Model Selection and Estimation

A short course authored by the Actuarial Community

19 Jan 2021

Nonparametric Estimation Tools

Nonparametric Estimation

Basic Assumption

- $X_1, \ldots, X_n$ is a random sample (with replacement) from F(.)
- Sometimes we say that $X_1, \ldots, X_n$ are independent and identically distributed (*iid*)

We will not assume a parametric form for cdf $F(.)$ and proceed with a nonparametric analysis

Nonparametric estimation is also referred to as empirical estimation

Moment Estimators

- kth raw moment is $E(X^k) = \mu'_k$
- Empirically estimated by the corresponding statistic

$$\frac{1}{n}\sum_{i=1}^{n} X_i^k$$

- kth central moment is $E(X-\mu)^k = \mu_k$
- Empirically estimated by the corresponding statistic

$$\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^k$$

Empirical Cumulative Distribution Function

- Define the empirical cumulative distribution function

$$\begin{aligned} F_n(x) &= \frac{\text{number of observations less than or equal to } x}{n} \\ &= \frac{1}{n}\sum_{i=1}^{n} I\left(X_i \leq x\right). \end{aligned}$$

  Here, $I(\cdot)$ is an indicator function, it returns 1 if the event $(\cdot)$ is true and 0 otherwise

- When the random variable is discrete, estimate the pmf $f(x) = \Pr(X = x)$ using

$$f_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i = x)$$

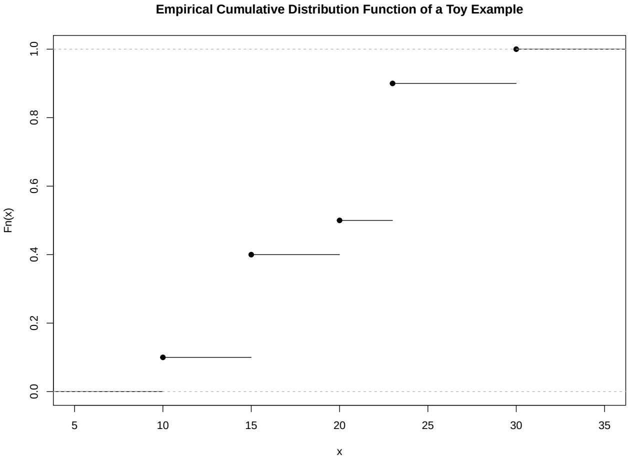
Empirical Example

- **Example – Toy**. Consider $n = 10$ observations:

i	1	2	3	4	5	6	7	8	9	10
X_i	10	15	15	15	20	23	23	23	23	30

- Empirical estimate of the mean (**sample mean**) is $\bar{x} = 19.7$, and empirical estimate of the second central moment (**biased sample variance**) is 31.01

Empirical Cumulative Distribution Function of a Toy Example



Quantiles

- Special Case
  - Median is that number so that half of a data set is below (or above) it
- A $100q$ quantile is that number so that $100 \times q$ percent of data is below it
- In general, for a given $0 < q < 1$, define the $100q$th quantile, q_F, to be any number that satisfies

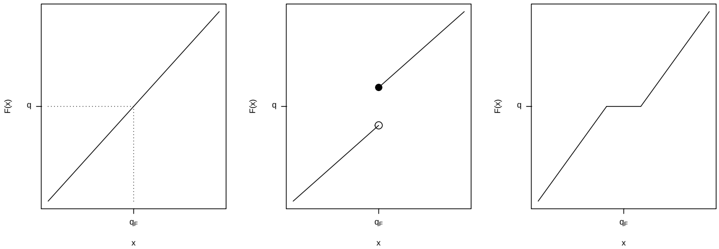
$$F(q_F-) \leq q \leq F(q_F)$$

Here, $F(x-)$ means to evaluate $F(\cdot)$ as a left-hand limit

- If $F(\cdot)$ is continuous at q_F, then $F(q_F-) = F(q_F)$
- Quantile is the most general term for $0 < q < 1$. If $q = 0.01$, 0.02, 0.03, ..., a quantile can be called a percentile

Quantiles

- If F is smooth or there is a jump at q, the quantile q_F is unique
- If F is flat at q, then there are many definitions of q_F



Smoothed Empirical Percentiles

Example – Toy. Consider $n = 10$ observations:

i	1	2	3	4	5	6	7	8	9	10
X_i	10	15	15	15	20	23	23	23	23	30

- Median can defined to be any number between 20 and 23 (many software packages use the average 21.5)
- The smoothed empirical percentile is

$$\hat{\pi}_q = (1 - h)X_{(j)} + hX_{(j+1)}$$

where $j = [(n+1)q]$ and, $h = (n+1)q - j$, and $X_{(1)}, \ldots, X_{(n)}$ are the ordered values (order statistics) corresponding to $X_1, \ldots, X_n$

Smoothed Empirical Percentiles

Example – Toy. Take $n = 10$ and $q = 0.5$. Then,

- $j = [(11)(0.5)] = [5.5] = 5$ and, $h = (11)(0.5) - 5 = 0.5$
- $$\hat{\pi}_{0.5} = (1-0.5)X_{(5)} + (0.5)X_{(6)} = (0.5)(20) + (0.5)(23) = 21.5$$

Take $n = 10$ and $q = 0.2$. Then,

- $j = [(11)(0.2)] = [2.2] = 2$ and $h = (11)(0.2) - 2 = 0.2$
- $$\hat{\pi}_{0.2} = (1-0.2)X_{(2)} + (0.2)X_{(3)} = (0.2)(15) + (0.8)(15) = 15$$

Density Estimators

- When the random variable is discrete, estimate the probability mass function $f(x) = \Pr(X = x)$ is using

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

- Observations may be "grouped" in the sense that they fall into intervals of the form $[c_{j-1}, c_j)$, for $j = 1, \ldots, k$. The constants $\{c_0 < c_1 < \cdots < c_k\}$ form some partition of the domain of F(.).

- Then, use

$$f_n(x) = \frac{n_j}{n \times (c_j - c_{j-1})} \qquad c_{j-1} \leq x < c_j,$$

where n_j is the number of observations (X_i) that fall into the interval $[c_{j-1}, c_j)$.

- Another way to write this is

Uniform Kernel Density Estimator

- Let $b > 0$, known as a "bandwidth,"

$$f_n(x) = \frac{1}{2nb} \sum_{i=1}^{n} I(x - b < X_i \leq x + b).$$

- The estimator is the average over n *iid* realizations of a random variable with mean

$$\begin{aligned} \mathrm{E}\left[\frac{1}{2b} I(x - b < X \leq x + b)\right] &= \frac{1}{2b}\left(F(x+b) - F(x-b)\right) \\ &\to F'(x) = f(x), \end{aligned}$$

as $b \to 0$. That is, $f_n(x)$ is an asymptotically unbiased estimator of $f(x)$.

Kernel Density Estimator

- More generally, define the **kernel density estimator**

$$f_n(x) = \frac{1}{nb} \sum_{i=1}^{n} k\left(\frac{x - X_i}{b}\right).$$

where k is a probability density function centered about 0.

Special Cases

- uniform kernel, $k(y) = \frac{1}{2} I(-1 < y \leq 1)$, .
- triangular kernel, $k(y) = (1 - |y|) \times I(|y| \leq 1)$,
- Epanechnikov kernel, $k(y) = \frac{3}{4}(1 - y^2) \times I(|y| \leq 1)$, and
- Gaussian kernel $k(y) = \phi(y)$, where $\phi(\cdot)$ is the standard normal density function.

Kernel Density Estimator of a Distribution Function

- The kernel density estimator of a **distribution function** is

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x - X_i}{b}\right).$$

  where K is a probability distribution function associated with the kernel density k.

- To illustrate, for the uniform kernel, we have $k(y) = \frac{1}{2} I(-1 < y \leq 1)$ so

$$K(y) = \begin{cases} 0 & y < -1 \\ \frac{y+1}{2} & -1 \leq y < 1 \\ 1 & y \geq 1 \end{cases}$$

Grouped Data

- Observations (X) may be grouped in the sense that they fall into intervals of the form $(c_{j-1}, c_j]$, for $j = 1, \ldots, k$
- Let n_j denote the number of observations that fall into $(c_{j-1}, c_j]$. Total number of observations is $n = \sum_{j=1}^{k} n_j$
- Constants $\{c_0 < c_1 < \cdots < c_k\}$ form some partition of the domain of F(.)
- Empirical cdf at boundaries is defined in the usual way:

$$F_n(c_j) = \frac{\text{number of observations } \le c_j}{n}$$

- For $c_{j-1} < x < c_j$, one could use the ogive, where one connects the values of the boundaries with a straight line:

$$F_n(x) = \frac{c_j - x}{c_j - c_{j-1}} F_n(c_{j-1}) + \frac{x - c_{j-1}}{c_j - c_{j-1}} F_n(c_j)$$

Grouped Data

- Derivative of the ogive is called the histogram:

$$f_n(x) = F_n'(x) = \frac{F_n(c_j) - F_n(c_{j-1})}{c_j - c_{j-1}} = \frac{n_j}{n \times (c_j - c_{j-1})} \text{for } c_{j-1} < x \leq c_j$$

- Another way to write this is

$$f_n(x) = \frac{1}{n(c_j - c_{j-1})} \sum_{i=1}^{n} I(c_{j-1} < X_i \leq c_j)$$

- Histogram assumes a uniform distribution within each interval

REVIEW

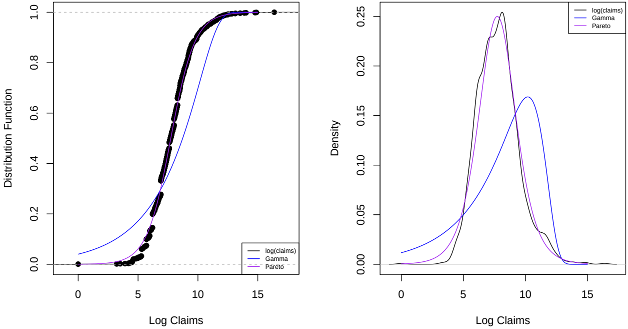
In this section, you learned how to:

- Estimate moments, quantiles, and distributions without reference to a parametric distribution

Tools for Model Selection

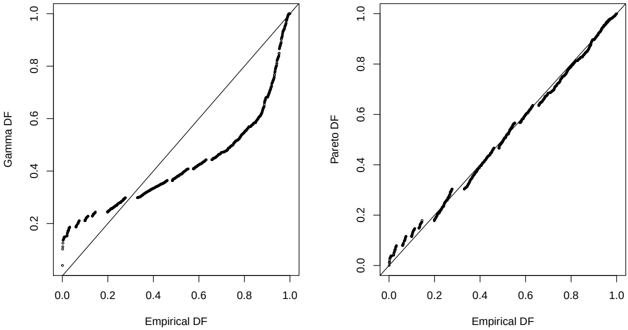
Comparing Distribution and Density Functions

- Left-hand panel compares cdfs, with dots corresponding to the empirical distribution, blue to the fitted gamma, and purple to the fitted Pareto
- Right hand panel compares these three distributions summarized using pdfs



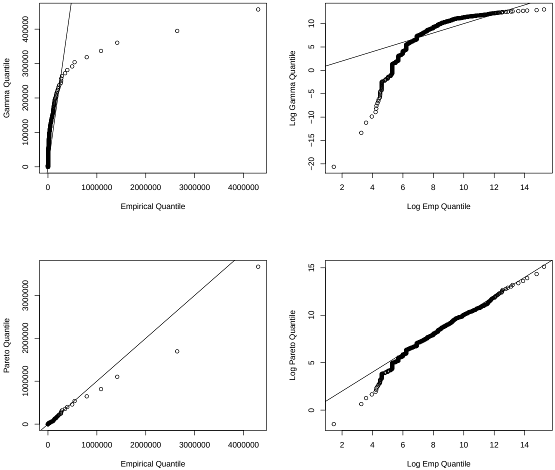
PP Plot

- Horizontal axes gives the empirical cdf at each observation
- In the left-hand panel, the corresponding cdf for gamma is shown in vertical axis
- Right-hand panel shows fitted Pareto distribution. Lines of $y = x$ are superimposed



QQ Plot

- Horizontal axes gives the empirical quantiles at each observation
- Vertical axis gives the quantiles from the fitted distributions
- Pareto distribution fits large observations well, and fits small observations poorly



Goodness-of-Fit Test

- One important type of inference is to determine whether a probability distribution fits a random sample (data) from a certain population well
- We want to determine the goodness-of-fit of a candidate probability model to these data - Do these data represent what you would expect to observe if the population had model cdf $F(.)$?
- Consider the following goodness-of-fit hypothesis test:
  - H_0: Data come from a population with cdf $F(.)$
  - H_1: Data did not come from such a population
  - To determine whether or not to reject H_0, compare a test statistic to a critical value
  - If the test statistic for a one-sided test (absolute value of test statistic for a two-sided test) is smaller than the critical value, you fail to reject H_0: $F(.)$ is an acceptable model for population
  - Otherwise you reject H_0: $F(.)$ is not an acceptable model

Kolmogorov-Smirnov Test

- Consider these data: x_1, x_2, $\ldots$, x_n
- Let $F_n(.)$ denote the empirical cdf
- Model cdf $F(.)$ is assumed to be for a continuous distribution
- Kolmogorov-Smirnov test statistic:

$$\max_x \left(|F_n(x-) - F(x)|, |F_n(x) - F(x)| \right)$$

- Commonly used critical values, where α is significance level:
  - $\frac{1.22}{\sqrt{n}}$ if $\alpha = 0.10$
  - $\frac{1.36}{\sqrt{n}}$ if $\alpha = 0.05$
  - $\frac{1.63}{\sqrt{n}}$ if $\alpha = 0.01$
- Often different critical values are used for really small sample sizes, if there are estimable parameters in $F(.)$, or if data over a certain value cannot be observed (censoring)

Chi-Square (χ^2) Test

- Consider n observations grouped in intervals of the form $(c_{j-1}, c_j]$, for $j = 1, \ldots, k$
  - n_j is the number of observations in $(c_{j-1}, c_j]$
- Assume the model cdf $F(.)$ has r estimable parameters (r can be zero)
  - Let $p_j = F(c_j) - F(c_{j-1}) = \Pr(X \text{ in } (c_{j-1}, c_j])$
  - $E_j = np_j$, the expected number of observations in $(c_{j-1}, c_j]$ under $F(.)$
- Chi-Square (χ^2) test statistic:

$$Q = \sum_{j=1}^{k} \frac{(n_j - E_j)^2}{E_j}$$

- Critical value is the 100(1 - α)% quantile of a chi-square distribution with degrees of freedom equal to (k - 1 - r)

REVIEW

In this section, you learned how to:

- Summarize the data graphically without reference to a parametric distribution
- Determine measures that summarize deviations of a parametric from a nonparametric fit
- Use nonparametric estimators to approximate parameters that can be used to start a parametric estimation procedure

Model Selection: Likelihood Ratio Tests and Goodness of Fit

Likelihood Ratio Test

One important type of inference is to select one of two candidate models, where one model (reduced model) is a special case of the other model (full model)

In a Likelihood Ratio Test, we conduct the hypothesis test:

- H_0: Reduced model is correct
- H_1: Full model is correct

Likelihood Ratio Test Process

To conduct the Likelihood Ratio Test:

- Determine the maximum likelihood estimator for full model, $\widehat{\theta}_{\textbf{Full}}$
- Now assume that p restrictions are placed on the parameters of the full model to create the reduced model; determine the maximum likelihood estimator for the reduced model, $\hat{\boldsymbol{\theta}}_{\boldsymbol{Reduced}}$
- $LRT = 2\left(l(\widehat{\theta}_{\textbf{Full}}) - l(\widehat{\theta}_{\textbf{Reduced}})\right)$ is the likelihood ratio. Under the null hypothesis, the likelihood ratio has a chi-square distribution with degrees of freedom equal to p. LRT is the test statistic
- Critical value is a quantile ($100(1-\alpha)\%$ for significance level α) from a chi-square distribution with degrees of freedom equal to p - If LRT is large relative to the critical value, then we reject the reduced model in favor of the full model

Information Criteria: Exam STAM Version

- Following statistics can be used when comparing several candidate models that are not necessarily nested (as in the Likelihood Ratio Test). One picks the model that maximizes the criterion
- *Akaike's Information Criterion* (AIC)

$$AIC = l(\widehat{\theta}_{\mathbf{MLE}}) - (number\ of\ parameters)$$

- Additional term (*number of parameters*) is a penalty for the complexity of the model
- Other things equal, a more complex model means more parameters, resulting in a smaller value of the criterion
- *Bayesian Information Criterion* (BIC)

$$\begin{aligned} BIC &= l(\widehat{\theta}_{\mathbf{MLE}}) \\ &\quad -(0.5)(number\ of\ parameters)\ln(number\ of\ observations) \end{aligned}$$

- This measure gives greater weight to the number of parameters, resulting in a larger penalty
- Other things being equal, *BIC* will suggest a more parsimonious model than *AIC*

Information Criteria: Alternative Version

- One picks the model that minimizes the criterion
- *Akaike's Information Criterion* (AIC)

$$AIC = -2 \times l(\widehat{\theta}_{\mathbf{MLE}}) + 2 \times (number\ of\ parameters)$$

- *Bayesian Information Criterion* (BIC)

$$\begin{aligned} BIC &= -2 \times l(\widehat{\theta}_{\mathbf{MLE}}) \\ &\quad +(number\ of\ parameters) \ln(number\ of\ observations) \end{aligned}$$

Information Criteria Example

Distribution	*AIC*	*BIC*
Gamma	$28,305.2$	$28,315.6$
Lognormal	$26,837.7$	$26,848.2$
Pareto	$26,813.3$	$26,823.7$
GB2	$26,768.1$	$26,789.0$

Estimation using Modified Data: Nonparametric Approach

Grouped Data

- Observations may be “grouped” in the sense that they fall into intervals of the form $[c_{j-1}, c_j)$, for $j = 1, \ldots, k$.
- The constants $\{c_0 < c_1 < \cdots < c_k\}$ form some partition of the domain of F(.).
- Define the empirical distribution function at the boundaries is defined in the usual way:

$$F_n(c_j) = \frac{\text{number of observations } \leq c_j}{n}$$

- For other values of x, one could use the

Ogive: connect values of the boundaries with a straight line. - For another way of smoothing, recall the kernel density estimator of the distribution function

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} K\left(\frac{x - X_i}{b}\right).$$

- For densities, use

$$f_n(x) = \frac{n_j}{n \times (c_j - c_{j-1})} \qquad c_{j-1} \leq x < c_j$$

Censored Data

- Censoring occurs when we observe only a limited value of an observation.
- Suppose that X represents a loss due to an insured event and that u is a known censoring point.
- If observations are censored from the **right** (or from above), then we observe
$$Y = \min(X, u).$$
- In this case, u may represent the upper limit of coverage for an insurer. The loss exceeds the amount u but the insurer does not have in its records the amount of the actual loss.
- If observations are censored from the **left** (or from below), then we observe
$$Y = \max(X, u).$$
- Let u represents the upper limit of coverage but now $Y - u$ represents the amount that a *reinsurer* is responsible for. If the loss $X < u$, then $Y = 0$, no loss for the reinsurer. If the loss $X \geq u$, then $Y = X - u$ represents the reinsurer's retained claims.

Kaplan-Meier Product Limit Estimator

- Let $t_1 < \cdots < t_c$ be distinct points at which an event of interest occurs, or non-censored losses, and let s_j be the number of events at time point t_j .
- Further, the corresponding "risk set" is the number of observations that are active at an instant just prior to t_j . Using notation, the risk set is $R_j = \sum_{i=1}^{n} I(x_i \geq t_j)$.
- With this notation, the **product-limit estimator** of the distribution function is

$$\hat{F}(x) = \begin{cases} 0 & x < t_1 \\ 1 - \prod_{j:t_j \leq x} \left(1 - \frac{s_j}{R_j}\right) & x \geq t_1. \end{cases}$$

- Greenwood (1926) derived the formula for the estimated variance

$$\widehat{Var}(\hat{F}(x)) = (1 - \hat{F}(x))^2 \sum_{j:t_j \leq x} \frac{s_j}{R_j(R_j - s_j)}.$$

REVIEW

In this section, you learned how to:

- Describe grouped and censored truncated data
- Estimate distributions nonparametrically based on grouped and censored data

Estimation using Modified Data: Parametric Approach

Truncated Data

- An outcome is potentially **truncated** when the availability of an observation depends on the outcome.

- In insurance, it is common for observations to be truncated from the **left** (or below) at d when the amount observed is

$$Y = \begin{cases} \text{we do not observe X} & X < d \\ X - d & X \geq d. \end{cases}$$

- In this case, d may represent the deductible associated with an insurance coverage. If the insured loss is less than the deductible, then the insurer does not observe the loss. If the loss exceeds the deductible, then the excess $X - d$ is the claim that the insurer covers.

- Observations may also truncated from the **right** (or above) at d when the amount observed is

$$Y = \begin{cases} X & X < d \\ \text{we do not observe X} & X \geq d \end{cases}$$

- Classic examples of truncation from the right include X as a measure of distance of a star. When the distance exceeds a certain level d, the star is no longer observable.

Censored Data

- Suppose that X represents a loss due to an insured event and that u is a known censoring point
- If observations are right censored (from above), we observe

$$Y = \min(X, u)$$

- u may represent the upper limit of coverage for an insurer. Loss exceeds u but the insurer does not have in its records the amount of the actual loss
- If observations are left censored (from below), we observe

$$Y = \max(X, u)$$

- Let u represent the upper limit of coverage but now $Y - u$ represents the amount that a reinsurer is responsible for. If the loss $X < u$, then $Y = 0$ for the reinsurer. If the loss $X \geq u$, then $Y = X - u$ represents the reinsurer's retained claims

Truncated Data

- In insurance, it is common for observations to be left truncated (from below) at d when amount observed is

$$Y = \begin{cases} \text{we do not observe X} & X \leq d \\ X - d & X > d. \end{cases}$$

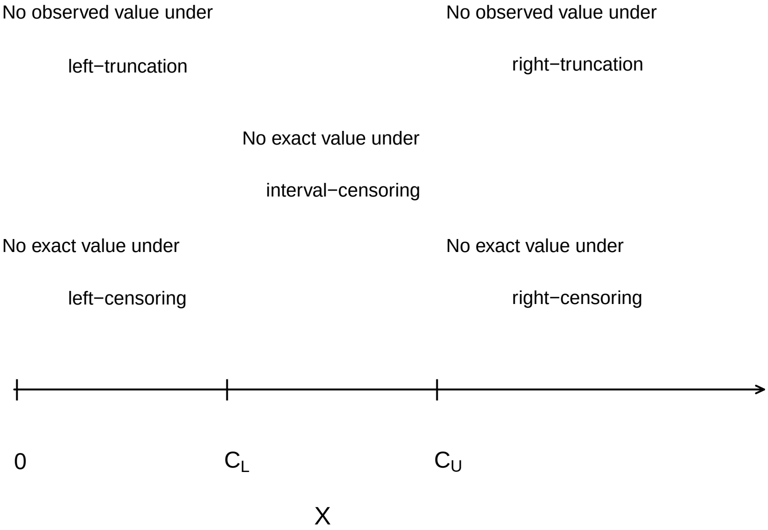
- d may represent the deductible associated with an insurance coverage. If the insured loss is less than deductible, then the insurer does not observe the loss. If the loss exceeds deductible, then the excess $X - d$ is the claim that the insurer covers

- Observations may also be right truncated (from above) at d when the amount observed is

$$Y = \begin{cases} X & X \leq d \\ \text{we do not observe X} & X > d \end{cases}$$

- Consider X as a measure of distance of a star. When the distance exceeds a certain level d, the star is no longer observable

Censored and Truncated Data



No observed value under

left−truncation

No observed value under

right−truncation

No exact value under

interval−censoring

No exact value under

left−censoring

No exact value under

right−censoring

Maximum Likelihood Estimation with Grouped Data

- Probability of an observation X falling in the jth interval is

$$\Pr\left(X \in (c_{j-1}, c_j]\right) = F(c_j) - F(c_{j-1})$$

- Corresponding pmf is

$$\begin{aligned} f(x) &= \begin{cases} F(c_1) - F(c_0) & \text{if } x \in (c_0, c_1] \\ \vdots & \vdots \\ F(c_k) - F(c_{k-1}) & \text{if } x \in (c_{k-1}, c_k] \end{cases} \\ &= \prod_{j=1}^{k} \left\{F(c_j) - F(c_{j-1})\right\}^{I(x \in (c_{j-1}, c_j])} \end{aligned}$$

- Likelihood is

$$L(\boldsymbol{\theta}) = \prod_{j=1}^{n} f(x_i) = \prod_{j=1}^{k} \left\{F(c_j) - F(c_{j-1})\right\}^{n_j}$$

- Log-likelihood is

$$l(\boldsymbol{\theta}) = \ln \prod_{j=1}^{n} f(x_i) = \sum_{j=1}^{k} n_j \ln\left\{F(c_j) - F(c_{j-1})\right\}$$

Censored Data Likelihood

- Suppose that X represents a loss due to an insured event and that u is a known censoring point
- If observations are censored from the **right** (or from above), then we observe $Y = \min(X, u)$ and $\delta_u = \mathrm{I}(X \geq u)$
- If censoring occurs so that $\delta_u = 1$, then $X \geq u$ and the likelihood is $\Pr(X \geq u) = 1 - \mathrm{F}(u)$
- If censoring does not occur so that $\delta_u = 0$, then $X < C_U$ and the likelihood is $\mathrm{f}(y)$

$$\begin{aligned} \mathit{Likelihood} &= \begin{cases} \mathrm{f}(y) & \text{if } \delta = 0 \\ 1 - \mathrm{F}(u) & \text{if } \delta = 1 \end{cases} \\ &= \left(\mathrm{f}(y)\right)^{1-\delta} \left(1 - \mathrm{F}(u)\right)^{\delta}. \end{aligned}$$

Censored Data Likelihood

- For a single observation, we have

$$Likelihood = (\mathrm{f}(y))^{1-\delta} (1 - \mathrm{F}(u))^{\delta} .$$

- Consider a random sample of size n, $\{(y_1, \delta_1), \ldots, (y_n, \delta_n)\}$ with potential censoring times $\{u_1, \ldots, u_n\}$
- Likelihood is

$$\prod_{i=1}^{n} (\mathrm{f}(y_i))^{1-\delta_i} (1 - \mathrm{F}(u_i))^{\delta_i} = \prod_{\delta_i=0} \mathrm{f}(y_i) \prod_{\delta_i=1} \{1 - \mathrm{F}(u_i)\},$$

- Here, notation "$\prod_{\delta_i=0}$" means take the product over the uncensored observations, and similarly for "$\prod_{\delta_i=1}$"
- Log-likelihood is

$$l(\theta) = \sum_{i=1}^{n} \{(1 - \delta_i) \ln \mathrm{f}(y_i) + \delta_i \ln (1 - \mathrm{F}(u_i))\}$$

Maximum Likelihood Estimation Using Censored and Truncated Data

- Truncated data are handled in likelihood inference via conditional probabilities
- Adjust the likelihood contribution by dividing by the probability that the variable was observed
- Summarizing, we have the following contributions to the likelihood for six types of outcomes

Outcome	Likelihood Contribution
exact value	$f(x)$
right-censoring	$1 - F(C_U)$
left-censoring	$F(C_L)$
right-truncation	$f(x)/F(C_U)$
left-truncation	$f(x)/(1 - F(C_L))$
interval-censoring	$F(C_U) - F(C_L)$

Maximum Likelihood Estimation Using Censored and Truncated Data

- For known outcomes and censored data, the likelihood is

$$\prod_E \mathrm{f}(x_i) \prod_R \{1 - \mathrm{F}(C_{Ui})\} \prod_L \mathrm{F}(C_{Li}) \prod_I (\mathrm{F}(C_{Ui}) - \mathrm{F}(C_{Li})),$$

where "$\prod_E$" is product over observations with *E*xact values, and similarly for *R*ight-, *L*eft- and *I*nterval-censoring

- For right-censored and left-truncated data, the likelihood is

$$\prod_E \frac{\mathrm{f}(x_i)}{1 - \mathrm{F}(C_{Li})} \prod_R \frac{1 - \mathrm{F}(C_{Ui})}{1 - \mathrm{F}(C_{Li})},$$

- Similarly for other combinations

REVIEW

In this section, you learned how to:

- Describe grouped, censored, and truncated data
- Estimate parametric distributions based on grouped, censored, and truncated data

Bayesian Inference

Bayesian Inference

- In the **frequentist interpretation**, one treats the vector of parameters $\boldsymbol{\theta}$ as fixed yet unknown, whereas the outcomes X are realizations of random variables.
- With Bayesian statistical models, one views both the model parameters and the data as random variables.
- Use probability tools to reflect this uncertainty about the parameters $\boldsymbol{\theta}$.
- For notation, we will think about $\boldsymbol{\theta}$ as a random vector and let $\pi(\boldsymbol{\theta})$ denote the distribution of possible outcomes.

Bayesian Inference Strengths

There are several advantages of the Bayesian approach.

- One can describe the entire distribution of parameters conditional on the data. This allows one, for example, to provide probability statements regarding the likelihood of parameters.
- This approach allows analysts to blend information known from other sources with the data in a coherent manner. This topic is developed in detail in the credibility chapter.
- The Bayesian approach provides for a unified approach for estimating parameters. Some non-Bayesian methods, such as least squares, required a approach to estimating variance components. In contrast, in Bayesian methods, all parameters can be treated in a similar fashion. Convenient for explaining results to consumers of the data analysis.
- Bayesian analysis is particularly useful for forecasting future responses.

Bayesian Model

- **Prior Distribution.** $\pi(\boldsymbol{\theta})$ is called the *prior distribution*.
  - Typically, it is a regular distribution and so integrates to one.
  - We may be very uncertain (or have no clue) about the distribution of $\boldsymbol{\theta}$; the Bayesian machinery allows this situation

$$\int \pi(\theta) d\theta = \infty$$

in which case $\pi(\cdot)$ is called an **improper prior**.

- **Model Distribution.** The distribution of outcomes given an assumed value of $\boldsymbol{\theta}$ is known as the *model distribution* and denoted as $f(x|\boldsymbol{\theta}) = f_{X|\boldsymbol{\theta}}(x|\boldsymbol{\theta})$. This is the (usual frequentist) mass or density function.
- **Joint Distribution**
- **Marginal Outcome Distribution**
- **Posterior Distribution of Parameters**

Bayesian Model

- **Prior Distribution**
- **Model Distribution**
- **Joint Distribution.** The distribution of outcomes and model parameters is, not surprisingly, known as the *joint distribution* and denoted as $f(x, \boldsymbol{\theta}) = f(x|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$.
- **Marginal Outcome Distribution.** The distribution of outcomes can be expressed as

  $$f(x) = \int f(x|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}.$$

  This is analogous to a frequentist mixture distribution.
- **Posterior Distribution of Parameters**

Bayesian Model

- **Prior Distribution**

- **Model Distribution**

- **Joint Distribution**

- **Marginal Outcome Distribution**

- **Posterior Distribution of Parameters.** After outcomes have been observed (hence the terminology "posterior"), one can use Bayes theorem to write the distribution as

  $$\pi(\boldsymbol{\theta}|x) = \frac{f(x,\boldsymbol{\theta})}{f(x)} = \frac{f(x|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(x)}$$

  The idea is to update your knowledge of the distribution of $\boldsymbol{\theta}$ ($\pi(\boldsymbol{\theta})$) with the data x.

  - We can summarize the distribution using a confidence interval type statement.
  - **Definition**. $[a, b]$ is said to be a $100(1-\alpha)\%$ **credibility interval** for $\boldsymbol{\theta}$ if

    $$\Pr(a \le \theta \le b|\mathbf{x}) \ge 1 - \alpha.$$

Two Examples

Exam C Question 157. You are given:

(i) In a portfolio of risks, each policyholder can have at most one claim per year.

(ii) The probability of a claim for a policyholder during a year is q.

(iii) The prior density is

$$\pi(q) = q^3/0.07, \quad 0.6 < q < 0.8$$

A randomly selected policyholder has one claim in Year 1 and zero claims in Year 2. For this policyholder, calculate the posterior probability that $0.7 < q < 0.8$.

Exam C Question 43. You are given:

(i) The prior distribution of the parameter Θ has probability density function:

$$\pi(\theta) = 1/\theta^2, \quad 1 < \theta < \infty$$

(ii) Given $\Theta = \theta$, claim sizes follow a Pareto distribution with parameters $\alpha = 2$ and θ.

A claim of 3 is observed. Calculate the posterior probability that Θ exceeds 2.

Decision Analysis

- In classical decision analysis, the loss function $l(\hat{\theta}, \theta)$ determines the penalty paid for using the estimate $\hat{\theta}$ instead of the true θ.
- The **Bayes estimate** is that value that minimizes the expected loss $\mathrm{E}\ l(\hat{\theta}, \theta)$.
- Some important special cases include:

Loss function $l(\hat{\theta}, \theta)$	Descriptor	Bayes Estimate
$(\hat{\theta} - \theta)^2$	squared error loss	$\mathrm{E}(\theta \vert X)$
$\vert\hat{\theta} - \theta\vert$	absolute deviation loss	median of $\pi(\theta \vert x)$
$I(\hat{\theta} = \theta)$	zero-one loss (for discrete probabilities)	mode of $\pi(\theta \vert x)$

- For new data y, the predictive distribution is

$$f(y|x) = \int f(y|\theta)\pi(\theta|x)d\theta.$$

- With this, the Bayesian prediction of y is

$$\begin{aligned}\mathrm{E}(y|x) &= \int yf(y|x)dy = \int y\left(\int f(y|\theta)\pi(\theta|x)d\theta\right)dy \\ &= \int \mathrm{E}(y|\theta)\pi(\theta|x)d\theta.\end{aligned}$$

Posterior Distribution

How to calculate the posterior distribution?

- **By hand** - can do this in special cases
- **Simulation** - uses modern computational techniques. **KPW** (Section 12.4.4) mentions Markov Chain Monte Carlo (MCMC) simulation
- **Normal Approximation**. Theorem 12.39 of **KPW** provides a justification
- **Conjugate distributions**. Classical approach. Although this approach is available only for a limited number of distributions, it has the appeal that it provides closed-form expressions for the distributions, allowing for easy interpretations of results. We focus on this approach.

To relate the prior and posterior distributions of the parameters, we have

$$\begin{array}{ccc} \pi(\boldsymbol{\theta}|x) & = & \dfrac{f(x|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(x)} \\ & \propto & f(x|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \\ \text{Posterior} & \text{is proportional to} & \text{likelihood} \times \text{prior} \end{array}$$

For **conjugate distributions**, the posterior and the prior come from the same family of distributions.

Poisson–Gamma Conjugate Family

- Assume a Poisson(λ) model distribution so that

$$f(\mathbf{x}|\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}$$

- Assume λ follows a gamma(α, θ) prior distribution so that

$$\pi(\lambda) = \frac{(\lambda/\theta)^{\alpha} \exp(-\lambda/\theta)}{\lambda \Gamma(\alpha)}.$$

- The posterior distribution is proportional to

$$\begin{aligned} \pi(\lambda|\mathbf{x}) &\propto f(\mathbf{x}|\theta)\pi(\lambda) \\ &= C\lambda^{\sum_i x_i + \alpha - 1} \exp(-\lambda(n + 1/\theta)) \end{aligned}$$

  where C is a constant.
- We recognize this to be a gamma distribution with new parameters $\alpha_{new} = \sum_i x_i + \alpha$ and $\theta_{new} = 1/(n + 1/\theta)$.

REVIEW

In this section, you learne how to:

- Describe the Bayesian model as an alternative to the frequentist approach and summarize the five components of this modeling approach.
- Summarize posterior distributions of parameters and use these posterior distributions to predict new outcomes.
- Use conjugate distributions to determine posterior distributions of parameters.