Theory of Statistical Practice

Derek Li

Contents

1 Probability Review

1.1 Basic Definition

Definition 1.1. ***Random experiment*** is a mechanism producing an outcome (result) perceived as random or uncertain.

Definition 1.2. ***Sample space*** is a set of all possible outcomes of the experiment:

$$\mathcal{S} = \{\omega_1, \omega_2, \cdots\}.$$

Example 1.1. Waiting time until the next bus arrives: $\mathcal{S} = \{t : t \geqslant 0\}$.

1.2 Probability Function/Measure

Definition 1.3. Given a sample space $\mathcal{S}$, define $\mathcal{A}$ to be a collection of subsets (events) of $\mathcal{S}$ satisfying the following conditions:

1. $\mathcal{S} \in \mathcal{A}$;
2. $A \in \mathcal{A} \Rightarrow A^C \in \mathcal{A}$;
3. $A_1, A_2, \cdots \in \mathcal{A} \Rightarrow A_1 \cup A_2 \cup \cdots \in \mathcal{A}$.

If $\mathcal{S}$ is finite or countably infinite, then $\mathcal{A}$ could consist of all subsets of $\mathcal{S}$ including $\varnothing$.

Definition 1.4. The ***probability function*** (***measure***) P on $\mathcal{A}$ satisfies the following conditions:

1. $P(A) \geqslant 0, \forall A \in \mathcal{A}$;
2. $P(\varnothing) = 0$ and $P(\mathcal{S}) = 1$;
3. If $A_1, A_2, \cdots$ are disjoint (mutually exclusive) events, i.e., $A_i \cap A_j = \varnothing$ for $i \neq j$, then

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

Property 1.1. $P(A^C) = 1 - P(A)$.

Proof. $1 = P(\mathcal{S}) = P(A \cup A^C) = P(A) + P(A^C)$. □

Property 1.2. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Proof. $P(A) = P(A \cap B) + P(A \cap B^C)$ and $P(A \cup B) = P(B) + P(A \cap B^C)$. □

Property 1.3. $P(A \cup B) \leqslant P(A) + P(B)$.

Property 1.4. In general,

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) - \cdots - (-1)^n P(A_1 \cap \cdots \cap A_n).$$

Property 1.5 (Bonferroni's Inequality)**.** In general,

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

1.3 Conditional Probability

Definition 1.5. The probability of A ***conditional*** on B is defined by

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

if $P(B) > 0$. Note that if $P(B) = 0$, we can still define $P(A|B)$ but we need to be more careful mathematically.

Theorem 1.1 (Bayes Theorem)**.** If $B_1, \cdots, B_k$ are disjoint events with $B_1 \cup \cdots \cup B_k = \mathcal{S}$, then

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{\sum_{i=1}^{k} P(A|B_i)P(B_i)}.$$

1.4 Independence

Definition 1.6. Two events A and B are ***independent*** if

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

When $P(A), P(B) > 0$, we can also say

$$P(A|B) = P(A) \text{ and } P(B|A) = P(B).$$

Events $A_1, \cdots, A_k$ are independent if

$$P\left(\bigcap_{i=1}^{k} A_i\right) = \prod_{i=1}^{k} P(A_i).$$

1.5 Interpretation of Probability

- Long-Run frequencies: If we repeat the experiment many times, then $P(A)$ is the proportion of times the event A occurs.
- Degrees of belief (subjective probability): If $P(A) > P(B)$, then we believe that A is more likely to occur than B.
- Frequentist versus Bayesian statistical methods:
  - Frequentists: Pretend that an experiment is at least conceptually repeatable.
  - Bayesians: Use subjective probability to describe uncertainty in parameters and data.

1.6 Random Variable

Definition 1.7. ***Random variable*** is a real-valued function defined on a sample space $\mathcal{S}, X : \mathcal{S} \to \mathbb{R}$. In other words, for each outcome $\omega \in \mathcal{S}, X(\omega)$ is a real number.

Definition 1.8. The ***probability distribution*** of X depends on the probabilities assigned to the outcomes in $\mathcal{S}$.

Definition 1.9. The ***cumulative distribution function*** (CDF) of X is

$$F(x) = P(X \leqslant x) = P(\omega \in \mathcal{S} : X(\omega) \leqslant x).$$

We denote it $X \sim F$.

Property 1.6. CDF satisfies:

1. If $x_1 \leqslant x_2$, then $F(x_1) \leqslant F(x_2)$;
2. $F(x) \to 0$ as $x \to -\infty$ and $F(x) \to 1$ as $x \to \infty$;
3. F is right-continuous with left-hand limits:

$$\lim_{y \to x^+} F(y) = F(x), \lim_{y \to x^-} F(y) = F(x-) = P(X < x);$$

4. $P(X = x) = F(x) - F(x-)$.

Definition 1.10. If $X \sim F$ where F is a continuous function, then X is a ***continuous r.v.***, and we can typically find a non-negative ***probability density function*** (PDF) f s.t.

$$F(x) = \int_{-\infty}^{x} f(t)\mathrm{d}t.$$

Definition 1.11. If X takes only a finite or countably infinite number of possible values, then X is a ***discrete r.v.***, and F is a step function. We can define its ***probability mass function*** (PMF) by

$$f(x) = F(x) - F(x-) = P(X = x).$$

1.7 Expected Value

Definition 1.12. Suppose X with PDF $f(x)$ and Y with PMF $f(y)$. We can define the ***expected value*** of $h(X)$ and $h(Y)$ by

$$\mathbb{E}[h(X)] = \int_{-\infty}^{\infty} h(x)f(x)\mathrm{d}x \text{ and } \mathbb{E}[h(Y)] = \sum_{y} h(y)f(y).$$

We can also write $h(x) = h^+(x) - h^-(x)$ where $h^+(x) = \max\{h(x), 0\}$ and $h^-(x) = \max\{-h(x), 0\}$, then $\mathbb{E}[h(X)] = \mathbb{E}[h^+(X)] - \mathbb{E}[h^-(X)]$:

1. If $\mathbb{E}[h^+(X)]$ and $\mathbb{E}[h^-(X)]$ are finite, then $\mathbb{E}[h(X)]$ is well defined.
2. If $\mathbb{E}[h^+(X)] = \infty$ and $\mathbb{E}[h^-(X)]$ is finite, then $\mathbb{E}[h(X)] = \infty$.
3. If $\mathbb{E}[h^+(X)]$ is finite and $\mathbb{E}[h^-(X)] = \infty$, then $\mathbb{E}[h(X)] = -\infty$.
4. If $\mathbb{E}[h^+(X)]$ and $\mathbb{E}[h^-(X)]$ are infinite, then $\mathbb{E}[h(X)]$ does not exist.

Example 1.2 (Expected Values of Cauchy Distribution)**.** X is a continuous r.v. with

$$f(x) = \frac{1}{\pi(1 + x^2)}, -\infty < x < \infty.$$

We have

$$\mathbb{E}[X^+] = \mathbb{E}[X^-] = \int_0^{\infty} \frac{x}{\pi(1 + x^2)}\mathrm{d}x = \lim_{x \to \infty} \frac{1}{2\pi} \ln(1 + x^2) = +\infty.$$

Thus, $\mathbb{E}[X] = \mathbb{E}[X^+] - \mathbb{E}[X^-]$ does not exist.

1.8 Independent Random Variable

Definition 1.13. R.v.s. $X_1, X_2, \cdots$ are ***independent*** if the events $[X_1 \in A_1], [X_2 \in A_2], \cdots$ are independent events for any $A_1, A_2, \cdots$.

If $X_1, \cdots, X_n$ are independent r.v.s. with PDF or PMF $f_1, \cdots, f_n$, then the joint PDF or PMF of $(X_1, \cdots, X_n)$ is

$$f(x_1, \cdots, x_n) = \prod_{i=1}^{n} f_i(x_i).$$

Suppose $X_1, \cdots, X_n$ are independent r.v.s. with mean $\mu_1, \cdots, \mu_n$ and variance $\sigma_1^2, \cdots, \sigma_n^2$. Define $S = X_1 + \cdots + X_n$, then $\mathbb{E}[S] = \mu_1 + \cdots + \mu_n$ (which is true even if $X_1, \cdots, X_n$ are not independent) and $\mathrm{Var}[S] = \sigma_1^2 + \cdots + \sigma_n^2$.

1.9 Convergence of Random Variable

Theorem 1.2 (Markov's Inequality)**.** Suppose Y is a random variable with $\mathbb{E}[|Y|^r] < \infty$ for some $r > 0$, then

$$P(|Y| > \varepsilon) \leqslant \frac{\mathbb{E}[|Y|^r]}{\varepsilon^r}.$$

Proof. For any $\varepsilon > 0$,

$$\mathbb{E}[|Y|^r] = \mathbb{E}[|Y|^r I(|Y| \leqslant \varepsilon)] + \mathbb{E}[|Y|^r I(|Y| > \varepsilon)] \geqslant 0 + \varepsilon^r P(|Y| > \varepsilon),$$

then $P(|Y| > \varepsilon) \leqslant \frac{\mathbb{E}[|Y|^r]}{\varepsilon^r}$. □

Theorem 1.3 (Chebyshev's Inequality)**.**

$$P(|X - \mathbb{E}[X]| > \varepsilon) \leqslant \frac{\mathrm{Var}[X]}{\varepsilon^2}.$$

Proof. Take $r = 2, Y = X - \mathbb{E}[X]$ in Markov's Inequality. □

1.9.1 Convergence in Probability

Definition 1.14. A sequence of r.v.s. $\{Y_n\}$ ***converges in probability*** to a r.v. Y (denoted $Y_n \xrightarrow{p} Y$) if for each $\varepsilon > 0$,

$$\lim_{n \to \infty} P(|Y_n - Y| > \varepsilon) = 0.$$

Typically, the limiting r.v. Y is a constant.

Theorem 1.4 (Weak Law of Large Numbers)**.** If $X_1, X_2, \cdots$ are independent r.v.s. with finite mean μ, then

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

1.9.2 Convergence in Distribution

Definition 1.15. A sequence of r.v.s. $\{X_n\}$ ***converges in distribution*** to a r.v. X if

$$\lim_{n \to \infty} F_n(x) = F(x),$$

for every $x \in \mathbb{R}$ at which F is continuous. F_n and F are CDF of X_n and X, respectively.

Let $S_n = \sqrt{n}(\overline{X}_n - \mu)$, then we have $\mathbb{E}[S_n] = 0$ and $\mathrm{Var}[S_n] = \sigma^2$. $\{S_n\}$ is bounded in probability since

$$P(|S_n| > M) \leqslant \frac{\mathbb{E}[S_n^2]}{\varepsilon^2} = \frac{\sigma^2}{M^2} \to 0 \text{ as } M \to \infty.$$

Theorem 1.5 (Basic Central Limit Theorem)**.** If $X_1, X_2, \cdots$ are independent r.v.s. with common CDF F with finite mean and variance μ and σ^2, then

$$\lim_{n\to\infty} P(S_n \leqslant x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2\sigma^2}} \mathrm{d}t,$$

denoted as $S_n \xrightarrow{d} S \sim \mathcal{N}(0, \sigma^2)$.

As a consequence, the distribution of $\overline{X}_n$ is approximately $\mathcal{N}(\mu, \frac{\sigma^2}{n})$, when n is sufficiently large, denoted as $\overline{X}_n \dot{\sim} \mathcal{N}(\mu, \frac{\sigma^2}{n})$.

We can approximate $g(\overline{X}_n)$ by Taylor's Formula. We have

$$g(\overline{X}_n) = g(\mu) + g'(\mu)(\overline{X}_n - \mu) + o(\overline{X}_n - \mu)$$

and thus

$$\sqrt{n}(g(\overline{X}_n) - g(\mu)) = g'(\mu)\sqrt{n}(\overline{X}_n - \mu) + \sqrt{n}o(\overline{X}_n - \mu)$$

with $o(\overline{X}_n - \mu) \to 0$, suggesting that

$$\sqrt{n}(g(\overline{X}_n) - g(\mu)) \xrightarrow{d} \mathcal{N}(0, [g'(\mu)]^2\sigma^2).$$

Theorem 1.6 (General Central Limit Theorem)**.** Suppose $X_1, X_2, \cdots$ are independent with $\mathbb{E}[X_i] = \mu_i$ and $\mathrm{Var}[X_i] = \sigma_i^2$ and let

$$S_n = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n}\mu_i\right),$$

then

$$S_n \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{n}\sum_{i=1}^{n}\sigma_i^2\right)$$

provided that $\sum_{i=1}^{n}\sigma_i^2$ is not dominated by a small number of terms and the tails of the distributions of $\{X_i\}$ are not too dissimilar.

1.9.3 Quality of Normal Approximation

Definition 1.16. Define ***skewness*** of X_i as

$$\mathrm{Skew}(X_i) = \frac{\mathbb{E}[(X_i - \mu)^3]}{\sigma^3}.$$

Definition 1.17. Define ***kurtosis*** of X_i as

$$\mathrm{Kurt}(X_i) = \frac{\mathbb{E}[(X_i - \mu)^4]}{\sigma^4}.$$

Let $S_n = \sqrt{n}(\overline{X}_n - \mu)$, where $\overline{X}_n$ is the sample mean of independent $X_1, \cdots, X_n$ with CDF F with mean μ and variance σ^2. The normal approximation works better for fixed n if $\mathrm{Skew}(X_i)$ and $\mathrm{Kurt}(X_i)$ are close to the values for normal distribution (0 and 3, respectively).

1.9.4 Distribution Approximation

Theorem 1.7 (Slutsky's Theorem)**.** Suppose $X_n \xrightarrow{d} X \sim G$ and $Y_n \xrightarrow{p} \theta$, where θ is a constant, then as $n \to \infty, \psi(X_n, Y_n) \xrightarrow{d} \psi(X, \theta)$, where ψ is continuous.

Theorem 1.8 (Delta Method)**.** Suppose $a_n(X_n - \theta) \xrightarrow{d} Z$ where $a_n \uparrow \infty$. If $g(x)$ is differentiable at $x = \theta$, then

$$a_n(g(X_n) - g(\theta)) \xrightarrow{d} g'(\theta)Z.$$

Proof. By Taylor's Formula,

$$g(X_n) = g(\theta) + g'(\theta)(X_n - \theta) + \Delta(X_n)(X_n - \theta),$$

where $\Delta(X_n) \xrightarrow{p} 0$. Therefore,

$$a_n(g(X_n) - g(\theta)) = g'(\theta)a_n(X_n - \theta) + \Delta(X_n)a_n(X_n - \theta) \xrightarrow{d} g'(\theta)Z + 0 = g'(\theta)Z.$$

□

We can use the Delta Method to estimate standard errors of parameter estimators, and extend the Delta Method to functions of several sample means: $g(\overline{X}_n, \overline{Y}_n, \overline{Z}_n, \cdots)$.

Another application with the Delta Method is the variance stabilizing transformations for some distributions with $\text{Var}[X_i] = \phi(\mu)$.

Example 1.3. For the Poisson distribution, $\text{Var}[X_i] = \phi(\mu) = \mu$, then by CLT,

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \mu).$$

We find g s.t.

$$\sqrt{n}(g(\overline{X}_n) - g(\mu)) \xrightarrow{d} \mathcal{N}(0, 1),$$

i.e.,

$$[g'(\mu)]^2\mu = 1 \Rightarrow g'(\mu) = \frac{1}{\sqrt{\mu}}.$$

Thus, $g(x) = 2\sqrt{x} + C$ and

$$2\sqrt{\overline{X}_n} \dot{\sim} \mathcal{N}\left(2\sqrt{\mu}, \frac{1}{n}\right).$$

2 Statistical Models

2.1 Probability versus Statistics

Suppose $X_1, \cdots, X_n$ are independent r.v.s., with some CDF F. For probability, F is known and we can calculate probabilities involving the r.v.s. $X_1, \cdots X_n$. Knowledge of the population F gives information about the nature of samples from the population. For statistics, F is unknown and we observe outcomes of $X_1, \cdots X_n : x_1, \cdots, x_n$ (data).

Definition 2.1. ***Statistical inference*** uses the information in the data to estimate of infer properties of the unknown F.

Definition 2.2. Assume that the data $x_1, \cdots, x_n$ are outcomes of r.v.s. $X_1, \cdots, X_n$ whose joint distribution is F (which is assumed to be unknown to some degree). A ***statistical model*** is a family $\mathcal{F}$ of probability distributions of $(X_1, \cdots, X_n)$.

We assume that true distribution $F \in \mathcal{F}$ but in practice, $\mathcal{F}$ typically represents only an approximation to the truth, i.e., $F \notin \mathcal{F}$ but F is close to some $F_0 \in \mathcal{F}$.

Definition 2.3. $\mathcal{F}$ is called a ***parametric model*** if

$$\mathcal{F} = \{F_\theta : \theta \in \Theta\}, \Theta \subset \mathbb{R}^p,$$

where θ is the parameter and Θ is the parameter space. We can write $\theta = (\theta_1, \cdots, \theta_p)$.

Definition 2.4. The model is said to be ***non-parametric*** if the parameter space Θ is not finite dimensional.

In practice, we often approximate the infinite dimensional parameter by a finite dimensional parameter. For example, we assume $g(x) \approx \sum_{k=1}^{p} \beta_k \phi_k(x)$ for some functions ϕ_k's and unknown parameters β_k's.

Definition 2.5. The model is said to be ***semi-parametric*** if non-parametric model has a finite dimensional parametric component.

2.2 Bayesian Models

Assume we have a parametric model with parameter space $\Theta \subset \mathbb{R}^p$. For each $\theta \in \Theta$, the joint CDF F_θ is the conditional distribution of $(X_1, \cdots, X_n)$ given θ. ***Bayesian inference*** is the process that we put a probability distribution on Θ - ***prior distribution***, and then after observing $X_1 = x_1, \cdots, X_n = x_n$, we can use Bayes Theorem to obtain a ***posterior distribution*** of θ.

Note that we can take Bayesian inference for non-parametric models.

2.3 Empirical Distribution Function (EDF)

We are often interested in estimating characteristics of $F, \theta(F)$, called statistical functionals (mathematically, $\theta : \mathcal{F} \to \mathbb{R}$).

2.3.1 Substitution Principle

Given $X_1, \cdots, X_n$ from F, we first estimate F by $\widehat{F}$ and substitute $\widehat{F}$ for F into $\theta(F)$:

$$\widehat{\theta}(F) = \theta(\widehat{F}).$$

If $\theta(\cdot)$ is continuous and $\widehat{F} \approx F$, then $\theta(\widehat{F}) \approx \theta(F)$. A simple estimator of F is the ***empirical distribution function*** (EDF)

$$\widehat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leqslant x).$$

Note that $\widehat{F}$ is a discrete CDF putting mass $\frac{1}{n}$ at $X_1, \cdots, X_n$. In R, we can plot by `plot(ecdf(x))`.

2.3.2 Properties of EDF

Note that the EDF is a sample mean for each x and so the WLLN and CLT hold as $n \to \infty$, and $I(X_1 \leqslant x), \cdots, I(X_n \leqslant x)$ are independent Bernoulli random variables.

Property 2.1. EDF satisfies:

1. $\mathbb{E}[\widehat{F}(x)] = F(x), \mathrm{Var}[\widehat{F}(x)] = \frac{F(x)(1-F(x))}{n}$;
2. (WLLN)

$$\widehat{F}(x) = \widehat{F}_n(x) \xrightarrow{p} F(x) \text{ for each } x \Leftrightarrow \sup_{-\infty<x<\infty} |\widehat{F}(x) - F(x)| \xrightarrow{p} 0;$$

3. (CLT) $\sqrt{n}(\widehat{F}_n(x) - F(x)) \xrightarrow{d} \mathcal{N}(0, F(x)(1-F(x))$.

2.3.3 Example and Counterexample

If $\theta(F)$ is based on expected values, the substitution principle using the EDF works well.

Example 2.1. $\theta(F) = \mathbb{E}_F[X_i]$:

$$\widehat{\theta}(F) = \theta(\widehat{F}) = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}.$$

Example 2.2. $\theta(F) = \mathbb{E}_F[h(X_i)]$:

$$\widehat{\theta}(F) = \theta(\widehat{F}) = \frac{1}{n}\sum_{i=1}^{n} h(X_i).$$

Example 2.3 (Theil Index)**.** $\theta(F) = \mathbb{E}_F\left[\frac{X_i}{\mu(F)} \ln\left(\frac{X_i}{\mu(F)}\right)\right]$:

$$\widehat{\theta}(F) = \theta(\widehat{F}) = \frac{1}{n}\sum_{i=1}^{n} \frac{X_i}{\overline{X}} \ln\left(\frac{X_i}{\overline{X}}\right).$$

Example 2.4. Suppose F is continuous and $\theta(F) = f(x) = F'(x)$, and application of the substitution principle with EDF gives

$$\widehat{f}(x) = \widehat{F}'(x) = \begin{cases} 0, & x \neq X_i, \forall i \\ \text{undefined}, & x = X_i \end{cases}.$$

The substitution principle fails.

2.4 Order Statistics

Suppose $X_1, \cdots, X_n$ are independent with unknown CDF F, we order $X_1, \cdots, X_n$ from smallest to largest:

$$X_{(1)} \leqslant \cdots \leqslant X_{(n)}.$$

Due to the independence assumption, the order statistics carry the same information about F as the unordered data.

Order statistics can be used to estimate the quantiles $F^{-1}(\tau)$ of F.

Example 2.5 (Sample Median)**.**

$$M = \begin{cases} \frac{X_{(n/2)}+X_{(n/2+1)}}{2}, & n \text{ is even} \\ X_{((n+1)/2)}, & n \text{ is odd} \end{cases}.$$

M is an estimator of $F^{-1}(\frac{1}{2})$.

Likewise, we can estimate $F^{-1}(\tau)$ by $X_{(k)}$ where $k \approx \tau n$.

2.4.1 Distribution of $X_{(k)}$

Now suppose $X_1, \cdots, X_n$ are independent with continuous CDF F and PDF f.

We first find CDF of sample minimum and maximum. For the minimum,

$$P(X_{(1)} > x) = P(X_1 > x, \cdots, X_n > x) = [1 - F(x)]^n$$

and thus

$$P(X_{(1)} \leqslant x) = 1 - [1 - F(x)]^n.$$

For the maximum,

$$P(X_{(n)} \leqslant x) = P(X_1 \leqslant x, \cdots, X_n \leqslant x) = F(x)^n.$$

Therefore, the densities of $X_{(1)}$ and $X_{(n)}$ are

$$g_1(x) = n[1 - F(x)]^{n-1} f(x), g_n(x) = nF(x)^{n-1} f(x).$$

Now define

$$Z(x) = \sum_{i=1}^{n} I(X_i \leqslant x) \sim \text{Bin}(n, F(x)),$$

and note that $[X_{(k)} \leqslant x] = [Z(x) \geqslant k]$. Thus,

$$P(X_{(k)} \leqslant x) = P(Z(x) \geqslant k) = \sum_{i=k}^{n} \binom{n}{i} F(x)^i (1 - F(x))^{n-i}.$$

Wherefore, the PDF of $X_{(k)}$ is

$$g_k(x) = \frac{\mathrm{d}}{\mathrm{d}x} \sum_{i=k}^{n} \binom{n}{i} F(x)^i [1 - F(x)]^{n-i} = \frac{n!}{(k-1)!(n-k)!} F(x)^{k-1} [1 - F(x)]^{n-k} f(x).$$

2.4.2 Convergence in Distribution of Central Order Statistics

Definition 2.6. Suppose $k = k_n \approx \tau n$ for some $\tau \in (0,1)$ (but not too close to 0 or 1). $X_{(k)}$ is called a ***central order statistic***.

For example, if $k \approx \frac{1}{2}$, then $X_{(k)}$ is the sample median. **Intuitively, $X_{(k)}, k \approx \tau n$ is an estimator of the τ-quantile $F^{-1}(\tau)$.**

Theorem 2.1. If $\{k_n\}$ is a sequence of integers with $\sqrt{n}\left(\frac{k_n}{n} - \tau\right) \to 0$ for some $\tau \in (0,1)$ and $f(F^{-1}(\tau)) > 0$, then

$$\sqrt{n}(X_{(k_n)} - F^{-1}(\tau)) \xrightarrow{d} \mathcal{N}\left(0, \frac{\tau(1-\tau)}{[f(F^{-1}(\tau))]^2}\right).$$

Proof. Suppose $U_1, \cdots, U_n$ are independent $\text{Unif}(0,1)$ r.v.s. and $U_{(1)} \leqslant \cdots \leqslant U_{(n)}$ are the corresponding order statistics.

Take $E_1, \cdots, E_{n+1}$ to be independent exponential r.v.s. with mean 1, then

$$(U_{(1)}, \cdots, U_{(n)}) \overset{d}{=} \left(\frac{E_1}{S}, \cdots, \frac{E_1 + \cdots + E_n}{S}\right),$$

where $S = E_1 + \cdots + E_{n+1}$. Note that

$$\frac{S}{n} = \underbrace{\frac{E_1 + \cdots + E_{n+1}}{n+1}}_{\xrightarrow{p} 1 \text{ (WLLN)}} \cdot \frac{n+1}{n} \xrightarrow{p} 1.$$

Therefore, we can approximate the distribution of $U_{(k)}$ by a distribution of a sum of exponential r.v.s.:

$$U_{(k)} = \frac{(E_1 + \cdots + E_k)/n}{(E_1 + \cdots + E_{n+1})/n} \approx \frac{1}{n}(E_1 + \cdots + E_k).$$

Assume $\sqrt{n}\left(\frac{k_n}{n} - \tau\right) \to 0$, then

$$\sqrt{n}(U_{(k_n)} - \tau) \overset{d}{=} \sqrt{n}\left(\frac{E_1 + \cdots + E_{k_n} - \tau S}{S}\right).$$

Now we need to show:

$$\sqrt{n}\left(\frac{1}{n}(E_1 + \cdots + E_{k_n} - \tau S)\right) \xrightarrow{d} \mathcal{N}(0, \tau(1-\tau)).$$

Let

$$E_1 + \cdots + E_{k_n} - \tau S = \sum_{i=1}^{n+1} a_i E_i,$$

where $a_i = 1 - \tau$ for $i = 1, \cdots, k_n$ and $a_i = -\tau$ for $i = k_n + 1, \cdots, n+1$.

We have

$$\mathbb{E}\left[\frac{1}{\sqrt{n}} \sum_{i=1}^{n+1} a_i E_i\right] = \frac{1}{\sqrt{n}}(k_n(1-\tau) - (n - k_n + 1)\tau) \to 0$$

and

$$\text{Var}\left[\frac{1}{\sqrt{n}} \sum_{i=1}^{n+1} a_i E_i\right] = \frac{1}{n}(k_n(1-\tau)^2 + (n - k_n + 1)\tau^2) \to \tau(1-\tau).$$

Recall that if $U \sim \text{Unif}(0,1)$ and F is a continuous CDF, $0 < F(x) < 1$ with PDF f, $f(x) > 0$ for all x, then $X = F^{-1}(U) \sim F$.

Thus if $U_{(1)} \leqslant \cdots \leqslant U_{(n)}$, then $F^{-1}(U_{(1)}) \leqslant \cdots \leqslant F^{-1}(U_{(n)})$ are order statistics from F. In other words,

$$\sqrt{n}(X_{(k_n)} - F^{-1}(\tau)) \stackrel{d}{=} \sqrt{n}(F^{-1}(U_{(k_n)}) - F^{-1}(\tau))$$

and we can use the Delta method with $g(\tau) = F^{-1}(\tau)$.

Note that $F(F^{-1}(\tau)) = \tau$ and

$$\frac{\mathrm{d}}{\mathrm{d}\tau} F(F^{-1}(\tau)) = \frac{\mathrm{d}}{\mathrm{d}\tau}\tau \Rightarrow f(F^{-1}(\tau))\frac{\mathrm{d}}{\mathrm{d}\tau}F^{-1}(\tau) = 1,$$

and thus

$$g'(\tau) = \frac{\mathrm{d}}{\mathrm{d}\tau}F^{-1}(\tau) = \frac{1}{f(F^{-1}(\tau))}.$$

Applying Delta method, we have

$$\sqrt{n}(X_{(k_n)} - F^{-1}(\tau)) \xrightarrow{d} \mathcal{N}\left(0, \frac{\tau(1-\tau)}{[f(F^{-1}(\tau))]^2}\right)$$

□

Example 2.6 (Sample Median versus Sample Mean from Normal Distribution)**.** Suppose $X_i, \cdots, X_n$ are independent $\mathcal{N}(\mu, \sigma^2)$ r.v.s., and $\mathbb{E}[X_i] = F^{-1}(\frac{1}{2}) = \mu$. Both sample mean and sample median can be used to estimate μ. However, $\text{Var}(\overline{X}) = \frac{\sigma^2}{n}$, and

$$\text{Var}[X_{(k)}] \approx \frac{\tau(1-\tau)}{n[f(F^{-1}(\tau))]^2} = \frac{\pi\sigma^2}{2n}, k \approx \frac{n}{2}.$$

Thus,

$$\frac{\text{Var}[X_{(k)}]}{\text{Var}[\overline{X}]} = \frac{\pi}{2} > 1,$$

and $\overline{X}$ is a better estimator of μ.

Example 2.7 (Sample Median versus Sample Mean from Laplace Distribution)**.** Note that

$$f(x) = \frac{1}{2}e^{-|x-\mu|}.$$

We have

$$\text{Var}[\overline{X}] = \frac{2}{n} \text{ and } \text{Var}[X_{(k)}] \approx \frac{1}{(4n[f(\mu)]^2)} = \frac{1}{n}, k \approx \frac{n}{2}.$$

Therefore, $X_{(k)}$ is a better estimator of μ.

2.5 Plots

2.5.1 Boxplot

Boxplot is the simplified graphical representation of the data. R function is `boxplot`.

2.5.2 Quantile-Quantile Plot

Quantile-quantile plot is the graphical tool to check if data come from a particular location or scale family of distributions.

Suppose we check whether data $x_1, \cdots, x_n$ are well-modeled by some specified F_0, we plot ordered value $x_{(k)}$ against $F_0^{-1}(\tau_k)$ for $k = 1, \cdots, n$, where $\tau_k = \frac{k-1/2}{n}, \frac{k}{n+1}$ or $\frac{k-3/8}{n+1/4}$. If F_0 is a good model then the points should fall close to a straight line.

Suppose we check whether data $x_1, \cdots, x_n$ are well-modeled by $F(x) = F_0\left(\frac{x-\mu}{\sigma}\right)$, where F_0 is specified but μ and σ are unknown. We have $F^{-1}(\tau) = \mu + \sigma F_0^{-1}(\tau)$ and can plot $x_{(k)}$ against $F_0^{(-1)}(\tau_k)$ for $k = 1, \cdots, n$.

By theorem, if the data come from a distribution of this form, then

$$x_{(k)} = \mu + \sigma F_0^{-1}(\tau) + \varepsilon_k, k = 1, \cdots, n,$$

where

$$\varepsilon_k \dot{\sim} \mathcal{N}\left(0, \frac{\sigma^2 \tau_k(1-\tau_k)}{n[f_0(F_0^{-1}(\tau_k))]^2}\right).$$

For each fixed τ_k, $\mathrm{Var}[\varepsilon_k] \to 0$ as n increases. Behavior of $\mathrm{Var}[\varepsilon_k]$ as $\tau_k \to 0$ or 1 is less obvious.

Note that quantile-quantile plots are most useful for lighter tailed distributions (such as the normal distribution).

Example 2.8 (Normal Quantile-Quantile Plot)**.** $F_0 = \mathcal{N}(0, 1)$ and so

$$F_0(x) = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} \mathrm{d}t.$$

R function is `qqnorm` - uses $\tau_k = \frac{k-3/8}{n+1/4}$ for $n \leqslant 10$ and $\frac{k-1/2}{n}$ for $n > 10$.

Also, we can use Shapiro-Wilk test that is based on correlation between $\{x_{(k)}\}$ and normal scores $\{\Phi^{-1}(\tau_k)\}$: if the data are normal, then the correlation should be very close to 1. R function is `shapiro.test`. If p-value is close to 0, then a normal model is not good for the data.

2.5.3 Line-Up Plot

Line-up plot compares the Q-Q plot of $x_1, \cdots, x_n$ to other Q-Q plot constructed from normally distributed data.

2.5.4 Weibull Plot

Weibull distribution is

$$f(x; \alpha, \lambda) = \left(\frac{\alpha}{\lambda}\right)\left(\frac{x}{\lambda}\right)^{\alpha-1} e^{-\left(\frac{x}{\lambda}\right)^{\alpha}}, x \geqslant 0,$$

where $\lambda, \alpha > 0$ (α is the Weibull modulus). The quantiles of the Weibull distribution are

$$F^{-1}(\tau; \alpha, \lambda) = \lambda(-\ln(1-\tau))^{\frac{1}{\alpha}}$$

or

$$\ln(F^{-1}(\tau; \alpha, \lambda)) = \frac{1}{\alpha}\ln(-\ln(1-\tau)) + \ln(\lambda).$$

The Weibull plot is $\ln(x_{(k)})$ against $\ln(-\ln(1-\tau_k))$ for $k = 1, \cdots, n$.

2.5.5 Histogram

A histogram is a very simple density estimator. Given data $x_1, \cdots, x_n$ and bins $B_k = [u_{k-1}, u_k)$ for $k = 1, \cdots, m$, we define for $x \in B_k$,

$$\text{Hist}(x) = \frac{1}{n(u_k - u_{k-1})} \sum_{i=1}^{n} I(x_i \in B_k).$$

Note that $\text{Hist}(x)$ is constant for x in each B_k and Hist is a density function since $\text{Hist}(x) \geqslant 0, \forall x$ and

$$\int_{-\infty}^{\infty} \text{Hist}(x)\mathrm{d}x = \sum_{k=1}^{m} \int_{B_k} \text{Hist}(x)\mathrm{d}x = \frac{1}{n} \sum_{i=1}^{n} \underbrace{\sum_{k=1}^{m} I(x_i \in B_k)}_{=1} \underbrace{\int_{B_k} \frac{1}{u_k - u_{k-1}} \mathrm{d}x}_{=1} = 1.$$

The appearance of the histogram depends on two factors:

1. Number of bins (m);
2. Boundaries of the bins ($u_0, \cdots, u_m$).

To find optimal bin width (i.e., $u_k - u_{k-1}$), we need the knowledge of the true f. If we assume f is clos to normal then one proposal is

$$u_k - u_{k-1} = 3.49 \cdot \text{SD} \cdot n^{-\frac{1}{3}},$$

where SD is the sample standard deviation of the data.

2.6 Spacings

Definition 2.7. Given $X_{(1)} \leqslant \cdots \leqslant X_{(n)}$, we define $(n-1)$ ***spacings*** (or ***first-order spacings***) by

$$D_k = X_{(k+1)} - X_{(k)}, k = 1, \cdots, n-1.$$

Intuitively, the spacings should carry some information about the PDF f.

2.6.1 Exponential Spacings

Suppose $X_1, \cdots, X_n$ are independent exponential r.v.s. with

$$f(x; \lambda) = \lambda e^{-\lambda x}, x \geqslant 0.$$

Given the order statistics, define

$$\begin{aligned}
Y_1 &= nX_{(1)} \\
Y_2 &= (n-1)(X_{(2)} - X_{(1)}) = (n-1)D_1 \\
Y_3 &= (n-2)(X_{(3)} - X_{(2)}) = (n-2)D_2 \\
&\vdots \\
Y_n &= X_{(n)} - X_{(n-1)} = D_{n-1}
\end{aligned}$$

Theorem 2.2. $Y_1, \cdots, Y_n$ are independent exponential r.v.s. with $f(x; \lambda)$, i.e., spacings from an exponential sample are themselves exponential and independent.

Proof. The joint PDF of $(X_{(1)}, \cdots, X_{(n)})$ is

$$f(x_1, \cdots, x_n) = n!\lambda^n e^{-\lambda \sum_{i=1}^n x_i}$$

for $0 \leqslant x_1 < \cdots < x_n$.

By definition of Y_k, we have

$$X_{(1)} = \frac{Y_1}{n}, X_{(k)} = \frac{Y_1}{n} + \cdots + \frac{Y_k}{n-k+1}$$

for $k = 2, \cdots, n$. Thus,

$$g(y_1, \cdots, y_n) = f\left(\frac{y_1}{n}, \cdots, \frac{y_1}{n} + \cdots + y_n\right) |J(y_1, \cdots, y_n)|,$$

where

$$J(y_1, \cdots, y_n) = \begin{pmatrix} \frac{1}{n} & 0 & 0 & \cdots & 0 \\ \frac{1}{n} & \frac{1}{n-1} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \cdots & \vdots \\ \frac{1}{n} & \frac{1}{n-1} & \frac{1}{n-2} & \cdots & 1 \end{pmatrix},$$

which is a lower triangular and thus $|J(y_1, \cdots, y_n)| = \frac{1}{n!}$. Therefore,

$$g(y_1, \cdots, y_n) = \lambda^n e^{-\lambda \sum_{i=1}^n y_i}$$

for $y_1, \cdots, y_n \geqslant 0$. □

2.6.2 Spacings from Continuous F

Now we assume that $\tau \approx \frac{k}{n} \approx \frac{k+1}{n}$ and thus if τ is not too close to 0 or 1, then $X_{(k+1)} \approx X_{(k)} \approx F^{-1}(\tau)$.

Theorem 2.3. If $\frac{k_n}{n} \to \tau, \tau \in (0,1)$ and $f(F^{-1}(\tau)) > 0$, then

$$nD_{k_n} \xrightarrow{d} \exp(f(F^{-1}(\tau))),$$

i.e.,

$$P(D_{k_n} \leqslant x) \approx 1 - e^{-nf(F^{-1}(\tau))x}$$

for $x \geqslant 0$.

Proof. Recall that

$$X_{(k_n+1)} \overset{d}{=} F^{-1}(U_{(k_n+1)}) \overset{d}{=} F^{-1}\left(\frac{E_1 + \cdots + E_{k_n+1}}{E_1 + \cdots + E_{n+1}}\right)$$

and

$$X_{(k_n)} \overset{d}{=} F^{-1}\left(\frac{E_1 + \cdots + E_{k_n}}{E_1 + \cdots + E_{n+1}}\right),$$

where $E_1, \cdots, E_{n+1}$ are independent exponential r.v.s. with mean 1. Thus

$$\begin{aligned} nD_{k_n} &\overset{d}{=} n\left[F^{-1}\left(\frac{E_1 + \cdots + E_{k_n+1}}{E_1 + \cdots + E_n}\right) - F^{-1}\left(\frac{E_1 + \cdots + E_{k_n}}{E_1 + \cdots + E_{n+1}}\right)\right] \\ &\approx \frac{1}{f(F^{-1}(\tau))} \cdot \frac{nE_{k_n+1}}{E_1 + \cdots + E_{n+1}} \\ &= \frac{1}{f(F^{-1}(\tau))} \cdot \frac{E_{k_n+1}}{\frac{1}{n}(E_1 + \cdots + E_{n+1})}. \end{aligned}$$

By WLLN, we have

$$\frac{E_1 + \cdots + E_{n+1}}{n} = \frac{E_1 + \cdots + E_{n+1}}{n+1} \cdot \frac{n+1}{n} \xrightarrow{p} 1,$$

and thus

$$nD_{k_n} \approx \frac{1}{f(F^{-1}(\tau))} E_{k_n+1} \sim \exp(f(F^{-1}(\tau))).$$

□

Note that if $\frac{k_n}{n} = \tau$, then D_{k_n} is approximately exponential with mean $\frac{1}{nf(F^{-1}(\tau))}$ and variance $\frac{1}{[nf(F^{-1}(\tau))]^2}$. Spacings have a variety of applications in statistics, such as goodness-of-fit.

2.7 Density Estimation

2.7.1 Spacings Estimation

We can use the spacings to estimate the density f. Suppose $D_1, \cdots, D_{n-1}$ are independent exponential r.v.s. with

$$\mathbb{E}[nD_k] = e^{g(V_k)}, V_k = \frac{X_{(k+1)} + X_{(k)}}{2}.$$

Note that $V_k \approx F^{-1}(\tau)$ for $\tau \approx \frac{k}{n} \approx \frac{k+1}{n}$. Then the density function is

$$f(x) = e^{-g(x)}.$$

We estimate $g(x)$ by B-spline function

$$g(x) = \beta_0 + \sum_{j=1}^{p} \beta_j \psi_j(x),$$

where $\psi_j(x)$ is the B-spline, and $\beta_0, \cdots, \beta_p$ are unknown parameters. The implementation below uses R function `glm`:

```
den.splines <- function(x, p = 5){
            library(splines)
            n <- length(x)
            x <- sort(x)
            x1 <- c(NA, x)
            x2 <- c(x, NA)
            sp <- (x2 - x1)[2:n]
            mid <- 0.5 * (x1 + x2)[2:n]
            y <- n * sp
            xx <- bs(mid, df = p)
            r <- glm(y ~ xx, family = quasi(link = "log", variance = "mu^2"))
            density <- exp(-r$linear.predictors)
            r <- list(x = mid, density =density)
            r
            }
```

Note that there are several issues with this method: $\widehat{f}(x)$ does not necessarily integrate to 1 and it has problems with discretized observations where 2 or more observations may be equal (e.g., due to rounding). Better density estimation methods exist such as kernel density estimation. Also note that the spacings method is similar in spirit to nearest neighbor density estimation.

2.7.2 Example: Hazard Functions

Definition 2.8. If X is a positive continuous r.v. with CDF F and PDF f, its ***hazard function*** is

$$h(x) = \frac{f(x)}{1 - F(x)}.$$

If X is a survival time,

$$\begin{aligned}\frac{1}{\delta} P(x < X \leqslant x + \delta | X > x) &= \frac{1}{\delta} \cdot \frac{P(x < X \leqslant x + \delta)}{P(X > x)} \\ &= \frac{1}{\delta} \cdot \frac{F(x + \delta) - F(x)}{1 - F(x)} \\ &\to \frac{f(x)}{1 - F(x)} = h(x)\end{aligned}$$

as $\delta \to 0$, i.e., $h(x)$ is the instantaneous death rate given survival to time x.

Since

$$h(x) = \frac{f(x)}{1 - F(x)} = -\frac{\mathrm{d}}{\mathrm{d}x} \ln(1 - F(x)),$$

then given $h(x)$, we can define CDF and PDF by

$$F(x) = 1 - \exp\left(-\int_0^x h(t)\mathrm{d}t\right)$$

and

$$f(x) = h(x) \exp\left(-\int_0^x h(t)\mathrm{d}t\right).$$

We also assume

$$\int_0^\infty h(x)\mathrm{d}x = \infty.$$

The shape of $h(x)$ gives information not immediately apparent in f and F:

- $h(x)$ is increasing: new better than used.
- $h(x)$ is decreasing: used better than new.

Suppose $X_1, \cdots, X_n$ are independent continuous positive r.v.s. with $h(x)$. Define normalized spacings $nX_{(1)}, (n-1)(X_{(2)} - X_{(1)}), \cdots, X_{(n)} - X_{(n-1)}$. If $X_1, \cdots, X_n$ are $\exp(h(x) = C)$, where C is a constant, then these normalized spacings are exponential.

In general, if $\frac{k}{n} \approx \tau \in (0, 1)$, then

$$(n - k)(X_{(k+1)} - X_{(k)}) \approx (1 - \tau)n(X_{(k+1)} - X_{(k)}).$$

Since

$$h(F^{-1}(\tau)) = \frac{f(F^{-1}(\tau))}{1 - \tau},$$

then $(n - k)(X_{(k+1)} - X_{(k)})$ is approximately exponential with mean $\frac{1}{h(F^{-1}(\tau))}$. We can use it to estimate the hazard function using a similar approach to that used to estimate the density.

2.7.3 Kernel Density Estimation

Definition 2.9. Let a (usually symmetric) density function $w(x)$ be a ***kernel***. Given the kernel w and a ***bandwidth*** parameter h, we defined the ***kernel density estimator*** as

$$\widehat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} w\left(\frac{x - X_i}{h}\right).$$

In R, the function is `density` that scales its kernels to have variance 1 and mean 0. The bandwidth parameter h controls the amount of smoothing: as h increases, the estimator becomes smoother. The choice of kernel is much less important than the choice of bandwidth.

The choice of h depends on what we believe the underlying density looks like: if we believe f is smooth then we should take larger h; if we believe f has a number of modes then we should take smaller h. There are methods for choosing h but not always reliable. In R, the default is

$$h_0 = 0.9 \cdot \min\left\{\text{SD}, \frac{\text{IQR}}{1.34}\right\} \cdot n^{-\frac{1}{5}}.$$

There is a bias-variance trade-off:

$$\text{MSE}(\widehat{f}_h(x)) = \mathbb{E}[(\widehat{f}_h(x) - f(x))^2] = \text{Var}[\widehat{f}_h(x)] + (\mathbb{E}[\widehat{f}_h(x)] - f(x))^2 = \text{Var}[\widehat{f}_h(x)] + (\text{Bias}(\widehat{f}_h(x)))^2.$$

As h increases, bias increases and variance decreases; as h decreases, bias decreases and variance increases.

Example 2.9 (Gaussian Kernel)**.**

$$w(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}},$$

which is the default kernel in R.

Example 2.10 (Epanechnikov Kernel)**.**

$$w(x) = \frac{3}{4\sqrt{5}}\left(1 - \frac{x^2}{5}\right), |x| \leqslant \sqrt{5}.$$

Example 2.11 (Rectangular Kernel)**.**

$$w(x) = \frac{1}{2\sqrt{3}}, |x| \leqslant \sqrt{3}.$$

Example 2.12 (Triangular Kernel)**.**

$$w(x) = \frac{1}{\sqrt{6}}\left(1 - \frac{|x|}{\sqrt{6}}\right), |x| \leqslant \sqrt{6}.$$

Example 2.13 (Property of Rectangular Kernel)**.** Suppose n is large and h is small. Take $w(x) = \frac{1}{2}$ for $|x| \leqslant 1$, then

$$\widehat{f}_h(x) = \frac{1}{2nh}\sum_{i=1}^{n} I(x - h \leqslant X_i \leqslant x + h).$$

The mean of $\widehat{f}_h(x)$ is

$$\mathbb{E}[\widehat{f}_h(x)] = \frac{F(x+h) - F(x-h)}{2h} \approx f(x) + \frac{h^2}{6} f''(x)$$

and so the squared bias is

$$(\mathbb{E}[\widehat{f}_h(X)] - f(x))^2 \approx \frac{h^4}{36}[f''(x)]^2.$$

The variance is

$$\mathrm{Var}[\widehat{f}_h(x)] = \frac{1}{4h^2n}\mathrm{Var}[I(x-h \leqslant X_i \leqslant x+h)] \approx \frac{1}{4h^2n} \cdot 2hf(x) = \frac{f(x)}{2hn}$$

and thus the mean square error is

$$\mathrm{MSE}(\widehat{f}_h(x)) \approx \frac{f(x)}{2hn} + \frac{h^4}{36}[f''(x)]^2$$

and the latter term is minimized at $h^* = \gamma(x)n^{-\frac{1}{5}}$.

Here are two motivations: redistribution and convolution.

- **Redistribution**: The empirical distribution function $\widehat{F}$ puts probability mass $\frac{1}{n}$ at each of the points $X_1, \cdots, X_n$. We use the kernel with bandwidth h to redistribute the mass around each X_i :

  $$\frac{1}{nh}w\left(\frac{x-X_i}{h}\right),$$

  where

  $$\int_{-\infty}^{\infty} \frac{1}{nh}w\left(\frac{x-X_i}{h}\right)\mathrm{d}x = \frac{1}{n}\int_{-\infty}^{\infty} w(t)\mathrm{d}t = \frac{1}{n}.$$

  The density estimate is now simply the sum of these densities over all observations.

- **Convolution**: The distribution $Y_h = U + hV$ where $U \sim \widehat{F}$ and V has density w with $V \perp U$. Y_h is a continuous r.v. for each $h > 0$:

  $$P(Y_h \leqslant x) = \sum_{i=1}^{n} P(U + hV \leqslant x | U = X_i)P(U = X_i) = \frac{1}{n}\sum_{i=1}^{n} P\left(V \leqslant \frac{x-X_i}{h}\right).$$

  Differentiating we get the density estimate.