Qualifying Exam Prep Session #1: Probability Fundamentals

I will use the following shorthand to refer to various textbooks. Let me know if you want access and can't find them on your own.

- C&B: Casella & Berger: Statistical Inference (2^{nd} edition)
- G&S: Grimmett & Stirzaker: One Thousand Exercises in Probability (3^{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 statement like $\mathbb{P}(X \in A) = \int Af(x)dx$.

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 $(X_1, \ldots, X_n)$ *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 $A_1, \ldots, A_n \subseteq \mathbb{R}$.

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

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

Definition 3 $(X_1, \ldots, X_n)$ *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 $(X_1, \ldots, X_n)$ *are independent if and only if, for all measurable functions* $\phi_1, \ldots, \phi_n$*:*

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

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 (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

$$\mathbb{E}\, g(X)|Y = y = \sum_{x \in \mathcal{X}} g(x) f_{X|Y}(x|y) \text{ (for discrete r.v.)}$$
$$\mathbb{E}\, g(X)|Y = y = \int_{-\infty}^{\infty} g(x) f_{X|Y}(x|y) dx \text{ (for continuous r.v.)}$$

Conditional expectation is well-defined if either $\mathbb{E}\, g(X)^+|y < \infty$ or $\mathbb{E}\, g(X)^-|y < \infty$.

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

Theorem 2.1 (Law of Total Expectation)**.** *Let X be a r.v. with expected value $\mathbb{E}\, X$ and Y be a r.v. Then,*

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

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

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

Theorem 2.2 (Law of Total Variance)**.** *Let X be a r.v. with finite variance and Y be a r.v. Then,*

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

Exercise #3: 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 $\text{Exp}(\lambda)$ 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 $\text{Exp}(\lambda)$ are $(\lambda^{-1}, \lambda^{-2})$.

3 Inequalities (Module 9 of BST 230 Notes)

3.1 Basic Inequalities

We begin with a few inequalities immediately implied by the probability axioms.

Theorem 3.1 (Boole's Inequality (Union Bound)). *For any events $A_1, A_2, \ldots$,*

$$\mathbb{P}\left(\cup_i^\infty A_i\right) \leqslant \sum_i^\infty \mathbb{P}(A_i).$$

In general, we use Boole's Inequality to show that an event E has small probability. The general strategy is to identify $\{A_i\}$ such that $E \subseteq \cup_i^\infty A_i$ and where you can reason about $\mathbb{P}(A_i)$ individually for each i.

Theorem 3.2 (Bonferroni's Inequality). *For any events $A_1, A_2, \ldots$,*

$$\mathbb{P}\left(\cap_i^\infty A_i\right) \geqslant 1 - \sum_{i=1}^\infty \mathbb{P}(A_i^c).$$

Bonferroni's Inequality is useful for showing that a set of events all occur with high probability.

Exercise #4:
Use Boole's inequality to prove Bonferroni's inequality.

3.2 Tail Inequalities

We often want to reason about the "tails" of a r.v. X, making statements of the form $\mathbb{P}(|X| \geqslant a) \leqslant c$. If we know the distribution of X exactly, we can (at least in principle) give a very tight characterization of the tails. What can we do if we don't know the distribution of X exactly but, instead, know only some moments of X and/or moments of functions of X?

Theorem 3.3 (Markov's Inequality). *Let X be a non-negative r.v. with finite expectation and $a > 0$. Then,*

$$\mathbb{P}(X \geqslant a) \leqslant a^{-1}\,\mathbb{E}\,X.$$

Markov's provides the basis for other commonly used tail inequalities:

Theorem 3.4 (Chebyshev's Inequality)**.** *Let X be a r.v. with finite variance and $a > 0$. Then,*

$$\mathbb{P}(|X - \mathbb{E}\,X| \geqslant a) \leqslant \frac{Var(X)}{a^2}$$

Chebyshev's essentially takes $|X - \mathbb{E}\,X|$ to be the non-negative r.v. in Markov's and assumes bounded second moment (instead of first).

The Chernoff bound takes the idea from above a step further, using Markov's on the moment-generating function of X, $M_X(t) = \mathbb{E}\exp(tX)$.

Theorem 3.5 (Chernoff Bound)**.** *Let X be a r.v. and $a \in \mathbb{R}$. Then,*

$$\mathbb{P}\left(X \geqslant a\right) \leqslant \inf_{t>0} \exp(-ta)\,\mathbb{E}\exp(tX).$$

Exercise #5
Let $X \sim \text{Pois}(\lambda)$. Show that, for any $a > \lambda$:

$$\mathbb{P}(X \geqslant a) \leqslant \exp(-\lambda)\left(\frac{\exp(1)\lambda}{a}\right)^a.$$

Hint: $M_X(t) = \exp\left(\lambda(\exp(t) - 1)\right)$

3.3 Jensen's Inequality and Friends

Theorem 3.6 (Jensen's Inequality)**.** *Let X be a r.v. with range $\mathcal{X}$. If $g : \mathcal{X} \to \mathbb{R}$ is convex, then*

$$g(\mathbb{E}\,X) \leqslant \mathbb{E}\,g(X).$$

Theorem 3.7 (Weighted AM-GM Inequality)**.** *For any $x_{1:n} \geqslant 0$ and $w_{1:n} \geqslant 0$ s.t. $\sum_{i=1}^n w_i = 1$:*

$$w_{1:n}^T x_{1:n} \geqslant \prod_{i=1}^n x_i^{w_i}.$$

Theorem 3.8 (Hoeffding's Inequality)**.** *Suppose we have independent r.v. $X_{1:n}$ where $X_i \in [r_i, s_i]$. If we define $S_n = \sum_{i=1}^n X_i$, then for all $a > 0$:*

$$\mathbb{P}\left(|S_n - \mathbb{E}\,S_n| \geqslant a\right) \leqslant 2\exp\left(-\frac{2a^2}{\sum_{i=1}^n (s_i - r_i)^2}\right).$$

Note the similarity between Hoeffding's and the Chernoff bound; each provides a tail bound which is exponential in a. For Chernoff we know the first moment of the moment-generating function, whereas for Hoeffding's we need the r.v. to be bounded.

Exercise #6: Azuma's Inequality Let (Y_n) be a symmetric random walk such that $Y_0 = 0$ and, for all $n > 0$:

$$Y_n = \begin{cases} Y_{n-1} - 1, & \text{w prob } 1/2 \\ Y_{n-1} + 1, & \text{w prob } 1/2. \end{cases}$$

Show that

$$\mathbb{P}\left(|Y_n| > a\right) \leqslant 2\exp\left(-\frac{a^2}{2n}\right).$$

3.4 L^p Inequalities

Definition 3.9 (L^p norm of a r.v.)**.** *For $p \geqslant 1$, we define the L^p norm of a r.v. X as $\left(\mathbb{E}\left|X\right|^p\right)^{1/p}$. We say that $X \in L^p$ if X has a finite L^p norm.*

Theorem 3.10 (Hölder Inequality)**.** *Let $p, q > 1$ be such that $p^{-1} + q^{-1} = 1$. Then for r.v. X, Y:*

$$\mathbb{E}\left|XY\right| \leqslant \left(\mathbb{E}\left|X\right|^p\right)^{1/p}\left(\mathbb{E}\left|Y\right|^q\right)^{1/q}.$$

Many useful inequalities on L^p norms follow from Hölder.

- Cauchy-Schwarz: $\mathbb{E}\left|XY\right| \leqslant \left(\mathbb{E}\left|X\right|^2\right)^{1/2}\left(\mathbb{E}\left|Y\right|^2\right)^{1/2}$
- Lyapunov's: For $1 \leqslant r < s < \infty : X \in L^s \implies X \in L^r$
- Minkowski's: For any $p \geqslant 1 : \left(\mathbb{E}\left|X+Y\right|^p\right)^{1/2} \leqslant \left(\mathbb{E}\left|X\right|^p\right)^{1/p} + \left(\mathbb{E}\left|Y\right|^p\right)^{1/p}$

Minkowski's is essentially the triangle inequality for L^p norms.

Exercise #7: Paley Zygmund Inequality
Let X be a non-negative r.v. with finite variance. Prove that, for all $\theta \in (0, 1)$:

$$\mathbb{P}\left(X \geqslant \theta\, \mathbb{E}\, X\right) \geqslant (1-\theta)^2 \frac{(\mathbb{E}\, X)^2}{\mathbb{E}\, X^2}.$$

Hint: Use the law of total expectation and Cauchy-Schwarz

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

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

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 $a > 0$.