Making statistical inferences means to learn about what you do not observe, which is called parameters, from what you do observe, which is called data. We learn the basic principles of statistical inference from a perspective of causal inference, which is a popular goal of political science research. Namely, we study statistics by learning how to make causal inferences with statistical methods.

1 Statistical Framework of Causal Inference

What do we exactly mean when we say “An event A causes another event B”? Whether explicitly or implicitly, this question is asked and answered all the time in political science research. The most commonly used statistical framework of causality is based on the notion of counterfactuals (see [Holland, 1986]). That is, we ask the question “What would have happened if an event A were absent (or existent)”? The following example illustrates the fact that some causal questions are more difficult to answer than others.

Example 1 (Counterfactual and Causality) Interpret each of the following statements as a causal statement.

1. A politician voted for the education bill because she is a democrat.
2. A politician voted for the education bill because she is liberal.
3. A politician voted for the education bill because she is a woman.

In this framework, therefore, the fundamental problem of causal inference is that the counterfactual outcomes cannot be observed, and yet any causal inference requires both factual and counterfactual outcomes. This idea is formalized below using the potential outcomes notation.

Definition 1 (Potential Outcomes) Let $T_i$ be the causal (or treatment) variable of interest for unit $i$ where $i = 1, 2, \ldots, n$. $T_i$ is a random variable which takes a value in a set $T$. The potential outcome $Y_i(t)$ represents the outcome that would be observed for unit $i$ if it receives the treatment whose value is $t$, i.e., $T_i = t$ for $t \in T$.

We use $Y_i$ to denote the observed outcome for unit $i$. The treatment variable determines which of potential outcomes will be revealed. This can be seen, for example, from the fact that if the treatment is binary, the observed outcome is given by, $Y_i = T_iY_i(1) + (1 - T_i)Y_i(0)$. 
The framework described above makes an implicit but important assumption that the treatment status of one unit does not affect the potential outcomes of another unit. This can be formalized as follows,

**Assumption 1 (No Interference Between Units)** Formally, let \( T \) and \( \tilde{T} \) be an \( n \) dimensional vector of treatment assignment, whose \( i \)th element represents the treatment value of unit \( i \) where \( i = 1, 2, \ldots, n \). Let \( Y_i(T) \) be the potential outcome of unit \( i \) given the treatment assignment for all units, i.e., \( T \). Then, the assumption implies that \( Y_i(T) = Y_i(\tilde{T}) \) whenever \( T_i = \tilde{T}_i \).

This assumption is sometimes called the Stable Unit Treatment Value (SUTVA) assumption. Now, consider the following examples.

**Example 2** Is the assumption of no interference between units violated in the following examples?

1. (Flu vaccine) Units: individuals, Treatment: flu shot, Potential outcomes: hospitalization with and without the shot.

2. (Negative campaign in elections) Units: candidates, Treatment: Use of negative ads, Potential outcomes: Vote shares with and without the negative ads.

Of course, the potential outcomes framework described above is “a” model of causal inference, which has proven to be effective in many applied settings. The philosophical discussion of what is causality is an interesting topic, but is beyond this course.

Now, any causal quantities of interest for each unit can be written as a function of these potential outcomes. For the notational simplicity, we consider the situation where the treatment variable \( T_i \) is binary (i.e., \( T = \{0, 1\} \)), which implies that there are two potential outcomes for each unit, i.e., \( Y_i(1) \) and \( Y_i(0) \). However, the arguments presented here can be extended directly to the causal inference with a multi-valued treatment variable. We first give the definition of some frequently used causal effects for each unit.

**Definition 2 (Unit Causal Effects)** Let \( T_i \) be a binary treatment variable. The following causal effects can be defined for each unit.

1. Difference: \( Y_i(1) - Y_i(0) \)

2. Ratio: \( Y_i(1)/Y_i(0) \)

3. Percentage Increase: \( 100 \cdot \frac{[Y_i(1) - Y_i(0)]}{Y_i(0)} \)

Because the potential outcomes, \( Y_i(1) \) and \( Y_i(0) \), are never jointly observed, the joint distribution of the potential outcomes, \( P(Y_i(1), Y_i(0)) \), cannot be directly inferred from the data. This implies that the distribution of unit causal effects, e.g., \( P(Y_i(1) - Y_i(0)) \), also cannot be estimated directly from the data without additional assumptions. However, as we see later, the average causal effects can be identified in some situations and are often quantities of interest (e.g., [Imbens] 2004).

We first consider the sample average differences of potential outcomes.

**Definition 3 (Sample Average Causal Effects)** Let \( T_i \) be a binary (random) treatment variable for unit \( i \) where \( i = 1, 2, \ldots, n \). Consider fixed (i.e., non-random) but possibly unknown potential outcomes, \( Y_i(1) \) and \( Y_i(0) \), for each \( i \). Then, the following sample average causal effects of interest can be defined.

1. (Sample Average Causal Effect) \( \frac{1}{n} \sum_{i=1}^{n} Y_i(1) - Y_i(0) \).
2. (Sample Average Causal Effect for the Treated) \( \frac{1}{\sum_{i=1}^{T_i}} \sum_{i=1}^{n} T_i (Y_i(1) - Y_i(0)) \).

The distinction between the sample and population causal quantities is important.

**Definition 4 (Population Average Causal Effects)** Let \( T_i \) be a binary (random) treatment variable for unit \( i = 1, 2, \ldots, n \). Let \((Y_i(0), Y_i(1))\) be a simple random sample of potential outcomes from a population. Then, the following population average causal effects of interest can be defined.

1. (Population Average Causal Effect) \( E[Y(1) - Y(0)] \).
2. (Population Average Causal Effect for the Treated) \( E[Y(1) - Y(0) \mid T = 1] \).

The subscript \( i \) can be dropped because it is a simple random sample. We can also define the conditional average causal effects given the observed characteristics of each unit in the sample.

**Definition 5 (Conditional Average Causal Effects)** Let \( T_i \) be a binary (random) treatment variable for unit \( i = 1, 2, \ldots, n \). Let \((Y_i(0), Y_i(1), X_i)\) be a simple random sample from a population where \( Y_i(0) \) and \( Y_i(1) \) denote potential outcomes and \( X_i \) is a vector of characteristics for each unit \( i \). Then, the following conditional average causal effects of interest can be defined.

1. (Conditional Average Causal Effect) \( \frac{1}{n} \sum_{i=1}^{n} E[Y_i(1) - Y_i(0) \mid X_i] \).
2. (Conditional Average Causal Effect for the Treated) \( \frac{1}{\sum_{i=1}^{T_i}} \sum_{i=1}^{T_i} T_i E[Y_i(1) - Y_i(0) \mid X_i] \).

Here, the subscript \( i \) is retained because the conditional expectation is taken with respect to a particular unit \( i \) whose characteristics are represented by \( X_i \).
2 Statistical Analysis of Classical Randomized Experiments

In this section, we first consider statistical analysis of classical randomized experiments as a way to motivate the general theory of statistical inference.

2.1 Fisher’s Hypothesis Testing

Ronald A. Fisher was the first to come up with the idea that randomized experiments can be used to test a scientific hypothesis. Before him, scientists were using controlled experiments where they tried to minimize the differences between the treatment and control groups (except the fact that the former receives the treatment and the latter does not), as much as possible. However, controlled experiments had two problems. First, researchers can never make the conditions completely identical for each group. Second and more importantly, when these differences are not eliminated, there is no way for researchers to quantify the error that result caused by those uncontrolled differences.

To overcome these problems, Fisher (1935) proposed the use of randomized experiments and illustrated its use with the following famous example,

Example 3 (Lady Tasting Tea) In one summer afternoon in 1919, Cambridge, England, a group of university dons, their wives, and some guests were sitting around an outdoor table for afternoon tea. A lady declared, “Tea tastes different depending on whether the tea was poured into the milk or whether the milk was poured into the tea.” How should one test this proposition using a randomized experiment?

This simple example can be generalized into the method called, randomization (or permutation) inference. In the potential outcomes framework described earlier, the randomization of treatment guarantees the independence between the treatment and potential outcomes.

Definition 6 (Randomization of Treatment) The treatment is said to be randomized if the treatment variable $T_i$ is independent of all potential outcomes, $Y_i(t)$, for all units, i.e., $Y_i(t) \perp T_i$ for all $t$ and all $i$.

Note that there are many ways of randomizing the treatment. For example, simple random assignment assigns the treatment to each unit independently with equal probability, while completely random assignment (what Fisher did) randomly selects the predetermined number of units which receive the treatment. Other designs include matched pair design, randomized blocks, and Latin square.

We now formalize Fisher’s randomization inference as follows,

Definition 7 (Randomization Inference) Let $T_i$ be a binary treatment variable for unit $i$ where $i = 1, 2, \ldots, n$, and $T$ be an $n$ dimensional vector whose $i$th element is $T_i$. Then, $P(T = t)$ defines the randomized treatment assignment mechanism. Suppose that $Y(t)$ represents an $n$ dimensional vector of fixed (but possibly unknown) potential outcomes when $T = t$, and $T_{obs}$ represents the observed treatment assignment. Then, randomization inference is defined as follows,

1. (Sharp Null Hypothesis) $H_0 : Y_i(1) - Y_i(0) = \tau_0$ for all $i$ with some fixed $\tau_0$.

2. (Test Statistic) $S(Y, T)$.

3. (p-value) $P(S(Y, T_{obs}) \leq S(Y, T))$ computed under the sharp null hypothesis.

A smaller value of the p-value indicates more (statistically) significant evidence against $H_0$. 
Randomization inference described here is inference about the sample rather than the population. The only source of randomness, therefore, comes from the randomized assignment of the treatment, and this creates the **reference distribution** of the test statistic under the null hypothesis. Randomization inference is **distribution-free** because it does not make a distributional assumption about potential outcomes. It is also **exact** because it does not make any approximation. Moreover, randomization inference respects the randomization procedure that was actually conducted in the experiment, as the following application of randomization inference shows,

**Example 4 (California Alphabet Lottery)** Since 1975, California law has mandated that the Secretary of State draw a random alphabet for each election to determine the order of candidates for the first assembly district. The law further requires the candidate order to be systematically rotated throughout the remaining assembly districts. Ho and Imai (2006) applies randomization inference to the California alphabet lottery to investigate the ballot order effects in elections.

Alternatively, one can interpret randomization inference as inference about the infinite population from which a simple random sample is obtained. In this case, we assume that potential outcomes, \((Y_i(1), Y_i(0))\), are sampled at random from a population, which is characterized by marginal distributions \(P(Y(1))\) and \(P(Y(0))\), respectively. Now, the potential outcomes are random variables. Then, the null hypothesis of no treatment effect is given by,

\[
H_0 : P(Y(1)) = P(Y(0)).
\]

Under this null hypothesis, if each unit \(i\) gets assigned at random to the treatment or control group, then the distribution of the observed outcome vector \(Y\) will be still the same for any treatment assignment pattern. Therefore, each value of any test statistic, which corresponds to each treatment assignment pattern, is equally likely. This argument suggests that the randomization inference as described in Definition 7 can be viewed as inference about a population.

Finally, the scientific significance cannot be judged from the \(p\)-value, which can only tell us statistical significance. To do this, we need to go beyond hypothesis testing.

### 2.2 Neyman’s Analysis of Randomized Experiments

Randomization inference described above is concerned about unit causal effects as defined in Definition 2. The null hypothesis says that the causal effect is zero for every unit, and the randomization confidence interval is derived under the assumption of constant treatment effect. In contrast, Neyman (1923) considered inference about the **sample average causal effect** as defined in Definition 3. Neyman showed that the difference-in-means estimator is unbiased for the sample average causal effect and derived the expression for its variance.

**Theorem 1 (Estimation of Sample Average Causal Effect)** Consider a completely randomized experiment where \(2n\) units are randomly selected into the treatment and control groups of equal size. Let \(T_i\) be the binary treatment variable and \(Y_i\) is the outcome. Consider the following estimator of the sample average causal effect \(\tau\),

\[
\hat{\tau} \equiv \frac{1}{n} \sum_{i=1}^{2n} [T_iY_i - (1 - T_i)Y_i].
\]

Then,

\[
E(\hat{\tau}) = \tau, \quad \text{and} \quad \text{var}(\hat{\tau}) = \frac{S_1^2}{2n} + \frac{S_0^2}{2n} + \frac{S_{01}}{n},
\]
where $S^2_1$ and $S^2_0$ are the (sample) variance of the potential outcomes $Y_i(1)$ and $Y_i(0)$, respectively, and $S_{01}$ is their sample covariance.

Under randomization, the sample variances of $Y_i(1)$ and $Y_i(0)$ can be estimated without bias using the sample variances of the observed outcomes for the treatment and control groups. However, the sample covariance between the two potential outcomes cannot be estimated directly because we never observe them jointly. Neyman (1923) further demonstrated that the standard estimator of the variance of the average treatment effect is too conservative (i.e., too large).

**Theorem 2 (Bounds for Variance of Sample Average Causal Effect Estimator)** If $\hat{\tau}$ represents the estimator of the average treatment effect defined in Theorem 1, then its variance satisfy the following inequality,

$$\text{var}(\hat{\tau}) \leq \frac{S^2_1}{n} + \frac{S^2_0}{n},$$

where the upper bound is obtained under the constant treatment effect assumption.

So far we have focused on the estimation of sample causal quantities. Alternatively, we can also consider the estimation of the population average causal effect as defined in Definition 4 by thinking that the sample at hand comes from a population. It turns out that in this situation the variance can be identified from the data.

**Theorem 3 (Estimation of Population Average Causal Effect)** Consider the same experiment and estimator, $\hat{\tau}$, as defined in Theorem 1 except that the potential outcomes $(Y_i(1), Y_i(0))$ are a simple random sample from the population with marginal means $(\mu_1, \mu_0)$ and marginal variances $(\sigma^2_1, \sigma^2_0)$. Consider the population average causal effect as the estimator, i.e., $\tau = \mu_1 - \mu_0$. Then,

$$E(\hat{\tau}) = \tau, \quad \text{and} \quad \text{var}(\hat{\tau}) = \frac{\sigma^2_1}{n} + \frac{\sigma^2_0}{n}.$$

Therefore, we can estimate the variance of $\hat{\tau}$ directly from the data without bias using the sample variance of the observed outcomes for the treatment and control groups. Note that the variance of the population estimator is greater than the variance of the sample estimator. This makes sense because the former has an extra variability induced by random sampling from a population.

All the properties derived above are finite sample properties because they hold regardless of one’s sample size. We are also interested in asymptotic (large sample) properties of a given estimator. “How does a particular estimator behave as the sample size goes to infinity?” A “good” estimator should converge to the true value of its estimand. We would also want to derive the asymptotic distribution of the estimator so that an approximate variance of the estimator can be obtained.

**Theorem 4 (Asymptotic Properties of the Difference-in-Means Estimator)** Consider the same setting as in Theorem 3 where we denote the difference-in-means estimator as $\hat{\tau}_n$. Then,

1. (Consistency) $\hat{\tau}_n \overset{p}{\rightarrow} \tau$.

2. (Asymptotic Normality) $\sqrt{n}(\hat{\tau}_n - \tau) \overset{d}{\rightarrow} N(0, \sigma^2_1 + \sigma^2_2)$.

Given the intuitions we developed through two particular examples, we now turn to the general discussion of point estimation, hypothesis testing, and interval estimation.
3 Point Estimation

Building on the intuition we developed from Neyman’s approach, we study general principles of the estimation of scientific quantities of interest. The estimation of any quantity involves uncertainty, which needs to be quantified in every statistical estimation. One common way to quantify the uncertainty of one’s estimate is to estimate its variance.

3.1 Nonparametric Estimation

The difference-in-means estimator is a simple and special case of so called nonparametric plug-in estimators. The key idea of nonparametric estimation is to avoid as many assumptions as possible. Neyman’s estimator is such an example because it does not make any assumption about the distribution of potential outcomes. A simple but important nonparametric estimator is the empirical CDF, which is discrete and puts $1/n$ probability mass at each realization of $X_i$. The key properties of the empirical CDF are given now,

**Theorem 5 (Properties of Empirical CDF)** Let $X_i$ with $i = 1, 2, \ldots, n$ be a simple random sample from a population, which is characterized by the distribution function $F$, and let $\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{\{X_i \leq x\}}$ be the empirical CDF. Then, for any fixed $x$,

1. $E(\hat{F}_n(x)) = F(x)$.
2. $\text{var}(\hat{F}_n(x)) = \frac{F(x)(1-F(x))}{n}$.
3. $\hat{F}_n(x) \xrightarrow{p} F(x)$.

Using the empirical CDF, we can come up with a class of nonparametric estimators, called nonparametric plug-in estimator,

**Definition 8 (Nonparametric Plug-in Estimator)** Let $X_i$ with $i = 1, 2, \ldots, n$ be a simple random sample from a population, which is characterized by the distribution function $F$. If we define a statistical functional $\theta = S(F)$, then the nonparametric plug-in estimator of $\theta$ is given by,

$$\hat{\theta}_n = S(\hat{F}_n),$$

where $\hat{F}_n(x)$ is the empirical distribution function.

In this section, we focus on a special type of statistical functionals.

**Definition 9 (Linear Statistical Functional)** Let $F$ be an unknown distribution function, which characterizes the data generating process. Then, a statistical functional of the form,

$$\theta = \int g(x) dF(x) = \left\{ \begin{array}{ll} \int g(x)f(x) \, dx & \text{if } X \text{ is continuous} \\ \sum_x g(x)f(x) & \text{if } X \text{ is discrete} \end{array} \right.,$$

is called a linear functional where $f$ is the probability density or mass function corresponding to $F$.

Some examples will help understand the above definitions.

**Example 5** Write mean, variance, and correlation as linear statistical functionals and then derive their nonparametric plug-in estimators. Do they equal sample counterparts?
Given a linear functional, \( S(aF + bG) = aS(F) + bS(G) \) holds for any distribution functions \( F, G \) and constants \( a, b \) (hence, it’s name). The above definition also implies the nonparametric plug-in estimator for a linear functional is in general given by

\[
S(\hat{F}_n) = \int g(x) \, d\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} g(x_i).
\]

Under certain regularity conditions, it can be shown that a nonparametric plug-in estimator of a linear statistical functional converges in probability to the true statistical functional. Furthermore, one can also show the asymptotic normality of this plug-in estimator.

**Theorem 6 (Asymptotic Properties of Plug-in Estimator for a Linear Functional)**  Let \( X_i \) with \( i = 1, 2, \ldots, n \) be a simple random sample from a population, which is characterized by the distribution function \( F(\cdot; \theta) \) with unknown parameter \( \theta \) whose parameter space \( \Theta \) is finite-dimensional. This means that in parametric estimation, we only need to estimate \( \theta \) and there is no need to estimate the distribution function itself, \( F \) (because \( \theta \) completely characterizes \( F \)). One can imagine that this will simplify the estimation problem significantly in many problems.

A formal way to consider the distinction between parametric and nonparametric estimation is to say that in the former, we consider the data generating process to be characterized by the distribution function \( F(\cdot; \theta) \) with unknown parameter \( \theta \) whose parameter space \( \Theta \) is finite-dimensional. This means that in parametric estimation, we only need to estimate \( \theta \) and there is no need to estimate the distribution function itself, \( F \) (because \( \theta \) completely characterizes \( F \)). One can imagine that this will simplify the estimation problem significantly in many problems.

Here, we consider two general ways of conducting a parametric analysis using randomized experiments as an example. The first is called the method of moments. This method is often suboptimal, but its advantage is the ease of computation.

**Definition 10 (Method of Moments)**  Let \( X_i \) with \( i = 1, 2, \ldots, n \) be a simple random sample from a population whose distribution function is \( F(\cdot; \theta) \). The method of moments estimator, \( \hat{\theta}_n \), of the \( J \)-dimensional parameter vector \( \theta \), is defined to be the value of \( \theta \) which is a solution to the following system of \( J \) equations,

\[
\frac{1}{n} \sum_{i=1}^{n} X_i^j = \int x^j \, dF(x; \theta).
\]

for \( j = 1, 2, \ldots, J \).
Note that \( E(X_i^j) = \int x^j \, dF(x; \theta) \). Let’s consider some examples.

**Example 7** In the randomized classical experiment described in Theorem 1, construct the method of moments estimator of the population average causal effect based on the following assumptions about the marginal distributions of the potential outcomes.

1. \( Y_i(1) \sim N(\mu_1, \sigma_i^2) \) and \( Y_i(0) \sim N(\mu_2, \sigma_i^2) \).

2. \( Y_i(1) \sim \text{Binom}(k_1, \pi_1) \) and \( Y_i(0) \sim \text{Binom}(k_2, \pi_2) \).

In many cases, the method of moments estimator is poor and can be improved upon. But, the method of moments estimator is consistent and asymptotically normal.

**Theorem 7 (Asymptotic Properties of Method of Moments Estimator)** Let \( \hat{\theta}_n \) be the method of moments estimator as defined in Definition 10. Let \( m^{(j)}(\theta) = \int x^j \, dF(x; \theta) \) represent the \( j \)th moment of a random variable \( X_i \) for \( j = 1, 2, \ldots, J \). Then, \( \hat{\theta}_n \) is a solution to the system of \( J \) equations, \( m(\theta) = \hat{m}_n \), where \( \hat{m}_n^{(j)} \) is the sample \( j \)th moment.

Assume that \( \theta \in \Theta \) and \( \Theta \subset \mathbb{R}^J \) is an open set. Further suppose that \( m : \Theta \to \mathbb{R}^J \) has non-zero Jacobian at \( \theta_0 \) and is continuous and differentiable at \( \theta_0 \) where \( \theta_0 \) is the true value of \( \theta \). Then,

1. (Existence) \( \hat{\theta}_n \) exists with probability tending to one.

2. (Consistency) \( \hat{\theta}_n \xrightarrow{p} \theta_0 \)

3. (Asymptotic Normality)

\[
\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, g(\theta_0)\text{var}(Y_i)g(\theta_0)^\top)
\]

where \( Y_i = (X_i, X_i^2, \ldots, X_i^J)^\top \) and \( g(\theta) = (g_1(\theta), g_2(\theta), \ldots, g_J(\theta))^\top \) with \( g_j(\theta) = \frac{\partial}{\partial \theta} m_j^{-1}(\theta) \).

It is immediate that we can consistently estimate the asymptotic variance-covariance matrix by

\[
\tilde{v}_n = g(\hat{\theta}_n) \left[ \frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})(Y_i - \bar{Y})^\top \right] g(\hat{\theta}_n)^\top,
\]

where \( \bar{Y} = (\sum_{i=1}^n X_i/n, \sum_{i=1}^n X_i^2/n, \ldots, \sum_{i=1}^n X_i^J/n)^\top \). Then, it follows that in a sufficiently large sample, the method of moments has the following approximate sampling distribution,

\[
\hat{\theta}_n \xrightarrow{\text{approx.}} N\left(\theta_0, \frac{\tilde{v}_n}{n}\right).
\]

Finally, this implies that each parameter, i.e., each element of \( \hat{\theta}_n \), also has the approximately normal sampling distribution in a sufficiently large \( n \). Formally,

\[
\hat{\theta}_n^{(i)} \xrightarrow{\text{approx.}} N\left(\theta_0^{(i)}, \frac{\tilde{v}_n^{(i,i)}}{n}\right),
\]

where the superscript denotes \( i \)th element of a vector or \( (i, i) \) element of a matrix. Note that \( \sqrt{\tilde{v}_n^{(i,i)}/n} \) is called the estimated asymptotic standard error of \( \theta_n^{(i)} \). It is important to note that this derivation of asymptotic standard error can be applied to any \( \sqrt{n} \)-consistent estimators. Let’s derive the estimator for the asymptotic variance of the method of moments estimators.
**Example 8** Derive the asymptotic variance of the method of moments estimators from Example 7 and consistent estimators of the resulting variances.

Another commonly used estimator maximizes the “likelihood” of observing the data you actually observed.

**Definition 11 (Maximum Likelihood Estimator)** Let \( X_i \) with \( i = 1, 2, \ldots, n \) be a random sample from the population with the probability density or mass function, \( f(x \mid \theta) \). Then, the likelihood function is defined as,

\[
L(\theta \mid X_n) = \prod_{i=1}^{n} f(X_i \mid \theta),
\]

where \( X_n = (X_1, X_2, \ldots, X_n) \). The log-likelihood function is given by \( l(\theta \mid X_n) = \sum_{i=1}^{n} \log f(X_i \mid \theta) \). Finally, the maximum likelihood estimator is given by,

\[
\hat{\theta}_n = \sup_{\theta \in \Theta} l(\theta \mid X_n).
\]

**Example 9** Find the maximum likelihood estimator in Example 7(1).

The maximum likelihood estimators have many desirable large-sample properties.

**Theorem 8 (Asymptotic Properties of Maximum Likelihood Estimator)** Consider the maximum likelihood estimator defined in Definition 11 and suppose that \( \theta_0 \) is the true value of \( \theta \). Under certain regularity conditions,

1. (Consistency) \( \hat{\theta}_n \xrightarrow{p} \theta_0 \).

2. (Invariance) \( g(\hat{\theta}_n) \) is also the maximum likelihood estimator of \( g(\theta_0) \) for any function \( g \).

3. (Asymptotic Normality)

\[
\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, I(\theta_0)^{-1}),
\]

where \( I(\theta_0) \) is the expected Fisher information and is defined as \( E \left[ \frac{\partial}{\partial \theta} l(\theta \mid X_i) \big|_{\theta=\theta_0} \right] \left[ \frac{\partial}{\partial \theta} l(\theta \mid X_i) \big|_{\theta=\theta_0} \right]^\top \).

4. (Asymptotic Efficiency) Let \( \hat{\theta}_n \) be any estimator of \( \theta \). Then,

\[
\text{var}(\hat{\theta}_n) \geq - \left[ \frac{\partial}{\partial \theta} E(\hat{\theta}_n) \right] E \left[ \frac{\partial^2}{\partial \theta \partial \theta^\top} l(\theta \mid X_i) \big|_{\theta=\theta_0} \right] \left[ \frac{\partial}{\partial \theta} E(\hat{\theta}_n) \right]^\top
\]

Using the same argument as we have done for the method of moments estimator, we can show that for a sufficiently large sample, the maximum likelihood estimator has the following sampling distribution,

\[
\hat{\theta}_n \approx_{\text{approx.}} N \left( \theta_0, - \left\{ \frac{\partial^2}{\partial \theta \partial \theta^\top} l(\theta \mid X_n) \big|_{\theta=\hat{\theta}_n} \right\}^{-1} \right),
\]

where the variance is called observed Fisher information matrix and is equal to the minus inverse of Hessian matrix. To prove this, you need to also show the information matrix equality,

\[
E \left[ \left\{ \frac{\partial}{\partial \theta} l(\theta \mid X_i) \big|_{\theta=\theta_0} \right\} \left\{ \frac{\partial}{\partial \theta} l(\theta \mid X_i) \big|_{\theta=\theta_0} \right\}^\top \right] = -E \left[ \frac{\partial^2}{\partial \theta \partial \theta^\top} l(\theta \mid X_i) \big|_{\theta=\theta_0} \right].
\]

**Example 10** Derive the asymptotic variance of the maximum likelihood estimator in the previous example. Take the log transformation of the variance so that it is not bounded.
4 Interval Estimation

So far, we have used the variance as a measure of uncertainty about our estimates. However, we also have seen that in many cases we know the asymptotic sampling distribution of our estimators. Therefore, we can go a step further to produce another measure of uncertainty called a confidence set (region), which covers the true value of the parameter with some probability.

**Definition 12 (Confidence Sets)** Let \( X_n = (X_1, X_2, \ldots, X_n) \) represent the data of sample size \( n \) and \( \theta_0 \) be the true value of the parameter of interest, \( \theta \in \Theta \). The \((1 - \alpha)\) confidence set is a set \( C(X_n) \), which satisfies the following equality,

\[
\inf_{\theta_0 \in \Theta} P_{\theta_0}(\theta_0 \in C(X_n)) = 1 - \alpha,
\]

where \( 0 \leq \alpha \leq 1 \) and \( P_{\theta_0}(\theta_0 \in C(X_n)) \) is called the coverage probability.

If \( C(X_n) \) is an interval, then it is called the \((1 - \alpha)\) confidence interval. One may construct \((1 - \alpha)\) asymptotic confidence interval \( C(X_n) \) such that \( \inf_{\theta_0 \in \Theta} \lim_{n \to \infty} P_{\theta_0}(\theta_0 \in C(X_n)) = 1 - \alpha \). The interpretation of confidence sets requires a caution.

- The \((1 - \alpha)\) confidence set is the set which contains the true value of the parameter with the probability at least \((1 - \alpha)\).
- It is incorrect to say that the true value of the parameter lies in the \((1 - \alpha)\) confidence set obtained from a particular data set at least \((1 - \alpha) \times 100\) percent of time.

Note that what is random is the confidence set and not the parameter, which is unknown but fixed.

The mean and variance alone do not give the confidence intervals. But, we can use an asymptotic distribution of the estimator to come up with the confidence interval,

**Theorem 9 (Normal-based Asymptotic Confidence Interval)** Let \( X_n = (X_1, X_2, \ldots, X_n) \) represent the data of sample size \( n \) and \( \theta_0 \) be the true value of the parameter of interest, \( \theta \in \Theta \). Suppose that the asymptotic distribution of the estimator, \( \hat{\theta}_n \), of the parameter \( \theta \) is given by,

\[
\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, \sigma^2),
\]

where \( \theta_0 \) is the true value of \( \theta \). Then, the \((1 - \alpha)\) asymptotic confidence interval is given by

\[
C(X_n) = (\hat{\theta}_n - z_{\alpha/2}s.e., \hat{\theta}_n + z_{\alpha/2}s.e.),
\]

where \( z_{\alpha/2} = \Phi^{-1}(1 - \alpha/2) \) and \( \Phi(.) \) is the distribution function of the standard normal random variable. This confidence interval satisfies the following property,

\[
\lim_{n \to \infty} P(\theta_0 \in C(X_n)) = 1 - \alpha.
\]

Note that for \( \alpha = 0.05 \), \( z_{0.025} \approx 1.96 \). Applying this theorem, one can immediately derive the confidence interval of the nonparametric plug-in, the method of moments, and maximum likelihood estimators we studied earlier. Let’s apply it to Neyman’s estimator.

**Example 11 (Confidence Intervals for the Population Average Causal Effect)** Construct the \((1 - \alpha)\) asymptotic confidence intervals for the estimator of the population average causal effect.
5 Statistical Hypothesis Testing

5.1 General Concepts

Using Fisher’s randomization inference as a motivating example, we consider statistical hypothesis tests more generally. Fisher used his $p$-value as a measure of evidence against the null hypothesis. We can push his argument a bit further, and come up with a procedure which we use to reject or retain the proposed null hypothesis.

**Definition 13 (Hypothesis Test)** Let $\theta$ be a parameter of interest and $\Theta$ be its parameter space. Suppose that we wish to test the null hypothesis $H_0 : \theta \in \Theta_0$ against the alternative $H_1 : \theta \in \Theta \setminus \Theta_0$, using the test statistic $S(X_n)$ where $X_n = (X_1, X_2, \ldots, X_n)$ represents the data of sample size $n$. Then, the hypothesis test is defined by specifying the rejection region $R$ where if $S(X_n) \in R$, then we reject $H_0$, and if $S(X_n) \notin R$, then we retain $H_0$. Typically, the rejection region is defined as $R = (c, \infty)$ where $c$ is called a critical value.

A null hypothesis of the form $\theta = \theta_0$ is called a *simple null hypothesis*, whereas $\theta > \theta_0$, $\theta < \theta_0$, etc. are called a *composite null hypothesis*. While the choice of the null hypothesis should be based on one’s scientific research question of interest (Fisher used $\tau_0 = 0$), the choice of test statistic should be governed by its statistical properties. Fisher used $S(Y_n, T_n) = \sum_{i=1}^n T_i Y_i + (1 - T_i)(1 - Y_i)$ as his test statistic, but other choices are also possible.

We can further investigate the consequence of choosing a particular test statistic on the performance of the hypothesis testing procedure. In particular, when conducting a hypothesis test, one can make two types of mistakes, which are called *Type I error* and *Type II error*:

<table>
<thead>
<tr>
<th>$H_0$ is true</th>
<th>$H_0$ is false</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reject $H_0$</td>
<td>Type I error</td>
</tr>
<tr>
<td>Retain $H_0$</td>
<td>Correct</td>
</tr>
</tbody>
</table>

Of course, we would like our hypothesis testing procedure to minimize those two types of errors. But there is an inherent trade-off between the two errors. If you always reject $H_0$, then you eliminate Type II error but maximize the possibility of Type I error. A way to get around this is to minimize one type of error without increasing the other type. As in Fisher’s example, hypothesis tests are conducted typically by deriving the distribution of test statistics under the null hypothesis. Therefore, the probability of committing Type I error can be chosen by data analysts. Given this fixed probability of Type I error, we may try to minimize the probability of committing Type II error. To formalize this idea, we introduce the following concepts,

**Definition 14 (Power and Size of Hypothesis Tests)** Consider a hypothesis test defined by the null hypothesis, $H_0 : \theta \in \Theta_0$, the test statistic $S(X_n)$, and the rejection region $R$ where $X_n = (X_1, X_2, \ldots, X_n)$ is the data of sample size $n$. Then,

1. The *power function* of the test is $\beta(\theta) = P_\theta(S(X_n) \in R)$.

2. The *size* of the test is $\alpha = \sup_{\theta \in \Theta_0} \beta(\theta)$.

3. A *level $\alpha$ test* is a hypothesis test whose size is less than or equal to $\alpha$.

In words, the power of a test is the probability that one rejects the null, while the size of a test is the largest probability of rejecting $H_0$ when $H_0$ is true.

According to the logic described above, we wish to find a test which has the largest power under the alternative hypothesis among all size $\alpha$ test. If you can find such a test, which is often a
difficult thing to do, a test is called *most powerful*. If the test is most powerful at all values of $\alpha$, then it is called *uniformly most powerful*. We do not go into the details of how to find such tests, but the above discussion offers the following lesson about the general interpretation of hypothesis testing,

- A failure to reject the null hypothesis may arise from the lack of power of your hypothesis testing procedure rather than the fact that the null hypothesis is true.

Let’s make sure that we understand the concepts using the following simple example.

**Example 12 (One sample $t$-test)** Assume that $X_i \sim N(\mu, \sigma^2)$ for $i = 1, 2, \ldots, n$. In our causal inference example, we may assume that the marginal distribution of each of the two potential outcomes is Normal and test whether its mean is less than some specified value. Derive a level $\alpha$ test and its power function for each of the following cases,

1. to test $H_0: \mu = \mu_0$ against $H_1: \mu \neq \mu_0$.
2. to test $H_0: \mu \leq \mu_0$ against $H_1: \mu > \mu_0$.

Finally, we give the general definition and interpretation of $p$-value.

**Definition 15 ($p$-value)** Let $X_n = (X_1, X_2, \ldots, X_n)$ represent the data of sample size $n$. Consider a test statistic $S(X_n)$ and its observed value $S(X_n^{obs})$ given the null hypothesis $H_0: \theta \in \Theta_0$ where $\Theta_0$ is a subset of the parameter space $\Theta$ of $\theta$. Then, the $p$-value, $p(X_n^{obs})$, is equal to,

$$p(X_n^{obs}) = \sup_{\theta \in \Theta_0} P_\theta(S(X_n) \geq S(X_n^{obs})).$$

If the null hypothesis is simple, i.e., $\Theta_0 = \{\theta_0\}$, then the $p$-value equals $P_{\theta_0}(S(X) \geq S(X_n^{obs}))$. In general, one needs to be careful about the interpretation of the $p$-value.

- The $p$-value is the probability, computed under the null hypothesis, of observing a value of the test statistic at least as extreme as the value actually observed.
- The $p$-value is not the probability that the null hypothesis is true.
- A large $p$-value can occur either because the null hypothesis is true or because the null hypothesis is false but the test is not powerful.
- The statistical significance indicated by the $p$-value does not necessarily imply scientific significance.

The first statement corresponds exactly to Fisher’s formulation and this is what exactly the definition of the $p$-value says. We also realize that the $p$-value is a function of the data and so can be seen as a statistic. Then, we can derive the distribution of the $p$-value under the null hypothesis,

**Theorem 10 (Distribution of the $p$-value)** Consider a size $\alpha$ test, which is defined by the null hypothesis $H_0: \theta \in \Theta_0$, the rejection region, $R_\alpha = (c_\alpha, \infty)$, and the test statistic, $S(X_n)$, where $X_n = (X_1, X_2, \ldots, X_n)$ represents the data of sample size $n$. Then, the distribution of the $p$-value under the null is stochastically greater than or equal to Uniform$(0,1)$. That is,

$$P(p(X_n) \leq \alpha) \leq \alpha,$$

for $\alpha \in [0, 1]$. 

13
An important special case is where the null hypothesis is simple and the test statistic is continuous. In that case, the reference distribution of the \(p\)-value is Uniform(0,1). To understand the implication of this theorem, consider rejecting/retaining the null hypothesis based on the \(p\)-value. That is, we use the \(p\)-value itself as a test statistic and reject \(H_0\) when the \(p\)-value is less than or equal to \(\alpha\). Then, the probability of committing Type I error is also less than or equal to \(\alpha\). (Why?) This implies,

- the \(p\)-value is the smallest level at which we can reject the null hypothesis.

or more formally the \(p\)-value is equal to \(\inf\{\alpha : S(X_n) \in R_\alpha\}\).

5.2 Finding Tests

There are many ways of finding a test. Fisher’s randomization test is an example of nonparametric tests where no distributional assumption is made. It is also an example of exact tests because no approximation is made to derive the reference distribution of a test statistic. Other nonparametric tests include \(\chi^2\) tests of independence. Alternatively, one can also consider parametric tests where the reference distribution of test statistic is derived based on some parametric assumptions. For example, one sample \(t\) test we studied earlier is such an example. Let’s review it in the context of causal inference with randomized experiments, which is often called paired \(t\)-test.

**Example 13 (Paired \(t\)-test)** Let \(\tau\) be the population average causal effect in a randomized experiment with a matched-pair design where two units are paired on the basis of observed covariates and the randomization of the treatment is conducted within each pair. Suppose that we wish to test \(H_0 : \tau = \tau_0\) against \(H_1 : \tau \neq \tau_0\). Derive a level \(\alpha\) test assuming the differences of the observed outcomes between each pair of the treated and control units are normally distributed.

Another example of parametric tests is the two-sample \(t\)-test.

**Example 14 (Two-sample \(t\)-test)** Consider two independent random samples from different Normal distributions, i.e., \(X_i \sim N(\mu_X, \sigma_X^2)\) for \(i = 1, \ldots, n_X\) and \(Y_i \sim N(\mu_Y, \sigma_Y^2)\). Construct the level \(\alpha\) test for each of the following cases,

1. to test \(H_0 : \mu_X \leq \mu_Y\) against \(H_1 : \mu_X > \mu_Y\).
2. to test \(H_0 : \mu_X = \mu_Y\) against \(H_1 : \mu_X \neq \mu_Y\).

In the context of causal inference with randomized experiments, one may assume that independent random samples are drawn from the population of the treated group and that of the control group.

These tests may give you misleading results if the underlying distributional assumptions are violated. One can easily check from the observed data whether this normality assumption is reasonable (and there is a way to test this formally (e.g., Kolmogorov-Smirnov Test). Now, we may wonder if we can avoid this distributional assumption and conduct a hypothesis test by using the Neyman’s nonparametric estimator. This is easy to do because we know the asymptotic distribution of the estimator. In fact, the use of asymptotic sampling distribution is a very general way to construct a hypothesis test and works for both nonparametric and parametric tests.

**Definition 16 (Wald Test)** Consider a simple null hypothesis of the form, \(H_0 : \theta = \theta_0\), against the alternative, \(H_1 : \theta \neq \theta_0\). Assume that an estimator \(\hat{\theta}_n\) is asymptotically normal.

\[
\sqrt{n}(\hat{\theta}_n - \theta_0) \to N(0, v^2),
\]
where $v^2$ is the asymptotic variance. Then, the size $\alpha$ Wald test rejects $H_0$ if and only if

$$\left| \frac{\hat{\theta}_n - \theta_0}{\text{s.e.}} \right| > z_{\alpha/2},$$

where $z_{\alpha} = \Phi^{-1}(1 - \alpha)$ and $\Phi(\cdot)$ is the inverse of the standard normal distribution function and s.e. is the estimated standard error, i.e., $\sqrt{v/n}$.

We now apply the Wald test to causal inference with randomized experiments.

**Example 15 (Wald Test based on the Neyman’s Estimator)** In the context of classical randomized experiments, construct a $(1 - \alpha)$ level hypothesis test for the null hypothesis $H_0: \tau = \tau_0$ where $\tau$ is the population average causal effect.

Note that Wald tests rely on the asymptotic sampling distribution of an estimator. This means that Wald tests are valid level $\alpha$ test only asymptotically.

**Theorem 11 (Asymptotic Property of Wald Test)** Consider Wald test as defined in Definition 16. Then, under the null hypothesis,

$$\lim_{n \to \infty} P_{\theta_0} \left( \left| \frac{\hat{\theta}_n - \theta_0}{\text{s.e.}} \right| > z_{\alpha/2} \right) = 1 - \alpha,$$

where s.e. is the estimated asymptotic standard error of $\hat{\theta}_n$.

For likelihood inference, there is another way of constructing a hypothesis test by using the likelihood itself rather than using the asymptotic sampling distribution of the ML estimator,

**Definition 17 (Likelihood Ratio Test)** Consider a hypothesis test where the null hypothesis $H_0: \theta \in \Theta_0$ is tested against the alternative hypothesis $H_0: \theta \in (\Theta \setminus \Theta_0)$. The likelihood ratio statistic is given by,

$$\lambda(X_n) = 2 \log \sup_{\theta \in \Theta} L(\theta \mid X_n) \sup_{\theta \in \Theta_0} L(\theta \mid X_n) = 2[\ell(\hat{\theta}_n \mid X_n) - \ell(\tilde{\theta}_n \mid X_n)],$$

where $X_n$ is the data of sample size $n$, and $\hat{\theta}_n$ and $\tilde{\theta}_n$ represent the unrestricted and restricted maximum likelihood estimates, respectively.

One can derive the asymptotic distribution of the likelihood ratio test statistic,

**Theorem 12 (Asymptotic Distribution of Likelihood Ratio Test Statistic)** Consider the likelihood ratio test defined in Definition 17 where $\Theta_0 = \{ \theta: \theta^{(i)} = \theta_0^{(i)} \text{for some } i \}$. Then, under the null hypothesis $H_0: \theta \in \Theta_0$, we have

$$\lambda(X_n) \xrightarrow{d} \chi^2_{\nu},$$

where $\nu$ is the dimension of $\Theta$ minus the dimension of $\Theta_0$. The p-value is given by $P(Z \geq \lambda(X_n))$ where $Z$ is distributed as $\chi^2_{\nu}$.

Finally, there is a close relationship between hypothesis testing and confidence interval. Indeed, one way to obtain a $(1 - \alpha)$ confidence interval is to “invert” a $(1 - \alpha)$ level test.
Theorem 13 (Inverting the Test) Consider a level $\alpha$ hypothesis test defined by the null hypothesis $H_0 : \theta = \theta_0$ and the rejection region $R_{\theta_0}$. Define a set $C(X) = \{\theta_0 : X \notin R_{\theta_0}\}$ where $X$ represents the data. Then, $C(X)$ is a $(1 - \alpha)$ confidence set. The converse also holds.

The theorem implies that randomization-based confidence sets can be constructed by inverting randomization test.

Example 16 (Randomization-Based Confidence Sets) Assume the constant treatment effect, i.e., $Y_i(1) - Y_i(0) = \tau$ for all $i = 1, 2, \ldots, n$ where $\tau$ is a fixed but unknown parameter. How does one obtain the $(1 - \alpha)$ confidence set about $\tau$?
References


