Three Important Distributions

Bruce A. Desmarais

Stats I Lab

November 16, 2009

Three Important Distributions

Three Important Distributions

1. $f_Y(y|\theta)$: The distribution of the raw data. This is the (possibly joint ...multivariate) distribution of the data in the rows and columns of the dataset collected by the analyst. The objective of statistical inference is to find $f_Y(y|\theta)$ or properties of it.

Three Important Distributions

1. $f_Y(y|\theta)$: The distribution of the raw data. This is the (possibly joint ...multivariate) distribution of the data in the rows and columns of the dataset collected by the analyst. The objective of statistical inference is to find $f_Y(y|\theta)$ or properties of it.
2. $f_T(T(Y)|\theta, N)$: The *finite sample* distribution of some statistic $(T(\cdot))$, which is a function of the data in a sample of size N from $f_Y(y|\theta)$.

Three Important Distributions

1. $f_Y(y|\theta)$: The distribution of the raw data. This is the (possibly joint ...multivariate) distribution of the data in the rows and columns of the dataset collected by the analyst. The objective of statistical inference is to find $f_Y(y|\theta)$ or properties of it.
2. $f_T(T(Y)|\theta, N)$: The *finite sample* distribution of some statistic ($T(\cdot)$), which is a function of the data in a sample of size N from $f_Y(y|\theta)$.
3. $f_T(T(Y)|\theta, \infty) = \lim_{N\to\infty}(f_T(T(Y)|\theta, N))$: The limiting distribution of $T(Y)$. It is often difficult to derive $f_T(T(Y)|\theta, N)$ for finite N. The Central Limit Theorem often eases the derivation of the limiting distribution. In *large* samples, the limiting distribution can be used instead of the sampling distribution.

Points on $f_Y(y|\theta)$

Points on $f_Y(y|\theta)$

- The distribution of the data is *invariant* to N. No matter how big N, data that is non-normal at $N = 10$ will always be non-normal.

Points on $f_Y(y|\theta)$

- The distribution of the data is *invariant* to N. No matter how big N, data that is non-normal at $N = 10$ will always be non-normal.
- You will only know with certainty the distribution of your data if $N = \infty$, so you must be prepared to justify assumptions about the form of $f_Y(\cdot)$.

Points on $f_Y(y|\theta)$

- The distribution of the data is *invariant* to N. No matter how big N, data that is non-normal at $N = 10$ will always be non-normal.
- You will only know with certainty the distribution of your data if $N = \infty$, so you must be prepared to justify assumptions about the form of $f_Y(\cdot)$.
- In regression analysis you make assumptions about the *conditional* distribution of Y given X, not the *marginal* distribution of Y.

Points on $f_T(T(Y)|\theta, N)$

Points on $f_T(T(Y)|\theta, N)$

- Mostly this is only available if a fully parametric assumption about $f_Y(\cdot)$ is made.

Points on $f_T(T(Y)|\theta, N)$

- Mostly this is only available if a fully parametric assumption about $f_Y(\cdot)$ is made.
- Can generally only be derived if $T(\cdot)$ is arithmetically simple.

Points on $f_T(T(Y)|\theta, N)$

- Mostly this is only available if a fully parametric assumption about $f_Y(\cdot)$ is made.
- Can generally only be derived if $T(\cdot)$ is arithmetically simple.
- In the case that the analyst is comfortable making parametric assumptions and there exists a formula for $f_T(T(Y)|\theta, N)$, the finite sample distribution should be used for inference tasks. WARNING: This is a choice that is often not made by software. Since asymptotic arguments apply more generally, defaults are often set to them.

Points on $f_T(T(Y)|\theta, \infty)$

Points on $f_T(T(Y)|\theta, \infty)$

- Only exact, but is a good approximation if N is large.

Points on $f_T(T(Y)|\theta, \infty)$

- Only exact, but is a good approximation if N is large.
- Is usually a normal distribution, derived with aid of the central limit theorem

Points on $f_T(T(Y)|\theta, \infty)$

- Only exact, but is a good approximation if N is large.
- Is usually a normal distribution, derived with aid of the central limit theorem
- Typically one result applies to a class, or family of parametric conditions...can be more *Robust*

The CLT
If $S_n = X_1 + \cdots X_n$

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} \ c.d. \ N(0,1)$$

Examples