Statistics

Washington University Math Circle

Jimin Ding, 11/13/2016

Nowadays, predictions and decisions are often made based on data numbers in context. Statistics is an art and science of collecting, summarizing, analyzing and interpreting data. Today we will taste statistics through some interesting examples.

1 Sampling

When the population of interest (may think as all the possible data) is too big or too expensive to study, one may "sample" a portion of it to study. Just like when you are in the ice cream store, you may ask the seller to taste some flavor, and you can know whether you like the flavor of the bucket, although you did not take the whole bucket. This process of selecting units from the whole set is called "sampling". Most people have a sense about how to sample to draw a conclusion. But sometimes that sense can confuse us.

1. In 1990s, a medical study was conducted to compare several treatments of kidney stones, and found that 273 out of 350 patients who underwent open surgery were successfully cured, while 289 out of 350 who underwent noninvasive percutaneous nephrolithotomy.

   (a) If you have a friend or relative who has kidney stone problem, what treatment would you suggest him/her?

   (b) Actually for kidney stones, the treatment is often assigned based on the size of stones instead of random. The following table summarizes data from the two subgroups of small and large stones. Now rethink the previous question. If you have changed your recommendations above, please discuss what makes you flip.

	New Treatment	Standard Treatment
Small Stones	81 out of 87 (%)	234 out of 270 (%)
Large Stones	192 out of 263 (%)	55 out of 80 (%)
Total	273 out of 350 (%)	289 out of 350 (%)

2. Based on an exit poll of 2016 presidential election, the poor were less likely than the rich to vote for Trump. Is this really true or possible to be just a similar example of above? Suppose we have the following data (toy-example), what will you conclude?

	Rich	Poor
Black	25 out of 500 (%)	35 out of 500 (%)
White	4750 out of 9500 (%)	350 out of 500 (%)
Total	4775 out of 10000 (%)	385 out of 1000 (%)

Comments:
This is an example of **Simspon's paradox**. There are many real-life examples of Simspon's paradox. For example, in 1970's, UC Berkeley was sued for bias against female applicants in graduate school admission. Statistical professors in Berkeley later showed that women tended to apply to more competitive departments with low acceptance rates, whereas men tended to apply the less-competitive departments high acceptance rates, and in fact, most department admissions are in favor of women.

In real life, a much harder question of Simspon's paradox is: How do we know the comparison wont flip again if we condition on more (e.g., the surgeons skill, patient's age or gender)? This is one of the important questions that triggers the field of individual medicine/precision medicine.

More reading:
`https://en.wikipedia.org/wiki/Simpson%27s_paradox`
`http://www.stat.columbia.edu/~gelman/stuff_for_blog/LiuMengTASv2.pdf`

2 Conditional Probability

The examples of Simpson's paradox suggests that we are often more interested on the likelihood of something happening under certain conditions. This is called **conditional probability**. With the information that the condition A has already occured, we may know better about the event B if they are dependent. It is crutial to understand what is the right kind of probability (or conditional probability) in the context that you are trying to explain.

1. A hundred players were playing a fair game. Only 10 of them were cheaters and used a trick to win the game at 80% of chance, while others have only 50% of chance of winning.

   (a) What is the chance of a randomly choosen player is a cheater?

   (b) If all players played the game once, 8 cheaters and 45 honest players won the game. What is the chance of a randomly selected winner is a cheater?

   (c) Suppose all players played the game six times. What is the probability that a honest player can win all six games? What is the probability that a cheater can win all six games? If you have seen a player who won all six games, will you think he /she is a cheater?

2. A patient saw his medical diagnostic report of some cancer is "+", which indicates the presence of cancer. He immediately searched online and found this type of medical diagnose is very accurate: the sensitivity of the test is 99% (if a person has this cancer, with 99% of chance, his test result will "+") and the specificity of the test is 95% (if a person is normal, with 95% of chance, his test result will be "-"). Of course, the patient was really scared and worried. But his doctor told him that the type of cancer is very rare with only 0.1% prevalence rate, so he does not need to worry too much. Why does the doctor say that? Knowing his diagnostic result is "+", what is the chance (probability) that he actually has this type of cancer?

Comments:

In this question, we used the idea of conditional probability and saw how it is different from marginal probability. Actually the calculation of the conditional probability can be summarized by the **Bayes's Rule**. Consider two events, A and B . Suppose we know that A has occurred (medical diagnostic report shows +). This knowledge may change the probability that B will occur (the person has cancer), which is called "the conditional probability of event B given that A has occurred" and denoted by $P(B|A)$. The Bayes' rule says that

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A|B)P(B)}{P(A)} = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)}.$$

3 Randomness and Inference

It is hard to draw a decision on the pattern of population or predict a future event based on data because of the uncertainty in the data collection process and the selected units. To make an informed decision, it is essential to understand such a randomness and quantify the variability. For example, like in the previous section, a player who won the game once will be think as normal, a player who won the game twice or three times might be viewed as lucky, but a player who won the game too many times in a row will be very suspicious. But how many times of winning in a row will trigger your decision of calling him a cheater?

1. Consider tossing a coin. A coin is called a fair coin if it has equal chance of having a head or tail facing up.

   (a) If you toss a fair coin 10 times, how many heads will you get? What is the probability that you have exactly 5 heads? What is the probability that you have 4 or 6 heads? Is it rare to observe less than 4 or more than 6 heads when you toss a coin?

   (b) If you toss a fair coin 40 times, is it rare to observe less than (40%) or more than 24(60%) of heads?

   (c) If you toss a fair coin 100 times, what is the probability that the number of heads facing up is less than 40 or more than 60?

   Comments: This a phenomenon of central limit theorem. The more times you tossing the coin, the distribution of the proportion of heads is closer to a normal distribution. It always centers at 50%, but the variability is getting smaller and smaller.