Coursera Stat Notes – Chapter 4 Foundations for Inference

*Introduction*

1. Inference’s main concern is the quality of the parameter estimates in statistics
   1. such as, how close is the sample mean to the true population mean *µ*?

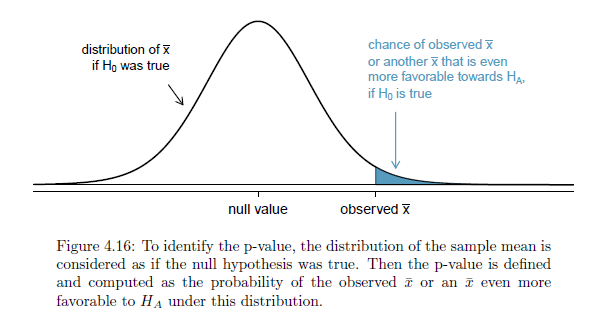
*Variability in estimates*

1. Population VS Sample Mean
   1. Population mean *µ* can be approximated by taking the mean of the sample and inferring the population mean. The sample must be selected randomly for this approach to be acceptable.
   2. Simplest way to find the population mean: just taking the mean of the whole population, if this is possible.
   3. **Point Estimate**: the best guess of the true variable.
      1. taking several estimates from the population and comparing them to each other will show **sampling variation**. If they are good point estimates, the sampling variation will typically converge.
      2. Point estimates of **population parameters** can be approximated too. For instance, the median and standard deviation can be estimated.
2. Point Estimates are not exact
   1. **Running Mean:** the sequence of means where each mean uses one more observation in its calculation than the mean directly before it in the sequence.
      1. RunningMeans tend to converge on the true mean as more data becomes available.
3. Standard Error of Mean
   1. **Sampling Distribution**: taking repeated samples of a population and calculating the same variable.
      1. Formal Definition: The Sampling Distribution represents the distribution of the point estimates based on a fixed size from a certain population. It is useful to think of a particular point estimate as being drawn from such a distribution. Understanding the concept of a sampling distribution is central to understanding statistical inference.
4. Standard Error of an Estimate
   1. The standard deviation associated with an estimate is called the standard error. It describes the typical error or uncertainty associated with the estimate.
   2. The standard error of the sample mean can be computed as the population mean divided by the square root of the of the sample size:
   3. This equation assumes the sample population provided is less than 10% of the total population
   4. When the population standard deviation is unknown, using the population st deviation is typically “good enough” given the sample is random, contains at least 30 observations, and is not strongly skewed.

*Confidence Intervals*

1. Confidence Interval: The possible range of values for the population parameter estimate.
   1. the confidence interval is based on a point estimate, and should contain the true population parameter.
2. 95% confidence interval
   1. a confidence interval that has a 95% chance of containing the true population parameter can be expressed as
   2. very strong approximation when the sampling space takes on a normal distribution shape
3. Sampling Distribution of the Mean
   1. Informal Definition of Central Limit Theorem: If a sample consists of at least If a sample consists of at least 30 independent observations and the data are not strongly skewed, then the distribution of the sample mean is well approximated by the normal model.
   2. If a sample meets the criteria of the Central Limit Theorem, the 95% confidence interval can be amended to be
   3. if a sample meets the criteria of the Central Limit Theorem, the 99% confidence interval can be described as
4. Conditions for the sample mean being nearly normal and the *SE* being accurate:
   1. The sample observations are independent
      1. that is, the sample is random and is less than 10% of the total population
      2. or, the subjects of an experiment were randomly assigned to a group
      3. or, if the sample is from a seemingly random process
   2. The sample size is larger than
   3. The population distribution is not strongly skewed, according to your best judgment.
      1. a strong skew is indicated by prominent outliers
      2. include at least 100 observations if there are prominent outliers
   4. the larger the sample size, the more lenient one can be with the skew
5. Confidence Level for any Interval
   1. to find a confidence level of any percentage (51.8%, 25%, 93%, etc) the confidence level for a given parameter can be expressed as where *z* represents the standard deviation that corresponds to the area under the normal distribution curve.
   2. *z*\**SE* is known as the **margin of error**

*Hypothesis Testing*

1. Null Hypothesis *H0*: the null hypothesis is the claim to be tested. It is also called the skeptical perspective.
   1. often represents no change
   2. Null Hypotheses are not rejected unless there is overwhelming evidence in favor of the Alternative Hypothesis
   3. **Null Value** is the values we test in hypothesis testing. It is the value of the parameter assuming that the null hypothesis is true.
2. Alternative Hypothesis *HA*: the alternative hypothesis is the alternative claim to the null hypothesis. It often represents a wide range of possible parameter values.
   1. often represents the possibility of change
3. A **one sided hypothesis test** only considers a portion of the possible outcomes within the null hypothesis and the alternative hypothesis.
   1. example: Testing to see if students at a school sleep on average the same amount as the national average, or more than the national average, but not considering if they sleep less than the national average.
   2. In Stats, a one-sided hypothesis is not preferred
   3. If the data was collected with a two-sided test in mind then is switched to a one-sided test, more Type 1 errors can be inadvertently introduced. (accidentally rejecting the null hypothesis).
4. A **two-sided hypothesis tests** all possible outcomes
   1. example: testing to see if a new pain medication gives more relief (*H0*) or the same or less relief (*HA*) than an existing pain medication
   2. in Stats, a two sided hypothesis test is preferred.
5. **Decision Errors**
   1. Type 1 Error: when the null hypothesis is rejected but was actually true.
      1. example: convicting an innocent person
   2. Type 2 Error: accepting the null hypothesis when the alternative was actually true.
      1. example: finding a guilty person innocent at trial
   3. Reducing one type of error often means increasing the other type of error.
   4. **Significance Level**: to minimize errors, we do not want to incorrectly reject the null hypothesis more than 5% of the time. This is called the significance level and it is symbolized by α = 0.05
      1. the significance level corresponds to the confidence interval. If we want a confidence interval of 95%, we’ll have to accept a significance level of α = 5%
6. **Formal Testing with p-Values** 
   1. a conditional probability
   2. formal definition: the p-value is the probability of observing data at least as favorable to the alternative hypothesis as our current data set, assuming the null hypothesis is true. We typically use a summary statistic of the data (such as the sample mean) to help compute the p-value and evaluate the hypothesis.  
      
   3. Visually, the p-value is the area under the tail end of the plot showing expected null value, and the actual sample variable when plotted along with the null value. This visualization depends on your assumption that the null hypothesis is true
   4. **Step to evaluating a hypothesis test with p-values**
      1. the Null Hypothesis represents a skeptic’s position or a position of no difference. This position is rejected if the evidence strongly favors the alternative Hypothesis
      2. A small p-value means that if the null hypothesis is true, there is a low probability of seeing an extreme point estimate. This is interpreted as strong favorable evidence of the Alternative Hypothesis.
      3. The null hypothesis is rejected if the p-value is smaller than the significance level α, usually assumed to be α = 0.05. Otherwise, we fail to reject the null hypothesis.
      4. The results of the hypothesis test should be communicated in plain English so that non-statisticians will easily understand the result.
   5. p-values that are less than or equal to the significance level α greatly reduce the number of type 1 errors (erroneously rejecting the null hypothesis)
      1. the null hypothesis is rejected if p-value < α

*Examining the Central Limit Theorem*

1. Informal definition: The distribution of the sample mean is approximately normal. This approximation is poor if the sample size is small, but improves with larger sample sizes.
2. **Confidence Interval:** a confidence interval based on an unbiased and nearly normal point estimate distribution is where *z\** is selected to correspond to the confidence level, and *SE =* is the standard error. is known as the *margin of error*.
3. **How to test the hypothesis using a normal model**
   1. write the hypothesis in plain language. Then, follow with the mathematical notation
   2. identify an appropriate point estimate for the parameter of interest
   3. verify conditions to ensure the standard error estimate is reasonable and that the point estimate is nearly normal and unbiased.
   4. Compute the standard error then draw a picture of the normal distribution with the estimate assuming the null hypothesis *H0* is true. Shade the area that represents the pvalue
   5. using the picture of the normal model compute the test statistic (*Z*-score) and identify the p-value to evaluate the hypotheses.
   6. Write your conclusion in plain language