Chapter 7 - Sampling Distribution

In this chapter, we talk about the techniques of collecting data and drawing samples that represent the distribution of values in a population.

Examples of data sets that use samples

1. Decennial Census
2. Current Population Survey (CPS)
3. Consumer Price Index (CPI)

Decennial Census

1. Began August, 1790. Conducted every 10 years.
2. Motivation - estimate population that can be taxed and assess the country's industrial and military potential.
3. Data are kept confidential for 75 years.

Data items collected

1. 1790 - number of persons in household and counts of persons in the following categories, free white males and females, all other free persons, slaves, ethnicity.
2. Added items - agricultural, mining, government activities, religious bodies (1810), taxes, education, wages, value of property

(1840-1850). farm and home mortgages (1880), unemployment (1930), business, housing and transportation (1940), place of work and means to work (1960), occupation history (1970)

Data items collected

1. 1980 - use and creation of TIGER files - data can be mapped geographically.

Methodology

1. 1790 - 1880 - Marshals interview household.
2. 1830 - Standard survey form used, prior to this, marshals used whatever paper was available, rule it, write in headings and bind the sheets together.
3. 1870 - Rudimentary tallying device used to help clerks.
4. 1890 - Herman Hollerith introduced punchcards and electric tabulating machines.

Methodology

1. 1910 - Census office organized as a permanent agency.
2. 1950 - First full use of computer support

3. 1960 - Devices set up to read data on returns, use of the Postal System to distribute surveys.
4. 1990 - Counts of the homeless instituted.

Use of samples

1. 1880 - Basic counts took almost until 1890 census to tabulate and publish.
2. 1890 - Supplemental survey - some subjects were covered in more detail.
3. 1940 - Sampling introduced. 5% of the population were asked an additional set of questions.

Consumer Price Index (CPI)

1. Collected monthly
2. Inititated during World War I when rapid increases in prices, particularly in shipbuilding centers made such an index essential for calculating cost-of-living adjustments in wages.
3. Hypothetical "bundle of goods" is defined for the "typical family". Changes to the bundle and the family composition occurred over time.
4. Samples of prices collected from establishments are used to estimated value of bundles.

Current Population Survey (CPS)

1. 12 monthly survey - topics include employment, temporary workers, job tenure and occupational mobility, school enrollment, race and ethnicity, voting and voter registrations, food security, work schedules, computer ownership, fertility and marital history.
2. set up in the late 1930's to provide direct measurements of unemployment each month.
3. Probability sampling as used from the beginning. Early samples of 50,000 households used to estimate employment activities of the general population.
4. Supplements decennial census data.

In this chapter, we will concern ourselves with samples and sampling distributions.

1. Previously, when we looked at distributions, we examined data distributions.
2. In this chapter, the distributions are made of parameters from samples (mean distribution, p distribution).
3. Before we talk about sampling distributions, we need to talk about how a sample is derived.

In this chapter, certain terms that we used before are given a new name.

1. **Parameters of interest** $\rightarrow \mu$, σ^2 and p are values or parameters that are ones we wish to derive from the samples.
2. **Statistics** With each sample, we derive statistics or the set of parameters of interest.

Methods of extracting a sample

1. Data on population is available - extract a subset.
2. Data on population is not available - determine how many people (n) are needed to obtain a representative sample - survey n individuals.

Types of samples

1. Random sample
2. Non-random methods (convenience sample, judgement sample, quota sampling). Data sets created by any of these methods cannot be used for making inferences.

Random Samples

1. In the simple version, each element of the population has same chance of being selected.

2. Other versions can also provide an unbiased sample.
   (a) Stratified random sample
   (b) Cluster sample
   (c) 1 in k systematic random sample.

Two methods of data collecting

1. Sampling from existing database (e.g., stock market activity, price data from a random sample of grocery stores, researcher making use of data collected by the government) - secondary source data collection
2. Retrieving data directly from the respondents using a survey designed by the researcher - primary source data collection.

Secondary data source

1. Benefits, low cost, data is typically high quality, takes less time to obtain.
2. Cost, variables might not be a close fit to the variables desired by the researcher.

Primary data source

1. Benefits, variables come close to fitting the type of variables desired.

2. Costs, added cost of survey design, data collection and processing the data.

Data collection problems that can result in an biased sample

1. Distributing surveys to a random sample and accepting a low response rate.
2. Collection techniques reaches only a subset of the full sample. Even with a 100% response rate, the methods will produce a biased sample.
3. Wording/Interviewer bias. The choice of words used in the survey and the choice of interviewers can bias the results.

Sampling Distribution

Sampling distribution of a statistics is the probability distribution for all possible values of the statistics that results when random sample of size n are repeatedly drawn from the population.

Methods of obtaining a sampling distribution

1. Derive the distribution mathematically using the laws of probability (Examples 7.3 and tables 7.5 are examples of this method).
2. Approximate the distribution empirically by drawing a large number of samples.

3. Use statistical theorems (such as the Central Limit theorem) to derive exact or approximate distributions.

Central Limit Theorem

If random samples of n observations are drawn from a non-normal population with finite mean μ and standard deviation σ, then, when n is large, the sampling distribution of the same mean $\bar{x}$ is approximately normally distributed, with mean and standard deviation (also known as the standard error of the mean (SE)).

$$\mu_{\bar{x}} = \mu$$
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Conditions

Under certain conditions, the means of random samples drawn from a population tend to approximate a normal distribution.

Conditions

1. If the population can be represented by a normal distribution, the sampling distribution of $\bar{x}$ will be normal.

2. If the population can be represented by a symmetric distribution, the sampling

distribution of $\bar{x}$ becomes normal for small values of n (for samples that are small relative to the population).

3. If the population can be represented by a skewed distribution, the sampling distribution of $\bar{x}$ becomes normal for large values of n(for samples that are large relative to the population).

Tools to assess $\bar{x}$ given $\mu_{\bar{x}} = \mu$

1. Compute the mean and standard deviation of the sample distribution ($\mu_{\bar{x}}$, $\sigma_{\bar{x}}$).
2. Determine condition to test (e.g., P($\bar{x} < 7$)).
3. Convert $\bar{x}$ to a z score using the following function
$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}}$$
4. Use the table in the back of the book to test the probability condition. region of the distribution.

Sampling Distribution of sample proportion

1. Recall from the previous chapter $\rightarrow$ Let x be a binomial random variable with n trials and probability p of success. The probability

distribution of x approximates the normal with $\mu = np$ and $\sigma = \sqrt{npq}$.

2. There is a similar outcome in sampling. Let's assume that the sampling distribution has the following characteristics. For a sample, the probability of successes is equal to the number of person with this characteristics over the total number of persons in the sample (or)
$$\hat{p} = \frac{x}{n}$$
where $\hat{p}$ is the probability of success derived from the sample.

3. For the sampling distribution
$$\mu_{\hat{p}} = p$$
$$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}}$$
where q = 1 - p.
If np > 5 and nq > 5, the sampling distribution can be approximated by the normal distribution.

Tools to assess $\hat{p}$ given that $\mu_{\hat{p}} = p$

1. Convert $\hat{p}$ into a z score and calculate the probability.
$$z = \frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}}$$