Statistics

Ye Wang
University of North Carolina at Chapel Hill

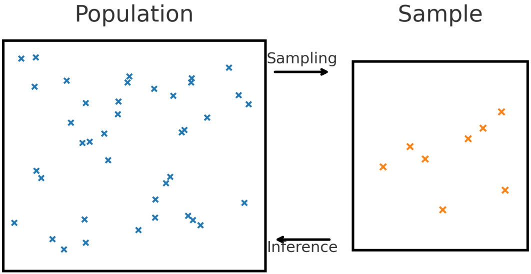
Mathematics and Statistics For Political Research
POLI783

From probability to statistics

- Probability starts from known distributions.
- If $X \sim \mathcal{N}(0,1)$, we know how to calculate $\mathbb{P}(X > 0.5)$ or its moments.
- In reality, we observe data rather than distributions.
- E.g., the height of 100 different individuals.
- How can we know whether the r.v. height obeys the normal distribution?
- How do we infer the probability $\mathbb{P}(X > 6)$ or the average height of Americans?

From probability to statistics

- This is the goal of statistics.
- Statistics starts from data and aims to learn the distribution underlying the data.
- We assume that the data are generated by sampling from a fixed distribution.



Basic ideas of statistics: estimand

- We refer to this distribution as the population distribution.
- It describes the randomness of a variable in a population.
- The population is defined by our research question.
- E.g., the ideology of voters in U.S.; a policy's effect on GDP growth across counties in China.
- The quantity of interest is often a functional of the distribution.
- As the functional is a real number, it is also called a "parameter" or the "estimand."
- We denote an estimand/parameter with $\tau = \tau(F)$.
- E.g., the average ideology of Hispanic female voters in America; the average effect of a policy on the poorest 10% of counties in China.
- The estimand can also be a real vector or a function.

Basic ideas of statistics: sampling

- The data contain N observations or units (sample size), each of which is a vector of r.v.s.
- E.g., ($Gender_i$, $Education_i$, $Income_i$, $Ethnicity_i$, $Ideology_i$).
- We use the subscript i to denote variables for the ith observation.
- Each observation is randomly and independently drawn from the population.
- The observations are thus i.i.d.
- The data compose a sample from the population (a joint distribution).
- The process to generate the sample from the population is known as sampling.
- The population distribution combined with the sampling process is known as the data-generating process (DGP).
- The reverse process to learn about the population from the sample is known as statistical inference.

Estimator and estimate

- How do we learn the estimand from the sample?
- We construct an estimator that maps the data to a number:

$$\hat{\tau} = \Psi\left(\mathbf{O}_1, \mathbf{O}_2, \ldots, \mathbf{O}_N\right),$$

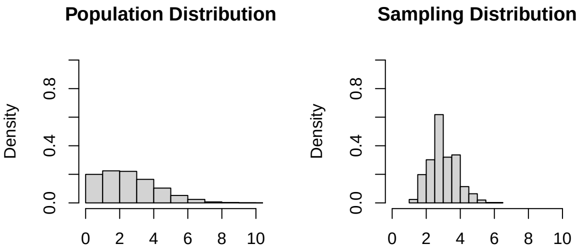
  where $\mathbf{O}_i$ refers to variables from observation i.
- The number generated by the estimator is called an estimate.
- Any arbitrary function of the sample/data is known as a statistic.
- A statistic becomes an estimator if we want to use it to infer an estimand.
- We say an estimand is identifiable if there exists an estimator $\hat{\tau}$ for it and $\hat{\tau} = \tau$ when $N = \infty$.

An example

- A common estimand is the expectation of some r.v.: $\tau = \mathbb{E}[Y_i]$.
- The data: $(Y_1, Y_2, \ldots, Y_N)$.
- There are multiple estimators of it.
- $\hat{\tau}_1 = \frac{1}{N}\sum_{i=1}^{N} Y_i$ (sample average).
- $\hat{\tau}_2 = \frac{1}{N/2}\sum_{i=1}^{N/2} Y_i$ (sample average across half of the observations).
- $\hat{\tau}_3 = Y_1$ (the first observation).
- $\hat{\tau}_3 = 3$ (constant).
- These are all estimators, but their performance differs by a lot.

Sampling distribution

- The estimator $\hat{\tau}$ is a function of N r.v.s thus a r.v. itself.
- The distribution of the estimator is known as the sampling distribution, which we denote as $F_{\hat{\tau}}(x)$.
- E.g., the average height of individuals in the sample depends on who are sampled.



Bias and variance

- As $\hat{\tau}$ is a r.v., we can define its expectation and variance.
- We define the bias of an estimator (relative to the estimand) as

$$Bias = \mathbb{E}[\hat{\tau}] - \tau.$$

- An estimator is unbiased if its bias equals zero.
- Which of the three estimators are unbiased?
- An estimator's variance is known as the sampling variance:

$$\sigma_{\hat{\tau}}^2 = \mathrm{Var}[\hat{\tau}].$$

- $\sigma_{\hat{\tau}}$ is known as the standard error of $\hat{\tau}$.
- Consider the sample average $\bar{X}$.
- It is unbiased with $\sigma_{\hat{\tau}} = \sigma/\sqrt{N} \to 0$ as $N \to \infty$.
- We refer to $N\sigma_{\hat{\tau}}^2$ as the normalized variance of $\hat{\tau}$.

Mean squared error

- We often use the mean squared error (MSE) to measure the prediction performance of the estimator.
- The MSE is defined as

$$
\begin{aligned}
&\mathbb{E}\left[(\hat{\tau}-\tau)^2\right] = \mathbb{E}\left[\hat{\tau}^2\right] - 2\mathbb{E}\left[\hat{\tau}\right]\tau + \tau^2 \\
=&\mathbb{E}\left[\hat{\tau}^2\right] - \left(\mathbb{E}\left[\hat{\tau}\right]\right)^2 + \left(\mathbb{E}\left[\hat{\tau}\right]\right)^2 - 2\mathbb{E}\left[\hat{\tau}\right]\tau + \tau^2 \\
=&\,\mathrm{Var}[\hat{\tau}] + (\mathbb{E}[\hat{\tau}] - \tau)^2 = \sigma_{\hat{\tau}}^2 + \mathit{Bias}^2.
\end{aligned}
$$

- It penalizes larger deviations of $\hat{\tau}$ from τ more severely.
- The MSE's magnitude is decided by both the bias and the variance.
- A common pheonomenon: an estimator with a larger bias often has a smaller variance.
- This is known as the bias-variance trade-off.

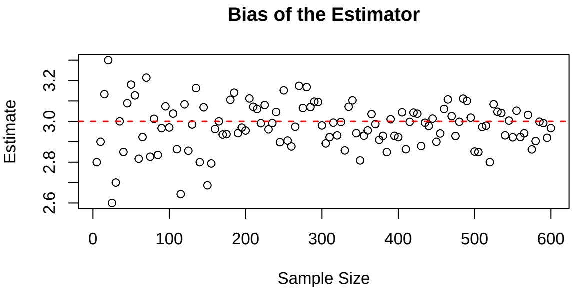
Consistency

- Unbiasedness does not guarantee that the difference between $\hat{\tau}$ and τ is small in a given sample.
- We hope that $|\hat{\tau} - \tau|$ shrinks as N grows.
- As $\hat{\tau}$ depends on N, we can define the sequence, $\{\hat{\tau}_N\}_{N=1}^{\infty}$.
- We say $\hat{\tau}_N$ is consistent for τ if

$$\mathbb{P}\left(|\hat{\tau}_N - \tau| > \varepsilon\right) \to 0$$

  for any $\varepsilon > 0$ as $N \to \infty$.
- Formally, it means $\hat{\tau}_N$ converges to τ in probability.
- We can denote it as $\hat{\tau} \xrightarrow{p} \tau$ as N grows or $\operatorname{plim}_{N \to \infty} \hat{\tau} = \tau$.

Consistency



Markov's equality

- We need some inequalities to show the consistency of any estimator.
- Markov's equality: for $X \geq 0$ and $a > 0$,

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$

- This is a concentration inequality: the probability for X to exceed a is bounded by its expectation divided by a.
- Larger values of X are less likely to occur.
- If the average income in this country is \$5k per month, what is the largest possible proportion of individuals whose monthly income is above \$10k?
- First note that $\mathbb{E}\left[X \mid X \geq a\right] \geq a$.
- Then, we can see that

$$\begin{aligned}\mathbb{E}[X] =& \mathbb{E}\left[X \mid X \geq a\right]\mathbb{P}(X \geq a) + \mathbb{E}\left[X \mid X < a\right]\mathbb{P}(X < a) \\ \geq& \mathbb{E}\left[X \mid X \geq a\right]\mathbb{P}(X \geq a) \geq a\mathbb{P}(X \geq a).\end{aligned}$$

Chebyshev's inequality

- From Markov's equality, we can derive Chebyshev's inequality.
- If $\mu = \mathbb{E}[X]$ and $\sigma^2 = \mathrm{Var}(X)$, then for any $t > 0$,

$$\mathbb{P}\big(|X - \mu| \geq t\big) \;\leq\; \frac{\sigma^2}{t^2}$$

- We apply Markov's inequality to $Y = (X - \mu)^2 \geq 0$,

$$\mathbb{P}\big(|X - \mu| \geq t\big) = \mathbb{P}\big(Y \geq t^2\big) \leq \frac{\mathbb{E}[Y]}{t^2} = \frac{\mathbb{E}[(X - \mu)^2]}{t^2} = \frac{\sigma^2}{t^2}.$$

- Consider again the income distribution in this country and assume that $\mu = \$5k$ and $\sigma = \$0.5k$.
- What is the largest possible proportion of individuals whose monthly income is at least $\$5k$ away from the average?

Consistency

- Let's assume that $\mathbb{E}\left[\hat{\tau}_N\right] = \tau$.
- Recall the definition of consistency and apply Chebyshev's inequality:
$$\mathbb{P}\left(|\hat{\tau}_N - \tau| > \varepsilon\right) \leq \frac{\sigma^2_{\hat{\tau}_N}}{\varepsilon^2} \to 0,$$
if $\sigma^2_{\hat{\tau}_N} \to 0$ as $N \to \infty$.
- Consistency holds for unbiased estimators if their variances shrink to zero as N grows.
- For the sample average, we know that $\sigma_{\hat{\tau}_N} = \sigma/\sqrt{N} \to 0$, thus it is consistent for μ.
- Biased estimators converge to $\mathbb{E}\left[\hat{\tau}_N\right] \neq \tau$ if their variances shrink to zero as N grows.
- Consistency requires at least asymptotic unbiasedness.
- Which of the three estimators are consistent?

Parametric estimators

- How do we construct estimators?
- Sometimes, the observations are drawn from a known distribution that can be represented by several unknown parameters: $F(\cdot) \in \mathcal{F} = \{F(\cdot;\theta)\}$.
- E.g., each $Y_i \sim \mathcal{N}\left(\mu, \sigma^2\right)$ with μ and σ^2 unknown to researchers.
- We first estimate θ from the data.
- Then, a natural estimator is $\hat{\tau} = \tau\left(F(\cdot;\hat{\theta})\right)$ and known as a parametric estimator.
- We can show that the estimator is consistent under general conditions.

Parametric estimators

- A common approach to estimate θ is maximum likelihood estimation (MLE) proposed by Ronald Fisher.
- Intuition: what are the parameters that can maximize the probability/likelihood for us to see the collected data?
- E.g., we have results from N independent coin flips: $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_N)$.
- $Y_i = 1$ if heads and $Y_i = 0$ if tails.
- Each coin flip follows the same distribution $Bern(p)$.
- How do we estimate the parameter p?

Parametric estimators

- The likelihood to observe the data:

$$L = p^{\sum_{i=1}^{N} Y_i}(1-p)^{\sum_{i=1}^{N}(1-Y_i)}.$$

- We often work with the log-likelihood function:

$$\log L = \log p \sum_{i=1}^{N} Y_i + \log(1-p)\sum_{i=1}^{N}(1-Y_i).$$

- We find $\hat{p}$ to maximize $\log L$.
- The first-order condition is

$$\frac{\partial \log L}{\partial p}\Big|_{p=\hat{p}} = \frac{1}{\hat{p}}\sum_{i=1}^{N} Y_i - \frac{1}{1-\hat{p}}\sum_{i=1}^{N}(1-Y_i) = 0.$$

- Therefore, $\hat{p} = \frac{\sum_{i=1}^{N} Y_i}{N}$.

Parametric estimators

- Parametric models are more convincing with guidance from substantive theory.
- E.g., legislator i has an ideal point θ_i, bill m has an ideological position η_m, and the status quo is at 0.
- Legislator i votes Yea for bill m ($Y_{im} = 1$) if

$$-(\theta_i - \eta_m)^2 \geq -(\theta_i - 0)^2 + \varepsilon_{im},$$

  where ε_{im} follows the logistic distribution with a variance σ_m^2.
- We can prove that in this setting,

$$\mathbb{P}(Y_{im} = 1) = \frac{e^{(2\theta_i \eta_m - \eta_m^2)/\sigma_m}}{1 + e^{(2\theta_i \eta_m - \eta_m^2)/\sigma_m}}.$$

- This is known as the item response theory (IRT) model and serves as the foundation for DW-NOMINATE.
- The connection between utility maximization and IRT was established by Daniel McFadden.

Structural estimation

- One extreme of parametric estimation is known as “structural estimation” in econometrics.
- Economists develop complex models to describe the behavior of multiple agents.
- These models are built upon “micro foundations” such as utility maximization and profit maximization.
- They fit these models on economic data to estimate a large number of parameters.
- Some scholars believe that this approach helps us understand underlying mechanisms and make predictions.
- E.g., how would tariffs affect consumption, investment, and the stock market in the US?
- But how can we know whether the model provides us with the correct likelihood function?
- Why would production follow the Cobb-Douglas production function?

Structural estimation

- Structural estimation is still common in fields such as macro economics and industrial organization.
- In general, it has been less popular since the "credibility revolution."
- Less common in political science but do exist.
- The results can be sensitive to the selected functional form.
- It can be the only approach when there aren't enough data.

Non-parametric estimators

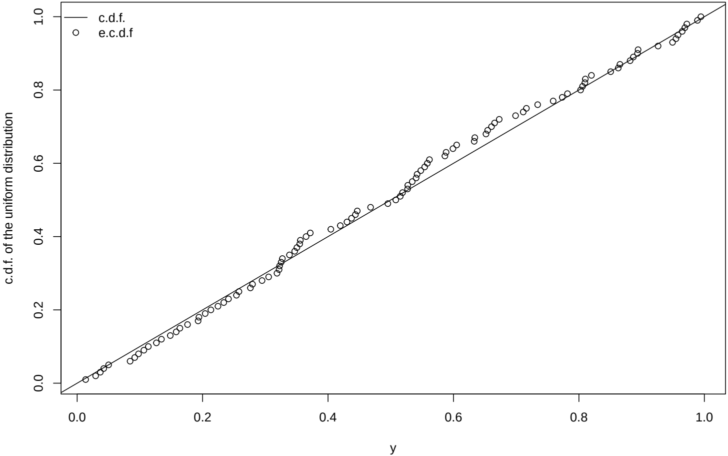
- In non-parametric estimation, we do not assume that $\mathcal{F}$ can be indexed by a finite-dimensional parameter.
- Some general restrictions are still necessary (e.g., smooth or integrable).
- One can understand $\mathcal{F}$ as indexed by an infinite-dimensional parameter (e.g., coefficients in the Taylor expansion).
- Two common options: we can use estimators that do not explicitly depend on $F(\cdot)$.
- E.g., if the observations are i.i.d., we can estimate the expectation μ with the sample average.
- The estimator is unbiased and consistent regardless the underlying distribution.
- Or, we use parametric models that can converge to $F(\cdot)$ when $N \to \infty$.
- E.g., we approximate any Taylor expansion with polynomials whose power grows with N.

Sample analogues

- If the estimand takes the form of $\tau = \mathbb{E}\left[g(\mathbf{O}_i)\right]$, a natural estimator is
$$\hat{\tau} = \frac{1}{N}\sum_{i=1}^{N} g(\mathbf{O}_i).$$
- The sample analogue or "plug-in" estimator of τ.
- It is often unbiased and consistent for i.i.d. data.
- E.g., how do we estimate the c.d.f. underlying the data?
- Remember that $F_X(x) = \mathbb{P}(X \leq x) = \mathbb{E}[\mathbf{1}\{X \leq x\}]$.
- Using its sample analogue, we have
$$\hat{F}_X(x) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{X_i \leq x\},$$
which is known as the "empirical cumulative distribution function" (e.c.d.f) of X.

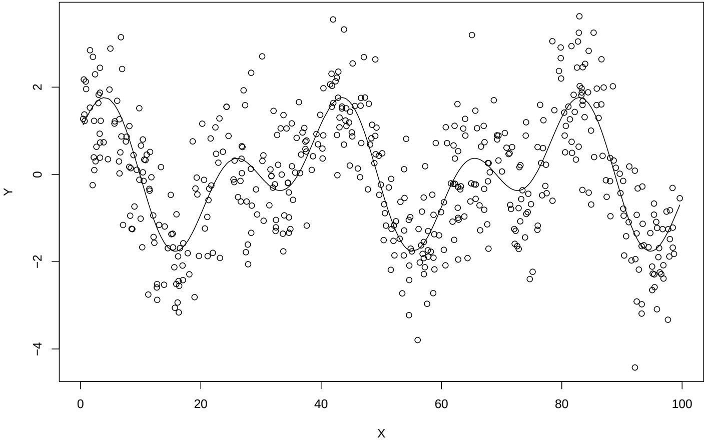
Sample analogues

- A sample analogue estimates $\tau(F)$ with $\tau\left(\hat{F}\right)$.



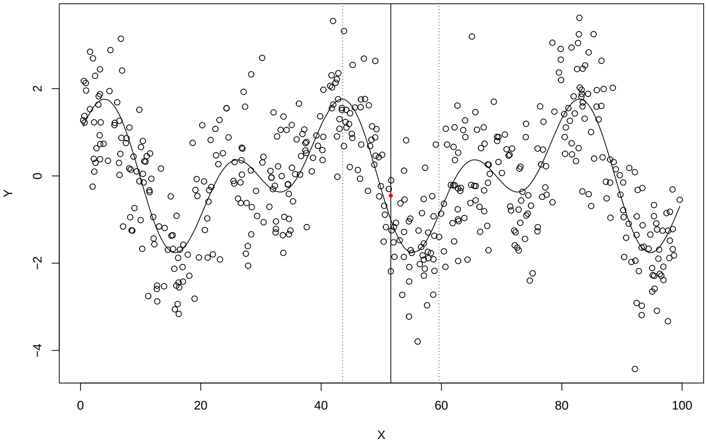
Kernel estimators

- How do we estimate the conditional expectation at x, $\mathbb{E}[Y \mid X = x]$?



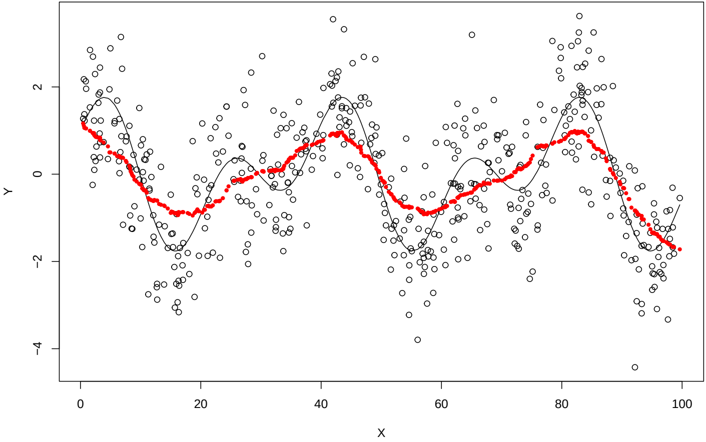
Kernel estimators

- We average Y within a small neighborhood of x's.
- The neighborhood has a width of 8 (the bandwidth).



Kernel estimators

- This is known as a kernel estimator, and we can use it to estimate the CEF non-parametrically.
- The estimator converges to $\mathbb{E}[Y \mid X]$ if $h \to 0$ as N grows.



Machine learning and A.I.

- A broad category of non-parametric estimators is known as machine learning (ML) algorithms.
- Kernel regression, splines, LASSO, ridge, regression tree, random forest, neural network, etc.
- They work well even with high-dimensional data, where the number of variables exceeds that of observations.
- They allow us to select variables to construct estimators with the smallest MSE.
- Artificial intelligence (A.I.) can learn the population from unstructured data such as texts, images, and videos.
- Non-parametric estimators may converge to the estimand very slowly and often lack intepretability.
- Many ML algorithms are blamed for being black boxes (e.g., recommendation algorithms).

Semi-parametric estimators

- Many problems in statistics involve two parameters.
- A low-dimensional one we care about (the estimand) and a high-dimensional one we need to estimate.
- The former is called the target parameter and the latter is known as the nuisance parameter.
- If we can estimate the nuisance parameter non-parametrically, the target parameter can be estimated via a parametric estimator.
- Such approaches are known as semi-parametric estimators.

Semi-parametric estimators

- Suppose we are interested in the relationship between ideology Y and having a college degree $X \in \{0, 1\}$.
- We know ideology can be affected by many other factors, such as age, gender, ethnicity, etc.
- Denoting these other factors as $\mathbf{Z}$, we can assume that

$$Y_i = \beta X_i + f(\mathbf{Z}_i) + \varepsilon_i.$$

- We first estimate $f(\cdot)$ non-parametrically and then estimate β conditioning on $\hat{f}(\mathbf{Z}_i)$.
- The boundary between non- and semi-parametric estimators is often blurry.
- Many consider the sample analogue as a semi-parametric estimator.