MH2500 Probability and Introduction to Statistics

Handout 11 - Survey Sampling

Introduction

Survey sampling

- Sample surveys are used to obtain information about a large population by examining only a small fraction of that population.
- In statistics, a **population** is a set of similar items or events which is of interest for some question or experiment.
- A fraction of the population is a **sample** of the population.
- The sampling techniques are probabilistic in nature—each member of the population has a specified probability of being included in the sample, and the actual composition of the sample is ramdom (ramdom sampling).

Population Parameters

Assume that the population is of size N and that associated with each member of the population is a numerical value of interest denoted by x_1, x_2, $\cdots$, x_N. (Note that x_i's are not random variables but fixed here.)

population mean

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

.

population total

$$\tau = \sum_{i=1}^{N} x_i = N\mu$$

.

Population Parameters

Assume that the population is of size N and that associated with each member of the population is a numerical value of interest denoted by x_1, x_2, $\cdots$, x_N. (Note that x_i's are not random variables but fixed here.)

population variance

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

.

population standard deviation

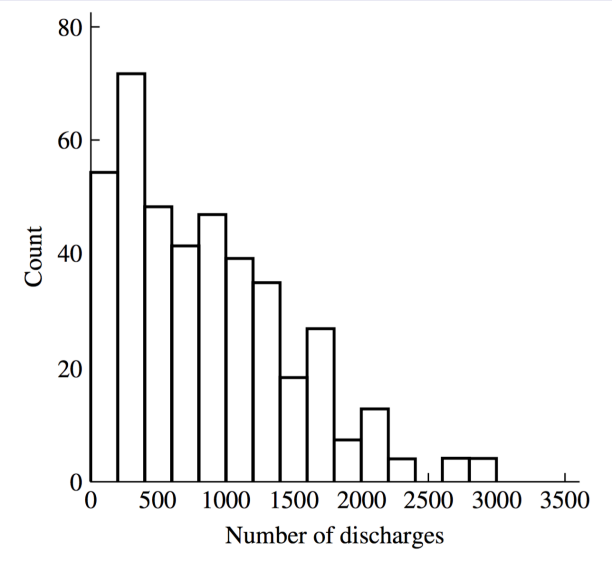
$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

A measure of how spread out, dispersed, or scattered the individual values are.

Population Parameters

Example A

The population consists of $N = 393$ short-stay hospitals. x_i denotes the number of patients discharged from the i-th hospital during January 1968. The number of hospitals that dischrged $0 - 200, 201 - 400, ..., 2801 - 3000$ patients were graphed as horizontal lines above the respective intervals.



- $\mu = 814.6$
- $\tau = 320,138$
- $\sigma^2 = 347,766$
- $\sigma = 589.7$

Simple Random Sampling

Simple random sampling

Each particular sample of size n has the same probability of occurrence; that is, each of the $\binom{N}{n}$ possible samples of size n taken without replacement has the same probability.

- The composition of the sample is random.
- We denote the sample size by n ($n \leq N$) and the values of the sample members by $X_1, X_2, ..., X_n$. X_i is the value of the ith member of the sample, which is a random variable. (x_i is the ith member of the population, which is fixed.)

The Expectation and Variance of the Sample Mean

Definition

We consider the **sample mean**,

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

as an estimate of the population mean μ, and

$$T = N\overline{X}$$

as an estimate of the population total τ.

The sample mean $\overline{X}$ is a random variable; its probalility distribution is called its **sampling distribution**. The sampling distribution of $\overline{X}$ determines how accurately $\overline{X}$ estimates μ; the more tightly the sampling distribution is centered on μ, the better the estimate.

The Expectation and Variance of the Sample Mean

LEMMA A

Denote the distinct values assumed by the population members by ξ_1, ξ_2, ..., ξ_m, and denote the number of population members that have the value ξ_j by N_j, $j = 1, 2, ..., m$. Then X_i is a discrete random variable with probability mass function

$$P(X_i = \xi_j) = \frac{N_j}{N}$$

Also,

$$E(X_i) = \mu$$

$$Var(X_i) = \sigma^2$$

The Expectation and Variance of the Sample Mean

THEOREM A

With simple random sampling, $E(\overline{X}) = \mu$.

COROLLARY A

With simple random sampling, $E(T) = E(N\overline{X}) = \tau$.

Definition

In general, if we wish to estimate a population parameter, θ say, by a function $\hat{\theta}$ of the sample, $X_1, X_2, ..., X_n$, and $E(\hat{\theta}) = \theta$, whatever the value of θ may be, we say that $\hat{\theta}$ is **unbiased**.

Remark: $\overline{X}$ and T are unbiased estimates of μ and τ.

The Expectation and Variance of the Sample Mean

LEMMA B

For simple random sampling without replacement,

$$Cov(X_i, X_j) = -\sigma^2/(N-1) \quad \text{if } i \neq j$$

THEOREM B

With simple random sampling,

$$\begin{aligned} Var(\overline{X}) &= \frac{\sigma^2}{n}\frac{N-n}{N-1} \\ &= \frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right) \end{aligned}$$

$\left(1 - \frac{n-1}{N-1}\right)$ is the **finite population correction**. n/N is the **sampling fraction**. The sampling fraction is frequently very small, in which case the **standard error** (standard deviation) of $\overline{X}$ is $\sigma_{\overline{X}} \approx \sigma/\sqrt{n}$.

The Expectation and Variance of the Sample Mean

Example B

If the population of hospitals is sampled without replacement and the sample size is $n = 32$,

$$\begin{aligned}
\sigma_{\overline{X}} &= \sqrt{\frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right)} \\
&= \frac{\sigma}{\sqrt{n}}\sqrt{1 - \frac{n-1}{N-1}} \\
&= \frac{589.7}{\sqrt{32}}\sqrt{1 - \frac{31}{392}} \\
&= 104.2 \times 0.96 \\
&= 100.0
\end{aligned}$$

Estimation of the Population Variance

The average squared deviation from the sample mean

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

THEOREM B

With simple random sampling,

$$E(\hat{\sigma}^2) = \sigma^2 \left(\frac{n-1}{n}\right) \frac{N}{N-1}.$$

$\hat{\sigma}^2$ tends to underestimate σ^2. $\frac{n}{n-1}\frac{N-1}{N}\hat{\sigma}^2$ is an unbiased estimate of σ^2.

Estimation of the Population Variance

Corollary A

An unbiased estimate of $Var(\overline{X})$ is

$$s^2_{\overline{X}} = \frac{\hat{\sigma}^2}{n}\frac{n}{n-1}\frac{N-1}{N}\frac{N-n}{N-1}$$
$$= \frac{s^2}{n}\left(1-\frac{n}{N}\right),$$

where s^2 is the sample variance

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

The quantity $S_{\overline{X}}$, also called the **estimated standard error of** $\overline{X}$, gives a rough idea of how accurate the value of $\overline{X}$ is.

Estimation of the Population Variance

Example A

A simple random sample of 50 of the 393 hospitals was taken. From this sample, $\overline{X} = 938.5$ (recall that, in fact, $\mu = 814.6$) and $s = 614.53$ ($\sigma = 589.7$). An estimate of the variance of $\overline{X}$ is

$$s^2_{\overline{X}} = \frac{s^2}{n}\left(1 - \frac{n}{N}\right) = 6592.$$

The estimated standard error of $\overline{X}$ is

$$s_{\overline{X}} = 81.19$$

(Note that the true value is $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{50}}\sqrt{1 - \frac{49}{392}} = 78.$) This estimated standard error gives a rough idea of how accurate the value of $\overline{X}$ is; in this case, we see that the magnitude of the error is of the order 80, as opposed to 8 or 800, say. In fact, the error was 123.9, or about $1.5 s_{\overline{X}}$

Normal Approximation to the Sampling Distribution of $\overline{X}$

Normal Approximation

Use the central limit theorem, if n is large but still small relative to N, then $\overline{X}_n$, the mean of a simple random sample, is approximately normally distributed

$$P\left(\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \leq z\right) = P\left(\frac{\overline{X}_n - \mu}{\sigma_{\overline{X}_n}} \leq z\right) \approx \Phi(z)$$

$$P(|\overline{X} - \mu| \leq \delta) = P(-\delta \leq \overline{X} - \mu \leq \delta) = P(-\frac{\delta}{\sigma_{\overline{X}}} \leq \frac{\overline{X} - \mu}{\sigma_{\overline{X}}} \leq \frac{\delta}{\sigma_{\overline{X}}})$$

$$\approx \Phi\left(\frac{\delta}{\sigma_{\overline{X}}}\right) - \Phi\left(-\frac{\delta}{\sigma_{\overline{X}}}\right)$$

$$= 2\Phi\left(\frac{\delta}{\sigma_{\overline{X}}}\right) - 1$$

Normal Approximation to the Sampling Distribution of $\overline{X}$

Example B

For a sample of size 50, the standard error of the sample mean number of discharges is $\sigma_{\overline{X}} = 78$. For the particular sample of size 50 discussed in Example A in Section 7.3.2, we found $\overline{X} = 938.35$, so $\overline{X} - \mu = 123.9$. We now calculate an approximation of the probability of an error this large or larger:

$$\begin{aligned} P(|\overline{X} - \mu| \geq 123.9) &= 1 - P(|\overline{X} - \mu| < 123.9) \\ &\approx 1 - \left[2\Phi\left(\frac{123.9}{78}\right) - 1\right] \\ &= 2 - 2\Phi(1.59) \\ &= 0.11 \end{aligned}$$

Thus, we can expect an error this large or larger to occur about 11% of the time.

Confidence Interval

A **confidence interval** for a population θ, is a random interval, calculated from the sample that contains θ with some specific probability. For example, a 95% confidence interval for μ is a random interval that contains μ with probability 0.95.

If we were to take many random samples and form a confidence interval from each one, about 95% of these intervals would contain μ.

If the coverage probability is $1-\alpha$, the interval is called a $100(1-\alpha)\%$ confidence interval.

E.g. A 95% confidence interval for the population average is given by

$$(\overline{X} - 1.96 s_{\overline{X}}, \overline{X} + 1.96 s_{\overline{X}}).$$

Confidence Interval

We derive a **confidence interval** for the population mean, μ.

If Z follows a standard normal distribution, then

$$P\left(-z(\alpha/2) \leq Z \leq z(\alpha/2)\right) = 1 - \alpha$$

where function $z(\alpha)$ is defined as

$$1 - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z(\alpha)} e^{-\frac{t^2}{2}} dt = \alpha.$$

From the central limit theorem, $(\overline{X} - \mu)/\sigma_{\overline{X}}$ has approximately a standard normal distribution, so

$$P\left(-z(\alpha/2) \leq \frac{\overline{X} - \mu}{\sigma_{\overline{X}}} \leq z(\alpha/2)\right) \approx 1 - \alpha$$

Confidence Interval

We derive a **confidence interval** for the population mean, μ.

$$P\left(-z(\alpha/2) \leq \frac{\overline{X}-\mu}{\sigma_{\overline{X}}} \leq z(\alpha/2)\right) \approx 1-\alpha$$

gives

$$P\left(\overline{X}-z(\alpha/2)\sigma_{\overline{X}} \leq \mu \leq \overline{X}+z(\alpha/2)\sigma_{\overline{X}}\right) \approx 1-\alpha.$$

That is, the probability that μ lies in the interval

$$\left(\overline{X}-z(\alpha/2)\sigma_{\overline{X}}, \overline{X}+z(\alpha/2)\sigma_{\overline{X}}\right)$$

is approximately $1-\alpha$. The interval is thus called a $100(1-\alpha)\%$ **confidence interval**.

Example D

A particular area contains 8000 condominium units. In a survey of the occupants, a simple random sample of size 100 yields the information that the average number of motor vehicles per unit is 1.6 with a sample standard deviation of 0.8.
Give a 95% confidence interval for the total number of motor vehicles.

. . . . . .

The estimated standard error of $\overline{X}$ is

$$s_{\overline{X}} = \frac{s}{\sqrt{n}}\sqrt{1-\frac{n}{N}} = \frac{0.8}{10}\sqrt{1-\frac{100}{8000}} = 0.08.$$

The estimated total number of motor vehicles is $T = 8000 \times 1.6 = 12,800$.

The estimated standard error of T is $s_T = 8000 \times s_{\overline{X}} = 640$.

A 95% confidence interval is

$$(T - 1.96s_T, T + 1.96s_T) = (11,546, 14,054).$$