Probability with Continuous Random Variables

Ben Goodrich

February 05, 2019

Obligatory Disclosure

- Ben is an employee of Columbia University, which has received several research grants to develop Stan
- Ben is also a manager of GG Statistics LLC, which uses Stan for business purposes
- According to Columbia University policy, any such employee who has any equity stake in, a title (such as officer or director) with, or is expected to earn at least $\$5,000.00$ per year from a private company is required to disclose these facts in presentations
- But you should install the C++ toolchain and RStan R package anyway; see https://github.com/stan-dev/rstan/wiki/RStan-Getting-Started

Hypergeometric Confusion

- The hypergeometric distribution corresponds to sampling WITHOUT replacement and has PMF $\Pr\left(x \middle| N, K, n\right) = \binom{K}{x}\binom{N-K}{n-x}\binom{N}{n}^{-1}$ where
  - N is the (finite) size of the set being drawn from
  - K is the number of successes in that finite set
  - n is the number of times you draw without replacement
- The `dhyper` function in R parameterizes the hypergeometric PMF differently:
  - `x` is the number of successes sought in `k` draws
  - `m` is the number of successes in the set (deck, urn, etc.)
  - `n` is the number of failures in the set (deck, urn, etc.)
  - `k` is the number of times you draw from the set
- The probability of being dealt, for example, two tens from a deck is just `dhyper(x = 2, m = 4, n = 52 - 4, k = 2)` ≈ 0.004525

Probability and Cumulative Mass Functions

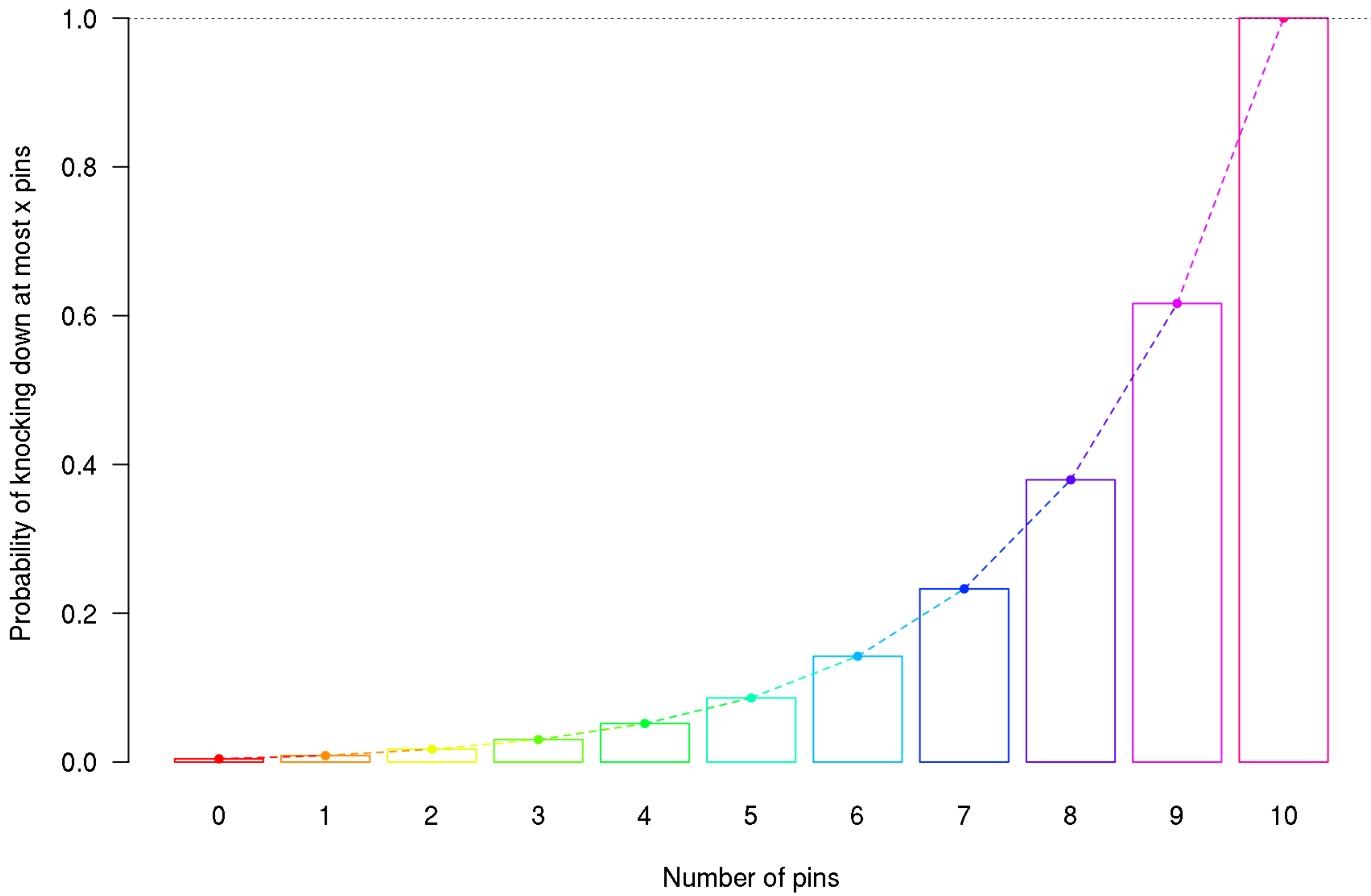
- $\Pr\left(x \mid \boldsymbol{\theta}\right)$ is a Probability Mass Function (PMF) over a discrete Ω that may depend on some parameter(s) $\boldsymbol{\theta}$ and the Cumulative Mass Function (CMF) is $\Pr\left(X \leq x \mid \boldsymbol{\theta}\right) = \sum\limits_{i = \min\{\Omega\}}^{x} \Pr\left(i \mid \boldsymbol{\theta}\right)$
- In the first roll of bowling, some simplification implies $\Pr\left(X \leq x\right) = \frac{-1 + \mathcal{F}_{x+2}}{-1 + \mathcal{F}_{n+2}}$

```
source("https://tinyurl.com/y93srfmp") # code from week 1 to define F() and Omega
CMF <- function(x, n = 10) return( (- 1 + F(x + 2)) / (- 1 + F(n + 2)) )
round(CMF(Omega), digits = 4)
```

```
##      0      1      2      3      4      5      6      7      8      9     10
## 0.0043 0.0086 0.0172 0.0302 0.0517 0.0862 0.1422 0.2328 0.3793 0.6164 1.0000
```

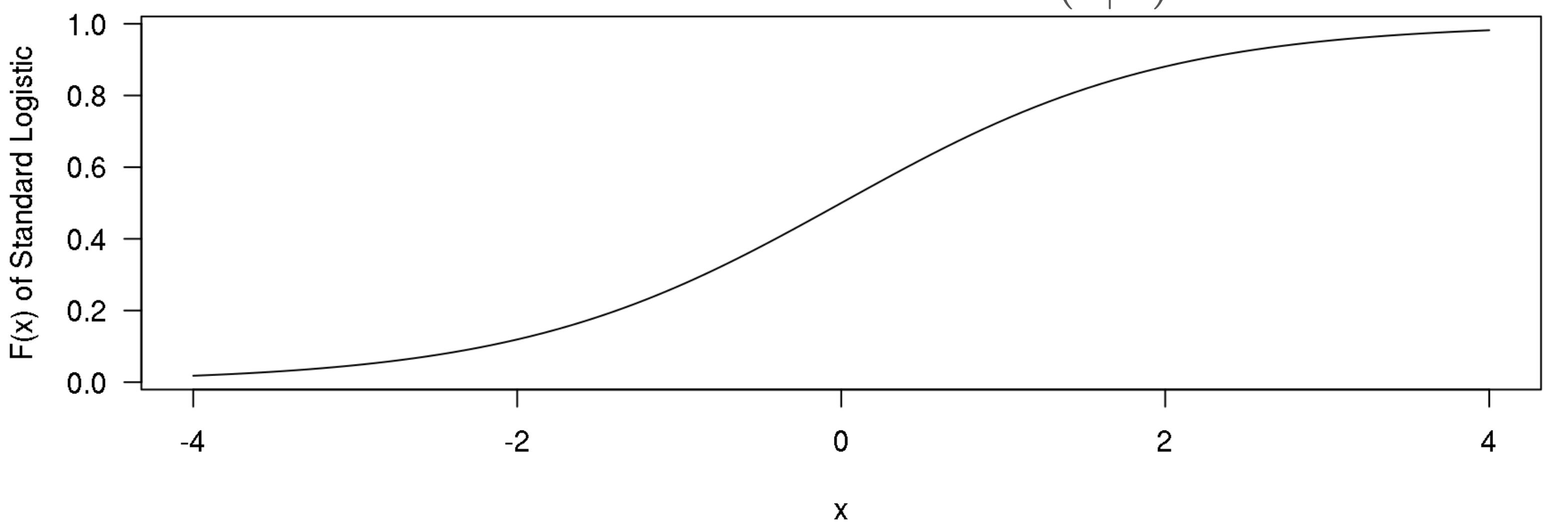
- How do we know this CMF corresponds to our PMF $\Pr\left(x \mid n\right) = \frac{\mathcal{F}_x}{-1 + \mathcal{F}_{n+2}}$?

PMF is the Rate of Change in the CMF



Cumulative Density Functions

- Now Ω is an interval; e.g. $\Omega = \mathbb{R}$, $\Omega = \mathbb{R}_+$, $\Omega = (a, b)$, etc.
- Ω has an infinite number of points, so $\Pr(X = x) \downarrow 0$
- $\Pr(X \leq x)$ is called the Cumulative Density Function (CDF) from Ω to $(0, 1)$
- No conceptual difference between a CMF and a CDF except emphasis on whether Ω is discrete or continuous so we use $F(x|\boldsymbol{\theta})$ for both



The Standard Logistic CDF and PDF

- E.g., CDF of the standard logistic distribution over $\Omega = \mathbb{R}$ is $F(x) = \frac{1}{1+e^{-x}}$
- $\Pr(a < X \leq x) = F(x|\,\boldsymbol{\theta}) - F(a|\,\boldsymbol{\theta})$ as in the discrete case
- If $x = a + h$, $\frac{F(x|\boldsymbol{\theta}) - F(a|\boldsymbol{\theta})}{x-a} = \frac{F(a+h|\boldsymbol{\theta}) - F(a|\boldsymbol{\theta})}{h}$ is the slope of a line segment
- If we then let $h \downarrow 0$, $\frac{F(a+h|\boldsymbol{\theta}) - F(a|\boldsymbol{\theta})}{h} \rightarrow \frac{\partial F(a|\boldsymbol{\theta})}{\partial a} \equiv f(x|\,\boldsymbol{\theta})$ is still the RATE OF CHANGE in $F(x|\,\boldsymbol{\theta})$ at x
- The derivative of the CDF $F(x)$ is called the Probability Density Function (PDF) and denoted $f(x)$, which is always positive because the CDF increases
- $f(x)$ is NOT a probability but is used like a PMF
- What is slope of $F(x) = \frac{1}{1+e^{-x}}$ at x?
- Answer: $\frac{\partial}{\partial x} F(x) = \frac{-1}{(1+e^{-x})^2} \times \frac{\partial}{\partial x} e^{-x} = \frac{-e^{-x}}{(1+e^{-x})^2} \times \frac{\partial - x}{\partial x} = \frac{e^{-x}}{(1+e^{-x})^2} \geq 0$

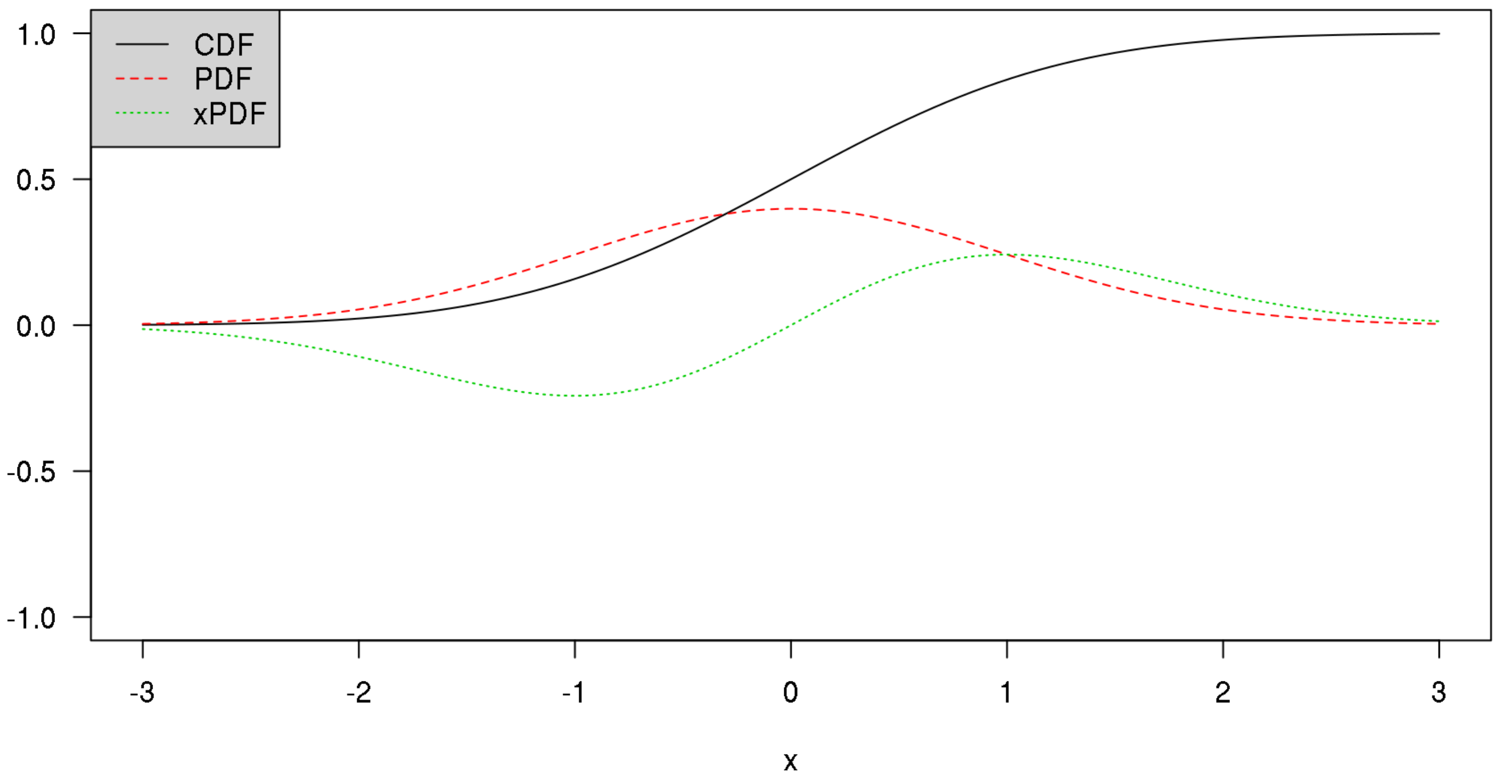
The Standard Normal CDF and Its Slope

Standard normal CDF over $\Omega = \mathbb{R}$ is $\Phi(x) = \frac{1}{2} + \phi(x)\, S(x)$ where $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$, $S(x) = \sum_{n=0}^{\infty} \frac{x^{2n+1}}{(2n+1)!!}$, and $a!!$ is the “double factorial” function for a non-negative integer a such that $0!! = 1$, $1!! = 1$, and else $a!! = a \times (a-2)!!$. What is the slope of $\Phi(x)$ at x?

- $$\frac{\partial}{\partial x}\Phi(x) = \phi(x)\, S'(x) + \phi'(x)\, S(x) = \phi(x) \sum_{n=0}^{\infty} \frac{(2n+1)\, x^{2n}}{(2n+1)!!} - \phi(x)\, x \sum_{n=0}^{\infty} \frac{x^{2n+1}}{(2n+1)!!} = \phi(x) \sum_{n=0}^{\infty} \frac{(2n+1)\, x^{2n} - x^{2n+2}}{(2n+1)!!} = \phi(x) \left(\frac{1-x^2}{1} + \frac{3x^2 - x^4}{3 \times 1} + \frac{5x^4 - x^6}{5 \times 3 \times 1} \ldots \right) = \phi(x)$$

CDF and PDF of the Standard Normal Distribution

```
curve(pnorm(x), from = -3, to = 3, ylim = c(-1,1), ylab = "") # CDF; what is the median?
curve(dnorm(x), add = TRUE, col = "red", lty = "dashed")      # PDF; what is the mode?
curve(x * dnorm(x), add = TRUE, col = 3, lty = "dotted")      # g function being integrated
legend("topleft", legend = c("CDF", "PDF", "xPDF"), col = 1:3, lty = 1:3, bg = "lightgrey")
```



Expectations of Functions of a Continuous RV

- Let $g\left(X\right)$ be a function of a continuous $X \in \Omega$
- The probability that X is in the interval $\left[x, x + dx\right]$ is $f\left(x|\boldsymbol{\theta}\right)dx$ where dx is essentially the smallest non-neglible piece of X
- The expectation of $g\left(X\right)$, if it exists (which it may not), is defined as

$$\mathbb{E}g\left(X\right) = \int_{\min\Omega}^{\max\Omega} g\left(x\right)f\left(x|\boldsymbol{\theta}\right)dx = G\left(\boldsymbol{\theta}\right)\Big|_{x=\min\Omega}^{x=\max\Omega}$$

- Integrals are usually impossible but we can use simulations to approximate them arbitrarily well. Still need to understand integrals conceptually as area.
- Columbia students can download Mathematica for free
- If $g\left(X\right) = X$, $\mathbb{E}X = \mu$ is "the" expectation and if $g\left(X\right) = \left(X - \mu\right)^2$, $\mathbb{E}\left[\left(X - \mu\right)^2\right] = \mathbb{E}\left[X^2\right] - \mu^2 = \sigma^2$ is the variance

Moments of a Standard Normal Distribution

- Note that the Standard Normal PDF only depends on the square of x, so

$$\mu = \int_{-\infty}^{\infty} x\phi(x)\,dx = \int_{-\infty}^{0} x\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\,dx + \int_{0}^{\infty} x\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\,dx = 0$$

- Let $y = \frac{x^2}{2}$ so that $\sqrt{2y} = x$ and $dy = xdx$. Then, the variance is given by

$$\sigma^2 = \int_{-\infty}^{\infty} (x-0)^2\phi(x)\,dx = 2\int_{0}^{\infty} x\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}xdx =$$

$$\frac{2}{\sqrt{2\pi}}\int_{0}^{\infty} \sqrt{2y}e^{-y}dy = \frac{2}{\sqrt{\pi}}\int_{0}^{\infty} y^{\frac{3}{2}-1}e^{-y}dy = \frac{2}{\sqrt{\pi}}\Gamma\left(\frac{3}{2}\right) = 1$$

- $\Gamma(z) = \int_0^{\infty} y^{z-1}e^{-y}dy$ is a very important special function that is a continuous generalization of $(z+1)!$ and is implemented as `gamma(z)` in R

Shift and Scale Transformations

- If Z is distributed standard normal & $X\left(Z\right)=\mu+\sigma Z$, what's the PDF of X?
- Answer: Note that $Z\left(X\right)=\frac{X-\mu}{\sigma}$. Since this is a monotonic transformation

$$\Pr\left(X\leq x\right)=\Pr\left(Z\leq z\left(x\right)\right)=\Phi\left(z\left(x\right)\right)$$

$$\frac{\partial}{\partial x}\Phi\left(z\left(x\right)\right)=\frac{\partial\Phi\left(z\right)}{\partial z}\times\frac{\partial z\left(x\right)}{\partial x}=\phi\left(z\left(x\right)\right)\frac{1}{\sigma}=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^{2}}$$

- $\mathbb{E}X=\mu+\sigma\mathbb{E}Z=\mu$ and $\mathbb{E}\left[\left(X-\mu\right)^{2}\right]=\mathbb{E}\left[\left(\sigma Z\right)^{2}\right]=\sigma^{2}\mathbb{E}\left[Z^{2}\right]=\sigma^{2}$
- Thus, $f\left(x\right|\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^{2}}$ is the PDF of the general normal distribution with expectation μ and standard deviation σ as parameters
- The normal distribution is one of several in the location-scale family, where such transformations only change the location and scale of the distribution

Nonlinear but Monotonic Transformations

- If Z is distributed normal with expectation μ and standard deviation σ and $X\left(Z\right) = e^{Z}$, what is the PDF of X? Hint: $\Pr\left(X \leq x\right) = \Pr\left(Z \leq z\left(x\right)\right)$
- Answer: Note that $Z\left(X\right) = \ln X$ and $\frac{\partial}{\partial x} z\left(x\right) = \frac{1}{x}$ so $f_X\left(x \middle| \mu, \sigma\right) = f_Z\left(z\left(x\right) \middle| \mu, \sigma\right) \times \frac{\partial}{\partial x} z\left(x\right) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\ln(x)-\mu}{\sigma}\right)^2}$ is the PDF of the lognormal distribution over $\Omega = \mathbb{R}_+$
- μ and σ are parameters but NOT the expectation and standard deviation of X, due to the nonlinearity of the antilog transformation. It can be shown that $\mathbb{E}X = e^{\mu + \frac{1}{2}\sigma^2}$ and $\mathrm{Var}\left(X\right) = \left(-1 + e^{\sigma^2}\right) e^{2\mu + \sigma^2}$.

Scale Transformations of Exponential Variates

- Standard exponential distribution over $\Omega = \mathbb{R}_+$ has CDF $F(x) = 1 - e^{-x}$
- Its PDF is obviously $f(x) = \frac{\partial}{\partial x} F(x) = e^{-x}$, which must integrate to 1
- $\mathbb{E}X = \int_0^\infty x e^{-x} dx = -(x+1) e^{-x} \Big|_{x=0}^{x \to \infty} \to -\infty e^{-\infty} + e^0 \to 1$
- What is $\mathrm{Var}(X)$?
- $\int_0^\infty (x-1)^2 e^{-x} dx = \int_0^\infty x^2 e^{-x} dx - 2 \int_0^\infty x e^{-x} dx + \int_0^\infty e^{-x} dx =$
  $-\left(x^2 + 2x + 2\right) e^{-x} \Big|_{x=0}^{x \to \infty} - 2 \times 1 + 1 \to 1$
- If X is distributed standard exponential and $Y = \mu X$, what is the PDF of Y?
- Answer: $\Pr(X \leq x) = \Pr(Y \leq y(x))$, so $f(y|\mu) = \frac{\partial 1 - e^{-\frac{y}{\mu}}}{\partial y} = \frac{1}{\mu} e^{-\frac{y}{\mu}}$
- You will often see this with the substitution $\lambda = \frac{1}{\mu}$. What are $\mathbb{E}Y$ & $\mathrm{Var}(Y)$?

Bivariate Normal Distribution

If $\Pr\left(X \leq x \bigcap Y \leq y \middle| \boldsymbol{\theta}\right) = F\left(x, y \middle| \boldsymbol{\theta}\right)$ is a biviariate CDF, then the bivariate PDF is $\frac{\partial^2}{\partial x \partial y} F\left(x, y \middle| \boldsymbol{\theta}\right)$. This generalizes beyond two dimensions. The PDF of the bivariate normal distribution over $\Omega = \mathbb{R}^2$ has five parameters:

$$f\left(x, y \middle| \mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho\right) =$$

$$\frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} e^{-\frac{1}{2\left(1-\rho^2\right)}\left(\left(\frac{x-\mu_X}{\sigma_X}\right)^2 + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2 - 2\rho\frac{x-\mu_X}{\sigma_X}\frac{y-\mu_Y}{\sigma_Y}\right)} =$$

$$\frac{1}{\sigma_X\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu_X}{\sigma_X}\right)^2} \times \frac{1}{\color{blue}{\sigma_Y\sqrt{1-\rho^2}}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-\left(\color{red}{\mu_y + \frac{\sigma_X}{\sigma_Y}\rho\left(x-\mu_x\right)}\right)}{\color{blue}{\sigma_Y\sqrt{1-\rho^2}}}\right)^2},$$

where the first term is a marginal normal PDF and the second is a conditional normal PDF w/ parameters $\color{red}{\mu = \mu_Y + \frac{\sigma_X}{\sigma_Y}\rho\left(x - \mu_x\right)}$ & $\color{blue}{\sigma = \sigma_Y\sqrt{1-\rho^2}}$.

Where Is this Class Going?

- We usually think of parameters as being continuous but contained in a parameter (sub)space Θ
- If you cut a continuous RV into a categorical RV, you could apply Bayes Rule
- If you take the limit as the number of cuts $\uparrow \infty$ you get Bayes Rule for continuous random variables

$$f\left(\theta \mid y, a, b\right) = \frac{f\left(\theta \mid a, b\right) f\left(y \mid \theta\right)}{f\left(y \mid a, b\right)}$$

- Iff you have data, y, then $L\left(\theta; y\right)$ is the same expression as $f\left(y \mid \theta\right)$ but is a mathematical function of θ called the likelihood function that can be evaluated at any $\theta \in \Theta$ but only at the OBSERVED y
- By choosing functions for the numerator, you can (in principle) work out what $f\left(y \mid a, b\right) = \int_\Theta f\left(\theta \mid a, b\right) L\left(\theta; y\right) d\theta$ evaluates to complete Bayes Rule