LECTURE 2B: PROBABILITY DISTRIBUTIONS (2)

--- R programming for Biostatistics and Bioinformatics

OUTLINE

- Exponential Distribution
- Gamma Distribution
- Beta Distribution

- Hands-on Exercises

EXPONENTIAL DISTRIBUTION (1)

- Continuous distribution for X ≥ 0
- Given X ~ Exp(λ), then the pdf for X

$$f(x;\lambda) = \lambda e^{-\lambda x}$$

- the cdf for X

$$\mu = \frac{1}{\lambda} \qquad \sigma^2 = \frac{1}{\lambda^2}$$

$$F(x;\lambda) = 1 - e^{-\lambda x}$$

EXPONENTIAL DISTRIBUTION (2)

```
mydexp <- function(x, lambda) {
   lambda * exp(-lambda*x)
}
>>> x <- seq(0, 10, by=0.1)
>>> y1=mydexp(x,2)
>>> y2=mydexp(x,0.5)
>>> plot(x,y1,type='l', col='red')
>>> lines(x,y2, col='blue')
>>> text(locator(2),  col=c('red', 'blue'),
   c(expression(lambda==2),expression(lambda==0.
   5)))
```

PROPERTIES

- Let $X_1, X_2, \ldots, X_n \sim \text{Exp}(\lambda)$ are i.i.d, then

$$\sum_i \boldsymbol{X}_i \sim \boldsymbol{\Gamma}(\boldsymbol{n}, \lambda)$$

- If $X \sim \text{Exp}(\lambda)$ and $Y = \exp(-\lambda X)$, then

$$\boldsymbol{Y} \sim \boldsymbol{Uniform}(\mathbf{0},\mathbf{1})$$

- Use histogram to test the above two properties.

GAMMA DISTRIBUTION (1)

- Gamma distribution is usually used for modeling highly skewed parameters, especially in the context of Bayesian statistics

- If X~Gamma(α), then the probability density function for X is

$$f(x;\alpha)=\frac{1}{\Gamma(\alpha)}x^{\alpha-1}e^{-x}$$

GAMMA DISTRIBUTION (2)

- If X~Gamma(α,ß), then the probability density function for X is

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

$$\mu=\alpha\beta \qquad \sigma^2=\alpha\beta^2$$

BETA DISTRIBUTION (1)

- The Beta density function is a very versatile way to represent outcomes like proportions or probabilities.
- The standard beta distribution:
  - pdf

$$f(x;\alpha,\beta)=\frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$$

$$B(\alpha,\beta)=\int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

PROPERTIES

- Expected value: $$\mu = \frac{\alpha}{\alpha + \beta}$$

- Variance: $$\sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{\beta}{(\alpha+\beta)(\alpha+\beta+1)}\mu$$

EXERCISE 1
HIDDEN MARKOV CHAIN (HMM)

- Initial State: begin=1
- State Space: s = {1,2}
- Observation space: o = {“a”,”b”,”c”}
- State Transition Matrix:
  - t = matrix(c(0.7,0.4,0.3,0.6), ncol=2, nrow=2)
  - t[1,2] denotes the transition probability from state 1 to state 2
- Emission Matrix:
  - e = matrix(c(0.3, 0.4, 0.3,0.5, 0.1, 0.4),ncol=3,nrow=2,byrow=T)
  - e[1,3] is the conditional probability of observation c under the state 1

EXERCISE 1
THE TASK FOR HMM

- Simulate 1 Markov chain of the states with length n=1000 based on the above given information
- Summarize the frequencies of the state 1, 2 in the Markov Chain
- Simulate 10 observation based on the above Markov chain and the given emission matrix

EXERCISE 2
PROOF FOR GAMMA FUNCTION

- Here is the Gamma function

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt$$

- Prove that

$$\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$$

$$\int_0^\infty \frac{1}{\Gamma(\alpha)} t^{\alpha-1} e^{-t} dt = 1$$

EXERCISE 3
PROOF FOR GAMMA DISTRIBUTION

- If $X \sim Gamma(\alpha, \beta)$ and $Y = \beta X$, prove that $Y \sim Gamma(\alpha)$

- In the density function for Gamma distribution, α is called the shape parameter; and β is called the inverse scale parameter, also known as rate parameter.

- Gamma(α) is named the basic Gamma distribution, which is the special form for Gamma(α,β) when $\beta=1$.

EXERCISE 4
GAMMA DENSITY FUNCTION

- Plot the density curves in a single diagram for Gamma distribution when (dgamma())
  a) α=2, β=4
  b) α=4, β=4
  c) α=2, β=1
  d) α=4, β=1

  or

  a) α=1, β=1
  b) α=1, β=2
  c) α=1, β=4

EXERCISE 5
BETA DISTRIBUTION

- If X1 and X2 are independently distributed and X1 ~ Gamma(α1,β), X2 ~ Gamma(α2,β), then X1/(X1+X2) ~ Beta(α1, α2)
- Write a function to simulate a Beta sample with two parameters (α,β respectively) using the rgamma() function (sample size n=100).

- Plot the density curve of beta distribution when (dbeta())
  a) α=1, β=1
  b) α=1, β=1.5
  c) α=1, β=2
  d) α=1, β=4
  e) α=2, β=1
  f) α=4, β=1
  g) α=1.5, β=1

EXERCISE 5 (CONT.)

- We have two prior distributions for modeling the probability that a specific event occurs, Beta(10,10) and Beta(1000,1000), which one lends you more confidence that the probability p=0.5? Why?

EXERCISE 6
DIRICHLET DISTRIBUTION

- If X_i ~ Gamma(α_i,ß), then (X_i/sum(X)) ~ Dirichlet(α_i)
- Write a function to simulate a Dirichlet sample with 4 parameters (10,10,10,10 respectively) using the rgamma() function (sample size n=100).

  - For the given two distributions Dirichlet(1,1,1,1) or Dirichlet(1000,1000,1000,1000), which is more confident that the four numbers are equal? why?

SOLUTION

```
rdirichlet <- function (n, alpha)
# n: number of samples
# alpha: a vector, the parameters for each gamma distribution
{
    p <- length(alpha)
    x <- matrix(rgamma(p * n, alpha), ncol = p, byrow = TRUE)
    sm <- x %*% rep(1, p)
    x/as.vector(sm)
}
```

EXERCISE 7
BAYESIAN STATISTICS FOR FLIPPING COINS

- Experiment: Coin tossing
- Prior knowledge: the chance of getting head p ~ beta(10, 10)
- We tossed the coin 100 times, and get 45 heads
- Estimate the posterior distribution for p using Bayesian theorem

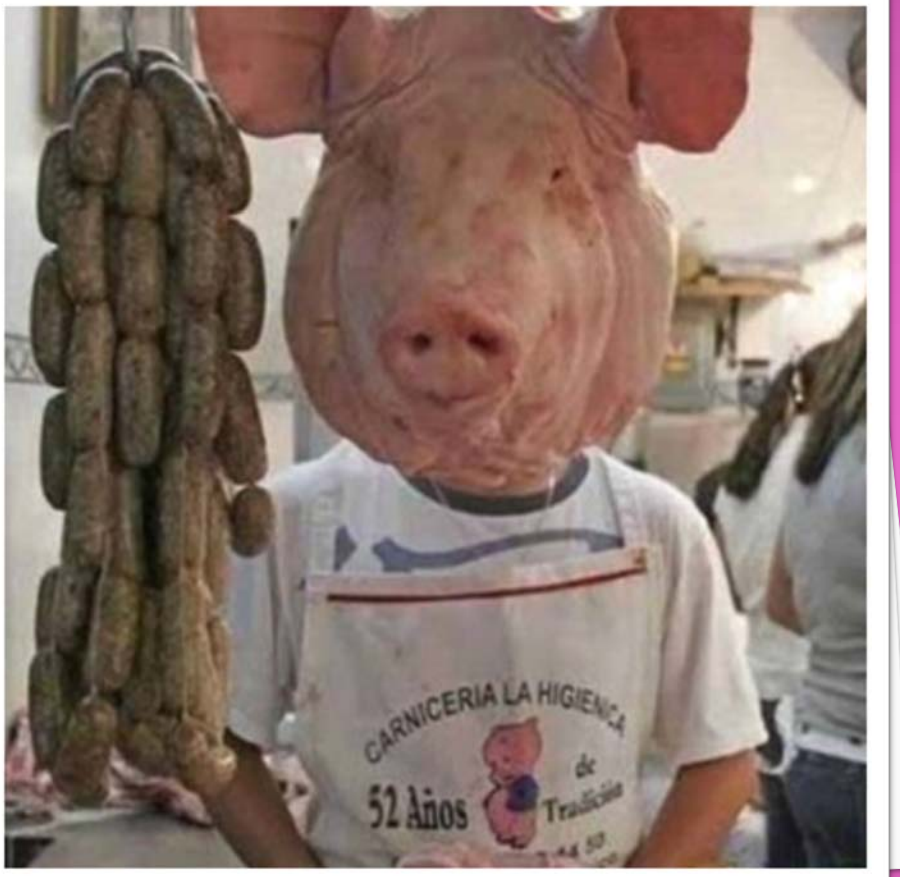
$$p(\theta_i \mid D) = \frac{p(\theta_i) L(D \mid \theta_i)}{\sum_{j \in J} p(\theta_j) L(D \mid \theta_j)}$$

EXERCISE 8 RANDOM SEQUENCE

- Create a function that generates a single random nucleotide *X* where P(*X* = “G”) = 0.30, P(*X* = “A”) = 0.20, P(*X* = “C”) = 0.25, and P(*X* = “T”) = 0.25

  Hint: You may want to use the runif() function to do this.
- Using the function you have created, create another function that generates a random nucleotide sequence of length *n.*
- Generate a random nucleotide sequence of length 100 using the sample() function, where the probability of each nucleotide is given as above.

  Hint: type ‘?sample’ for more information.



THAT'S ALL, THANKS

TOP TEN REASONS TO BE A STATISTICIAN

- Estimating parameters is easier than dealing with real life.
- Statisticians are significant
- I always wanted to learn the entire Greek alphabet.
- The probability a statistician major will get a job is > .9999.
- If I flunk out I can always transfer to Engineering.
- We do it with confidence, frequency, and variability.
- You never have to be right - only close.
- We're normal and everyone else is skewed.
- The regression line looks better than the unemployment line.
- No one knows what we do so we are always right.

STATISTICS HAS THE ADVANTAGE THAT YOU NEVER HAVE TO SAY THAT YOU ARE CERTAIN

Three professors (a physicist, a chemist, and a statistician) are called in to see their dean. Just as they arrive the dean is called out of his office, leaving the three professors there. The professors see with alarm that there is a fire in the wastebasket. The physicist says, "I know what to do! We must cool down the materials until their temperature is lower than the ignition temperature and then the fire will go out.“

The chemist says, "No! No! I know what to do! We must cut off the supply of oxygen so that the fire will go out due to lack of one of the reactants."

While the physicist and chemist debate what course to take, they both are alarmed to see the statistician running around the room starting other fires. They both scream, "What are you doing?"

To which the statistician replies, "Trying to get an adequate sample size."

ANOTHER JOKE

I read that there is about one chance in one million that someone will board an airplane carrying a bomb, and I started carrying a bomb with me on every flight I take. The way I figure it, the odds against two people having a bomb on the same plane are 1 in a trillion.