Week 2: Random Variables
Lecture Outline

Random Variables:

Random Variables Notation/Symbols Cheat Sheet	
Symbol	**Meaning**
$X(s)$	X is a random variable function which assigns a numerical value to a specific outcome s *e.g.* $s_1 = \{Tails\}, s_2 = \{Heads\}$ $X(s_1) = 0, X(s_2) = 1$
$\in$	"is an element of"; "is one possible outcome within the range of"
S	The range of possible outcomes resulting from an experiment *e.g. "$S = \{H, T\}$"* means that in our experiment (which flips a coin) we can produce outcomes of either heads or tails
$s \in S$	s represents just one specific, possible outcome which could occur within the range of possible outcomes represented by S *e.g. "$s = \{H\} \in S$"* means getting heads is one possible outcome (s) of the two possible outcomes (S) from flipping a coin

1. **Random Variables (r.v.)**- The single number value associated w/ a specific outcome (realization) of one random process which can realize a different number value if the random process produces a different outcome when repeated
    a. In English: A variable that takes on specific number values depending on which outcome happens
    b. *E.g.* Tails = 0, Heads = 1
        i. $S = \{s_1 = T, s_2 = H\}$
            1. $s_1 =(\{T\} \in S), s_2 =(\{T\} \in S)$
        ii. $X(s_1) = 0, X(s_2) = 1$
2. **Discrete Random Variable**- Random variables which can take on a *finite* number of distinct values
    a. *E.g.* Dice rolls can only result in 6 possible outcomes
3. **Continuous Random Variable**- Random variables which can take on an *infinite* number of distinct values
    a. *E.g.* The exact crop yield of a given year can result in an infinite number depending on decimal precision
    b. *NOTE*: A random variable can still be discrete despite being constructed on a continuous sample

        i. *E.g.* A random variable could be assigned to whether a crop yield is greater than or less than a certain value (only 2 outcomes)

Distribution Functions:

Random Variables Notation/Symbols Cheat Sheet	
Symbol	**Meaning**
$p(X = xi)$ *OR* $p(xi)$	The probability distribution function (pdf) which gives the probability of realizing outcome x_i from all possible outcomes in r.v. X *e.g.* "$p(3)$" is shorthand for "$p(X = 3) = p(S_j \in S : X(S_j) = 3)$"
$f(x)$	The pdf for continuous variables (described by "likelihood", not probability)
$P(x)$	The cumulative distribution function (cdf) describing the probability of realizing outcomes with values less than or equal to x *e.g.* "$P(x)$" is shorthand for "$P(X \leq x_i) = P(Sj \in S : X(Sj) \leq x_i)$"
$F(x) = \int_{lb}^{ub} f(x)dx$	The cdf ($F(x)$) of a continuous random variable (X) is the integral from a lower bound (lb) to an upper bound (ub) of a continuous pdf ($f(x)$)
$\frac{dF(x)}{dx} = f(x)$	The pdf ($f(x)$) of a continuous random variable (X) is the derivative of a continuous cdf ($F(x)$)

1. **Probability Distribution Function (pdf)-** the probability (p) that a discrete random variable (X) takes on the value of x_i
    a. In English: What's the probability of getting a specific outcome(s) considering the range of all possible outcomes?
    b. *i.e.* $p(X = x_i)$ for the probability of realizing discrete variables
        i. Given $H = 2$ means that two of three flipped coins came up heads…
        ii. $p(H = 2) = p(2) = 3/8$
            1. 3 of the 8 equally possible outcomes result in two heads (HHT, HTH, THH)
    c. *NOTE:* pdf becomes $f(x)$ when describing the likelihood of realizing continuous variables
2. **Cumulative Distribution Function (cdf)-** the corresponding cumulative probability (P) that a discrete random variable (X) takes on a value less than or equal to x_i

a. In English: What's the probability of getting outcomes less than or equal to a specific value considering the range of all possible outcomes?
b. A function is a cdf if and only if it…
    i. Equals zero near negative infinity
    ii. Equals one near positive infinity
    iii. Is a non-decreasing function of x
    iv. Is right continuous
c. *i.e.* $P(x)$ for the probability of realizing discrete variables
    i. Given $H = 2$ means that two of three flipped coins came up heads…
    ii. $P(2) = p(X = H \leq 2) = 7/8$
        1. 7 of the 8 equally possible outcomes results in less than or exactly two heads (TTT, TTH, THT, HTT, THH, HTH, HHT)
d. *NOTE:* pdf becomes *F(x)* when describing the likelihood of realizing continuous variables

3. **Continuous PDFs and CDFs**
    a. Characteristics
        i. A random variable is continuous if its range includes an interval on the real number line with an infinite number of outcomes
        ii. The likelihood of any given outcome occurring is a positive measure but with zero value
        iii. pdf sums to infinity given all outcomes have a positive measure
    b. Calculus
        i. We take the derivative to move from a cdf to a pdf of a continuous r.v. (the pdf is the derivative of the cdf)
            1. *i.e.* $\frac{dF(x)}{dx} = f(x)$
        ii. We take the integral to move from a pdf to a cdf of a continuous r.v. (the cdf is the integral of the pdf)
            1. *i.e.* $F(x) = \int_{lb}^{ub} f(x)dx$
        iii. $[lb, ub]$- the range over which the continuous r.v. has positive measure
    c.

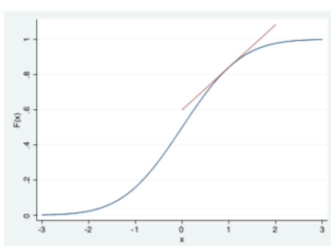


Figure: CDF

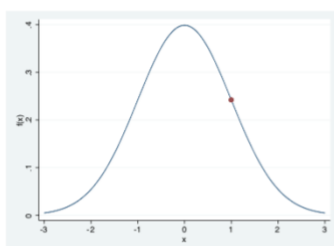


Figure: PDF

Moments:

Random Variables Notation/Symbols Cheat Sheet	
Symbol	**Meaning**

$E[Y] = \sum_{y\in Y(S)} yp(y)$	The calculation for the expected value of a discrete random variable *e.g.* The expected value of a dice roll: $E[D] = 1\times\frac{1}{6} + 2\times\frac{1}{6} + 3\times\frac{1}{6} + 4\times\frac{1}{6} + 5\times\frac{1}{6} + 6\times\frac{1}{6} = \frac{7}{2}$
$E[Z] = \int_{lb}^{ub} zf(z)dx$	The calculation for the expected value of a continuous random variable
$E[(X - E[X])^2] = E[X^2] - E[X]^2$	The calculation for the variance of a random variable *e.g.* The variance of random variable (H) that counts the number of heads on 3 coin tosses: $\left(0-\frac{3}{2}\right)^2\left(\frac{1}{8}\right) + \left(1-\frac{3}{2}\right)^2\left(\frac{3}{8}\right) + \left(2-\frac{3}{2}\right)^2\left(\frac{3}{8}\right) + \left(3-\frac{3}{2}\right)^2\left(\frac{3}{8}\right) = \frac{3}{4}$
σ_x	The standard deviation (square root of the variance)

1. **Expected Value-** The average or expected value (central tendency) of a random variable given a large sample of its realizations
   a. For a discrete random variable (Y): $E[Y] = \sum_{y\in Y(S)} yp(y)$
      i. In English: The expected value of a discrete random variable is equal to the sum of each possible multiplied by its probability
   b. For a continuous random variable (Z): $E[Z] = \int_{lb}^{ub} zf(z)dx$
      i. In English: The expected value of a continuous random variable is equal to the integral of the product of each differential outcome and its probability
   c. Properties:
      i. $E[ag(X) + b = aE[g(X)] + b$
2. **Moment-** n^th moment of a random variable is $E[X^n]$
   a. Expected value is a special case of the first moment
   b. Second moment is useful in calculating variance
3. **Variance-** summarizing measure of a random variable's dispersion
   a. *i.e.* $E[(X - E[X])^2] = E[X^2] - E[X]^2$
   b. Properties
      i. It is always greater than or equal to zero
      ii. The **standard deviation** (σ_x) is the square root of the variance
      iii. $var(aX + b) = a^2 var(X)$
      iv. When the distribution of the outcomes of a random variable is independent and identical (uniform), $var(\sum_{i=1}^{n} Z_i = \sum_{i=1}^{n} var(Z_i)$

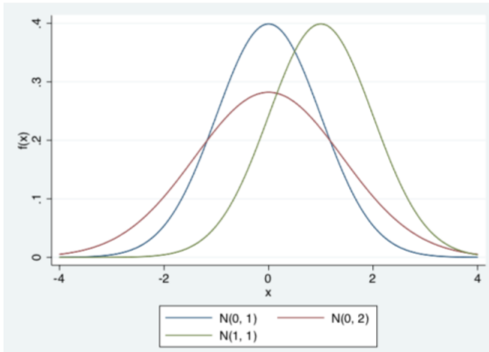
Distributions:

Random Variables Notation/Symbols Cheat Sheet	
Symbol	**Meaning**
$p(1) = \pi$ $0 < \pi < 1$ $p(0) = 1 - \pi$	The probability of returning the outcome with the value of one (success) in a Bernoulli distribution; This probability is always greater than zero and less than one in a Bernoulli distribution; The probability of returning the outcome with the value of zero
n	The number of Bernoulli random variables being summed together to form a binomial distribution
$Z = \sum_{i=1}^{n} X_i$ OR $Z(n, \pi)$	The calculation for the binomial distribution (Z) of a random variable; Can also be written in shorthand
$E[Z] = n\pi$	The calculation for the expected value of a binomial distribution
$var(Z) = n\pi(1 - \pi)$	The calculation for the variance of a binomial distribution
$p(z) = \binom{n}{z} \pi^z (1 - \pi)^{n-z}$ $= \frac{n!}{z!\,(n-z)!} \pi^z (1 - \pi)^{n-Z}$ OR `dbinom(z, n, π)`	The calculation for the probability of returning a value (z) from a binomial distribution; Can be solved in R
$P(z)$ OR `pbinom(z, n, π)`	The cumulative probability of returning a value (z) from a binomial distribution ($\leq z$); Can be solved in R
$X \sim U[a, b]$	Notation for a uniform distribution of a random variable (X) across range ($[a, b]$)
$f(y) = c$ if $a \leq y \leq b$ and 0 if $y < a$ or $y > b$ OR `dunif(# quantiles, a, b)`	Pdf of a uniform distribution; Can be solved in R

$c = \frac{1}{b-a}$	Probability of returning a value in a uniform distribution
$F(y) = 0 \text{ if } y \leq a, \frac{1}{b-a} \text{ if } a < y < b, 1 \text{ if } y \geq b$ *OR* `punif(# quantiles, a, b)`	Cdf of a uniform distribution; Can be solved in R
$\mu = \int_{-\infty}^{\infty} xf(x)dx$	Central tendency/mean/expected value of a normal distribution
$\sigma^2 = \int_{-\infty}^{\infty} x^2 f(x)dx - \mu^2$	Variance of a normal distribution
$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$ *OR* $X \sim N(\mu, \sigma^2)$	Conditions for a normal distribution of a random variable; Can also be written in shorthand
$a + bQ \sim N(a + \mu, b^2 q^2)$	Scalar transformations of a normal distribution
$Z = \frac{X-\mu}{\sigma} \sim N(0,1)$	Scalar transformations of a normal distribution to convert into a standard normal distribution
$\phi(z)$	Pdf for normal distribution
$\Phi(z)$	Cdf for normal distribution
$\phi(z) = \phi(-z)$	Probability symmetry property of normal distribution
$\Phi(c) = 1 - \Phi(-c)$; $P(c > Z) = P(Z < -c) = \Phi(-c) = 1 - \Phi(c)$	Inverse probability equality principle of normal distribution
`pnorm(x, μ, σ)`	R-code for solving for the cumulative probability of a random variable to the value (x)

1. **Bernoulli Distribution of a Random Variable-** Probability distributions which model random variables with binary situations (outcome of either 0 or 1) with $p(1) = \pi,$, and $p(0) = 1 - \pi$.
   a. In English: Bernoulli distributions model random variables that have only two possible results. Given that the total probability is equal to one and the probability of returning one outcome, represented by π is $>$ 0 and $<$ 1, the probability of the alternate outcome is $1-\pi$.
   b. *E.g.* coin flips, win/loss, success/failure, above/below
   c. **Variance for Bernoulli Distributions**- $\pi(1-\pi)$

2. **Binomial Distribution of a Random Variable**- Probability distribution defined as a series of independently and identically distributed (iid) Bernoulli random variables
    a. *i.e.* $Z = \sum_{i=1}^{n} X_i$ where $X_1, X_2, \dots, X_n$ is a series of iid Bernoulli random variables
        i. **Independently distributed**- the realization of one outcome (X_i) has no bearing on the value of another outcome (X_j)
        ii. **Identically distributed**- the probability that $X_1, X_2, \dots, X_n = 1$ have the same π
    b. Shorthand: $Z(n, \pi)$
    c. Facts about Z:
        i. $E[Z] = n\pi$
        ii. $var(Z) = n\pi(1-\pi)$
        iii. $p(z) = \binom{n}{z}\pi^z(1-\pi)^{n-z} = \frac{n!}{z!(n-z)!}\pi^z(1-\pi)^{n-Z}$
            1. Can be solved in R using "`dbinom(z, n, π)`"
        iv. $P(z)$ can be solved in R using "`pbinom(z, n, π)`" and used to solve segmented probabilities (e.g. finding probability of returning $40 \leq z \leq 50$ through $P(50) - P(39)$.
3. **Uniform Distribution**- The simplest distribution of a random variable (X) defined by a pdf that puts equal probability (U) on all outcomes over some interval on the number line ($[a, b]$)
    a. *i.e.* $X \sim U[a, b]$
        i. In English: Random variable (X) can result in all values between a and b with equal probability
        ii. *E.g.* dice roll
    b. Facts about X:
        i. Pdf of uniform random variable Y: $f(y) = c$ if $a \leq y \leq b$ and 0 if $y < a \; or \; y > b$
            1. $c = \frac{1}{b-a}$
        ii. Cdf of uniform random variable Y: $F(y) = 0$ if $y \leq a$, $\frac{1}{b-a}$ if $a < y < b$, 1 if $y \geq b$
4. **Normal Distribution**- the most important distribution function defined for any value between $-\infty$ and $+\infty$, with parameters being the expected value/mean (μ) and variance (σ^2)
    a. Notation:
        i. Normal distribution for random variable X: $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$
            1. Shorthand notation: $X \sim N(\mu, \sigma^2)$
        ii. Expected value: $\mu = \int_{-\infty}^{\infty} x f(x) dx$
        iii. Variance: $\sigma^2 = \int_{-\infty}^{\infty} x^2 f(x) dx - \mu^2$
    b. Visualization of bell curve:

i. 

c. Features:
   i. $a + bQ \sim N(a + \mu, b^2 q^2)$
   ii. **Standard normal random variable**- often we want to use scalars a and b to transform normal distributions ($F(x)$) into standard normal distributions ($F(z)$) where $\mu = 0$ and $\sigma^2 = 1$
      1. *E.g.* if $X \sim N(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma} \sim N(0, 1)$ where $a = -\mu$ and $b = \frac{1}{\sigma}$
         a. Z: the number of standard deviations (1) an outcome's value lies away from the mean (0)
      2. Use z-chart to find z-value for $p(Z)$ (
   iii. Pdf: $\phi(z)$ and Cdf: $\Phi(z)$
      1. $\phi(z) = \phi(-z)$ due to symmetry around zero
      2. $\Phi(c) = 1 - \Phi(-c)$
      3. $P(c > Z) = P(Z < -c) = \Phi(-c) = 1 - \Phi(c)$
d. Converting from normal random variable distribution $F(x)$ to standard normal random variable $F(z)$ to solve for $P(x)$ using R: `pnorm(x, μ, σ)`
   i. Can also use conversion formulas in conjunction with z-table

---

Confidence Intervals:

Random Variables Notation/Symbols Cheat Sheet	
Symbol	**Meaning**
$1 - \alpha$; α	The probability that the true expected value of a random variable is contained within the bounds of a confidence interval (*e.g.* a 95% confidence interval); The probability that the expected value of a random variable is not contained within the bounds of a confidence interval (*e.g.* a 95% confidence interval has a 5% chance of failing to capture the true expected value)
lb and ub	The lower and upper bound, respectively, of a confidence interval

$\Phi(ub) - \Phi(lb)$	Calculation for finding the probability/degree of confidence $(1-\alpha)$ of a confidence interval
$p(Q < lb) = p(Q > ub) \rightarrow \alpha_{lb} = \alpha_{ub} = \frac{\alpha}{2}$	Calculation for the tail-end probability of a symmetric, two-sided confidence interval
$p(\Phi^{-1}\left(\frac{\alpha}{2}\right) < Z < \Phi^{-1}\left(1-\frac{\alpha}{2}\right) = 1-\alpha$ *OR* `qnorm(1-α, μ, σ)`	Calculation for the lower and upper bounds of a symmetric, two-sided confidence interval given $1-\alpha$; Can also be solved in R

`rbinom(n = 1, size = 100, prob = 0.5);`
 	
`runif(n = 10, min = 3, max = 7);`
 	
`rnorm(n = 1000, mean = 3, prob = 7)`
	R code for Monte Carlo simulations of a binomial cdf distribution;
	
	R code for Monte Carlo simulations of a uniform cdf distribution;
	
	R code for Monte Carlo simulations of a normal cdf distribution;

$Y \sim F()$	The cdf distribution of a random variable
$X \sim U[0,1]$	Notation for running a Monte Carlo simulation of a cdf
$y = F^{-1}(x)$	Notation for indirectly/inversely sampling random variable values by matching value to their corresponding cdf probability when an inverse function ($F^{-1}()$) does exist
$y = \min(y)$ such that $F(y) \geq x)$	Method for inverse transform sampling which matches random variable values to probability along probability intervals for when a formal inverse function ($F^{-1}()$) does not exist

1. **Confidence Interval (CI)-** the probability that the realization of a random variable will occur within a given range
2. Two approaches to constructing confidence intervals:
    a. Given the value bounds, lb and ub, on the CI we find out the probability, $1-\alpha$, that the random variable is contained in the CI
        i. *E.g.* What is the probability that approval rate for the president is between 40% to 60%?
        ii. *NOTE*: α is the probability that the random variable is *not* contained within the CI, therefore, $1-\alpha$ is the probability that the random variable *is* contained within the CI
            1. *E.g.* an $\alpha = 0.05$ would produce a 95% confidence interval $(1 - 0.05 = 0.95)$
        iii. Solved using $\Phi(ub) - \Phi(lb)$

1. *i.e.* subtracting the cumulative probabilities of realizing each bound to find the enclosed confidence interval probability

b. Given the probability, $1-\alpha$, that the random variable is contained in the CI, we find out the bounds, lb and ub, on the CI
   i. *E.g.* Between which percentages is there a 95% probability does approval for the president lie?
   ii. More complicated solution: need inverse cdf function which can provide multiple CI's with the same probability
      1. Usually, we are finding a **symmetric two-sided CI** which presents an equal probability of realizing the random variable above and below a statistic
         a. $p(Q < lb) = p(Q > ub) \rightarrow \ \alpha_{lb} = \alpha_{ub} = \frac{\alpha}{2}$
            i. *i.e.* A 95% CI leaves 2.5% unincluded on either side of the CI below the lower bound and above the upper bound
         b. $p(\Phi^{-1}\left(\frac{\alpha}{2}\right) < Z < \Phi^{-1}\left(1-\frac{\alpha}{2}\right) = 1-\alpha$
            i. *E.g.* when $1-\alpha = 0.99,\ p(\Phi^{-1}(0.005) < Z < \Phi^{-1}(0.995) = p(-2.576 < Z < 2.576)$
         c. Can be solved in R using `qnorm(1-α, μ, σ)`

3. **Monte Carlo Simulation**- drawing realizations of random variables to simulate outcomes in a random process (cdf)
   a. In English: Compiling the results from drawing a large number of random, possible samples in order to approximate the probability of producing an outcome
   b. Performing Monte Carlo simulations in R:
      i. "`rbinom(n = 1, size = 100, prob = 0.5)`" draws one random variables from a binomial distribution of 100 trials with each trial having a probability of success of 0.5
      ii. "`runif(n = 10, min = 3, max = 7)`" draws 10 random variables from a uniform distribution between sd and 7
      iii. "`rnorm(n = 1000, mean = 3, prob = 7)`" draws 1000 random variables from a normal distribution with a mean of 2 and a standard deviation of 4
      iv. *NOTE:* Because these functions are simulations, we will get different values each time we run them unless we use the "`set.seed()`" function prior to drawing the random variables

4. **Inverse Transform Sampling Method**- using a Monte Carlo simulation to approximate the pdf of a distribution of a random variable with a known cdf
   a. We can apply the inverse transform sampling method for $Y \sim F()$ when $F^{-1}()$ exists by first performing a Monte Carlo simulation of known cdf probabilities ($X \sim U[0,1]$) and then matching corresponding random variable values with their respective cdf probabilities from the cdf simulation ($y = F^{-1}(x)$). to produce a pdf distribution.
      i. *E.g.* inverse transform sampling a random variable with a normal distribution

b. We can also apply the inverse transform sampling method for $Y \sim F()$ when $F^{-1}()$ does *not* exist by first performing a Monte Carlo simulation of known cdf probabilities ($X \sim U[0, 1]$) and then matching corresponding random variable values with their cdf probabilities from the cdf simulation (set $y = \min(y)$ such that $F(y) \geq x$) to produce a pdf distribution
   i. *E.g.* to create a sampling distribution of a die roll we can use the inverse transform sampling method