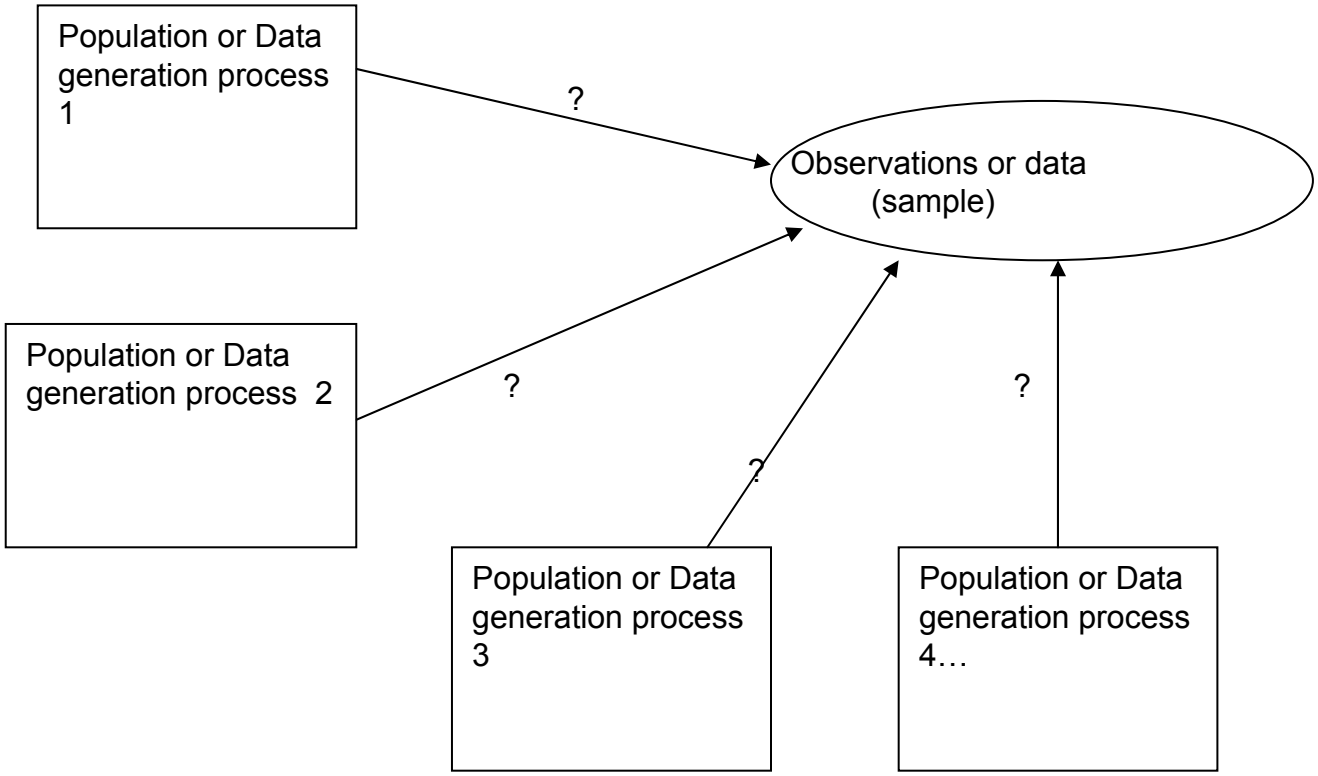
Lecture 4:
Sampling Distributions

Quantitative Methods for Regulation and Competition

Statistical Inference

- So far, we have been looking at how different data generating process can generate data
- *Statistical inference* is the art of looking at the data and making guesses about the underlying population or data generating process

Where does our data come from?



Today's lecture

- Population
- Samples and sample selection
- What is the sample mean? A RV!!!
- Point and interval estimates
- Sample variance

Populations

Definition: “Set of all things under consideration”

However, the term is used in two slightly different senses:

a) the set of individual items in which we are interested (group)

b) one or more measured attributes of that set of items

Group	**Attribute e.g.**
People entitled to vote in the next election	Height; voting intention
All past and future students of City University	Ownership of laptops
Potential applicants to course X	Current country of residence
Grains of sand in a bucket	Mass; hardness
Present and future members of the human species	Length of life; ability to jump
Outcomes of random number generating process	Value

Samples

- Definition:
  - “Subset of the population”
  - All that we observe!
  - We want to make inferences about an attribute of the population from a measured attribute of the sample
- Examples:
  - Voting intention of 1,000 people called
  - Heights of the people in the class
  - Mass of 10 grains of the bucket

Sample Selection

- Simple random sampling:
  - “each member of the population has the same chance of being chosen” and
  - “the selection of a member has no effect on the probability of another element being selected”
  - Example: select at random people from phone book
- Selective sampling:
  - Example: ask for home Internet connection at the doors of the university
  - Potential selection bias!
- Quota sampling:
  - Random sampling difficult to achieve in practice
  - Used to stratify and reduce selection bias

What is a “sample”?

- Take the first element of the sample:
  - e.g. Labour
  - This can be thought as the *outcome* of a RV X_1
  - Step back…. What is the probability distribution of this random variable? Or Prob(X_1=0) and Prob(X_1=1) where 0=Labour and 1=Conservative?

    Empirical frequency of the population!
  - E.g. if the population can be characterised by Bernouilli with p=0.55, then X_1 has the same distribution!
- A sample can be viewed as a…
  - Set of random variables!
  - Independently and identically distributed (and equal to the distribution of the population!)

Sample Mean

- The sample mean is…another random variable!

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- Objective: find its "sampling" distribution
  - First, what is the expected value (mean)?

$$E\left(\overline{X}\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E\left(X_i\right) = \frac{1}{n}\sum_{i=1}^{n} \mu_x = \mu_x$$

    where μ_x is the population mean

  - Second, what is the variance?

$$Var\left(\overline{X}\right) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var\left(X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma_x^2 = \frac{\sigma_x^2}{n}$$

    where σ^2_x is the population variance

Finally… the Central Limit Theorem

- If X_1, X_2, X_3,…, constitute a simple random sample from a population with mean μ_x and variance σ^2_X, then, for a large n…

$$\overline{X} \sim N\left(\mu_X, \frac{\sigma_X^2}{n}\right)$$

- Or alternatively…

$$Z = \frac{\overline{X} - \mu_X}{\sqrt{\frac{\sigma_X^2}{n}}} = \frac{\overline{X} - \mu_X}{\frac{\sigma_X}{\sqrt{n}}} \sim N(0,1)$$

What is the point?

- We want to *infer* a characteristic of the population (e.g. the mean) from the sample
- Point estimate:
  - e.g. sample mean is a sample *statistic* that "estimates" the population mean
- Interval estimate:
  - We need the distribution of the statistic to give *confidence intervals*, in which the characteristic will "most likely" lie on

Looking ahead: a procedure and an example

- We are interested in IQ level of all students in a given course. Our sample mean of 16 students is 120.
- Point estimate:
  - Sample mean: 120
- Interval estimate. Suppose that we want 95% of confidence and we know that the standard deviation of the population is 10.
  1. Obtain sample mean Xbar (120)
  2. Divide the standard deviation of the sample mean, std, by dividing the population std deviation by n (std =10/$\sqrt{16}$=2.5)
  3. Find the critical values *z* of the standard Normal distribution from the tables, e.g. Appendix B1 in Ashenfelter et al (1.96).
  4. Then the 95% confidence interval goes from Xbar - z. std (115.1) to Xbar + z. std (124.9)

Portion of Table B.1 in Ashenfelter

Table B.1 Cumulative Areas Under the Standard Normal Distribution

z	0	1	2	3	4	5	6	7	8	9
−3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
−2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
−2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
−2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
−2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
−2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
−2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
−2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
−2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
−2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
−2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
−1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
−1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
−1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
−1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
−1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
−1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
−1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823

Sample Variance

- The sample variance is…

$$S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

- What is its "sampling" distribution?

$$\frac{(n-1)S_X^2}{\sigma_X^2} \sim \chi_{n-1}^2$$

where chi-squared (with n-1) degrees of freedom is another distribution

- Finally, one can show that…

$$\frac{\overline{X} - \mu_X}{\frac{S_X}{\sqrt{n}}} \sim t_{n-1}$$

where t (with n-1 degrees of freedom) is another distribution