Qualifying Exam Prep Session #1: Probability Fundamentals

Christian's Note: Welcome to the first qualifying exam review session! Some of these questions are, in my opinion, quite challenging, so you shouldn't feel discouraged if you're not sure how to approach them. I suggest first attempting them on your own, then with others/using external resources.

I will use the following shorthand to refer to various textbooks, which I think are useful for studying. Let me know if you want access and can't find them on your own.

- C&B: Casella & Berger: Statistical Inference (2$^{\text{nd}}$ edition)
- G&S: Grimmett & Stirzaker: One Thousand Exercises in Probability (3$^{\text{rd}}$ edition)

1 Basics (Modules 2/3/7 of BST 230 Notes)

1.1 Motivating Axiomatic Probability

When we think about data, we consider it to be the result of some *experiment.* We refer to the set of all possible outcomes of this experiment Ω as the *sample space.* For example, if our experiment is a single flip of a coin, $\Omega = \{H, T\}$; if our experiment involves flipping two coins, $\Omega = \{HH, HT, TH, TT\}$.

Let $A \subseteq \Sigma$ be any set of possible outcomes; we call A an event. If we have a bunch of events A_i we call the set of all these events $\mathcal{F} = \{A_i\}$, an *event space.* The purpose of a probability function $\mathbb{P} : \mathcal{F} \to [0, 1]$ is to assign each event A_i its "probability".

Andrey Kolmogorov introduced what is now the dominant notion of the *axioms of probability* in the 1930s. These axioms effectively constrain both the event space $\mathcal{F}$ and the probability function $\mathbb{P}$ to exhibit the behavior we've come to take for granted.

First, we require that the event space $\mathcal{F}$ be something called a *σ-algebra.* We say that $\mathcal{F}$ is a σ-algebra on Ω if $\mathcal{F} \subseteq 2^{\Omega}$ (i.e. $\mathcal{F}$ is a set of subsets of Ω) with the following properties:

- $\emptyset \in \mathcal{F}$
- $A \in \mathcal{F} \implies A^c \in \mathcal{F}$
- $A_1, A_2, \ldots \in \mathcal{F} \implies \cup_{i=1}^{\infty} A_i \in \mathcal{F}$.

Kolmogorov's axioms also require the following conditions on the probability function $\mathbb{P}$:

- Axiom #1: $\mathbb{P} : \mathcal{F} \to [0, 1]$
- Axiom #2: $\mathbb{P}(\Omega) = 1$
- Axiom #3: If $\{A_i\}_{i \in [n]}$ are pairwise disjoint and $A_i \in \mathcal{F}$ for all $i \in [n]$, then $\mathbb{P}\left(\cup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mathbb{P}(A_i)$

We put the elements $(\Omega, \mathcal{F}, \mathbb{P})$ together to form a *probability space.* Whenever you reason about probabilities, you are doing so with respect to a probability space.

1.2 The Calculus of Probability

These axioms will give rise to the fundamental calculus of probabilities, but first we need some basic set theory:

Theorem 1.1 (C&B Theorem 1.1.4)**.** *Let A, B, C be events in an event space $\mathcal{F}$ defined on a sample space Ω. Then,*

(a) Commutativity

$$A \cup B = B \cup A,$$

$$A \cap B = B \cap A;$$

(b) Associativity

$$A \cup (B \cup C) = (A \cup B) \cup C,$$
$$A \cap (B \cap C) = (A \cap B) \cap C;$$

(c) Distributive Laws

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C),$$
$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C);$$

(d) DeMorgan's Laws

$$(A \cup B)^c = A^c \cap B^c,$$
$$(A \cap B)^c = A^c \cup B^c.$$

We use these facts, along with the probability axioms, to produce fundamental rules for the probability function:

Theorem 1.2 (C&B Theorems 1.8 & 1.9)**.** *Let A, B be events in an event space $\mathcal{F}$ defined on a sample space Ω, and $\mathbb{P}$ an associated probability function. Then,*

(a) $\mathbb{P}(\emptyset) = 0$

(b) $\mathbb{P}(A) \leqslant 1$

(c) $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$

(d) $\mathbb{P}(B \cap A^c) = \mathbb{P}(B) - \mathbb{P}(A \cap B)$

(e) $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$

(f) $A \subseteq B \implies \mathbb{P}(A) \leqslant \mathbb{P}(B)$

Exercise #1 (C&B Theorem 1.2.11): Let $A \in \mathcal{F}$ and suppose $\{C_i\}$ is a (potentially countably infinite) partition of Ω such that $C_i \cap C_j = \emptyset$ for $i \neq j$ and $\cup_i C_i = \Omega$.
Prove the law of total probability; i.e. that

$$\mathbb{P}(A) = \sum_i \mathbb{P}(A \cap C_i)$$

1.3 Random Variables

1.3.1 Discrete/Continuous/Neither

A random variable (r.v.) $X : \Omega \to \mathbb{R}$ is a function from the sample space to the real numbers.[1]. We call $\{X(\omega) : \omega \in \Omega\}$ the *range* of X. X is said to be *discrete* if the range is countable. Discrete r.v. have a *probability mass function (pmf)* $f_X(x) = \mathbb{P}(X = x)$ and *cumulative distribution function (cdf)* $F_X(x) = \mathbb{P}(X \leqslant x)$. The cdf of a discrete r.v. is always discontinuous.

If X is a r.v. with continuous cdf $F_X(x) = \mathbb{P}(X \leqslant x)$ we say that X is a *continuous r.v.* Continuous r.v. do not have pmfs and the $\mathbb{P}(X = x) = 0$ for all x. We instead define the *probability density function (pdf)* as the function $f : \mathbb{R} \to [0, \infty)$ such that

$$\forall x \in \mathbb{R} : F_X(x) = \int_{-\infty}^{x} f_X(t)dt.$$

By the fundamental theorem of calculus, this also tells us that $f_X(x) = \frac{d}{dx}F_X(x)$.

For continuous r.v. it's not useful (as stated above) to think about $\mathbb{P}(X = x)$ as it's always 0. Instead, we consider statements like $\mathbb{P}(X \in A) = \int_A f(x)dx$.

The fact that we define discrete and continuous r.v. in qualitatively different ways (and not, say, "any r.v. with an uncountably infinite range is continuous") suggests that there are r.v. which are neither discrete nor continuous. In these cases, we cannot define either a pmf or pdf, but we can still reason about the cdf. In particular, we can always make statements of the kind "what is the probability that X is in A" for any $A \subseteq \mathbb{R}$. We calculate this by defining[2]

$$\begin{aligned}\mathbb{P}(X \in A) &= \mathbb{P}\left(X^{-1}(A)\right) \\ &= \mathbb{P}\left(\{\omega \in \Omega : X(\omega) \in A\}\right).\end{aligned}$$

1.3.2 Independence and Identical Distributions

When we reason about relationships between random variables we will often use the notions of "independence" and "identical (equal in) distribution".

Definition 1.3 (Independence)**.** *We will present four different definitions:*

Definition 1 $(X_1, \dots, X_n)$ *are independent if and only if*

$$\mathbb{P}(\cap_{i=1}^{n} X_i \in A_i) = \prod_{i=1}^{n} \mathbb{P}(X_i \in A_i)$$

for all measurable subsets $A_1, \dots, A_n \subseteq \mathbb{R}$.

Definition 2 $(X_1, \dots, X_n)$ *are independent if*

$$f_{X_{1:n}}(x_{1:n}) = \prod_{i=1}^{n} f_{X_i}(x_i)$$

Definition 3 $(X_1, \dots, X_n)$ *are independent if and only if*

$$F_{X_{1:n}}(x_{1:n}) = \prod_{i=1}^{n} F_{X_i}(x_i),$$

where $F_{X_{1:n}}(x_{1:n}) = \mathbb{P}\left(\cap_{i=1}^{n} \{X_i \leqslant x_i\}\right)$.

Definition 4 $(X_1, \dots, X_n)$ *are independent if and only if, for all measurable functions* $\phi_1, \dots, \phi_n$*:*

$$\mathbb{E} \prod_{i=1}^{n} \phi_i(X_i) = \prod_{i=1}^{n} \mathbb{E}\, \phi_i(X_i).$$

First, note that every definition (except for #2) is an "if and only if" statement, so they can be used both to reason about properties of independent r.v. and to prove that a set of r.v. with said property are independent.

Definition 1.4 (Equality in distribution)**.** *We say that r.v. X and Y are "equal in distribution" (i.e. $X \overset{d}{=} Y$) if*

$$\forall x \in \mathbb{R} : F_X(x) = F_Y(x).$$

[1] We're ignoring conditions on the "measurability" of X which are irrelevant for you now, but will be relevant to those who take BST 240

[2] This formulation is given on Slide 17 of the BST 230 Lecture 3 notes, but there's a typo in that version.

1.4 Combinatorics

In combinatorics we are considering, roughly, how many ways we can choose k things from a group of n. Two features of the process (other than k and n) dictate how we think about this. The first is whether or not the sampling happens with or without replacement; this is simply whether or not an item, after being selected, is placed back in the pool and able to be reselected. The second is whether or not order matters in our selection; for example do we consider drawing a blue ball and then a red ball qualitatively different than drawing a red ball and then a blue ball. A selection of items in which order doesn't matter is often called a *combination*; likewise a selection in which order does matter is called a *permutation.*

	without replacement	**with replacement**
ordered	$\frac{n!}{(n-k)!}$	n^k
unordered	$\binom{n}{k}$	$\binom{n+k-1}{k}$

Exercise #2 (G&S Exercise 3.4.7) Let $G = (V, E)$ be a finite graph with a set of vertices V and set of edges E between the vertices. That is, for any two vertices $v_1, v_2 \in V$, they have an edge between them if the pair $(v_1, v_2) \in E$.
For any $W \subseteq V$ and edge $e \in E$, define

$$I_W(e) = \mathbb{1}(\exists v \in W, v' \in W^c \text{ s.t. } e = (v, v')).$$

That is, $I_W(e)$ is an indicator for whether or not W and W^c are connected by e.
Let $N_W = \sum_{e \in E} I_W(e)$ and show that there exists $W \subseteq V$ such that $N_W \geqslant \frac{1}{2}|E|$.
Hint: To show that such a W exists, try constructing random W and showing that $\mathbb{P}\left(N_W \geqslant \frac{1}{2}|E|\right) > 0$; therefore, such a W must exist.

2 Transformations of Random Variables

Given knowledge of a random variable X, we are often interested in studying $Y := g(X)$. We can always characterize this distribution with

$$\mathbb{P}(Y \in A) = \mathbb{P}(g(X) \in A) = \mathbb{P}(X \in g^{-1}(A)).$$

We define $g^{-1}(A) = \{x : g(x) \in A\}$ and call this the *preimage* of A under g. This does not require that g is invertible, and is well-defined for any measurable g.

Transforming Discrete Random Variables Let X be a discrete random variable. Transforming the pmf for X to a pmf for Y is straightforward,

$$\begin{aligned} f_Y(y) &= \mathbb{P}(Y = y) \\ &= \mathbb{P}(g(X) = y) \\ &= \mathbb{P}(X \in g^{-1}(y)) \\ &= \sum_{x : g(x) = y} f_X(x). \end{aligned}$$

Transforming Continuous Random Variables It is **not** true in general, even for invertible g, that $f_Y(y) = f_X(g^{-1}(y))$, as these fs are densities, not probabilities. We do have a nice way to transform continuous random variables, but it requires some extra conditions.

Theorem 2.1. *Suppose X is a continuous r.v. and let $\mathcal{X} = \{x : f_X(x) > 0\}$. Suppose $Y = g(X)$ where $g : \mathcal{X} \to \mathbb{R}$ is strictly monotone and $g^{-1}(y)$ has a continuous derivative. Then,*

$$f_Y(y) = \begin{cases} f_X\left(g^{-1}(y)\right) \left|\frac{d}{dy} g^{-1}(y)\right|, & \text{for } y \in \mathcal{Y} := \{g(x) : x \in \mathcal{X}\} \\ 0, & \text{elsewhere} \end{cases}$$

In cases where g is non-monotonic, you generally have two different strategies for finding the density. The first is to work directly with the cdf and differentiate. The second is to partition X into subsets on which g is piecewise monotone and then apply Theorem 2.1. Theorem 2.2 shows how to do this

Theorem 2.2. *Suppose X is a continuous r.v. and let $\mathcal{X} = \{x : f_X(x) > 0\}$. Suppose $Y = g(X)$ where $g : \mathcal{X} \to \mathbb{R}$, and that there exists a set of disjoint subsets $\{I_1, \ldots, I_k\}$ of $\mathcal{X}$ such that $\mathbb{P}\left(X \in \bigcup_{i=1}^{k} I_k\right) = 1$. Moreover, assume that g is strictly monotonic, with differentiable inverse, on each I_j. We refer to $g_j := g|_{I_j}$ as the restriction of g to I_j.*

$$f_Y(y) = \sum_{j=1}^{k} f_X\left(g_j^{-1}(y)\right) \left|\frac{d}{dy} g_j^{-1}(y)\right| \mathbb{1}(y \in g(I_j)).$$

Exercise #3: Density of log-normal
Let $X \sim N(\mu, \sigma^2)$ and define $Y := \exp(X)$; we say that Y is a log-normal random variable. Find the density $f_Y(y)$.

Exercise #4: Density of Transformed Normal Let $X \sim N(0,1)$ and $Y := X^2$. Find the density $f_Y(y)$.

Multivariate Transformations Suppose $X = (X_1, \dots, X_n)$ is a random vector with density $f_X(x)$ and $Y := g(X) = \begin{pmatrix} g_1(X) \\ \vdots \\ g_n(X) \end{pmatrix}$ where $g : \mathbb{R}^n \to \mathbb{R}^n$ is bijective and differentiable. Then,

$$f_Y(y) = f_X\left(g^{-1}(y)\right) \left|\det\left(J_{g^{-1}}(y)\right)\right|,$$

where $J_{g^{-1}}$ is the $n \times n$ Jacobian of g^{-1} given by

$$J_{g^{-1}} = \begin{pmatrix} \frac{\partial g_1^{-1}}{\partial y_1} & \dots & \frac{\partial g_1^{-1}}{\partial y_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_n^{-1}}{\partial y_1} & \dots & \frac{\partial g_n^{-1}}{\partial y_n}. \end{pmatrix}$$

Exercise #5: Sums/Differences of Independent Normals Suppose X, Y are independent $N(0,1)$ random variables. Find the distributions of $X + Y$ and $X - Y$ and show that they are independent.

2.1 Probability Integral Transform

Suppose X is a random variable with continuous cdf F_X. Then $F_X(X) \sim \text{Uniform}(0,1)$. This is called the *probability integral transform.*

Now let $F_X^{-1}(u) = \inf\{x \in \mathbb{R} : F_X(x) \geqslant u\}$, we call F^{-1} the *generalized inverse* of F_X; it is also often called the *quantile function* of X. If $U \sim \text{Uniform}(0,1)$, then $F_X^{-1}(U) \overset{D}{=} X$; that is, $F_X^{-1}(U)$ is a random variable identical in distribution to X. This is called the *inverse probability integral transform.*

The inverse probability integral transform is a useful tool for using the uniform distribution as a basis for generating samples of random variables.

Exercise #6: (G&S Exercise 4.11.16)
Part (a): Let $U \sim \text{Uniform}(0,1)$ and $G : (0,1) \to \infty$ be continuous and monotone increasing. Show that the quantile function of $G(U)$ is G.
Part (b): For a r.v. X with continuous distribution function F and density f, define

$$H_X(v) = \mathbb{E}\left[F^{-1}(U)|U > v\right].$$

For $V \sim \text{Uniform}(0,1)$ we call $H_X(V)$ the *Hardy-Littlewood transform* of X.
Show that the quantile function of $H_X(V)$ is given by

$$(1-v)^{-1}\int_v^\infty F^{-1}(w)dw$$

3 (Conditional) Expectation (Modules 4/7 of BST 230 Notes)

Let X be a r.v. with pdf/pmf f_X. Then, for any measurable function g we write the *expectation* of $g(X)$ as

$$\mathbb{E}\, g(X) = \sum_{x \in \mathcal{X}} g(x) f_X(x) \;\; \text{(for discrete r.v.)}$$

$$\mathbb{E}\, g(X) = \int_{-\infty}^{\infty} g(x) f_X(x) dx \;\; \text{(for continuous r.v.)}$$

Likewise, for a random vector (X, Y) we write the *conditional expectation* of $g(X)$ given $Y = y$ as

$$\mathbb{E}\left[g(X)|Y = y\right] = \sum_{x \in \mathcal{X}} g(x) f_{X|Y}(x|y) \;\; \text{(for discrete r.v.)}$$

$$\mathbb{E}\left[g(X)|Y = y\right] = \int_{-\infty}^{\infty} g(x) f_{X|Y}(x|y) dx \;\; \text{(for continuous r.v.)}$$

Conditional expectation is well-defined if either $\mathbb{E}\left[g(X)^+|Y = y\right] < \infty$ or $\mathbb{E}\left[g(X)^-|Y = y\right] < \infty$.

Conditional expectation is itself a r.v. e.g. define the function $h(y) = \mathbb{E}\left[g(X)|Y = y\right]$, then $h(Y) = \mathbb{E}\left[g(X)|Y\right]$ is a r.v.

If we define $g(X) = \mathbb{1}(X \in A)$, then we can define the conditional distribution of X given $Y = y$ as

$$Q(A) = \mathbb{E}\left[\mathbb{1}(X \in A)|Y = y\right],$$

for any measurable set A.

(Conditional) expectations have a few useful arithmetic properties:

$$\mathbb{E}\left[cg(X)|Y = y\right] = c\,\mathbb{E}\left[g(X)|Y = y\right] \quad \text{(constants come out)}$$

$$\mathbb{E}\left[g(X) + h(X)|Y = y\right] = \mathbb{E}\left[g(X)|Y = y\right] + \mathbb{E}\left[h(X)|Y = y\right] \quad \text{(linearity)}$$

$$g(x) \leqslant h(x) \text{ for all } x \in \mathcal{X} \implies \mathbb{E}\left[g(X)|Y = y\right] \leqslant \mathbb{E}\left[h(X)|Y = y\right] \quad \text{(monotonicity)}$$

Theorem 3.1 (Law of Total Expectation)**.** *Let X be a r.v. with expected value $\mathbb{E}\, X$ and Y be a r.v. Then,*

$$\mathbb{E}\, X = \mathbb{E}\left[\mathbb{E}\left[X|Y\right]\right].$$

One particularly useful application of the law of total expectation is to let Y be a categorical r.v. taking values over a countable partition $\{A_i\}_{i \in \mathbb{N}}$ of the sample space Ω. This yields

$$\begin{aligned} \mathbb{E}\, X &= \mathbb{E}\left(\mathbb{E}\left[X|Y\right]\right) \quad \text{(law of total expectation)} \\ &= \sum_{i \in \mathbb{N}} \mathbb{E}(X|A_i)\, \mathbb{P}(A_i). \end{aligned}$$

Theorem 3.2 (Law of Total Variance)**.** *Let X be a r.v. with finite variance and Y be a r.v. Then,*

$$Var(X) = \mathbb{E}\left[Var(X|Y)\right] + Var\left(\mathbb{E}\left[Y|X\right]\right)$$

The laws of total expectation and variance are extremely useful when you want to find the expectation or variance of random variables defined with parameters which themselves are random variables.

Exercise #7: Suppose our company sells n widgets, each which breaks (independently) with probability p. We give independent random payouts Y_i to the buyers whose widgets break. Each payout follows an Exp(λ) distribution.

Let X be a r.v. representing the number of widgets that break and Y a r.v. representing our total payout to buyers.

Show that

$$\mathbb{E}\, Y = \frac{np}{\lambda}$$

$$\text{Var}(Y) = \frac{np(2-p)}{\lambda^2}.$$

Hint: Recall that the mean/variance of a Binomial(n, p) are $(np, np(1-p))$ and the mean/variance of an Exp(λ) are $(\lambda^{-1}, \lambda^{-2})$.

4 Extra Exercises

Exercise #8: Does marginal independence imply conditional independence? Suppose X, Y are r.v. such that $X \perp Y$. Does this imply that $(X|Z) \perp (Y|Z)$?
If so, prove it. If not, give a counterexample.

Exercise #9 (G&S 4.6.7): Let X, Y be r.v. with correlation ρ. Show that

$$\mathbb{E}\left(\mathrm{Var}(Y|X)\right) \leqslant (1-\rho^2)\mathrm{Var}(Y).$$

Hint: In addition to the techniques we've covered in this document, you may need to use Cauchy-Schwarz.

Exercise #10 (G&S 4.9.5): Let $X \sim N(0,1)$ and $a > 0$ be arbitrary. Define

$$Y = \begin{cases} X, & \text{if } |X| < a \\ -X, & \text{if } |X| \geqslant a. \end{cases}$$

Show that $Y \sim N(0,1)$ and find an expression for $\rho(a) = \mathrm{Cor}(X,Y)$ in terms of the density ϕ of X. Are (X,Y) bivariate normal for all $a > 0$?
Hint: You can assume that $\rho(a) = 0$ for some $a > 0$.