Lecture 23

We have been spending time talking about joint distributions for the last several lectures. Building on that work will lead us to sampling distributions today and finally to formal statistical inference in a few weeks/lectures time.

Recall from early on we used

$x_1, x_2, \ldots, x_n$ to represent observations of sample data (eg formulas for mean and standard deviation). In practise the exact values of $x_1, x_2, \ldots, x_n$ are going to depend on the exact sample we get. To account for this variation, before data becomes available we should denote the sample as r.v.s

$X_1, X_2, X_3, \ldots, X_n$

Note this implies that quantities such as the sample mean $\bar{x}$ and sample standard deviation s are

also r.v.

More specifically these are examples of statistics.

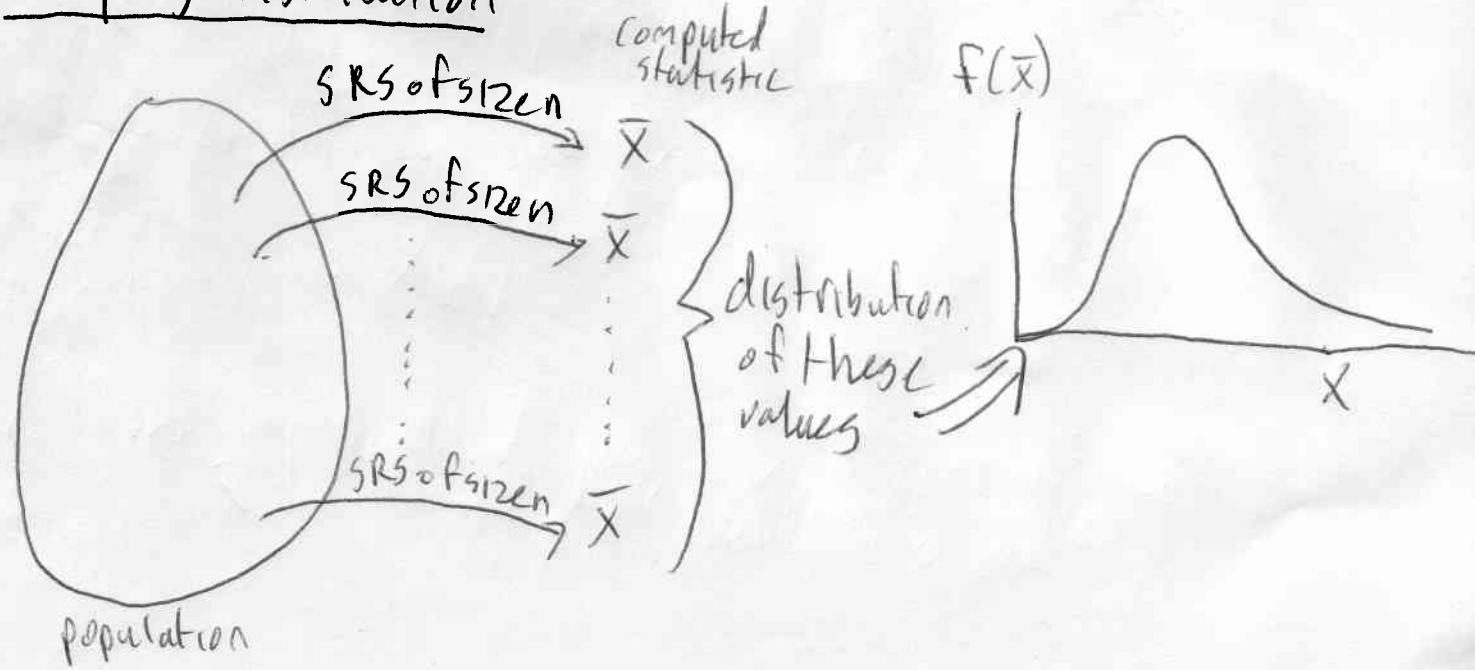
Statistic: a quantity whose value may be computed using sample data.

Statistics are r.v prior to obtaining the data.

what we would like is to know something about the probability distribution of a statistic.

More specifically we refer to this as the sampling distribution.

Sampling Distribution



More specifically the sampling distribution depends on

1. the population distribution
2. the sample size n
3. the sampling method

The sampling method we discuss here is called Simple Random Sampling (SRS).

In an SRS every member of the population has an equally likely chance of being selected for the sample.

For the purposes of discussing sampling distributions we will assume that $X_1, \ldots X_n$ are a SRS of size n where

1. the X_i's are all independent
2. Each X_i comes from the same distribution.

We use the term "independent and identically distributed" or the abbreviation "iid"

How do we derive a sampling distribution?

Two ways

1. Theoretically based on probability rules
2. Empirically based on simulations.

An example of by theory

Suppose that X_1 represents the number of traffic lights that the bus from BART to SFSU stops at on my morning ride to campus and X_2 represents the number of lights stopped at on the way from SFSU back to Bart in the afternoon. Suppose that X_1 and X_2 are independent with same distribution and that there are 3 traffic lights between SFSU and BART.

x_1	0	1	2	3
$p(x_1)$	.1	.2	.4	.3

Let $T = X_1 + X_2$ represent the total number of stops I make on the way to and from campus.

What is distribution of T?

x_1	x_2	t	P(T=t)
0	0	0	(.1)(.1) = .01
0	1	1	(.1)(.2) = .02
0	2	2	(.1)(.4) = .04
0	3	3	(.1)(.3) = .03
1	0	1	(.2)(.1) = .02
1	1	2	(.2)(.2) = .04
1	2	3	(.2)(.4) = .08
1	3	4	(.2)(.3) = .06
2	0	2	(.4)(.1) = .04
2	1	3	(.4)(.2) = .08
2	2	4	(.4)(.4) = .16
2	3	5	(.4)(.3) = .12
3	0	3	(.3)(.1) = .03
3	1	4	(.3)(.2) = .06
3	2	5	(.3)(.4) = .12
3	3	6	(.3)(.3) = .09

So

t	0	1	2	3	4	5	6
P(T=t)	.01	.04	.12	.22	.28	.24	.09

↑ Sampling distribution of T

Note that

$$E[T] = E[X_1] + E[X_2]$$ (exercise: check)

$$Var(T) = Var(X_1) + Var(X_2)$$ (exercise: check)

By Simulation

Sometimes we can do this by manual simulation other times it is easier using a computer. What do we need to do the simulation?

1. To know the statistic of interest
2. the population distribution
3. Sample size (n)
4. number of times to repeat simulation. (k)

Basic Simulation Algorithm

for rep = 1 to K

draw n observations at random from population distribution

Evaluate sample statistic based on sample and store this result

endfor

Draw a histogram of stored sample statistic values.

Online you will find sample R and minitab code which carries out a simulation to find the sampling distribution discussed above.