Probability Theory

Patrick Lam

Outline

Most Basic Definition of Probability:

$$\frac{\text{number of successes}}{\text{number of possible occurrences}}$$

Three Axioms of Probability

Let S be the sample space and A be an event in S.

1. For any event A, $P(A) \geq 0$.
2. $P(S) = 1$.
3. If $A_1, A_2, \ldots, A_n$ are mutually disjoint, then

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = P(A_1) + P(A_2) + \cdots + P(A_n)$$

The three axioms imply:

- $P(\emptyset) = 0$
- $P(A^c) = 1 - P(A)$
- For any event A, $0 \leq P(A) \leq 1$.
- If $A \subset B$, then $P(A) \leq P(B)$.
- For any two events A and B, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Conditional Probability

$$P(A|B) \;=\; \frac{P(A \cap B)}{P(B)}$$

Multiplicative Law of Probability:

$$P(A \cap B) = P(B|A)P(A) = P(A|B)P(B)$$

Bayes' Rule:

$$P(A|B) \;=\; \frac{P(B|A)P(A)}{P(B)}$$

Law of Total Probability:

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$$

Independence

A and B are independent if

$$P(AB) = P(A)P(B)$$

If A and B are independent, then

$$\begin{aligned} P(A|B) &= \frac{P(AB)}{P(B)} \\ &= \frac{P(A)P(B)}{P(B)} \\ &= P(A) \end{aligned}$$

Conditional Independence:

$$P(AB|C) = P(A|C)P(B|C)$$

Outline

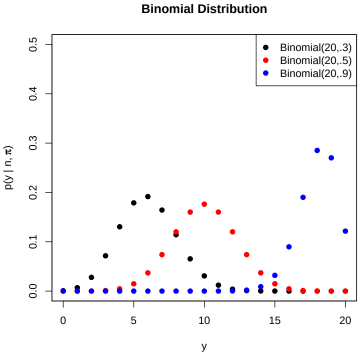
Random Variables

A **random variable** is a function that takes a random experiment and assigns a number to the outcome of the experiment.

Outcome values are assigned probabilities by a probability mass function (for discrete RV) or probability density function (for continuous RV).

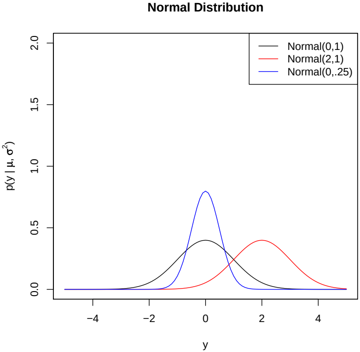
Probability Mass Function

$$P(Y = y)$$



Probability Density Function

$$P(Y \in A) = \int_A f(y)dy$$



Characteristics of all PDFs and PMFs:

- Area under the curve must integrate to 1
- $P(Y = y) \geq 0$ for all Y
- The support is all y's where $P(Y = y) > 0$.

Marginal, Conditional, and Joint Densities

$$
\begin{aligned}
f(x) &= \int f(x,y)dy \\
f(x,y) &= \int f(x,y,z)dz \\
f(x|y) &= \frac{f(x,y)}{f(y)} \\
f(x|y,z) &= \frac{f(x,y,z)}{f(y,z)} \\
f(x,y) &= f(x|y)f(y) \\
&= f(y|x)f(x) \\
f(x,y,z) &= f(x|y,z)f(y|z)f(z)
\end{aligned}
$$

Expectation

Discrete Case:

$$E(X) = \sum_i x_i P(X = x_i)$$

where $P(X = x)$ is the probability mass function (PMF).

Continuous Case:

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

where $f(x)$ is the probability density function (PDF).

Expectation of a Function of a Random Variable

$$E[g(X)] = \sum_i g(x_i)P(X = x_i)$$

for discrete random variables and

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx$$

for continuous random variables.

Variance

$$\mathrm{Var}(X) = E[(X - E(X))^2]$$

$$\begin{aligned}
\mathrm{Var}(X) &= E[(X - E(X))^2] \\
&= E[X^2 - 2XE(X) + (E(X))^2] \\
&= E(X^2) - 2E(X)E[E(X)] + E([E(X)]^2) \\
&= E(X^2) - 2[E(X)]^2 + [E(X)]^2 \\
&= \mathbf{E(X^2) - [E(X)]^2}
\end{aligned}$$

Outline

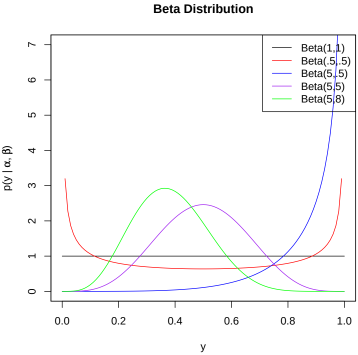
Monte Carlo Simulation

All the simulations we will be doing in class is what we call **Monte Carlo simulation**.

Fancy way of saying we will simulate random draws to calculate quantities of interest.

Simulating from a Random Variable

Suppose we have a random variable X that follows a Beta(α, β) distribution.



$$f(x|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{(\alpha-1)}(1-x)^{(\beta-1)}$$

Let $\alpha = 2$ and $\beta = 3$. Find $E(X)$.

$$\begin{aligned} E(X) &= \int_0^1 xf(x)dx \\ &= \int_0^1 x \, \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{(\alpha-1)}(1-x)^{(\beta-1)} \\ &= \int_0^1 x \, \frac{\Gamma(2+3)}{\Gamma(2)\Gamma(3)} x^{(2-1)}(1-x)^{(3-1)}dx \end{aligned}$$

We can ask R to help us integrate.

```
> ex.beta.func <- function(x, alpha, beta) {
+     x * gamma(alpha + beta)/(gamma(alpha) * gamma(beta)) * x^(alpha -
+         1) * (1 - x)^(beta - 1)
+ }
> e.x <- integrate(Vectorize(ex.beta.func), lower = 0, upper = 1,
+     alpha = 2, beta = 3)$value
> e.x
```

```
[1] 0.4
```

Or we can use simulation, which is much easier.

```
> x.draws <- rbeta(10000, shape1 = 2, shape2 = 3)
> sim.e.x <- mean(x.draws)
> sim.e.x

[1] 0.3987049
```

We can find all kinds of quantities of interest (variance, quantiles, etc.) by just doing it on the simulated draws rather than doing complicated integrals.

Why does this work?

Monte Carlo Integration

What we just did was called **Monte Carlo Integration**, which means exactly what it sounds like (doing integrals via Monte Carlo simulation).

If we need to take an integral of the following form:

$$I = \int g(x)p(x)dx$$

Monte Carlo Integration allows us to approximate it by simulating M values from $f(x)$ and calculating:

$$\hat{I}_M = \frac{1}{M}\sum_{i=1}^{M} g(x^{(i)})$$

By the Strong Law of Large Numbers, our estimate $\hat{I}_M$ is a simulation consistent estimator of I as $M \to \infty$ (our estimate gets better as we increase the number of simulations).

Let Y be another random variable where $Y = e^X$. Find $E(Y)$.

$$E(Y) = \int_0^1 e^x \; \frac{\Gamma(2+3)}{\Gamma(2)\Gamma(3)} x^{(2-1)}(1-x)^{(3-1)}dx$$

Via simulation:

```
> e.y <- mean(exp(x.draws))
> e.y

[1] 1.520348
```

Note that $E(g(X)) \neq g(E(X))$. In fact, $E(g(X)) \geq g(E(X))$ by Jensen's Inequality.

Monte Carlo Integration tells us we need $E(g(X))$.

Take home point:

If we can somehow generate random draws from the distribution of a random variable, we can calculate complicated integrals (mean, variance, functions of RV) easily by simulation.

This seems trivial but is one of the foundations of statistics, especially Bayesian statistics.

Simulating Probability Problems

A related application of simulation is to help us solve probability problems.

General idea:

1. Suppose we have an experiment where we want to know the probability of success. Simulate from the population many times.
2. For each simulation, conduct the experiment and see whether there is success.
3. The proportion of simulations that achieve success is the probability of success.

An Example

Suppose we have two urns containing marbles. The first urn contains 6 red marbles and 4 green marbles and the second urn contains 9 red marbles and 1 green marble. Take one marble from the first urn (without looking at it) and put it in the second urn. Then take one marble from the second urn (again without looking at it) and put it in the first urn. What is the probability of now drawing a red marble from the first urn?

```
> urn.func <- function(n.sims, urn1, urn2) {
+     final.draws <- c()
+     for (i in 1:n.sims) {
+         draw1 <- sample(urn1, 1)
+         draw2 <- sample(c(urn2, draw1), 1)
+         final.draws[i] <- sample(c(urn1, draw2), 1)
+     }
+     prob <- mean(final.draws)
+     return(prob)
+ }
> urn.func(n.sims = 10000, urn1 = c(rep(1, 6), rep(0, 4)), urn2 = c(rep(1,
+     9), 0))

[1] 0.6293
```

$$\left(\frac{6}{10}\right)\left(\frac{10}{11}\right)\left(\frac{6}{10}\right)+\left(\frac{6}{10}\right)\left(\frac{1}{11}\right)\left(\frac{5}{10}\right)+\left(\frac{4}{10}\right)\left(\frac{9}{11}\right)\left(\frac{7}{10}\right)+\left(\frac{4}{10}\right)\left(\frac{2}{11}\right)\left(\frac{6}{10}\right)\approx 0.63$$

Another Example

Suppose we have two urns containing marbles. The first urn contains $10 - g$ red marbles and g green marbles and the second urn contains 9 red marbles and 1 green marble. Take one marble from the first urn (without looking at it) and put it in the second urn. Then take one marble from the second urn (again without looking at it) and put it in the first urn. What is the minimum g such that the probability of now drawing a red marble is less than 0.5?

```
> urn.func2 <- function(n.sims, urn2, p) {
+     final.draws <- c()
+     g <- 0
+     urn1 <- c(rep(1, 10 - g), rep(0, g))
+     prob <- 1
+     while (prob >= p) {
+         for (i in 1:n.sims) {
+             draw1 <- sample(urn1, 1)
+             draw2 <- sample(c(urn2, draw1), 1)
+             final.draws[i] <- sample(c(urn1, draw2), 1)
+         }
+         prob <- mean(final.draws)
+         g <- g + 1
+         urn1 <- c(rep(1, 10 - g), rep(0, g))
+     }
+     g <- g - 1
+     return(g)
+ }
> urn.func2(1000, urn2 = c(rep(1, 9), 0), p = 0.5)

[1] 6
```

Outline

Outline

The Bernoulli Distribution

$Y \sim \text{Bernoulli}(\pi)$

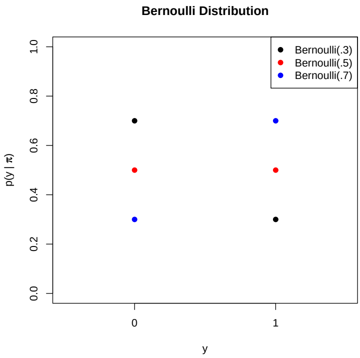
$y = 0, 1$

probability of success: $\pi \in [0, 1]$

$p(y|\pi) = \pi^y (1 - \pi)^{(1-y)}$

$E(Y) = \pi$

$\text{Var}(Y) = \pi(1 - \pi)$



The Binomial Distribution

$Y \sim \text{Binomial}(n, \pi)$

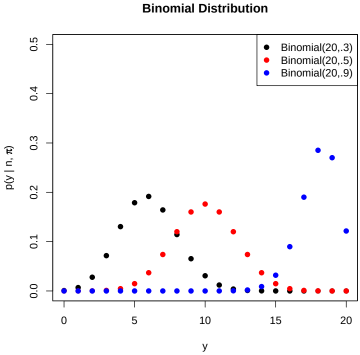
$y = 0, 1, \ldots, n$

number of trials: $n \in \{1, 2, \ldots\}$
probability of success: $\pi \in [0, 1]$

$p(y|\pi) = \binom{n}{y} \pi^y (1 - \pi)^{(n-y)}$

$E(Y) = n\pi$

$\text{Var}(Y) = n\pi(1 - \pi)$



The Multinomial Distribution

$Y \sim \text{Multinomial}(n, \pi_1, \ldots, \pi_k)$

$y_j = 0, 1, \ldots, n; \;\; \sum_{j=1}^{k} y_j = n$

number of trials: $n \in \{1, 2, \ldots\}$
probability of success for j: $\pi_j \in [0, 1]; \;\; \sum_{j=1}^{k} \pi_j = 1$

$p(\mathbf{y}|n, \boldsymbol{\pi}) = \frac{n!}{y_1! y_2! \ldots y_k!} \pi_1^{y_1} \pi_2^{y_2} \ldots \pi_k^{y_k}$

$E(Y_j) = n\pi_j$

$\text{Var}(Y_j) = n\pi_j(1 - \pi_j)$

$\text{Cov}(Y_i, Y_j) = -n\pi_i\pi_j$

The Poisson Distribution

$Y \sim \text{Poisson}(\lambda)$

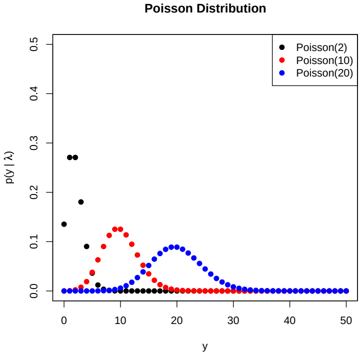
$y = 0, 1, \ldots$

expected number of occurrences: $\lambda > 0$

$p(y|\lambda) = \frac{e^{-\lambda}\lambda^y}{y!}$

$E(Y) = \lambda$

$\text{Var}(Y) = \lambda$



The Geometric Distribution

How many Bernoulli trials until success?

$Y \sim \text{Geometric}(\pi)$

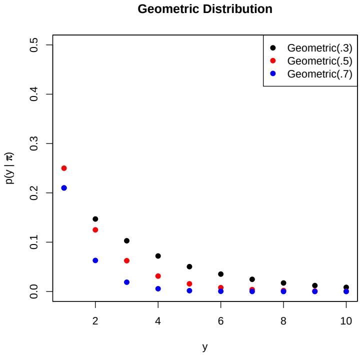
$y = 1, 2, 3, \ldots$

probability of success: $\pi \in [0, 1]$

$p(y|\pi) = (1 - \pi)^{(y-1)}\pi$

$E(Y) = \frac{1}{\pi}$

$\text{Var}(Y) = \frac{1-\pi}{\pi^2}$



Outline

The Univariate Normal Distribution

$$Y \sim \text{Normal}(\mu, \sigma^2)$$

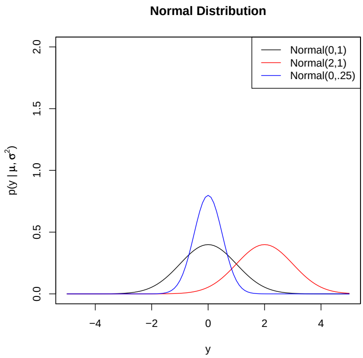
$$y \in \mathbb{R}$$

mean: $\mu \in \mathbb{R}$
variance: $\sigma^2 > 0$

$$p(y|\mu, \sigma^2) = \frac{\exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)}{\sigma\sqrt{2\pi}}$$

$$E(Y) = \mu$$

$$\text{Var}(Y) = \sigma^2$$



The Multivariate Normal Distribution

$Y \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

$\mathbf{y} \in \mathbb{R}^k$

mean vector: $\boldsymbol{\mu} \in \mathbb{R}^k$
variance-covariance matrix: $\boldsymbol{\Sigma}$ positive definite $k \times k$ matrix

$p(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\pi}) = (2\pi)^{-k/2}|\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{\mathbf{-1}}(\mathbf{y} - \boldsymbol{\mu})\right)$

$E(Y) = \boldsymbol{\mu}$

$\mathrm{Var}(Y) = \boldsymbol{\Sigma}$

The Uniform Distribution

$Y \sim \text{Uniform}(\alpha, \beta)$

$y \in [\alpha, \beta]$

Interval: $[\alpha, \beta]$; $\beta > \alpha$

$p(y|\alpha, \beta) = \frac{1}{\beta - \alpha}$

$E(Y) = \frac{\alpha + \beta}{2}$

$\text{Var}(Y) = \frac{(\beta - \alpha)^2}{12}$

The Beta Distribution

$Y \sim \text{Beta}(\alpha, \beta)$

$y \in [0, 1]$

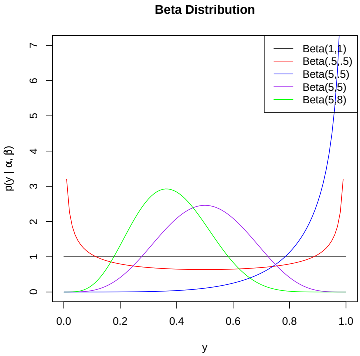
shape parameters:

$\alpha > 0; \quad \beta > 0$

$p(y|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{(\alpha-1)}(1-y)^{(\beta-1)}$

$E(Y) = \frac{\alpha}{\alpha+\beta}$

$\text{Var}(Y) = \frac{\alpha\beta}{(\alpha+\beta)^2)\alpha+\beta+1)}$



The Gamma Distribution

$Y \sim \text{Gamma}(\alpha, \beta)$

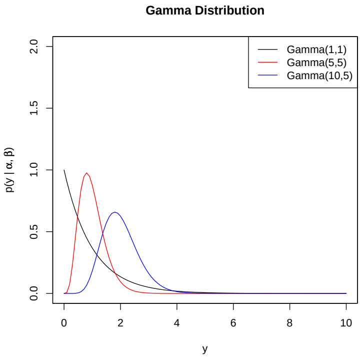
$y > 0$

shape parameter: $\alpha > 0$
inverse scale parameter: $\beta > 0$

$p(y|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{(\alpha-1)} \exp\left(-\beta y\right)$

$E(Y) = \frac{\alpha}{\beta}$

$\text{Var}(Y) = \frac{\alpha}{\beta^2}$



The Inverse Gamma Distribution

Distribution of the Inverse of a Gamma Distribution: If $X \sim \text{Gamma}(\alpha, \beta)$, then $\frac{1}{X} \sim \text{Invgamma}(\alpha, \beta)$.

$Y \sim \text{Invgamma}(\alpha, \beta)$

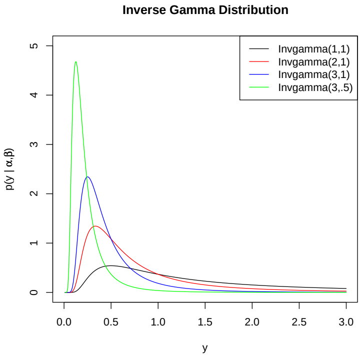
$y > 0$

shape parameter: $\alpha > 0$
scale parameter: $\beta > 0$

$$p(y|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{-(\alpha+1)} e^{-\frac{\beta}{y}}$$

$$E(Y) = \frac{\beta}{\alpha - 1} \text{ for } \alpha > 1$$

$$\text{Var}(Y) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)} \text{ for } \alpha > 2$$



The Dirichlet Distribution

$Y \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_k)$

$y_j \in [0, 1]; \;\; \sum_{j=1}^{k} y_j = 1$

α parameters: $\alpha_j > 0; \;\; \sum_{j=1}^{k} \alpha_j \equiv \alpha_0$

$p(\mathbf{y}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_k)}{\Gamma(\alpha_1) \ldots \Gamma(\alpha_k)} y_1^{\alpha_1 - 1} \ldots y_k^{\alpha_k - 1}$

$E(Y_j) = \frac{\alpha_j}{\alpha_0}$

$\text{Var}(Y_j) = \frac{\alpha_j(\alpha_0 - \alpha_j)}{\alpha_0^2(\alpha_0 + 1)}$

$\text{Cov}(Y_i, Y_j) = -\frac{\alpha_i \alpha_j}{\alpha_0^2(\alpha_0 + 1)}$