Lectures 1-4 Review

Quantitative Political Science

Prof. Bisbee

Vanderbilt University

Lecture Date: 2023/09/12

Slides Updated: 2023-09-11

Agenda

1. Three discrete probability distributions
2. Two continuous probability distributions
3. Recap of what we know

Theoretical Probability Models

- Three **discrete** examples
  - the Bernoulli
  - the Binomial
  - the Poisson

Bernoulli

- A Bernoulli experiment is the *observation of an experiment consisting of one trial with two outcomes: zero or one*
  - $Y = \{0, 1\}$
  - I.e., coin toss, whether someone approves of Biden's performance, whether a country signs a treaty
- A Bernoulli random variable is characterized by one parameter π: the probability of "success"
- A Bernoulli probability distribution is:
  - $p(y = 1) = \pi$
  - $p(y = 0) = 1 - \pi$
  - Or $p(y) = \pi^y (1 - \pi)^{(1-y)}$
- Practice proof: show that $E(Y) = \pi$ and $VAR(Y) = \pi(1 - \pi)$

The Binomial

- A Binomial experiment is the *observation of an experiment consisting of a sequence of identical and independent Bernoulli trials*
  - Y is the number of successes observed during the n trials
  - I.e., # of heads observed in n coin tosses, # of people approving of Biden's performance out of n people, # of countries signing a treaty out of n eligible countries
- Let's find the Binomial probability distribution!
  - Let our event of interest be $Y = y$ where y is either success or failure (S or F)
  - One event might be $S, S, F, S, F, F, F, S, F, S, \ldots, F, S$
  - Reorder to $S_1, S, S, S, \ldots, S, S_y$ and $F_1, F, F, \ldots, F, F_{n-y}$
  - The number of successes is simply y, and the number of failures is $n - y$

The Binomial contd

- This event can be expressed with set notation as the **intersection** of n simple events: $S_1 \cap S_2 \cap \ldots S_y \cap F_1 \cap F_2 \cap \ldots F_{n-y}$
    - These are **independent** events, meaning $P(S_1 \cap S_2 \cap \ldots S_y \cap F_1 \cap F_2 \cap \ldots F_{n-y}) = P(S_1)P(S_2)\ldots P(S_y)P(F_1)P(F_2)\ldots P(F_{n-y})$
    - This is just $\pi^y(1-\pi)^{n-y}$...same as Bernoulli!
    - BUT! Not probability of $Y = y$ because the event $Y = y$ can happen in many different ways than the above order.
- How many different ways are there to order y S's and $n - y$ F's?
    - Number of different ways we can choose y elements out of a total of n elements
    - $\binom{n}{y}$ or $\frac{n!}{y!(n-y)!}$
- Thus the Binomial probability distribution is $p(y) = \frac{n!}{y!(n-y)!}\pi^y(1-\pi)^{n-y}$

Example

- 9 students in class, 5 males. If I pick 6 at random with replacement, what is the chance I pick the same number of males and females?
    - Call "success" a female: $\pi = \frac{4}{9}$
    - $n = 6$ (number of trials)
    - $y = 3$ (number of successes)
- Thus $p(Y = 3) = \frac{6!}{3!(6-3)!}\left(\frac{4}{9}\right)^3\left(1 - \frac{4}{9}\right)^{6-3} \approx 0.30$
- What if I draw six students at random with replacement, on average how many females will I pick? And how much will this number vary over repeated draws of six?
- Expectations: $E(Y) = n\pi$, $VAR(Y) = n\pi(1 - \pi)$

The Poisson

- A Poisson experiment is *the observation of a count of events that occur in an interval, broadly defined*.
    - An **interval**: a given space, time period, or any other dimension
    - I.e., environmental laws per Congressional session
    - Errors per page
    - Government shutdowns per decade
    - Homeless centers per census tract
- A Poisson can be understood as a Binomial experiment as the number of trials approaches infinity

The Poisson contd

- Split the interval into n subintervals, each so small that at most one event could occur in it
    - Thus, each subinterval can be thought of as a Bernoulli trial: $p(y) = \pi^y(1-\pi)^{1-y}$
    - And n subintervals can be thought of as a Binomial: $p(y) = \frac{n!}{y!(n-y)!}\pi^y(1-\pi)^{n-y}$
- How many subintervals are required? Who knows. But we can make them infinitely small by taking the limit of the Binomial as $n \to \infty$
    - Interested in the number of successes over the interval: $\lambda = n\pi$
    - $\lim_{n\to\infty} \frac{n!}{y!(n-y)!}\pi^y(1-\pi)^{n-y}$
    - $\lim_{n\to\infty} \frac{n!}{y!(n-y)!}\left(\frac{\lambda}{n}\right)^y\left(1-\frac{\lambda}{n}\right)^{n-y}$

The Poisson contd

- Note that, by the definition of e, $\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n = e^{-\lambda}$
    - $\lim_{n\to\infty} \frac{n!}{y!(n-y)!}\left(\frac{\lambda}{n}\right)^y\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-y}$
- Thus:
    - $\lim_{n\to\infty} \frac{n!}{y!(n-y)!}\frac{\lambda^y}{n^y}e^{-\lambda}(1)$
    - $\frac{e^{-\lambda}\lambda^y}{y!}\lim_{n\to\infty} \frac{n!}{(n-y)!n^y}$
    - $\frac{\lambda^y}{y!}e^{-\lambda}\lim_{n\to\infty} \frac{n(n-1)(n-2)\ldots(n-y+1)}{n^y}$
    - $\frac{\lambda^y}{y!}e^{-\lambda}\lim_{n\to\infty} \frac{n}{n}\frac{(n-1)}{n}\frac{n-2}{n}\ldots\frac{n-y+1}{n}$

The Poisson contd

- And finally
  - $\frac{\lambda^y}{y!} e^{-\lambda} \lim_{n \to \infty} 1 \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) \dots \left(1 - \frac{(y+1)}{n}\right)$
  - $\frac{\lambda^y}{y!} e^{-\lambda} (1)$
- For proving:
  - $E(Y) = \lambda$
  - $VAR(Y) = \lambda$

Continuous Random Variables

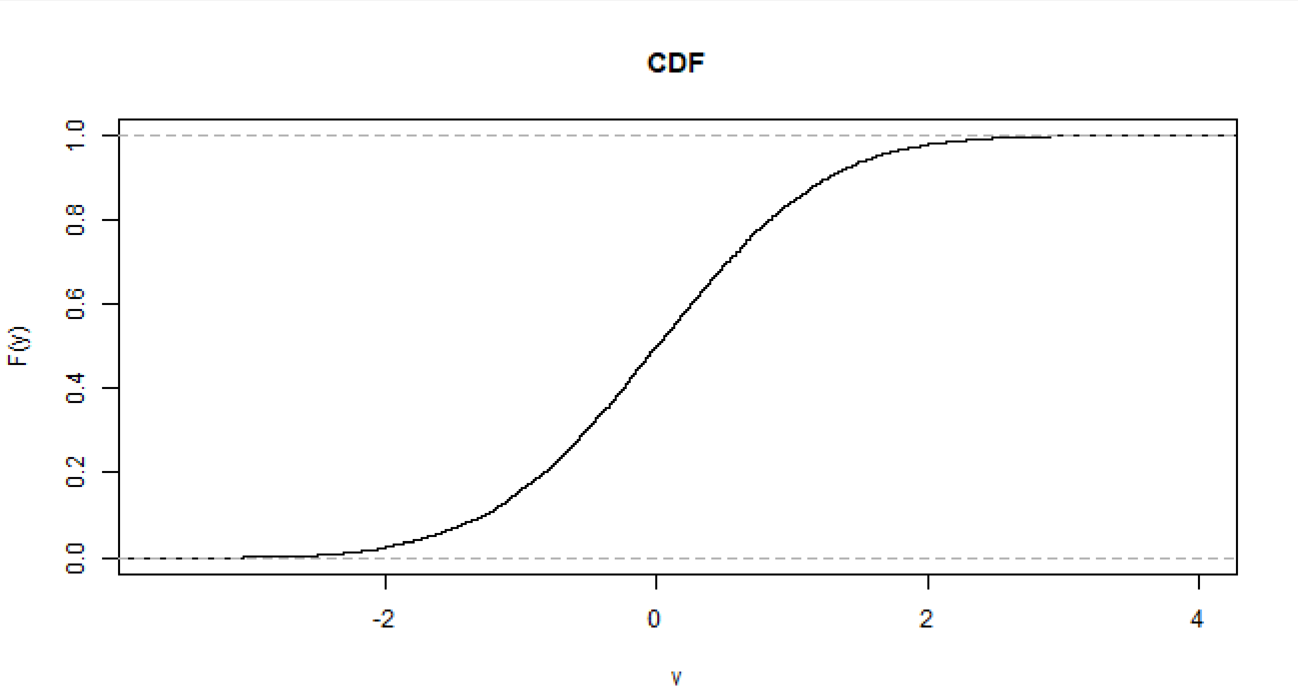
- Often dealing with RVs that take on uncountably infinite values. These are **continuous** random variables.
    - It is impossible to assign nonzero probabilities to all the uncountably infinite points on an interval while satisfying that they sum to 1.
    - Thus the notion of $p(y)$ from the discrete world doesn't work with continuous RVs
- Need a different approach to describing the probability distribution of a continuous RV
    - Define the cumulative distribution function (CDF) as $F(y)$ where $F(y) \equiv P(Y \leq y)$ for $-\infty < y < \infty$

CDFs

- CDFs have the following properties
  - $F(-\infty) \equiv \lim_{y \to -\infty} F(y) = 0$
  - $F(\infty) \equiv \lim_{y \to \infty} F(y) = 1$
  - $y_1 < y_2 \Rightarrow F(y_1) \leq F(y_2)$
- Note that discrete random variables also have CDFs
  - If $F(y)$ is continuous for $-\infty < y < \infty$, then Y is continuous
  - Discrete CDFs are always **step** functions: meaning they have discontinuities separating the possible values of y

CDF

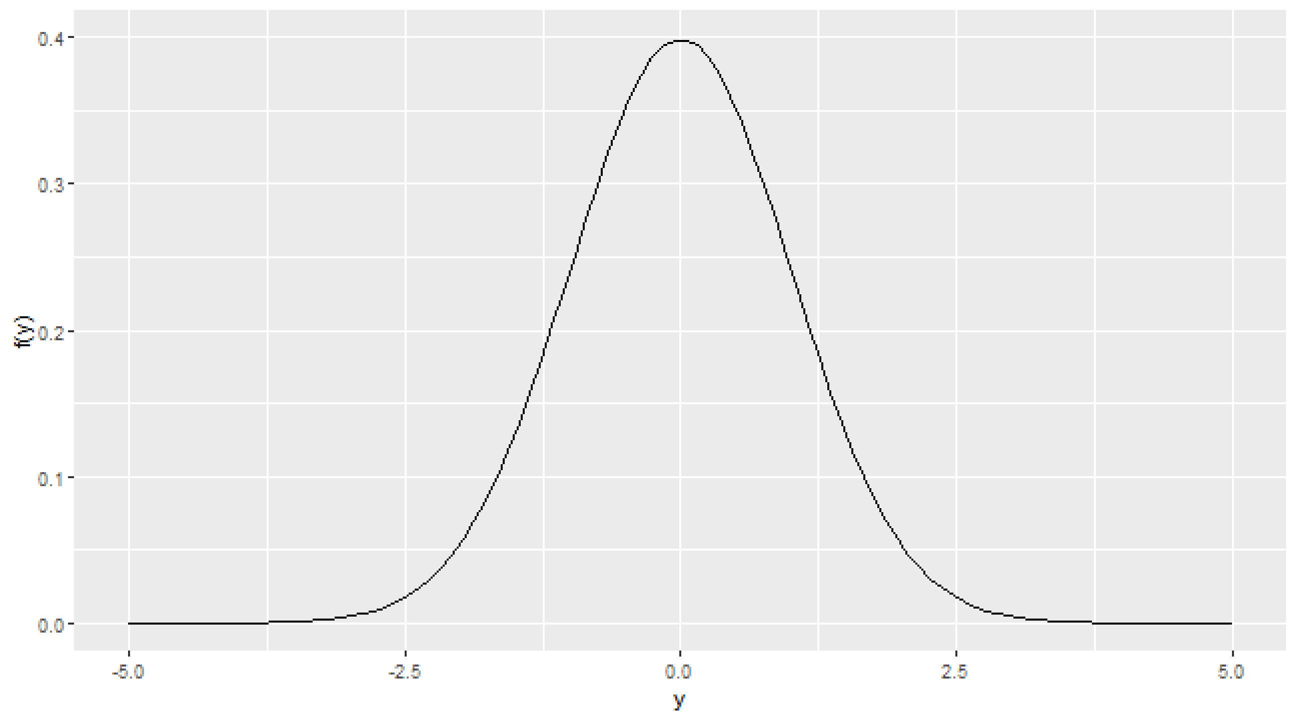
```
# create sample data
sample_Data = rnorm(5000)
# calculate CDF
CDF <- ecdf(sample_Data )
# draw the cdf plot
plot(CDF,main = 'CDF',ylab = 'F(y)',xlab = 'y')
```



Density

- NB: $P(Y = y) = 0 \; \forall \; y$
  - Weird? Imagine calculating the probability of observing the temperature of 50.71351309 degrees F. Now add 10 additional digits to this number.
- Instead, we think about probability for continuous random variables in terms of **density**
- Define $f(y)$ as the derivative of F
  - $f(y) \equiv \frac{dF(y)}{dy} = F'(y)$
- $f(y)$ is the probability density function (PDF)

PDF



PDF and CDF

- Having defined $f(y) \equiv \frac{dF(y)}{dy}$, we can write $F(y) = \int_{-\infty}^{y} f(t)dt$ where t is a placeholder.
- The pdf $f(\cdot)$ has the following properties
  - $f(y) \geq 0 \; \forall \; y, -\infty < y < \infty$
  - $\int_{-\infty}^{\infty} f(y)dy = 1$
- How do we work with probabilities in this setting?
  - What is the probability that Y takes on values y that fall between a and b?
  - $P(a < Y \leq b) = P(Y \leq B) - P(Y \leq a)$
  - $P(a < Y \leq b) = F(b) - F(a)$
  - $P(a < Y \leq b) = \int_{a}^{b} f(y)dy$
- NOTE: $P(a < Y < b) = P(a < Y \leq b) = P(a \leq Y < b) = P(a \leq Y \leq b)$. Why?

Expectations

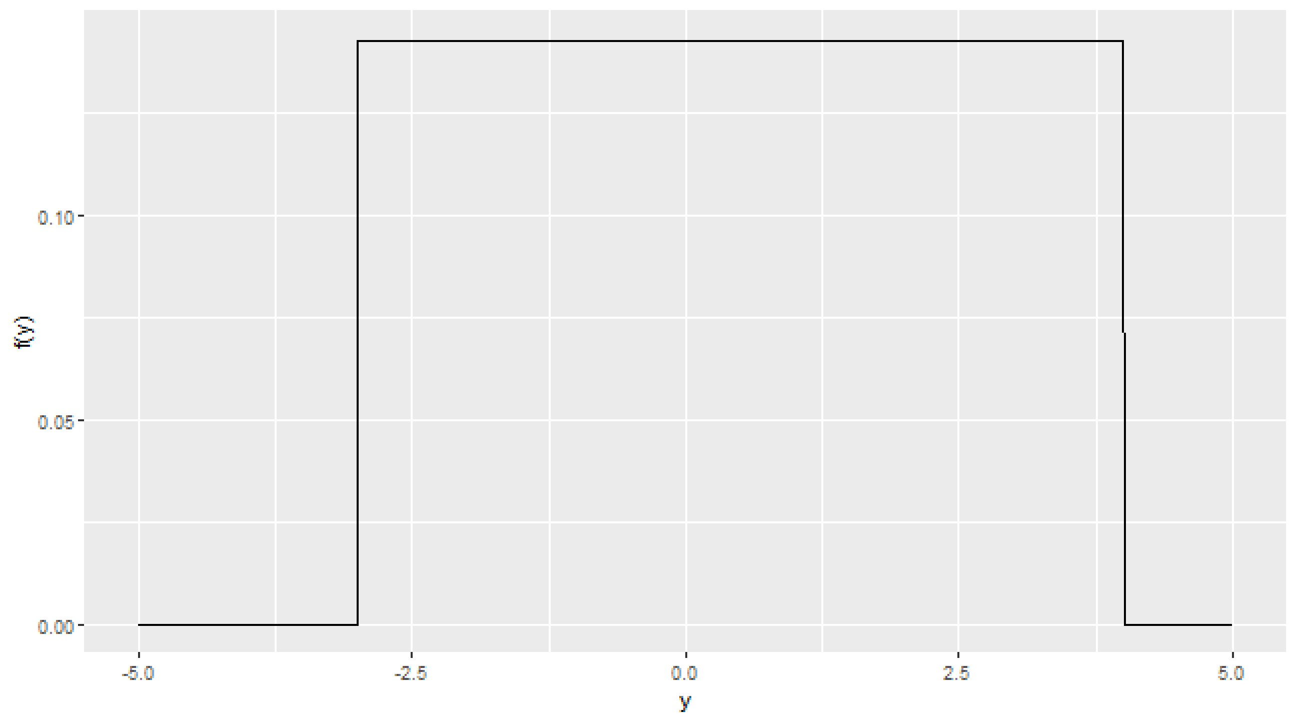
- Recall that the expectation of a discrete random variable is $E(Y) \equiv \sum_y yp(y)$
- For continuous RVs, the intuition is similar
  - $E(Y) \equiv \int_{-\infty}^{\infty} yf(y)dy$
  - $E[g(Y)] = \int_{-\infty}^{\infty} g(y)f(y)dy$
  - $VAR(Y) \equiv \int_{-\infty}^{\infty} (y-\mu)^2 f(y)dy$
- Prove $VAR(Y) = E(Y^2) - \mu^2$

Theoretical models

- We'll look at two commonly used to describe continuous random variables
    - The **uniform**
    - The **Normal**
- And three distributions related to the Normal that we will use constantly in statistical tests
    - The **Chi-squared** (χ^2) distribution
    - The **t-distribution**
    - The **F-distribution**

The Uniform

- A random variable that can take on any value in an interval between two other values, and the chances are equal for every value
- We can visualize the density function like this:



The Uniform

- The pdf is thus:
  - $f(y) = \frac{1}{\theta_2 - \theta_1}$ for $\theta_1 \leq y \leq \theta_2$
  - $f(y) = 0$ otherwise
- Proof? Geometry!
- The CDF can be derived:
  - $F(y) = \int_{-\infty}^{y} f(t)dt$
  - $F(y) = \int_{\theta_1}^{y} \frac{1}{\theta_2 - \theta_1} dt$
  - $F(y) = \left. \frac{t}{\theta_2 - \theta_1} \right|_{\theta_1}^{y}$
  - $F(y) = \frac{y - \theta_1}{\theta_2 - \theta_1}$

The Uniform

- What is $E(Y)$?
- What is $VAR(Y)$?
- What are some examples of uniformly distributed continuous random variables?

The Normal

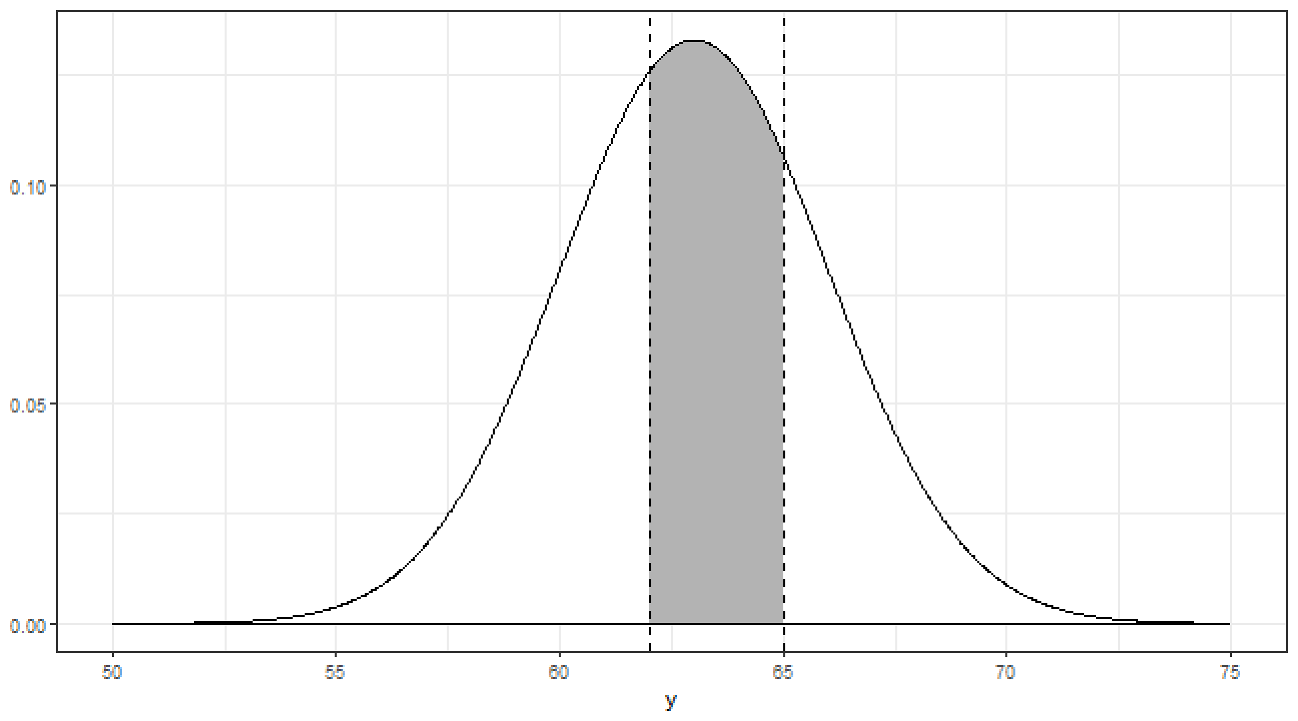
- Many empirical distributions are closely approximated by a distribution that is:
    1. symmetric
    2. has non-zero probability for all possible values of y
    3. is "bell shaped"
- These characteristics are embodied in the **normal distribution**

The Normal

- We won't get into the math of the normal, but it is an essential part of quantitative analysis!
- SO just trust me:
  - $f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(y-\mu)^2}{2\sigma^2}}$ for $-\infty < y < \infty$
- Two parameters: μ and σ
  - $E(Y) = \mu$
  - $VAR(Y) = \sigma^2$

The Normal

- What is the probability Y takes on some value y within an interval between $a = 62$ and $b = 65$?
  - $P(a \leq Y \leq b) = \int_a^b f(y)dy = \int_{62}^{65} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(y-\mu)^2}{2\sigma^2}} dy$



The Normal

- We typically **standardize** a normally distributed variable
  - Units measured in terms of standard deviations (instead of inches or whatever else)
- $Z \equiv \frac{Y-\mu}{\sigma}$
  - Z is a random variable with mean zero and standard deviation one
  - PDF simplifies to $f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$
- We use these so frequently in statistics we denote them with special symbols!
  - "Little phi of z" is the PDF of the standardized normal evaluated at $Z = z$: $\phi(z)$
  - "Big phi of z" is the CDF of the standardized normal evaluated at $Z = z$: $\Phi(z)$

Three Associated Distributions

- We use the normal a **ton**
- But we also use it with three other distributions
  1. The **Chi-squared** (χ^2): Y is the sum of squares of a series of standard normal RVs
  2. The **t-distribution**: Y is the ratio of the standard normal RV / the square root of the chi-squared RV
  3. The **F distribution**: Y is the ratio of two chi-squared RVs
- We will return to these later, but I'm signposting them here

What do we now know?

- **Definitions**:
  - Research questions and statistics
  - Units and variables
  - Summarizing data
- **Probability**:
  - Experiments, observations and events
  - Set theory
  - Event composition method (4 tools)
- **Random Variables**:
  - Summarizing theoretical distributions
  - Expectations