EKT-816 Lecture 1

Probability Review (1)

Jesse Naidoo

University of Pretoria

PDFs, CDFs, and Quantiles

- Discrete distribution:
- mass functions: $f(x) = P(X = x)$.
- cumulative distribution functions: $F(x) = P(X \leq x)$.
- Examples: Bernoulli(p); binomial(n, p); Poisson(λ).
- continuous distributions:
- density function $f_X(x)$ such that

$$\int_{-\infty}^{\infty} f_X(x)dx = 1.$$

- CDF $F_X(x)$ is increasing and such that $\lim_{x \to -\infty} F_X(x) = 0$, $\lim_{x \to \infty} F_X(x) = 1$.
- $F_X'(x) = f_X(x)$, or

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt$$

- the τ-th quantile of the distribution F is the value x_τ such that

$$F(x_\tau) = \tau$$

Moments

- the *mean* of a distribution is

$$E[X] = \int_{-\infty}^{\infty} x f_X(x) dx$$

- the *variance* of the distribution is

$$V[X] = E[(X - \mu)^2]$$

  where μ is the mean of the distribution
- note, these moments may not exist!
  - but, if $V[X] < \infty$, the mean will exist (why?)
  - also notice that $V[X] = E[X^2] - E[X]^2$
  - third (centered) moment is called *skewness*
  - fourth (centered) moment is called *kurtosis*

Example: Pareto Distributions

- let $\alpha > 0$ be some constant
- density is

$$f_X(x) = \begin{cases} \alpha x^{-(\alpha+1)} & \text{if } x > 1 \\ 0 & \text{else} \end{cases}$$

  - what is the CDF, $F_X(x)$?
  - what is the mean, $E[X]$? do we have to impose any conditions to ensure the mean exists?
  - what is the variance, $V[X]$?

Inverse CDF Trick

- suppose we want to generate random numbers from some distribution with CDF F
- we can compute F and F^{-1}
- we can generate uniformly distributed random numbers, $U \sim U(0, 1)$
- then, you can generate $X \sim F$ as follows:

$$X = F^{-1}(U)$$

- proof: let's find the CDF of such Xs.
- let x be an arbitrary number; we're going to show that $P(X \leq x) = F(x)$
- $P(X \leq x) = P(F^{-1}(U) \leq x) = P(U \leq F(x)) = F(x)$

Marginal and Conditional Distributions

- take a joint density $f_{XY}(x, y)$ that integrates to 1 over $\mathbb{R}^2$
- the marginal density of X is

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y) dy$$

- analogous for marginal of Y
- the *conditional* density of Y *given that* $X = x$ is

$$f_{Y|X}(y|X = x) = \frac{f_{XY}(x, y)}{f_X(x)}$$

Basic Rules

- expectations are linear: $E[aX + Y] = aE[X] + E[Y]$
- $V[aX] = a^2 V[X]$
- $V[X + Y] = V[X] + V[Y] + 2\,\mathrm{cov}(X, Y)$

Example: Zero Correlation, But Not Independent

- consider the following distribution:

$$f_{XY}(x, y) = \begin{cases} 3/4 & \text{if } x \in (-1, 1) \text{ and } y \in (0, 1 - x^2) \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

- show that $\text{cov}(X, Y) = 0$
- yet, the two are not independent!
- to see this, compute the conditional expectation $E[Y|X]$

Law of Iterated Expectations and Variance Decomposition

- law of iterated expectations:

$$E[E[Y|X]] = E[Y]$$

- variance decomposition:

$$V[Y] = V[E[Y|X]] + E[V[Y|X]]$$

Classical Statistical Paradigm

- sample data $X_1, \ldots X_n$ are draws from some *data-generating process* $f(x|\theta)$
- θ: a vector of parameters - unknown to us
- our goal is to learn about θ from the sample
- a *statistic* is any function of the data (or *known* parameters)
- as such, they are themselves random variables
- and, they have a distribution
- which we would like to characterize as much as possible
- why do we care about this? want to answer two questions
- Given enough data, will our estimate "eventually" get "close" to θ_0?
- For any fixed sample, how "close" is our estimate "likely" to be to the truth θ_0?
- asymptotic theory is useful because it allows us to answer these questions using a precise meaning for the words "close", "eventually", and "likely"

Modes of Convergence

- $(X_n)_{n=1}^{\infty}$ converges *in probability* to the random variable X if for all $\varepsilon > 0$,

$$\lim_{n\to\infty} P(|X_n - X| > \varepsilon) \quad = \quad 0 \tag{2}$$

  - We write $X_n \longrightarrow_p X$ as shorthand.
- $(X_n)_{n=1}^{\infty}$ converges *in mean square* to the random variable X if

$$\lim_{n\to\infty} E[(X_n - X)^2] \quad = \quad 0 \tag{3}$$

  - We write $X_n \longrightarrow_{m.s.} X$ as shorthand.
- Convergence in mean square implies convergence in probability, but *not* vice versa

Law(s) of Large Numbers

- consider $(X_i)_{i=1}^{\infty}$ i.i.d. with mean μ and variance σ^2
- let $\overline{X}_n = n^{-1} \sum_{k=1}^{n} X_k$
- $V[\overline{X}_n] = n^{-1}\sigma^2$, so $\overline{X}_n \longrightarrow_{m.s.} 0$
- thus, $\overline{X}_n \longrightarrow_p 0$ also
- we say "the sample mean is *consistent* for the population mean"
- consistency is a necessary condition for an estimator to be useful
- if you're never going to get the truth out of this calculation, why bother?

Central Limit Theorems

- besides consistency we would like to know about the *precision* of our estimates
- it is good to know that we get to the truth "eventually", but how close are we right now?
- we need a different notion of convergence to characterize the asymptotic approximation here
- *convergence in distribution*: we say $(X_n)_{n=1}^{\infty} \longrightarrow_d X$ if,
- for every point a where $F_X(\cdot)$ is continuous, $P(X_n \leq a) \longrightarrow P(X \leq a)$
- *Central Limit Theorem*:
    - if $(X_i)_{i=1}^{\infty}$ is an i.i.d. sample from a distribution with $V[X] = \sigma^2 < \infty$ and $E[X] = \mu$, then

$$\sqrt{n}\left(\frac{n^{-1}\sum_{k=1}^{n} X_k - \mu}{\sigma}\right) \quad \longrightarrow_d \quad N(0,1) \tag{4}$$

    - this suggests we may use the approximation

$$P(\overline{X}_n \leq a) \quad \approx \quad \Phi\left(\frac{a-\mu}{\sigma/\sqrt{n}}\right) \tag{5}$$

Monte Carlo Simulations

- we need a way to examine the sampling distribution of some estimator $\widehat{\theta}(X_1, \dots X_n)$
- except for special cases (e.g. the mean of normal observations), we will not be able to calculate the distribution of a general function of the data for arbitrary sample sizes
- solution: approximate the distribution of $\widehat{\theta}(X_1, \dots X_n)$ by simulating some large number of datasets (each of size n)
- pseudocode:
- B is number of simulated datasets
- for each $b = 1, \dots B$ we generate a dataset of size n; compute $\widehat{\theta}_{n,b}$
- treat the resulting sample $\widehat{\theta}_{n,1}, \widehat{\theta}_{n,2}, \dots \widehat{\theta}_{n,B}$ as the population distribution

Monte Carlo Simulations

for n = 10, 100, 1000, 10000 {
  for b = 1, ... B {

    draw $X_1 \dots X_n$ from $f(X|\theta_0)$;
    calculate $\widehat{\theta}_{n,b}$ from $X_1 \dots X_n$;
    store $\widehat{\theta}_{n,b}$;
    discard $X_1 \dots X_n$;

  }
}

- then, using $\widehat{\theta}_{n,1}, \widehat{\theta}_{n,2}, \dots \widehat{\theta}_{n,B}$:
- plot the density of $(\widehat{\theta}_{n,b})_{b=1}^{B}$ for $n = 10, 100, 1000, 10000$
- compute the variance of $(\widehat{\theta}_{n,b})_{b=1}^{B}$ for $n = 10, 100, 1000, 10000$
- etc.

References

Table of Contents