**Agenda**

* This session will address traditional and contemporary statistical methods.
* Since 1960 there have been many methods developed to deal with non-normal distributions. This is largely due to modern computing capability.

**Statistical Hypothesis Testing**

* The scientific method is based on accumulating evidence in order to make an informed choice between contending beliefs.
* You could believe all swans were white. Once a black one shows up the rational action would be to change your beliefs however cherished they are.
* Statistical hypothesis testing formalizes this process as a choice between a null hypothesis and an alternative hypothesis.
* **The null hypothesis (H0) is a hypothesis which a researcher tries to disprove, reject or nullify. Hence the term, “null” hypothesis.**
* **The alternative hypothesis is an alternative view of what the true state of nature or affairs may be.**
* Clinical trials: Drug has no effect, drug has effect.
* **From a decision making point of view, things are boiled down to a binary comparison leading to one of two alternative courses of action.**
* You are never correct 100% of the time. You could reject the null hypothesis when it is true, and vice versa. These are the Type 1 and Type 2 errors respectively. They are easy to get confused.
* Students have given me examples they use to remember.

**Type I, Type II and Sample Size**

* When dealing with variable and imperfect data, you can never be right 100% of the time. Statistical tests are not mathematical theorems: true or false.
* **Cutoffs have to be established for decision making.**
* The fire alarm example illustrates this. “How sensitive do you want the alarm to be?” What is a suitable tradeoff of receiving false alarms versus missing a fire? This may mean some hard thinking.
* Alpha is used to designate the type 1 error rate. Beta denotes the type 2 error rate. These quantities are related and affected by the sample size.
* **Suppose you adjust the fire alarm setting to be more sensitive. More false alarms mean the chances of not missing a fire are reduced.**
* Hold onto this page. In the next sync session I will provide some graphical displays that visually show the connection between these quantities.

**Sampling Distributions from Normal Population**

* The normal distribution provides a useful starting point for statistical hypothesis testing and confidence interval estimation.
* Here are two examples with n = 10 and n = 40.
* The value of this is the following. If you have a random sample from a normal population and calculate the mean value, you can assume that value is going to have a normal sampling distribution with reduced variance. .
* The z statistic takes advantage of this property.

**Normal Distribution Point and Interval Estimation**

* The first statement is a mathematical statement based on the knowledge the z statistic follows a particular distribution. The second is a statistical inference statement in the form of a confidence interval.
* This example illustrates a process for constructing confidence intervals. A probability statement is used to derive a confidence interval statement.
* Knowledge of the sampling distribution allowed determination of specific values for the **QUANTILES.**
* **Note: The parameter value is fixed. The confidence interval will vary sample to sample. The confidence interval captures the parameter.**

**Central idea** Every time a hypothesis is tested or a parameter is estimated, a sampling distribution is needed for the statistic of interest so that a confidence statement can be made. With the z statistic, it is the normal distribution.

**z-test Examples**

* First example of a statistical hypothesis test.
* Here we have data that is not in accord with the null hypothesis.
* Using a statistical test, a decision results: do not reject the null hypothesis or reject the null hypothesis. The null hypothesis is never accepted.
* More data or a superior designed study may disprove the null hypothesis. Until then, the correct phrasing is that **“null hypothesis was not rejected”**.
* The z-test is not a standard R function. Gives us a chance to practice R.

**Eight-step Process**

* This shows a rigorous approach to hypothesis testing in a research setting. Don’t frame the statistical hypotheses after collecting data or you may be applying a test to some spurious feature of the data.
* There should be investigative work done on historical or other data in advance which results in a study design and formal hypothesis testing.
* Good procedure to follow when doing problems. You can expect this type of hypothesis testing problem on exams.

**The t statistic**

* The t statistic was a step ahead. It allows interval estimation when the variance is unknown. **Normality is still assumed in the definition of the t statistic. This is sometimes overlooked by people.**
* The main difference is that s is used and the t table is used. **QUANTILES**
* Have to take into account the degrees of freedom for the statistic.
* We no longer have the population variance. We have a sample.
* What are degrees of freedom? In this situation, they indicate the extent to which the sample variance is constrained.
* **When using the t tables, enter them using the degrees of freedom.**
* What if the original population is not normally distributed and the variance is unknown? With a sufficiently large sample size, the sample variance converges to the population variance and the t-statistic may be used. **Keep in mind this is an approximation.**

**t-test Example**

* First example shows how to do this in R. Unaware of a function for this.
* Note that in both cases I followed the convention in Black to split alpha since we are doing a two-sided test and did not know beforehand which tail the test statistic would fall into.
* One-sided uses one quantile. Alpha is not split.
* Note the terminology and calculations.
* t.test()

**What has been shown so far are traditional statistical methods based on the normal distribution. Most data encountered in practice are not normally distributed. The central limit theorem is a great help, but by itself is not sufficient in all cases.**

**Central Limit Theorem Convergence**

* What do we do if the parent population is not normally distributed?
* If the distribution is well behaved, the classical central limit theorem applies.
* What does “well behaved” mean? Limited asymmetry, few if any outliers.
* There is no precise answer regarding the sample size, n, due to the diversity of situations that can arise and the degree of convergence needed.
* **The statements made by Black and Wilcox are starting points.**

**So, the z-test can be used if you have a large enough sample size and know the standard deviation. Large n will allow for Student’s t test to be used as well.**

**The exponential distribution**

* To illustrate how this works consider the exponential distribution.
* The variance is the mean squared, which here is 1.0.
* 1,000 random samples N=30, skewness = 0.4890. How well does the tail probability match a normal distribution for inference purposes?

**t statistic Performance—exponential distribution**

* Not knowing the standard deviation requires a larger sample size. A sample size of 30 is not sufficient for the t statistic to converge to the t distribution.
* Note the left skew. Why? It is a consequence of sampling from an asymmetric distribution on the RHS. LHS leads to means tucking in.
* What are the tabled t values? +-2.05 If we had a nice symmetric histogram.
* Boosting the sample size to 100 would make a big difference.
* The recommendations by Black and Wilcox are minimums.
* Large sample size or an alternative method.

**Four Types of Distributions**

* Wilcox introduces this topic in Section 4.7 page 70 of his book. Each cell in this matrix leads to a different analytical approach.
* Black addresses the top two cells of this matrix. Wilcox discusses the bottom two cells. We will look at each of these this evening.
* Use EDA to determine in which cell you are operating.

**Example of an asymmetric distribution. Transformations and Bootstrapping**

**Transformations**

* One method for dealing with situations like this is to transform the data.
* This goes back to the 1940s as a way to apply normal distribution theory.
* There is an extensive literature dealing with transformations. Here are four common ones. The first two have been used with discrete variables. The latter two with continuous variables.
* The use of a transformation raises a number of questions.
  + Can the results be interpreted if not on the original scale of measurement? This can be the case but not always.
  + Keep in mind the mean of transformed data does not correspond to the mean on the original scale of measurement.
  + You may not be testing the parameter of interest.
* Check the course reserves for references dealing with transformations.

**Box-Cox Power Transformations**

* The curvature allows for re-expression of data on a different scale.

**Power Transformation Results**

* The Box Cox Power Transformation is popular. R provides software for identifying the best power transformation for your data.
* AID package with boxcoxnc() function. Variety of diagnostic metrics.
* Lilliefors test is a normality test based on the Kolmogorov–Smirnov test. It tests the null hypothesis that data come from a normal distribution.
* The function in R does a numerical evaluation. For each value of lambda, the statistic is calculated on the transformed scale. Results are plotted.
* You can see the resulting transformation with lambda = 0.15.
* We obtain reasonable symmetry. Note that the outliers are still present.
* The usual next step is to use the t statistic.
* Recognize that when this transformation is applied to generate a symmetric distribution, half of the data falls to the left and half to the right of the mean. On the transformed scale the mean and median are about the same. In effect, you are doing an analysis of the median.
* Note the reverse transformation of the mean value—median. Perhaps the analysis should have been done on the median to begin with.
* In summary, transformations developed so that normal distribution theory could be applied to data. They can be useful depending on the problem.
* We will use the log10() transformation in the second data analysis.

**Bootstrapping**

* **What do you do with a non-normal distribution if your sample size is small and outliers are present? Sampling distribution?**
* **The CLT does not pertain.**
* **Sampling distribution quantiles for a t statistic need to be estimated.**
* **The idea behind all bootstrap methods is to use the data obtained from a study to approximate the sampling distribution of the test statistic so that confidence intervals may be determined and hypotheses tested.**

**Asymmetric Distribution Sampling Distributions n = 40**

* Using the example distribution from the public health paper, let’s take a look at the sampling distributions for the mean value and t statistic with samples of size 40. I’m going to construct the true sampling distributions.
* Resample with replacement from the original population 1,000.
* Calculate the mean and t-statistic from each sample.
* The t-statistic is much affected by the skewed distribution. This is due to the impact of the denominator or standard deviation for the mean.
  + RHS—If a sample has one or more outliers in it, the s is inflated which reduces the size of the t statistic.
  + LHS--- If the sample is drawn from the smaller values in the population, the s is small relative to the mean and the t-statistic is negative and in some cases very negative.
* QQ plot is using the theoretical quantiles for a t-distribution.
* The vertical blue lines are at the 2.5% and 97.5% quantiles of the sampling distribution for this statistic based on the population being sampled.
* The red lines show the theoretical values if you had sampled from normality.
* **Depending on the precision you desire, probability statements involving a traditional t statistic using the tabled quantiles are inaccurate. What you really want to use are the quantiles from the true sampling distribution.**

**What to Do with a Single Sample?**

* The bootstrap approach is to use the data in hand. Resample from it with replacement, and construct an approximation to the true but unknown sampling distribution.
* This can be done with any summary statistic.
* I will discuss two methods for bootstrapping.
* **Do not drop the outliers! Otherwise a bias would be introduced.**

**Bootstrap t Method n = 40**

* A similar approach is used for the bootstrap t method.
* The quantiles are taken from the simulated sampling distribution.
* Note that the bootstrap methods adjust for the skewing of the sampling distribution as compared to the traditional t statistic approach.
* Most people when they first see this think it is cheating.
  + It is true we are limited by the first sample of size 40. However, keep in mind all possible resamples with replacement, 40^40, could have been drawn from the original population. We are using a subset of these to approximate the true sampling distribution for the statistic.
  + If the traditional t statistic is applied only to the 40 values in our sample, it too is limited by these values.
  + Head to head simulation studies of these two methods shows the bootstrap t method to work well in comparison.
  + **When doing an analysis, you can always do both. If the two agree, use the traditional t statistic. If not, go with the bootstrap.**

**Bootstrapping is not a panacea. Need representative data.**

**https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4784504/**

**“**For skewed data, confidence intervals should reach longer in the direction of the skewness; the bootstrap *t* does this well, the percentile method makes about 1*/*3 of that correction.”

**Heavy Tailed Distribution**

* This takes us to the fourth cell in Wilcox’s matrix of distribution types.
* Do not always know the reasons for heavy tails but they do occur.
* One example that comes to mind is with drug testing in which there are two groups of subjects mixed together unbeknown to the investigator. Both groups respond on average the same, but there is a difference in the variability of the response to treatment between the groups.
* Some subjects may be affected by extraneous factors leading to outliers for unknown reasons. That appears to be the case with these data..
* Detecting this type of thing is not easily done with a histogram.
* The QQ plot and boxplot help reveal the heavy tails.
* The main problem is that with the heavy tails, the standard deviation is inflated thus widening a CI and reducing the power of a hypothesis test.
* This is frequently overlooked which affects statistical inferences. May end up making a type 2 error not finding something statistically significant.

**Robust Estimation**

* The nature of the distribution begs for a 20% trimmed mean.
* **What estimate do we use for the standard deviation of the 20% trimmed mean? It is not the sample standard deviation you have studied.**
* **And, if we use it, what sampling distribution is appropriate for confidence interval construction? Bootstrapping.**

**Trimmed t-statistic for bootstrapping**

* This represents a major advance in contemporary statistics.
* The variance for the 20% trimmed mean was difficult to derive.
* The asterisk indicates a resampled mean and variance.

Some background is in order here. Some people will trim data and then calculate the variance of the remaining values. This is theoretically wrong. The remaining values after trimming are no longer independent, so the usual sample variance formula is incorrect. The proof of this goes beyond the scope of the course. The 20% winsorized variance has been shown to be the correct sample variance for a 20% trimmed mean. That is main purpose.

**Comparison of methods**

* No trimming vs bootstrap
* By using the trimmed mean in this situation you get the equivalent of a sample of size 135, an improvement of a factor of 2.25.

**Confidence Interval Construction**

**Every time you construct a confidence interval, you are assuming a specific sampling distribution applies to the test statistic.**

**There are many situations with different statistics for which no formula exists for the variance. Or, no table exists for sampling distribution quantiles. In these situations bootstrapping provides one worthwhile solution.**

**Sync Session Learning points**

**Comments on Test #3**

**Extra Credit Problem #2**

* Skewness and kurtosis statistics appear in the literature and I use them in this course as descriptive measures when discussing convergence to normality.
* See how these statistics behave with different sample sizes from normal distributions by generating their sampling distributions.
* Their sampling distributions differ, and so does their convergence.
* I do not perform statistical tests using skewness and kurtosis since the literature indicates it is difficult to obtain good standard error estimates.
* There are superior statistical methods. Second data analysis.
* During EDA, I only use them as descriptive statistics in conjunction with histograms, boxplots and QQ charts when it appears the distribution is basically symmetric and well behaved.

**Extra Credit Problem #3**

* The third extra credit problem will give you experience with bootstrapping.
* You will use the earthquake magnitude data to estimate a variance.
* The traditional method using the chi square statistic assumes a normal population. That is not the case with these data.
* Have a chance to compare results, and also try out the boot() package.

**Final Exam Reminders and Suggestions**