STAT 330 - Mathematical Statistics

Fantastic Theorems and How to Prove Them

Jose Luis Avilez
Faculty of Mathematics
University of Waterloo

Chapter 1

Random Variables

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. Let S be a sample space with power set $\mathcal{P}(S)$. The collection of sets $\mathcal{B} \subseteq \mathcal{P}(S)$ is called a σ-field (or σ-algebra) on S if:

1. $\emptyset \in \mathcal{B}$ and $S \in \mathcal{B}$
2. $\mathcal{B}$ is closed under complementation
3. $\mathcal{B}$ is closed under countable unions

The pair $(S, \mathcal{B})$ is called a **measurable space**.

Definition 1.3. Let S be a sample space with a sigma field $\mathcal{B} = \{A_1, A_2, \ldots\}$. A **probability set function** or **probability measure** is a function $P : \mathcal{B} \to [0, 1]$ that satisfies:

1. $P(A) \geq 0, \forall A \in \mathcal{B}$
2. $P(S) = 1$
3. If $A_1, A_2, \ldots \in \mathcal{B}$ are disjoint events, then

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

We call the triple $(S, \mathcal{B}, P)$ a **probability space**.

Definition 1.4. Consider the probability space $(S, \mathcal{B}, P)$. The function $X : S \to \mathbb{R}$ is called a **random variable** if

$$P(X \leq x) = P(\{\omega \in S : X(\omega) \leq x\})$$

is defined for all $x \in \mathbb{R}$.

Definition 1.5. The **cumulative distribution function** of a random variable X is defined as

$$F(x) = P(X \leq x)$$

for all $x \in \mathbb{R}$.

Definition 1.6. X is said to be a **discrete random variable** if its domain of values form a countable set $D(X) = \{x_1, x_2, \ldots\}$ and its probability function is defined as:

$$f(x) := P(X = x) = F(x) - \lim_{\epsilon \to 0^+} F(x - \epsilon)$$

The set $A = \{x : f(x) > 0\}$ is called the **support** of X and

$$\sum_{x \in A} f(x) = \sum_{i=1}^{\infty} f(x_i) = 1$$

Definition 1.7. A random variable X is said to be **continuous** if its cumulative distribution function is a continuous function on $\mathbb{R}$ and is differentiable everywhere except possibly at countably many points. The set $\{x : f(x) > 0\}$ is called the **support** of X and

$$\int_{x\in A} f(x)dx = 1$$

If X is continuous then the probability density function is defined to be

$$f(x) = \frac{d}{dx}F(x)$$

Definition 1.8. The **gamma function** is defined as

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

Theorem 1.9. *If X and $Y = h(X)$ are both discrete random variables, then the probability distribution of Y is given by*

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

Theorem 1.10. *If X is continuous and Y is discrete with $A = \{x : h(x) = y\}$, then*

$$P(Y = y) = \int_{x\in A} f(x)dx$$

Theorem 1.11. *If X and $Y = h(X)$ are both continuous, then*

$$F_Y(y) = P(Y \le y) = P(h(X) \le y)$$

Theorem 1.12. *Suppose h is a monotone differentiable function on the support of X, with continuous random variables X and $Y = h(X)$. Then,*

$$f_Y(y) = f_x\left(h^{-1}(y)\right)\left|\frac{d}{dy}h^{-1}(y)\right|$$

Definition 1.13. If X is a discrete random variable with p.m.f. $f(x)$ and support A, then the **expectation** or **expected value** of X is defined by:

$$\mathrm{E}(X) = \sum_{x\in A} xf(x)$$

provided that the sum converges absolutely; that is, $\mathrm{E}(|X|) < \infty$. Otherwise, we say that $\mathrm{E}(X)$ does not exist.

Definition 1.14. If X is a continuous random variable with p.d.f. $f(x)$ and support A, then the **expectation** or **expected value** of X is defined by:

$$\mathrm{E}(X) = \int_{x\in A} xf(x)$$

provided that the integral converges absolutely; that is, $\mathrm{E}(|X|) < \infty$.

Theorem 1.15. ***Probability Integral Transformation.*** *Suppose X is continuous random variable with c.d.f. F. Then $Y = F(X) \sim \mathrm{Unif}(0, 1)$.*

Proof. Since X is continuous, then F is a monotonically increasing continuous function, and is thus injective. It is thus surjective onto its range, and thus bijective and an inverse F^{-1} exists. Thus, we have,

$$P(Y \le y) = P(F(X) \le y) = P(X \le F^{-1}(y)) = F(F^{-1}(y)) = y$$

Now, we may obtain the pmf by taking the derivative

$$\frac{d}{dy}(y) = 1$$

Since this holds over $0 \le y \le 1$ it follows that Y follows a uniform distribution on the desired range. ∎

Theorem 1.16. *Suppose X is a nonnegative continuous random variable with c.d.f. $F(x)$ and finite expectation. Then*

$$\mathrm{E}\left(X\right) = \int_0^\infty \left[1 - F(x)\right] dx$$

If X is a discrete random variable with finite expectation, where $R(X) = \{1, 2, 3, \ldots\}$, then

$$\mathrm{E}\left(X\right) = \sum_{i=1}^{\infty} P(X \geq x)$$

Theorem 1.17. *Suppose that $h(X)$ is a real-valued function.*

1. *If X is a discrete random variable with p.m.f. $f(x)$ and support A, then*

$$\mathrm{E}\left(h(X)\right) = \sum_{x \in A} h(x) f(x)$$

   provided that the sum converges absolutely.

2. *If X is a continuous random variable with p.d.f. $f(x)$, then*

$$\mathrm{E}\left(h(X)\right) = \int_{-\infty}^{\infty} h(x) f(x) dx$$

   provided that the integral converges absolutely.

Proof. This is the law of the unconscious statistician. It is left as an exercise for all statisticians who used it without proof. ∎

Theorem 1.18. *Expectation is linear.*

Proof. Follows trivially from the fact that summation and integration are linear. ∎

Example 1.19. Although expectation is linear, it usually does not commute as an operator with transformations. That is, in general $\mathrm{E}\left(g(X)\right) \neq g(\mathrm{E}\left(X\right))$

Definition 1.20. The following are special cases of the expectation of transformations of X:

1. The **variance** of X is $\mathrm{Var}\left(X\right) = \mathrm{E}\left[(X - \mu)^2\right] = \mathrm{E}\left(X^2\right) - \mathrm{E}\left(X\right)^2$.
2. The **k-th moment** of X is $\mathrm{E}\left(X^k\right)$.
3. The **k-th moment about the mean** is $\mathrm{E}\left[(X - \mu)^k\right]$.
4. The **k-th factorial moment about the mean** is $\mathrm{E}\left[X(X-1)\ldots(X-k+1)\right] = \mathrm{E}\left(X^{(k)}\right) = \mathrm{E}\left[\frac{X!}{(X-k)!}\right]$.

Theorem 1.21. *Suppose X is a random variable, then* $\mathrm{Var}\left(aX + b\right) = a^2 \mathrm{Var}\left(X\right)$.

Proof. Follows from the definition of variance. ∎

Example 1.22. If $X \sim Po(\theta)$ then $\mathrm{E}\left(X^{(k)}\right) = \theta^k$. The calculation is as follows:

$$\begin{aligned}
\mathrm{E}\left[X^{(k)}\right] &= \sum_{k}^{\infty} \frac{x!}{(x-k)!} \frac{e^{-\infty}\theta^x}{x!} \\
&= \theta^k \sum_{k}^{\infty} \frac{x!}{(x-k)!} \frac{e^{-\infty}\theta^{(x-k)}}{x!} \\
&= \theta^k
\end{aligned}$$

Theorem 1.23. *If X is a random variable and $u(X)$ is a nonnegative real-values function such that $\mathrm{E}[u(X)]$ exists, then for any positive constant $c > 0$,*

$$P[u(X) \geq c] \leq \frac{\mathrm{E}[u(X)]}{c}$$

Proof. We argue as follows:

$$\begin{aligned}
\mathrm{E}[u(X)] &= \int_{x \in A} u(x) f(x) dx + \int_{x \notin A} u(x) f(x) \\
&\geq \int_{x \in A} u(x) f(x) dx \\
&\geq \int_{x \in A} c f(x) dx \\
&= c \int_{x \in A} f(x) dx \\
&= cP(X \in A) \\
&= cP(u(X) \geq c)
\end{aligned}$$

Which completes the proof. ■

Theorem 1.24. ***Markov's Inequality.*** *Suppose that X is a random variable and $k > 0$ is a constant. Then*

$$P(|X| \geq c) \leq \frac{\mathrm{E}\left[|X|^k\right]}{c^k}$$

Proof. Follows from Theorem 1.23. ■

Theorem 1.25. ***Chebyshev's Inequality.*** *Suppose X is a random variable with a finite mean μ and finite variance σ^2. Then for any $k > 0$,*

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

or, equivalently,

$$P(|X - \mu| \geq k) \leq \frac{\sigma^2}{k^2}$$

Proof. Follows from Markov's Inequality. ■

Definition 1.26. If X is a random variable then the moment generating function of X is given by

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

provided that this expectation exists for all $t \in (-h, h)$ for some $h > 0$.

Example 1.27. Let $X \sim \Gamma(\alpha, \beta)$. Then, after some algebraic mumbo-jumbo, we can find that $M_X(t) = (1 - t\beta)^{-\alpha}$ for $t < \beta^{-1}$.

Theorem 1.28. *Some properties of the moment generating function of X:*

1. $M_X(0) = 1$
2. *If the m.g.f. exists, then the k-th moment is given by* $\mathrm{E}\left[X^k\right] = M_X^{(k)}(0)$

Proof strategy. The first property is trivial. The second property can be proven by taking a Taylor expansion of e^{tX}, taking the k-th derivative of $\mathrm{E}\left[e^{tX}\right]$, using the Lebesgue Dominated Convergence Theorem to commute the differentiation and summation operators, and observe that when evaluating the expression at $t = 0$, we obtain the expectation of the k-th moment. I might post a complete proof at a later date. Might. ■

Theorem 1.29. *Suppose the random variable X has m.g.f. $M_X(t)$ defined for $t \in (-h, h)$. Let $Y = aX + b$ where $a, b \in \mathbb{R}$. Then,*

$$M_Y(t) = e^{bt} M_X(at) \qquad |t| < \frac{h}{|a|}$$

Proof. Follows from the definition of moment generating functions. ■

Theorem 1.30. ***Uniqueness theorem.*** *Suppose that X and Y have the same moment generating function over the same domain. Then X and Y have the same distribution, modulo a set of Lebesgue measure zero.*

Proof. Stay tuned for PMATH 352!

Example 1.31. Suppose $X \sim Unif(0, 1)$ and let $Y = -2 \log X$. Then using the uniqueness theorem, we can prove that $Y \sim \chi^2_2$.

Chapter 2

Joint Distributions

Definition 2.1. Suppose X and Y are random variables defined on a sample space S. Then (X, Y) is a **random vector** whose **joint cdf** is

$$F(x, y) = P(X \leq x, Y \leq y) = P[X \leq x \cap Y \leq y] \qquad (x, y) \in \mathbb{R}^2$$

Theorem 2.2. *The following are cool facts of life related to joint cdfs:*

1. *For fixed x, F is non-decreasing in y.*
2. *For fixed y, F is non-decreasing in x.*
3. $\lim_{x \to -\infty} F(x, y) = 0$ *and* $\lim_{y \to -\infty} F(x, y) = 0$
4. $\lim_{(x,y) \to (-\infty,-\infty)} F(x, y) = 0$ *and* $\lim_{(x,y) \to (\infty,\infty)} F(x, y) = 1$

Proof. Each of these follows using properties of cdfs. ■

Definition 2.3. The **marginal cdf** of X given a joint cdf $F(x, y)$ is

$$F_X(x) = P(X \leq x) = \lim_{y \to \infty} F(x, y) \qquad x \in \mathbb{R}$$

Definition 2.4. Two random variables X and Y are said to be **jointly continuous** if there exists a function $f(x, y)$ such that the joint c.d.f. of X and Y can be written as

$$F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(t_1, t_2) dt_2 dt_1 \qquad \forall (x, y) \in \mathbb{R}^2$$

We define the **joint p.d.f.** as

$$\frac{\partial^2}{\partial x \partial y} F(x, y)$$

Definition 2.5. Suppose X and Y are both continuous random variables with joint p.d.f. $f(x, y)$. The **marginal p.d.f.** of X is given by

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy$$

and the marginal p.d.f. of Y is

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

Definition 2.6. Two random variables X and Y with joint c.d.f. $F(x, y)$ are **independent** if and only if

$$F(x, y) = F_X(x) F_Y(y)$$

Equivalently, two continuous random variables X and Y with p.d.f. $f_X(x)$ and $f_Y(y)$ are independent if and only if

$$f(x, y) = f_X(x) f_Y(y) \quad \forall (x, y) \in Supp(x, y)$$

Remark. A necessary, but not sufficient, condition for independence is that the support set be a rectangle.

Theorem 2.7. ***Factorisation theorem for independence.*** *Suppose X and Y are random variables with joint p.m.f./p.d.f. $f(x,y)$ and marginal distributions $f_X(x)$ and $f_Y(y)$, respectively. Suppose also that $A = \{(x,y) : f(x,y) > 0\}$ is the support of (X,Y), $A_X = \{x : f_X(x) > 0\}$ is the support of X, and $A_Y = \{x : f_Y(y) > 0\}$ is the support of Y.*

Then X and Y are independent if and only if $A = A_X \times A_Y$ and there exist non-negative functions $g(x)$ and $h(y)$ such that $f(x,y) = g(x)h(y)$ for all $(x,y) \in A_X \times A_Y$.

Proof. The result follows from a standard result in calculus where the integral of the product is the product of the integral in a hyperrectangle, as a consequence of Fubini's theorem allowing us to switch the order of integration[1]. ∎

Definition 2.8. The **conditional distribution** of X given $Y = y$ is

$$f(x|y) = \frac{f(x,y)}{f_Y(y)}$$

Theorem 2.9. *If X and Y are independent, then $f(x|y) = f_X(x)$.*

Proof. Duh. ∎

Theorem 2.10. *Suppose X and Y are two random variables with joint p.m.f./p.d.f. $f(x,y)$ and a_i, b_i, $i = 1, \dots, n$ are constants, and $g_i(x,y)$ are real valued functions. Then,*

$$\mathrm{E}\left[\sum_{i=1}^{n} (a_i g_i(X,Y) + b_i)\right] = \sum_{i=1}^{n} (a_i \mathrm{E}\left[g(X,Y))\right]) + \sum_{i=1}^{n} b_i$$

provided each $\mathrm{E}\left[g_i(X,Y)\right]$ exist.

Proof. We prove the existence of the linear combination of the expectation using the triangle inequality. The remainder follows from linearity of summation and integration. ∎

Theorem 2.11. *If X and Y are independent random variables and $g(x)$ and $h(y)$ are real valued functions, then*

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

Proof. It follows by a simple manipulation of the integral:

$$\begin{aligned}
\mathrm{E}\left[g(X)h(Y)\right] &= \int\int_S g(x)h(y)f(x,y)dxdy \\
&= \int\int_S g(x)h(y)f_X(x)f_Y(y)dxdy \\
&= \int h(y)f_Y(y)\int g(x)f_X(x)dxdy \\
&= \mathrm{E}\left[g(X)\right]\mathrm{E}\left[h(Y)\right]
\end{aligned}$$

∎

Definition 2.12. In general, for $X_1, \dots, X_n$ we say they are **mutually independent** if

$$f(x_1, \dots, x_n) = f(x_1) \dots f(x_n)$$

Remark. Mutual independence implies pairwise independence, but the converse is not true. See Hogg p.122 for a counterexample.

[1] See Wade's Introduction to Analysis, Chapter 12.3, Problem 6a, page 418.

Theorem 2.13. *If $X_1, \ldots, X_n$ are independent random variables and $h_1, \ldots, h_n$ are real valued functions, then*

$$\mathrm{E}\left[\prod_{i=1}^{n} h_i(X_i)\right] = \prod_{i=1}^{n} \mathrm{E}\left[h_i(X_i)\right]$$

Definition 2.14. The **covariance** of random variables X and Y is given by

$$\mathrm{Cov}\,(X, Y) = \mathrm{E}\left[(X - \mu_X)(Y - \mu_Y)\right] = \mathrm{E}\left[XY\right] - \mu_X \mu_Y$$

where $\mu_X = \mathrm{E}\left[X\right]$ and $\mu_Y = \mathrm{E}\left[Y\right]$. If $\mathrm{Cov}\,(X, Y) = 0$ we say X and Y are **uncorrelated**. Note that $\mathrm{Cov}\,(X, X) = \mathrm{Var}\,(X)$.

Theorem 2.15. *If X and Y are independent random variables, then* $\mathrm{Cov}\,(X, Y) = 0$. *The converse is not true.*

Proof. $\mathrm{Cov}\,(X, Y) = \mathrm{E}\left[XY\right] - \mathrm{E}\left[Y\right]\mathrm{E}\left[Y\right] = \mathrm{E}\left[Y\right]\mathrm{E}\left[Y\right] - \mathrm{E}\left[Y\right]\mathrm{E}\left[Y\right] = 0$. For a counterexample to the converse, $Y = X^2$ over a symmetric support probably works.

Theorem 2.16. *Suppose X and Y are random variables and a, b, c are real constants. Then:*

$$\mathrm{Var}\,(aX + bY + c) = a^2\mathrm{Var}\,(X) + b^2\mathrm{Var}\,(Y) + 2ab\mathrm{Cov}\,(X, Y)$$

Proof. It follows from the definition of variance. ∎

Theorem 2.17. *Suppose $X_1, \ldots, X_n$ are random variables with* $\mathrm{Var}\,(X_i) = \sigma_i^2$, *and $a_1, a_2, \ldots a_n$ are real constants. Then*

$$\mathrm{Var}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 \sigma_i^2 + 2\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} a_i a_j \mathrm{Cov}\,(X_i, X_j)$$

Proof. Follows from the Binomial Theorem. ∎

Definition 2.18. The **correlation coefficient** of random variables X and Y is given by

$$\rho(X, Y) = \frac{\mathrm{Cov}\,(X, Y)}{\sigma_X \sigma_Y}$$

where $\sigma_X = \sqrt{\mathrm{Var}\,(X)}$ and $\sigma_Y = \sqrt{\mathrm{Var}\,(Y)}$.

Theorem 2.19. *The following are properties of the correlation coefficient:*

1. $-1 \leq \rho(X, Y) \leq 1$.
2. $\rho(X, Y) = 1 \iff Y = aX + b$ *for some* $a > 0$.
3. $\rho(X, Y) = 1 \iff Y = aX + b$ *for some* $a < 0$.

Proof. Exercise. ∎