Computational Statistics - Seminar 1

Andrew Struthers

January 2023

Honor code: I pledge that I have neither given nor received help from anyone other than the instructor or the TAs for all work components included here. – Andrew

1 Questions

Please explain the following in detail from Chapter 1 and 2 with equations where appropriate.

1. Explain the difference between a **theory** and a **hypothesis**.

   - A theory is considered a set of general principles that could explain findings on a topic, and a hypothesis is a prediction from a theory [Chapter 1 Lecture]. For example, a theory could be the observation that worms always come out when it is raining, and a hypothesis extrapolated from that theory could be that, since birds avoid flying in the rain, the number of worms aboveground increases due to the lack of birds.

2. Explain the **two types of variations** in data collection methods.

   - The two types of variation are Systematic Variation and Unsystematic Variation. Systematic variation is the difference in performance as a result of experimental manipulation [Chapter 1 Lecture]. This could be something like measuring the time it takes to cook food while manipulating the heat during cooking. Unsystematic variation is a difference in performance as a result of unknown factors. An example of this, still using the cooking food example, would be the time it takes five people to cook the same meal. Their skill level with cooking, alertness, or many other factors could impact the results.

3. Explain what is a **skew** and **kurtosis** in terms of a data distribution curve.

   - Skew of a distribution curve is a way to measure the symmetry of the distribution. We expect a distribution to follow the “normal” distribution, but sometimes the distribution has many low values (positive skew) or many high values (negative skew). Kurtosis is the weight of each tail. The tails of a distribution are the values that fall outside of the mode of the distribution. A heavy tail (leptokurtic) means that there are very few datapoints outside of the main distribution, whereas a light tail (platykurtic) means that, while there is still a recognizable mode, there might be many values outside of the "obvious" peak of the graph. [Chapter 1 Lecture]

4. Explain the difference between the **alternate** and **null** hypothesis.

   - The null hypothesis (H_0) is that nothing happens. If we are using the example of the birds and worms from question one, the null hypothesis could be that worms don't come outside more often when it is raining. The alternative hypothesis (H_1) is the hypothesis that we are trying to prove, known as the experimental hypothesis [Chapter 1 Lecture]. Again using the birds and worms example, the alternative hypothesis is that less birds due to rain increases the amount of worms aboveground.

5. Explain the **Central Limit Theorem** with equations.

   - When we are taking samples from different populations, each sample will have its own mean value, known as the sample mean. Since we want to make conclusions about the population as a whole, we must find the standard error of the sample means, so that we can calculate the deviation in each sample mean. The Central Limit Theorem is a way of figuring out how to calculate this. If the sample has more than 30 datapoints, we can use the equation:
   $$\sigma_{\bar{X}} = \frac{s}{\sqrt{N}}$$
   to calculate the standard error of the sample means. The Central Limit Theorem also states that if there are 30 or less datapoints, the sampling distribution takes the shape of a t-distribution [Chapter 2 Lecture].

6. Explain how the **confidence interval** is calculated in *small sample sizes.*

   - When we have a small sample size, we calculate the confidence interval based off of the t-distribution instead of the z-value. We can still use similar equations:
   $$\text{lower boundary of CI} = \bar{X} - (t_{n-1} \times SE)$$
   $$\text{upper boundary of CI} = \bar{X} + (t_{n-1} \times SE)$$
   Notice, we used the value for t from the t-distribution instead of the z-value. The $n - 1$ in the t-value comes from the degrees of freedom, telling us which t-distribution we want to use. [Section 2.5.2.3: Calculating Confidence Intervals in Small Samples, Discovering Statistics]

7. Explain why we use a **95% confidence** (**0.05** probability) interval as a benchmark of statistical measure.

   - The reason that we use $p < 0.05$ as a benchmark of statistical measure is pretty arbitrary. Before there were tools for exact probability calculations (like R), probabilities were selected from tables of "critical values". These critical values usually incremented in such a way that they didn't contain every single possibility in order to save space. These tables influenced statistics in such a way that $p < 0.05$, $p < 0.02$, and $p < 0.01$ (or 95%, 98%, and 99% respectively) became the gold standard for statistical significance. [Page 52, Discovering Statistics]

8. Explain **Type 1** and **Type 2** errors with examples.

    - Type 1 error is when we believe something is an effect in the population we are studying, but in reality there isn't. Type 1 errors are also known as false positives. An example of a false positive test is taking a pregnancy test that says you are pregnant, when in reality you aren't. Type 2 errors are the opposite. Type 2 errors are when we believe that there isn't an effect in in the population, when there actually is. An example of this is you take a pregnancy test and it says you aren't pregnant, even though you are. Type 2 errors are also known as false negative errors [Chapter 2 Lecture].

9. Explain **effect sizes**.

    - Effect sizes are a general way to measure how much of an effect impacts the variance inside of a population. These generalized measurements are a way to explain variance in a more general way, that can be applied to a population that has a different sample size. A high effect size implies that the effect causes large amounts of variance in the data, and a small effect size implies the opposite. We don't want all effect sizes to be small, however. We typically want to reduce variance, but in a study like vaccine success rates or influence from advertisements, having a large effect size typically implies success [Chapter 2 Lecture].