STAT 330 - Mathematical Statistics

Fantastic Theorems and How to Prove Them

Jose Luis Avilez
Faculty of Mathematics
University of Waterloo

Contents

Chapter 1

Random Variables

1.1 Preliminaries

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*

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

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, $\mathrm{E}\,(|X|) < \infty$.

1.2 Moments, Transformations, and Inequalities

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

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,

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

Now, we may obtain the pmf by taking the derivative

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

Since this holds over $0 \leq y \leq 1$ 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*

$$\text{E}(X) = \int_0^\infty [1 - F(x)]\, dx$$

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

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

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*

$$\text{E}(h(X)) = \sum_{x \in A} h(x) f(x)$$

   provided that the sum converges absolutely.

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

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

   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 $\text{E}(g(X)) \neq g(\text{E}(X))$

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

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

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

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:

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

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

2.1 Bivariate 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. By independence,

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

■

Definition 2.10. Given a random variable X, the **expected** value of $g(X)$ is

$$\mathrm{E}\left[g(X)\right] = \sum_x g(x)f(x) \qquad \text{if } X \text{ is discrete}$$

$$\mathrm{E}\left[g(X)\right] = \int_{-\infty}^{\infty} g(x)f(x)dx \qquad \text{if } X \text{ is continuous}$$

Remark. In fact, this result requires proof starting for the definition of the expectation of a single random variable. For many years it was used without proof and it was jokingly named the "law of the unconscious statistician".

Definition 2.11. Suppose X and Y are random variables with joint distribution $f(x,y)$ with support S. Suppose $h(x,y)$ is a real-valued function. Then, the **joint expectation** under h is defined as:

$$\mathrm{E}\left[h(X,Y)\right] = \sum_{(x,y)\in S}\sum h(x,y)f(x,y) \qquad \text{if } X \text{ and } Y \text{ are discrete}$$

$$\mathrm{E}\left[h(X,Y)\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x,y)f(x,y)dxdy \qquad \text{if } X \text{ and } Y \text{ are continuous}$$

provided that the double sum/integral converge absolutely.

Theorem 2.12. *Suppose X and Y are two random variables with joint p.m.f./p.d.f. $f(x,y)$ and a_i, b_i, $i = 1, \ldots, n$ are constants, and $g_i(x,y)$ are real valued functions. Then,*

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

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. ■

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

Theorem 2.13. *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 2.14. In general, for $X_1, \ldots, X_n$ we say they are **mutually independent** if

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

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

Theorem 2.15. *If $X_1, \ldots, X_n$ are independent random variables and $h_1, \ldots, h_n$ are real valued functions, then*

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

Proof. We could use Theorem 2.13 and proceed by induction, or simply split the integral over a hyper-rectangle again. ■

Definition 2.16. The **covariance** of random variables X and Y is given by

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

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

Theorem 2.17. *If X and Y are independent random variables, then* $\mathrm{Cov}\left(X,Y\right) = 0$. *The converse is not true.*

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

Theorem 2.18. *Suppose X and Y are random variables and a, b, c are real constants. Then:*

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

Proof. It follows from the definition of variance. ■

Theorem 2.19. *Suppose $X_1, \ldots, X_n$ are random variables with* $\mathrm{Var}\left(X_i\right) = \sigma_i^2$, *and $a_1, a_2, \ldots a_n$ are real constants. Then*

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

Proof. Follows from the Binomial Theorem. ■

Definition 2.20. The **correlation coefficient** of random variables X and Y is given by

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

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

Theorem 2.21. *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. ■

Theorem 2.22. *If X and Y are independent random variables, then* $\mathrm{E}[g(Y)|x] = \mathrm{E}[g(Y)]$ *and* $\mathrm{E}[h(X)|y] = \mathrm{E}[h(X)]$.

Proof. We show the continuous case.

$$\begin{aligned}
\mathrm{E}[g(Y)|X=x] &= \int_{-\infty}^{\infty} g(y) f(y|X=x)dy \\
&= \int_{-\infty}^{\infty} g(y)\frac{f(x,y)}{f_X(x)}dy \\
&= \int_{-\infty}^{\infty} g(y)\frac{f_X(x)f_Y(y)}{f_X(x)}dy \qquad \text{(since } X \text{ and } Y \text{ are independent)} \\
&= \int_{-\infty}^{\infty} g(y) f_Y(y)dy \\
&= \mathrm{E}[g(Y)]
\end{aligned}$$

■

Definition 2.23. Let g be a real valued function, and X and Y be random variables. The **conditional expectation** of $g(Y)|X = x$ is given by,

$$\mathrm{E}[g(Y)|X=x] = \begin{cases} \sum_y g(y) f(y|X=x) & \text{if } Y|X=x \text{ is discrete} \\ \int_{-\infty}^{\infty} g(y) f(y|X=x)dy & \text{if } Y|X=x \text{ is continuous} \end{cases}$$

Remark. Using the definition of expectation, we can define the conditional expectation and variance of $Y|X = x$. Thus $\mathrm{E}[Y|X=x]$ is a number. However, $\mathrm{E}[Y|X]$ is a random variable. And a pretty useful one if you ask me.

Theorem 2.24. *Suppose X and Y are random variables, then*

$$\mathrm{E}[\mathrm{E}[g(Y)|X]] = \mathrm{E}[g(Y)]$$

If g is the identity function we obtain the ***law of total expectation****.*

Proof. We prove the continuous case.

$$\begin{aligned}
\mathrm{E}[\mathrm{E}[g(Y)|X]] &= \int_{-\infty}^{\infty} \mathrm{E}[g(Y)|X]\, f_X(x)dx \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (g(y)f(y|x)dy)\, f_X(x)dx \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left(g(y)\frac{f(x,y)}{f_X(x)}dy\right) f_X(x)dx \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (g(y)f(x,y)dy)\, dx \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (g(y)f(x,y)dx)\, dy \qquad \text{(by Fubini's Theorem)} \\
&= \int_{-\infty}^{\infty} g(y)\int_{-\infty}^{\infty} (f(x,y)dx)\, dy \\
&= \int_{-\infty}^{\infty} g(y) f_Y(y)dy \\
&= \mathrm{E}[g(Y)]
\end{aligned}$$

■

Theorem 2.25. *Suppose X and Y are random variables. Then*

$$\operatorname{Var}(Y) = \operatorname{E}[\operatorname{Var}(Y|X)] + \operatorname{Var}(\operatorname{E}[X|Y])$$

Proof. First, note that

$$\begin{aligned}
\operatorname{E}[\operatorname{Var}(X|Y)] &= \operatorname{E}\left[\operatorname{E}\left[X^2|Y\right] - \operatorname{E}[X|Y]^2\right] \\
&= \operatorname{E}\left[\operatorname{E}\left[X^2|Y\right]\right] - \operatorname{E}\left[\operatorname{E}[X|Y]^2\right] \\
&= \operatorname{E}\left[X^2\right] - \operatorname{E}\left[W^2\right] \qquad \text{where } W^2 = \operatorname{E}[X|Y]^2 \qquad (*)
\end{aligned}$$

Likewise,

$$\begin{aligned}
\operatorname{Var}(\operatorname{E}[X|Y]) &= \operatorname{Var}(W) \\
&= \operatorname{E}\left[W^2\right] - \operatorname{E}[W]^2 \\
&= \operatorname{E}\left[W^2\right] - \operatorname{E}[\operatorname{E}[X|Y]]^2 \\
&= \operatorname{E}\left[W^2\right] - \operatorname{E}[X]^2 \qquad \text{by the law of total expectation} \qquad (**)
\end{aligned}$$

Adding $(*)$ and $(**)$ yields the result. ■

Definition 2.26. The **joint moment generating function** of two random variables X and Y is defined as

$$M(t_1, t_2) = \operatorname{E}\left[e^{t_1 X + t_2 Y}\right]$$

if this expectation exists for all $t_1 \in (-h_1, h_1)$ and $t_2 \in (-h_2, h_2)$ for some $h_1, h_2 > 0$.

More generally, if $X_1, \ldots, X_n$ are random variables then

$$M(t_1, \ldots, t_n) = \operatorname{E}\left[\exp\left(\sum_{i=1}^{n} t_i X_i\right)\right]$$

is called the **joint moment generating function** of $X_1, \ldots, X_n$ if this expectation exists for all $t_i \in (-h_i, h_i)$ for some $h_i > 0$ for $i = 1, \ldots, n$.

Theorem 2.27. *Let X and Y be random variables. Then X and Y are independent if and only if their joint moment generating function is the product of their individual moment generating functions.*

Proof. This follows from the fact that, for functions of independent variables, the product of the expectation is the expectation of the product. ■

Theorem 2.28. *Given the joint moment generating function of X and Y, we have that*

$$\operatorname{E}\left[X^j Y^k\right] = \frac{\partial^{j+k}}{\partial t_1^j \partial t_2^k} M(t_1, t_2)|_{(t_1,t_2)=(0,0)}$$

Proof. The proof is similar to the case for the single variable m.g.f.; we simply use the multivariable Taylor's theorem for this case.

2.2 Multivariate Distributions

Now we transition from bivariate to multivariate. To be frank, studying bivariate distributions as special cases of bivariate distributions is a distraction. Everything is the same. So we repeat everything below, probably without proof.

Definition 2.29. The k-variate **cumulative density function** of $X_1, \ldots, X_k$ is

$$F(x_1, \ldots, x_k) = P(X_1 \leq x_1, \ldots, X_k \leq x_k)$$

In the continuous case, the **joint probability density function** is[2]

$$f(x_1, \ldots, x_k) = \frac{\partial^k}{\partial x_1 \ldots \partial x_k} F(x_1, \ldots, x_k)$$

The usual properties of c.d.f.s. and p.d.f.s. hold.

Definition 2.30. Random variables $X_1, \ldots, X_n$ are said to be **independent** if and only if

$$f(x_1, \ldots, x_n) = f_{X_1}(x_1) \ldots f_{X_n}(x_n)$$

Theorem 2.31. *Independence implies pair-wise independence. However, pairwise independence for all $i \neq j$ does not imply independence.*

Proof. Exercise. ■

Definition 2.32. The **joint moment generating function** of $X_1, \ldots, X_k$ is

$$M(t_1, \ldots, t_k) = \mathrm{E}\left[\exp\left(t_1 X_1 + \ldots + t_k X_k\right)\right]$$

Example 2.33. The multinomial distribution and the multivariate normal distribution are canonical examples of multivariate distributions.

Definition 2.34. A random vector $X = (X_1, \ldots, X_k)$ is said to follow a **multinomial distribution** if it has the p.d.f.

$$f(x_1, x_2, \ldots, x_k) = \frac{n!}{x_1! x_2! \ldots x_k! (n - x_1 - \ldots - x_k)!} p^{x_1} p^{x_2} \ldots p^{x_k} p^{(n - x_1 - \ldots x_k)}$$

We say that $X = (X_1, \ldots, X_k) \sim \mathrm{MULT}(n, p_1, \ldots, p_k)$.

Theorem 2.35. *Let $X = (X_1, \ldots, X_k)$ follow a multinomial distribution. Denote $p_{k+1} = 1 - (p_1 + \ldots p_k)$. Then*

1. *The joint moment generating function for X is*

$$M(t_1, \ldots, t_k) = \mathrm{E}\left[\exp\left(t_1 X_1 + \ldots + t_k X_k\right)\right] = \left(p_1 e^{t_1} + \ldots + p_k e^{t_k} + p_{k+1}\right)^n$$

2. *Any combination of the random variables $X_{i_1}, \ldots X_{i_m}$ also follows a multinomial distribution. In particular, $X_i \sim B(n, p_i)$.*

3. *If $T = X_i + X_j$ with $i \neq j$ then $T \sim B(n, p_i + p_j)$.*

4. *$\mathrm{Cov}\left(X_i, X_j\right) = -n p_i p_j$ whenever $i \neq j$*

5. *$X_i | X_j = x_j \sim Bin\left(n - x_j, \frac{p_i}{1 - p_j}\right)$*

6. *$X_i | (X_i + X_j = t) \sim Bin\left(t, \frac{p_i}{p_i + p_j}\right)$*

[2] For the purposes of this course we assume that F is a $\mathcal{C}^k$ function for some large k. This allows us to reorder the partial differentiation operator.

Proof. (1) − (3), (6) should be obvious. We prove the remaining.

(4) To compute the covariance, we use the definition; hence, we first compute $\mathrm{E}[X_iX_j]$ using the moment generating function's second order partials:

$$\frac{\partial M}{\partial t_j} = np_je^{t_j}\left(p_1e^{t_1}+\ldots+p_ke^{t_k}+p_{k+1}\right)^{n-1} \qquad \frac{\partial^2 M}{\partial t_i\partial t_j} = n(n-1)p_ip_j\left(p_1e^{t_1}+\ldots+p_ke^{t_k}+p_{k+1}\right)^{n-2}$$

Evaluating the second order partial at $(t_i,t_j)=(0,0)$, we obtain $\mathrm{E}[X_iX_j] = n(n-1)p_ip_j$. Using the fact that each component of the random vector behaves binomially we obtain,

$$\mathrm{Cov}(X_i,X_j) = n(n-1)p_ip_j\mathrm{E}[X_i]\,\mathrm{E}[X_j] \;= n(n-1)p_ip_j - n^2p_ip_j = -np_ip_j$$

(5) We use the definition of conditional probability:

$$\begin{aligned} f(x_i|x_j) &= \frac{f(x_i,x_j)}{f_{X_j}(x_j)} \\ &= \frac{\frac{n!}{x_i!x_j!(n-x_i-x_j)!}p_i^{x_i}p_j^{x_j}(1-p_i-p_j)^{n-x_i-x_j}}{\frac{n!}{x_j!(n-x_j)!}p_j^{x_j}(1-p_j)^{n-x_j}} \\ &= \frac{(n-x_j)!}{x_i!(n-x_j-x_i)!}\left(\frac{p_i}{1-p_j}\right)^{x_i}\left(\frac{1-p_i-p_j}{1-p_j}\right)^{n-x_j-x_i} \end{aligned}$$

and the result follows. ∎

Definition 2.36. Let X_1 and X_2 be random variables with p.d.f.

$$f(x_1,x_2) = \frac{1}{2\pi|\Sigma|^{1/2}}\exp\left(-\frac{1}{2}(x-\mu)^t\Sigma^{-1}(x-\mu)\right)$$

where $(x_1,x_2)\in\mathbb{R}^2$ and

$$x = \begin{bmatrix}x_1\\x_2\end{bmatrix} \qquad \mu = \begin{bmatrix}\mu_1\\\mu_2\end{bmatrix} \qquad \Sigma = \begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}$$

and Σ is invertible. Then $X=(X_1,X_2)^t$ is said to have a **bivariate normal distribution**. We write $X\sim BVN(\mu,\Sigma)$. The multivariate normal distribution is similarly written using a mean vector and a variance-covariance matrix and we say that $X\sim MVN(\mu,\Sigma)$

Theorem 2.37. *For a bivariate normal distribution, if $\rho=0$, then its p.d.f. is the product of two independent normal distributions.*

Proof. Follows from factorisation theorem. ∎

Theorem 2.38. *The following are properties of the bivariate normal distribution:*

1. *The joint m.g.f. is given by*

$$M(t_1,t_2) = \exp\left(\mu^tt+\frac{1}{2}t^t\Sigma t\right) \quad \forall t=(t_1,t_2)\in\mathbb{R}^2$$

2. *$X_1\sim N(\mu_1,\sigma_1^2)$ and $X_2\sim N(\mu_2,\sigma_2^2)$*
3. *$\mathrm{Cov}(X_1,X_2)=\rho\sigma_1\sigma_2$ and $\mathrm{Corr}(X_1,X_2)=\rho$ where $-1\le\rho\le1$.*
4. *X_1 and X_2 are independent random variables if and only if $\rho=0$.*
5. *If $c=(c_1,c_2)^t$ is a non-zero vector of constants, then*

$$c^tX\sim N(c^t\mu, c^t\Sigma c)$$

6. *If A is a 2×2 invertible constant matrix and b is a 2×1 constant vector, then*

$$Y = AX + b \sim BVN(A\mu + b, A\Sigma A^t)$$

7. *The conditional distributions are:*

$$X_2|X_1 = x_1 \sim N\left(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x_1 - \mu_1), \sigma_2^2(1-\rho^2)\right)$$

and

$$X_1|X_2 = x_2 \sim N\left(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(x_2 - \mu_2), \sigma_1^2(1-\rho^2)\right)$$

8. *We can relate the normal distribution with the chi-squared distribution.*

$$(X-\mu)^t\Sigma^{-1}(X-\mu) \sim \chi^2(2)$$

Proof. Exercise. ■

Definition 2.39. Random variables $X_1, \ldots, X_n$ are said to form a **simple random sample** or are said to be **independent and identically distributed** (shortened to IID) if $X_1, \ldots, X_n$ are independent and $f_{X_i} = f_{X_j}$ for all $i \neq j$.

Chapter 3

Functions of Random Variables

In this chapter we are concerned with two main questions. Suppose X and Y are two continuous variables with joint probability distribution $f(x, y)$. Then, we ask:

1. What is the distribution of $U = h(X, Y)$?
2. What is the joint distribution of $U = h_1(X, y)$ and $V = h_2(X, Y)$?

3.1 Transformations

Example 3.1. Suppose $X_1, \dots X_n$ are an i.i.d. sample from continuous distribution, each with p.d.f. $f(X)$ and c.d.f. $F(x)$. We try to find:

1. $T = \min(X_1, \dots, X_n) = X_{(1)}$
2. $Y = \max(X_1, \dots, X_n) = X_{(n)}$

We find the c.d.f. for T:

$$\begin{aligned} P(T \leq t) &= 1 - P(T > t) \\ &= 1 - P(\min(X_1, \dots, X_n) > t) \\ &= 1 - P(X_1 > t, X_2 > t, \dots, X_n > t) \\ &= 1 - \prod_{i=1}^{n} P(X_i > t) \qquad \text{since } X_i\text{'s are independent} \\ &= 1 - \prod_{i=1}^{n} P(X_1 > t) \qquad \text{since } X_i\text{'s are identical} \\ &= 1 - [1 - F_X(t)]^n \end{aligned}$$

To find the p.d.f., we take the derivative and arrive at:

$$f_Y(t) = \frac{d}{dt} F_T(t) = n f_X(t) \left(1 - F_X(t)\right)^{n-1}$$

where the support of T is the same as the support of X_i.

For the maximum, we obtain:

$$\begin{aligned} P(Y \leq y) &= P(\max(X_1, \dots, X_n) \leq y) \\ &= P(X_1 < y, X_2 < y, \dots, X_n < y) \\ &= \prod_{i=1}^{n} P(X_i < y) \qquad \text{since } X_i\text{'s are independent} \\ &= \prod i = 1^n P(X_1 < y) \qquad \text{since } X_i\text{'s are identical} \\ &= [F_X(y)]^n \end{aligned}$$

Likewise, the p.d.f. is obtained from the derivative

$$f_Y(y) = \frac{d}{dy}\left[F_X(y)\right] = n\left[F_x(y)\right]^{n-1} f_X(y)$$

Exercise. Find the joint distribution of T and Y.

Definition 3.2. Let $S : (x, y) \to (u, v)$ be a one-to-one function such that $u = h_1(x, y)$ and $v = h_2(x, y)$. Since S is one-to-one, there exists an inverse function $T = S^{-1}$ such that

$$x = w_1(u, v) \qquad y = w_2 = w_2(u, v)$$

The **Jacobian of the transformation** T is

$$|J| = \det\left(\frac{\partial(x, y)}{\partial(u, v)}\right) = \det\begin{pmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{pmatrix}$$

Theorem 3.3. ***One-to-one bivariate transformations.*** *Given continuous variables X and Y with joint p.d.f. $f(x, y)$ and support $R_{XY} = \{(x, y) : f(x, y) > 0\}$, let $U = h_1(X, Y)$ and $V = h_2(X, Y)$, with the transformation being one-to-one with inverse*

$$X = w_1(U, V) \qquad Y = w_2(U, V)$$

Suppose also that S maps R_{XY} to $\mathbb{R}_{UV}$. Then the joint distribution $g(u, v)$ is given by

$$g(u, v) = f(w_1(u, v), w_2(u, v))|J|$$

where J is the Jacobian of the transformation.

Proof. The proof hinges on the proof for the change of variables formula for integration, which is particularly hard. We leave this without proof. ■

Theorem 3.4. ***Linear combinations of independent random variables.*** *If $X_i \sim N(\mu_i, \sigma_i^2)$ with $i = 1, 2, \ldots, n$ independently, then*

$$\sum_{i=1}^{n} a_i X_i \sim N\left(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2\right)$$

Proof. We argue by using their moment generating functions. The details are left as an exercise. ■

Theorem 3.5. *Assume $X_1, \ldots X_n \sim N(\mu, \sigma^2)$ independently. Define $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2)}{n-1}$. Then,*

1. $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$
2. $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$
3. $\frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$

Proof. We prove each sub-theorem independently.

1. Denote $Y_i = \frac{X_i}{n}$ and using the theorem above note that $Y_i \sim N(\frac{\mu}{n}, \frac{\sigma^2}{n^2})$; furthermore, $\overline{X} = \sum_{i=1}^{n} Y_i$. Using the moment generating function of the sum of independent variables, we obtain,

$$M_{\overline{X}}(t) = \prod_{i=1}^{n} M_{Y_i}(t) = e^{\mu t + \frac{t\sigma^2}{2n}}$$

   Thus, by the uniqueness of the MGF, the result follows.

2. We use a little trick:

$$\sum_{i=1}^{n}(X_i-\mu)^2=\sum_{i=1}^{n}(X_i-\overline{X}+\overline{X}\mu)^2$$
$$=\sum_{i=1}^{n}(X_i-\overline{X})^2+n(\overline{X}-\mu)^2 \qquad \text{(exercise)}$$

We divide through by σ^2 to obtain,

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2=\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\overline{X})^2+\frac{n}{\sigma^2}(\overline{X}-\mu)^2$$
$$\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2=\sum_{i=1}^{n}\left(\frac{X_i-\overline{X}}{\sigma}\right)^2+\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right)^2$$

Now, we have that $\frac{X_i-\mu}{\sigma}\sim N(0,1)$, so the sum of its squares follows a $\chi^2(n)$ distribution. Likewise, $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$, so its square follows a $\chi^2(1)$ distribution. Thus, by Cochran's Theorem[1], the independence of the LHS and the right term in the RHS, implies that the middle term follows a $\chi^2(n-1)$ distribution.

3. From the sub-theorem above, we have that $\overline{X}$ and S^2 are independent. We can write,

$$\frac{\overline{X}-\mu}{S/\sqrt{n}}=\frac{\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}}$$

The numerator follows a $N(0,1)$ distribution and the denominator is the quotient of a $\chi^2(n-1)$ distribution when divided by $n-1$. This is by definition a $t(n-1)$ distribution, which completes the proof. ■

Definition 3.6. A random variable is said to follow a F_{v_1,v_2} distribution if

$$F_{v_1,v_2}=\frac{\chi^2_{v_1}/v_1}{\chi^2_{v_2}/v_2}$$

for independent $\chi^2_{v_1}$ and $\chi^2_{v_2}$ distributions.

Theorem 3.7. *Suppose $X_1,\ldots,X_n$ is a random sample from a $N(\mu_1,\sigma_1^2)$ distribution independently. Suppose that $Y_1,\ldots,Y_m$ is a random sample from a $N(\mu_2,\sigma_2^2)$ distribution. Let $S_1^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2$ and $S_2^2=\frac{1}{n-1}\sum_{i=1}^{m}(Y_i-\overline{Y})^2$. Then,*

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}\sim F(n-1,m-1)$$

Proof. We can write the following,

$$\frac{\sum_{i=1}^{n}\left(\frac{X_i-\overline{X}}{\sigma_1}\right)^2}{\sum_{i=1}^{m}\left(\frac{Y_i-\overline{Y}}{\sigma_2}\right)^2}=\frac{(n-1)S_1^2/\sigma_1^2}{(m-1)S_2^2/\sigma_2^2}$$

Where we use $S_1=\sum_{i=1}^{n}(X_i-\overline{X})$ and similarly for Y. Then, we have,

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}=\frac{\frac{(n-1)S_1^2}{\sigma_1^2}/(n-1)}{\frac{(m-1)S_2^2}{\sigma_2^2}/(m-1)}\sim\frac{\chi^2(n-1)/(n-1)}{\chi^2(m-1)/(m-1)}=F_{n-1,m-1}$$

as required. ■

[1] I might add a section about it later. It is not necessary material

3.2 Special Distributions

In this subsection we state results for special distributions, and leave the proofs as an exercise using the methods described in this chapter.

Theorem 3.8. *Suppose $X \sim \Gamma(\alpha, \beta)$ where α is a positive integer. Then,*

$$\frac{2X}{\beta} \sim \chi^2(2\alpha)$$

Proof. Follows from the uniqueness theorem for moment generating functions. ■

Theorem 3.9. *Suppose $X_i \sim \Gamma(\alpha_i, beta)$ for $i = 1, \ldots, n$ independently. Then,*

$$Y = \sum_{i=1}^{n} X_i \sim \Gamma\left(\sum_{i=1}^{n} \alpha_i, \beta\right)$$

Proof. Simple application of moment generating functions. ■

Theorem 3.10. *Suppose $X_i \sim Exp(\beta)$ for $i = 1, \ldots, n$ independently. Then*

$$Y = \sum_{i=1}^{n} X_i \sim \Gamma(n, \beta)$$

Proof. Note that $Exp(\beta) = \Gamma(1, \beta)$. The result the follows from the above theorem. ■

Theorem 3.11. *Suppose $X_i \sim \Gamma\left(\frac{k_i}{2}, 2\right) = \chi^2(k_i)$ independently. Then*

$$Y = \sum_{i=1}^{n} X_i \sim \chi^2\left(\sum_{i=1}^{n} k_i\right)$$

Proof. Use distributions above. ■

Theorem 3.12. *Suppose $X_i \sim N(\mu, \sigma^2)$ independently. Then*

$$Y = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi^2(n)$$

Proof. The shifted normal follows a standard normal distribution. Its square follos a chi-squared distribution with one degree of freedom (independently). Use the uniqueness theorem to complete the result.

Theorem 3.13. *Suppose $X_i \sim Poi(\lambda_i)$. Then,*

$$Y = \sum_{i=1}^{n} X_i \sim Po\left(\sum_{i=1}^{n} \lambda_i\right)$$

Proof. Use MGF method. ■

Theorem 3.14. *Suppose $X_i \sim NB(k_i, p)$ independently. Then,*

$$\sum_{i=1}^{n} X_i \sim NB\left(\sum_{i=1}^{n} k_i, p\right)$$

Theorem 3.15. *Suppose $X_i \sim Bin(k_i, p)$ independently. Then,*

$$\sum_{i=1}^{n} X_i \sim Bin\left(\sum_{i=1}^{n} k_i, p\right)$$

Chapter 4

Limiting or Asymptotic Distributions

We shift our interest to discussing the distribution of $g(X_1, \ldots, X_n)$ whenever we do not have enough information. That may be that g or the X_i's might be too complicated. Thus, we try to approximate the distribution $g(X_1, \ldots, X_n) \approx Y$ under mild assumptions on each X_i and if n is large.

4.1 Convergence in Distribution

Definition 4.1. Let (X_n) be a sequence of random variables such that X_n has a c.d.f. $F_n(x)$ for all n. Let X be a random variable with c.d.f. $F(x)$. We say that X_n **converges in distribution** to X and we write

$$X_n \xrightarrow{D} X$$

if

$$\lim_{n\to\infty} F_n(x) = F(x) \qquad \forall x \in \mathbb{R}$$

at all points x at which F is continuous.

Remark. Just because the c.d.f.s converge, this does not mean that the random variables converge. Also, this definition of convergence holds for both continuous and discrete variables.

Theorem 4.2. *If $b, c \in \mathbb{R}$ are constants and $\lim_{n\to\infty} \psi(n) = 0$, then*

$$\lim_{n\to\infty} \left[1 + \frac{b}{n} + \frac{\psi(n)}{n}\right]^{cn} = e^{bc}$$

As a corollary,

$$\lim_{n\to\infty} \left(1 + \frac{b}{n}\right)^{cn} = e^{bc}$$

Proof. Follows from MATH 147.[1] ■

Definition 4.3. The function $F(y)$ is the c.d.f. of a **degenerate distribution** at a value $y = c$ if

$$F(y) = \begin{cases} 0 & y < c \\ 1 & y \geq c \end{cases}$$

Example 4.4. Let $X_i \sim U(0, \theta)$ for $i \in \mathbb{N}^*$. Define the sequence $(Y_n)_{n=1}^{\infty}$ with $Y_n = \max(X_1, \ldots X_n)$. The limiting distribution of Y is degenerate.

[1] Warning. I think the proof that was given in class is circular.

4.2 Convergence in Probability

Definition 4.5. A sequence of random variables (X_n) is said to **converge in probability** to a random variable X if, for all $\epsilon > 0$, we have

$$\lim_{n\to\infty} P(|X_n - X| \geq \epsilon) = 0$$

or equivalently,

$$\lim_{n\to\infty} P(|X_n - X| < \epsilon) = 1$$

If so, we write

$$X_n \xrightarrow{P} X$$

Definition 4.6. A sequence of random variables (X_n) is said to **converge in probability to a constant** c to a random variable X if, for all $\epsilon > 0$, we have

$$\lim_{n\to\infty} P(|X_n - c| \geq \epsilon) = 0$$

or equivalently,

$$\lim_{n\to\infty} P(|X_n - c| < \epsilon) = 1$$

If so, we write

$$X_n \xrightarrow{P} c$$

Theorem 4.7. *Convergence in probability implies convergence in distribution. Notationally,*

$$X_n \xrightarrow{P} X \quad \Longrightarrow \quad X_n \xrightarrow{D} X$$

Proof. We take this proof from Hogg. Let x be a point of continuity of $F_X(x)$. Fix $\epsilon > 0$. Then,

$$\begin{aligned} F_{X_n}(x) &= P[X_n \leq x] \\ &= P[(X_n \leq x) \cap (|X_n - X| < \epsilon)] + P[(X_n \leq x) \cap (|X_n - X| \geq \epsilon)] \\ &\leq P[X \leq x + \epsilon] + P[|X_n - X| \geq \epsilon] \end{aligned}$$

Since X_n converges in probability to X, we have

$$\limsup F_{X_n}(x) \leq F_X(x + \epsilon)$$

For a lower bound, we have

$$P[X_n > x] \leq P[X \geq x - \epsilon] + P[|X_n - X| \geq \epsilon]$$

and thus,

$$\liminf F_{X_n}(x) \geq F_X(x - \epsilon)$$

Combining the two inequalities, we obtain

$$F(x - \epsilon) \leq \liminf F_{X_n}(x) \leq \limsup F_{X_n}(x) \leq F_X(x + \epsilon)$$

Applying the squeeze theorem we obtain the desired result. ■

Theorem 4.8. ***Weak Law of Large Numbers.*** *Let (X_n) be a sequence of independent and identically distributed random variables with common mean μ and finite variance σ^2. Define the random variable $\overline{X}_n = n^{-1}\sum_{i=1}^{n} X_i$. Then,*

$$\overline{X}_n \xrightarrow{P} \mu$$

Proof. We use Chebyshev's Inequality. Note that the mean and variance of $\overline{X}_n$ are μ and $\frac{\sigma^2}{n}$, respectively. Fix $\epsilon > 0$. Then,

$$\begin{aligned} P[|\overline{X}_n - \mu| \geq \epsilon] &= P\left[|\overline{X}_n - \mu| \geq k\frac{\sigma}{\sqrt{n}}\right] \qquad \text{where } k = \frac{\epsilon\sqrt{n}}{\sigma} \\ &\leq \frac{\sigma^2}{n\epsilon^2} \to 0 \end{aligned}$$

as required. ■

Remark. What the weak law of large numbers is saying is that the more we sample a dataset, the mass of the distribution of $\overline{X}_n$ will tend to be μ.

Theorem 4.9. *Convergence in probability is closed under linear combinations. That is, if $a \in \mathbb{R}$, $X_n \xrightarrow{P} X$, and $Y_n \xrightarrow{P} Y$ then $X_n + Y_n \xrightarrow{P} X + Y$ and $aX_n \xrightarrow{P} aX$.*

Proof. We want to show that for all $\epsilon > 0$, we have

$$\lim_{n\to\infty} P[|(X_n + Y_n) - (X + Y)| \geq \epsilon] = 0$$

By the triangle inequality,

$$\epsilon \leq |(X_n + Y_n) - (X + Y)| = |(X_n - X) + (Y_n - Y)| \leq |X_n - X| + |Y_n - Y|$$

Then, using the fact that P respects sets inclusion monotonically, we have,

$$P[|(X_n + Y_n) - (X + Y)| \geq \epsilon] \leq P[|X_n - X| + |Y_n - Y| \geq \epsilon] \geq P\left[|X_n - X| \geq \frac{\epsilon}{2}\right] + P\left[|Y_n - Y| \geq \frac{\epsilon}{2}\right]$$

which go to zero by assumption. The proof of the second part is immediate. ■

Theorem 4.10. *Suppose $X_n \xrightarrow{P} a$ and the real function $g(x)$ is continuous at a. Then*

$$g(X_n) \xrightarrow{P} g(a)$$

Proof. Fix $\epsilon > 0$. Since g is continuous at a, for all $\epsilon > 0$, there exists a $\delta > 0$ such that whenever $|x - a| < \delta$ then $|f(x) - f(a)| < \epsilon$. The contrapositive of this last implication is that, for g, if $|g(x) - g(a)| \geq \epsilon$ then $|x - a| \geq \delta$. Thus, we have that

$$P[|g(X_n) - g(a)| \geq \epsilon] \leq P[|X_n - a| \geq \delta]$$

By hypothesis the last term goes to zero as $n \to \infty$ which completes the proof. ■

Theorem 4.11. *Suppose $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$. The $X_nY_n \xrightarrow{P} XY$.*

Proof. We use an algebraic trick.

$$\begin{aligned} X_nY_n &= \frac{1}{2}X_n^2 + \frac{1}{2}Y_n^2 - \frac{1}{2}(X_n - Y_n)^2 \\ &\xrightarrow{P} \frac{1}{2}X^2 + \frac{1}{2}Y^2 - \frac{1}{2}(X - Y)^2 \\ &= XY \end{aligned}$$

■

Theorem 4.12. *Let (Y_n) be a sequence of random variables with moment generating functions $(M_n(t))$ defined on a common neighbourhood around 0. Then Y_n converges in distribution to Y if and only if*

$$\lim_{n\to\infty} M_n(t) = M(t), \qquad t \in (-h, h)$$

where $M(t)$ is the moment generating function of the limiting variable Y.

Proof. Will provide later. ■

Theorem 4.13. *Suppose $X_1, X_2, \ldots$ is a sequence of independent random variables with $\mathrm{E}[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2 < \infty$. Then*

$$\frac{\sqrt{n}\left(\overline{X}_n - \mu\right)}{\sigma} \xrightarrow{P} N(0,1)$$

Proof. The proof is complicated. Instead, we prove the special case where the moment generating function of X_i exists.

Let $M_j(t)$ be the sequence moment generating functions for the sequence $X_i - \mu$. Note that

$$M_j(0) = 1 \qquad M_j'(0) = 0 \qquad M_j''(0) = \sigma^2 \qquad (*)$$

We look at its Taylor expansion:

$$M_j(t) = \sum_{i=0}^{k} \frac{M_j^{(i)}(0)t^i}{i!} + \frac{M_j^{(i+1)}(c)t^{i+1}}{(i+1)!} \qquad 0 < c < t$$

We will focus on the second order expansion of the Taylor series using . Now look at

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu)$$

And the moment generating function of Z_n is

$$\begin{aligned} M_{Z_n}(t) &= \mathrm{E}\left[\exp\left(t\left(\frac{1}{\sigma\sqrt{n}}\sum_{i=1}^{n}(X_i-\mu)\right)\right)\right] \\ &= M_{\sum_{i=1}^{n}(X_i-\mu)}\left(\frac{t}{\sigma\sqrt{n}}\right) \\ &= \left[1 + \frac{t^2}{2n} + \frac{\left(M^{(2)} - \sigma^2\right)}{2n\sigma^2}t^3\right]^n \end{aligned}$$

Which converges to $e^{\frac{t^2}{2}}$ as $n \to \infty$[2]. Since this is the moment generating function for the standard normal distribution, the result holds. ■

Theorem 4.14. *Let (X_n) be a sequence of random variables such that*

$$n^b(X_n - a) \xrightarrow{D} X$$

for some $b > 0$. Suppose the function $g(x)$ at a and $g'(a) \neq 0$. Then

$$n^b\left[g(X_n) - g(a)\right] \xrightarrow{D} g'(a)X$$

Remark. We state this without proof, but we do prove a corollary.

Theorem 4.15. *Let (X_n) be a sequence of i.i.d. random variables with mean μ and variance σ^2. Suppose the function $g(x)$ is differentiable at μ and $g'(\mu) \neq 0$. Then,*

$$\sqrt{n}\left[g(\overline{X}_n) - g(\mu))\right] \xrightarrow{D} Z \sim N(0, g'(\mu)^2\sigma^2)$$

Proof. Will provide later.

[2] I'll fill in the details later.

Chapter 5

Estimation

Suppose $X_1, \ldots, X_n$ are i.i.d. following a p.d.f. $f(x;\theta)$ where θ is a vector of parameters. In this chapter we are interested in estimating the value of θ (**estimation**) and checking the validity of claims about θ (**inference**). This set-up is called a **parametric model**.

Definition 5.1. A **statistic** $T = T(X) = T(X_1, \ldots, X_n)$ is a function of the data which does not depend on any unknown parameter.

Example 5.2. The sample mean and sample variance are examples of statistics. In fact, any sample central moment is a statistic.

Example 5.3. Order statistics (such as max or min) are statistics.

Example 5.4. The quantity $\frac{\overline{X}-\mu}{s/\sqrt{n}}$ is not a statistic because μ is an unknown parameter.

Definition 5.5. A statistic T which is used to estimate an unknown parameter θ, or a function of theta, say $\tau(\theta)$, is called an **estimator**. If $T(X)$ estimates $\tau(\theta)$, an observed value of the statistic $t = t(x) = t(x_1, \ldots, x_n)$ is called an **estimate** of $\tau(\theta)$.

Example 5.6. The function

$$T(X) = \frac{1}{n}(X_1 + \ldots + X_n)$$

is an estimator (since it is a random variable, whereas,

$$t(x) = \frac{1}{n}(x_1 + \ldots x_n)$$

is an estimate.

Remark. As a convention at the University of Waterloo, we denote an estimator for θ as $\tilde{\theta}$.
Note. Some desirable properties of an estimator $\tilde{\theta}$ are

1. Unbiasedness: $\mathrm{E}\left[\tilde{\theta}\right] = \theta$
2. Small variability: $\mathrm{Var}\left(\tilde{\theta}\right)$ is small.
3. Consistency: $\tilde{\theta} \xrightarrow{P} \theta$

Definition 5.7. Maximum Likelihood Principle. Suppose $X_1, \ldots, X_n$ form an i.i.d. random sample from a discrete distribution $f(x;\theta)$. In this case, the joint distribution of $(X_1, \ldots, X_n)$ is

$$\prod_{i=1}^{n} f(y_i;\theta)$$

If $x_1, \ldots, x_n$ is the observed sample, then we define the **likelihood function**

$$L(\theta) = P(X_1 = x_1, X_2 = x_2, \ldots xX_n = x_n) = \prod_{i=1}^{n} f(x_i; \theta)$$

We call $\hat{\theta}$ the **maximum likelihood estimate** whenever

$$\hat{\theta} = \text{argmax}_{\theta \in \Omega} L(\theta)$$

Theorem 5.8. *The maximum likelihood principle holds for a continuous random variable X.*

Proof. Suppose the random variable X has a p.d.f. $f(x, \theta)$ and suppose we observe the value x. Then, for a small enough $\delta > 0$, we have

$$P(x - \delta/2 < X < x + \delta/2) \approx \delta f(x; \theta)$$

Since maximising the probability that observing a point in $(x - \delta/2, x + \delta/2)$ is observed given θ does not depend on θ, the principle holds. ■

Definition 5.9. The **log-likelihood function** is

$$l(\theta) = \log(L(\theta)) = \sum_{i=1}^{n} \log\left(f(x_i; \theta)\right)$$

Definition 5.10. The **score function** is defined as the first derivative of the log-likelihood function:

$$S(\theta) = \frac{dl}{d\theta} = \frac{d}{d\theta} l(\theta)$$

Definition 5.11. Suppose $X_1, \ldots, X_n$ are i.i.d. random variables with probability function $f(x; \theta)$ and log-likelihood $l(\theta)$. The **information function** of θ, denoted $I(\theta)$, is a (random) function defined by:

$$I(\theta) = -\frac{d^2}{d\theta^2} l(\theta) = -\frac{d}{d\theta} S(\theta)$$

Definition 5.12. If θ is a scalar, then the **expected** or **Fisher information** is given by:

$$J(\theta) = \mathrm{E}\left[I(\theta; X)\right] = \mathrm{E}\left[-\frac{\partial^2}{\partial \theta^2} I(\theta; X)\right]$$

Example 5.13. For any distribution from an exponential family, the equality $J(\theta) = \frac{1}{\text{Var}(MLE(\theta))}$ holds. However, this is not true in general.

Remark. The observed information depends on the sample. However, the Fisher information is the same provided that the distribution remains the same.

Example 5.14. We look at an example where the support set depends on the parameter being estimated. Suppose $X_1, \ldots, X_n$ is a random sample from UNIF$(0, \theta)$. Find the maximum likelihood estimator of θ.

Note that

$$f(x_i; \theta) = \frac{1}{\theta} \qquad 0 \leq x_i \leq \theta$$

For the likelihood function, we obtain

$$L(\theta) = \frac{1}{\theta^n} \qquad \text{where } 0 \leq x_i \leq \theta$$

or, equivalently for the support $x_{(n)} \leq \theta$. Thus,

$$MLE(\theta) = \text{argmax}_{\theta \in \Omega} L(\theta) = \text{argmax}_{\theta \geq x_{(n)}} L(\theta) = x_{(n)}$$

That is, the maximum likelihood estimate for θ is simply $\max(X_1, \ldots, X_n)$.

Example 5.15. Suppose $X_1, \ldots, X_n$ is a random sample from UNIF$(\theta, \theta + 1)$. Find the MLE of θ.

Using the same reasoning as above, we observe that θ must satisfy $\theta \geq x_{(1)}$ and $\theta \leq x_{(n)} - 1$. Note, furthermore, that the likelihood function is constant. We conclude that the maximum likelihood of θ is not unique and can take any values between $x_{(n)} - 1$ and $x_{(1)}$.

Theorem 5.16. ***Invariance property of MLE****. Suppose $\tau = h(\theta)$ is a one-to-one function of θ. Suppose also that $\hat{\theta}$ is the M.L. estimator of θ. Then, $\hat{\tau} = h(\hat{\theta})$ is the M.L. estimator of τ.*

Proof. Maybe later. ■

Definition 5.17. Suppose $X_1, \ldots, X_n$ are independently and identically distributed with distribution $f(x; \theta)$ where the likelihood function is $L(\theta)$ and the MLE of θ is $\hat{\theta}$. The **relative likelihood function** $R(\theta)$ is defined by

$$R(\theta) = R(\theta; x) = \frac{L(\theta)}{L(\hat{\theta})} \qquad \theta \in \Omega$$

Definition 5.18. The set of θ values for which $R(\theta) \geq p$ is called a $100p\%$ **likelihood region** for θ. If the region is an interval of real values then it is called a $100p\%$ **likelihood interval** for θ.

Remark. Values inside a 10% L.I. are called plausible; those outside are implausible.

Remark. If the underlying distribution is multimodal, then the region will be disconnected. This is why don't call these intervals. Naturally, if we have unimodality then we will obtain intervals.

We extend our study of MLEs to the multiparameter case.

Definition 5.19. Suppose $X_1, \ldots, X_n$ are i.i.d. with distribution $f(x; \theta)$ where $\theta = (\theta_! \ \ldots \ \theta_k)^T$. The **score vector** is defined as

$$S(\theta) = \left[\frac{\partial l}{\partial \theta_1}, \ \ldots \ , \frac{\partial l}{\partial \theta_k} \right]^T \qquad \theta \in \Omega$$

The maximum likelihood estimate of θ is

$$\hat{\theta} = (\hat{\theta}_1 \ \ldots \ \hat{\theta}_l) = \text{argmax}_{\theta \in \Omega} l(\theta_1 \ \ldots \ \theta_k)$$

which is calculated by simultaneously setting the k estimating equations in $S(\theta)$ to zero.

Definition 5.20. Suppose $X_1, \ldots, X_n$ are i.i.d. with distribution $f(x; \theta)$ where $\theta = (\theta_1 \ \ldots \ \theta_k)^T$. We define:

1. The **information matrix** $I(\theta)$ is a $k \times k$ symmetric matrix whose (i, j)th entry is given by

$$-\frac{\partial^2}{\partial \theta_i \partial \theta_j} l(\theta)$$

   and $I(\hat{\theta})$ is the **observed information matrix**.

2. The **expected** or **Fisher information matrix** $J(\theta)$ is a $k \times k$ symmetric matrix whose (i, j) entry is given by

$$\text{E}\left[-\frac{\partial^2}{\partial \theta_i \partial \theta_j} l(\theta; X) \right]$$

3. The set of θ values for which $R(\theta) \geq p$ is called a $100p\%$ **likelihood region** for θ which is a region in $\mathbb{R}^k$.

Remark. Once we obtain candidates for $\hat{\theta}$, then we proceed using our usual multivariable second derivative test.

Theorem 5.21. *Suppose $X = (X_1, \ldots, X_n)$ be a random sample from $f(x; \theta)$. Let $\tilde{\theta}_n = \tilde{\theta}_n(Z_1, \ldots, X_n)$ be the ML estimator of θ based on X. Then under certain (regularity) conditions, we have:*

1. ***Consistency:***

$$\tilde{\theta}_n \xrightarrow{P} \theta$$

2. ***Asymptotic Normality:***

$$\sqrt{J(\theta)}(\tilde{\theta}_n - \theta) \xrightarrow{D} Z \sim N(0,1)$$

3. ***Asymptotic Distribution of Relative Likelihood:***

$$-2\log R(\theta; X) = 2[l(\tilde{\theta}_n; X) - l(\theta, X)] \xrightarrow{D} W \sim \chi^2(1)$$

where θ is the true but unknown parameter of the distribution.

Proof. I might post later. ■

Definition 5.22. Suppose X is a random variable whose distribution depends on θ. Suppose that $A(x)$ and $B(x)$ are functions with $A(x) \leq B(x)$ for all x in the support of X and $\theta \in \Omega$. Let x be the observed data. Then $(A(x), B(x))$ is an **interval estimate** for θ and $(A(X), B(X))$ is an **interval estimator** for θ. The two most common interval estimators are confidence and likelihood interval.

Definition 5.23. The random variable $Q(X; \theta)$, which is a function of the data X and the unknown parameter θ, is a **pivotal quantity** if the distribution of Q does not depend on θ.

The random variable Q is called an **asymptotic pivotal quantity** if the limiting distribution of Q does not depend on θ.

Chapter 6

Hypothesis Tests

In this chapter we formalise the idea of "finding an effect".

Definition 6.1. We introduce a bit of jargon first:

1. A **statistical hypothesis** is a statement about the population parameters.
2. A **test of hypothesis** is the procedure to check the validity of a statistical hypothesis based on evidence found in collected data.

As a general procedure, we must formulate a statistical hypothesis and then test the relevant parameters using the observed data. Suppose we have a model $f(x;\theta)$ where $\theta = (\theta_1, \theta_2, \ldots, \theta_n)^T \in \Omega$, where $\Omega \subseteq R^k$ is the space of all possible parameters. A hypothesis is usually formulated as

$$H_0 \,:\, \theta \in \Omega_0$$

where $\Omega_0 \subseteq \Omega$.

Definition 6.2. A **null hypothesis** is the default position under the model assumptions and is usually denoted as H_0. Opposite to it, we have the **alternative hypothesis**.

Definition 6.3. A **discrepancy measure** or a **test statistic** is a measure that evaluates if the data supports H_0. We want to test if the discrepancy measure is extreme enough to reject the null hypothesis.

Definition 6.4. A **Type I Error** is the case where we reject the null hypothesis given that it is true. A **Type II Error** is the case where we do not reject the null hypothesis given that it is false.

We seek to develop a general algorithm to test a hypothesis. One such procedure is the **likelihood ratio test** for simple hypotheses.

Theorem 6.5. *Under regularity conditions, one of which is that the support of X distribution does not depend on θ, we have*

$$\Lambda(\theta_0) = -2\log[R(\theta_0; X)] \xrightarrow{D} W \sim \chi^2(k)$$

where k is the number of parameters under the general hypothesis mins the number of parameters under H_0:

$$k = \dim\Omega - \dim\Omega_0$$

Proof. Perhaps for a different course? ■

Definition 6.6. For the likelihood ratio test, the **asymptotic p-value** is defined as

$$p = P\left(W \geq \lambda(\theta_0)\right)$$

Example 6.7. I refer the reader to the course notes for examples. None of them are different from the ones seen in STAT 231.

That's all folks!