Statistics and Sampling

Washington University Math Circle

Jimin Ding, 10/11/2015

Statistics is everywhere in your Life! Have you ever average your midterm and final exam scores? Have you ever read the weather prediction from the news? Have you draw Pie Chart in your science class? If so, you have seen and used statistics, even if we have not studied it. This makes statistics so interesting – you don't need to know a lof of math to use statistics to understand life and solve problems.

1 Average

Statistics is about "statistics", which is a summary of data. There are two major groups of statistics: descriptive statistics, which simply summarize and describe data in a meaningful way to understand the pattern, and inferential statistics, which infers from the sample data to the properties of population data and helps judgements of the probability models. Today, let's focus on the former, descriptive statistics.

Let's first consider how to measure the central tendency (or the location or the average) of the data. For example, if we observe some numbers: $1, 1, 2, 3, 4$, can you use one number (does not have to be one of the observed numbers) to describe the average or the center of the observations? There are more than one ways. For example,

1. **mean** (arithmetic): sum up all the observations and divide by the number of observations, $(1 + 1 + 2 + 3 + 4)/5 = 2.2$.

2. **median**: the middle number, 2.

3. **mode** : the most frequently observed number, 1. They all somehow give you some insights about the central tendency of the data without telling you all individual observations.

Activity plan:

(a) Change 4 to 100, and re-calculate mean, median, and mode. Question: which one may give a better description of central tendency?

(b) Evenly split students into two groups. Each student picks a number

$$A = \{1, 3, 5, 7, ...\}, B = \{22, 24, 26, ...\}.$$

Let each group report their mean. Then let a student with small number in group B to move to group A. Ask group A and B report their new mean. Question: how can both group has increased means? Who from group B move to group A can increase both means again?

(c) After moving a few students from group B to A, Question: is the mean of two group means same as the mean of all students (numbers)?

(d) Let's guess the temperature tomorrow. For simplicity, let's just consider the mean temperature of the day. Checking the weather history, we know the temperature of the past week (Oct. 1-11) is:

$$59, 60, 54, 55, 63, 69, 69, 69, 72, 64, 70, 80$$

and also know the temperature of tomorrow in the past few years (Oct 12, 2005-2014):

$$65, 44, 56, 73, 50, 71, 68, 58, 68, 58.$$

Can you use these data to forcast tomorrow's mean temperature?

Comments:

(a) Mean is sensitive to outliers, but median and mode are not.

(b) This is called **Will Rogers phenomenon** based on his quote: when Okies left Oklahoma and move to California, they raised the average intelligence level of both states. A real life example is medical stage migration. For example, early detection of a cancer leads to the movement of cancer patients from the set of healthy people to the set of unhealthy people, and raises the average lifespan of both sets, even if early detection may not lead to better treatment.

(c) Mean of the means may not be the grand mean. Extension: **Stein's paradox**: there exist combined estimators more accurate than any method that estimate parameters separately, on average (having lower expected mean square error).

(d) The process of using observed data to guess unobserved ones is called **prediction** in statistics (or forecast for future observations). The prediction is is not unique, and different models and data can be used for prediction. What is the "best" prediction?

2 Spread

Besides the average or the central tendency of the data, we often are also interested in spread of the data and want to know how much the data differ away from its central tendency. As a measurement of the spread, we would like it to satisify some intuitive conditions, such as:

- It does not change as the mean of the data shifts.
- It should be proportional to scale change of the data.

- It should not depend on the number of observations.

If we look at the average distance between all observations to the mean, then all above conditions are satisified. For example, if we have data $1, 1, 2, 3, 4$, and the mean of these number is 2.2 (was calculated above), then we can calculate the **variance** as the following.

$$\frac{(1-2.2)^2+(1-2.2)^2+(2-2.2)^2+(3-2.2)^2+(4-2.2)^2}{5-1} = 1.7.$$

Due to the squares, the variance is not in the same magnitude as the original data. So we often take square root of the variance, which is called the **standard deviation**. The standard deviation of above data is $\sqrt{1.7} \approx 1.30$.

Activity plan:

(a) Change 4 to 100, and re-calculate variance. Question: is variance sensitive to outliers?

(b) If we increase one of the 1 to 4, is the mean increased or decreased? Is the variance increased or decreased? What about if we change all numbers to 4?

(c) Does above calculation satisify the intuitive conditions we listed?
Add all numbers by 2 and recalculate the variance.
Multiply all numbers by 10 and recalculate the variance and the standard deviation.
Duplicate the observations: $1, 1, 2, 3, 4, 1, 1, 2, 3, 4$. What is the new variance?

(d) Can you add a number to the list to keep mean and variance unchanged?

Comments:

(a) Variance is sensitive to outliers as mean.

(b) Increaseing the value of observations or increasing the mean may not lead to variance increasing.

(c) The variance (and standard deviation) is invariant of mean shift and proportional to scale change. The variance describes the spread of the data but does not depend on the number of data.

(d) In general, it is impossible. But you may add two numbers.

3 Sampling

Statistics is a branch of mathematics that deals with collection, analysis, interpretation and presentation of numerical data. 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. The following table reports the success records of two treatments for kidney stones in a real-life medical study.

	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 (%)

   Activity plan:

   (a) Split students to 6 groups and let each group report the success rate of each cell.

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

   (c) Form two groups based on the different answers to debate.

   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.

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?

   Activity plan: Start with a simpler experiement of conditional probability.

   (a) Let three students play the cancer patients and other students play normal people. Let two cancer patients get "+" (diagnosed with cancer) and most of normal people get "-"(diagnosed without cancer).

   (b) Draw a Venn diagram to represent the probabilities and conditional probabilities.

   (c) Group all students with "+". Question: What is the percentage of cancer patient given "+" diagnosis? Why does it different from the percentage of cancer patients in the classroom? Is it same with the percentage of "+" diagnosis? Is it same with the percentage of diagnosed cancer within the cancer patients?

   (d) Let students calculate the chance that the poor guy actually has this rare 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. **First Digit Law (Benford's Law)**.

   Activity plan: Find the numbers of pages of all your books and keep the first digits. The first digit could be $1, 2, \cdots, 9$. Count how many 1s you get, how many 2s you get ... Record the percentage of each of nine numbers in a table. Do you observe 1s more frequently than 9s? And in general more frequently observe smaller first digits than larger first digits? Does your table looks like

d	1	2	3	4	5	6	7	8	9
%	30.1	17.6	12.5	9.7	7.9	6.7	5.8	5.1	4.6

   Benford's law says the probability of the first digit being d is

$$P(d) = \log(d+1) - \log(d).$$

4. **Friendship paradox**. Most people have fewer friends than their friends have, on average. (Scott L. Feld 1991.)

   Activity plan:

   Draw a graph to present a symmetric social network. Put each participants' name in a node and draw a edge between two nodes if they are friends. For each node (person), count the number of edges (friends) it has, and find the mean of the counts (the number of friends a person has on average). For each edge (friendship) and an end-node of the edge (a friend), count the number edges (friends of this friend) it has, and find the mean of the counts (the number of friends a typical friend has on average).