Lecture 2: Probability and statistics

Mitch Downey[1]

January 21, 2024

[1]IIES, Stockholm University. Slides heavily based on those developed by Markus Jäntti based on the textbook by Bruce Hansen.

Today

- Key question of Estimation (course part 1):
  What is a "good" guess for some parameter's value?
- Today/tomorrow: What does "good" mean?
- First, properties of an estimator
- Second, approaches to generating an estimator

- ***NOTE***: This is the only lecture with no corresponding chapter in Hansen

σ-fields

- A σ**-field**, called $\mathcal{F}$ is a collection of subsets of Ω that satisfies
  1. $\emptyset \in \mathcal{F}$
  2. $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$
  3. $A_1, A_2, \ldots \in \mathcal{F} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$
- The best way to think about a σ-field is as the set of possible states of the world
- Events within this set can be "impossible" in the sense of being zero probability
- But it must be conceptually possible to define them
- These properties can be translated as:
  1. It's possible nothing happens
  2. If A might have happened, then it's also possible that A didn't happen
  3. If A_1 and A_2 are both possible, then any event that's part of either of them is also possible
- σ-fields are the domains of probability measures

Random variables

- Let $\mathcal{F}$ be a σ-field
- We are interested in the random event f which is an element of $\mathcal{F}$
- The *random variable* y is a *function* $y(\omega)$ from the set Ω on to the set of numeric values y.
    - ω is the unobservable state of the world
    - $y(\cdot)$ is the function that maps the unobservable state of the world to something we can observe
    - y can be vector valued: $y(\omega) \in \mathbb{R}^k$
- In econometrics, we are interested in probability measures
- $\mathcal{F}$ is the collection of subsets of Ω for which a probability measure is defined
    - Defining measurability and formalizing this is technical, and we will not do it
- We call the rules governing the random variable $y()$ its *law*.

σ-field example

Suppose we flip a coin twice and record the outcome. We may define $\Omega = \{(H,H),(H,T),(T,H),(T,T)\}$ and let $\mathcal{F}$ be the set of all subsets of Ω including $\emptyset$ and Ω. Let $X : \Omega \to \mathbb{R}$ be the number of heads tossed. That is,

$$X(H,H) = 2, \;\; X(H,T) = 1, \;\; X(T,H) = 1, \;\; X(T,T) = 0.$$

The σ-field $\sigma(X)$ is given by

$$\sigma(X) = \Big\{\emptyset, \Omega, \big\{(H,H)\big\}, \big\{(H,T),(T,H)\big\}, \big\{(T,T)\big\}, \big\{(H,H),(T,T)\big\} \\ \big\{(H,H),(H,T),(T,H)\big\}, \big\{(T,H),(H,T),(T,T)\big\}\Big\}$$

This σ-field represents the information learned about Ω by observing X. Note that it does not allow us to distinguish between (H,T) and (T,H) since they both correspond to $X = 1$. (Later, this observation will be formalized as identification.)

Probability measures

- Given a sample space Ω and a sigma field $\mathcal{F}$, a *probability measure* P is a function $P : \mathcal{F} \longrightarrow [0, 1]$ that satisfies:
  - $P(\Omega) = 1$
  - If $A_1, A_2, \ldots \in \mathcal{F}$ are mutually disjoint then $P\big(\cup A_i\big) = \sum P(A_i)$
    - A_1 and A_2 are mutually disjoint if $a \in A_1 \Rightarrow a \notin A_2$ and $a \in A_2 \Rightarrow a \notin A_1$
- The following statements are true:
  - $P(\emptyset) = 0$
  - For any $A \in \mathcal{F}$ we have $P(A^c) = 1 - P(A)$
  - For any $A, B \in \mathcal{F}$ with $A \subset B$, then $P(A) \leq P(B)$
  - For any $A, B \in \mathcal{F}$, then $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
  - If $A_1, A_2, \ldots \in \mathcal{F}$ satisfy $A_n \subset A_{n+1} \forall n$, then $\lim_{n\to\infty} P(A_n) = P(\cup A_n)$
  - If $A_1, A_2, \ldots \in \mathcal{F}$ satisfy $A_{n+1} \subset A_n \forall n$, then $\lim_{n\to\infty} P(A_n) = P(\cap A_n)$
  - For $A_1, A_2, \ldots \in \mathcal{F}$, we have $P(\cup A_n) \leq \sum P(A_n)$

Distribution and density functions

- For some random variable $Y : \Omega \longrightarrow \mathbb{R}^k$, the *cumulative distribution function* (CDF) denoted $F_Y(\cdot)$ is defined as: $F_Y(y) = P(-\infty, y] = P(Y \leq y), y \in \mathbb{R}^k$
- Theorem: A random variable is uniquely identified by its CDF
    - Talk to me if you want this formalized
- The definition of the *probability density function* (pdf) – denoted $f_y(\cdot)$ – depends on whether the variable is discrete or continuous (defined below):
    - If a random variable is *discrete*:
        - Definition: There exists a countable set M such that $P(M) = 1$
        - Then $f(y) = P(y)$
    - If a random variable is *continuous*:
        - Definition: There is no countable set M such that $P(M) = 1$
        - Then $f(y)$ is the "derivative" of $F(y)$, which we will not formalize
        - See the Radon-Nikodym Theorem, more measure theory, and Lebesgue Integration

Independence

- Given some probability measure P and σ-field $\mathcal{F}$, two **sets** $A, B \in \mathcal{F}$ are *independent* if $P(A \cap B) = P(A)P(B)$
  - Generalization: Sets $A_1, A_2, \ldots \in \mathcal{F}$ are *mutually independent* if $P(\cap A_i) = \prod P(A_i)$
- Given some probability measure P and σ-field $\mathcal{F}$, two **σ-fields** $\mathcal{A}_1, \mathcal{A}_2 \subset \mathcal{F}$ are independent if for any $A_1 \in \mathcal{A}_1$ and $A_2 \in \mathcal{A}_2$, A_1 and A_2 are independent
- Given some probability measure P and σ-field $\mathcal{F}$, two **random variables** X, Y are independent if $\sigma(X)$ and $\sigma(Y)$ are independent
- Factorization of distribution functions: Consider two random vectors $X : \Omega \to \mathbb{R}^k$ and $Y : \Omega \to \mathbb{R}^\ell$. Define $Z : \Omega \to \mathbb{R}^{k+\ell}$ to be the stacked random vector given by $Z(\omega) = \big(X(\omega), Y(\omega)\big)$. Then X and Y are independent if and only if $F_Z(x, y) = F_X(x)F_Y(y)$.
  - Also implies $f_Z(x, y) = f_X(x)f_Y(y)$ (very important later)
- More intuitive definition: Two random variables X, Y are independent (written $X \perp Y$) if $F_{X|Y=y}(x) = F_X(x) \ \ \forall x, y$

Random sample

- A random sample is a collection of random variables $Y_1, Y_2, ..., Y_n$ with sample values $y_1, y_2, ..., y_n$
  - The former are random numbers, whereas the latter are not
- For each Y_i, we generally think of its law as being determined by a set of parameters θ. Learning about these parameters is the main goal of econometrics.
  - Example: $y_i = \alpha + \beta x_i + \varepsilon_i$
- To learn about θ, we often rely on *independently and identically distributed* (iid) random samples
- A random sample is iid if
  - Y_i, Y_j are independent (already defined) for all i, j
  - $F_i(y) = F_j(y) \;\; \forall i, j, y$ (where F_i is the CDF of Y_i and F_j is the CDF for Y_j)
- Observation: Time series econometrics is really hard...

Estimator

- An *estimator* is in general a function $\widehat{\theta}$ of the random variable(s) we are interested in, whose purpose is to provide us with an *estimate* of the value of θ.
- An estimator is a function of random variables and is therefore itself random. (This is why probability theory is an important tool.)
- There are usually very many possibile estimators for any particular problem. To assist in choosing a useful estimator we have a set of criteria.

Moments

- For some random variable Y with density $f_Y(y)$, its n^{th} moment is given by:

$$E(Y^n) \equiv \int_{-\infty}^{\infty} y^n f_Y(y) dy$$

- First moment is called the mean of Y or the expected value:

$$E(Y) \equiv \int_{-\infty}^{\infty} y f_Y(y) dy$$

- What's going on here?
  - We're integrating over all possible values of y
  - If discrete: Summing over possible values, each weighted by their probability
  - Integrating is continuous version, although $f_Y(y)$ can be greater than 1

Moments

- For some random variable Y with density $f_Y(y)$, its n^{th} moment is given by:

$$E(Y^n) \equiv \int_{-\infty}^{\infty} y^n f_Y(y)dy$$

- First moment is called the mean of Y or the expected value:

$$E(Y) \equiv \int_{-\infty}^{\infty} y f_Y(y)dy$$

- What's going on here?
  - We're integrating over all possible values of y
  - If discrete: Summing over possible values, each weighted by their probability
  - Integrating is continuous version, although $f_Y(y)$ can be greater than 1
- More generally:

$$E\big(g(Y)\big) = \int_{-\infty}^{\infty} g(y) f_Y(y)dy$$

- Useful fact:

$$E\big(E(Y)\big) = E(Y) \quad \text{(See law of iterated expectations)}$$

Centered moments

- For some random variable Y with density $f_Y(y)$, its n^{th} moment is given by:

$$E(Y^n) \equiv \int_{-\infty}^{\infty} y^n f_Y(y) dy$$

- The centered n^{th} moment is given by:

$$E\big(Y - E(Y)\big)^n \equiv \int_{-\infty}^{\infty} \big(y - E(y)\big)^n f_Y(y) dy$$

- The centered second moment is called the variance
- It is a symmetric measure of dispersion:
  Expected squared distance between a realization and its expected value

Properties of variance

- The centered 2^{nd} moment is the variance:

$$E\big(Y - E(Y)\big)^2 \equiv \int_{-\infty}^{\infty} \big(y - E(y)\big)^2 f_Y(y)dy$$

- Property 1:

$$\begin{aligned} Var(Y) &= E\big(Y - E(Y)\big)^2 = E\Big[Y^2 - 2YE(Y) + \big(E(Y)\big)^2\Big] \\ &= E\big[Y^2\big] - 2E(Y)E(Y) + \big[E(Y)\big]^2 \\ &= E\big[Y^2\big] - \big[E(Y)\big]^2 \end{aligned}$$

- Property 2:

$$Var(aY) = a^2 Var(Y)$$

  (proof is simple)

- Property 3:

$$Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$$

  (definition, next slide; proof in problem set 1)

Covariance

- The centered second moment is called the variance:

$$Var(Y) = E\left[Y - E(Y)\right]^2 = E\left[\left(Y - E(Y)\right)\left(Y - E(Y)\right)\right]$$

- We often care about the relation between two random variables
- The linear notion of this is covariance

$$Cov(X, Y) = E\left[\left(Y - E(Y)\right)\left(X - E(X)\right)\right]$$

  - Note $Cov(Y, Y) = Var(Y)$
- This only captures mean independence, true independence is stronger
  - Theorem: $X \perp Y \Rightarrow Cov(X, Y) = 0$ (proof is simple)
  - $Cov(X, Y) = 0 \nRightarrow X \perp Y$: See problem set 1
  - Matters for structural vs. reduced form methods, and linear vs. non-linear models

Properties of an estimator: bias

- An *unbiased* estimator is such that

$$\mathrm{E}[\widehat{\theta}] = \theta. \tag{1}$$

- A biased estimator is one for which this is not the case:

$$\mathrm{E}[\hat{\theta}] - \theta = \text{bias} \neq 0. \tag{2}$$

Properties of an estimator: efficiency

- Since an estimator is a random variable it has not only an expectation but also a variance. For two alternative unbiased estimators $\widehat{\theta}_E$ and $\widehat{\theta}_I$, we say that $\widehat{\theta}_E$ is *more efficient* if
$$\text{Var}[\widehat{\theta}_E] \leq \text{Var}[\widehat{\theta}_I]. \tag{3}$$
- Note 1: If θ is a vector, we have to be more precise about this comparison
- Note 2:
  - $\widehat{\theta}$ is a random variable because it is a function of some random variable Y
  - Properties (e.g., variance) of $\widehat{\theta}$ depend on properties (e.g., variance) of Y
  - Without knowing the distribution of Y we cannot know the distribution of $\widehat{\theta}$
  - Debates about the *robustness* of an estimator are often about how sensitive the properties of $\widehat{\theta}$ are with respect to different distributions of Y

Mean squared error (MSE)

- Should we ignore biased estimators?

Mean squared error (MSE)

- Should we ignore biased estimators?
- No, version 1
- Most econometricians compare estimators based on mean squared error (MSE):

$$\begin{aligned} \mathrm{MSE}[\hat{\theta}|\theta] &= E[(\hat{\theta}-\theta)^2] \\ &= \mathrm{Var}[\hat{\theta}] + (\mathrm{bias}[\hat{\theta}|\theta])^2. \end{aligned} \tag{4}$$

- Problem set 1: A biased estimator that is clearly better than an unbiased one
  - Also basic properties of the sample mean and sample variance

Convergence in probability (plim)

- A random variable y_n is said to converge in probability to a random variable y if

$$\lim_{n\to\infty} \Pr(|y_n - y| > \epsilon) = 0, \forall \epsilon > 0. \tag{5}$$

- We denote this as

$$\operatorname{plim} y_n = y \tag{6}$$

- A special case is when $y = c$ is a constant.

Convergence in probability (plim)

- *The weak law of large numbers*
  - Suppose we have an iid sample from a population with a finite mean
  - Then as $n \to \infty$, the sample mean converges in probability to the population mean:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \xrightarrow{pr} \mathrm{E}[y]. \tag{7}$$

  - Most underrated theorem in economics (my opinion), see problem set 1 for examples
  - Note: Other LLN's apply under other conditions
- When

$$\operatorname{plim} \widehat{\theta} = \theta \tag{8}$$

  we say that the estimator $\widehat{\theta}$ is *consistent*.
- We will, for reasons to become clear later, often be much better placed to know if estimators are consistent than if they are unbiased.

Convergence in probability (plim)

In case we have a vector-valued variable, y, things are a little more complicated.

- expectation: is a vector of expectations of each of the elements of y (i.e., the marginal means):

$$E[y] = \mu = \begin{bmatrix} E[y_1] \\ \vdots \\ E[y_K] \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_K \end{bmatrix}. \tag{9}$$

The variance matrix of y is

$$\mathrm{Var}[y] = \boldsymbol{V} = E[(y - \mu)(y - \mu)'] \tag{10}$$

Convergence in distribution

- A random variable y_n is said to converge in distribution to a random variable y if

$$\lim_{n\to\infty} |F_n(y) - F(y)| = 0 \tag{11}$$

  at all continuity points of y.
- Example: Take $y_n \in \{1, 2\}$ with distribution

$$\begin{aligned} \Pr(y_n = 1) &= 1/2 + 1/(n+1) \\ \Pr(y_n = 2) &= 1/2 - 1/(n+1) \end{aligned} \tag{12}$$

  which converges in distribution to the random variable $y \in (1, 2)$ with $\Pr(y = 1) = \Pr(y = 2) = 1/2$.
- More interesting and useful example in problem set 1.

Central limit theorem

- *Lindberg-Levy Central Limit Theorem*: Let $y_1, y_2, ..., y_n$ be an iid sequence of random variables with mean $E(y) = \mu$ and finite variance σ^2. Then

$$\sqrt{n}(\bar{y} - \mu) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (y_i - \mu) \xrightarrow{d} N(0, \sigma^2) \tag{13}$$

  - There are other CLT's that can be proved without iid
- If y_n converges in distribution to y with $F_y(y)$, denoted $y_n \xrightarrow{d} y$, then $F_y(y)$ is called the *limiting distribution*.

Central limit theorem

- *Lindberg-Levy Central Limit Theorem*: Let $y_1, y_2, ..., y_n$ be an iid sequence of random variables with mean $E(y) = \mu$ and finite variance σ^2. Then

$$\sqrt{n}(\bar{y} - \mu) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (y_i - \mu) \xrightarrow{d} N(0, \sigma^2) \tag{13}$$

  - There are other CLT's that can be proved without iid
- If y_n converges in distribution to y with $F_y(y)$, denoted $y_n \xrightarrow{d} y$, then $F_y(y)$ is called the *limiting distribution*.
- Note 1: Also holds for multivariate Y converging to multivariate normal.
- Note 2: This is **incredibly** powerful. No matter what the distribution of y is, the limiting distribution of the sample mean is a distribution we understand. This is the backbone of hypothesis testing.
- Note 3: What if you can't use a CLT because *i*) the sample isn't truly iid in the sense that it's not "i" (serial dependence or clustering)? *ii*) the sample isn't truly iid in the sense that it's not "id" $\big($for some $i, j, y,\ F_i(y) \neq F_j(y)\big)$? *iii*) you don't have infinite sample? These are what bootstrapping and randomization inference are for. More in Econometrics II.

Properties of convergence:

- in probability:
  Let $\text{plim}\, X_n = c$, $\text{plim}\, y_n = d$ be two stochastic variables and their probability limits, and $\text{plim}\, \boldsymbol{V}_n = \Sigma$, $\text{plim}\, \boldsymbol{W}_n = \Omega$ two (conforming) matrices that limit the matrices.

$$\begin{array}{ccr}
\text{plim}(X_n + y_n) & = \text{c + d} & \text{sum} \\
\text{plim}(X_n y_n) & = \text{c d} & \text{product} \\
\text{plim}(X_n / y_n) & = \text{c / d, d} \neq 0 & \text{ratio} \\
\text{plim}\, W_n^{-1} & = \Omega^{-1} & \text{matrix inverse} \\
\text{plim}\, V_n W_n & = \Sigma\Omega & \text{matrix product}
\end{array}$$

- in distribution:
  For $X_n \xrightarrow{d} X$, $\text{plim}\, y_n = c$, and a continuous function $g()$,

$$\begin{array}{cc}
\text{X}_n y_n \xrightarrow{d} & \text{c X} \\
\text{X}_n + y_n \xrightarrow{d} & \text{X+ c} \\
\text{X}_n / y_n \xrightarrow{d} & \text{X/c, c} \neq 0 \\
\text{g}(\text{X}_n) \xrightarrow{d} & \text{g(X)}
\end{array}$$