PhD Bootcamp Day 3: Distributions and Inference

Alex Dombowsky and Jennifer Kampe

August 13, 2021

Duke University

Hello!

Welcome to the department! Today's bootcamp session is structured as follows:

- Basic distribution theory review.
- A review of concepts that you should be comfortable with before starting classes.
- A list of review exercises that *you should do* to warm-up your stats knowledge before the school year begins.

Cumulative Distribution Functions

- The distribution of a real-valued random-variable X is defined by its **cumulative distribution function**, $F_X(x) = P(X \leq x)$. The CDF is right continuous, non-decreasing, and has limits 0 and 1 as x tends to $-$ or $+\infty$.
- The CDF is usually represented as an integral over another function, so that

$$F_X(x) = \int_{-\infty}^{x} dF_X(t).$$

Probability Density and Mass Functions

- While the dF_X notation may be unfamilar, it is defined as

$$\int_{-\infty}^{x} dF_X(t) = \begin{cases} \int_{-\infty}^{x} f_X(t)dt : & \text{continuous rv} \\ \sum_{t=-\infty}^{x} f_X(t) : & \text{discrete rv} \end{cases}.$$

- $f_X(x)$ is the **probability mass** (discrete) or **density** (continuous) function. It is often convenient to write

$$f_X(x) = \frac{h(x)}{c}$$

for **kernel** h and **normalizing constant** $c < \infty$. We assume $h(x) \geq 0$ and

$$c = \begin{cases} \int_{-\infty}^{\infty} h(x)dx : & \text{continuous rv} \\ \sum_{x=-\infty}^{\infty} h(x) : & \text{discrete rv} \end{cases}.$$

Multivariate Random Variables

- Random variables $\boldsymbol{X}$ can be defined on $\mathbb{R}^m$. The multivariate CDF is

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_m \leq x_m)$$

$$= \begin{cases} \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_m} f_{\boldsymbol{X}}(\boldsymbol{t}) dt_1 \cdots dt_m : & \text{continuous rvs} \\ \sum_{t_1=-\infty}^{x_1} \cdots \sum_{t_m=-\infty}^{x_m} f_{\boldsymbol{X}}(\boldsymbol{t}) : & \text{discrete rvs} \end{cases}.$$

- This can be generalized to random vectors consisting of discrete and continuous rvs. To recover the PDF/PMF of, say, X_1, we merely integrate out all other variables:

$$f_{X_1}(x_1) = \begin{cases} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\boldsymbol{X}}(x_1, t_2, \ldots, t_m) dt_2 \cdots dt_m : & \text{cont.} \\ \sum_{t_2=-\infty}^{\infty} \cdots \sum_{t_m=-\infty}^{\infty} f_{\boldsymbol{X}}(x_1, t_2, \ldots, t_m) : & \text{disc.} \end{cases}.$$

Independence and Covariance

- Two random variables are **independent** if their joint density/mass function factorizes into the product of their marginal distributions, i.e.

$$f_{X,Y}(x,y) = f_X(x) f_Y(y) \, \forall \, x, y.$$

- The **covariance** of two random variables is

$$\text{Cov}(X,Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

- Covariance says something about the relationship between X and Y (note that the outer expectation is with respect to their *joint* distribution).

Does covariance tell us anything about independence?

- We can roughly describe how two random variables affect each other with **correlation**:
$$\rho_{XY} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} \in [-1,1].$$
- $\rho_{XY} > 0$ implies X and Y are *positively correlated*, ie. an increase in X *tends to result* in an increase in Y (and X and Y are dependent).
- $\rho_{XY} < 0$ implies X and Y are *negatively correlated*, ie. an increase in X *tends to result* in an decrease in Y (and X and Y are dependent).
- What about $\rho_{XY} = 0$?

Zero Correlation

- **The problem with correlation**: it describes (approximately) linear relationships.
- In a sense, ρ_{XY} may be interepreted as the sign of a in a linear equation $Y = aX + \epsilon$.
- But what if the relationship between X and Y is *not linear* (eg. quadratic, cubic, sinusoidal, step functions, etc.).
- As it turns out,

$$\rho_{XY} = 0 \not\Rightarrow X \text{ and } Y \text{ are independent.}$$

- So, correlation can only tell us about the dependence structure if it is non-zero.

Conditional Distributions

- The conditional PDF/PMF of $X \mid Y = y$ is

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{f_{X,Y}(x,y)}{\int f_{X,Y}(x,y)dx}.$$

- **Bayes theorem** gives us a way to do "backward conditioning"

$$f_{Y|X}(y \mid x) = \frac{f_{X|Y}(x \mid y) f_Y(y)}{f_X(x)} = \frac{f_{X|Y}(x \mid y) f_Y(y)}{\int f_{X|Y}(x \mid y) f_Y(y) dy}.$$

- Note that the denominator does not depend on y.

Conditional Expectations

- The conditional expectation of $X \mid Y = y$ is

$$\mathbb{E}[X \mid Y = y] = \begin{cases} \int_{-\infty}^{\infty} x f_{X|Y}(x \mid y) dx \\ \sum_{x=-\infty}^{\infty} x f_{X|Y}(x \mid y) \end{cases}$$

  and will be a function of y.
- As such, we can define the random variable $\mathbb{E}[X \mid Y]$.
- **Law of Total Expectation**:

$$\mathbb{E}[\mathbb{E}[X \mid Y]] = \mathbb{E}[X].$$

- **Law of Total Variance**:

$$\mathbb{E}[\mathrm{Var}(X \mid Y)] + \mathrm{Var}(\mathbb{E}[X \mid Y]) = \mathrm{Var}(X).$$

Example: Bivariate Normal Distribution

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{bmatrix}\right)$$

- Describes a two-dimensional vector that takes values in $\mathbb{R}^2$.
- $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$.
- $\mathrm{Cov}(X, Y) = \sigma_{XY}$.
- $X \mid Y$ and $Y \mid X$ are also normal.
- For any $a, b \in \mathbb{R}$,

$$aX + bY \sim \mathcal{N}(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\sigma_{XY}).$$

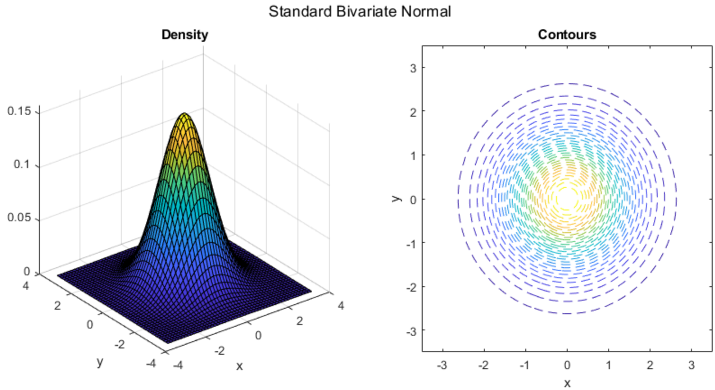

Figure 1: Density and contours of the standard bivariate normal distribution.

Example: Gamma Distribution

A positive random variable $X \sim \text{Gamma}(\alpha, \beta)$ with PDF:

$$f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-x\beta};$$
$$E[X] = \frac{\alpha}{\beta};$$
$$V[X] = \frac{\alpha}{\beta^2};$$

where $x \in (0, \infty)$, $\alpha, \beta > 0$.
Note: this is referred to as the shape-rate parameterization.
You may also see the shape-scale parameterization with scale $\theta = 1/\beta$

Gamma Distribution - Important Properties

Here are some properties that will come in handy throughout the first year:

- If $X \sim \text{Gamma}(\alpha, \beta)$ with $\alpha = 1$, $X \sim \text{Exponential}(\lambda = \beta)$
- If $X \sim \text{Gamma}(v/2, 1/2)$, then $X \sim \chi^2_v$
- If $X \sim \text{Gamma}(\alpha_1, \beta)$ and $Y \sim \text{Gamma}(\alpha_2, \beta)$, then $X + Y \sim \text{Gamma}(\alpha_1 + \alpha_2, \beta)$
- If $X \sim \text{Gamma}(\alpha, \beta)$ (shape-rate parameterization), then $1/X \sim \text{Inverse Gamma}(\alpha, \beta)$ with expectation $\frac{\beta}{\alpha-1}$
- If $X \sim \text{Gamma}(\alpha, \theta)$ (shape-scale parameterization), then $1/X \sim \text{Inverse Gamma}(\alpha, 1/\theta)$ with expectation $\frac{\beta}{\alpha-1}$
- If $X \sim \text{Gamma}(\alpha, \beta)$, then $X/n \sim \text{Gamma}(\alpha, n\beta)$

Miscellaneous Useful Facts about Distributions

- If $X_1, \cdots, X_n$ are iid with CDF $F(x)$, then $X_{(1)}$ has CDF $1-(1-F(x))^n$
- If $X_1, \cdots, X_n$ are iid with CDF $F(x)$, then $X_{(n)}$ has CDF $F(x)^n$
- If $X_1, \cdots, X_n \overset{iid}{\sim}$ Poisson(λ), then $\sum_{i=1}^n X_i \sim$ Poisson$(n\lambda)$
- f $X_1, \cdots, X_n \overset{iid}{\sim}$ Exponential$(\lambda) \leftrightarrow$ Gamma$(1, \lambda)$, then $\sum_{i=1}^n X_i \sim$ Gamma(n, λ)
- If $\beta|\phi \sim N(m, \Sigma/\phi)$ and $\phi \sim$ Gamma$(v/2, v\sigma^2/2)$ then the marginal distribution of β is $t_v(m, \sigma^2\Sigma)$
- Mins, maxes, and CDF counts of random variables are binomial random variables

Change of Variables

Motivation: Let X be a real-valued random variable with pdf $f_X(x)$ and let $Y = g(X)$ for some one-to-one differentiable function $g(x)$. Then Y will also have a continuous distribution - what is it?

One Dimension: let $Y = g(X)$, g monotone with $X = g^{-1}(Y) = h(Y)$, then

$$X \sim f_X(x) \implies f_Y(y) = f_X(h(y))|dh/dy|$$

Change of Variables: d-Dimensions

Let $\boldsymbol{X} = (X_1, \cdots, X_{d1})$ be a collection of random variables with support $\mathbb{X}^{(d_1)}$ and joint pdf $f_{X_1,\cdots,X_{d_1}}$, and let

$$\boldsymbol{Y} = g(\boldsymbol{X}) \leftrightarrow (Y_1, \cdots, Y_{d_2}) = (g_1(\boldsymbol{X}), \cdots, g_{d_2}(\boldsymbol{X})),$$

where $g : \mathbb{X}^{d_1} \to \mathbb{R}^{d_2}$ and $h = g^{-1} : \mathbb{R}^{d_1} \to \mathbb{X}^{d_2}$
Then $\boldsymbol{Y}$ has joint pdf:

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = f_{\boldsymbol{X}}\left(h_1(\boldsymbol{Y}), \cdots, h_{d1}(\boldsymbol{Y})\right) \times |J(\boldsymbol{Y})|$$

Change of Variables: Step-by-Step

1. Note the set of transformation functions $g = (g_1, \cdots, g_{d_2})$:

$$Y_1 = g_1(X_1, \cdots, X_{d_1})$$
$$\vdots$$
$$Y_{d_2} = g_{d_2}(X_1, \cdots, X_{d_1})$$

2. Find the set of inverse functions, $h = g^{-1}(\boldsymbol{X})$:

$$X_1 = h_1(Y_1, \cdots, Y_{d_2})$$
$$\vdots$$
$$X_{d_1} = h_{d_1}(Y_1, \cdots, Y_{d_2})$$

3. Identify the joint support of the new variables, $\mathbb{Y}^{d_2}$

4. Compute the Jacobian of the inverse transformation $h(\boldsymbol{Y})$ in Step 2: form the matrix of partial derivatives and take its determinant.

$$D_y = \left[\frac{\partial x_i}{\partial y_j}\right]_{ij} = \begin{bmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \cdots & \frac{\partial x_1}{\partial y_{d_2}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial x_{d_1}}{\partial y_1} & \frac{\partial x_{d_1}}{\partial y_2} & \cdots & \frac{\partial x_{d_1}}{\partial y_{d_2}} \end{bmatrix}$$

Set $J(y_1, \cdots, y_{d_2}) = \det D_y$. Alternately, note $J(\boldsymbol{Y}) = \frac{1}{J(\boldsymbol{X})}$

5. The joint pdf of $(Y_1, \cdots, Y_{d_1})$ is

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = f_{\boldsymbol{X}}\left(h_1(\boldsymbol{Y}), \cdots, h_{d1}(\boldsymbol{Y})\right) \times |J(\boldsymbol{Y})|$$

What if g is not one-one?

Make it one-to-one! **For example**:

1. Let $g : \mathbb{R}^2 \to \mathbb{R}$ and suppose we know the distribution of (X_1, X_2) (and at least one of the marginal distributions).
2. Set up a one-to-one transformation:
$$Y_1 = g(X_1, X_2) \text{ and } Y_2 = X_1 \text{ (or } X_2)$$
and find the distribution of (Y_1, Y_2).
3. Then use marginalization:
$$f_{Y_1}(y_1) = \int_{-\infty}^{\infty} f_{Y_1,X_1}(y_1, x_1) dx_1 = \int_{-\infty}^{\infty} f_{Y_1,X_2}(y_1, x_2) dx_2.$$

Moment Generating Functions

For a random variable X, the **moment generating function** (MGF) is the real-valued function

$$M_X(t) = \mathbb{E}[e^{tX}]$$

for all $t \in \mathbb{R}$. If the MGF is finite for an open interval around 0,

$$\mathbb{E}[X^n] = \left. \frac{dM_X(t)}{dt^n} \right|_{t=0}.$$

MGF Properties

1. **Uniqueness property**: If $M_X(t) = M_Y(t)$ for all $t \in \mathbb{R}$, then $F_X(x) = F_Y(x)$ for all $x \in \mathbb{R}$ (ie, $X \stackrel{d}{=} Y$).
2. **Linear transformations**: For all $a, b \in \mathbb{R}$,
$$M_{aX+b} = e^{bt} M_X(at).$$
3. **Linear combinations**: Let $X_1, \ldots, X_n$ be *independent*, $a_i \in \mathbb{R}$, and $S_n = \sum_{i=1}^n a_i X_i$. Then
$$M_{S_n}(t) = \prod_{i=1}^n M_{X_i}(a_i t).$$

Characteristic Functions

Similarly, the **characteristic function** (CF) is the complex function

$$\varphi_X(t) = \mathbb{E}[e^{itX}] = \mathbb{E}[\cos(tX) + i\sin(tX)]$$

for $t \in \mathbb{R}$. For all t such that $M_X(t)$ is finite,

$$\varphi_X(-it) = M_X(t).$$

The CF has many of the same properties as the MGF. However, the CF *always exists* for all $t \in \mathbb{R}$ and, in some cases, is easier to calculate than the MGF.

The Likelihood Function

- If $X_1, \cdots, X_n$ are and i.i.d. sample from a population with pdf/pmf $f(x \mid \boldsymbol{\theta})$ the **likelihood function** is

$$L(\boldsymbol{\theta}|x_1 \cdots, x_n) = \prod_{i=1}^{n} f(x_i \mid \boldsymbol{\theta})$$

- Density function versus likelihood: the density function $f(x \mid \boldsymbol{\theta})$ is a non-negative function of the data x that integrates to 1. The likelihood function is a function of the parameters $\boldsymbol{\theta}$ and typically will not integrate to 1

Maximum Likelihood Estimation

- **Maximum likelihood estimation** finds values of the parameters that maximize the likelihood function: $\hat{\theta} = \arg\max_{\theta \in \Theta} L(\boldsymbol{\theta}|\boldsymbol{x})$
- If the likelihood function is differentiable, then candidates for the MLE satisfy $\frac{\partial}{\partial \theta_i} L(\boldsymbol{\theta}|\boldsymbol{X}) = 0, i = 1, \cdots, k$.
- Since $log(t)$ is a monotonically increasing function of t, for any positive valued function f, $\arg\max_\theta f(x) = \arg\max_\theta \log f(x)$.
- Verify that the identified root is a local max by checking that the Hessian matrix is negative semi-definite at $\hat{\boldsymbol{\theta}}$.
- **Invariance property**: if $\hat{\theta}$ is the MLE for θ, then $g(\hat{\theta})$ is the MLE for $g(\theta)$

Convergence in Probability and Distribution

- Suppose we have an infinite sequence of random variables $X_1, X_2, \ldots$. What happens as $n \to \infty$? Can it "converge" like a sequence of real numbers? It turns out it can... in several ways!
- The sequence X_n **converges in probability** to an rv X (denoted $X_n \xrightarrow{p} X$) if for all $\epsilon > 0$,

$$P(|X_n - X| > \epsilon) \to 0 \text{ as } n \to \infty.$$

- The sequence X_n with corresponding sequence of CDFs F_n **converges in distribution** to an rv X (denoted $X_n \xrightarrow{d} X$) with cdf F if

$$F_n(x) \to F(x) \text{ for all continuity points } x \text{ of } F.$$

Large Sample Theory: Key Theorems

Under some conditions, the sample mean $\bar{X} = n^{-1} \sum_{i=1}^{n} X_i$ has some interesting properties as the sample size gets arbitrarily large.

1. **The Central Limit Theorem**: Let $X_1, X_2, \ldots$ be an infinite sequence of *iid* rvs, with $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}[X_i] = \sigma^2 < \infty$. Then
$$\sqrt{n}(\bar{X} - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2) \text{ as } n \to \infty.$$
2. **Weak Law of Large Numbers**: Let $X_1, X_2, \ldots$ be an infinite sequence of *iid* rvs, with $\mathbb{E}[X_i] = \mu < \infty$. Then
$$\bar{X} \xrightarrow{p} \mu \text{ as } n \to \infty.$$

Large Sample Theory: Useful Tools

1. **Slutsky's Theorem**: Let X_n, Y_n be sequences of rvs with $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$, a constant. Then:
   - $X_n + Y_n \xrightarrow{d} X + c$;
   - $X_n Y_n \xrightarrow{d} Xc$;
   - $X_n / Y_n \xrightarrow{d} X/c$ if $c \neq 0$.
2. **Continuous Mapping Theorem**: Let $X_n \xrightarrow{p} X$ and h be any continuous function on $\mathbb{R}$. Then

$$h(X_n) \xrightarrow{p} h(X).$$

3. $X_n \xrightarrow{p} X \implies X_n \xrightarrow{d} X$ and $X_n \xrightarrow{p} c \iff X_n \xrightarrow{d} c$.