STAT 333 - Applied Probability

Fantastic Theorems and How to Prove Them

Jose Luis Avilez
Faculty of Mathematics
University of Waterloo

Chapter 1

Review of Probability

Definition 1.1. A **sample space** S is the set of all possible outcomes in an experiment. In an experiment one, and only one, outcome can occur (that is, they are mutually disjoint). An event A is a subset of the sample space $A \subseteq S$.

Definition 1.2. Kolmogorov's Axioms. For each event A, $P(A)$ is defined as the probability of A satisfying the following properties:

1. $0 \leq P(A) \leq 1$
2. $P(S) = 1, P(\emptyset) = 0$
3. For $n \in \mathbb{Z}^+$,

$$P(A_1 \cup \ldots \cup A_n) = \sum_{i=1}^{n} P(A_i)$$

if the sequence $\{A_i\}_{i=1}^{n}$ is mutually exclusive.

Definition 1.3. Events A and B are said to be **independent** if and only if $P(A \cap B) = P(A)P(B)$. The events are said to be **dependent** if they are not independent.

Theorem 1.4. *If $A \subseteq B$, then $P(A) \leq P(B)$.*

Proof. Create partition of B through A and then use property 1 of Definition 1.2. ∎

Definition 1.5. The **conditional probability** of A given B is defined as

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

provided that $P(B) > 0$.

Theorem 1.6. *If A and B are independent then $P(A|B) = P(A)$.*

Proof. Duh. ∎

Theorem 1.7. *If A and B are dependent, then either (i) $P(A|B) > P(A)$ and $P(B|A) > P(B)$ or (ii) $P(A|B) < P(A)$ and $P(B|A) < P(B)$.*

Definition 1.8. We say that a collection of events $A_1, A_2, \ldots, A_k$ is a **partition** of S if it satisfies:

1. $A_i \cap A_j = \emptyset$ for all $i \neq j$
2. $\bigcup_{i=1}^{k} A_i = S$

Theorem 1.9. *For any event $B \subseteq S$ and partition $\{A_i\}_{1 \leq i \leq k}$, we have*

$$B = (B \cap A_1) \cup (B \cap A_2) \cup \ldots \cup (B \cap A_k)$$

Proof. Follows trivially from Definition 1.8 and set algebra. ■

Theorem 1.10. ***Law of total probability***. *For some event B and partition $\{A_i\}_{i=1}^k$, we have:*

$$P(B) = \sum_{i=1}^{k} P(A_i)P(B|A_i)$$

Proof. By Theorem 1.9, we can express B as the disjoint union

$$B = (B \cap A_1) \cup (B \cap A_2) \cup \ldots \cup (B \cap A_k)$$

By Axiom 3 and the definition of conditional probability, we have that

$$P(B) = P(B \cap A_1) + \ldots + P(B \cap A_k) = \sum_{i=1}^{k} P(A_i)P(B|A_i)$$

■

Theorem 1.11. ***Bayes' Rule.***

$$P(A_i|B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(A_i)P(B|A_i)}{\sum_{i=1}^{k} P(A_i)P(B|A_i)}$$

Proof. Follows from Theorem 1.10 and the definition of conditional probability. ■

Example 1.12. In the Monty Hall problem, we can use Bayes' rule to prove that the optimal strategy is choosing to switch doors.

Chapter 2

Random Variables

Definition 2.1. A **random variable** $X : S \to \mathbb{R}$ is a function that maps points on the sample space to real numbers.

Definition 2.2. A random variable X is said to be **discrete** if the range of X is countable.

Definition 2.3. A random variables X is said to be **continuous** if the range is uncountable.

Definition 2.4. We say that a process is a **Bernouilli trial** if it satisfies the following three conditions:

1. There are two possible outcomes.
2. The trials are independent.
3. The probability of a success remains constant over time.

Definition 2.5. We define a **probability mass function** using the diabolical notation:

$$p(x) = P(X = x) = P(\{e \in S | X(e) = x\})$$

Definition 2.6. The **cumulative distribution function** of a random variable X is

$$F(x) = P(X \leq x) = P(\{e \in S | X(e) \leq x\})$$

Definition 2.7. A **Bernouilli random variable** is defined as

$$X = \begin{cases} 1 & \text{if there is a success} \\ 0 & \text{if there is a failure} \end{cases}$$

with p.m.f. $p(x) = p^x(1-p)^x$.

Theorem 2.8. *For a Bernouilli random variable X, $\mathrm{E}(X) = p$ and $\mathrm{Var}(X) = p(1-p)$.*

Proof. Trivial ∎

Definition 2.9. A **binomial random variable** is defined as the number of successes in n Bernouilli trials. We say $X \sim \mathrm{Bin}(n, p)$. Notice that this is the sum of n Bernouilli random variables. The p.m.f. is given by $p(x) = \binom{n}{x} p^x(1-p)^x$.

Theorem 2.10. *For a binomial random variable X, $\mathrm{E}(X) = np$ and $\mathrm{Var}(X) = np(1-p)$*

Proof. Easy piecey. ∎

Definition 2.11. We say that X is a geometric random variable if it records the number of trials required until a first success. We say that $X \sim \mathrm{Geo}(p)$. It has p.m.f. $p(x) = (1-p)^{x-1}p$.

Theorem 2.12. *For a geometric random variable X, $\mathrm{E}(X) = \frac{1}{p}$ and $\mathrm{Var}(X) = \frac{(1-p)}{p^2}$.*

Proof. I'll post it later. ∎

Definition 2.13. A **negative binomial random variable** is defined as the number of trials until the k-th success is observed. The range is $\{k, k+1, k+2, \ldots\}$. Its p.m.f. is given by $p(x) = \binom{x-1}{k-1} p^k (1-p)^{x-k}$

Theorem 2.14. *A negative binomial random variable X has $\mathrm{E}(X) = \frac{k}{p}$ and $\mathrm{Var}(X) = \frac{k(1-p)}{p^2}$.*

Proof. Use linearity over sum of geometric random variables. ∎

Definition 2.15. We say that a random variable X is **Poisson** if it counts the number of events occurring randomly through time t at constant rate λ. We say that $X \sim \mathrm{Po}(\lambda t)$ which has a (provable) p.m.f. $p(x) = \frac{e^{\lambda}(\lambda t)^x}{x!}$.

Theorem 2.16. *For a Poisson random variable X, $\mathrm{E}(X) = \mathrm{Var}(X) = \lambda t$.*

Proof. Use the Changbao tricks from STAT 240. ∎

Definition 2.17. The probability density function of a random variable X is defined to be $f(x) = \frac{d}{dx}F(x)$ where F is the cumulative distribution of X.

Definition 2.18. We say that a random variable X is uniform, and denote $X \sim \mathrm{U}(a, b)$ if it has p.d.f. $f(x) = \frac{1}{b-a}$ with $x \in (a, b)$.

Theorem 2.19. *The c.d.f. of a uniform random variable X is*

$$F(x) = \begin{cases} \frac{x-a}{b-a} & x \in (a, b) \\ 0 & x \leq a \\ 1 & x \geq b \end{cases}$$

the expectation of X is $\mathrm{E}(X) = \frac{a+b}{2}$; the variance is $\mathrm{Var}(X) = \frac{(b-a)^2}{12}$.

Proof. Follows from STAT 240. ∎

Definition 2.20. A random variable X is said to be **exponential** if it records the amount of time elapsed between events in a Poisson process with rate λ. Its range is $(0, \infty)$.

Theorem 2.21. *An exponential random variable X has p.d.f. $f(x) = \lambda e^{-\lambda x}$, $\mathrm{E}(X) = \frac{1}{\lambda}$; $\mathrm{Var}(X) = \frac{1}{\lambda^2}$; and has the memoryless property: $P(X > t + s | X > s) = P(X > t)$.*

Proof. The c.d.f. of X is given by

$$F(x) = 1 - P(\text{no events in (0,x)}) = 1 - e^{-\lambda x}$$

where the second equality follows since it is the probability of no events in a Poisson distribution with rate λ and time x. Taking its derivative yields the desired result. The remaining facts follow from STAT 240. ∎

Definition 2.22. We say that a random variable X follows a gamma distribution if its p.d.f. is

$$f(x) = \frac{e^{-\lambda x} \lambda^{\alpha} x^{\alpha-1}}{\Gamma(\alpha)}$$

where,

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} dx$$

Note that $\Gamma(\alpha) = (\alpha - 1)!$ if $\alpha \in \mathbb{Z}^+$.

Example 2.23. The gamma distribution can be used to model the waiting time for α events in a Poisson process with rate λ if $\alpha \in \mathbb{Z}^+$. If $\alpha = 1$, the gamma distribution reduces to the exponential distribution.

Definition 2.24. We say that a random variable X follows a normal distribution if its p.d.f. is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

Definition 2.25. For two random variables X, Y we can define the following:

1. The **joint cumulative distribution** of X and Y is $F(x, y) = P(X \leq x, Y \leq y)$.
2. The **joint probability mass function** of x, y is $p(x, y) = P(X = x, Y = y)$. The **joint probability density function** of x, y is $\frac{\partial^2}{\partial x \partial y} F(x, y)$ (for now assume that $F \in C^2$, ask on Piazza later).
3. The **marginal probability mass function** is $p_X(x) = \sum_y p(x, y)$. The **probability density function** is $f_X(x) = \int_y f(x, y) dy$.

Definition 2.26. We say that X and Y are **independent** if and only if $f(x, y) = f_X(x) f_Y(y)$ for all x, y.

Definition 2.27. We define the expectation of a transformation g of X as $\mathrm{E}(g(X)) = \int_x g(x) f(x)$.

Definition 2.28. The **variance** of a random variable is defined as $\mathrm{Var}(X) = \mathrm{E}\left(X^2\right) - \mathrm{E}(X)^2$.

Theorem 2.29. *Expectation is linear.*

Proof. Follows from the linearity of summation and integration . ■

Definition 2.30. For multiple variables, we say:

1. $\mathrm{E}(g(X, Y)) = \int \int g(x, y) f(x, y) dx dy$
2. $\mathrm{Cov}(X, Y) = \mathrm{E}(XY) - \mathrm{E}(X) \mathrm{E}(Y)$

Theorem 2.31. ***Linear combinations.*** *Say $X_1, \ldots, X_n$ have means $\mu_1, \ldots, \mu_n$ and variances $\sigma_1^2, \ldots, \sigma_n^2$, respectively. Let $Y = \sum_{i=1}^n a_i X_i$ where $a_i \in \mathbb{R}$. Then*

1. $\mathrm{E}(Y) = \sum_{i=1}^n a_i \mu_i$
2. $\mathrm{Var}(Y) = \sum_{i=1}^n a_i^2 \sigma_i^2 + 2 \sum \sum_{i<j} a_i a_j Cov(X_i, X_j)$

Proof. Part 1 follows by linearity of expectation. Part 2 follows by Definition 2.30 and induction. ■

Definition 2.32. We say that I_A is an indicator variable if

$$I_A = \begin{cases} 1 & \text{if } A \\ 0 & \text{otherwise} \end{cases}$$

where A is an event.

Theorem 2.33. *The expectation of an indicator variable I_A is $\mathrm{E}(I_A) = P(A)$; the variance of an indicator variable is $\mathrm{Var}(I_A) = P(A)[1 - P(A)]$; the covariance of I_A and I_B is*

$$\mathrm{Cov}(I_A, I_B) = \mathrm{E}[I_A I_B] - \mathrm{E}[I_A] \mathrm{E}[I_B] = P(A \cap B) - P(A)P(B)$$

Proof. Expectation and variance follow from the Bernouilli distribution. For the covariance, drawing a joint distribution will convince us of that. ■

Example 2.34. Suppose a fair 6-sided die is rolled n times. Let X be the number of unrolled faces after n rolls. Find the mean and variance of X.

If we let X_i be an indicator variable signalling whether the number i has been rolled after n rolls. Then $\mathrm{E}[X_i] = \left(\frac{5}{6}\right)^n$. Thus, by linearity of expectation, $\mathrm{E}[X] = 6 \times \mathrm{E}[X_i] = 6 \times \left(\frac{5}{6}\right)^n$.

For the variance of the indicator we don't have to do any work: $\mathrm{Var}(X_i) = \left(\frac{5}{6}\right)^n\left[1-\left(\frac{5}{6}\right)^n\right]$. For the variance of X, we do, unfortunately. We begin by tackling the covariance of two indicator variables. We have, for $i \neq j$,

$$\mathrm{Cov}(X_i, X_j) = \mathrm{E}[X_i X_j] - \mathrm{E}[X_i]\,\mathrm{E}[X_j] = \left(\frac{2}{3}\right)^n - \left(\frac{5}{6}\right)^{2n}$$

Thus, we obtain,

$$\mathrm{Var}(X) = \sum_{i=1}^{6} \mathrm{Var}(X_i) + 2\sum\sum_{i<j} \mathrm{Cov}(X_i, X_j) = 6 \times \left(\frac{5}{6}\right)^n \left[1 - \left(\frac{5}{6}\right)^n\right] + 2 + \binom{6}{2}\left(\frac{2}{3}\right)^n - \left(\frac{5}{6}\right)^{2n}$$

Ta-da!

Definition 2.35. We say that a waiting time random variable is **proper** if $P(X < \infty) = 1$. An **improper** random variable is one where $P(X < \infty) < 1$.

Theorem 2.36. *An improper random variable has non-finite expectation.*

Proof. Duh! ∎

Remark. Note that a proper random variable does not necessarily have a finite mean.

Definition 2.37. A **short proper** random variable is a proper waiting time variable with finite mean. A **long proper** random variable is a proper waiting time variable with infinite mean.

Example 2.38. Examples for short proper variables are a dime a dozen. For long proper variables, we can use $f(x) = \frac{c}{x^2}$ for some $c \in \mathbb{R}^+$ and this works both in the continuous and discrete case.

Definition 2.39. The moment generating function (m.g.f.) of a random variable X is

$$\phi_X(t) = \mathrm{E}\left[e^{tX}\right]$$

Theorem 2.40. *We can use the moment generating function to generate moments! That is,* $\phi^{(n)}(0) = \mathrm{E}[X^n]$.

Proof. Using the Taylor series expansion for e^{tX}, we find that

$$\phi_X(t) = \mathrm{E}\left[e^{tX}\right] = \mathrm{E}\left[\sum_{k=0}^{\infty} \frac{(tX)^k}{k!}\right]$$

Since our probability function is bounded, by the Lebesgue Dominated Convergence Theorem, we can commute the differentiation operator and the infinite sum to obtain,

$$\phi_X^{(n)}(t) = \mathrm{E}\left[\sum_{k=n}^{\infty} \frac{k^{(n)} X^n (tX)^{(k-n)}}{k!}\right] = \mathrm{E}[X^n + t(\ldots)]$$

Thus $\phi_X^{(n)}(0) = \mathrm{E}[X^n]$. ∎

Theorem 2.41. *The moment generating function, under some mild regularity conditions, uniquely determines the pdf.*

Proof. Stay tuned for PMATH 352! ∎

Theorem 2.42. *If X and Y are independent random variables, then* $\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t)$.

Proof. The independence of X and Y implies the independence of e^{tX} and e^{tY}; the remainder of the proof follows from expectation algebra. ∎

Definition 2.43. The **probability generating function** (p.g.f.) of a discrete random variable on $\{0, 1, 2, \ldots\}$ is

$$G_X(s) = \mathrm{E}\left[s^X\right] = \phi_X(\log(S)) = \sum_{x=0}^{\infty} s^x p(x)$$

Chapter 3

Conditional Probability and Conditional Expectation

Definition 3.1. If X and Y are both discrete random variables with joint p.m.f. $p(x, y)$ and marginal p.m.f.s $p_X(x)$ and $p_Y(y)$, then, we denote the conditional distribution of X given Y as $X|Y = y$ and its conditional p.m.f. is:

$$P_{X|Y}(x|y) = P(X = x|Y = y) = \frac{P(X = x \cap Y = y)}{P(Y = y)}$$

Theorem 3.2. *The following are facts of life related to conditional distributions:*

1. $p_{X|Y}(x|y) \geq 0$
2. $\sum_x p_{X|Y}(x|y) = 1$
3. *If X and Y are independent, then the conditional distributions are simply the parent distributions.*

Proof. Duh. ■

Definition 3.3. The **conditional mean** of $X|(Y = y)$ is

$$\mathrm{E}\left[X|Y = y\right] = \sum_x x p_{X|Y}(x|y)$$

Theorem 3.4. *If g, h are arbitrary real valued functions, then,*

1. $\mathrm{E}\left[g(X)|Y = y\right] = \sum_x g(X) p_{X|Y}(x|y)$
2. *Conditional expectation is linear.*
3. $\mathrm{E}\left[g(X)h(Y)\right] = h(y)\mathrm{E}\left[g(X)|Y = y\right]$

Proof. The first point is the law of the unconscious statistician. Point 2 is trivial. Point three is proven as follows:

$$\mathrm{E}\left[g(X)h(Y)\right] = \mathrm{E}\left[g(X)h(y)\right] = h(y)\mathrm{E}\left[g(X)|Y = y\right]$$

because Y is fixed and thus $h(Y) = h(y)$, a constant. ■

Theorem 3.5. *If X and Y are independent, then* $\mathrm{E}\left[X|Y = y\right] = \mathrm{E}\left[X\right]$.

Proof. Trivial. ■

Definition 3.6. We define the random variable $\mathrm{E}\left[X|Y\right] = \mathrm{E}\left[X|Y = y\right]_{y=Y} = v(Y)$. We thus define the expectation of $\mathrm{E}\left[X|Y\right]$ as

$$\mathrm{E}\left[v(Y)\right] = \mathrm{E}\left[\mathrm{E}\left[X|Y\right]\right] = \sum_y v(y)p_y(y) = \sum_y \mathrm{E}\left[X|Y = y\right] p_y(y)$$

Theorem 3.7. ***Law of total expectation.*** *For random variables X and Y, we have* $\mathrm{E}\left[X\right] = \mathrm{E}\left[\mathrm{E}\left[X|Y\right]\right]$.

Proof. This follows by a few algebraic tricks:

$$\begin{aligned}
\mathrm{E}\left[\mathrm{E}\left[X|Y\right]\right] &= \sum_y \mathrm{E}\left[X|Y=y\right] p_Y(y) \\
&= \sum_y \left(\sum_x x p(x|y) \right) p_Y(y) \\
&= \sum_y \sum_x x p(x|y) p_Y(y) \\
&= \sum_x \sum_y x p(x|y) p_Y(y) \\
&= \sum_x x \sum_y p(x|y) p_Y(y) \\
&= \sum_x x \sum_y p(x,y) \\
&= \sum_x x p_X(x) \\
&= \mathrm{E}\left[X\right]
\end{aligned}$$

■