Probability distributions

Susan Thomas
IGIDR, Bombay

August 7, 2008

Recap

- Economists problem: creating a quantifiable hypothesis
- Econometricians problem: validating the hypothesis
- Focus of the problem: an economic RV and it's probability distribution.
- Some concepts in probability: PDs, PDFs, CDs, CDFs

Moment generating functions of PDs/PDFs

Expectation of RVs

- E(x)
- For a discrete RV:

$$\mathrm{E}(x) = \sum_{i=1}^{n} x_i \mathrm{Pr}(x_i)$$

  where Pr(x) is the probability distribution of x.
- For a continuous RV:

$$\mathrm{E}(x) = \int_{x=-\infty}^{x=\infty} x f(x) dx$$

  where $f(x)$ is the probability density function of x.
- Example: bernoulli RV.
  x = 0, 1; Pr(0) = p, Pr(1) = 1 - p

$$E(x) = 0 * p + 1 * (1 - p) = 1 - p$$

Testing concepts: expectation of a discrete variable

RV x, can take the following discrete values, each with equal probability:

x	-1	2	5	7	10	11	12	15	20	30

What is E(x)?

Testing concepts: expectation of a discrete variable

With each value taking equal probability, the PD for *x* is:

x	-1	2	5	7	10	11	12	15	20	30
x*Pr(x)	-0.1	0.2	0.5	0.7	1.0	1.1	1.2	1.5	2.0	3.0

$\text{E(x)} = \sum \text{x*Pr(x)} = 11.1$

Example: expectation of continuous variables

- Uniform Continuous RV: x = [L,U]; $\Pr(x_i) = p = 1/(U - L)$

$$\begin{aligned} E(x) &= \int_L^U \frac{x}{U-L}\, d(x) \\ &= \frac{1}{U-L} \int_L^U x\, d(x) = \frac{1}{U-L} \left(\frac{x^2}{2} \right)_L^U \\ &= \frac{U+L}{2} \end{aligned}$$

Example: expectation of continuous variables

- Normal RV: $x = [-\infty, \infty]$; $\Pr(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}(x-\mu/\sigma)^2}$

$$
\begin{aligned}
E(x) &= \int_{-\infty}^{\infty} x\, f(x)\, d(x) \\
&= \frac{e^{1/2\sigma^2}}{\sqrt{2\pi\sigma}} \int_{-\infty}^{\infty} x\, e^{-\frac{1}{2}(x-\mu)^2}\, d(x) \\
\text{Set } y &= x - \mu \\
E(x) &= C \int_{-\infty}^{\infty} y\, e^{-\frac{1}{2}y^2}\, d(y) - \mu C \int_{-\infty}^{\infty} e^{-\frac{1}{2}y^2}\, d(y) \\
&= C \int_{-\infty}^{\infty} y\, e^{-\frac{1}{2}y^2}\, d(y) - \mu \int_{-\infty}^{\infty} f(x)\, d(x)
\end{aligned}
$$

Example: expectation of normal distribution

- First integral on the RHS:

$$\begin{aligned} \int y\, e^{-\frac{1}{2}y^2}\, d(y) &= \left(-e^{-\frac{y^2}{2}}\right)_{-\infty}^{\infty} \\ &= 0 \end{aligned}$$

- Then the expectation becomes:

$$\begin{aligned} E(x) &= 0 + \mu \int_{-\infty}^{\infty} f(x)\, d(x) \\ E(x) &= \mu \end{aligned}$$

Testing concepts: Expectation of a uniform RV

- Uniform Continuous RV: x = [0,10]. What is E(x)?

$$
\begin{aligned}
E(x) &= \int_0^{10} x\, f(x)\, d(x) \\
&= \int_0^{10} x\, \frac{1}{10}\, d(x) \\
&= \frac{1}{10} \int_0^{10} x\, d(x) \\
&= \frac{1}{10} \left(\frac{x^2}{2} \right)_0^{10} \\
&= 5
\end{aligned}
$$

Expectations of functions of discrete RVs

- Function $g(x)$ of a random variable x is a random variable. Then, g(x) has a probability distribution based on Pr(x).
- For any **discrete RV,** x, with a known PD, Pr(x), the expectation of any function $g()$ of x is calculated as:

$$E(g(x)) = \sum_{\text{min}}^{\text{max}} g(x) Pr(x)$$

- For any **continuous RV** y, with a known PDF, f(y), the expectation of any function $g()$ of y is calculated as:

$$E(g(y)) = \int_{\text{min}}^{\text{max}} g(y) f(y) dy$$

Example: E(x^2) for a binary variable

- Bernoulli RV: x = 0, 1, Pr(x) = p, (1-p)
- g(x) = x^2 = 0, 1, Pr(g(x)) = p, (1-p)
- E(g(x)) = E(x^2) = 0*p + 1*(1-p) = (1-p)

Testing concepts: binary variable

RV x is binary with the following probability distribution:

x	Pr(x)	x^2	x^2*Pr(x)
2	0.3	4	1.2
5	0.7	25	17.5

Questions:

1. What is E(x^2)? 18.7

Example: E(x^2) for a continuous uniform variable

- Uniform Continuous RV: x = [L,U]; $f(x_i) = p = 1/(U - L)$
- g(x) = x^2

$$
\begin{aligned}
E(x^2) &= \int_L^U \frac{x^2}{U - L}\, d(x) \\
&= \frac{1}{U - L} \int_L^U x^2\, d(x) = \frac{1}{U - L} \left(\frac{x^3}{3} \right)_L^U \\
&= \frac{U^3 - L^3}{3 * (U - L)}
\end{aligned}
$$

- Example: L=0, U=10; What is E(x^2)?
- Pr(x_i) = 1/10, E(x^2) = 1000/30 = 33.3333

Moment generating functions

- For any distribution, there can be a series of "moments" calculated as follows:

$$\begin{aligned}\textbf{(Discrete rv)} \quad E(x^i) &= \sum_{\min}^{\max} x^i Pr(x) \\ \textbf{(Continuous rv)} \quad E(x^i) &= \int_{-\infty}^{\infty} x\, f(x)\, d(x)\end{aligned}$$

- Each moment describes a feature of the distribution.

The unique **moments** of a distribution

- The moments are functions of the parameters of the distribution.
- Every distribution has as many **unique** moments as parameters.
  The remainder of the moments can be expressed as functions of the parameters.
- For example, the bernoulli distribution had E(x) = E(x^2) = (1-p), the probability of success.
- For example, every moment of the normal distribution can be expressed as a function of the first two moments, the mean (μ) and the variance (σ^2).

Numerical tools to describe a distribution

Statistical measures for a distribution

- Measure of location: Mean, mode, median
- Measure of dispersion: Variance, range, quartiles

Measures of location

- **Mean**: An expected value on a random draw from the dataset.
- **Mode**: The value that occurs with the maximum frequency.
  Easily interpreted for discrete variables.
  The mode for the continuous RV datasets is interpreted in terms of the “range/set” of values that are most often observed.
- **Median**: The value of the RV at which 50% of the dataset is observed.

Examples: Mean, $\bar{x}$

- Find the mean of the data: 5, 1, 6, 2, 4:

$$\bar{x} = \frac{\sum x}{n} = \frac{18}{5} = 3.6$$

Examples: Median

- Find the median of: 1, 7, 3, 1, 4, 5, 3.
- First step is to order the data: 1, 1, 3, 3, 4, 5, 7.
- The median is 3, the midway point, for an odd number of data.
- When the data has an even number of points, the median is calculated as the midpoint between the two choices.
- For a dataset: 9, 5, 7, 3, 1, 8, 4, 6, ordered as 1, 3, 4, 5, 6, 7, 8, 9, the median is 5.5.

Pros and cons of location measures

- Typically, all three measures tend to cluster together - the differences are not very large.
- **However** the mean is most sensitive to the presence of **outliers**.
  (For example, a day on which a trader places a buy limit order for 100 million shares of Reliance instead of a a thousand shares.)
- The median is less sensitive to the mean. It is not influenced by the value of the observations, just their number.
  Thus, it can be a more robust measure of location than the mean.

Measures of dispersion

- Range: The difference between the highest and the lowest value of the RV in the dataset.
  Example: In data, 3, 7, 2, 1, 8 the range = 8 - 1 = 7.
- Variance: The value of the RVs as differences from the average value, squared and summed up. It is denoted by $\sigma(x)^2$.

$$\sigma(x) = \frac{\sum x_i - \bar{x}}{(n-1)}$$

  Example: $\bar{x} = 4.2, \sigma(x)^2 = (-1.2^2 + 2.8^2 + -2.2^2 + -3.2^2 + 3.8^2)/4 = 9.7$.

Calculating the data dispersion using $\bar{x}, \sigma^2$

- Question: what is the range of values of the RV between which we can find 95% of the data?
- Answer:
  1. Upper range value = $\bar{x} + 1.96 * \sigma$
  2. Lower range value = $\bar{x} - 1.96 * \sigma$

Empirical rules

- $\bar{x} \pm \sigma = 85\%$ of the dataset
  The percentage will be larger for more skewed distributions. The percentage will be closer to 70% for distributions that are more symmetric.
- $\bar{x} \pm 2\sigma = 97\%$ of the dataset
- $\bar{x} \pm 3\sigma = 99\%$ of the dataset

Measures of dispersion: Percentiles, Quartiles

- Percentiles: Denoted as p^{th} percentile. The value of RV, x, such that p% of the dataset falls below the value x, and $(100 - p)$% is above.
- Quartiles: A set of three specific percentiles at the $25^{th}, 50^{th}, 75^{th}$ percentiles. They are the lower, median and upper quartile values.
  The median is the 2^{nd} quartile and the 50^{th} percentile.
- Inter-quartile range (IQR): The distance between the lower and the upper quartile values.

Setting up likelihood framework for estimation

Syntax: Population distributions

- Given a population prob. distribution, $P(x)$, the frequency of x in the population is denoted as $f(x)$.
- Given a sample of size n, the frequency of x in the sample is denoted as $\hat{f}(x)$.
- $f(x)$ is a deterministic function of the PD/PDF.
  $\hat{f}(x)$ is a random variable, which is the function of the sample!
- $f(x)$ is always the same for a given x.
  $\hat{f}(x)$ varies depending upon the sample.

Recap on PDFs vs. CDFs

- Theory about distributions focuses on CDFs.

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

  This is well defined, irrespective of the type of rv.
- From the CDFs we can calculated the joint distribution of X, Y as:

$$F(x, y) = P(X \leq x \textbf{ and } Y \leq y)$$

- From the CDF we calculate the marginal distribution function as:

$$F(y) = P(Y \leq y) = P(X \leq \infty \textbf{and} Y \leq y)$$

- If X, Y are independent, the joint distribution function is the product of the marginals:

$$P(X \leq x \text{ and } Y \leq y) = P(X \leq x)P(Y \leq y)$$

Developing a full statistical model

The Bernoulli model

- Question: What is the population frequency of girls among new born children?
- Economists hypothesis: If Y_i is the gender of child i, then:
  1. Y_i are independent across all i.
  2. Y_i come from an *identical* distribution
  3. Y_i are Bernoulli distributed, with $p = \theta$
  4. θ will take values between 0 and 1.

  (Note: Are all these reasonable assumptions?)
- Econometrician's task: Find the correct θ
- Get a data: same dataset as the UK dataset of newborn children where $P(Y = \text{boy}) = 0.513\%$ and $N \sim 715,000$ observations.

The likelihood function approach

- We analyse how **probable** the different outcomes observed are for a given θ.
- Start from scratch: if we know θ, then for a sample of size N, can we write the probability of observing $Y_1 = y_1, Y_2 = y_2, \ldots, Y_N = y_N$?
- Assuming independence (assumption #1), it is:

$$\begin{aligned}
P(Y_1 \leq y_1, \ldots, Y_N \leq y_N) &= P(Y_1 \leq y_1) \ldots P(Y_N \leq y_N) \\
&= \Pi_{i=1}^{i=N} P(Y_i \leq y_i) \\
&= \Pi_{i=1}^{i=N} \theta^{y_i} (1-\theta)^{(1-y_i)} \\
&= \theta^{\sum_1^N y_i} (1-\theta)^{\sum_1^N (1-y_i)} \\
\text{But } \bar{y} &= \frac{1}{N} \sum_1^N y_i \\
P(Y_1 \leq y_1, \ldots, Y_N \leq y_N) &= \theta^{n\bar{y}} (1-\theta)^{n(1-\bar{y})}
\end{aligned}$$

The likelihood function, L

- $P(Y_1 \leq y_1, Y_2 \leq y_2, \ldots, Y_N \leq y_N | \theta) = f_\theta(y_1, y_2, \ldots, y_N)$
  Here, the probability is a function of a known θ. The $(y_1, y_2, \ldots, y_N)$ is a set of outcomes in a sample of size N generated by parameter θ.
- We flip this around to asking: given an N-sized sample, can we use the likelihood of a given value of θ to have generated the observed sample of size N? Or,
- What is the likelihood of θ given the sample:

$$L_{Y_1, Y_2, \ldots, Y_N(\theta)} = f_\theta(Y_1, Y_2, \ldots, Y_N)$$

  Here, $Y_1, \ldots, Y_N$ is a fixed set of random observations in the dataset.
- The likelihood function depends **only** on the observations.

The likelihood function for the Bernoulli model

- Given the sample, we say:

$$L_{(Y_1,\ldots,Y_N)}(\theta) = \theta^{\bar{Y}}(1-\theta)^{(}1-\bar{Y})^{N}$$

- For the Bernoulli model, $L(\theta)$ depends only on $\bar{Y}$
- Therefore, $\bar{Y}$ becomes a sufficient statistic for θ.
- Estimation is about how to find the best value of θ given $\bar{Y}$: Find θ such that L is maximised.
- Which brings us to optimisation theory given a function.