Lecture 1: Statistics Review

Charles J. Saunders

University of Western Ontario

EC 2123: Econometrics II

Definition of a Distribution

A distribution provides information on the relative number of times (probability or share or proportion) each possible outcome for a variable will occur in a number of trials.

- The *probability density function* (pdf) gives the probability of observing a given value.
  - The integral of the pdf from $-\infty$ to ∞ must equal one.
- The *cumulative distribution function* (cdf) is the cumulative probability of observing a value less than or equal to a given value of the variable.
  - The cdf is monotonic (cannot decrease).

Notation

- Let $F_X(x)$ represent the **cdf** of a random variable X, where

$$F_X(x) = P(X \leq x)$$

- Let $f_x(x)$ represent the **pdf** of a random variable X, where the pdf is the derivative of the cdf:

$$\begin{aligned} f_X(x) &= F'_X(x) \\ &= P(X = x) \end{aligned}$$

- The reverse relationship:

$$F_X(x) = \int_{-\infty}^{x} f_X(w)dw$$

Conditional Distribution

The cdf and pdf can be conditional on other variables

- The conditional pdf is then written as:

$$f_Y(y|x) = P(Y = y|X = x)$$

- The conditional cdf is:

$$\begin{aligned} F_Y(y|x) &= \int_{-\infty}^{y} f_Y(w|x)dw \\ &= P(Y \leq y|X = x) \end{aligned}$$

- The probability that y occurs given x is known (observed)

Joint Distribution

A joint distribution describes the probability that two events occur, X and Y,

- Joint pdf:

$$f_{Y,X}(y,x) = P(Y = y, X = x)$$

- Joint cdf:

$$\begin{aligned} F_{Y,X}(y,x) &= \int_{-\infty}^{x}\int_{-\infty}^{y} f_{Y,X}(w,z)dw\, dz \\ &= P(Y \le y, X \le x) \end{aligned}$$

Marginal Distribution

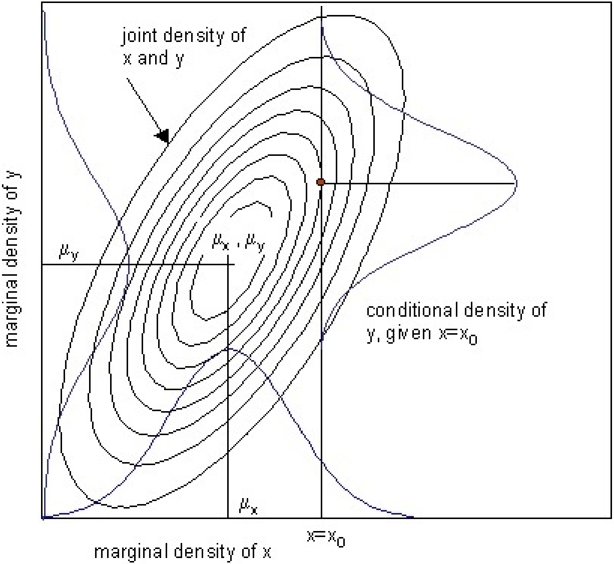
- The marginal distribution of Y is obtained by integrating over all possible values of X.

$$
\begin{aligned}
F_{Y,X}(y, x = \infty) &= \int_{-\infty}^{y} \int_{-\infty}^{\infty} f_{Y,X}(w, z) dw\, dz \\
&= P(Y \leq y, X \leq \infty)
\end{aligned}
$$

- The probability that X is less than infinity is one, so

$$
\begin{aligned}
F_{Y,X}(y, \infty) &= F_Y(y) \\
&= P(Y \leq y)
\end{aligned}
$$

Visual representation of Joint, Marginal, and Conditional



Bayes Theorem

Formal relationship between the probability of observing two events (joint) and the conditional and marginal probabilities.

$$
\begin{aligned}
P(Y = y, X = x) &= P(X = x|Y = y)P(Y = y) \\
&= P(Y = y|X = x)P(X = x)
\end{aligned}
$$

The pdf form is used extensively in econometrics and statistics:

$$
f_{Y,X}(y, x) = f_X(x|y)f_Y(y) = f_Y(y|x)f_X(x) = f_{X,Y}(x, y)
$$

Discrete Random Variables

- Similar to the continuous random variables
- Let $F_X(x)$ represent the **cdf** of a discrete random variable X, where

$$F_X(x) = P(X \leq x)$$

- Let $f_x(x)$ represent the probability mass function (**pmf**) of a discrete random variable X, where the pdf is the derivative of the cdf:

$$f_X(x) = P(X = x)$$

- The cdf-pmf relationship is given by the sum:

$$F_X(x) = \sum_{-\infty}^{x} f_X(w)$$

- What do the cdf and pmf look like?

Moments of Random Variables

- The 1^{st} moment of a random variable (mean) is given by:

$$m_1 = E[X] = \int_{-\infty}^{\infty} x f_X(x) dx$$

  and commonly denoted as μ_x.
- Other central moments are defined as:

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

  for $\ell \geq 2$.

Population Moments and Sample Moments

- Population:
  - Mean:

$$m_1 = E[X] = \int_{-\infty}^{\infty} x f_X(x) dx = \mu_x$$

  - Variance:

$$m_2 = V[x] = E[(X - \mu_x)^2] = \int_{-\infty}^{\infty} (X - \mu_x)^2 f_X(x) dx = \sigma_x^2$$

- Sample:
  - Average:

$$\hat{\mu}_x = \frac{1}{N} \sum_{i=1}^{N} x_i$$

  - Variance:

$$\hat{\sigma}_x^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \hat{\mu}_x)^2$$

Conditional Expectation

- The expectation of a random variable (Y) for a given or known value of another random or nonrandom variable ($X = x$)
$$E[Y|x] = \int_{-\infty}^{\infty} y f_{Y,X}(y|x) dy$$
- What does this look like on our graphs?

Expectation Rules

1. Expectation of the sum is the sum of the expectations

$$E[X + Y + Z] = E[X] + E[Y] + E[Z]$$

2. Expected value of a constant is that constant

$$E[b] = b$$

3. Expectation of a constant times a random variable is that constant times the expectation of the random variable

$$E[bX] = bE[X]$$

Covariance

- The expectation of Y times X

$$\begin{aligned} cov(Y,X) &= \sigma_{YX} = E[(Y-\mu_Y)(X-\mu_X)] \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(y-\mu_Y)(x-\mu_X)f_{Y,X}(y,x)dy\ dx \end{aligned}$$

- If Y and X are **independent** then the covariance is zero

$$\begin{aligned} cov(Y,X) &= E[(Y-\mu_Y)]E[(X-\mu_X)] = 0 \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(y-\mu_Y)(x-\mu_X)f_{Y,X}(y,x)dy\ dx \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(y-\mu_Y)(x-\mu_X)f_Y(y)f_X(x)dy\ dx \\ &= \int_{-\infty}^{\infty}(y-\mu_Y)f_Y(y)dy\int_{-\infty}^{\infty}(x-\mu_X)f_X(x)dx \end{aligned}$$

Covariance Rules

1. Covariance of X and a sum $Y = V + W$

$$cov(X, Y) = cov(X, V) + cov(X, W)$$

2. Covariance of X and a random variable times a constant ($Y = bW$)

$$cov(X, Y) = b\, cov(X, W)$$

3. Covariance of a random variable and a constant is zero

$$cov(X, b) = 0$$

Variance Rules

1. Variance of the sum ($Y = V + W$)

$$V[Y] = V[V] + V[W] + 2cov(V, W)$$

2. Variance of the difference ($Y = V - W$)

$$V[Y] = V[V] + V[W] - 2cov(V, W)$$

3. Variance of constant times random variable ($Y = bV$)

$$V[Y] = b^2 V[V]$$

4. Variance of constant ($Y = b$)

$$V[Y] = 0$$

5. Variance of constant plus random variable ($Y = b + V$)

$$V[Y] = V[V]$$

Unbiased Estimators

- An estimator is unbiased if the expected value is equal to the population characteristic
- Consider the following estimators of the mean:

$$\hat{\mu}_x = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\hat{\mu}_x = \frac{1}{N} + \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\hat{\mu}_x = 0.1 + \frac{1}{N}\sum_{i=1}^{N} x_i$$

Probability Limit

- The **probability limit** of a sequence of random variables (X_N), written as $plim(X_N) = a$

$$\lim_{N \to \infty} P(|X_N - a| > \epsilon) \longrightarrow 0$$

- plim Rules
  1. plim(X+Y+Z) = plim(X) +plim(Y) + plim(Z)
  2. plim(XY) = plim(X)plim(Y)
  3. plim(X/Y) = plim(X)/plim(Y)
  4. plim(b) = b
  5. plim(f(X)) = f(plim(X))

Consistency

- An estimator is said to be consistent if
  1. The estimator collapses to a 'spike' as $N \longrightarrow \infty$
  2. The spike is located at the true value of the population
- The plim is used to prove consistency

Central Limit Theorem

- The mean of a random variable ($\bar{X}$) converges to a *normal distribution* with variance, σ_X^2

$$\sqrt{N}(\bar{X} - \mu_X) \xrightarrow{d} \mathcal{N}(0, \sigma_X^2) \tag{1}$$

- This is important because CLT does not require that X is normally distributed

Summary Statistics

- **ALWAYS EXAMINE YOUR DATA!!**
- **Statistics**: mean, median, mode, standard deviation, variance, skewness, kurtosis
- **Characteristics**: minimum, maximum, range, sum, count (number of observations)
- In reports, provide a summary statistics table
- **Graph** data to look for outliers or oddities

Logarithm and Exponent Rules

- Logarithms and exponents are used a lot in economics and econometrics
- Log Rules:

$$\begin{aligned} log(XY) &= log(X) + log(Y) \\ log(X/Y) &= log(X) - log(Y) \\ \frac{\partial log(X)}{\partial X} &= \frac{1}{X} \end{aligned}$$

- Exponent Rules:

$$\begin{aligned} exp(log(X)) &= log(exp(X)) = X \\ \frac{\partial exp(X)}{\partial X} &= exp(X) \end{aligned}$$

Summation Rules

- Summations are used extensively in this course
- c is constant and x_i is random

$$\begin{aligned}
\sum_{i=1}^{N} c &= c + c + c + \ldots + c = Nc \\
\sum_{i=1}^{N} cx_i &= cx_1 + cx_2 + cx_3 + \ldots + cx_N = c\sum_{i=1}^{N} x_i \\
E\left[\sum_{i=1}^{N} x_i\right] &= \sum_{i=1}^{N} E\left[x_i\right] \\
\frac{\partial}{\partial z}\left(\sum_{i=1}^{N} zx_i\right) &= \sum_{i=1}^{N} \frac{\partial}{\partial z}\left(zx_i\right) = \sum_{i=1}^{N} x_i
\end{aligned}$$

Simple Derivatives

- This course uses many derivatives

$$
\begin{aligned}
\frac{\partial c}{\partial z} &= 0 \\
\frac{\partial z}{\partial z} &= 1 \\
\frac{\partial z^2}{\partial z} &= 2z \\
\frac{\partial}{\partial z}\left(\sum_{i=1}^{N} z^2\right) &= \sum_{i=1}^{N} \frac{\partial}{\partial z}\left(z^2\right) = \sum_{i=1}^{N} 2z = 2zN \\
\frac{\partial}{\partial z}\left(\sum_{i=1}^{N} (zx_i)^2\right) &= \sum_{i=1}^{N} \frac{\partial}{\partial z}\left(z^2 x_i^2\right) = \sum_{i=1}^{N} 2zx_i^2
\end{aligned}
$$

Geometric series

$$\begin{aligned}
(1+b+b^2+\ldots+b^N) &= \frac{(1-b)}{(1-b)}(1+b+b^2+\ldots+b^N) \\
&= \frac{1+b+b^2+\ldots+b^N}{(1-b)} \\
&\quad -\frac{b+b^2+b^3+\ldots+b^{N+1}}{(1-b)} \\
&= \frac{1-b^{N+1}}{1-b}
\end{aligned}$$

- Three cases when $N \to \infty$:

$$\begin{cases} |b|<1 & \Rightarrow \frac{1}{1-b} \\ b=1 & \Rightarrow +\infty \\ |b|>1 & \Rightarrow +\infty \ or \ -\infty \end{cases}$$