Quantitative Social Science Methods, I, Lecture Notes: Introduction

Gary King[1]
Institute for Quantitative Social Science
Harvard University

August 17, 2020

[1]GaryKing.org

Overview and Logistics

Statistical Models

Data Generation Processes (with Simulation)

Probability as a Model of the Data Generation Process

What is the Field of Statistical Analysis?

- An Old Field (circa 1662) Tied to Government and Politics
  - "*state*-istics"
  - Reformers getting the goods on the state
  - The state hunting for tax revenue from the people
  - Everyone collecting info like mortality data
- Also a New field: with fast progress
  - Mid-1930s: Experiments and random assignment
  - 1950s: The modern theory of inference
  - In your lifetime: Modern causal inference
  - Since you started school: A monumental societal change: "big data", the march of quantification through academic, professional, commercial, government and policy fields.
- The number of new methods: increasing fast
- Most important methods: originate *outside* the discipline of statistics, near the applications, the substance, the problem to be solved (random assignment, experimental design, survey research, machine learning, MCMC methods, ...).
- Statistics: abstracts, proves formal properties, generalizes, and distributes results. The core, but a very small core.

What's the Subfield of *Political Methodology*

- Political science's methodological subfield
- One of the largest APSA sections
- Akin to: econometrics within economics, psychometrics in psychology, sociological methodology within sociology, biostatistics in public health and medicine, ...
- Union of Methods Subfields: known as data science, applied statistics, or computational social science
- Part of a massive change in the evidence base of the social sciences, from: (a) surveys, (b) end of period government stats, and (c) one-off studies of people, places, or events $\rightsquigarrow$ numerous new types and huge quantities of (big) data
  - Previously: quantitative dissertations: download from ICPSR
  - Now: most dissertations: collect your own data, learn or develop methods to suit
- Political science: an unusually diverse discipline, cross-roads of other fields, great place for a broad perspective on methods

What's *This* Course in Political Methodology About?

1. Understanding statistical inference
   - "Using facts you know to learn about facts you don't know"
   - Understand statistical theory underlying specific methods
   - Feel comfortable using these methods in your own work
   - Learn, implement, apply, interpret, and explain methods invented after class ends
   - The practical tools of research: theory, applications, simulation, programming — whatever is useful
2. Making novel contributions to a scholarly literature.
   - Sounds hard, but almost everyone gets there
   - Revised versions of class papers (by grads and undergrads) $\rightsquigarrow$ articles, books, senior theses, dissertations, blog posts, conference presentations, job talks, many awards
3. Learning how to choose a topic for research, solve hard problems by *changing the question*, and identifying the big idea
   - Biggest factor impacting success: the question
   - Nursery school until today: questions are fixed, immutable
   - Today: you can change or choose the question!

Who takes this course?

- Gov2001
  - Most Gov Dept grad students doing empirical work
  - Grad students from other Harvard departments & schools
- Gov1002 Undergrads
- StatE-200 Non-Harvard students, visitors, faculty, & others (online through the Harvard Extension school)

Some of the best experiences here: getting to know people in other fields

Requirements

1. Engage in the collective experience
   - Many assignments: carefully choreographed dances
   - Other students depend on you, and you on them
   - Come to class, meet all deadlines, support your classmates
   - We learn better collectively
2. Weekly assignments
   - Readings & Lectures in Perusall
   - Annotate, help others, read slowly
   - Problem sets & assessment questions: try it yourself, work in groups of 3-4, write it up yourself
   - All questions sent via Perusall or in class
   - Peer grading and peer instruction ("teaching teaches the teacher")
3. One "publishable" coauthored paper (Easier than you think!)
   - Draft submission and replication exercise helps a lot.
   - See "Publication, Publication"
   - You won't be alone: you'll work with each other and us
   - Coauthored, peer grading, peer instruction

Course Information

- All class materials: j.mp/G2001
- Find Study Group of 3–6, and Coauthor Group of 2–3
- When are Gary's office hours?
  - (Big secret: Office hours prevent visits at other times)
  - Come when you like; if you can't find me or I'm in a meeting, come back, talk to my assistant in the office next to me, or send chat via Perusall
- Focus, like I will, on learning (not grades): Especially when we work on papers, I will treat you like a colleague, not a student

Course Strategy

- We could teach you:
  - the latest and greatest methods, but when you graduate *they* will be old
  - all the methods that might prove useful for you, but when you graduate *you* will be old
- Instead, we teach the fundamentals, the underlying theory of inference, from which statistical models are developed
- Then you can learn new methods invented after class
- Balance of emphasis: proofs v. concepts & intuition
  - Baby Stats: dumbed down proofs, vague intuition
  - Math Stats: rigorous mathematical proofs, no data experience
  - Practical Stats: math when needed, deep concepts and intuition
    - Ultimate Goal: how to do empirical research, in depth
    - Procedure: Traverse from theoretical foundations to practical application (including "how to" computations)
- Do you have the background? (1) data analysis experience, (2) R, (3) some probability (particularly distributions). See the class website.

The Big Picture: Inference

Just because it's impossible, doesn't mean you shouldn't do it!

- Choose: substantive question of interest
- Formalize: quantity of interest (QOI), given question
- Collect: data, given QOI, question
- Assume: class of models, given data, QOI, question
- Estimate from data: best model in class, given of all the above
- Present results: given all of the above
  - QOI Estimates: interpretable by anyone
  - Uncertainty estimates: CIs, SEs, etc.
- Open questions we'll answer:
  - Better ways of presenting results
  - Model dependence
  - Model class dependence
  - Data problems
  - More interesting QOIs
  - More impactful questions to avoid problems in the first place

[Change of subject] What is this?

- Now you know what a model is: an abstraction
- Is this model true or false? Crazy question!
- Are models realistic or not? OK for features we care about
- Are models useful or not? Always a good question
- E.g.: Models of dirt on airplanes v. aerodynamics

Overview and Logistics

Statistical Models

Data Generation Processes (with Simulation)

Probability as a Model of the Data Generation Process

Statistical Models: Notation

- Dependent (or "outcome") variable
  - Y is $n \times 1$.
  - y_i, a number (after we know it)
  - Y_i, a random variable (before we know it)
  - Commonly misunderstood: a "dependent variable" can be
    - a column of numbers in your data set
    - the random variable *for each unit i*.
- Explanatory variables
  - aka "covariates," "independent," or "exogenous" variables
  - $X = \{x_{ij}\}$ is $n \times k$ (observations by variables)
  - A set of columns (variables): $X = \{x_1 \ldots, x_k\}$
  - Row (observation) i: $\{x_{i1}, \ldots, x_{ik}\}$
  - X is fixed (not random)

Equivalent Linear Regression Notation

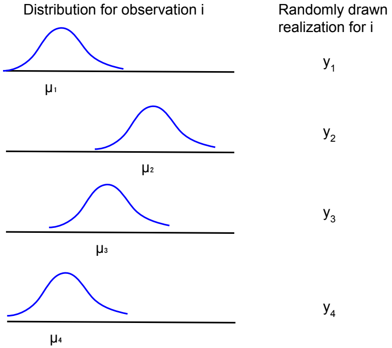
- Standard Notation

$$
\begin{aligned}
Y_i &= x_i\beta + \epsilon_i && = \text{systematic} + \text{stochastic} \\
\epsilon_i &\sim N(0, \sigma^2)
\end{aligned}
$$

- Alternative Notation

$$
\begin{aligned}
Y_i &\sim N(\mu_i, \sigma^2) && \text{stochastic} \\
\mu_i &= x_i\beta && \text{systematic}
\end{aligned}
$$

Understanding Alternative Regression Notation



where $\mu_i = x_i\beta$.

Quiz: Is a histogram of y a test of normality?

Generalized Alternative Notation for Most Models

$$
\begin{aligned}
Y_i &\sim f(\theta_i, \alpha) && \text{stochastic} \\
\theta_i &= g(x_i, \beta) && \text{systematic}
\end{aligned}
$$

where

Y_i random outcome variable

$f(\cdot)$ probability density

θ_i a systematic feature of the density that varies over i

α ancillary parameter (feature of f constant over i)

$g(\cdot)$ functional form

x_i explanatory variable vector for i

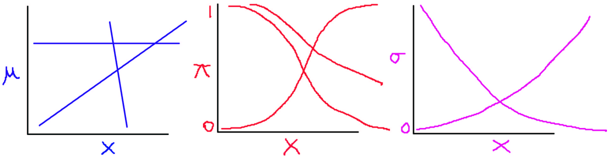
β effect parameters

Forms of Uncertainty

$$\begin{aligned} Y_i &\sim f(\theta_i, \alpha) && \text{stochastic} \\ \theta_i &= g(X_i, \beta) && \text{systematic} \end{aligned}$$

- Estimation uncertainty: Lack of knowledge of β and α. Vanishes as n gets larger.
- Fundamental uncertainty: Represented by the stochastic component (Exists no matter what you do, or size of n)
- Model dependence: Maybe the whole specification is wrong?
- Quiz: If you know the model, including α and β, is $R^2 = 1$?

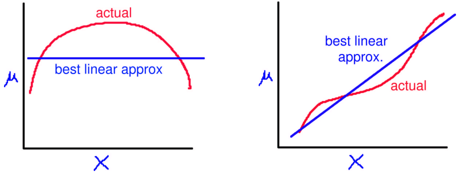
Systematic Components: Examples



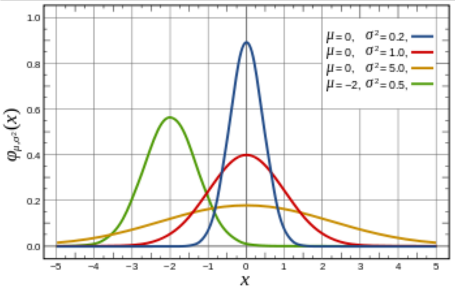
- $E(Y_i) \equiv \mu_i = x_i\beta = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}$
- $\Pr(Y_i = 1) \equiv \pi_i = \frac{1}{1+e^{-x_i\beta}}$
- $V(Y_i) \equiv \sigma_i^2 = e^{x_i\beta}$
- Interpretation:
  - Each is a class of functional forms
  - β in each is an "effect parameter" vector, with a different meaning
  - Set β and it picks out one member of the class

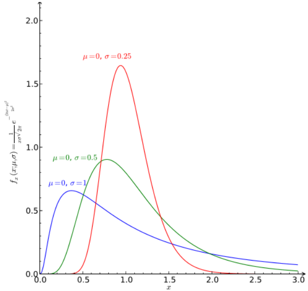
Systematic Components: Examples

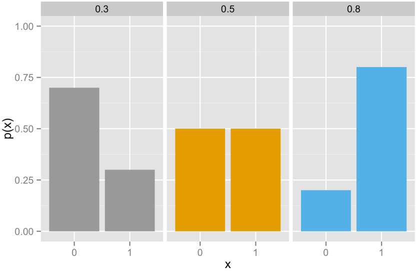
- Standard procedure:
  - Use theory: Assume a class of functional forms (that delineate all *remaining* possible relationships, indexed by parameters)
  - Use data: Estimate the values of the parameters (to pick out one member of the class)
  - Remain uncertain: Because of (1) estimation uncertainty, (2) fundamental uncertainty, (3) model dependence
- If we choose the wrong family of functional forms, we:
  - Have specification error, and potentially bias
  - Still get the best [linear,logit,etc] approximation to the correct functional form.
  - May be close or far from the truth:

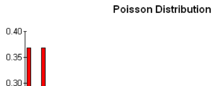


Overview of Stochastic Components









Choosing systematic and stochastic components

- If one is bounded, so is the other
- If the stochastic component is bounded, the systematic component must be globally nonlinear (although possibly locally linear)
- All modeling decisions: about the data generation process
    - The chain of evidence from the world to our observation of it
    - The first question you ask of every empirical paper!
- What if we don't know the DGP (& we usually don't)?
    - The problem: model dependence
    - Our first approach: make "reasonable" assumptions and check fit (& other observable implications of the assumptions)
    - Later:
        - Avoid it: relax assumptions (functional form, distribution, etc)
        - Detect remaining model dependence
        - Remove model dependence: preprocess data (via matching, etc.)

Overview and Logistics

Statistical Models

Data Generation Processes (with Simulation)

Probability as a Model of the Data Generation Process

What's the probability that these goats are real?

What's the probability that these goats are real?

[Hey! What are you looking at?]

- How I can convince you the goats are real? The DGP!
  - Here's one: I took the photos (Should you trust me?)
  - On a family vacation (The story helps some)
  - Here's some evidence (How good is Gary at photoshop?)
  - Would it help you if I told you why the goats are in the trees?
- Key points:
  - DGP Uncertainty $\rightsquigarrow$ Substantive Uncertainty
  - For data to be useful, you need the DGP
  - Theory is helpful for data; data is helpful for theory
- How will we use DGPs?
  - Probability: Assumes the DGP
  - Statistical Inference: Learns the DGP

Simulation is used to:

1. Understand the assumed DGP
2. Solve probability problems
3. Evaluate estimators
4. Calculate features of probability densities
5. Transform statistical results into quantities of interest
6. Get the right answer: easier than mathematical calculations

What is simulation?

Survey Sampling	Simulation
1. Learn about a population by taking a random sample from it	1. Learn about a distribution by taking random draws from it
2. Use the random sample to estimate a feature of the population	2. Use the random draws to approximate a feature of the distribution
3. The estimate is arbitrarily precise for large n	3. The approximation is arbitrarily precise for large M
4. Example: estimate the mean of the population	4. Example: Approximate the mean of the distribution

Monte Hall's Let's Make a Deal

Choose 1 of 3 doors (two with goats; one with a car). Monte peeks at the 2 unchosen doors and opens the one (or one of the two) with the goat and asks if you'll switch to the remaining door? Should you switch?

```
sims<-1000; WinNoSwitch<-0; WinSwitch<-0; doors<-c(1,2,3)
for (i in 1:sims) {
  WinDoor <- sample(doors, 1)
  choice <- sample(doors, 1)
  if (WinDoor == choice)                    # no switch
     WinNoSwitch <- WinNoSwitch + 1
  doorsLeft <- doors[doors != choice]       # switch
  if (any(doorsLeft == WinDoor))
     WinSwitch <- WinSwitch + 1
}
cat("Prob(Car | no switch)=", WinNoSwitch/sims, "\\n")
cat("Prob(Car | switch)=", WinSwitch/sims, "\\n")
```

Monte Hall's Let's Make a Deal

Pr(car|No Switch)	Pr(car|Switch)
.324	.676
.345	.655
.320	.680
.327	.673

The Birthday Problem

Given a room with 24 randomly selected people, what is the probability that at least two have the same birthday?

```
sims <- 1000
people <- 24
alldays <- seq(1, 365, 1)
sameday <- 0
for (i in 1:sims) {
   room <- sample(alldays, people, replace = TRUE)
   if (length(unique(room)) < people) \# same birthday
   sameday <- sameday+1
 }

 cat("Probability >=2:", sameday/sims, "\\n")
```

Four runs: 0.538, 0.550, 0.547, 0.524

Overview and Logistics

Statistical Models

Data Generation Processes (with Simulation)

Probability as a Model of the Data Generation Process

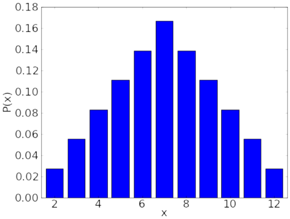
Probability

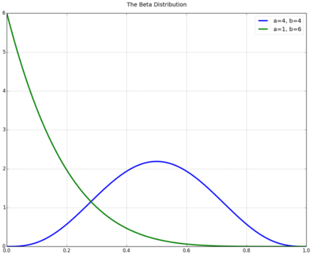
- A function $\Pr(y|M) \equiv \Pr(\text{data}|\text{Model})$, where $M = (f, g, X, \beta, \alpha)$.
- for simplicity: $\Pr(y|M) \equiv \Pr(y)$
- 3 axioms define the function $\Pr(\cdot)$
  1. $\Pr(z) \geq 0$ for some event z
  2. $\Pr(\text{sample space}) = 1$
  3. If $z_1, \dots, z_k$ are mutually exclusive events,

$$\Pr(z_1 \cup \cdots \cup z_k) = \Pr(z_1) + \cdots + \Pr(z_k),$$

- 1& 2 imply: $0 \leq \Pr(z) \leq 1$
- Axioms are not assumptions; they can't be wrong.
- From the axioms come all rules of probability theory.
- Quiz: what happens if $\Pr(\text{sample space}) = 2$
- Rules can be applied analytically or via simulation.

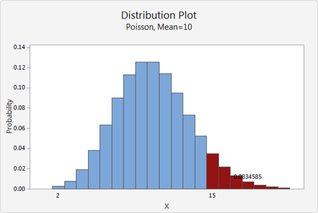
PDFs: Probability Density Functions

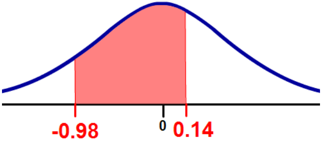




- defined for any y (outcome of the experiment)
- assigns probability to every possible y (or range of y)
- a function, $P(y)$ or $f(y)$, such that
  - $\mathrm{P}(y) \geq 0$ for any y
  - for discrete y: $\sum_{\text{all } y} \mathrm{P}(y) = 1$
  - for continuous y: $\int_{-\infty}^{\infty} f(y)dy = 1$
- Quiz: Are the curves above PDFs?

Computing Probabilities from PDFs



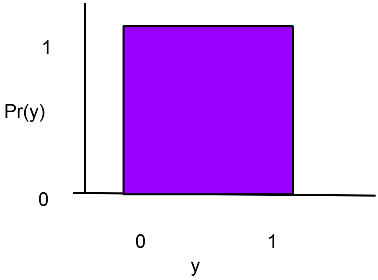


- $$\Pr(a \leq Y \leq b) = \begin{cases} \sum_{a \leq y \leq b} \mathrm{P}(y)dy & \text{discrete} \\ \int_a^b \mathrm{P}(y)dy & \text{continuous} \end{cases}$$
- $$\Pr(Y = y) = \begin{cases} \mathrm{P}(y) & \text{discrete} \\ 0 & \text{continuous} \end{cases}$$
- Quiz: why?

What you should know about every pdf

- The assignment of a probability or probability density to every conceivable value of Y_i
- The first principles
- How to use the final expression (but not necessarily the full derivation)
- How to simulate from the density
- How to compute features of the density such as its "moments"
- How to verify that the final expression is indeed a proper density

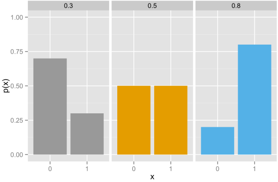
Uniform Density on the interval $[0, 1]$



First Principles about the process that generates Y_i is such that

- Y_i always falls in the "unit" interval: $\int_0^1 \mathrm{P}(y)dy = 1$
- $\Pr(Y \in (a, b)) = \Pr(Y \in (c, d))$ if $a < b$, $c < d$, and $b - a = d - c$.
- Quiz: How do you know it's a pdf?
- Quiz 2: How to simulate? runif(1000)
- Quiz 3: This PDF has no parameters. Could we add some?

Bernoulli pdf (or pmf)



- First principles about the process that generates Y_i:
  - Y_i has 2 mutually exclusive outcomes; and
  - The 2 outcomes are exhaustive
- Quiz: What's an example that violates these rules
- In this simple case, we'll compute features analytically and by simulation.
- Mathematical expression for the pmf
  - $\Pr(Y_i = 1|\pi_i) = \pi_i, \quad \Pr(Y_i = 0|\pi_i) = 1 - \pi_i$
  - The parameter π happens to be interpretable as a probability
  - $\implies \Pr(Y_i = y|\pi_i) = \pi_i^y(1 - \pi_i)^{1-y}$
  - Alternative notation: $\Pr(Y_i = y|\pi_i) = \text{Bernoulli}(y|\pi_i) = f_b(y|\pi_i)$

Features of the Bernoulli: analytically

- Expected value:

$$
\begin{aligned}
E(Y) &= \sum_{\text{all } y} y \mathrm{P}(y) \\
&= 0\,\Pr(0) + 1\,\Pr(1) \\
&= \pi
\end{aligned}
$$

- Variance:

$$
\begin{aligned}
V(Y) &= E[(Y - E(Y))^2] && \text{(The definition)} \\
&= E(Y^2) - E(Y)^2 && \text{(An easier version)} \\
&= E(Y^2) - \pi^2
\end{aligned}
$$

- How do we compute $E(Y^2)$?

Expected values of functions of random variables

$$E[g(Y)] = \sum_{\text{all } y} g(y)P(y)$$

or

$$E[g(Y)] = \int_{-\infty}^{\infty} g(y)P(y)dy$$

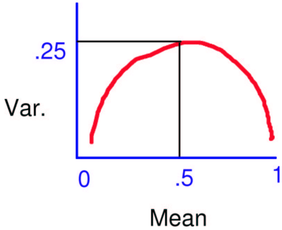
For example,

$$\begin{aligned} E(Y^2) &= \sum_{\text{all } y} y^2 P(y) \\ &= 0^2 \Pr(0) + 1^2 \Pr(1) \\ &= \pi \end{aligned}$$

Variance of the Bernoulli (uses above results)

$$
\begin{aligned}
V(Y) &= E[(Y - E(Y))^2] && \text{(The definition)} \\
&= E(Y^2) - E(Y)^2 && \text{(An easier version)} \\
&= \pi - \pi^2 \\
&= \pi(1 - \pi)
\end{aligned}
$$

This makes sense:



How to Simulate from the Bernoulli with parameter π

- Take one draw u from a uniform density on the interval $[0,1]$
- Set π to a particular value
- Set $y = 1$ if $u < \pi$ and $y = 0$ otherwise
- In R:

```
sims <- 1000                    # set parameters
bernpi <- 0.2
u <- runif(sims)                # uniform sims
y <- as.integer(u < bernpi)
y                               # print results
```

- Running the program gives:

```
0 0 0 1 0 0 1 1 0 0 1 1 1 0 ...
```

- Quiz: What can we do with the simulations?

Binomial Distribution

First principles:

- N iid Bernoulli trials, $y_1, \ldots, y_N$
- The trials are independent
- The trials are identically distributed
- We observe $Y = \sum_{i=1}^{N} y_i$

Density:

$$\mathrm{P}(Y = y|\pi) = \binom{N}{y}\pi^{y}(1-\pi)^{N-y}$$

Explanation:

- $\binom{N}{y}$ because (1 0 1) and (1 1 0) are both $y = 2$.
- π^y because y successes with π probability each (product taken due to independence)
- $(1-\pi)^{N-y}$ because $N - y$ failures with $1 - \pi$ probability each
- Moments: Mean $E(Y) = N\pi$; Variance $V(Y) = \pi(1-\pi)/N$

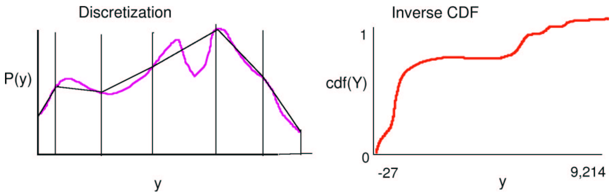
How to simulate from the Binomial distribution

- To simulate from the Binomial(π; N):
  - Simulate N independent Bernoulli variables, $Y_1, \ldots, Y_N$, each with parameter π
  - Add them up: $\sum_{i=1}^{N} Y_i$
- What can you do with the simulations?

Where to get uniform random numbers

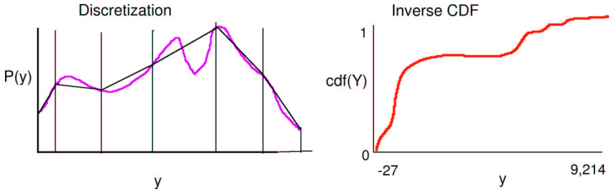
- Random is not haphazard (e.g., Benford's law)
- Random number generators are perfectly predictable (what?)
- We use pseudo-random numbers which have (a) digits that occur with 1/10th probability, (b) no time series patterns, etc.
- How to create real random numbers?

Discretization for random draws from discrete pmfs



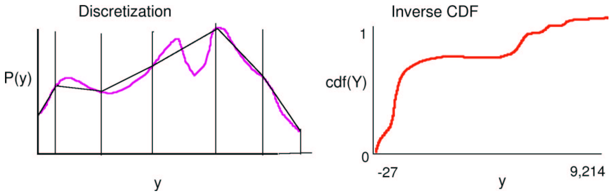
- Divide up PDF into a grid
- Approximate probabilities by trapezoids
- Map [0,1] uniform draws to the proportion area in each trapezoid
- Return midpoint of each trapezoid
- More trapezoids $\rightsquigarrow$ better approximation
- (Works for a few dimensions, but infeasible for many)

Inverse CDF: drawing from arbitrary continuous pdfs



- From the pdf $f(Y)$, compute the cdf:
  $\Pr(Y \le y) \equiv F(y) = \int_{-\infty}^{y} f(z)dz$
- Define the inverse cdf $F^{-1}(y)$, such that $F^{-1}[F(y)] = y$
- Draw random uniform number, U
- Then $F^{-1}(U)$ gives a random draw from $f(Y)$.

Using Inverse CDF to Improve Discretization Method



- Refined Discretization Method:
    - Choose interval randomly as above (based on area in trapezoids)
    - Draw a number within each trapezoid by the inverse CDF method applied to the trapezoidal approximation.
- Drawing random numbers from arbitrary multivariate densities: now an enormous literature

Normal Distribution

- Many different first principles
- A common one is the central limit theorem
- The univariate normal density (with mean μ_i, variance σ^2)

$$N(y_i|\mu_i, \sigma^2) = (2\pi\sigma^2)^{-1/2} \exp\left(\frac{-(y_i - \mu_i)^2}{2\sigma^2}\right)$$

- The stylized normal: $f_{stn}(y_i|\mu_i) = N(y|\mu_i, 1)$

$$f_{stn}(y|\mu_i) = (2\pi)^{-1/2} \exp\left(\frac{-(y_i - \mu_i)^2}{2}\right)$$

- The standardized normal: $f_{sn}(y_i) = N(y_i|0, 1) = \phi(y_i)$

$$f_{sn}(y_i) = (2\pi)^{-1/2} \exp\left(\frac{-y_i^2}{2}\right)$$

Reminder: Equivalent Regression Notation

- Standard version

$$\begin{aligned} Y_i &= x_i\beta + \epsilon_i &&= \text{systematic} + \text{stochastic} \\ \epsilon_i &\sim f_N(0, \sigma^2) \end{aligned}$$

- Alternative version

$$\begin{aligned} Y_i &\sim f_N(\mu_i, \sigma^2) && \text{stochastic} \\ \mu_i &= x_i\beta && \text{systematic} \end{aligned}$$

- Generalized version

$$\begin{aligned} Y_i &\sim f(\theta_i, \alpha) && \text{stochastic} \\ \theta_i &= g(x_i, \beta) && \text{systematic} \end{aligned}$$

Multivariate Normal Distribution

- Let $Y_i \equiv \{Y_{1i}, \ldots, Y_{ki}\}$ be a $k \times 1$ vector, jointly random:

$$Y_i \sim N(y_i|\mu_i, \Sigma)$$

where μ_i is $k \times 1$ and Σ is $k \times k$. For $k = 2$,

$$\mu_i = \begin{pmatrix} \mu_{1i} \\ \mu_{2i} \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$$

- Mathematical form:

$$N(y_i|\mu_i, \Sigma) = (2\pi)^{-k/2}|\Sigma|^{-1/2} \exp\left[-\frac{1}{2}(y_i - \mu_i)'\Sigma^{-1}(y_i - \mu_i)\right]$$

- Simulating *once* from this density produces k numbers. Special algorithms are used to generate normal random variates (in R, `mvrnorm()`, from the MASS library).

Multivariate Normal Distribution

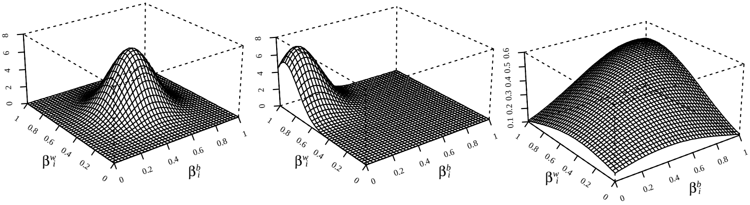
- Moments:
  - $E(Y) = \mu_i$
  - $V(Y) = \Sigma$
  - $\text{Cov}(Y_1, Y_2) = \sigma_{12} = \sigma_{21}$
- Correlation (standardized covariance):

$$\text{Corr}(Y_1, Y_2) = \frac{\sigma_{12}}{\sigma_1 \sigma_2}$$

- Marginals:

$$N(Y_1|\mu_1, \sigma_1^2) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} N(y_i|\mu_i, \Sigma) dy_2 dy_3 \cdots dy_k$$

Truncated bivariate normal examples (for β^b and β^w)



(a) 0.5 0.5 0.15 0.15 0 (b) 0.1 0.9 0.15 0.15 0 (c) 0.8 0.8 0.6 0.6 0.5

Parameters are μ_1, μ_2, σ_1, σ_2, and ρ.