Data analysis

Ben Graham

MA930, University of Warwick

October 15, 2015

Statistics

Def 5.2.1

- Let $X_1, \ldots, X_n \sim f(x \mid \theta)$ independent
- $f(x_1, \ldots x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$
- Let $T(x_1, \ldots, x_n)$ denote some function of the data, i.e. $T : \mathbb{R}^n \to \mathbb{R}$.
- T is a statistic
  - $T(x_1, \ldots, x_n) = x_1$
  - $T(x_1, \ldots, x_n) = \max_i x_i$
  - mean, median, etc.
- θ,$\mathbb{E}X$, $\mathrm{Var}(X)$, etc, are not statistics.
- The probability distribution of T is called the sampling distribution of T.

Common statistics

- $X_1, \ldots, X_n$ i.i.d.r.v.
- Sample mean

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n} = \sum_{i=1}^{n} X_i$$

- Sample variance

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 = \frac{1}{n-1} \left[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2 \right]$$

- If the mean $\mathbb{E}X_i$ and variance $\mathrm{Var}(X_i)$ exist, then these are *unbiased* estimates.

Trimmed mean

- Trimmed/truncated mean
  - $p \in [0, 1]$
  - Remove the smallest and biggest np items from the sample
  - Take the mean of what is left
  - i.e. LIBOR, 18 banks, top and bottom 4 removed
  - Cauchy location parameter

Def 5.4.1 Order Statistics

- The order statistics of a random sample $X_1, \ldots, X_n$ are the values placed into increasing order: $X_{(1)}, \ldots, X_{(n)}$.
  - $X_{(1)} = \min_i X_i$
  - $X_{(2)}$ =second smallest
  - ...
  - $X_{(n)} = \max_i X_i$
- The sample range is $R = X_{(n)} - X_{(1)}$
- The sample median is

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

The median may give a better sense of what is typical than the mean.

Quantiles

- For $p \in [0,1]$, the p quantile is (R, method 7 of 9)

  $$(1-\gamma)x_{(j)}+\gamma x_{(j+1)}, \quad (n-1)p < j \leq (n-1)p+1, \quad \gamma = np+1-p-j$$

- Quantiles are a continuous function of the data

```
data=c(1,3,7)
q=seq(0,1,length.out = 1000)
plot(quantile(data,q))
```

- Quartiles: $i/4$ quantiles
- Deciles: $i/10$ quantiles
- Percentiles: $i/100$ quantiles

Theorem 5.4.4 Distribution of the order statistics

- Sample of size n with c.d.f. F_X and pdf f_X
- Binomial distribution:

$$F_{X_{(j)}}(x) = \sum_{k=j}^{n} \binom{n}{k} [F_X(x)]^k [1 - F_X(x)]^{n-k}$$

- Differentiate:

$$f_{X_{(j)}}(x) = \frac{n!}{(j-1)!(n-j)!} f_X(x) [F_X(x)]^{j-1} [1 - F_X(x)]^{n-j}.$$

QQ-plots

- $X_{(1)} \leq \cdots \leq X_{(n)}$ order statistics
- suspected c.d.f. F
- The j-th order statistic is expected to be $\approx F^{-1}(\frac{2j-1}{2n})$
- Plot $X_{(1)} \leq \cdots \leq X_{(n)}$ against $F^{-1}(\frac{1}{2n}), F^{-1}(\frac{3}{2n}) \ldots, F^{-1}(\frac{2n-1}{2n})$
- fit line $y = x \rightarrow F$ is the right c.d.f
  N.B. expect some noise
  plot(apply(replicate(10000,sort(runif(11))),1,sd))
  plot(apply(replicate(10000,sort(rnorm(11))),1,sd))
- fit a line $\rightarrow F$ is correct c.d.f. for some $aX + b$, $a, b \in \mathbb{R}$.
- not a line $\rightarrow F$ is not really the c.d.f.

More sample mean

- Characteristic function

$$\phi_{\bar{X}}(t) = [\phi_X(t/n)]^n \approx \left[1 + \frac{it\mathbb{E}X}{n} + o(t/n)\right]^n$$

- Thm 5.3.1: If $X_i \sim N(\mu, \sigma^2)$ are independent then
  - $\bar{X}$ and S^2 are independent
  - $\bar{X} \sim N(\mu, \sigma^2/n)$.
  - $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$
- Proof: For simplicity, assume $\mu = 0$ and $\sigma^2 = 1$.

Ingredients for the proof

- If $e_1, \ldots, e_n$ form an orthonormal basis for $\mathbb{R}^n$. Then $(e_i \cdot X)_{i=1}^n$ are i.i.d.r.v $N(0,1)$.
- Let $e_1 = (\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}})$
- The $\Gamma(\alpha, \beta)$ distribution is defined

$$f(x) = C(\alpha, \beta) x^{\alpha - 1} \exp(-\beta x).$$

- The χ^2_k distribution is the $\Gamma(k/2, 1/2)$ distribution.
- If $X \sim N(0,1)$, then $X^2 \sim \chi^2_1$
- The sum of k independent χ^2_1 r.v. is χ^2_k.

Derived distributions

- If $Z \sim N(0,1)$ and $V \sim \chi^2_k$ then $Z/\sqrt{V/k} \sim t_k$ (Student's t)
- $$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$
- If $U \sim \chi^2_k$ and $V \sim \chi^2_\ell$ are independent, then $\frac{U/k}{V/\ell} \sim F_{k,\ell}$
- Suppose $X_1, \ldots, X_m \sim N(\mu_X, \sigma^2_X)$ and $Y_1, \ldots, Y_n \sim N(\mu_Y, \sigma^2_Y)$. Then

$$\frac{S^2_X/\sigma^2_X}{S^2_Y/\sigma^2_Y} \sim F_{m-1,n-1}$$

- These distributions are also used for linear regression/ANOVA.

Convergence in probability

- Def 5.5.1: A sequence of random variables $X_1, X_2, \ldots$ converge in probability to a random variable X if for every $\epsilon > 0$,

$$\lim_{n\to\infty} P(|X_n - X| \geq \epsilon) = 0.$$

- Thm 5.5.2 (Weak Law of Large Numbers) Let $X_1, \ldots, X_n$ be i.i.d.r.v. with mean μ and variance $\sigma^2 < \infty$ (or $\mathbb{E}|X| < \infty$).
  Define $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.
  $\bar{X}_n$converges in probability to μ.
  Proof: Characteristic functions:

$$\phi_{\bar{X}_n}(t) = [\phi_X(t/n)]^n \approx \left[1 + \frac{it\mathbb{E}X}{n} + o(t/n)\right]^n$$

Ex 5.5.3 Consistency of S^2

Sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2\right] - \frac{n}{n-1}\bar{X}_n^2$$

Converges to σ^2 in probability if $\mathbb{E}|X^2| < \infty$.

Def 5.5.6 Almost sure convergence

- A sequence of random variables $X_1, \ldots, X_n$ converges almost surely to a random variable X if, for all $\epsilon > 0$,

$$\mathbb{P}\left(\lim_{n\to\infty} |X_n - X| < \epsilon\right) = 1.$$

- N.B. the limit is now on the inside.
- $X_n \sim Bernoulli(1/n)$ independent: converge in probability to 0, not almost surely.
- Theorem 5.5.9 Strong Law of Large Numbers:
  Let X_1, X_2 be iidrv with mean μ and variance $\sigma^2 < \infty$
  (or even better: if $\mathbb{E}|X| < \infty$).
  Define $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.
  Then $\bar{X}_n$ converges almost surely to μ.

6.2 Sufficient statistics

- Def 6.2.1 A statistic $T(X)$ is a sufficient statistic for θ if the conditional distribution of the sample X given the value of $T(X)$ does not depend on θ.
- Thm 6.2.2 If $f(x \mid \theta)/f(T(x) \mid \theta)$ is constant, then $T(X)$ is sufficient.
- Thm 6.2.6 $T(X)$ is sufficient iff $f(x \mid \theta) = g(T(x) \mid \theta)h(x)$ for some g, h
- Example: Independent $X_i \sim$Bernoulli(θ), $\theta \in (0, 1)$.
- Example: Independent $X_1, \ldots, X_N \sim$Uniform$(0, \theta)$, $\theta > 0$.
- Example: Independent $X_1, \ldots, X_n \sim N(\theta, \sigma^2)$, $\theta \in \mathbb{R}$
- Example: Independent $X_1, \ldots, X_n \sim N(\theta_1, \theta_2^2)$, $\theta_1 \in \mathbb{R}, \theta_2 > 0$
- Minimal sufficient statistics

6.3 Likelihood principle

- Random sample $X = (X_1, \ldots, X_n)$
- $X_i \sim f(x_i \mid \theta)$ pmf or pdf
- $X \sim \prod_i f(x_i \mid \theta) = f(x \mid \theta)$
- Likelihood function

$$L(\theta \mid x) = f(x \mid \theta)$$

- Likelihood principle: if

$$L(\theta \mid x)/L(\theta \mid y)$$

is independent of θ, then the conclusions drawn from x and y should be identical.

Chapter 7 Point estimation

7.2.2 Maximum Likelihood Estimator

- $L(\theta \mid x) = \prod_i f(x_i \mid \theta)$, $\theta \in \mathbb{R}^k$
- MLE: Statistic $\hat{\theta}(x) = \arg\max_\theta L(\theta \mid x)$