STAT 516 Summer 2022

Faith Zhang
Department of Mathematics and Statistics
University of Massachusetts Amherst
yueqiaozhang@umass.edu

Contents

These notes are mainly based on Chapters 8-10 of our textbook: Wackerly, Mendenhall, and Scheaffer's *Mathematical Statistics with Applications*, 7th ed. However, the order of topics might differ significantly from the textbook. In the instances where a review of the concepts covered in STAT 515 is needed, it will be done in lectures. Please let me know if you find typos or mistakes, I will give extra credit as a reward.

1 Introduction

The objective of statistics often is to make inferences about unknown population parameters based on information contained in sample data. These inferences are phrased in one of two ways: as estimates of the respective parameters or as tests of hypotheses about their values. Therefore, this course will mainly cover estimation and hypothesis testing. We first introduce terminology and basic concepts.

1.1 Independent and identically distributed random variables

In statistics, we commonly deal with random samples. A random sample is a sequence of **independent and identically distributed** (i.i.d.) random variables (RVs). Independent means that the sample items are all independent events. Identically distributed means that there are no overall trends, i.e. the distribution does not fluctuate and all items in the sample are taken from the same probability distribution. As we already know from STAT 515 that the i.i.d. assumption has been used in the Central Limit Theorem, which states that the probability distribution of the sum (or average) of i.i.d. variables with finite variance approaches a normal distribution. In this course, we will always assume i.i.d. random variables, which will simplify the underlying mathematics.

Once we observe data, we can use the mathematics of probability and random variables to model the process that generates data. Now suppose we have n i.i.d. real random variables $X_1, X_2, \ldots, X_n$, where n is called the **sample size**. By independence of the X_i's, their joint cumulative distribution function (CDF) factors:

$$F_{\mathbf{X}}(x_1, x_2, \ldots, x_n) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n) = \prod_{i=1}^{n} P(X_i \leq x_i) = \prod_{i=1}^{n} F_{X_i}(x_i) \ .$$

By identical distributions of the X_i's: $F_{X_1} = \cdots = F_{X_n} = F_X$, we have that

$$F_{\mathbf{X}}(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} F_X(x_i) \ . \tag{1}$$

If the X_i's are discrete RVs, then their joint probability mass function (pmf) factors:

$$p_{\mathbf{X}}(x_1, x_2, \ldots, x_n) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \prod_{i=1}^{n} p_X(x_i) \ . \tag{2}$$

Similarly if the X_i's are continuous RVs, their joint probability density function (pdf) factors:

$$f_{\mathbf{X}}(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f_X(x_i) \ . \tag{3}$$

Note that (1) applies to both discrete and continuous RVs.

1.2 Population parameters

Again, the purpose of statistics with i.i.d. data is to use the information contained in a sample to make inferences about the population from which the sample is taken. Populations are characterized by numerical descriptive measures, which we call **population parameters**, **parameters of interest**, **target parameters**, or just parameters. The objective of many statistical investigations

is to estimate the value of one or more parameters. Some important population parameters are the population mean, variance or standard deviation (sd), and proportion.

We often denote a target parameter of interest by θ, which we may also call the **true parameter** value. For example, we may want to know the mean of the distribution

$$\theta = \mu_X = E[X_1] = E[X_2] = \cdots = E[X_n] \ .$$

This is a very common target parameter. Alternatively, we might want to know the variance of the distribution $\theta = \sigma_X^2 = \text{Var}(X_1)$, or the mean of some other function g, i.e. $\theta = E[g(X_1)]$, or even the entire CDF of the data in some cases, i.e. $\theta(x) = F_X(x)$ for all $x \in \mathbb{R}$. Estimation has many practical applications. Here's an example.

Example 1.1. Library or Student Union

Suppose you and your friend are deciding whether to go to the library or the Student Union to work on a class project. Your friend wants to go to the Student Union, but you are concerned that the Student Union could be too crowded, so you wish to go to the library instead. Your friend proposes that you flip a coin to decide. If the coin shows heads, you go to the Student Union; tails, you go to the library. But your friend has proposed this same scheme several times in the past, and it always seems to go their way. You are suspicious that they have a weighted/biased coin that they use for these occasions. How can you tell if the coin is biased? This time your friend agrees to let you flip the coin as many times as you like and see whether the coin is to your satisfaction. Say that $X_i = 1$ if the coin comes up heads. In this case, your target parameter is the probability that the coin shows heads, i.e. $\theta = P(X_i = 1)$. Say that after you flipping the coin 20 times and 15 of them coming up heads, you claim that $\theta = 0.75$, then 0.75 here is called a **point estimate** because a single value, or point, is given as an estimate of θ. But if instead if you claim that θ should fall between 0.7 and 0.8 (or between 0.6 and 0.9 to be more certain), then you are giving an **interval estimate** because the two values may be used to construct an interval $(0.7, 0.8)$ that encloses the parameter of interest. □

The information in a sample can be used to calculate the value of a point estimate, an interval estimate, or both. In any case, the actual estimation is accomplished by using an estimator for the target parameter.

1.3 Point estimators

Recall that a **statistic** is just a function $T(X_1, \ldots, X_n)$ of the observable random variables in a sample and known constants. For example, $X_1 + X_2$, the minimum $X_{(1)} = \min\{X_1, \ldots, X_n\}$ and maximum $X_{(n)} = \max\{X_1, \ldots, X_n\}$ are statistics. The **sample mean**

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

is a well-known and important statistic.

Any statistic can be used as an **estimator** of a population parameter θ. An estimator, often expressed as a formula, tells us how to calculate the value of an estimate based on sample data. The calculated result is our guess about what the value of the population parameter is. For example, the sample mean $\overline{X}$ mentioned above is one possible point estimator of the population mean μ.

We often use "hats" to denote estimators, so $\hat{\theta}_n = T(X_1, \ldots, X_n)$ indicates that $\hat{\theta}_n$ (pronounced "theta-hat n") is an estimator of θ. The subscript n in $\hat{\theta}_n$ explicitly indicates that the estimator is based on sample size n. Sometimes, we simply use the notation $\hat{\theta}$, instead of $\hat{\theta}_n$, as our estimator for θ and not explicitly display the dependence on n, to avoid overly complicated notation. The dependence of $\hat{\theta}$ on the sample size n is always implicit and should be used whenever necessary (e.g. when we talk about consistency later in the course).

Since an estimator $\hat{\theta}$ is a function of the RVs and hence is a random variable itself. Therefore an estimate will not be *exactly* right, meaning that it will not be exactly equal to the true population parameter θ for any given data set. However, if n is large enough, we would hope that our estimate is close to the true θ. We might also want to characterize how close we think our estimator likely is to the truth; this is the goal of **inferential statistics**. To do this, we will need to talk about **sampling distributions** of estimators. The sampling distribution of an estimator $\hat{\theta} = T(X_1, \ldots, X_n)$ is just the distribution of the random variable $\hat{\theta}$. For i.i.d. data, this distribution can be defined in terms of the distribution of X_1. In STAT 515, we talked about methods for finding the distribution of a transformation of a multivariate random variable; these methods can in some cases be used to find the sampling distribution of an estimator. We will need to use some sampling distributions and the Central Limit Theorem that we have learned before.

1.4 Confidence intervals

A point estimate does not tell us anything about how far *a particular* estimate $\hat{\theta}$ is from the true value. To give a sense of how uncertain our estimate is or how confident we are in our estimate for the sample we observed, we need to express uncertainty in statistics via a **confidence interval** (CI), generated from an **interval estimator**. Suppose we want to estimate a population parameter θ based on a random sample $X_1, X_2, \ldots, X_n$. A confidence interval is simply a random interval $[\hat{\theta}_L, \hat{\theta}_U]$ based on the data. We call $\hat{\theta}_L$ the **lower confidence limit**, and $\hat{\theta}_U$ the **upper confidence limit**.

Ideally, the resulting interval will have two properties: First, it will contain the target parameter θ; second, it will be relatively narrow. But like a point estimator $\hat{\theta}$, one or both of $\hat{\theta}_L$ and $\hat{\theta}_U$ are functions of the observed data $X_1, \ldots, X_n$, so they will vary randomly from sample to sample, i.e., they are also random. Thus, the length and location of the interval are random quantities, and we cannot be certain that the target parameter θ will fall between the endpoints of any single interval calculated from a single sample. This being the case, our objective is to find an interval estimator capable of generating narrow intervals that have a high probability of enclosing the true parameter θ. The probability that a (random) confidence interval will enclose θ (a fixed quantity) is called the **confidence coefficient**, denoted $1 - \alpha$.

$$P\left(\hat{\theta}_L \leq \theta \leq \hat{\theta}_U\right) = P\left(\theta \in [\hat{\theta}_L, \hat{\theta}_U]\right) = 1 - \alpha.$$

Here the randomness comes from $\hat{\theta}_L$ and $\hat{\theta}_U$, but θ is a fixed constant (unknown to us). In order for a confidence interval to be useful, the α is a small positive value and often chosen to be $0.1, 0.05$, or 0.01, resulting in 90%, 95%, or 99% CIs, respectively. From a practical point of view, the confidence coefficient identifies the fraction of the time, in repeated sampling, that the intervals constructed will contain the target parameter. Hence we can't be *certain* that the interval contains the true value. For example, if we use $\alpha = 0.1$, then we can say that 90% of all intervals constructed will contain the true θ. Note that any given interval either does or does not contain the true value; the 90% chance comes from repeated sampling. Although we do not know whether a particular interval

contains the truth, the procedure that generated it yields intervals that do capture the true mean in approximately 90% of all instances where the procedure is used.

Intervals of the form $[\hat{\theta}_L, \hat{\theta}_U]$ are called **two-sided confidence intervals**. We can also consider **one-sided confidence intervals**. In this case, we try to find $\hat{\theta}_L$ such that $P(\hat{\theta}_L \leq \theta) = P(\theta \in [\hat{\theta}_L, \infty)) = 1 - \alpha$ for a lower confidence interval, or $\hat{\theta}_U$ such that $P(\theta \leq \hat{\theta}_U) = P(\theta \in (-\infty, \hat{\theta}_U]) = 1 - \alpha$ for an upper confidence interval.

There is a tradeoff between having higher confidence and having a narrower confidence interval. Typically, the width $\hat{\theta}_U - \hat{\theta}_L$ of the confidence interval gets larger as α approaches 0. If we want to be more confident that our interval contains the truth, we have to make the interval wider. Our goal with confidence intervals is to have the smallest confidence interval possible while still maintaining the specified confidence coefficient $1 - \alpha$.

We will cover two main ways to construct confidence intervals. Sometimes we can derive an explicit formula for the sampling distribution of our estimator, and use this sampling distribution to quantify the uncertainty in our estimate; we will cover this in Section 3. In other cases, the explicit sampling distribution is not available, so we use the Central Limit Theorem to asymptotically approximate methods of defining confidence intervals; we will cover this in Section 4.

1.5 Statistical hypothesis testing

Hypothesis testing is another way to mathematically/probabilistically quantify statistical uncertainty. Sometimes, there is a particular value of the population parameter that is of special interest. In the context of confidence intervals, we frequently remarked on whether the interval contained the special value as an indicator of whether that special value was plausible given the data. Hypothesis testing is another way of formally assessing whether a particular value is plausible given the data. The formal procedure for hypothesis testing is to pose a hypothesis concerning one or more population parameters—that they equal specified values, then sample the population and compare the observations with the hypothesis. If the observations disagree with the hypothesis, we reject it; otherwise, we conclude either that the hypothesis is true or that the sample did not detect the difference between the real and hypothesized values of the population parameters. Let us consider the Library or Student Union example again.

Example 1.2. Library or Student Union

You are suspicious that your friend has a biased coin. However, in the absence of any other information, you would typically be inclined to believe that the coin is fair, i.e. 50% probability of showing heads. Unless someone provides convincing evidence, you wouldn't change that belief. Therefore, since most coins should be fair based on common sense, you propose the hypothesis that your friend's coin is fair, and then seek evidence to contradict it if the hypothesis is false. Then you flip the coin 20 times and 15 of them come up heads. If the coin was fair, you would expect approximately an equal number of heads and tails, not too many heads or too few heads. Then based on your data, you conclude that the hypothesis is not in agreement with the outcome, hence the coin is biased. Or you may argue that 15 heads in 20 flips is not enough evidence to say that the coin is biased. □

1.5.1 Elements of a statistical test

In a hypothesis test, we call the theory that we wish to support the **alternative hypothesis** (or **research hypothesis**), and the converse of the alternative hypothesis the **null hypothesis**.

Suppose we want to prove that some new drug is effective at treating some disease. We do this by attempting to show that the converse is false. That is, we start by assuming that the treatment is NOT effective, and we attempt to disprove our assumption. Thus, support for one theory is obtained by showing lack of support for its converse, similar to a proof by contradiction.

Often, the null hypothesis represents a sort of baseline about the world, and the alternative hypothesis is something that it would take convincing evidence to believe. In this way, there is an asymmetry between the null and alternative hypotheses. In the Library or Student Union example, your null hypothesis is that the coin is fair, and the alternative that you attempt to prove is that it is biased. Similarly, an overwhelming amount of scientific evidence indicates that the earth orbits the sun, and not the other way around. It would take extremely convincing evidence that the sun orbits the earth for you to change your mind. Hence, your null hypothesis is that the earth orbits the sun. Finally, we should start with the belief that a medical treatment is not effective for addressing some condition or disease, since most candidate treatments are not effective, so our null hypothesis is that the treatment is ineffective.

Therefore, the goal of hypothesis testing is to determine, based on observed data, whether there is evidence to reject the null hypothesis. That is, we are making a decision about whether the data support the alternative hypothesis when comparing the observed sample with theory. Strictly speaking, we can reject the null hypothesis, but **not accept** it. That's because of the way we set up the hypothesis testing problem. We are trying to see whether there is evidence that our alternative hypothesis holds, not whether there is evidence that the null hypothesis holds. So we will **fail to reject** the null hypothesis when we don not have enough evidence to reject the null hypothesis.

As always, we assume that we have n i.i.d. observations $X_1, \ldots, X_n$ with a common distribution function F_X, and that we are interested in some population parameter θ of this distribution. A statistical test has four elements:

1. **Null hypothesis** H_0. This is a value θ_0 or a set of values Θ_0 that represent the "uninteresting results". We write $H_0 : \theta = \theta_0$ to mean "the null hypothesis is that the population parameter θ equals θ_0" or $H_0 : \theta \in \Theta_0$ to mean "the null hypothesis is that the population parameter θ is in the set Θ_0".

2. **Alternative hypothesis** H_a. This is the hypothesis to be accepted in case H_0 is rejected, a value θ_a or a set of values Θ_a that represent what we are trying to prove. We write $H_a : \theta = \theta_a$ to mean "the alternative hypothesis is that the population parameter θ equals θ_a" or $H_a : \theta \in \Theta_a$ to mean "the alternative hypothesis is that the population parameter θ is in the set Θ_a".

3. **Test statistic** T. This is a function of the data $T = T(X_1, \ldots, X_n)$ on which the statistical decision will be based to reject the null hypothesis or not.

4. **Rejection region** RR. This is a set of values RR that typically depends on the sample size n where, if the test statistic T falls in RR, then there is enough evidence for us to **reject the null hypothesis** in favor of the alternative hypothesis. If T does not fall in RR, then we **fail to reject the null hypothesis**.

In a hypothesis testing problem, the functioning parts are the test statistic T and an associated rejection region RR. Usually the null and alternative hypotheses are obvious and our task is to identify the functioning parts, T and RR. That is, we need a decision rule to guide us to reject the null hypothesis based on the observed data.

Example 1.3. Library or Student Union

You are suspicious that the coin used is biased. Let $\theta = p$ denote the probability that the coin shows heads. You want to show that $p \neq 0.5$, so the null hypothesis is that $H_0 : p = 0.5$, here θ_0 is 0.5 and it represents that the coin is fair.

Scenario 1: If you only care about the fairness of the coin, then the alternative hypothesis should be that $H_a : p \neq 0.5$. This is called a **two-sided test**.

Scenario 2: If you really suspect that the coin is biased towards coming up heads (in your friend's favor), then the alternative hypothesis should be $H_a : p > 0.5$. This is called a **one-sided test**.

Suppose you flip the coin 20 times and observe $X_1, \ldots, X_{20}$, where $X_i = 1$ if the coin comes up heads. The question then is: given the results of 20 flips, how would you determine whether the coin is biased or not? The task is to come up with a test statistic T based on these flips and a set RR such that you will reject the null hypothesis and conclude the coin is biased if $T \in RR$. Otherwise, you will fail to reject the null hypothesis, and remain with the original presumption that the coin was fair.

Scenario 1: You can use $T = \sum_{i=1}^{20} X_i$ (representing the number of heads) as a test statistic. It seems reasonable that if you observe too few or too many heads, you tend to believe $p \neq 0.5$ rather than $p = 0.5$, so that H_a holds and you should reject H_0. In other words, if $p = 0.5$, then T is unlikely to be very small OR very large. Thus it makes sense to set $RR = \{T \leq 5 \text{ or } T \geq 15\}$ as the rejection region. You will reject the null if there are at most 5 heads or at least 15 heads and conclude that the coin is biased.

Scenario 2: You can again use the same statistic $T = \sum_{i=1}^{20} X_i$ but $RR = \{T \geq 15\}$ this time. You will only reject the null if there are at least 15 heads and conclude that the coin is biased. In this scenario, you will only reject the null if there is evidence that the coin is biased *in your friend's favor* but could not show that it is biased in your favor. □

There are of course infinitely many statistics T and rejection regions RR that we could follow for when to reject the null hypothesis. In order to come up with reasonable rules, we need some criteria about what a "good" testing procedure would look like.

1.5.2 Type I and Type II errors, test level, and test power

We have two hypotheses: the null $H_0, \theta \in \Theta_0$, the population parameter θ lies in the set Θ_0 representing the values of our null hypothesis and the alternative $H_a, \theta \in \Theta_a$, the population parameter θ lies in the set Θ_a representing the values of our alternative hypothesis. We also have two actions: reject the null H_0 when our test statistic T falls in RR and fail to reject the null H_0 when our test statistic T does not fall in RR. For any fixed rejection region, two types of errors can be made in reaching a decision. We use a two-by-two table (Table 1) to present all possible scenarios. In the upper left, we reject the null H_0 in favor of H_a when H_0 is true, we call this a **Type I error**. In the lower right, we fail to reject the null H_0 when H_0 is false, we call this a **Type II error**. In the upper right, we reject the null H_0 when in reality H_0 is false. Hence, we *correctly* reject the null hypothesis, which is good. In the bottom left, we fail to reject the null H_0 when H_0 is the truth, and again this is a success.

We want to be in the green cells where we correctly reject or fail to reject. However, due to the randomness in the data, it is rarely possible to construct a testing procedure that avoids *any* errors.

	H_0 is true ($\theta \in \Theta_0$)	H_0 is false (H_a is true, $\theta \in \Theta_a$)
Reject the null H_0	Type I error	Success!
Fail to reject the null H_0	Success!	Type II error

Table 1: The four possible scenarios based on the reality of the world (columns) and our decision about the null hypothesis (rows).

Instead, we can try to control the **probability** of committing a type I or type II error. We define

$$\alpha = P(\text{type I error}) = P(\text{reject } H_0 \text{ when } \theta \in \Theta_0) = P(T \in RR \text{ when } \theta \in \Theta_0).$$

Does this notation α look familiar to you? Yes, we have seen it when we introduced CIs. We will see the connection a lot along the way. Here we call α the **level** of the test. The level of the test measures how likely we are to reject a null hypothesis that is actually true. We also define

$$\beta = P(\text{type II error}) = P(\text{fail to reject } H_0 \text{ when } \theta \in \Theta_a) = P(T \notin RR \text{ when } \theta \in \Theta_a). \quad (4)$$

We call $1 - \beta$ the **power** of the test, since it tells us how likely we are to discover an effect given that the effect is actually present. Note that both α and β typically depend on the true value of θ and the sample size n.

Example 1.4. Library or Student Union

We will calculate the probabilities of type I and type II errors this time. We are suspicious that the coin is biased. Let $\theta = p$ denote the probability that the coin shows heads and our null hypothesis is that $H_0 : p = 0.5$.

Scenario 1: $H_0 : p = 0.5$ versus $H_a : p \neq 0.5$. We used $T = \sum_{i=1}^{20} X_i$ as our test statistic and $RR = \{T \leq 5 \text{ or } T \geq 15\}$ as the rejection region. What are α and β for this test? We have learned how to calculate this probability if H_0 is true. Using the fact that $T \sim \text{Binomial}(n, p)$, we have

$$\begin{aligned} \alpha &= P(T \in RR \text{ when } p = 0.5) = P\left(T \leq 5 \text{ or } T \geq 15 \text{ when } p = 0.5\right) \\ &= \sum_{k=0}^{5} \binom{20}{k} 0.5^k (1 - 0.5)^{20-k} + \sum_{k=15}^{20} \binom{20}{k} 0.5^k (1 - 0.5)^{20-k} \approx 0.0414. \end{aligned}$$

This value means that if the true probability of a heads is 0.5, then we have a roughly 4% chance of incorrectly rejecting the null hypothesis.

For β, we have

$$\beta = P(T \notin RR \text{ when } p \neq 0.5) = P\left(6 \leq T \leq 14 \text{ when } p \neq 0.5\right) = \sum_{k=6}^{14} \binom{20}{k} p^k (1 - p)^{20-k}. \quad (5)$$

In this case, we need to first assign a value to p in order to calculate β. Note if it were the case that $p = 0.5$, then $\beta = 1 - \alpha \approx 0.9586$. When $p = 0.4$ or 0.6, $\beta \approx 0.8728$ and when $p = 0.3$ or 0.7, $\beta \approx 0.58359$. We have some high type II error probabilities here.

Scenario 2: $H_0 : p = 0.5$ versus $H_a : p > 0.5$. We used the same statistic $T = \sum_{i=1}^{20} X_i$ but $RR = \{T \geq 15\}$. What are α and β now? We have

$$\alpha = P(T \in RR \text{ when } p = 0.5) = P\left(T \geq 15 \text{ when } p = 0.5\right) = \sum_{k=15}^{20} \binom{20}{k} 0.5^k (1 - 0.5)^{20-k} \approx 0.0207.$$

For β, we have

$$\beta = P(T \notin RR \text{ when } p > 0.5) = P\left(T \leq 14 \text{ when } p > 0.5\right) = \sum_{k=0}^{14} \binom{20}{k} p^k (1-p)^{20-k}.$$

This time, $\beta \approx 0.8744$ when $p = 0.6$ and $\beta \approx 0.58363$ when $p = 0.7$. Notice these two β values are only slightly greater than those corresponding values in Scenario 1. Intuitively this is because when p is over 0.5, it is unlikely to obtain fewer than 6 heads in 20 flips.

□

We now have at least a general goal: we want to avoid type I and type II errors as much as possible! Typically the more we try to reduce α and therefore the frequency of type I errors, the higher β will be, and so the more frequently we will make type II errors, and vice-versa. Thus, α and β are inversely related. Think of it like this: If a test rarely rejects true nulls, so α is close to 0, but we may very likely fail to reject false nulls, which leads to a high β value; If a test almost always rejects the null when it is false, so β is close to 0, but we may be frequently rejecting true null hypotheses and get a large α value. Ultimately, we have to make a trade-off between type I and type II errors. The most common way of doing this in statistics and science is to set a threshold for α that we want our test to achieve, typically $\alpha = 0.1, 0.05, 0.025$, or $\alpha = 0.01$. We then try to make β as small as possible given this threshold. The reason we do this is because we consider controlling the probability that we make type I errors a higher priority than limiting type II errors. We would rather have a test that is guaranteed to rarely incorrectly reject the null hypothesis but may not always detect true alternatives than a test that is very likely to detect true alternatives but may incorrectly reject the null too often.

Now we can formally state our hypothesis testing goal: we want a test statistic T and rejection region RR such that $P(T \in RR \text{ when } \theta \in \Theta_0) \leq \alpha$, where α is prespecified based on a desired level of the test, and $P(T \notin RR \text{ when } \theta \in \Theta_a)$ is as small as possible.