Key Ideas from Probability (Math 265)

(This study guide has been adapted with permission from a study guide created by Joseph Blitzstein)

First Principles

- Combinatorics: multiplication rule, binomial coefficients, permutations and combinations, sampling with/without replacement when order does/doesn't matter, inclusion-exclusion
- Basic Probability: sample spaces, events, axioms of probability, equally likely outcomes, inclusion-exclusion, unions, intersections, and complements
- Conditional Probability: definition and meaning, Bayes' Rule, Law of Total Probability, thinking conditionally, independence

Univariate Distributions

- Random Variables: definition and interpretations, stories, discrete vs. continuous, distributions, CDFs, PMFs, PDFs, functions of a RV, indicator RVs, memorylessness of the Geometric and Exponential, Poisson approximation, sums of independent RVs
- Expected Value: linearity, indicator RVs, variance, standard deviation, LOTUS
- Important Discrete Distributions: Bernoulli, Binomial, Geometric, Negative Binomial, Hypergeometric, Poisson, Uniform
- Important Continuous Distributions: Uniform, Normal, Exponential, Gamma

Multivariate Distributions

- Jointly Distributed Random Variables: joint, conditional and marginal distributions, independence
- Expected Value: linearity, covariance, correlation, 2D LOTUS
- Conditional Expectation: difference between $P(X|Y=y)$ and $P(X|Y)$, law of total expectation (Adam's Law), law of total variance (Eve's law)

Limit Laws and Inequalities

- Limit Theorems: Law of Large Numbers, Central Limit Theorem
- Inequalities: Markov, Chebyshev

Overall Strategies

- Complements
- Conditioning
- Symmetry
- Linearity
- Indicator RVs
- Checking whether answers make sense (e.g., looking at simple and extreme cases and avoiding category errors).

1 Important Distributions

1.1 Table of Distributions

Here $0 < p < 1$. The parameters for Gamma and Beta are positive real numbers; n, r, m, and N are positive integers.

Name	Param.	PMF	Mean	Variance
Discrete Uniform	n	$\frac{1}{n}$ $k \in \{1, \ldots, n\}$	$\frac{n+1}{2}$	$\frac{n^2-1}{12}$
Bernoulli	p	$P(X=1)=p$ $P(X=0)=1-p$	p	$p(1-p)$
Binomial	n, p	$\binom{n}{k}p^k(1-p)^{n-k}$ $k \in \{0, 1, \ldots, n\}$	np	$np(1-p)$
Geometric	p	$(1-p)^{k-1}p$ $k \in \{1, 2, \ldots\}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Negative Binomial	r, p	$\binom{k-1}{r-1}(1-p)^{k-r}p^r$ $n \in \{0, 1, 2, \ldots\}$	$\frac{r}{p}$	$\frac{r(1-p)}{p^2}$
Hypergeometric	$r, N-r, n$	$\frac{\binom{r}{k}\binom{N-r}{n-k}}{\binom{N}{n}}$ $k \in \{0, 1, \ldots n\}$	$\frac{nr}{N}$	$\frac{n(N-n)r(N-r)}{N^2(N-1)}$
Poisson	λ	$\frac{e^{-\lambda}\lambda^k}{k!}$ $k \in \{0, 1, 2, \ldots\}$	λ	λ
Uniform	$a < b$	$\frac{1}{b-a},\ x \in (a, b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Normal	μ, σ^2	$\frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/(2\sigma^2)}$ $x \in \mathbb{R}$	μ	σ^2
Exponential	λ	$\lambda e^{-\lambda x},\ x > 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Gamma	a, λ	$\frac{\lambda^a}{\Gamma(a)}x^{a-1}e^{-\lambda x}$ $x > 0$	$\frac{a}{\lambda}$	$\frac{a}{\lambda^2}$

1.2 Connections Between Distributions

The table above summarizes the PMFs/PDFs of the important distributions, and their means and variances, but it does not say where each distribution comes from (stories), or how the distributions interrelate. Some of these connections between distributions are listed below.

Also note that some of the important distributions are special cases of others. Bernoulli is a special case of Binomial; Geometric is a special case of Negative Binomial; and the Exponential is a special case of Gamma.

1. Binomial: If $X_1, \dots, X_n$ are i.i.d. Bernoulli(p), then $X_1 + \cdots + X_n \sim$ Binomial(n, p).
2. Negative Binomial: If $G_1, \dots, G_n$ are i.i.d. Geometric(p), then $G_1 + \cdots + G_n \sim$ NBin(n, p)
3. Location and Scale: If $X \sim \mathcal{N}(0,1)$, then $\mu + \sigma Z \sim \mathcal{N}\left(\mu, \sigma^2\right)$.
   If $U \sim$ Unif($0,1$) and $a < b$, then $a + (b-a)U \sim$ Unif(a, b).
   If $X \sim$ Exp(1), then $X/\lambda \sim$ Exp(λ).
   If $X \sim$ Gamma(a, λ), then $\lambda Y \sim$ Gamma($a, 1$).
4. Symmetry: If $X \sim$ Binomial($n, 1/2$), then $n - X \sim$ Binomial($n, 1/2$).
   If $U \sim$ Unif($0,1$), then $1 - U \sim$ Unif($0,1$).
   If $Z \sim \mathcal{N}(0,1)$, then $-Z \sim \mathcal{N}(0,1)$.
5. Gamma: If $X_1, \dots, X_n$ are i.i.d. Exp(λ), then $X_1 + \cdots + X_n \sim$ Gamma(n, λ).

2 Sums of Independent Random Variables

Let $X_1, \dots, X_n$ be independent random variables. The table below shows the distribution of their sum, $X_1 + \cdots + X_n$, for various important cases depending on the distribution of X_i. The central limit theorem says that a sum of a large number of i.i.d. RVs will be approximately Normal, while these are exact distributions.

X_i	$\sum_{i=1}^{n} X_i$
Bernoulli(p)	Binomial(n, p)
Binomial(n_i, p)	Binomial$(\sum_{i=1}^{n} n_i, p)$
Geometric(p)	NBin(n, p)
NBin(r_i, p)	NBin$(\sum_{i=1}^{n} r_i, p)$
Poisson(λ_i)	Poisson$(\sum_{i=1}^{n} \lambda_i)$
$\mathcal{N}\left(\mu_i, \sigma_i^2\right)$	$\mathcal{N}\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$
Exponential(λ)	Gamma(n, λ)
Gamma(α_i, λ)	Gamma$(\sum_{i=1}^{n} \alpha_i, \lambda)$

3 Review of Some Useful Results

3.1 De Morgan's Laws

$$(A_1 \cup A_2 \cup \cdots \cup A_n)^c = A_1^c \cap A_2^c \cap \cdots \cap A_n^c$$
$$(A_1 \cap A_2 \cap \cdots \cap A_n)^c = A_1^c \cup A_2^c \cup \cdots \cup A_n^c$$

3.2 Complements

$$P\left(A^c\right) = 1 - P(A)$$

3.3 Unions

$$P(A \cup B) = P(A) + P(B) - P(A \cup B)$$

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = \sum_{i=1}^{n} P\left(A_i\right), \text{ is the } A_i \text{ are disjoint;}$$

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = \sum_{k=1}^{n} \left((-1)^{k+1} \sum_{i_1 < i_2 < \cdots < i_n} P\left(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}\right) \right) \text{ (Inclusion-Exclusion)}$$

3.4 Intersections

$$P(A \cap B) = P(B|A)P(A) = P(A|B)P(B)$$

3.5 Law of Total Probability

If $E_1, E_2, \ldots, E_n$ are a partition of the sample space Ω (i.e., they are disjoint and their union is all of Ω) and $P(E_i) \neq 0$ for all i, then

$$P(A) = \sum_{i=1}^{n} P\left(A \middle| E_i\right) P\left(E_i\right)$$

An analogous formula holds for conditioning on a continuous RV X with PDF $f(x)$:

$$P(A) = \int\limits_{-\infty}^{\infty} P(A|X = x) f(x) dx$$

Similarly, to go from a joint PDF $f(x, y)$ for (X, Y) to the marginal PDF of Y, integrate over all values of x:

$$f_Y(y) = \int\limits_{-\infty}^{\infty} f(x, y) dx.$$

3.6 Bayes' Rule

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Often the denominator $P(B)$ is then expanded by the Law of Total Probability. For continuous RVs X and Y, Bayes' Rule becomes

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

3.7 Expected Value, Variance, and Covariance

Expected value is *linear*: for any random variables X and Y and constant c,

$$E(X + Y) = E(X) + E(Y)$$
$$E(cX) = cE(X)$$

Variance can be computed in two ways:

$$Var(X) = E\left[(X - EX)^2\right] = E\left(X^2\right) - (EX)^2$$

Constants come out from variance as the constant squared:

$$Var(cX) = c^2 Var(X).$$

For the variance of the sum, there is a covariance term:

$$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y),$$

where

$$Cov(X,Y) = E((X - EX)(Y - EY)) = E(XY) - (EX)(EY).$$

So if X and Y are uncorrelated, then the variance of the sum is the sum of the variances. Recall that independence implies uncorrelated but not vice versa. Covariance is symmetric:

$$Cov(Y,X) = Cov(X,Y),$$

and covariances of sums can be expanded as

$$Cov(X+Y, Z+W) = Cov(X,Z) + Cov(X,W) + Cov(Y,Z) + Cov(Y,W).$$

Note that for a constant c,

$$\begin{aligned} Cov(X,c) &= 0 \\ Cov(cX,Y) &= c \cdot Cov(X,Y) \end{aligned}$$

The correlation of X and Y , which is between -1 and 1, is

$$Corr(X,Y) = \frac{Cov(X,Y)}{SD(X)SD(Y)}.$$

This is also the covariance of the standardized versions of X and Y .

3.8 Law of the Unconscious Statistician (LOTUS)

Let X be a discrete random variable and h be a real-valued function. Then $Y = h(X)$ is a random variable. To compute EY using the definition of expected value, we would need to first find the PMF of Y and use $EY = \sum_y y \cdot P(Y = y)$. The Law of the Unconscious Statistician says we can use the PMF of X directly:

$$E[h(X)] = \sum_x h(x) \cdot P(X = x).$$

Similarly, for X a continuous RV with PDF $f_X(x)$, we can find the expected value of $Y = h(X)$ by integrating $h(x)$ times the PDF of X, without first finding $f_Y(y)$:

$$E[h(X)] = \int_{-\infty}^{\infty} h(x) \cdot f_X(x) dx$$

Remember that LOTUS also works in higher dimensions. For example, the two-dimensional version of LOTUS says that we can use the joint PMF or PDF to find the expected value of a function of

$g(X, Y)$:

$$E\left[h(X,Y)\right] = \sum_{y}\sum_{x} g(x,y) \cdot P(X = x, Y = y)$$

$$E\left[h(X,Y)\right] = \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} g(x,y) \cdot f_{X,Y}(x,y)dxdy$$

3.9 Indicator Random Variables

Let A and B be events. Indicator RVs form a bridge between probability and expectation: $P(A) = E(I_A)$, where I_A is the indicator RV for A. It is often useful to think of a "counting" RV as a sum of indicator RVs. Indicator RVs have many pleasant properties. For example, $(I_A)^k = I_A$ for any positive number k, so it's easy to handle moments of indicator RVs. Also note that

$$\begin{aligned} I_{A\cap B} &= I_A I_B \\ I_{A\cup B} &= I_A + I_B - I_A I_B \end{aligned}$$

3.10 Transformations

If you want to find the distribution of $Y = g(X)$, where $g : \mathbb{R} \to \mathbb{R}$. Remember that both the CDF and PDF/PMF uniquely define a distribution.

If X is continuous, then we can use the CDF method:

1. Determine the possible values of Y based on the values of X and the function g.
2. Begin with the CDF $F_Y(y) = P(Y \leq y) = P(g(X) \leq y)$. Express the CDF in terms of the original random variable X.
3. From $P(g(X) \leq y)$ obtain an expression of the form $P(X \leq \ldots)$. The righthand side of the expression will be a function of y. If f is invertible then $P(g(X) \leq y) = P(X \leq g^{-1}(y))$.
4. Differentiate with respect to y to obtain the density $f_Y(y)$.

If X is discrete, then we always use appeal to the original PMF of X:

$$P(Y = y) = \sum_{\{x:g(x)=y\}} P(X = x)$$

If g is invertible, then we can use a method similar to that for a continuous RV:

1. Determine the possible values of Y based on the values of X and the function g.
2. Begin with the PMF $P(Y = y) = P(g(X) = y)$. Express the PMF in terms of the original random variable X.
3. From $P(g(X) = y)$ obtain an expression of the form $P(X = g^{-1}(y))$.

3.11 Symmetry

There are many beautiful and useful forms of symmetry in probability and statistics. For example:

1. If X and Y are i.i.d., then $P(X < Y) = P(Y < X)$. More generally, if $X_1, \dots X_2$ are i.i.d., then $P(X_1 < X_2 < \dots < X_n) = P(X_n < X_{n-1} < \dots < X_1)$, and likewise all $n!$ orderings are equally likely (in the continuous case it follows that $P(X_1 < X_2 < \dots < X_n) = \frac{1}{n!}$, while in the discrete case we also have to consider ties).
2. If we shuffle a deck of cards and deal the first two cards, then the probability is 1/52 that the second card is the Ace of Spades, since by symmetry it's equally likely to be any card; it's not necessary to do a law of total probability calculation conditioning on the first card.
3. Consider the Hypergeometric, thought of as the distribution of the number of white balls, where we draw n balls from a jar with w white balls and b black balls (without replacement). By symmetry and linearity, we can immediately get that the expected value is $n\frac{w}{w+b}$, even though the trials are not independent, as the jth ball is equally likely to be any of the balls, and linearity still holds with dependent RVs.
4. $E(X_1|X_1 + X_2) = E(X_2|X_1 + X_2)$ by symmetry if X_1 and X_2 are i.i.d. So by linearity, $E(X_1|X_1 + X_2) + E(X_2|X_1 + X_2) = E(X_1 + X_2|X_1 + X_2) = X_1 + X_2$, which gives $E(X_1|X_1 + X_2) = (X_1 + X_2)/2$.

3.12 Conditional Expectation

The conditional expected value $E(Y|X = x)$ is a number (for each x) which is the average value of Y, given the information that $X = x$. The definition is analogous to the definition of $E(Y)$: just replace the PMF or PDF by the conditional PMF or conditional PDF.

It is often very convenient to just directly condition on X to obtain $E(Y|X)$, which is a random variable (it is a function of X). This intuitively says to average Y, treating X as if it were a known constant: $E(Y|X = x)$ is a function of x, and $E(Y|X)$ is obtained from $E(Y|X = x)$ by "changing x to X". For example, if $E(Y|X = x) = x^3$, then $E(Y|X) = X^3$.

Important properties of conditional expectation:

- $E(aY_1 + bY_2|X) = aE(Y_1|X) + bE(Y_2|X)$ where a and b are constants (Linearity)
- $E(Y|X) = E(Y)$ if X and Y are independent
- $E(Y) = E(E(Y|X))$ (Law of Total Expectation/Adam's Law)
- $Var(Y) = E(Var(Y|X)) + Var(E(Y|X))$ (Law of Total Variance/Eve's Law)

The last two identities are often useful for finding the mean and variance of Y: first condition on some choice of X where the conditional distribution of Y given X is easier to work with than the unconditional distribution of Y, and then account for the randomness of X.

3.13 Limit Laws

Let $X_1, X_2 \dots$ be i.i.d. random variables with mean μ and variance σ^2. The sample mean is defined as

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

The Strong Law of Large Numbers says that with probability 1, the sample mean converges to the true mean (i.e., it converges almost surely):

$$\overline{X}_n \to \mu \text{ with probability 1.}$$

The Weak Law of Large Numbers (which follows from Chebyshev's Inequality) says that $\overline{X}_n$ will be very close to μ with very high probability; that is for any $\epsilon > 0$,

$$P\left(|\overline{X}_n - \mu| > \epsilon\right) \to 0 \text{ as } n \to \infty.$$

The Central Limit Theorem says that the sum of a large number of i.i.d. random variables is approximately Normal in distribution. More precisely, standardize the sum $X_1 + \cdots + X_n$ (by subtracting its mean and dividing by its standard deviation); then the standardized sum approaches $\mathcal{N}(0,1)$ in distribution (i.e., the CDF of the standardized sum converges to Φ). So

$$\frac{(X_1 + \cdots + X_n) - n\mu}{\sigma\sqrt{n}} \to \mathcal{N}(0,1) \text{ in distribution.}$$

In terms of the sample mean,

$$\frac{\sqrt{n}}{\sigma}\left(\overline{X}_n - \mu\right) \to \mathcal{N}(0,1) \text{ in distribution.}$$

3.14 Inequalities

When probabilities and expected values are hard to compute exactly, it is useful to have inequalities. One simple but handy inequality is Markov's Inequality:

$$P(X > a) \le \frac{E(|X|)}{a},$$

for any $a > 0$. Let X have mean μ and variance σ^2. Using Markov's Inequality with $(X - \mu)^2$ in place of X gives Chebyshev's Inequality:

$$P(|X - \mu| > a) \le \frac{Var(X)}{a^2}.$$

4 Common Mistakes in Probability

4.1 Category Errors

A category error is a mistake that not only happens to be wrong, but also it is wrong in every possible universe. If someone answers the question "How many students are enrolled at Carleton?"

with “50” that is wrong, but there is no logical reason the enrollment couldn’t be 50. But answering the question with “-42” or “π” or “pink elephants” would be a category error. To help avoid being categorically wrong, always think about what type an answer should have. Should it be an integer? A nonnegative integer? A number between 0 and 1? A random variable? A distribution?

- Probabilities must be between 0 and 1
- Variances must be nonnegative
- Correlations must be between -1 and 1
- The range of possible values must make sense
- Units should make sense
- A number can’t equal a random variable (unless the RV is actually a constant). Quantities such as $E(X)$, $P(X > 1)$, $F_X(1)$, $Cov(X,Y)$ are numbers. We often use the notation “$X = x$”, but this is shorthand for an event (it is the set of all possible outcomes of the experiment where X takes the value x).
- Don’t replace a RV by its mean, or confuse $E(g(X))$ with $g(E(X))$.
- An event is not a random variable.
- Indicator variables in an integral can’t make their way out of the integral.
- A random variable is not the same thing as its distribution.

4.2 Notational Paralysis

Another common mistake is a reluctance to introduce notation. This can be both a symptom and a cause of not seeing the structure of a problem. Be sure to define your notation clearly, carefully distinguishing between constants, random variables, and events.

- Give objects names if you want to work with them.
  Example: Suppose that we want to show that

  $$E\left(\cos^4\left(X^2+1\right)\right) \geq \left(E\left(\cos^2\left(X^2+1\right)\right)\right)^2.$$

  The essential pattern is that there is a RV. on the right and its square on the left; so let $Y = \cos^2\left(X^2+1\right)$, which turns the desired inequality into the statement $E\left(Y^2\right) \geq (EY)^2$, which we know is true because variance is nonnegative.
- Introduce clear notation for events and RVs of interest.

4.3 Common Sense and Checking Answers

Whenever possible (i.e., when not under severe time pressure), look for simple ways to check your answers, or at least to check that they are plausible. This can be done in various ways, such as using the following methods.

1. *Miracle checks.* Does your answer seem intuitively plausible? Is there a category error? Did asymmetry appear out of nowhere when there should be symmetry?
2. *Checking simple and extreme cases.* What is the answer to a simpler version of the problem? What happens if $n = 1$ or $n = 2$, or as $n \to \infty$, if the problem involves showing something for all n?
3. Looking for alternative approaches and connections with other problems. Is there another natural way to think about the problem? Does the problem relate to other problems we've seen?

   - Probability is full of counterintuitive results, but not impossible results!
   - Check simple and extreme cases whenever possible.
   - Check that PMFs are nonnegative and sum to 1, and PDFs are nonnegative and integrate to 1 (or that it is at least plausible), when it is not too messy.

4.4 Random Variables vs. Distributions

A random variable is not the same thing as its distribution! Some people call this confusion sympathetic magic, and the consequences of this confusion are often disastrous. Every random variable has a distribution (which can always be expressed using a CDF, which can be expressed by a PMF in the discrete case, and which can be expressed by a PDF in the continuous case).

Every distribution can be used as a blueprint for generating RVs, but that doesn't mean that doing something to a RV corresponds to doing it to the distribution of the RV. Confusing a distribution with a RV with that distribution is like confusing a map of a city with the city itself, or a blueprint of a house with the house itself. The word is not the thing, the map is not the territory.

- A function of a RV is a RV
- Avoid sympathetic magic
- A CDF $F(x) = P(X \leq x)$ is a way to specify the distribution of X, and is a function defined for all real values of x. Here X is the RV, and x is any number; we could just as well have written $F(t) = P(X \leq t)$.

4.5 Conditioning

It is easy to make mistakes with conditional probability so it is important to think carefully about what to condition on and how to carry that out.

- Condition on *all* the evidence!
- Don't destroy information.
- Independence shouldn't be assumed without justification, and it is important to be careful not to implicitly assume independence without justification.
- Independence is completely different from disjointness!
- Independence is a symmetric property: if A is independent of B, then B is independent of A. *There's no such thing as unrequited independence.*
- The marginal distributions can be extracted from the joint distribution, but knowing the marginal distributions does not determine the joint distribution.
- Don't confuse $P(A|B)$ with $P(B|A)$.
- Don't confuse $P(A|B)$ with $P(A, B)$.