STAT 206B
Review: Probability Distributions

Winter 2022

† Probability

- Probability: A number between 0 and 1 assigned to an event A in the sample space, $\mathcal{S}$.

- A way to numerically express our belief and information about unknown quantities

- Axioms of Probability (Kolmogorov Axiom System): Given a sample space $\mathcal{S}$ and an associated sigma algebra $\mathcal{B}$, a *probability function* is a function Pr with domain $\mathcal{B}$ that satisfies;

  ⋆⋆ $\Pr(A_i) \geq 0$ for all $A_i \in \mathcal{B}$.

  ⋆⋆ $\Pr(\mathcal{S}) = 1$.

  ⋆⋆ If $A_1, A_2, \ldots \in \mathcal{B}$ are pairwise disjoint, then $\Pr(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \Pr(A_i)$.

† Interpretations of Probabilities

- (Frequency) An event's probability is the proportion of times that we expect the event to occur, if *the experiment were repeated a large number of times* – that is, relative frequencies.

  e.g. Roll a die repeatedly. Count how many times each face came up.

- (Classical) An event's probability is the ratio of the number of favorable outcomes and possible outcomes in a (**symmetric**) experiment.

  $**$ symmetric experiment: all single points in $\mathcal{S}$ are "equiprobable".

- (Subjectivist) A subject probability is an individual's degree of belief in the occurrence of an event.

† Interpretations of Probabilities – contd

- Any function Pr that satisfies the Axioms of Probability is called a probability function.
- For any sample space, many different probability functions can be defined.
- The axiomatic definition makes no attempt to tell what particular function Pr to choose.
- No single scientific interpretation of the term *probability* is accepted by all statisticians, philosophers, and other authorities.

† How to update our degree of belief? Bayes' Theorem

- If H denotes an hypothesis and D denotes data, the Bayes' theorem states

$$\Pr(H \mid D) = \frac{\Pr(D \mid H)\Pr(H)}{\Pr(D)}.$$

- $\Pr(H)$: a probabilistic statement of belief about H *before* obtaining data D.
- $\Pr(H \mid D)$: a probabilistic statement of belief about H *after* obtaining data D.
- Having specified $\Pr(D)$ and $\Pr(D \mid H)$, the mechanism of the theorem provides a solution to the problem of how to learn from data. i.e. modify the degrees of belief attached to the events when a real-world event occurs.

† Random Variables & Probability Distributions

- Definition (not rigorous): A random variable, X is a real-valued function from a sample space $\mathcal{S}$ into real numbers (range: $\mathcal{X}$, a new sample space).

  e.g.1 Toss a coin. Define a random variable $X(\{H\}) = 1$ and $X(\{T\}) = 0$

- We can define a probability function on $\mathcal{X}$. For example, suppose $\mathcal{S} = \{s_1, \ldots, s_n\}$ with a probability function Pr. We define a random variable X with range $\mathcal{X} = \{x_1, \ldots, x_m\}$. We can define a probability function Pr_X on $\mathcal{X}$ in the following way.

$$\mathrm{Pr}_X(X = x_i) = \mathrm{Pr}(\{s_j \in \mathcal{S} : X(s_j) = x_i\}).$$

  The function Pr_X is an induced probability function on $\mathcal{X}$, defined in terms of the original function Pr.

† Probability Distribution

- discrete distributions, continuous distributions, mixed distributions.

- The distribution of a random variable (X) is formally defined

$$F(t) \equiv F_X(t) \equiv \Pr(X \leq t) \equiv \Pr(\{s \in \mathcal{S}; X(s) \leq t\}).$$

  ⋆⋆ $F(\infty) = 1$, $F(-\infty) = 0$ and $F(a) \leq F(b)$ if $a < b$.

- *Descriptions of a distribution*: moments, mode, median, quantiles, variance, standard deviations, correlations...

- For more than one random variables: joint distributions, marginal distributions, conditional distributions....

- independent random variables, conditionally independent random variables, exchangeability...

• **Bayes Theorem for Random Variables** (D & S Th 3.6.4): If $f_2(y)$ is the marginal p.f. or p.d.f. of a random variable Y and $g_1(x \mid y)$ is the conditional p.f. or p.d.f. of X given $Y = y$, then the conditional p.f. or p.d.f. of Y given $X = x$ is

$$g_2(y \mid x) = \frac{g_1(x \mid y) f_2(y)}{f_1(x)},$$

where $f_1(x)$ is the marginal p.f. or p.d.f. of X;

⋆⋆ $f_1(x) = \sum_y g_1(x \mid y) f_2(y)$ if Y is discrete.

⋆⋆ If Y is continuous, $f_1(x) = \int_{-\infty}^{\infty} g_1(x \mid y) f_2(y) dy$.

† Some Important Distributions

See Appendix A of CR or Chapter 3 of Casella and Berger for more

- **Normal distribution**, $\mathrm{N}_p(\boldsymbol{\theta}, \Sigma)$.

  $\boldsymbol{\theta} \in \mathbb{R}^p$ and Σ is a $(p \times p)$ symmetric positive-definite matrix,

  $$f(\boldsymbol{x} \mid \boldsymbol{\theta}, \Sigma) = |\Sigma|^{-1/2}(2\pi)^{-p/2} \exp\left\{-(\boldsymbol{x} - \boldsymbol{\theta})'\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\theta})/2\right\}.$$

  ⋆⋆ $\mathrm{E}(\boldsymbol{X}) = \boldsymbol{\theta}$ and $\mathrm{E}((\boldsymbol{x} - \boldsymbol{\theta})(\boldsymbol{x} - \boldsymbol{\theta})') = \Sigma$.

  ⋆⋆ If Σ is not definite, the distribution has no density with respect to Lebesgue measure.

  ⋆⋆ Here $\boldsymbol{\theta}$ and Σ can be set to different values, producing different probability distributions $\Rightarrow$ $\boldsymbol{\theta}$ and Σ are called *parameters!*

- **Normal distribution**, $\mathrm{N}_p(\boldsymbol{\theta}, \Sigma)$ – contd.

  ⋆⋆ univariate ($p = 1$)

$$f(x \mid \theta, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\theta)^2}{2\sigma^2}\right\},$$

  where $\theta \in \mathbb{R}$ and $\sigma \in \mathbb{R}^+$.

  * $\mathrm{E}(X) = \theta$ and $\mathrm{Var}(X) = \sigma^2$.
  * $M_X(t) = \exp\left(\theta t + \frac{1}{2}\sigma^2 t^2\right)$
  * $\theta = 0$ and $\sigma = 1 \Rightarrow \mathrm{N}(0, 1)$, standard normal distribution
  * If $X \sim \mathrm{N}(\theta, \sigma^2)$, then $Y = \exp(X) \sim \text{log-N}(\theta, \sigma^2)$.

- **Uniform Distribution** Unif(a, b)

  $a, b \in \mathbb{R}$,
  $$f(x \mid a, b) = \frac{1}{b-a}, \qquad a < x < b.$$

  ⋆⋆ $\mathrm{E}(X) = (b-1)/2$ and $\mathrm{Var}(X) = (b-a)^2/12$

  ⋆⋆ If $X \sim \mathrm{Unif}(a, b)$, $X = (Y-a)/(b-a) \sim \mathrm{Unif}(0, 1)$.

  ⋆⋆ If $X \sim F$, where F is a continuous cdf, then $Y = F(X) \sim \mathrm{Unif}(0, 1)$.

- **Gamma Distribution** Gamma(α, β)

  $\alpha, \beta > 0$,

  $$f(x \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x), \qquad x > 0$$

  ⋆⋆ $E(X) = \alpha/\beta$ and $\text{Var}(X) = \alpha/\beta^2$ (α: shape, β: rate)

  ⋆⋆ Special cases:

  ∗ Erlang distribution: Gamma(k, β), $k = 1, 2, \ldots$ and $\beta \in \mathbb{R}$

  ∗ Exponential distribution: Gamma$(1, \beta)$

  ∗ χ^2 distribution: Gamma$(\nu/2, 1/2)$ (χ^2_ν)

  ⋆⋆ Sometimes it is parameterized as Gamma$(\alpha, 1/\beta)$ ($1/\beta$: scale).

- **Gamma Distribution** Gamma(α, β)–contd

  ⋆⋆ Inverse gamma distribution IG(α, β): when $X \sim$ Gamma(α, β), the distribution of $Y = X^{-1}$ is IG(α, β),

$$f(y \mid \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} y^{-(\alpha+1)} \exp(-\beta/y), \quad y > 0.$$

  ⋆⋆ $\mathrm{E}(Y) = \beta/(\alpha - 1)$ for $\alpha > 1$ and $\mathrm{Var}(Y) = \beta^2/\{(\alpha - 1)^2(\alpha - 2)\}$ for $\alpha > 2$

- **Student's t_n Distribution** t_n (n degrees of freedom)

  $n > 0$,

$$f(x \mid n) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}, \quad x \in \mathbb{R}$$

  ⋆⋆ $\mathrm{E}(X) = 0$ and $\mathrm{Var}(X) = n/(n-2)$ if $n > 2$.

  ⋆⋆ Let $X \mid W \sim \mathrm{N}(0, W)$ and $W \sim \mathrm{IG}(n/2, n/2)$. The marginal distribution $X \sim t_n$.

  ⋆⋆ Special cases:

  ∗ If $n = 1$, t_1 is the Cauchy distribution.

- **Beta Distribution** $\text{Be}(\alpha, \beta)$

$\alpha, \beta > 0$,

$$f(x \mid \alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}, \quad 0 < x < 1,$$

where

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

⋆⋆ $\text{E}(X) = \alpha/(\alpha+\beta)$ and $\text{Var}(X) = \alpha\beta/\{(\alpha+\beta)^2(\alpha+\beta+1)\}$

⋆⋆ $\text{Be}(1,1) \Rightarrow \text{Unif}(0,1)$

⋆⋆ Relationship: $Y_1 \sim \text{Gamma}(\alpha, \theta)$ and $Y_2 \sim \text{Gamma}(\beta, \theta)$, independently. Then the distribution of $X = Y_1/(Y_1 + Y_2)$ follows $\text{Be}(\alpha, \beta)$.

- **Dirichlet Distribution** $\mathrm{Dir}_k(\alpha_1, \ldots, \alpha_k)$

  $\alpha_1, \ldots, \alpha_k > 0$ and $\alpha_0 = \alpha_1 + \ldots + \alpha_k$,

  $$f(\boldsymbol{x} \mid \alpha_1, \ldots, \alpha_k) = \frac{\Gamma(k_0)}{\Gamma(\alpha_1) \ldots \Gamma(\alpha_k)} x_1^{\alpha_1 - 1} \ldots x_k^{\alpha_k - 1},$$

  for $0 < x_1, \ldots, x_k < 1 \quad \& \sum_{i=1}^k x_i = 1$.

  ⋆⋆ $\mathrm{E}(X_i) = \alpha_i / \alpha_0$ and $\mathrm{Var}(X_i) = (\alpha_0 - \alpha_i)\alpha_i / \{\alpha_0^2(\alpha_0 + 1)\}$ and $\mathrm{Cov}(X_i, X_j) = -\alpha_i \alpha_j / \{\alpha_0^2(\alpha_0 + 1)\}$, $i \neq j$.

  ⋆⋆ Special case: $k = 2$, $(X, 1 - X) \sim \mathrm{Dir}_2(\alpha_1, \alpha_2)$ is equivalent to $X \sim \mathrm{Be}(\alpha_1, \alpha_2)$.

- **Pareto Distribution** $\mathrm{Pa}(\alpha, x_0)$

  $\alpha > 0$ and $x_0 > 0$

  $$f(x \mid \alpha, x_0) = \alpha \frac{x_0^\alpha}{x^{\alpha+1}}, \quad x \geq x_0.$$

  ⋆⋆ $\mathrm{E}(X_i) = \alpha x_0/(\alpha - 1)$ $(\alpha > 1)$ and $\mathrm{Var}(X) = \alpha x_0^2/\{(\alpha - 1)^2(\alpha - 2)\}$ $(\alpha > 2)$.

- **Wishart Distribution** $W_m(\alpha, \Sigma)$

  $\alpha > 0$ and $\Sigma > 0$

  $$f(X \mid \alpha, \Sigma) = \frac{|X|^{\frac{\alpha-(m+1)}{2}} \exp\left(-tr(\Sigma^{-1}X)\right)/2}{\Gamma_m(\alpha)|\Sigma|^{\alpha/2}}, \quad X > 0.$$

  ⋆⋆ $\Gamma_m(\alpha)$ is a multivariate Gamma function.

  ⋆⋆ $\mathrm{E}(X) = \alpha\Sigma$

  ⋆⋆ $W = X^{-1}$ follows the inverse-Wishart distribution with parameters α and Σ^{-1} (careful with the parameterizations).

  $$f(W \mid \alpha, \Sigma) = \frac{|W|^{-\frac{\alpha+m+1)}{2}} \exp\left(-tr(\Sigma^{-1}W^{-1})\right)/2}{\Gamma_m(\alpha)|\Sigma|^{\alpha/2}}, \quad W > 0.$$

- **Point Mass Distribution** δ_a

  $a \in \mathbb{R}$

$$f(x \mid a) = \delta_a = \begin{cases} 1 & \text{if } x = a, \\ 0 & \text{if } x \neq a. \end{cases}$$

  $\star\star$ $\mathrm{E}(X) = a$ and $\mathrm{Var}(X) = 0$.

- **Bernoulli Distribution** $\mathrm{Ber}(p)$

  $0 \le p \le 1$

$$f(x \mid \lambda) = p^x (1-p)^{1-x}, \quad x \in \{0, 1\}.$$

  ⋆⋆ $\mathrm{E}(X) = p$ and $\mathrm{Var}(X) = p(1-p)$.

- **Binomial Distribution** $\mathrm{Bin}(n, p)$

  $0 \le p \le 1$

$$f(x \mid p) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x \in \{0, 1, \ldots, n\}.$$

$\star\star$ $\mathrm{E}(X) = np$ and $\mathrm{Var}(X) = np(1-p)$.

- **Poisson Distribution** $\mathrm{Poi}(\lambda)$

  $\lambda > 0$

  $$f(x \mid \lambda) = e^{-\lambda} \frac{\lambda^x}{x!}, \quad x \in \{0, 1, \ldots\}.$$

  $\star\star$ $\mathrm{E}(X) = \lambda$ and $\mathrm{Var}(X) = \lambda$.

- **Multinomial Distribution** $\text{Multinomial}_k(n, p_1, \ldots, p_k)$

  $0 \le p_i \le 1$, $i = 1, \ldots, k$ and $\sum p_i = 1$

  $$f(x_1, \ldots, x_k \mid p_1, \ldots, p_k) = \binom{n}{x_1 \ldots x_k} \prod_{i=1}^{k} p_i^{x_i},$$

  $x_i \in \{0, 1, \ldots, n\}$ with $\sum_{i=1}^{k} x_i = n$.

  ⋆⋆ $E(X_i) = np_i$, $\text{Var}(X_i) = np_i(1 - p_i)$ and $\text{Cov}(X_i, X_j) = -np_i p_j$ $(i \neq j)$.

  ⋆⋆ Special case: $(X, n - X) \sim \text{Multinomial}_2(n, p, 1 - p) \equiv X \sim \text{Bin}(n, p)$

- **Negative Binomial Distribution** Neg-Bin(n, p)

  $0 \le p \le 1$

  $$f(x \mid p) = \binom{n+x+1}{x} p^n (1-p)^x, \quad x \in \{0, 1, \ldots\}.$$

  ⋆⋆ random variable X= number of failures before the n-th success where n is fixed (the total # of trials: $X + n$)

  ⋆⋆ $\mathrm{E}(X) = n(1-p)/p$ and $\mathrm{Var}(X) = n(1-p)/p^2$.

  ⋆⋆ Can be defined in terms of the random variable Y the trials at which the n-th success occurs (i.e., $Y = n + X$).

  ⋆⋆ $n = 1 \Rightarrow$ Geometric distribution.

- **Hypergeometric Distribution** $\text{Hyp}(N, n, p)$

  $0 \le p \le 1$, $n < N$ and $pN \in \mathbb{N}$,

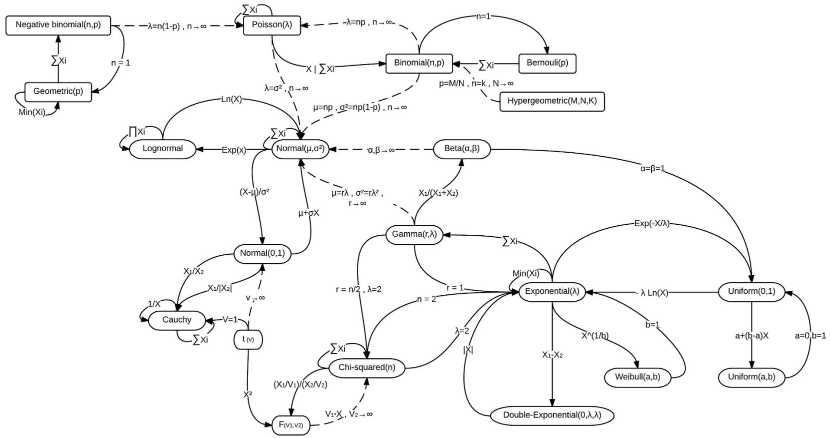
  $$f(x \mid p) = \frac{\binom{pN}{x}\binom{(1-p)N}{n-x}}{\binom{N}{n}},$$

  where $x \in \{n-(1-p)N, \dots, pN\}$ & $x \in \{0, 1, \dots, n\}$.

  ⋆⋆ $\mathrm{E}(X) = np$ and $\mathrm{Var}(X) = (N-n)np(1-p)/(N-1)$.

  ⋆⋆ N balls in total with pN in red and $(1-p)N$ in green. Select n balls at random (sampling without replacement) and random variable X denotes the number of red balls drawn.

- Relationship between Distributions



∗ *From wiki* (or page 627 of CB)

- A lot more not mentioned: ***t*-distribution, Laplace (double-exponential) distribution, *F*-distribution**

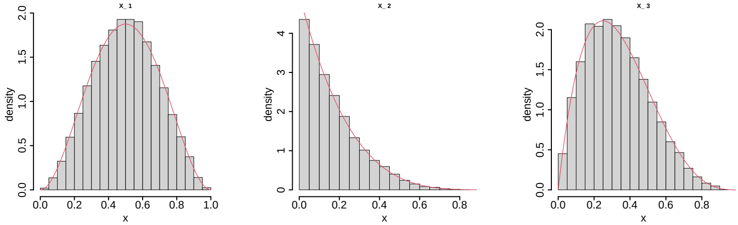
- Distributions can be parameterized in different ways. Please be careful when working on problems from JB since JB uses a parameterization different from that in CB.

- Simulating Random Samples from R
  - ⋆⋆ Use built-in functions. e.g.; `rnorm, dnorm, pnorm, qnorm...`
  - ⋆⋆ Use relationships between distributions.
  - ⋆⋆ Use relationship $p(x, y) = p(x)p(y \mid x)$ to simulation from a joint distribution when possible

- Example 1: Dirichlet distribution
  - ⋆⋆ Obtain a random sample from a Dirichlet distribution $\boldsymbol{x} = (x_1, \ldots, x_k) \sim \text{Dir}(a_1, \ldots, a_k)$.
  - ⋆⋆ (Step 1:) Simulate $\tilde{x}_p \sim \text{Gamma}(a_p, c)$, $p = 1, \ldots, k$, where $\tilde{x}_p$'s are independent and $c > 0$ is an arbitrary constant. Then let $x_p = \tilde{x}_p / \sum_{p'=1}^{k} \tilde{x}_{p'}$, $p = 1, \ldots, k$.
  - ⋆⋆ (Step 2:) Repeat until the target sample size is met.

- Example 1: Dirichlet distribution (contd)

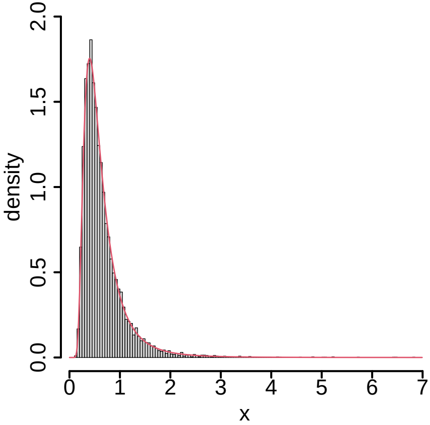
  ⋆⋆ Simulate $\boldsymbol{x} \sim \text{Dirichlet}(3, 1, 2)$



- Example 2: IG distribution
  - ⋆⋆ Obtain a random sample from an inverse Gamma distribution, $x \sim \text{IG}(a, b)$.
  - ⋆⋆ (Step 1:) Simulate $\tilde{x} \sim \text{Gamma}(a, b)$, where b is a rate parameter (so $\text{E}(\tilde{x}) = a/b$). Then let $x = 1/\tilde{x}$.
  - ⋆⋆ (Step 2:) Repeat until the target sample size is met.

- Example 2: IG (contd)

  ⋆⋆ Simulate $\boldsymbol{x} \sim \text{IG}(4, 2)$



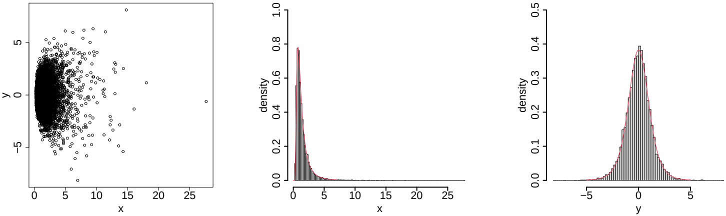
- Example 3: Normal $\times$ IG distribution
  - ⋆⋆ Suppose we have

$$\begin{aligned} p(x,y) &= p(x)p(y \mid x) \\ &= \underbrace{\frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{-\alpha-1} \exp\left(-\frac{\beta}{x}\right)}_{\mathrm{IG}(x \mid \alpha, \beta)} \underbrace{\frac{1}{\sqrt{2\pi x}} \exp\left(-\frac{(y-m)^2}{2x}\right)}_{\mathrm{N}(y \mid m, x)}. \end{aligned}$$

  - ⋆⋆ Obtain a random sample of (x, y) from their joint $p(x, y)$.
  - ⋆⋆ (Step 1:) Simulate $x \sim \mathrm{IG}(x \mid \alpha, \beta)$ and $y \mid x \sim \mathrm{N}(y \mid m, x)$.
  - ⋆⋆ (Step 2:) Repeat until the target sample size is met.

- Example 3: Normal $\times$ IG distribution (contd)

$\star\star$ Simulate $(x, y) \sim \mathrm{IG}(x \mid 3, 3)\mathrm{N}(y \mid 0, x)$



(a) (x, y) (b) x (c) y