S1211Q Introduction to Statistics
Lecture 16

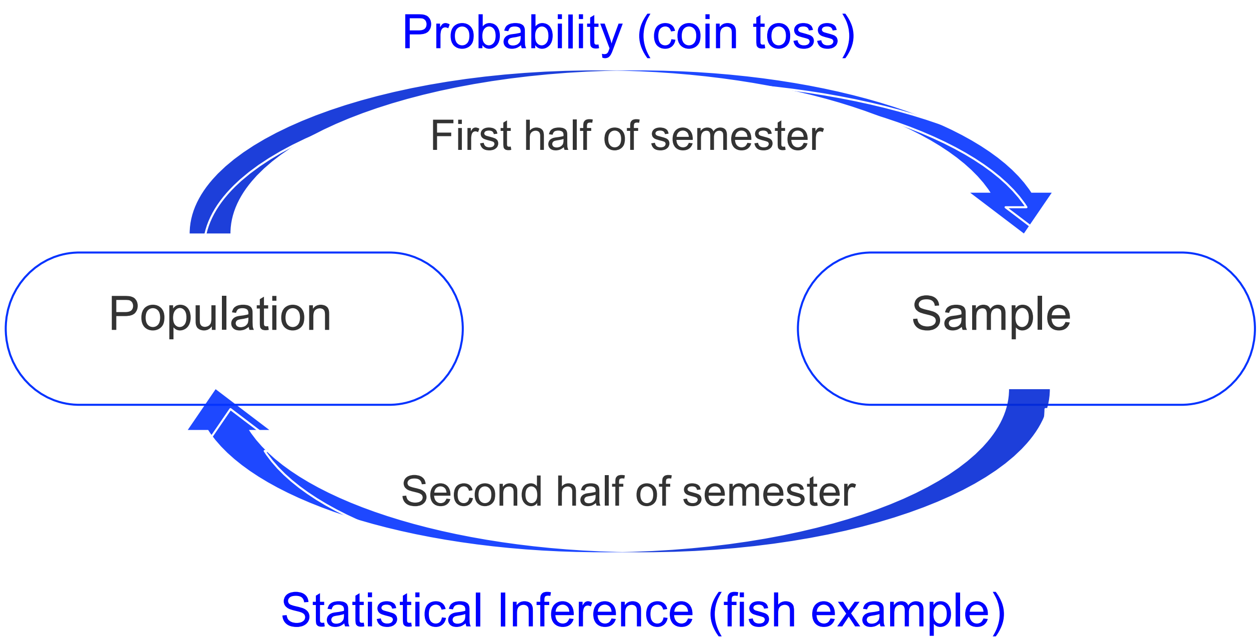
Wei Wang

Oct 29th, 2012

Population and Sample

- We will start changing our discussion from probability to statistics, which means we need to think about samples and how they relate to the underlying population.
- Recall the relationship between population and sample (probability and inference) that we visualized in the first lecture.

Probability and Inference



RV or a Particular Number

- In the first chapter, we use lowercase letters to represent the sample, $x_1, x_2, x_3, \ldots$. That means we have already observed the data and each of the letters can be replaced by a particular number.
- Before the data becoming available, there is uncertainty as to what value we will observe, so we view each observation as a RV, thus denoted by uppercase letter $X_1, X_2, X_3, \ldots$

Sample and Statistics

- A statistic is any quantity whose value can be calculated from sample data, such as Sample Mean and Sample Variance.
- Before obtaining data, a statistic is also a RV. The bulk of statistical inference is to find the distribution of the statistics, or the so-called *Sampling Distributions*.
- To make things easier, we often need to assume the observed data are *Simple Random Samples*, which means they are IID (Independently Identically Distributed).

Introduction to IID

- A sequence of random variables, X_1, X_2, …, X_n, is independent and identically distributed (i.i.d.) if each random variable has the same probability distribution as the others and all are mutually independent.

- In statistical analysis, we often assume the sampled data X_1, X_2, …, X_n, are i.i.d. from a common distribution $f(x)$. And usually, we end up analyzing a linear combination of the X_i's, that is

$$\mathrm{Y} = a_1\mathrm{X}_1 + \cdots + a_n\mathrm{X}_n = \sum_{i=1}^{n} a_i\mathrm{X}_i$$

Deriving a Sampling Distribution

- Probability rules can be used to obtain the distribution of a statistic that is a fairly simple function of X_i's and the distribution of X_i's are also of simple form. See Example 5.20 on page 215.
- However, this brute force method only works for a limited class of statistics. For most statistics, it is practically impossible to get the distribution (pmf/pdf).
- But for some of the most common statistics, such as Sample Mean, we have some useful results.

Sample Mean

- Let $X_1, X_2, \ldots, X_n$ be an IID sequence of random variables from a distribution with mean μ and variance σ^2. Sample Mean is $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$.
- Then we can derive the expectation and variance of sample mean $\bar{X}$

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

$$Var(\bar{X}) = Var(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Invariance of Normal RV under Linear Transformation

- When $X_1, X_2, X_3, X_4, \ldots$ are normal random variables, then the linear combination of them

$$a_1 X_1 + a_2 X_2 + \ldots + a_n X_n = \sum_{i=1}^{n} a_i X_i$$

  is still a normal random variable.
- In particular, sample mean $\bar{X}$ is still a random variables.
- Remark: 1. No IID assumption is necessary; 2. This property is for Normal only.

Sample Mean of IID Normal

- If $X_1, X_2, \ldots, X_n$ IID $\sim N(\mu, \sigma^2)$, then what is the distribution of $\bar{X}$?

Sample Mean of IID Normal

- If $X_1, X_2, \ldots, X_n$ IID $\sim N(\mu, \sigma^2)$, then what is the distribution of $\bar{X}$?
- $$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

Sample Mean of IID Normal

- If $X_1, X_2, \ldots, X_n$ IID $\sim N(\mu, \sigma^2)$, then what is the distribution of $\bar{X}$?
- $$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$
- But how can we derive the distribution of Sample Mean when the sample are not normal? We need to use Central Limit Theorem.

CLT

- Theorem:

The Central Limit Theorem (CLT)

Let X_1, X_2, …, X_n, be an i.i.d. sequence from a distribution with mean μ and variance σ^2. Then if n is sufficiently large, the sample mean $\bar{X}$ has approximately a normal distribution with $\mu_{\bar{X}} = \mu$ and $\sigma^2_{\bar{X}} = \sigma^2/n$;
And the sample total has approximately a normal distribution with $\mu_T = n\mu$, $\sigma_T^2 = n\sigma^2$. The larger the value of n, the better the approximation.

- Rule of Thumb: if $n>30$, the CLT can be used.

Distribution of a Linear Combination

- Sample mean is a particular case of linear combinations.
- The expectation and variance of a general linear combination

  $$a_1X_1 + a_2X_2 + \ldots + a_nX_n$$

  is given by the following result.

A key result ***

Let $X_1, X_2, \ldots, X_n$, have mean values $\mu_1, \mu_2, \ldots, \mu_n$, respectively, and variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$, respectively.

- Whether or not the Xi's are independent,

$$\begin{aligned} \mathrm{E}(a_1 \mathrm{X}_1 + a_2 \mathrm{X}_2 + \cdots + a_n \mathrm{X}_n) &= a_1 \mathrm{E}(\mathrm{X}_1) + a_2 \mathrm{E}(\mathrm{X}_2) + \cdots + a_n \mathrm{E}(\mathrm{X}_n) \\ &= a_1 \mu_1 + a_2 \mu_2 + \cdots + a_n \mu_n \end{aligned}$$

- For any $X_1, X_2, \ldots, X_n$,

$$\mathrm{Var}(a_1 \mathrm{X}_1 + a_2 \mathrm{X}_2 + \cdots + a_n \mathrm{X}_n) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \mathrm{Cov}(\mathrm{X}_i, \mathrm{X}_j)$$

  If they are independent, then

$$\begin{aligned} & \mathrm{Var}(a_1 \mathrm{X}_1 + a_2 \mathrm{X}_2 + \cdots + a_n \mathrm{X}_n) \\ = \; & a_1^2 \mathrm{Var}(\mathrm{X}_1) + a_2^2 \mathrm{Var}(\mathrm{X}_2) + \cdots + a_n^2 \mathrm{Var}(\mathrm{X}_n) \\ = \; & a_1^2 \sigma_1^2 + a_2^2 \sigma_2^2 + \cdots + a_n^2 \sigma_n^2 \end{aligned}$$

Special Cases

- $E(X+Y) = E(X) + E(Y)$;
- $E(X-Y) = E(X) - E(Y)$;
- $Var(X+Y) = Var(X) + Var(Y) + 2Cov(X, Y)$
- $Var(X-Y) = Var(X) + Var(Y) - 2Cov(X, Y)$
- If X and Y are independent, then $Cov(X, Y) = 0$, and
  $Var(X+Y) = Var(X) + Var(Y)$
  $Var(X - Y) = Var(X) + Var(Y)$

Example

Ex. Show that if X ~ Bin(n, p), then E(X) = np, and Var(X) = $np(1 - p)$.

Ex. Show that if X is a negative binomial rv with pmf $nb(x; r, p)$, then E(X) = $r(1–p)/p$,
Var(X) = $r(1 - p)/p^2$.