Model Selection and Inference

Fall 2017

Outline

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 identically and independently distributed (iid)

We will not assume a parametric form for the distribution function F() and so proceed with a *nonparametric* analysis

Nonparametric estimation is also referred to as “empirical estimation”

Moment Estimators

- The kth raw moment is $\mathrm{E}\, X^k = \mu'_k$
- It is nonparametrically estimated by the corresponding statistic

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

- The kth central moment is $\mathrm{E}\,(X-\mu)^k = \mu_k$
- It is nonparametrically estimated by the corresponding statistic

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

Empirical Distribution Function

- Define the **empirical distribution function** to be

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

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

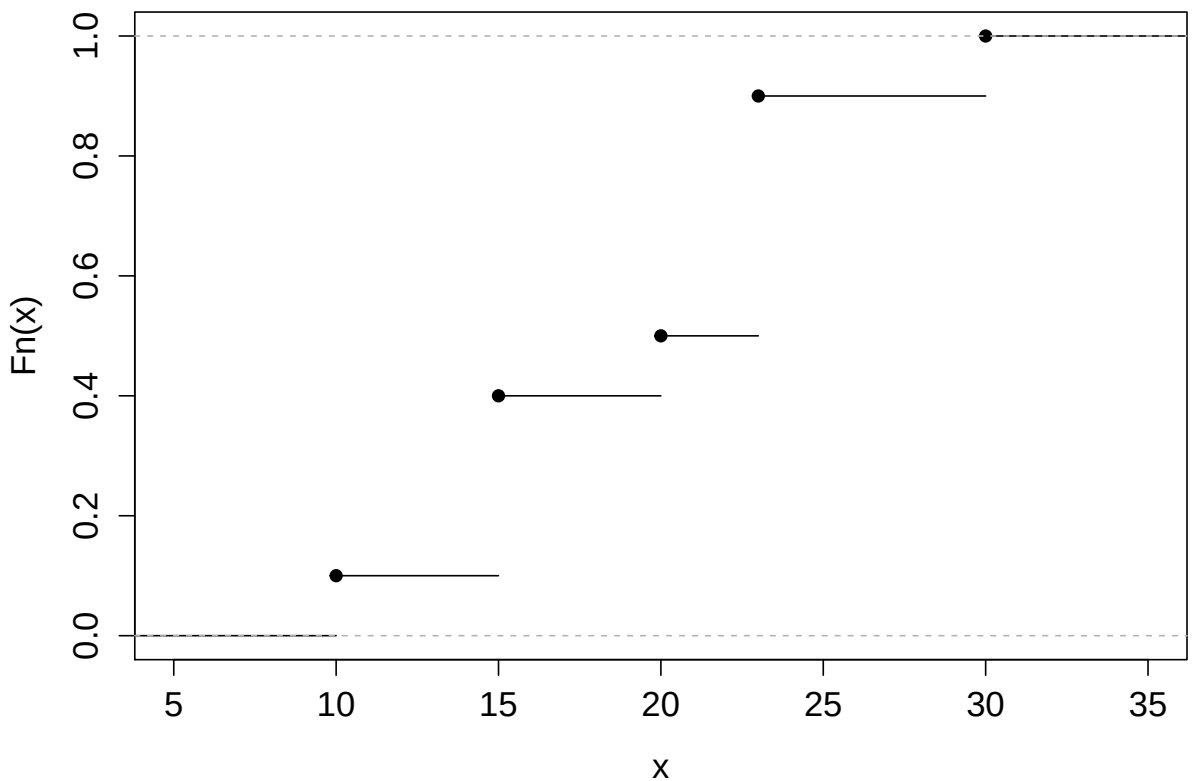
- **Example – Toy**. Consider $n = 10$ observations:

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

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

Empirical Distribution Function II

Figure: Empirical Distribution Function of a Toy Example



Percentiles I

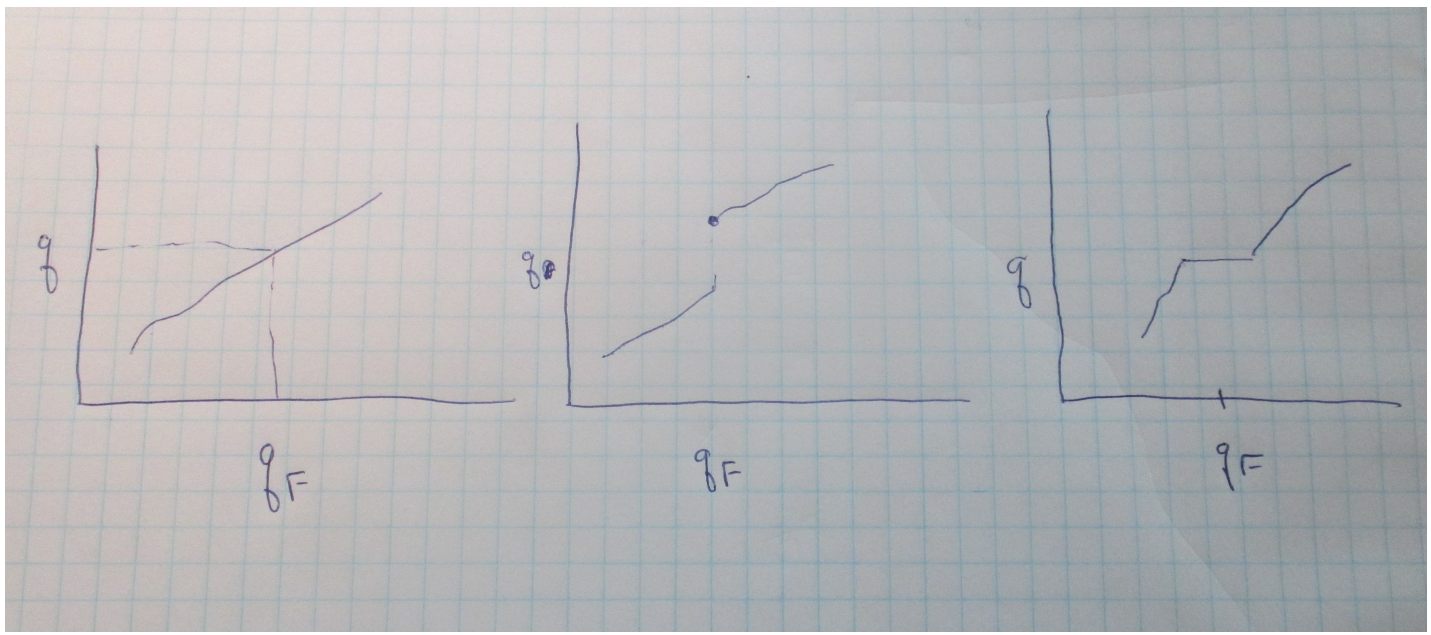
- Special Cases
  - The *median* is that number so that half of a data set is below (or above) it
  - The first *quartile* is that number so that 25% of the data is below it
  - A $100q$ *percentile* is that number so that $100 \times q$ percent of the data is below it
- In general, for a given $0 < q < 1$, define the $100q$**th percentile (or quantile)**, q_F, to be any number that satisfies

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

  Here, the notation $F(x-)$ means to evaluate the function $F(\cdot)$ as a left-hand limit
- If $F(\cdot)$ is continuous at q_F, then $F(q_F-) = F(q_F)$

Percentiles II

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



Smoothed Empirical Percentiles I

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

- The median might be defined to be any number between 20 and 23
  - Many software packages use the average 21.5
- KPW defines the *smoothed empirical percentile* to be

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

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

Smoothed Empirical Percentiles II

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

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

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

$$\begin{aligned} \mathrm{E}\, \frac{1}{2b} I(x - b < X \leq x + b) &= \frac{1}{2b} \left(F(x + b) - F(x - b)\right) \\ &= \frac{1}{2b} \left(\left\{F(x) + bF'(x) + b^2 C_1\right\}\right. \\ &\quad \left. - \left\{F(x) - bF'(x) + b^2 C_2\right\}\right) \\ &= F'(x) + b\frac{C_1 - C_2}{2} \to F'(x) = f(x), \end{aligned}$$

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: $k(x) = \frac{1}{2} I(|x| \leq 1)$,
  - triangular kernel: $k(x) = (1 - |x|) \times I(|x| \leq 1)$,
  - Epanechnikov kernel: $k(x) = \frac{3}{4}(1 - x^2) \times I(|x| \leq 1)$,
  - Gaussian kernel: $k(x) = \phi(x)$, 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

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

Comparing Distribution and Density Functions

- The left-hand panel compares distribution functions, with the dots corresponding to the empirical distribution, the thick blue curve corresponding to the fitted gamma and the light purple curve corresponding to the fitted Pareto.
- The right hand panel compares these three distributions summarized using probability density functions.

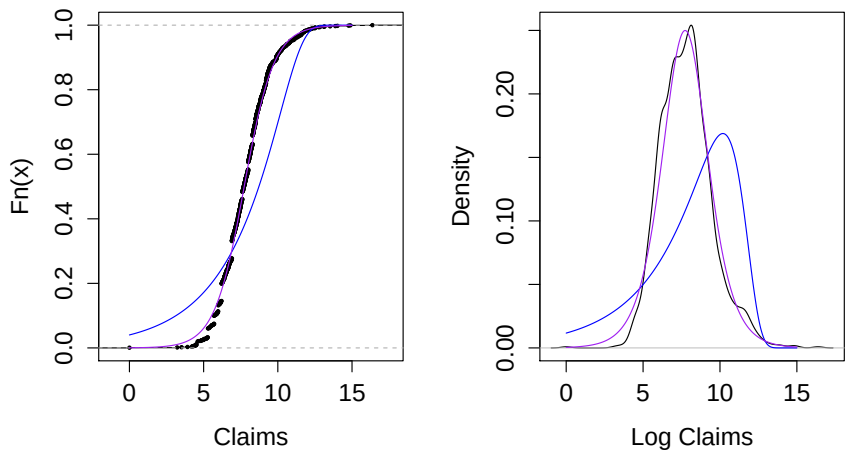


Figure: Nonparametric Versus Fitted Parametric Distribution and Density Functions.

PP Plot

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

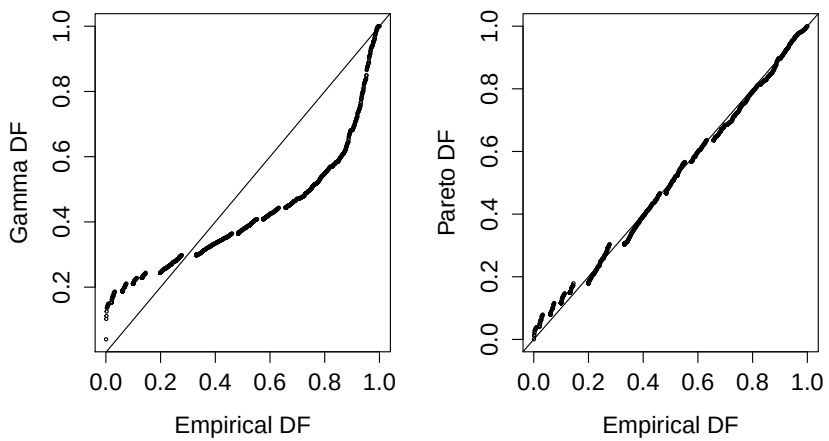
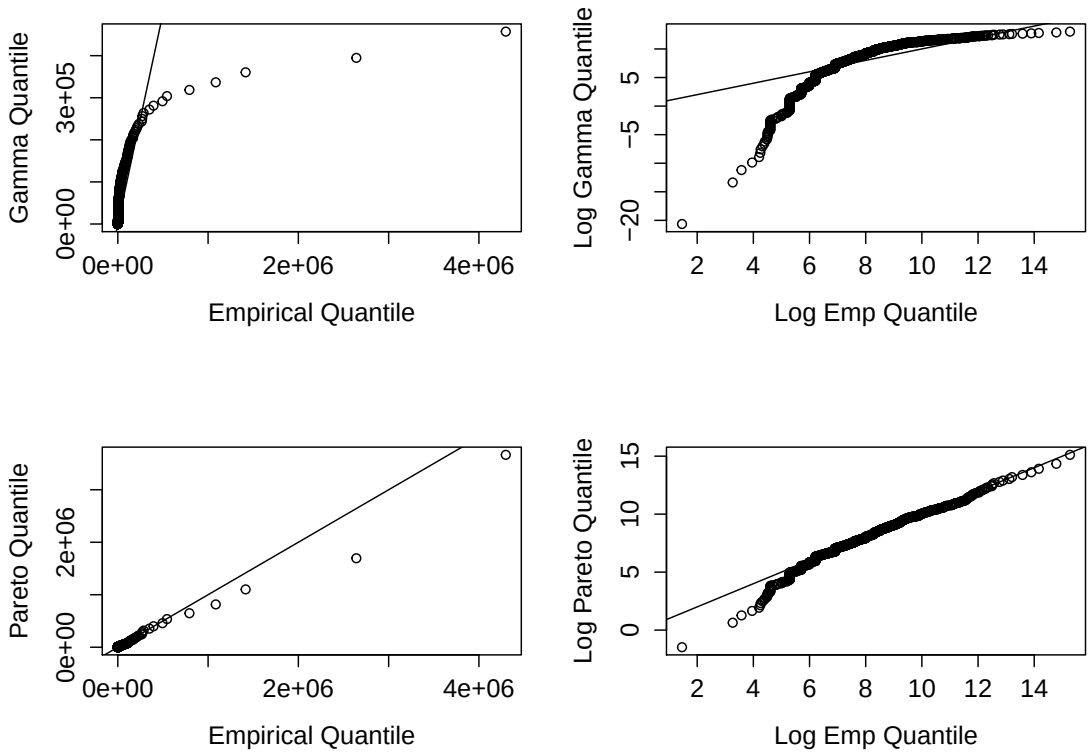


Figure: Probability-Probability (pp) Plots.

KPW also recommends plotting the difference $D(x) = F_n(x) - F^*(x)$ versus x. Here, $F^*(x)$ is the fitted model distribution function.

QQ Plot

- The horizontal axes gives the empirical quantiles at each observation.
- The right-hand panels they are graphed on a logarithmic basis.
- The vertical axis gives the quantiles from the fitted distributions; Gamma quantiles are in the upper panels, Pareto quantiles are in the lower panels.
- The lower-right hand panel suggests that the Pareto distribution does a good job with large observations but provides a poorer fit for small observations.



Three Goodness of Fit Statistics

Statistic	Definition	Computational Expression
Kolmogorov -Smirnov	$sup_x \lvert F_n(x) - F(x) \rvert$	$max(D^+ - D^-)$ where $D^+ = max_{i=1,\ldots,n}\left(\frac{i}{n} - F_i\right)$ $D^- = max_{i=1,\ldots,n}\left(F_i - \frac{i-1}{n}\right)$
Cramer -von Mises	$n\int (F_n(x) - F(x))^2 dx$	$\frac{1}{12n} + \sum_{i=1}^{n}\left(F_i - (2i-1)/n\right)^2$
Anderson -Darling	$n\int \frac{(F_n(x)-F(x))^2}{F(x)(1-F(x))} dx$	$-n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\log\left(F_i(1 - F_{n+1-i})\right)^2$

where F_i is defined to be $F(x_i)$.

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

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

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.

Right-Censored, Left-Truncated Empirical Distribution Functions

- Procedure from **KPW**. Notation:
  - For each observation i, let d_i be the lower truncation limit (0 if no truncation)
  - Let u_i be the upper censoring limit (=∞ if no censoring)
  - The recorded value is x_i in the case of no censoring, u_i if there is censoring.
  - For notation, let $t_1 < \cdots < t_k$ be k unique observations of x_i that are uncensored.
  - Define s_j to be the number of x_i's at t_j.
  - Define the risk set

$$R_j = \sum_{i=1}^{n} I(x_i \geq t_j) + \sum_{i=1}^{n} I(u_i \geq t_j) - \sum_{i=1}^{n} I(d_i \geq t_j)$$

- The product-limit estimator of the distribution function is

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

- The Nelson-Äalen estimator of the distribution function is

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

Starting Values

- Maximum likelihood is a desirable estimation technique because
  - It employs data efficiently (enjoys certain optimality properties)
  - It can be used in a variety of data sampling schemes (e.g.,*iid*, grouped, censored, regression, and so forth)
- However, maximum likelihood is a recursive estimation procedure that requires starting values to begin the recursion
- Two alternative estimation techniques are:
  - Method of moments
  - Percentile matching
- These are non-recursive techniques. Easy to implement and explain. Although less efficient than maximum likelihood, they can be employed to provide starting values for maximum likelihood.

Method of Moments

- Idea: Approximate the moments using a parametric distribution to the empirical (nonparametric) moments
- Assume we have a random sample X_1, ..., X_n, with all observations from the same (assumed) parametric distribution
- The parametric distribution has cdf $F(x;\theta)$, where $\theta = (\theta_1, ..., \theta_p)$ is a vector of p parameters to be estimated
- A **method of moments** estimate of θ is any solution of the p equations:

$$E[X^k] = \frac{1}{n}\sum_{i=1}^{n} X_i^k \text{ for } k = 1, 2, ..., p$$

- In theory, any p moments could be used (for example, negative moments to estimate the parameters of an inverse distribution)

Method of Moments - Example

- **Example - Property Fund.** For the 2010 property fund, there are $n = 1,377$ individual claims (in thousands of dollars) with

$$m_1 = \frac{1}{n}\sum_{i=1}^{n} X_i = 26.62259 \quad \text{and} \quad m_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 = 136154.6.$$

- Gamma Distribution
  - From theory, $\mu_1 = \alpha\theta$ and $\mu_2' = \alpha(\alpha+1)\theta^2$.
  - Equating the two yields the method of moments estimators, easy algebra shows that

$$\alpha = \frac{\mu_1^2}{\mu_2' - 2\mu_1^2} \quad \text{and} \quad \theta = \frac{\mu_2' - \mu_1^2}{\mu_1}.$$

  - The method of moment estimators are

$$\begin{aligned} \hat{\alpha} &= \frac{26.62259^2}{136154.6 - 26.62259^2} = 0.005232809 \\ \hat{\theta} &= \frac{136154.6 - 26.62259^2}{26.62259} = 5,087.629. \end{aligned}$$

  - In contrast, the maximum likelihood values turn out to be $\hat{\alpha}_{MLE} = 0.2905959$ and $\hat{\theta}_{MLE} = 91.61378$
  - Big discrepancies between the two estimation procedures, suggesting that the gamma model fits poorly.

Method of Moments - Example II

- **Example - Property Fund.** Recall the nonparametric estimates

$$m_1 = \frac{1}{n}\sum_{i=1}^{n} X_i = 26.62259 \quad \text{and} \quad m_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 = 136154.6.$$

- Pareto Distribution
  - From theory, $\mu_1 = \theta/(\alpha-1)$ and $\mu_2' = 2\theta^2/((\alpha-1)(\alpha-2))$.
  - Easy algebra shows

$$\alpha = 1 + \frac{\mu_2'}{\mu_2' - 2\mu_1^2} \quad \text{and} \quad \theta = (\alpha-1)\mu_1.$$

  - The method of moment estimators are

$$\begin{aligned} \hat{\alpha} &= 1 + \frac{136154.6}{136154.6 - 2 * 26,62259^2} = 2.01052 \\ \hat{\theta} &= (2.01052 - 1) \cdot 26.62259 = 26.9027 \end{aligned}$$

  - The maximum likelihood values turn out to be $\hat{\alpha}_{MLE} = 0.9990936$ and $\hat{\theta}_{MLE} = 2.2821147$.
  - Interesting that $\hat{\alpha}_{MLE} < 1$; for the Pareto distribution; recall that $\alpha < 1$ means that the mean is infinite.
  - Indicates that the property claims data set is a long tail distribution.

Percentile Matching

- Under percentile matching, one approximates the parametric distribution using the empirical (nonparametric) percentiles
- Assume we have a random sample X_1, ..., X_n, with all observations from the same (assumed) parametric distribution
- The parametric distribution has cdf $F(x;\theta)$, where $\theta = (\theta_1, ..., \theta_p)$ is a vector of p parameters to be estimated
- Let g_k denote an arbitrarily chosen value between 0 and 1 (0% and 100%), for $k = 1, 2, ..., p$
- Let $\hat{\pi}_{g_k}$ denote an empirical estimate of the percentile that corresponds to g_k
- A **percentile matching** estimate of θ is any solution of the p equations:

$$F[\hat{\pi}_{g_k};\theta] = g_k \text{ for } k = 1, 2, ..., p$$

Percentile Matching - Example

- **Example - Property Fund.**
- The 25th percentile (the first quartile) turns out to be 0.78853 and the 95th percentile is 50.98293 (both in thousands of dollars).
- Pareto Distribution
  - The Pareto distribution is particularly intuitively pleasing because of the closed-form solution for the quantiles.
  - The distribution function is $F(x) = 1 - (\theta/(x+\theta))^{\alpha}$.
  - Easy algebra shows that we can express the quantile as

$$F^{-1}(q) = \theta\left((1-q)^{-1/\alpha} - 1\right)$$

    for a fraction q, $0 < q < 1$.
  - With two equations

$$0.78853 = \theta\left((1-.25)^{-1/\alpha} - 1\right) \quad \text{and} \quad 50.98293 = \theta\left((1-.95)^{-1/\alpha} - 1\right)$$

    and two unknowns, the solution is $\hat{\alpha} = 0.9412076$ and $\hat{\theta} = 2.205617$.
  - A numerical routine was required for these solutions - no analytic solution available.
  - Recall that the maximum likelihood values are $\hat{\alpha}_{MLE} = 0.9990936$ and $\hat{\theta}_{MLE} = 2.2821147$.
  - The percentile matching provides a better approximation for the Pareto distribution than did the method of moments.

Parametric Estimation Using 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 n_j to be the number of observations that fall in the jth interval, $(c_{j-1}, c_j]$.
- The probability of an observation X falling in the jth interval is

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

Maximum Likelihood Estimation with Grouped Data

- The probability of an observation X falling in the jth interval is

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

- The corresponding mass function is

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

- The likelihood is

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

- The log-likelihood is

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

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 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 $\mathrm{f}(y)$.
- Summarizing, we have

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

The second right-hand expression allows us to present the likelihood more compactly.

Censored Data Likelihood II

- For a single observation, we have

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

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

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

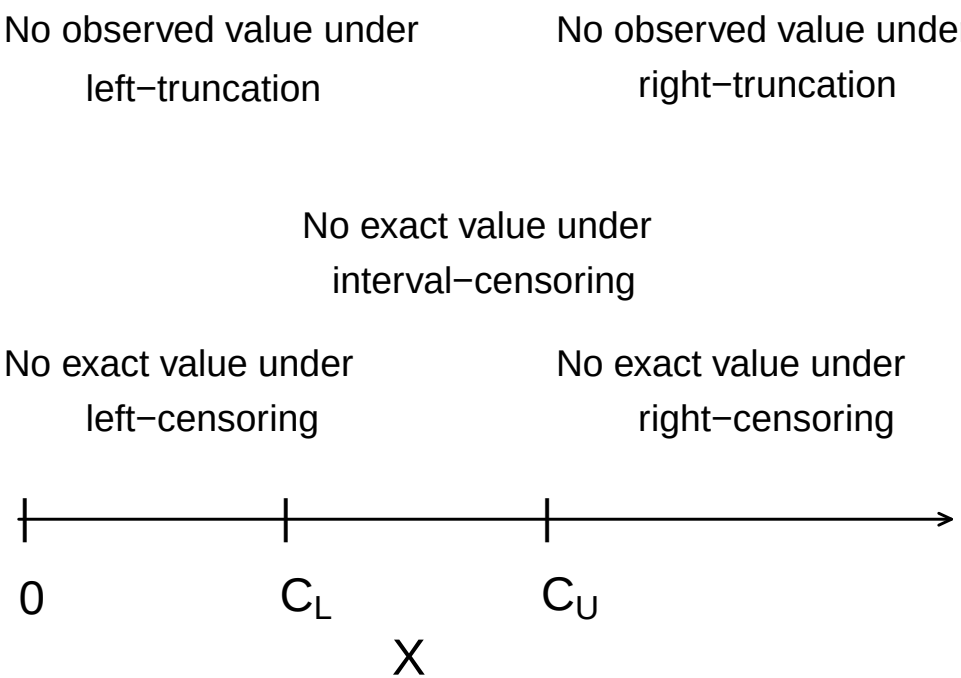
- Here, the notation "$\prod_{\delta_i=0}$" means take the product over uncensored observations, and similarly for "$\prod_{\delta_i=1}$."
- The log-likelihood is

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

Censored and Truncated Data

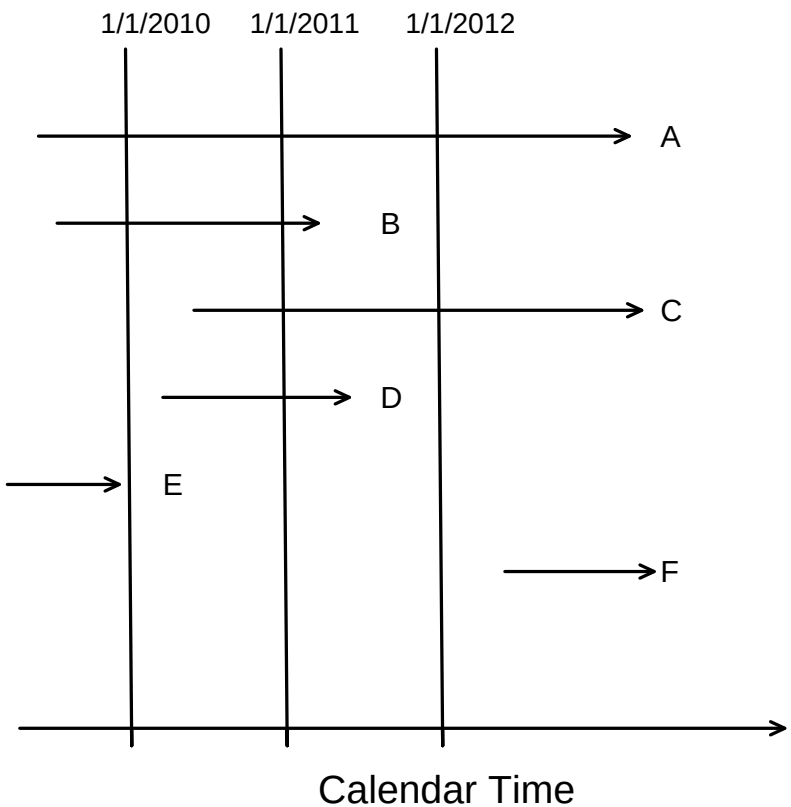
- Let X denote the outcome and let C_L and C_U be two constants.

Type	Limited Variable	Censoring Information
right censoring	$X_U^* = \min(X, C_U)$	$\delta_U = \mathrm{I}(X \geq C_U)$
left censoring	$X_L^* = \max(X, C_L)$	$\delta_L = \mathrm{I}(X \leq C_L)$
interval censoring	$\delta_{LU} = \mathrm{I}(C_L < X \leq C_U)$	
right truncation	X	observe X if $X < C_U$
left truncation	X	observe X if $X > C_L$



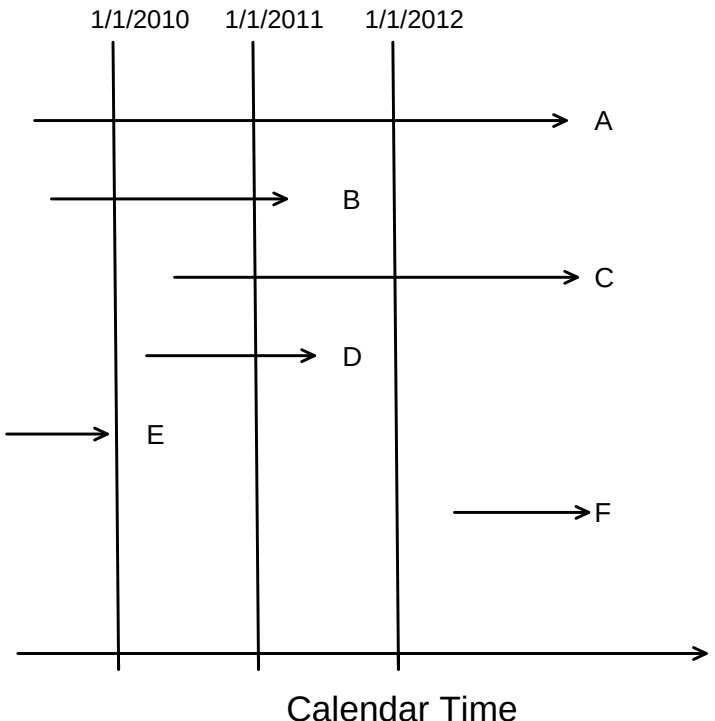
Example: Mortality Study

- Suppose that you are conducting a two-year study of mortality of high-risk subjects, beginning January 1, 2010 and finishing January 1, 2012.
- For each subject, the beginning of the arrow represents that the the subject was recruited and the arrow end represents the event time. Thus, the arrow represents exposure time.



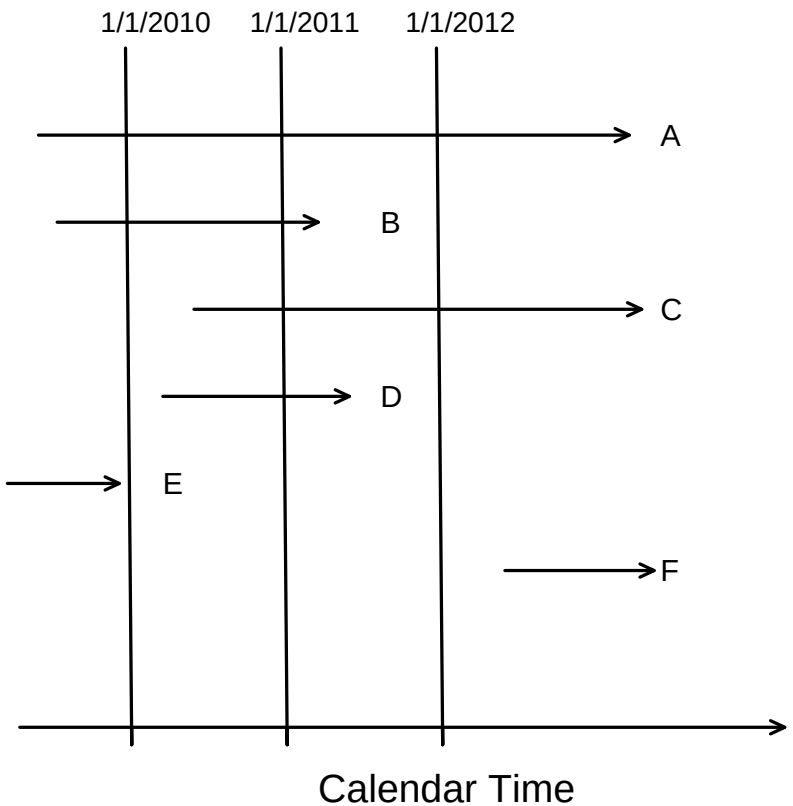
Example: Mortality Study

- Type A - **right-censored**. This subject is alive at the beginning and the end of the study. Because the time of death is not known by the end of the study, it is right-censored. Most subjects are Type A.
- Type B. **Complete information** is available for a type B subject. The subject is alive at the beginning of the study and the death occurs within the observation period.
- Type C - **right-censored** and **left-truncated**. A type C subject is right-censored, in that death occurs after the observation period. However, the subject entered after the start of the study and is said to have a *delayed entry time*. Because the subject would not have been observed had death occurred before entry, it is left-truncated.



Example: Mortality Study

- Type D - **left-truncated**. A type D subject also has delayed entry. Because death occurs within the observation period, this subject is not right censored.
- Type E - **left-truncated**. A type E subject is not included in the study because death occurs prior to the observation period.
- Type F - **right-truncated**. Similarly, a type F subject is not included because the entry time occurs after the observation period.



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 II

- 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 the 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

Special Case: Exponential Distribution

- Consider data that are right-censored and left-truncated, with random variables X_i that are exponentially distributed with mean θ.
- With these specifications, recall that $\mathrm{f}(x) = \theta^{-1}\exp(-x/\theta)$ and $\mathrm{F}(x) = 1 - \exp(-x/\theta)$.
- For this special case, the logarithmic likelihood is

$$\begin{aligned}\ln \mathit{Likelihood} &= \sum_E (\ln \mathrm{f}(x_i) - \ln(1-\mathrm{F}(C_{Li}))) - \sum_R (\ln(1-\mathrm{F}(C_{Ui})) - \ln(1-\mathrm{F}(C_{Li}))\\ &= \sum_E (-\ln\theta - (x_i - C_{Li})/\theta) - \sum_R (C_{Ui} - C_{Li})/\theta.\end{aligned}$$

- To simplify the notation, define $\delta_i = \mathrm{I}(X_i \geq C_{Ui})$ to be a binary variable that indicates right-censoring.
- Let $X_i^{**} = \min(X_i, C_{Ui}) - C_{Li}$ be the amount that the observed variable exceeds the lower truncation limit.
- With this, the logarithmic likelihood is

$$\ln \mathit{Likelihood} = -\sum_{i=1}^{n}\left((1-\delta_i)\ln\theta + \frac{x_i^{**}}{\theta}\right).$$

- Taking derivatives with respect to the parameter θ and setting it equal to zero yields the maximum likelihood estimator

Bayesian Inference

- In the **frequentist interpretation**, one treats the vector of parameters θ 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 θ.
- For notation, we will think about θ as a random vector and let $\pi(\theta)$ denote the distribution of possible outcomes.

Bayesian Inference Strengths

There are several advantages of the Bayesian approach.

1. 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.
2. 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.
3. 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.
4. 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

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

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	$\mathrm{E}(\theta \mid X)$
$\lvert\hat{\theta} - \theta\rvert$	absolute deviation loss	median of $\pi(\theta \mid x)$
$I(\hat{\theta} = \theta)$	zero-one loss (for discrete probabilities)	mode of $\pi(\theta \mid x)$

- 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

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

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

$$\begin{aligned} \pi(\boldsymbol{\theta}|x) &= \frac{f(x|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(x)} \\ &\propto f(x|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \end{aligned}$$

Posterior is proportional to likelihood $\times$ prior

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 $\theta_{new} = 1/(n + 1/\theta)$.