1 Properties of a Random Sample

Definition 1 (Random Sample)**.** A collection of random variables $X_1, \ldots, X_n$ is a *random sample of size n from population* $f(x)$ if they are iid with pdf/pmf f(x).

Definition 2 (Statistic)**.** $Y = T(X_1, \ldots, X_n)$ is a *statistic* if the domain of T contains the sample space of $(X_1, \ldots, X_n)$. The distribution of Y is the *sampling distribution of* Y.

Remark 1. A statistic is any function of the data. The only restriction is that the statistic is not also a function of some other parameters.

Definition 3 (Sample Mean and Sample Variance)**.** The *sample mean* $\bar{X}$ and *sample variance* S^2 of a random sample $X_1, \ldots, X_n$ are, respectively,

- $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$
- $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$.

The *sample standard deviation* is $S = \sqrt{S^2}$.

Theorem 1 (Some facts)**.** Let $X_1, \ldots, X_n$ be a random sample from a population with mean μ and finite variance σ^2, then

a. $\mathbb{E}[\bar{X}] = \mu$

b. $\mathrm{Var}[\bar{X}] = \frac{\sigma^2}{n}$

c. $\mathbb{E}[S^2] = \sigma^2$.

Theorem 2. Take $x_1, \ldots, x_n \in \mathbb{R}$ and let $\bar{x}$ be their mean. Then,

a. $\min_a \sum_{i=1}^n (x_i - a)^2 = \sum_{i=1}^n (x_i - \bar{x})^2$

b. $(n-1)s^2 = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2$.

Theorem 3. Let $X_1, \ldots, X_n$ be a random sample from population $f(x|\boldsymbol{\theta})$ belonging to an exponential family

$$f(x|\boldsymbol{\theta}) = h(x)c(\boldsymbol{\theta}) \exp\left(\sum_{i=1}^k w_i(\boldsymbol{\theta}) t_i(x)\right).$$

Define the statistics

$$T_i(X_1, \ldots, X_n) = \sum_{j=1}^n t_i(X_j), \quad i = 1, \ldots, k.$$

Then if the set $\{(w_1(\boldsymbol{\theta}), \ldots, w_n(\boldsymbol{\theta}), \theta \in \Theta\}$ contains and open subset of $\mathbb{R}^k$ then the distribution of $(X_1, \ldots, X_n)$ is an exponential family of the form

$$f(u_1, \ldots, u_n|\boldsymbol{\theta}) = H(u_1, \ldots, u_n)c(\boldsymbol{\theta})^n \exp\left(\sum_{i=1}^k w_i(\boldsymbol{\theta}) u_i\right).$$

Remark 2. The open condition eliminates curved exponential families from this result.

1.1 Sampling from the Normal Distribution

Theorem 4. Let $X_1, \ldots, X_n$ be a random sample from a $\mathcal{N}(\mu, \sigma)$ distribution. Then,

- $\bar{X}$ and S^2 are independent
- $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$
- $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$.

Lemma 1 (Covariance and independence)**.** In the case of samples from a multivariate normal

- Independence $\iff$ vanishing covariance
- Pairwise independence $\iff$ independence

1.2 Convergence Concepts

1.2.1 Convergence in Probability

Definition 4. A sequence of random variables $\{X_i : i \in \mathbb{N}\}$ *converges in probability* to a random variable X if $\forall \epsilon > 0$

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

(or equivalently if $\lim_{n\to\infty} \mathbb{P}(|X_n - X| < \epsilon) = 1$).

Theorem 5 (Weak Law of Large Numbers)**.** Let $X_1, X_2, \ldots$ be iid RVs with mean μ and $\mathbb{E}[|X_i|] < \infty$. Define $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n$. Then the sequence $\bar{X}_n$ converges in probability to μ. That is, $\forall \epsilon > 0$

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

Proof is using Chebychev's inequality.

Theorem 6. If $X_1, X_2, \ldots$ converges in probability to X and h is a continuous function, then $h(X_1), h(X_2), \ldots$ converges in probability to $h(X)$.

1.2.2 Almost Sure Convergence

Definition 5. A sequence of random variables $X_1, X_2, \ldots$ *converges almost surely* to a random variable X if $\forall \epsilon > 0$

$$\mathbb{P}\left(\lim_{n\to\infty} \left|\bar{X}_n - X\right| < \epsilon\right) = 1.$$

Comments.

+ Almost sure convergence is much stronger than convergence in probability. Convergence in probability states that the sequence of measures of the sets on which the sequence has finite difference from its the limit converges to 0. Almost sure convergence states that any place where the sequence has finite difference from its limit must have measure 0. It's like a sequence of integrals converging vs. whether the integrands converge.

+ Almost sure convergence implies convergence in probability but not the other way around.

Theorem 7. If a sequence converges in probability then it is possible to find a subsequence that converges almost surely.

Theorem 8 (Strong Law of Large Numbers)**.** Let $X_1, X_2, \ldots$ be iid RVs with mean μ and $\mathbb{E}[|X_i|] < \infty$. Define $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n$. Then the sequence $\bar{X}_n$ converges almost surely to μ. That is, $\forall \epsilon > 0$

$$\mathbb{P}\left(\lim_{n\to\infty} \left|\bar{X}_n - \mu\right| < \epsilon\right) = 1.$$

1.2.3 Convergence in Distribution

Definition 6. A sequence of random variables $X_1, X_2, \ldots$ *converges in distribution* to a random variable X if

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

at all points where F_X is continuous.

Remark 3. Here it is really the cdfs that converge, rather than the random variables. In this way convergence in distribution differs from the previous two concepts.

Theorem 9. Convergence in probability implies convergence in distribution

Theorem 10 (Central Limit Theorem)**.** Let $X_1, X_2, \ldots$ be a sequence of iid random variables with $\mathbb{E}[X_i] = \mu$ and finite variance $\mathrm{Var}[X_i] = \sigma^2 < \infty$. Define $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Let $G_n(x)$ denote the cdf of $\sqrt{n}(\bar{X}_n - \mu)/\sigma$. Then $\forall x \in \mathbb{R}$,

$$\lim_{n\to\infty} G_n(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2} \mathrm{d}y.$$

That is, $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ converges in distribution to the standard normal.

Theorem 11 (Slutsky's Theorem)**.** If $X_n \to X$ in distribution and $Y_n \to a$ in probability with a constant, then

a. $Y_n X_n \to aX$ in distribution

b. $X_n + Y_n \to X + a$ in distribution

Remark 4. This tells us, for instance, that

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} \to \mathcal{N}(0,1)$$

in distribution, since we know that $S_n \to \sigma$ in probability.

1.2.4 The Delta Method

If we are interested in the convergence of some function of a sequence of RVs, rather than the RVs themselves, then we can use the Delta Method (follows from an application of Taylor's theorem and Slutsky's theorem).

Theorem 12 (Delta Method)**.** Let Y_n be a sequence of random variables that satisfies $\sqrt{n}(Y_n - \theta) \to \mathcal{N}(0, \sigma^2)$ in distribution. For a given function g and a specific value of θ, suppose that $g'(\theta)$ exists and is non-zero. Then

$$\sqrt{n}[g(Y_n) - g(\theta)] \to \mathcal{N}(0, \sigma^2 g'(\theta)^2)$$

in distribution.

Remark 5. There exists a corresponding multivariate result.

If $g'(\theta) = 0$ then we take the next term in the Taylor series.

Theorem 13 (Second Order Delta Method)**.** Let Y_n be a sequence of random variables that satisfies $\sqrt{n}(Y_n - \theta) \to \mathcal{N}(0, \sigma^2)$ in distribution. For a given function g and a specific value of θ, suppose that $g'(\theta) = 0$ and $g''(\theta)$ exists and is non-zero. Then

$$\sqrt{n}[g(Y_n) - g(\theta)] \to \mathcal{N}(0, \sigma^2 g'(\theta)^2)$$

in distribution.

1.3 Generating A Random Sample

Definition 7 (Direct Method)**.** A *Direct Method* of generating a random sample uses the probability integral transform to map draws from a uniform$(0, 1)$ random variable to draws from the distribution of interest.

The Probability Integral Transform states that if X has continuous cdf $F_X(x)$ then

$$F_X(X) \sim \text{uniform}(0, 1).$$

Definition 8 (Accept-Reject Algorithm)**.** Let $Y \sim f_Y(y)$ and $V \sim f_V(v)$ where f_Y and f_V have common support with

$$M = \sup_y f_Y(y)/f_V(y) < \infty.$$

To generate a random variable $Y \sim f_Y$:

a. Generate $U \sim \text{uniform}(0, 1)$, $V \sim f_V$ independent.

b. If $U < \frac{1}{M} f_Y(V)/f_V(V)$, return V as a sample of Y; otherwise go back to (a.).

Remark 6.

- It is typical to call V the *candidate density* and Y the *target density.*
- One would normally try to choose a candidate density with heavier tails than the target density (e.g. Cauchy and normal) to ensure that the tails of the target are well represented. If the target has heavy tails, however, it can be hard to find a candidate that results in finite M. In this case people turn to MCMC methods.
- Note that $\mathbb{P}(\texttt{terminate}) = 1/M$. The number of trials to generate one sample of Y is therefore geometric$(1/M)$, with M the expected number of trials.
- The intuition behind this algorithm is that if we consider placing the density of a random variable Y in a box (2d for simplicity) with coordinates (v, u), we express the cdf of Y using $V, U \sim \text{uniform}(0, 1)$

  $$\mathbb{P}(Y \leq y) = \mathbb{P}(V \leq y | U \leq \frac{1}{c} f_Y(V))$$

  where $c = \sup_y f_Y(y)$. In the actual algorithm we take $U \sim \text{uniform}(0, 1)$ and V to be an RV that has common support with Y.