results would come from this group. That is, **80% of the positive test results would actually come from people who were not HIV-positive!**

Now, before we decide that diagnostic tests must be a waste of time, we should note that the test *has* dramatically increased the chance of detection; while only .0025% of all people with no known risk factors are HIV-positive, almost 20% of those people with a positive Elisa test are HIV-positive (an increase by a factor of almost 8000!). Still, there are many false positives. This is the problem with random testing for the presence

of a rare characteristic. Public policy must consider the potential damage to people who are falsely labeled as having a condition with negative public perception, such as drug abuse, alcohol abuse, or infectious diseases. This question will only become more important in the future; according to a survey of 630 major companies undertaken by the American Management Association in 1992, the proportion of companies conducting drug testing in the workplace rose from 22% in 1987 to 75% in 1991 to 85% in 1992. The survey also found that the proportion of people testing positive dropped during that time. This suggests that fewer drug users are in the workplace or applying for jobs, but it also suggests that therefore a larger proportion of the positive test results are actually false positives. Firms must be careful about how they apply such tests, however; in July 1993 a woman fired from her job because of a (subsequently determined to be false) positive drug test sued to get her job back, claiming that her firing was illegal under the Americans with Disabilities Act. For this reason, most private firms follow Department of Health and Human Services guidelines for testing federal employees, which require rigorous test standards with backup medical corroboration for someone who tests positive.

A study in the United Kingdom in the late 1980's confirms the numbers given here. Out of 3,122,556 blood samples taken from people without known risk factors for HIV, 373 tested positive for HIV based on the Elisa test. These samples were then retested using the much more specific (and expensive) Western Blot test, and 64 cases were confirmed. That is, 83% of the positive test results were false positives.

It should be reemphasized, by the way, that the HIV-positive incidence rate given above is for people without known risk factors (roughly speaking, people who are permitted to volunteer to give blood at most blood centers). The HIV-positive incidence rate is much higher in the general population; for example, the World Health Organization estimated it (in 1991) to be .00143 for women and .0133 for men, respectively, in the general population in North America; .002 for women and .008 for men, respectively, in Latin America; .0007 for women and .005 for men, respectively, in Western Europe; and .025 for both men and women in sub-Saharan Africa. Thus, there would be a smaller proportion of false positives in random testing from the general population than from testing only those people with no known risk factors. Still, false positives could still be a problem: in North America, in general random testing of 10,000,000 people, there would be about 73,650 true HIV-positive people, with approximately 73,134 being correctly identified as HIV-positive; there would be 9,926,350 true HIV-negative people, with approximately 993 false positives from this group. Thus, over 1% of the positives would still be false positives. Of course, it also would be very expensive to administer tests to the entire population, but that's not a directly probabilistic issue.

We also must recognize that the numbers presented here are for the United States and United Kingdom, two countries with relatively sophisticated medical establishments. The British medical journal *The Lancet* reported in June 1992 that the mass screening for HIV that is a major component of AIDS control strategy in Russia has led to gigantic false positive rates: approximately 99.4% in 1990, and 99.8% in 1991. Multiple administration of the less specific screening tests only worsens the problem (since a positive result in *any* of the tests is considered an overall positive result): in pregnant women, for whom two HIV tests during pregnancy are mandatory, the false positive rate exceeded 99.9%.

As medical technology advances, these kinds of issues will become more impor-

tant. For example, the February 4, 1993, issue of *The New England Journal of Medicine* included an article about a simple new test for HIV infection that is appropriate for newborn babies (the usual tests, based on looking for antibodies to the virus, are not accurate for infants; only one-third of infants born with maternal antibodies are actually infected, and it takes more than a year for the antibody tests to be accurate). The new test (called an *immune-complex-dissociated HIV p24 antigen assay*), is inexpensive (about \$80), and various states are considering whether to use it in routine testing of pregnant women and/or infants. Note, however, that unless the specificity of the test is **extraordinarily** high (or the test would only be applied to children of mothers who have been confirmed as being HIV-positive), there are likely to be far more false positives than true positives in this kind of testing. Considering that it might be impossible (or very expensive) to separate out accurately the true positives from the false positives (at least until the baby is a year old), routine application of the test could lead to great emotional distress for parents, and possibly incorrect (or dangerous) treatment of children. The article did refer to one study of the test conducted at UCLA, where 100% of babies in the study that were born of infected women were correctly classified as to whether or not they were infected. Unfortunately, the test only had 29 babies in it, so it cannot be considered to have resolved this question completely.

The issues here are not limited to HIV testing. For example, a study done at a Lyme disease clinic in Boston in 1992 found that of 800 patients who had been referred to the clinic because of positive results on blood tests, 45% were false positives — people who did not, in fact, have Lyme disease. Considering the virtual impossibility of preventing Lyme disease, short of eradicating it in the ticks that are its host, this translates into many patients receiving inappropriate medical care all over the country. Another example of this type of situation is the prostate specific antigen (PSA) blood test for prostate cancer in men. Autopsy studies suggest that about half of all men over 50 have cancerous cells in their prostate, but only about 2.4% die of prostate cancer. The PSA test has high false positive and false negative rates (e.g., about 50% of the men with high PSA readings do not have prostate cancer); when a person tests positive, there are other tests to do, which are often inconclusive. Given that even if the patient has cancerous prostate cells, it might never pose a health threat, it is a very difficult decision to make whether PSA tests should be given routinely or not. The current recommendation by the American Cancer Society is that men over the age of 50 should have an annual PSA test, along with a digital rectal exam, although a statistical analysis published in the September 14, 1994 issue of the *Journal of the American Medical Association* suggests that the benefits of such screening are marginal at best.

By the way, the same sort of confusion of conditional probabilities that occurs when mistaking the sensitivity of a test for the proportion of positive test results that correspond to people who actually have the disease (true positives) occurs in other contexts as well. Consider, for example, the use of DNA “fingerprinting” as evidence in a criminal trial. Laboratory (forensic) evidence addresses the question “What is the probability that the defendant’s DNA profile matches that of the crime sample, given that the defendant is not guilty?”. This is not, however, the question of interest to the jury, which is “What is the probability that the defendant is not guilty, given that the DNA profiles of the defendant and the crime sample match?”. Confusion of these two probabilities is commonly called “the prosecutor’s fallacy.” The statistician Peter Donnelly made the argument that the two probabilities can sometimes be quite different

in a successful appeal of a rape conviction in the United Kingdom in 1993.

Summary

Application of random drug and disease testing has been suggested in recent years as a way to address the presence of drug abuse and infectious disease in the population. Unfortunately, if the attribute being tested for is rare in the population, even very accurate tests are likely to produce a large proportion of false positive results. For example, it can be estimated that in random testing for the HIV virus in people with no known risk factors, over 80% of the positive test results are, in fact, false positives. This reinforces the notion that random testing for a rare characteristic can have important public policy implications, in terms of invasion of privacy, and needs to be carefully considered before implementation.

Technical terms

Conditional probability: a probability value that involves restricting attention to the subset of possibilities defined to be consistent with the occurrence of a specific condition. The notation $P(A | B)$, read “the probability of A given B ,” is defined to represent the probability of the event A occurring, given that the event B has occurred (or will occur).

Amniocentesis, blood tests, and Down's syndrome

Topics covered: Conditional probability.

Key words: False negative. False positive. Probability. Sensitivity. Specificity.

Down's syndrome is a genetic disorder, caused by the inheritance of an extra copy of chromosome 21, which leads to retarded mental and physical development. It is well established that the risk of having a Down's syndrome baby rises with the mother's age. While the overall risk is 1 in 700, it rises to 1 in 270 for mothers at age 35 (and is even more probable for older mothers).

Because of this, amniocentesis, a procedure used to withdraw amniotic fluid containing fetal cells from the mother's uterus, is routinely recommended for pregnant women over age 35. This procedure is virtually 100% sensitive in its ability to detect Down's syndrome, but it carries a small risk of causing miscarriage. It is also not inexpensive, with an average cost of about \$1000 in the United States. An alternative procedure, chorionic villus sampling, also carries the risk of miscarriage.

An article in the April 21, 1994, issue of *The New England Journal of Medicine* described the results of a study of the effectiveness of blood tests to detect Down's syndrome. These blood tests are not believed to create any risk of miscarriage. According to the article, a study in which 5385 pregnant women over age 35 were administered both the blood tests (which cost about \$75) and amniocentesis indicated that the sensitivity of the blood tests was .89, while the specificity was .75 (see the case "Random drug and disease testing" for the definitions of these terms). A "positive" result for the blood tests corresponds to an estimated probability of Down's syndrome exceeding 1 in 200 from the blood tests, which would then be followed by an amniocentesis.

We will use the definition of conditional probability to learn more about these blood tests. First, recall the definition of conditional probability:

$$P(A | B) = \frac{P(A \text{ and } B)}{P(B)},$$

for events A and B . From the definitions of sensitivity and specificity, we therefore know that

$$\begin{aligned} &P(\text{Positive test result} | \text{Presence of Down's syndrome}) \\ &= \frac{P(\text{Positive test result and Presence of Down's syndrome})}{P(\text{Presence of Down's syndrome})} \\ &= .89 \end{aligned}$$

and

$$\begin{aligned} &P(\text{Negative test result} | \text{Absence of Down's syndrome}) \\ &= \frac{P(\text{Negative test result and Absence of Down's syndrome})}{P(\text{Absence of Down's syndrome})} \\ &= .75 \end{aligned}$$

An important question is as follows: what proportion of women over age 35 would test positive on the blood tests? Note that such women fall into two mutually

exclusive groups: those who test positive and for whom Down's syndrome is present, and those who test positive and for whom Down's syndrome is absent. The first group comprises 0.33% of the population of women over age 35, while the second group comprises 24.91% of the population. (Why? Here's a hint: you'll need to use the overall incidence of Down's syndrome in the population, and the sensitivity [for the first group] and specificity [for the second group] of the blood tests.) Pooling these values, the probability of a positive test result is .2524.

What proportion of these positive results are *false* positives? That is, what is

$$P(\text{Absence of Down's syndrome} \mid \text{Positive test result})?$$

This can be calculated easily (based on what we've already calculated) as .98694. That is, almost 98.7% of the positive test results are false positives. Since any positive test would be followed by an amniocentesis, this is not necessarily a problem, as long as the mother knew that even after a positive blood test the risk of a Down's syndrome baby is still less than 1 in 75.

Another point of concern is the rate of false negative results. What proportion of women over age 35 who test negative on the blood tests (and might therefore not undergo an amniocentesis) would in fact give birth to a Down's syndrome baby? We know that 74.76% of all blood tests would be negative. Using the same kind of calculations that lead to the false positive rate, we can determine the false negative rate to be 0.0546%, or less than 1 in 1800. Thus the use of the blood test to "screen" women ensures that amniocentesis is only applied to higher risk women (a 1 in 75 chance of Down's syndrome) rather than the general population of women over age 35 (a 1 in 270 chance of Down's syndrome). The blood test reduces the health costs and risk of miscarriage for 3 out of 4 women, while leaving a 1 in 1800 chance for the birth of a Down's syndrome baby.

Perceptions of the New York City subway system: safety and cleanliness

Topics covered: Independence. Joint distribution. Marginal distribution.

Key words: Cross-classified data. Random variable. Sample proportions.

Data File: subway.dat

The New York City subway system is one of the largest and most complex rapid transit systems in the world. It provides service to 469 stations, 24 hours a day, seven days a week, using 714 miles of track. In recent years, the Metropolitan Transit Authority (MTA), the quasi-governmental organization that operates the subways (through its subsidiary agency the New York City Transit Authority) has spent millions of dollars to try to improve customer service and the perceptions of the subway system by New Yorkers. Considering that the operating deficit of the MTA was projected to be \$5 billion in 1994, the financial viability of the MTA is obviously predicated on its ability to attract and retain new and current riders. In order to do this, the agency needs to know what the perceived strengths and weaknesses of the subways are, so that it can develop appropriate strategies to address these perceptions.

A survey of public perception of the New York City rapid transit system engaged by the MTA is the basis of the data set analyzed here. The sample consists of 62 graduate students at the Leonard N. Stern School of Business at New York University, who were sampled in the Spring of 1994. The data were provided by Sundar Polavaram and L. Brooke Squire.

Attention will be focused here on four questions relating to the respondent's satisfaction level of the cleanliness of the stations (CLNSTAT), the cleanliness of the trains (CLNTRAIN), the safety in the stations (SAFSTAT) and the safety on the trains (SAFTRAIN). The responses are coded into a five-point scale, ranging from "Very unsatisfactory" (1) to "Very satisfactory" (5). Each of these variables is a **random variable**, in that it represents a rule that assigns a number to each outcome of a random process.

It might be expected that an individual's responses related to cleanliness would be similar to each other, as would be their responses related to safety. Is that the case? One way to answer this question is to construct the **marginal distribution** of each of the variables; that is, a table that reports the sample proportions that fall into each category of the variable.

Construct marginal distributions for the two cleanliness-related variables CLNSTAT and CLNTRAIN. Do the distributions look similar? Actually, there are noticeable differences, with many more people expressing dissatisfaction with the cleanliness of the stations than of the trains.

Now construct marginal distributions for the two safety-related variables SAFSTAT and SAFTRAIN. Do these distributions look similar? In fact, they do look similar to each other, and to that of CLNTRAIN.

Do these results imply that the safety-related variables are more associated with each other than are the cleanliness-related variables? Or that the safety-related variables are associated with CLNTRAIN? Not really. The reason is that the association between different variables cannot be seen from marginal distributions, which only examine the

variables one at a time. Rather, it is necessary to examine the **joint distribution** of the two variables, which summarizes their joint relationship (this is analogous to the difference in information that can be gleaned from two histograms versus a scatter plot).

Are the perceptions of cleanliness and safety related? Construct the joint distribution of the variables CLNSTAT and SAFSTAT. What do you notice about the pattern of sample proportions in the table? The probabilities are higher for pairs of values that are close to each other; that is, when the respondent rated cleanliness low (for example), he or she also tended to rate safety low. Respondent perceptions about cleanliness and safety in stations appear to be directly related to each other.

This pattern is suggestive, but it could be that it is actually not very surprising, even if there was no relationship between the variables. When two random variables are completely unrelated (that is, knowledge of the value of one gives no information about the value of the other), they are said to be **independent**; what would the joint distribution of CLNSTAT and SAFSTAT look like if they were independent?

Mathematical theory can help us to answer this question. Another way of saying that two events A and B are independent is that the conditional probability of A occurring given B occurs is the same as the unconditional probability that A occurs (without knowing anything about whether B occurs); that is, $P(A | B) = P(A)$. But, by the definition of conditional probabilities, if A and B are independent, then

$$P(A | B) \equiv \frac{P(A \text{ and } B)}{P(B)} = P(A)$$

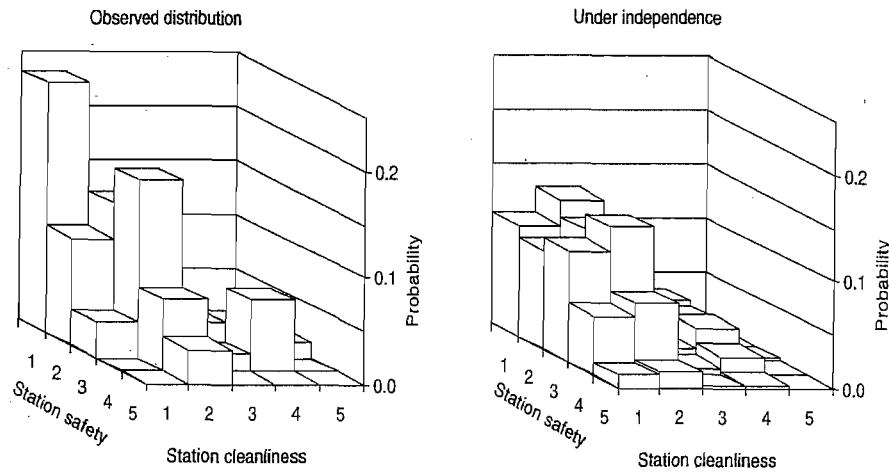
$$\Rightarrow P(A \text{ and } B) = P(A) \times P(B).$$

This mathematical identity can be used to compare the observed joint distribution of CLNSTAT and SAFSTAT to the distribution that would be expected under independence of the two random variables. For example, under the assumption of independence,

$$P(\text{CLNSTAT} = 1 \text{ and } \text{SAFSTAT} = 1) = P(\text{CLNSTAT} = 1) \times P(\text{SAFSTAT} = 1);$$

the two marginal probabilities on the right side of the equation can then be estimated from the appropriate sample proportions from the marginal distributions to yield an estimated joint probability under independence.

Construct the estimated joint distribution of CLNSTAT and SAFSTAT under independence. Does it look similar to the observed joint distribution of the two variables you constructed earlier? The answer is no. The observed probabilities are too large for close values of the two variables, and too small for values that are more different. The following two plots illustrate the difference in the two distributions. Note that under independence each row or column follows a similar pattern of increase or decrease, while the observed joint distribution shows the “piling up” of probabilities for similar values of CLNSTAT and SAFSTAT.



Thus, it appears that there is a real association between the perceptions of cleanliness and safety in subway stations for this sample. Are cleaner stations truly safer, or do riders just *feel* safer in cleaner stations? We cannot answer that question from these data, but subway crime statistics might shed some light on this question.

Now construct the observed joint distribution of CLNTRAIN and SAFTRAIN, the variables referring to trains (rather than stations). Compare this to the estimated joint distribution under independence of the two variables. Do the variables appear to be independent?

We might expect that the two cleanliness-related variables would be strongly related, as would the two safety-related variables. Is that the case here?

Technical terms

Independence: when knowledge of the occurrence of one event gives no information about the occurrence of another event. Two random variables are **independent** if knowledge of the observed value of one random variable gives no information about what the value of the other random variable is (or will be). The existence of independence of two events A and B can be defined as $P(A|B) = P(A)$, or, equivalently, $P(A \text{ and } B) = P(A) \times P(B)$. For contingency tables (tables of counts), independence is consistent with the probability in a cell equaling the product of the marginal probabilities of falling in that row and column; that is, $P(\text{Value falls in row } i \text{ and column } j) = P(\text{Value falls in row } i) \times P(\text{Value falls in column } j)$.

Joint distribution: a set of numbers that represents the probabilities of the possible outcomes for two random variables taken as a pair. That is, the joint distribution assigns probabilities to events of the form { Variable 1 equals x and Variable 2 equals y }.

Marginal distribution: a set of numbers that represents the probabilities of each possible outcome occurring for one random variable.

Random variable: a rule that assigns a number to each possible outcome of some random process. The numbers can have physical meaning (e.g., the number of people

arriving at a location in a given time period), or can be assigned arbitrarily (e.g., 1 if the person arriving is female, 0 if the person is male).
