Statistical Risk Analysis
Chapter 3 and 4: Classical Inference

Shaobo Jin

Department of Mathematics

Empirical Distribution

Definition (Definition 4.1 Empirical Probability)

Let x_1, ..., x_n be a sequence of measurements, then the fraction

$$F_n(x) = \frac{\text{number of } x_i \leq x,\ i = 1, 2, ..., n}{n} = \frac{1}{n} \sum_{i=1}^{n} 1(x_i \leq x)$$

is called the **empirical cumulative distribution function** (**ecdf**).

By definition,

1. $F_n(x)$ is a step function.
2. $F_n(x)$ is a non-decreasing function with $F_n(-\infty) = 0$ and $F_n(\infty) = 1$.

Law of Large Numbers

Theorem (Theorem 3.2, Law of Large Numbers (LLN))

Let X_1, ..., X_k be a sequence of independent and identically distributed (iid) random variables, all having the distribution $F_X(x)$. Denote the average of X_i by $\bar{X} = n^{-1}\sum_{i=1}^{k} x_k$. If $E[X]$ exists and is finite, then as k increases towards infinity, $\bar{X}$ is arbitrarily close to $E[X]$ with large probability.

By LLN, for any fixed x, $F_n(x)$ is arbitrary close to $F_X(x)$ with large probability.

Glivenko-Cantelli Theorem

Theorem (Glivenko-Cantelli Theorem)

Let X_1, ..., X_n be iid real valued random variables with distribution function $F_X(x) = P(X \leq x)$. Denote the empirical distribution function by

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1(x_i \leq x).$$

Then, $\sup_{x \in \mathbb{R}} |F_n(x) - F(x)|$ is arbitrary close to 0 with large probability.

Problem with $F_n(x)$

1. Despite the Glivenko-Cantelli Theorem, our sample size n can be small such that $F_n(x)$ is not good enough for rare events.
    - For example, earthquakes are rare events.
2. The empirical distribution contains no information about possible extreme values that have not been observed in a finite sequence
    - In earthquake example, the empirical probability of time interval being larger than 1901 is 0, even though those extreme events may occur with nonzero probability.

One way of solving this issue is the parametric approach: assume that $F_X(x)$ belongs to a family of distribution, indexed by some parameters.

Parametric Approach

To estimate a distribution, we often need three main steps: choice of a model, finding the parameters, and analysis of error (checking our assumptions)

1. Choose a model from one of the standard class of distribution $F_X(x)$, indexed by some parameter(s) θ.
2. Select a value of the parameter θ as $\theta^*(\boldsymbol{x})$ on the basis of our data. The book uses θ^*, it is more common in statistics to use $\hat{\theta}$.
3. Error analysis: the estimator $\Theta^*(\boldsymbol{X})$ is a random variable modelling the uncertainty of the value of an estimate due to the variability of data.

Choose F_X

The choice of a family of distributions $F(x;\theta)$ to model F_X often depends on experience from studies of similar experiments or by analysis of data.

We need to check whether $F(x;\theta)$ contradicts our data. One approach is the probability paper. Let $F(x;\theta^*)$ be the estimated cdf.

1. x-axis is the quantiles from the data.
2. y-axis is the quantiles obtained from the (estimated) distribution.
3. If they are close, the plot should be close to a straight line.

Example: Quantile of Exponential Distribution

Suppose that the estimated exponential distribution function is

$$F(x;\theta^*) = 1 - \exp\left(-\frac{x}{\theta^*}\right).$$

The $F_n(x)$ quantile is

$$x = -\theta^* \ln\left[1 - F_n(x)\right].$$

If the exponential distribution is reasonable, then we expect $(x, -\ln\left[1 - F_n(x)\right])$ is close to a straight line.

χ^2-Method for Discrete Distribution

The second method is the χ^2-method for goodness-of-fit tests.

Roll a Die

We roll a die $n = 20000$ times

Number of eyes i	1	2	3	4	5	6
Frequency n_i	3407	3631	3176	2916	3448	3422

If the die were fair, then

$$p_i \quad = \quad P\left(\text{the die shows number } i\right) = \frac{1}{6}.$$

We calculate

$$Q \quad = \quad \sum_{i=1}^{r} \frac{\left(n_i - np_i\right)^2}{np_i},$$

where p_i's are the probabilities under our hypothesis. We reject H_0 if $Q > \chi^2_{\alpha}\left(r-1\right)$, where α be the significance level.

χ^2-Method For Continuous Distribution

If the hypothesized distribution is a continuous distribution, we often need to introduce a partition of r groups such as

$$-\infty = c_0 < c_1 < c_2 < \cdots < c_{r-1} < c_{r=\infty}.$$

The observations $\boldsymbol{x}$ are classified into r groups based on the thresholds as $c_{i-1} < x \leq c_j$. Then the χ^2-method for discrete distribution is used.

Test Exponential Distribution

Suppose that our thresholds are $c_1 = 100$, $c_2 = 200$, $c_3 = 400$, $c_4 = 700$, $c_5 = 1000$, and $c_6 = \infty$. Test whether our earthquake data follows an exponential distribution with mean 400.

Properties of MLE

Suppose that we estimate the parameter θ in $F\left(x;\theta\right)$ by ML. We need to quantify the uncertainty of the MLE Θ^*. Let $\ell\left(\theta\right)$ be the log-likelihood function.

1. $V\left[\Theta^*\right]$ can be approximated by $-\left(\ddot{\ell}\left(\theta^*\right)\right)^{-1}$.
2. The distribution of $\left[\ddot{\ell}\left(\theta^*\right)\right]^{-1/2}\left(\Theta^*-\theta\right)$ can be approximated by $N\left(0,1\right)$, that is, $\left[\ddot{\ell}\left(\theta^*\right)\right]^{-1/2}\left(\Theta^*-\theta\right)\in AsN\left(\theta,1\right)$.

Exponential distribution

Suppose that we have a sample of size n from an exponential distribution with mean θ. Its MLE is $\Theta^*=\bar{X}$.

1. Approximate the variance of Θ^*.
2. Approximate $\mathrm{P}\left(\left[\ddot{\ell}\left(\theta^*\right)\right]^{-1/2}\left(\Theta^*-\theta\right)\leq 1\right)$.

Asymptotic Intervals

Denote $\sigma_\varepsilon^* = \sqrt{V\left[\Theta^*\right]}$. Since MLE is approximately normally distributed,

$$\mathrm{P}\left(\frac{\Theta^* - \theta}{\sigma_\varepsilon^*} \leq c\right) \quad \approx \quad \Phi\left(c\right),$$

for a constant c. For large n, a approximate $1-\alpha$ confidence interval is

$$\Theta^* - \lambda_{\alpha/2}\sigma_\varepsilon^* \leq \theta \leq \Theta^* + \lambda_{\alpha/2}\sigma_\varepsilon^*.$$

Exponential distribution

Suppose that we have a sample of size n from an exponential distribution with mean θ. Its MLE is $\Theta^* = \bar{X}$. Find a 95% confidence interval for θ.

Function of Parameter

Let $F_X(x;\theta)$ be the cdf of X. It is common that we want to estimate a function of parameter θ, say $g(\theta)$. The error becomes $g(\theta) - g(\Theta^*)$.

1. If we are interested in θ, then $g(\theta) = \theta$.
2. We are often interested in estimating
   1. the probability that some measured quantity exceeds a critical value u^{crt}, i.e., $p = P(X > u^{\text{crt}}) = 1 - F_X(u^{\text{crt}})$
   2. the α quantile, i.e., x_α such that $P(X > x_\alpha) = \alpha$.

The quantiles are essentially functions of parameter θ. Hence, we are estimating $g(\theta)$.

Delta Method

Let $g\left(r\right)$ possess a continuous derivative $\dot{g}\left(r\right)$. If

$$\theta-\Theta^{*} \quad \in \quad AsN\left(0,\left(\sigma_{\varepsilon}^{2}\right)^{*}\right),$$

then

$$g\left(\theta\right)-g\left(\Theta^{*}\right) \quad \in \quad AsN\left(0,\left[\dot{g}\left(\theta\right)\right]^{2}\left(\sigma_{\varepsilon}^{2}\right)^{*}\right).$$

Earthquake Data with Exponential Distribution

Suppose that we want to estimate $p=P\left(X>1500\right)=\exp\left(-\frac{1500}{\theta}\right)$.

1. Find the MLE of p.
2. Estimate the variance of the MLE.
3. Construct a confidence interval.

Alternative Approach: Parametric Bootstrap

We can use bootstrap to estimate the distribution of $\varepsilon = g\left(\theta\right) - g\left(\Theta^*\right)$.

Algorithm 1: Parametric Bootstrap

```
Specify a large $B$;
Obtain $\theta^*$ from data of size $n$;
for $b$ from $1$ to $B$ do
    Simulate a sample $x^*$ of size $n$ from the distribution $F\left(x; \theta^*\right)$ ;
    Estimate $\theta$ using the simulated sample $x^*$, denoted by $\theta_b^*$ ;
    Obtain the error $\varepsilon_b = \theta^* - \theta_b^*$ ;
end
Use the empirical distribution of $\{\varepsilon_b\}$ to estimate the distribution of
  $\varepsilon = g\left(\theta\right) - g\left(\Theta^*\right)$, denoted by $F_\varepsilon^B\left(e\right)$
```

Alternative Approach: Nonparametric Bootstrap

Algorithm 2: Nonparametric Bootstrap

```
Specify a large $B$;
Obtain $\theta^*$ from data of size $n$;
Obtain the empirical distribution $F_n$ from data of size $n$ ;
for $b$ from $1$ to $B$ do
    Simulate a sample $x^*$ of size $n$ from the empirical distribution ;
    Estimate $\theta$ using the simulated sample $x^*$, denoted by $\theta_b^*$ ;
    Estimate the error by $e_i^B = g\left(\theta^*\right) - g\left(\theta_b^*\right))$ ;
end
The distribution of the error is estimated by the empirical
  distribution of $e_i^B$.
```

Resampling From $F_n(x)$

Since $F_n(x)$ is a distribution function, we can construct a random variable with distribution $F_n(x)$.

- Let x_1, ..., x_n be a sequence of measurements and $F_n(x)$ be the ecdf.
- Let $\tilde{X}$ be a random number having distribution $F_n(x)$. Independent observations $\tilde{X}$ can be generated from $F_n(x)$.

Resample from $F_n(x)$

If our data are 1, 2, 1, 3, 4, explain how we can sample $\tilde{X}$ from $F_n(x)$.

Example: Bootstrap Interval

Let $F_{\varepsilon}^{B}\left(e_{1-\alpha/2}^{B}\right)=\alpha/2$ and $F_{\varepsilon}^{B}\left(e_{\alpha/2}^{B}\right)=1-\alpha/2$. The bootstrap confidence interval with $1-\alpha$ confidence is

$$g\left(\theta^{*}\right)+e_{1-\alpha/2}^{B}\leq g\left(\theta\right)\leq g\left(\theta^{*}\right)+e_{\alpha/2}^{B}.$$

Confidence interval: Earthquake data

Suppose that our data follow an exponential distribution. Find the asymptotic confidence interval, and the bootstrap confidence intervals.