Summarizing Data

MIT 18.443

Dr. Kempthorne

Spring 2015

Outline

Summarizing Data

Overview

- Batches of data: single or multiple

$$x_1, x_2, \ldots, x_n$$
$$y_1, y_2, \ldots, y_m$$
$$w_1, w_2, \ldots, w_l$$

  etc.

- Graphical displays
- Summary statistics:

$$\overline{x} = \frac{1}{n}\sum_1^n x_i, \quad s_x = \sqrt{\frac{1}{n}\sum_1^n (x_i - \overline{x})^2},$$
$$\overline{y},\ s_y,\ \overline{w},\ s_w$$

- Model: $X_1, \ldots, X_n$ independent with $\mu = E[X_i]$ and $\sigma^2 = Var[X_i]$.
  - Confidence intervals for μ (apply CLT)
  - If identical, evaluate GOF of specific distribution family(s)

Summarizing Data

Overview (continued)

- Cumulative Distribution Functions (CDFs)

  Empirical analogs of Theoretical CDFs
- Histograms

  Empirical analog of Theoretical PDFs/PMFs
- Summary Statistics
  - Central Value (mean/average/median)
  - Spread (standard deviation/range/inter-quartile range)
  - Shape (symmetry/skewness/kurtosis)
- Boxplots: graphical display of distribution
- Scatterplots: relationships between variables

Outline

Methods Based on Cumlative Distribution Functions

Empirical CDF

- Batch of data: $x_1, x_2, \ldots, x_n$
  (if data is a random sample, *Batch* $\equiv$ *i.i.d. sample*)
- **Def: Empirical CDF (ECDF)**
  $$F_n(x) = \frac{\#(x_i \leq x)}{n}$$
- Define ECDF Using Ordered batch:
  $$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$
  $$\begin{aligned} F_n(x) &= 0, \quad x < x_{(1)}, \\ &= \frac{k}{n}, \quad x_{(k)} \leq x \leq x_{(k+1)}, \quad k = 1, \ldots, (n-1) \\ &= 1, \quad x > x_{(n)}. \end{aligned}$$
- The ECDF is the CDF of the discrete uniform distribution on the values $\{x_1, x_2, \ldots, x_n\}$. (Values weighted by multiplicity)

Methods Based on CDFs

Empirical CDF

- For each data value x_i define indicator
$$\mathrm{I}_{(-\infty,x_i]}(x) = \begin{cases} 1, & \textit{if} \quad x \le x_i \\ 0, & \textit{if} \quad x > x_i \end{cases}$$
- $F_n(x) = \dfrac{1}{n}\sum_{i=1}^{n} \mathrm{I}_{(-\infty,x_i]}(x)$
- If $Batch \equiv sample$ from distribution with theoretical cdf $F(\cdot)$,
$\mathrm{I}_{(-\infty,x_i]}(x)$ are i.i.d. $Bernoulli(prob = F(x))$,
$nF_n(x) \sim Binomial(size = n, prob = F(x))$.

  Thus:
$$\begin{aligned} E[F_n(x)] &= F(x) \\ Var[F_n(x)] &= \frac{F(x)[1 - F(x)]}{n} \end{aligned}$$
- For samples, $F_n(x)$ is unbiased for $\theta = F(x)$
and maximum variance is at the median.

Estimating $\theta = F(x)$ Using the Empirical CDF

- Confidence Interval Based on Binomial $[nF_n(x)]$

$$\hat{\theta}_n(x) - z(\alpha/2)\hat{\sigma}_{\hat{\theta}} < \theta < \hat{\theta}_n(x) + z(\alpha/2)\hat{\sigma}_{\hat{\theta}}$$

  where

$$\begin{aligned} \hat{\theta}_n(x) &= F_n(x) \\ \hat{\sigma}_{\hat{\theta}} &= \sqrt{\frac{F_n(x)[1-F_n(x)]}{n}} \\ z(\alpha/2) &= \text{upper } \alpha/2 \text{ quantile of } N(0,1) \end{aligned}$$

  Applied to single values of $(x, \theta = F(x))$

- Kolmogorov-Smirnov Test Statistic: If the x_i are a random sample from a continuous distribution with true CDF $F(x)$, then

$$KSstat = \max_{-\infty < x < +\infty} |F_n(x) - F(x)| \sim T_n^*$$

  where the distribution of T_n^* has asymptotic distribution

$$\sqrt{n}T_n^* \xrightarrow{\mathcal{D}} K^*, \text{ the Kolmogorov distribution}$$

  (which does not depend on F!)

Survival Functions

Definition: For a r.v. X with CDF $F(x)$, the **Survival Function** is

$$S(x) = P(X > x) = 1 - F(x)$$

- X: Time until death/failure/"event"
- Empirical Survival Function

$$S_n(x) = 1 - F_n(x)$$

Example 10.2.2.A Survival Analysis of Test Treatements

- 5 Test Groups (I,II,III,IV,V) of 72 animals/group.
- 1 Control Group of 107 animals.
- Animals in each test group received same dosage of tubercle bacilli inoculation.
- Test groups varied from low dosage (I) to high dosage (V).
- Survival lifetimes measured over 2-year period.

Study Objectives/Questions:

- What is effect of increased exposure?

Hazard Function: Mortality Rate

Definition: For a lifetime distribution X with cdf $F(x)$ and surivival function $S(x) = 1 - F(x)$, the **hazard function** $h(x)$ is the instantaneous mortality rate at age x:

$$\begin{aligned} h(x) \times \delta &= P(x < X < x + \delta \mid X > x) \\ &= \frac{f(x)\delta}{P(X > x)} = \frac{\delta f(x)}{1 - F(x)} \\ &= \delta \frac{f(x)}{S(x)} \end{aligned}$$

$$\Longrightarrow \quad h(x) = f(x)/S(x).$$

Alternate Representations of the Hazard Function

$$\begin{aligned} h(x) &= \frac{f(x)}{S(x)} = -\frac{\frac{d}{dx}S(x)}{S(x)} \\ &= -\frac{d}{dx}(\log[S(x)]) = -\frac{d}{dx}(\log[1 - F(x)]) \end{aligned}$$

Hazard Function: Mortality Rate

Special Cases:

- $X \sim \textit{Exponential}(\lambda)$ with cdf $F(x) = 1 - e^{-\lambda x}$
  $h(x) \equiv \lambda$ (constant mortality rate!)
- $X \sim \textit{Rayleigh}(\sigma^2)$ with cdf $F(x) = 1 - e^{-x^2/2\sigma^2}$
  $h(x) = x/\sigma^2$ (mortality rate increases linearly)
- $X \sim \textit{Weibull}(\alpha, \beta)$ with cdf $F(x) = 1 - e^{-(x/\alpha)^\beta}$
  $h(x) = \beta x^{\beta-1}/\alpha^\beta$
  Note: value of β determines whether $h(x)$ is
  increasing ($\beta > 1$), constant ($\beta = 1$), or
  or decreasing ($\beta < 1$).

Hazard Function

Log Survival Functions

- Ordered times: $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$
- For $x = X_{(j)}$, $F_n(x) = j/n$ and $S_n(x) = 1 - j/n$,
- Plot Log Survival Function versus age/lifetime

  $$\log[S_n(x_{(j)})] \text{ versus } x_{(j)}$$

  Note: to handle $j = n$ case apply modifed definition of $S_n()$:

  $$S_n(x_{(j)}) = 1 - j/(n+1)$$
- In the plot of the log survival function, the hazard rate is the negative slope of the plotted function.

  (straight line $\equiv$ constant hazard/mortality rate)

Quantile-Quantile Plots

One-Sample Quantile-Quantile Plots

- X a continuous r.v. with CDF $F(x)$.
- pth quantile of the distribution: x_p:

$$\begin{aligned} F(x_p) &= p \\ x_p &= F^{-1}(p) \end{aligned}$$

- Empirical-Theoretical Quantile-Quantile Plot

  Plot $x_{(j)}$ versus $x_{p_j} = F^{-1}(p_j)$, where $p_j = j/(n+1)$

Two-Sample Empirical-Empirical Quantile-Quantile Plot

Context: two groups

- Control Group: $x_1, \ldots, x_n$ i.i.d. cdf $F(x)$
- Test Treatment: $y_1, \ldots, y_n$ i.i.d. cdf $G(y)$
- Plot order statistics of $\{y_i\}$ versus order statistics of $\{x_i\}$

Testing for No Treatment Effect

H_0: No treatment effect, $G() \equiv F()$

Two-Sample Empirical Quantile-Quantile Plots

Testing for Additive Treatment Effect

- Hypotheses:

  H_0: No treatment effect, $G() \equiv F()$

  H_1: Expected response increases by h units

  $$y_p = x_p + h$$

- Relationship between $G()$ and $F()$

  $$G(y_p) = F(x_p) = p$$
  $$\Longrightarrow G(y_p) = F(y_p - h).$$

  - CDF $G()$ is same as $F()$ but shifted h units to right
  - Q-Q Plot is linear with slope $= 1$ and intercept $= h$.

Two-Sample Empirical Quantile-Quantile Plots

Testing for Multiplicative Treatment Effect

- Hypotheses:

  H_0: No treatment effect, $G() \equiv F()$

  H_1: Expected response increases by factor of c (> 0)

  $$y_p = c \times x_p$$

- Relationship between $G()$ and $F()$

  $$G(y_p) = F(x_p) = p$$
  $$\Longrightarrow G(y_p) = F(y_p/c).$$

  - CDF $G()$ is same as $F()$ when plotted on log horizontal scale (shifted $\log(c)$ units to right on log scale).

  - Q-Q Plot is linear with slope $= c$ and intercept $= 0$.

Outline

1. Summarizing Data
   - Overview
   - Methods Based on CDFs
   - Histograms, Density Curves, Stem-and-Leaf Plots
   - Measures of Location and Dispersion

Histograms, Density Curves, Stem-and-Leaf Plots

Methods For Displaying Distributions (relevant R functions)

- Histogram: $hist(x, nclass = 20, probability = TRUE)$
- Density Curve: $plot(density(x, bw = "sj"))$
  - Kernel function: $w(x)$ with bandwidth h
    $$w_h(x) = \frac{1}{h} w\left(\frac{x}{h}\right)$$
    $$f_h(x) = \frac{1}{n} \sum_{i=1}^{n} w_h(x - X_i)$$
  - Options for Kernel function:
    Gaussian: $w(x) = Normal(0, 1)$ pdf
    $$\Longrightarrow w_h(x - X_i) \text{ is } N(X_i, h) \text{ density}$$
    Rectangular, triangular, cosine, bi-weight, etc.
  - See: Scott, D. W. (1992) Multivariate Density Estimation. Theory, Practice and Visualization. New York: Wiley
- Stem-and-Leaf Plot: $stem(x)$
- Boxplots: $boxplot(x)$

Displaying Distributions

Boxplot Construction

- Vertical axis = scale of sample $X_1, \ldots, X_n$
- Horizontal lines drawn at *upper-quartile*, *lower − quartile* and vertical lines join the box
- Horizontal line drawn at median inside the box
- Vertical line drawn up from *upper-quartile* to most extreme data point

  within 1.5 (IQR) of the upper quartile.

  (IQR=Inter-Quartile Range)

  Also, vertical line drawn down from *lower-quartile* to most extreme data point

  within 1.5 (IQR) of the *lower-quartile*

  Short horizontal lines added to mark ends of vertical lines
- Each data point beyond the ends of vertical lines is marked with $*$ or .

Outline

Measures of Location and Dispersion

Location Measures

- Objective: measure *center* of $\{x_1, x_2, \dots, x_n\}$
- **Arithmetic Mean**
  $\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$
  Issues: not robust to outliers.
- **Median**

$$median(\{x_i\}) = \begin{cases} x_{[j^*]}, & j^* = \frac{n+1}{2} \quad if \quad n \quad is \quad odd \\ \frac{x^*_{[j]} + x_{[j^*+1]}}{2}, & j^* = \frac{n}{2} \quad if \quad n \quad is \quad even \end{cases}$$

  Note: confidence intervals for $\eta = median(X)$ of form
  $[x_{[k]}, x_{[n+1-k]}]$
  Rice Section 10.4.2 applies binomial distribution for
  $\#(X_i > \eta)$

Measures of Location and Dispersion

Location Measures

- **Trimmed Mean**: $x.trimmedmean = mean(x, trim = 0.10)$
  - 10% of of lowest values dropped
  - 10% of highest values dropped
  - mean of remaning values computed

  ($trim =$ parameter must be less than 0.5)
- **M Estimates**
  - **Least-squares estimate**: Choose $\hat{\mu}$ to minimize
    $$\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2$$
  - **MAD estimate**: Choose $\hat{\mu}$ to minimize
    $$\sum_{i=1}^{n} \left| \frac{X_i - \mu}{\sigma} \right|$$
    (minimize mean absolute deviation)

Measures of Location and Dispersion

- **M Estimates** (continued)
  - **Huber Estimate**: Choose $\hat{m}u$ to minimize

$$\sum_{i=1}^{n} \Psi\left(\frac{X_i - \mu}{\sigma}\right)$$

where $\Psi()$ is sum-of-squares near 0 (within $k\ \sigma$) and sum-of-absolutes far from 0 (more than $k\ \sigma$)

In R:

$$library(mass)$$
$$x.mestimate = huber(x, k = 1.5)$$

Comparing Location Estimates

- For symmetric distribution: same *location* parameter for all methods!

  Apply Bootstrap to estimate variability
- For asymmetric distribution: *location* parameter varies

Measures of Dispersion

- **Sample Standard Deviation**: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$
  - s^2 unbiased for $Var(X)$ if $\{X_i\}$ are i.i.d.
  - s biased for $\sqrt{Var(X)}$.
  - For X_i i.i.d. $N(\mu, \sigma^2$: $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$.
- **Interquartile Range** : **IQR** $= q_{0.75} - q_{0.25}$
  where q_p is the p-th quantile of F_X
  - For X_i i.i.d. $N(\mu, \sigma^2)$, $IQR = 1.35 \times \sigma$
  - For Normal Sample: $\tilde{\sigma} = \frac{sample\ IQR}{1.35}$
- **Mean Absolute Deviation**: **MAD** $= \frac{1}{n}\sum_{i=1}^{n}|X_i - \overline{X}|$
  - For X_i i.i.d. $N(\mu, \sigma^2)$, $E[MAD] = 0.675 \times \sigma$
  - For Normal Sample: $\tilde{\sigma} = \frac{MAD}{0.675}$

18.443 Statistics for Applications
Spring 2015