

# Questions and Answers 1

Psychology 310

1. *Why does variance tend to be underestimated when sampling from a population?* It is difficult to answer this question at this stage of the course without sounding hopelessly technical or slightly mystical. But here is one “intuitive” way of thinking about it. The population variance  $\sigma^2$  is a “long run” quantity. It represents the long run average squared deviation score, across a potentially infinite number of observations. The sample variance, on the other hand, is computed on sample deviation scores. When you compute sample deviation scores, you reduce the dimensionality of the “score space” by 1, from  $n$  to  $n - 1$ . To see why that is true, consider any list of  $n$  scores. Once you have computed the first  $n - 1$  deviation scores, the last one is already determined, because they must sum to zero. So when computing a variance, the effective number of deviation scores is really  $n - 1$ , and a proper estimate needs to take this into account. Once we get deeper into the course material, I’ll give several more technical answers to this question.
2. *How do the different levels of measurement affect how I should analyze the data?* One could write entire chapters on this question, and I should state at the outset that the issue is controversial. However, at a very basic level, you need to ask whether the statistical questions you are asking are in fact supported by the level of measurement of your data. For example, if data are strictly ordinal, then differences need not be equivalent, and it may not make sense to compute averages. Consider, for example, data on errors in a cognitive task were converted

into ranks for 8 people. Suppose the actual data were as follows:

<i>Errors</i>	<i>Rank</i>	<i>Group</i>
0	1	1
17	2	2
18	3	2
19	4	1
21	5	2
22	6	1
24	7	1
36	8	2

If you compute the mean of the rankings, you find that both groups are equal. However, the mean number of errors is much lower for group 1 than for group 2. When you compute a mean, the implicit assumption is that all distances are equivalent in the data. Once you transform interval or ratio data to ranks, you lose that property, and strictly speaking a mean may not be meaningful.

3. *Have we covered order statistics/expected value of the range/skewness/kurtosis? Expected value of the range and other parts that we did not learn in undergraduate statistics course are still a bit difficult to understand. I have a general conceptual understanding of each of these topics, but not what I would consider an operational mastery. I fear that the test will demand a greater level of understanding. We talked briefly in class about the notion of an order statistic. Most undergraduate courses emphasize a statistical model in which a group of data points are reduced to statistics that are then analyzed. There is another model that is often more appropriate. In this model, we gather a set of data, then order the data in some way, and select data on the basis of that ordering. An example: you gather 100 observations from a normally distributed population, rank order them, and select the largest value. This value is the 100th order statistic. More generally, the  $j$ th order statistic is the  $j$ th smallest value.*

Order statistics, in general, have rather different distributions than a single observation taken at random from the same distribution. So it is *very* important to apply the correct model when you have, in fact, created an order statistic.

An example of the confusion that can be created when people fail to understand that they are analyzing order statistics occurs frequently

in exploratory correlational analysis. Suppose you have data on 20 variables, and you compute the correlations among all 20 of these variables. It turns out there are 190 different correlations that will be printed in a table of correlations. Commercial software such as SPSS will print these correlations in a table, with a  $p$ -value alongside each correlation. Many people sort through the correlations and pick out the largest ones for further analysis. But when you do this, you are now analyzing order statistics, and the  $p$ -values printed by SPSS are actually no longer correct for analyzing whether the sorted correlations are significant, in the sense that a correlation with a  $p$ -value of .05 should no longer be considered significant “at the .05 level.”

The expected value of the range asks the following question. Suppose I take a sample of size  $n$  from a normal distribution. I then rank order the data, and take the difference between the largest value (the  $n$ th order statistic) and the smallest value (the first order statistic) to get the sample range. If I do this over and over, what will be the long run average of that range?

This question can be answered using *order statistic theory* which is generally not taught in elementary statistics courses. However, as we saw in class, we can get a very close approximation to the answer using R to simulate the process described above.

If we realize that the normal distribution is symmetric, we realize that we can also use the same answer to gauge the expected value of the largest observation in our data, and the expected value of the smallest. We simply split the expected value of the range in half and put half the distance on either side of the mean.

As an example, consider the following question. Suppose you sample 200 observations from a normal distribution with a mean of 500 and a standard deviation of 100. What is the expected value of the range in your sample? What is the expected value of the largest score? The smallest score?

We can answer this using the table in the textbook on page 75. It says the expected value of the range is 55 when  $\sigma = 10$  and  $n = 200$ . To convert to  $\sigma = 100$ , we need to multiply this by 10, so the expected value of the range is 550. To compute the expected value of the minimum and maximum, we divide this range in half, getting  $550/2 = 275$ . If we put this distance on either side of the mean of 500, we find that the expected value of the maximum is  $500 + 275 = 775$ .

The expected value of the minimum is  $500 - 275 = 225$ .

4. *Measures of Variability, slide 22: expected value of the range, why is it called 'order statistic phenomena'? How is that related to calculating the expected range? See previous item.*
5. *Table on page 74 of textbook, calculating the expected range. Is this something we should know? Yes!*
6. *Are we expected to be able to use R to simulate a distribution of numbers and then find the expected range?*
7. *How can I find the standard deviation of a group when combining two separate groups together? There is a formula, with an example, given in section 5.12 of the textbook. However, we shall demonstrate an alternative approach in class.*
8. *Will we need to know how to use the computational formulas (e.g., SD, percentiles, skewness,  $\check{E}$ ) if we can do the same thing in R? In a sense, yes, because some questions asked on exams and/or homework may involve manipulation of those formulas.*
9. *What is the difference between discrete and continuous probability distributions? Discrete probability distributions assign probabilities to events in situations where the number of events is countable. Continuous probability distributions are used to model situations where the number of events is uncountably infinite. This was discussed in class, and I suggest you review the lecture recording.*
10. *Out of general curiosity: when we talk about location, shape, and sometimes spread, we are often given them in very 1-D terms. I can imagine ways that these concepts can generalize to N-dimensions but do some of the other ideas we have covered be generalized into multi-dimensions? Typically, we start in statistics courses with *univariate* approaches, considering one variable at a time. Ultimately, some of the ideas we talk about generalize quite nicely to *multivariate* situations, while others are more problematic. Later in the course, I'll briefly discuss the connection between statistics and geometry, and look at some multivariate generalizations.*