Sampling and Sampling Distribution

Sessions 9-10

Applying Sampling Distributions

- Suppose analysts decide to survey households to ascertain information about consumer spending on landscape services. Is it possible to take a census of all households or should just a sample be taken? If a sample is taken, what type of sampling technique would gain the most valid information? How can analysts be certain that the sample of households is representative of all households?
- The average household in the New Delhi spends Rs. 449 per year on their lawns and gardens. If a random sample of 40 households is taken, what is the probability that the sample mean amount spent on lawns and gardens per year is more than Rs. 500, assuming that the standard deviation is about Rs. 120?
- According to the National Association of Landscape Professionals, 35% of adults who have a lawn/landscape purchased lawn/landscaping services in the past year. Assuming this figure is true for this year, if a business analyst randomly samples 510 adults who have a lawn/landscape, what is the probability that less than 32% purchased lawn/landscaping services?

Terminology

- **Sampling Frame:** a list, map, directory, or some other source used in the sampling process to represent the population.
  - Register of students enrolled in a University
  - Register of voters in a village
- **Random Sampling (Probability Sampling)**: *every unit of the population has the same probability of being selected into the sample*
- **Non-Random Sampling:** Members of non-random samples are not selected by chance. For example, they might be selected because they are at the right place at the right time or because they know the people conducting the research.

Terminology – Random Sampling

- Simple Random Sampling
  - Number the population
  - Select the sample size
  - Generate random numbers / Use a pre-generated list of random numbers
  - Select entries in population as per numbers in the table of random numbers

Simple Random Sampling

Numbered Population of 30 Companies		
01 Alaska Airlines	11 DuPont	21 Lubrizol
02 Alcoa	12 ExxonMobil	22 Mattel
03 Ashland	13 General Dynamics	23 Merck
04 Bank of America	14 General Electric	24 Microsoft
05 Boeing	15 General Mills	25 Occidental Petroleum
06 Chevron	16 Halliburton	26 JCPenney
07 Citigroup	17 IBM	27 Procter & Gamble
08 Clorox	18 Kellogg	28 Ryder
09 Delta Air Lines	19 Kroger	29 Sears
10 Disney	20 Lowe's	30 Time Warner

R-Code

```
sample(1:100, 15, replace=F)
```

Draw a random sample of 15 from a dataset named 'df'

```
df[sample(nrow(df), 15),]
```

A Brief Table of Random Numbers							
91567	42595	27958	30134	04024	86385	29880	99730
46503	18584	18845	49618	02304	51038	20655	58727
34914	63974	88720	82765	34476	17032	87589	40836
57491	16703	23167	49323	45021	33132	12544	41035
30405	83946	23792	14422	15059	45799	22716	19792
09983	74353	68668	30429	70735	25499	16631	35006
85900	07119	97336	71048	08178	77233	13916	47564

Terminology – Random Sampling

- **Sampling Error:** By chance, the sample does not represent the population.
- **Stratified Random Sampling:** The population is divided into nonoverlapping subpopulations called strata. It has the potential for reducing sampling error.
- Internally, a stratum should be relatively homogeneous; externally, strata should contrast with each other.
- **Proportionate Stratified Random Sampling:** The percentage of the sample taken from each stratum is proportionate to the percentage that each stratum is within the whole population.
- **Disproportionate Stratified Random Sampling:** The proportions of the strata in the sample are different from the proportions of the strata in the population.

Terminology – Random Sampling

- **Systematic Sampling:** Every kth item is selected to produce a sample of size n from a population of size N.
  - Advantage: Convenience and ease of administration
  - Problem with systematic sampling can occur if the data are subject to any periodicity and the sampling interval is in sync with it.
- **Cluster (or Area) Sampling:** Dividing the population into nonoverlapping areas, or clusters. Cluster sampling identifies clusters that tend to be internally heterogeneous.

Terminology – Non-Random Sampling

- **Convenience Sampling:** Elements for the sample are selected for the convenience of the researcher – selection bias
- **Judgment Sampling:** Elements selected for the sample are chosen by the judgment of the researcher – Subjective choice
- **Snowball Sampling:** Survey subjects are selected based on referral from other survey respondents

Sample Statistics

- Population mean: μ
- Sample mean: $\bar{X}$
- Sampling distribution of $\bar{X}$
- Standard Deviation of sample: $s_x = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{(n-1)}}$
- Standard error of the mean: $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$

Central Limit Theorem

- For sufficiently large random samples from the population (μ, σ) with replacement, then the distribution of the sample means ($\bar{X}$) will be approximately normally distributed around the population mean ($\mu_{\bar{X}} = \mu$) and standard deviation ($\sigma_{\bar{X}} = \sigma/\sqrt{n}$).
- This will hold true regardless of the distribution of the source population, provided the sample size is sufficiently large (usually n $\geq$ 30).
- If the population is normal, then the theorem holds true even for samples smaller than 30
- Demo of CLT in R

Application of CLT

Suppose the population mean expenditure per customer at a store is Rs.125 and the population standard deviation is Rs.30. If a random sample of 40 customers is taken, what is the probability that the sample mean expenditure is more than Rs.133?

With $\mu = 125$ and a sample mean, of Rs.133, *z* can be computed as:

$$z = \frac{x - \mu}{\sigma/\sqrt{n}} = \frac{133 - 125}{30/\sqrt{40}} = 1.69$$

P(Z > 1.69) = 0.045

If population mean is 125, the probability that a random sample of 40 has a mean of 133 is 0.045

How much confidence do you have in your estimate?

Finite Sample Correction

Suppose the population mean expenditure per customer at a store is Rs.125 and the population standard deviation is Rs.30. If a random sample of 40 customers is taken for the total customer base of the store of 200, what is the probability that the sample mean expenditure is more than Rs.133?

With $\mu = 125$ and a sample mean, of Rs.133, *z* can be computed as:

$$z = \frac{x - \mu}{\frac{\sigma}{\sqrt{n}}\frac{N-n}{N-1}}$$

If sample is taken from a finite population, the chances of sample mean deviating from the population mean are reduced i.e. standard deviation of means is likely to be less

Hence a finite sample correction factor (N-n)/(N-1) is applied.

If sample size is less than 5% of the population, the FSC need not be applied

Application

Suppose that during any hour in a large department store, the average number of shoppers is 448, with a standard deviation of 21 shoppers. What is the probability that a random sample of 49 different shopping hours will yield a sample mean between 441 and 446 shoppers?

A production company's 350 hourly employees average 37.6 years of age, with a standard deviation of 8.3 years. If a random sample of 45 hourly employees is taken, what is the probability that the sample will have an average age of less than 40 years?

A batch of bolts purchased by your company is expected to have a diameter of 20mm. You will accept the batch if at least 95% of the bolts are between a tolerance limit of 19.5mm and 20.5mm. A random sample of 100 bolts is tested and is found to have a mean of 19.9mm with a standard deviation of 0.1mm. Will you accept the batch?

Sampling Distribution of $\hat{p}$

- Suppose 60% of the electrical contractors in a region use a particular brand of wire. What is the probability of taking a random sample of size 120 from these electrical contractors and finding that 0.50 or less use that brand of wire?
- Calculating z for sampling proportion for $n \cdot \mathrm{p} > 5$

$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$

Questions

If μ and σ are known what is the need to conduct a survey?

Will you ever know μ and σ?