Probabilistic Modeling and Statistical Computing Fall 2015

October 19, 2015

Sample Statistics

Given a sample $(x_1, x_2, \ldots, x_n)$ of size n (that is, n independent observations of the same random variable X), we can compute many functions ("statistics") from it:

- Mean
- Median
- Standard deviation
- Maximum and minimum
- Quantiles

What can be said about these random quantities?

Example: Maximum

Given a sample $(x_1, x_2, \ldots, x_n)$ of size n from an exponential distribution with $\lambda = 1$. Let
$M = \max_i x_i$.
Expect $M \to \infty$ as n becomes large.
Is that all that can be said?
How fast does $M \to \infty$?

How about if we use a Cauchy distribution?

Example: Law of Large Numbers for Means

Given a sample $(x_1, x_2, \ldots, x_n)$ of size n from a distribution for which $\mathcal{E}(X)$ exists. Let $\bar{x} = \frac{1}{n}\sum_i x_i$ be the sample mean.

Since $\bar{x}$ depends on all n observations, its distribution has a very complicated formula involving the joint pdf / pmf of all n observations.

LLN for Means

Amazing Simplification

As $n \to \infty$, $\bar{x} \to \mathcal{E}(X)$ with $\mathcal{P} = 1$.

The limit isn't random at all, it's a constant.

This is independent of the distribution of X.

Example: Law of Large Numbers for Medians

Given a sample $(x_1, x_2, \ldots, x_n)$ of size n from a **continuous** distribution with median μ, i.e. $F_X(\mu) = \frac{1}{2}$. Let m be the sample median.

LLN for Medians

Another Amazing Simplification:

As $n \to \infty$, $m \to \mu$ with $\mathcal{P} = 1$.

The limit isn't random at all, it's a constant.

This is independent of the distribution of X. We don't need to know anything about X (except that it is continuous).

Convergence Speed of $\bar{x} \to \mathcal{E}(X)$

Let $\bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i$ be the mean of the first n observations. Then $\bar{x}_n - \mathcal{E}(X) \to 0$.

Multiply this with n^α and check whether $n^\alpha(\bar{x}_n - \mathcal{E}(X))$ still converges to 0 or whether something else happens.

Central Limit Theorem for Means

Let $\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$ be the mean of the first n observations. Assume that $var(X) = \sigma^2$ exists. Then

$$\boxed{\sqrt{n}\,(\bar{x}_n - \mathcal{E}(X)) \sim N(0, \sigma^2)}$$

for large n.

Equivalently: Let $S_n = \sum_{i=1}^{n} x_i$. Then

$$\boxed{\frac{S_n - n\mathcal{E}(X)}{\sqrt{n}} \sim N(0, \sigma^2)}$$

Another amazing result: $\bar{x}_n = \mathcal{E}(X) + O(n^{-1/2})$, independent of the distribution of X.

CLT for Means

The sample size at which $\bar{x}$ is approximately normally distributed depends on the shape of the distribution of X.

- If the distribution of X is very skewed, it can take a large n, e.g. $n \approx 100$.
- If the distribution of X is symmetric, a smaller n is enough, e.g. $n \approx 10$.
- If X is normally distributed, then $\bar{x}$ is also normally distributed for any n.

Simulations

Central Limit Theorem for Medians

Let m_n be the median of the first n observations, coming from a continuous distribution with pdf f_X and true median μ. Also assume that $f_X(\mu) > 0$ and that f_X' exists. Then

$$\boxed{\sqrt{n}\,(m_n - \mu) \sim N\left(0, \frac{1}{4f_X(\mu)^2}\right)}$$

for large n.

Therefore $m_n = \mu + O(n^{-1/2})$, independent of the distribution of X.

Central Limit Theorem for Quantiles

Let $m_{\alpha,n}$ be the αth quantile of the first n observations, coming from a continuous distribution with pdf f_X and true quantile μ_α. Also assume that $f_X(\mu_\alpha) > 0$ and that f'_X exists. Then

$$\sqrt{n}\,(m_{\alpha,n} - \mu_\alpha) \sim N\left(0, \frac{\alpha(1-\alpha)}{f_X(\mu_\alpha)^2}\right)$$

for large n.
Therefore $m_{\alpha,n} = \mu_\alpha + O(n^{-1/2})$, independent of the distribution of X.

Central Limit Theorem for Standard Deviation?

Explore this with simulations.

Limit Theorems for Extrema

Survival function

$$S(x) = \mathcal{P}(X > x) = 1 - F(x)$$

where F is the cdf.

Example: Cauchy distribution:

$$F(x) = \frac{\arctan x}{\pi} + \frac{1}{2}, \; S(x) = \frac{1}{2} - \frac{\arctan x}{\pi}$$

Example: Exponential distribution:

$$F(x) = 1 - e^{-\lambda x}, \; S(x) = e^{-\lambda x}$$

Fréchet Limit Theorem for Extrema

Assume the survival function satisfies

$$\lim_{x\to\infty} S(x)\cdot x^{\alpha} = C$$

for some $\alpha > 0$, $C > 0$.
Cauchy distribution: True for $\alpha = 1$, $C = \frac{1}{\pi}$.
Let $x_1, x_2, \ldots$ be a sample. Then for large n

$$\mathcal{P}(\max_{i\leq n} X_i \leq x) \approx \exp(-nCx^{-\alpha})$$

or

$$\mathcal{P}\left(\frac{\max_{i\leq n} X_i}{(nC)^{1/\alpha}} \leq y\right) \approx \exp(-y^{-\alpha})$$