# Exam 2 Review

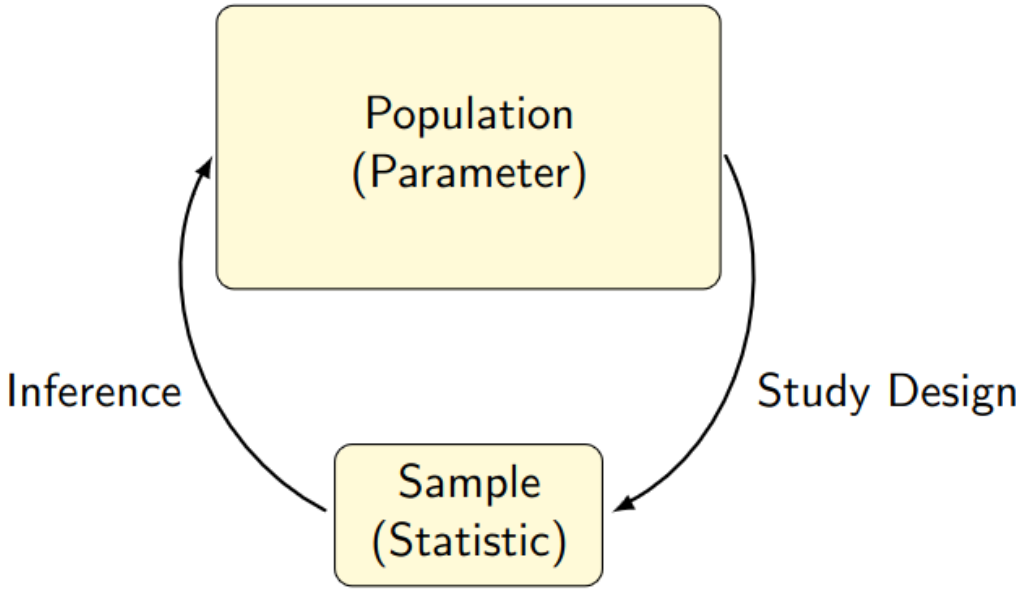
The primary focus of Exam 2 is going to be *statistical inference* – once we have collected data, how do we go about making statements about our population based on what we have seen in our sample. Critical topics will include:

1. Sampling and bias
2. Confidence intervals
    - Sampling distributions
    - Handling variability
3. Hypothesis testing
    - Drawing conclusions
    - Types of errors

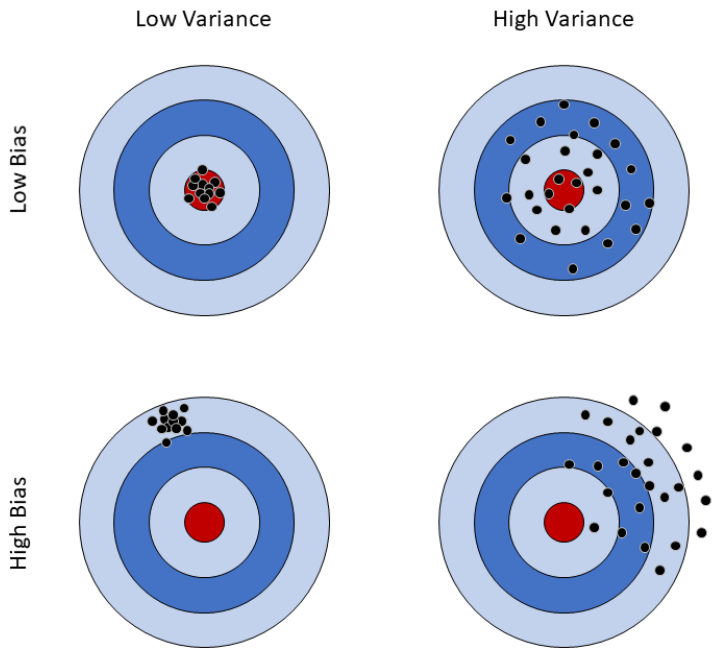
I have tried to cover most of the bigger topics, but everything from class or the slides and not explicitly removed (i.e., computing p-values) will be fair game for the exam.

## Population and Samples

Understand our goal in statistics with respect to the relationship between *Population (parameter)* and *Sample (statistic)*




**Statistical Inference** addresses the question: "how reliably can I expect trends in my sample to reflect what is true about the population?"




Ultimately, we want our sample to be *representative* of our population (what does this mean?)

Types of Sampling

- Convenience sampling
- Simple random sampling
- Stratified sampling

Examples of Bias

- Selection Bias
- Non-response Bias
- Survivor Bias

There is nothing we can do once a sample is collected, so it is critical to get it correct the first time. Additionally, once a sample is collected, there is generally no way to verify if it is representative of the population.

## Sampling distribution

Assuming no bias, we still must account for variability that is built into population. The variability that we see in collecting our sample is precisely the variability we see in our estimate of a statistic (i.e., the mean). In addition to this, variability in our statistic is also impacted by our sample size.

Having sampling distribution is important. We should understand that our point estimate highly likely to not exactly correct. So, based on:

- the point estimate for the mean that we found and
- the amount of variability in our data
- and our sample size

we should have a good sense of what kind of values we might expect, what is a reasonable range, how frequently values should appear, etc.,. This is precisely the information that is contained within the sampling distribution. Note that the standard deviation of a sampling distribution represents the *standard error*.

Ultimately, the sampling distribution will be used to construct *confidence intervals*, an interval containing our point estimate offering what might be considered a range of plausible values for our population parameter. You should be familiar with the myriad of ways in which confidence intervals might be used or how they can be modified.

We have several tools at our disposal for estimating what the sampling distribution of our statistic might look like.

**Bootstrapping**

Bootstrapping is the process of resampling from our sample *with replacement* and recomputing our statistic for each of the bootstrapped samples. You should be familiar with how this process works, how it can be used for constructing confidence intervals of varying size, and how it could be used for hypothesis testing.

**Central Limit Theorem**

Broadly speaking, for a population with mean $\mu$ and standard deviation $\sigma$, the central limit theorem says that the distribution of the population mean (or proportion) follows

$$\overline{X} \sim N(\mu, \sigma/\sqrt{n})$$

or

$$\hat{p} \sim N\left(p,\ \sqrt{\frac{p(1-p)}{n}}\right)$$

From this, it follows that we can create a (standardized) Z-score for our test statistic that will follow a standard normal $N(0,1)$ distribution

$$Z = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}}, \qquad Z \sim N(0,1)$$

It is a consequence of the normality assumption that permits us to use the 68-95-99 rule for computing confidence intervals. This is in reference to confidence intervals of the form:

$$\overline{x} \pm C\left(\frac{\sigma}{\sqrt{n}}\right)$$

where $C$ is the "calibration" value. For a standard normal distribution, the calibration values for the 68-95-99 intervals are $C = 1, 2, 3$, respectively.

In practice, however, we rarely know the true value of $\sigma$ and are required instead to use an estimate from our sample, $\hat{\sigma}$. For a sample of size $n$, this results in a $t$-statistic which follows a $t$-distribution with $n-1$ degrees of freedom:

$$t = \frac{\overline{x} - \mu}{\hat{\sigma}/\sqrt{n}}, \qquad t \sim t(n-1)$$

It is critical that you know how to create a t-statistic and understand how each of the terms will modify its value. Additionally, you should have a good understanding on how the sample size impacts the "fatness" of the tails of the t-distribution and what impact that has on the size of a given confidence interval (i.e., a 95% confidence interval will be wider in a $t$-distribution with fewer degrees of freedom)

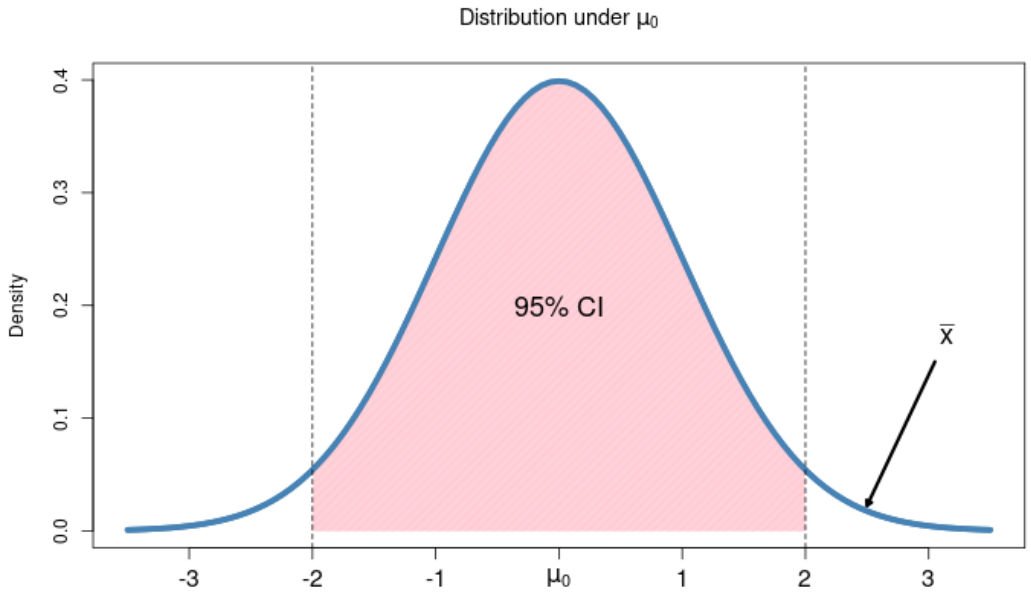
The $t$-distribution has fatter tails, depending on the degrees of freedom, and as such, the calibration value $C$ used in creating confidence intervals will change depending on the distribution. You do not need to know these, but you should understand how they work. For example, a calculated $t$-statistic of, say, $t = 1.8$ will

have a slightly different interpretation depending on the degrees of freedom. In general, the larger the value of $n$, the more "significant" will be a given $t$-statistic. This will be relevant when considering evidence in favor or against the null hypothesis.

As a final note, bootstrapping can *always* be used for finding a sampling distribution and confidence intervals, regardless of the statistic we are estiamting (i.e., centrality ratio, odds ratio, etc.,.). The CLT can be used so long as the estimated statistic is a mean or a proportion.

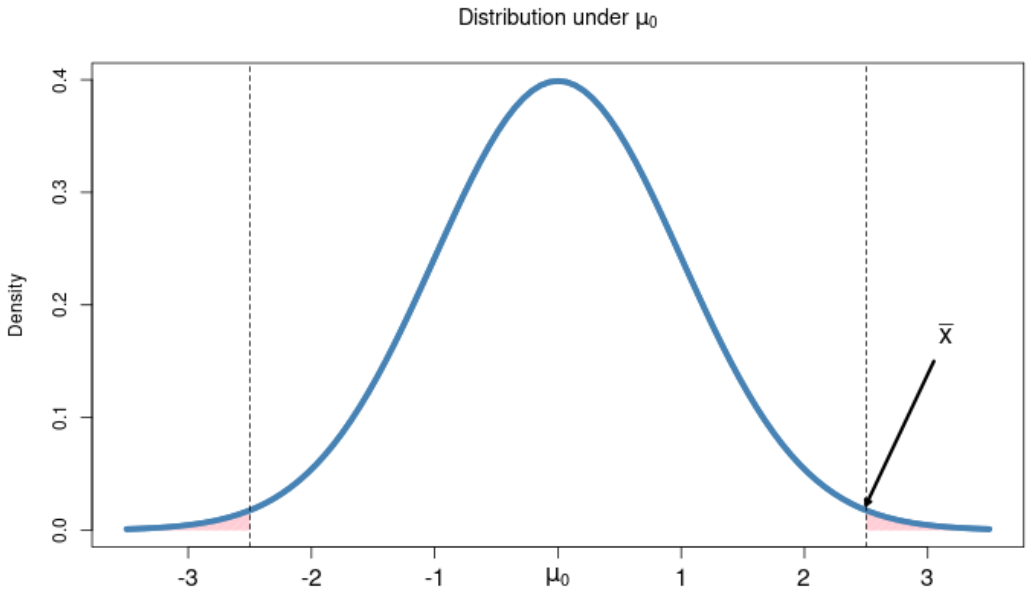
## Hypothesis Testing

Understand the basics of hypothesis testing. Under the null hypothesis ($H_0 : \mu = \mu_0$), we assume that our data follows a specific distribution (standard normal or $t$-distribution) under and determine if our observed test statistic is consistent with that assumption. This typically involves determining a range of values centered at $\mu_0$ and then determining if our observed statistic falls within those values.




### p-values

The $p$-value quantifies the strength of the evidence *assuming the null hypothesis is true.* That is, the $p$-value is given as the probability of having observed what we did (or something greater), assuming the null hypothesis is true. Below, the p-value is the area under the curve shaded in pink.




Understand that there is an intimate relationship between the p-value and a distribution: the more unlikely a value under the null hypothesis, the smaller the p-value should be. You should understand conceptual aspects of the p-value, but you will not be asked to compute it directly.

**Decision Errors**

Ultimately, we are required to make a decision regarding the null hypothesis to either reject or fail to reject, based on the evidence. Depending on the true state of nature (i.e., whether or not the null is correct), we may error in one of two ways. These are called the Type I and Type II errors

| | True State of Nature | |
|---|---|---|
| Test Result | $H_0$ True | $H_0$ False |
| Fail to reject $H_0$ | Correct | Type II Error |
| Reject $H_0$ | Type I Error | Correct |

Understand the relationship between the Type I error and confidence intervals, as well as how p-values would be used, based on our Type I error rate, to make a decision towards rejecting or not rejecting the null. Know what each of these types of errors represent and what the consequences of making them should be.

Given a number of tests with specified Type I and II error rates, you should also be able to fill in a table like the one above. We will have examples in class.

Finally, understand the implications of *multiple testing*, family-wise error, and how that might be mititgated.