J. Angrist

Lecture Note 1

Describing Distributions: Expectation and Moments

1 Random Variables, Distributions, and Samples

- X_i denotes a random variable. What's a random variable? An attribute we measure, most often of people. Subscript i on X_i reminds us that we see this for a particular person.
- The *probability distribution* of X_i is the relative frequency of values X_i assumes in the population over which it's defined. The *population* of interest contains all possible units we might see. Populations can be concrete, like the US population (X_i might be age or earnings). Although the US population is finite, we do no harm by thinking of it as infiinite.
  - X_i can also be generated by a stochastic process (X_i might encode heads or tails as we toss a coin repeatedly, or a daily stock return).
- We use samples to learn about populations. A *sample* includes info on X_i for a finite number of units. Sampled units are indexed by $i = 1, ..., n$, where n is the sample size and i is the order in which they're sampled.
  - Tricky point: X_i is random variable ... until it's not. If I tell you, say, that we observe $X_i = 6$, then it's no longer random (it's the number 6).

1.1 Expectation, $E[X_i]$

- *Expectation* is the population analog of a sample average
- discrete r.v.: $X_i \in \{x_1, ..., x_J\}$

$$E(X_i) = \sum_j x_j p(x_j)$$

  where x_j is one of $j = 1, ..., J$ values that X_i can take on and $p(x_j)$ is the probability that $X_i = x_j$
  - What's the expectation of Bernoulli X_i?
- continuous r.v.

$$E(X_i) = \int_{-\infty}^{\infty} t f_X(t) dt$$

  where $f_X(t)$ is the probability density function (*pdf*) of X_i (not "PDF", yo), sometimes referred to as simply "the density of X_i"

  - The probability any continuously distributed r.v. takes on a particular value is zero! But the probability that continuous X_i falls in the interval $[a, b]$ is the integral $\int_a^b f_X(t)dt$
  - What's the *pdf* and expectation of uniformly distributed X_i?

- Learning the lingo
  - The expectation of a random variable is a *parameter* that describes its distribution. Our people often label parameters in Greek, sometimes writing μ_X for $E(X_i)$. We also say: "μ_X is the population mean of X_i". When it's clear what you're talking about, ditch the subscript and just write μ.

- Draw a sample of n *observations* of r.v. X_i: we *estimate* $E(X_i)$ using the sample mean:

$$\bar{X} \equiv \frac{1}{n}\sum_{i=1}^{n} X_i$$

  - When it's important to keep track of sample size, we write $\bar{X}_n$ (e.g., when using the law of large numbers)

Exercise In a sample of size n, show that the sample mean of a discrete r.v. satisfies:

$$\bar{X} \equiv \frac{1}{n}\sum_{i=1}^{n} X_i = \sum_j x_j \hat{p}(x_j),$$

where $\hat{p}(x_j)$ is the sample proportion with $X_i = x_j$.

1.2 Moments

- The rth population moment of random variable X_i is defined as $E(X_i^r)$
- The rth central population moment of random variable X_i is $E[(X_i - \mu_X)^r]$
- The moments of X_i characterize its distribution
  - The mean, $E(X_i)$, is a *first moment*, sometimes said to be a measure of location
  - The variance, a *second moment*, measures the dispersion of X_i around the mean:

$$\sigma_X^2 = V(X_i) = E[(X_i - \mu_X)^2]$$

- The sample variance, s_X^2, replaces expectations with sample averages:

$$s_X^2 = \frac{1}{n}\sum (X_i - \bar{X})^2$$

(Sometimes we divide by $n-1$ instead of n, so that the resulting estimator is *unbiased*.)

- The 3rd central moment measures *skewness*, the extent to which a distribution is asymmetric; the 4th central moment measures *kurtosis*, or the likelihood of tail events (usually in comparison with tail probabilities for a Normal distribution). We're mostly concerned with first and second moments.

1.3 Expectation and Variance: Rules and Properties[1]

1. *Expectation of linear functions.* Let $Z_i = a + bX_i + cW_i$ for constants a, b, c and random variables X_i and W_i. Then

$$E(Z_i) = a + bE(X_i) + cE[W_i]$$

   The proof uses the law of the unconscious statistician, which tells us how to evaluate the expectation of a function of a random variable (discussed in recitation).

2. *Ways to write variance*

$$V(X_i) = \sigma_X^2 = E[(X_i - \mu_X)^2] = E(X_i^2) - \mu_X^2$$

   Proof:

$$\sigma_X^2 = E[(X_i - \mu_X)^2] = E[X_i^2 + \mu_X^2 - 2X_i\mu_X] = ...$$

   (now use #1). How many ways to write variance? Three, (3), (iii)!

3. *Variance of a linear function.* For any constants, a, b:

$$V(a + bX_i) = b^2\sigma_X^2.$$

   Be sure you can show this.

4. *Mean-squared error (*MSE*).* Suppose you'd like to predict the realization of random variable X_i. You get one chance: your prediction is a constant. The MSE of X_i around any constant, c, is the expectation of squared prediction errors:

$$\begin{aligned} MSE_X(c) &= E(X_i - c)^2 = \sigma_X^2 + (c - \mu_X)^2 \\ &= variance + bias^2 \end{aligned}$$

   Proof:

$$(X_i - c)^2 = [(X_i - \mu_X) + (\mu_X - c)]^2 = (X_i - \mu_X)^2 + (\mu_X - c)^2 + 2(X_i - \mu_X)(\mu_X - c)$$

   Now take expectations and use #1.

5. *Setting $c = \mu_X$ minimizes $MSE_X(c)$.* Proof: use #4. The fact that μ_X is the minimum MSE predictor of X_i is a good reason to be interested in it. (Sounds kinda like machine learning, but this idea is nothing new.)

6. These properties hold in samples, e.g., $s_X^2 = \frac{1}{n}\sum X_i^2 - \bar{X}^2$

[1]Important! We brush our teeth with these daily.

2 Bivariate Distributions: Characterizing Relationships Between Random Variables

2.1 Joint Moments

- An $r+s$ joint moment is $E(X_i^r Y_i^s)$
- An $r+s$ joint central moment is $E[(X_i - \mu_X)^r (Y_i - \mu_Y)^s]$

Moments of special importance:

Covariance $C(X_i, Y_i) = E[(X_i - \mu_X)(Y_i - \mu_Y)]$

(note that $r+s=2$, so this is a joint *second* moment)

Correlation $\rho_{XY} = \frac{C(X_i,Y_i)}{\sigma_X \sigma_Y} \in [-1, 1]$

Covariance and Correlation measure the extent of linear relationship between X_i and Y_i
(more on this soon).

2.2 Conditional Expectation

The conditional expectation of Y_i given X_i is the expected value of Y_i when X_i is fixed at a particular value.

- discrete r.v.: $Y_i \in \{y, ..., y_J\}$

$$E(Y_i|X_i = x) = \sum_j y_j f_2(y_j|X_i = x)$$

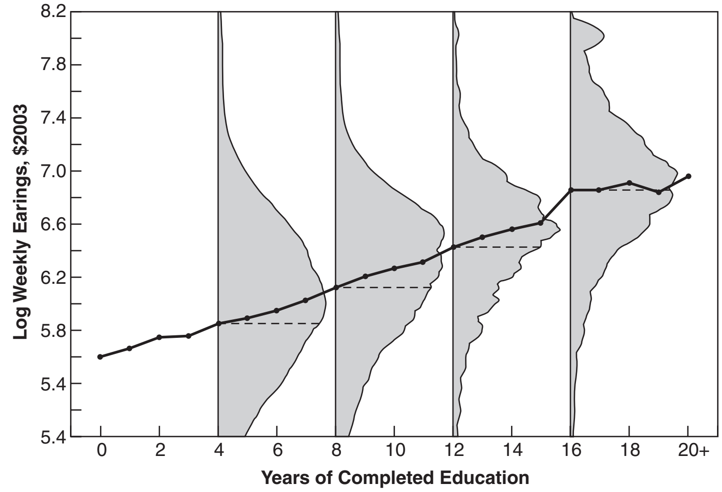
- continuous r.v.

$$E(Y_i|X_i = x) = \int_{-\infty}^{\infty} t f_2(t|X_i = x) dt$$

CEF The *Conditional Expectation* Function, written $E(Y_i|X_i)$, shows how mean Y_i varies as a function of X_i, without specifying which value of X_i we have in mind.

- $E(Y_i|X_i)$ is a random variable and therefore has a distribution. Why? Because the CEF is a function of random variable X_i, and functions of random variables are random variables too.

- Schooling and wages: conditional distributions



- Schooling and wages: a CEF and the regression line that fits it

- Interesting fig, but what do these conditional means *mean*? An important question . . . one we'll visit and revisit in the weeks to come

2.3 Covariance and CEFs: Rules and Properties[2]

1. *Ways to write covariance*

$$C(X_i, Y_i) = E[(X_i - \mu_X)(Y_i - \mu_Y)] = E(X_iY_i) - \mu_X\mu_Y$$

[2]Very important! *How* important, you ask? Very.

This means that if $E(X_i)$ or $E(Y_i) = 0$ then $C(X_i, Y_i) = E(X_iY_i)$. Also,

$$C(X_i, Y_i) = E[Y_i(X_i - E(X_i))] = E[X_i(Y_i - E(Y_i))]$$

How many ways to write covariance? Three; 3; iii! Or, maybe four.

- Lingo: Random variables that are uncorrelated, that is, $C(X_i, Y_i) = 0$, are said to be *orthogonal*

2. *Covariance of linear combinations of r.v.s.* Suppose

$$Z_{1i} = a_1 + b_1X_i + c_1Y_i$$
$$Z_{2i} = a_2 + b_2X_i + c_2Y_i$$

Then:

$$C(Z_{1i}, Z_{2i}) = b_1b_2V(X_i) + c_1c_2V(Y_i) + C(X_i, Y_i)(b_1c_2 + c_1b_2)$$

3. *Variance of sums and differences*

$$V(X_i + Y_i) = V(X_i) + V(Y_i) + 2C(X_i, Y_i)$$
$$V(X_i - Y_i) = V(X_i) + V(Y_i) - 2C(X_i, Y_i)$$

Show this using #2 above, or work out longhand.

- The variance of a sum of uncorrelated r.v.s is the sum of their variances

4. *Correlation measures the extent of linear relationship.*
   (a) If $Y_i = a + bX_i$ for some constants a and b, then $\rho_{XY} = 1$ when $b > 0$ and $\rho_{XY} = -1$ if $b < 0$.
   (b) In general, $-1 \leq \rho_{XY} \leq 1$. Much more on this later.

5. *The Law of Iterated Expectations (*LIE*).*[3] For any r.v.s, Z_i and X_i,

$$E(Z_i) = E[E(Z_i|X_i)]$$

In other words, "a marginal mean is the mean of conditional means." Proof: See pages 31-32 in MHE and Pset 1. Note: Z_i might be a function of other r.v.s, say $Z_i = h(X_i, Y_i)$.

6. Properties of CEF residuals

$$E[(Y_i - E(Y_i|X_i))X_i] = 0$$

Think of $E(Y_i|X_i)$ as a predictor for Y_i using information on X_i (e.g., predict wages using schooling). Prediction error is uncorrelated with the predictor, X_i. In fact, we can say something even stronger:

$$E[(Y_i - E(Y_i|X_i))g(X_i)] = 0,$$

for *any* function, $g(X_i)$. Prove this using the LIE.

[3] Very very important. Gotta be able to LIE in your sleep.

2.4 Bonus Properties (discussed in recitation)

More to know 'bout the CEF

1. Consider using a function of random variable X_i, denoted $g(X_i)$, to predict random variable Y_i. Then, $g(X_i) = E(Y_i|X_i)$ is the minimum MSE predictor of Y_i given X_i (Prove this using #6 above).

2. *Analysis of variance (ANOVA)*

$$\sigma_Y^2 = E[V(Y_i|X_i)] + V[E(Y_i|X_i)] \tag{1}$$

where $V(Y_i|X_i) = E\{(Y_i - E[Y_i|X_i])^2|X_i\}$ is the conditional variance function for Y_i given X_i. Equation (1) is called the analysis of variance (ANOVA) formula.

   - This is interpreted as follows: $V(Y_i|X_i)$ is "within-X_i" variance; i.e. variance in Y_i given X_i, while $V[E(Y_i|X_i)]$ is "between-X_i" variance, i.e., the variance in the CEF of Y_i given X_i (note that because X_i is random, $V(Y_i|X_i)$ and $E(Y_i|X_i)$ are also random). The total variance of Y_i is therefore the sum of (average) within-X_i variance and between-X_i variance.

Bounding probabilities using Chebyshev's Inequality

Pafnuty L. Chebyshev (b. May 16, 1821, Zhukovsky District, Kaluga Oblast, Russia) showed that for any random variable, X_i, and any positive constant, c:

$$P(|X_i - \mu_X| \geq c\sigma_X) \leq {}^1/_{c^2}.$$

In other words, the probability that X_i is more than c standard deviations from its mean is less than ${}^1/_{c^2}$.

- This inequality made PLC popular with his neighbors because it allowed them to bound the probability of extreme events, mostly disasters of various kinds

- We don't use this awesome inequality every day, but it's good to know and a good exercise to show. We will use it later to prove the *Law of Large Numbers*.