Lecture 5

5.1. Sampling and Sampling Distribution

Summary

In this lecture we presented the concepts of simple random sampling and sampling distributions. We demonstrated how a simple random sample can be selected and how the data collected for the sample can be used to develop point estimates of population parameters Because different simple random samples provide different values for the point estimators, point estimators such as $\bar{x}$ and $\bar{p}$ are random variables. The probability distribution of such a random variable is called a sampling distribution. In particular, we described the sampling distributions of the sample mean $\bar{x}$ and the sample proportion $\bar{p}$.

In considering the characteristics of the sampling distributions of $\bar{x}$ and $\bar{p}$, we stated that $E(\bar{x}) = \mu$ and $E(\bar{p}) = p$. After developing the standard deviation or standard error formulas for these estimators, we described the conditions necessary for the sampling distributions of $\bar{x}$ and $\bar{p}$ to follow a normal distribution. Other sampling methods including stratified random sampling, cluster sampling, systematic sampling, convenience sampling, and judgment sampling were discussed.

Glossary

Parameter A numerical characteristic of a population, such as a population mean μ, a population standard deviation σ, a population proportion *p*, and so on.

Simple random sampling Finite population: a sample selected such that each possible sample of size *n* has the same probability of being selected. Infinite population: a sample selected such that each element comes from the same population and the elements are selected independently.

Sampling without replacement Once an element has been included in the sample, it is removed from the population and cannot be selected a second time.

Sampling with replacement Once an element has been included in the sample, it is re turned to the population. A previously selected element can be selected again and therefore, may appear in the sample more than once.

Sample statistic A sample characteristic, such as a sample mean $\bar{x}$, a sample standard deviation *s*, a sample proportion $\bar{p}$, and so on. The value of the sample statistic is used to estimate the value of the corresponding population parameter.

Point estimator The sample statistic, such as $\bar{x}$, *s*, or $\bar{p}$, that provides the point estimation the population parameter.

Point estimate The value of a point estimator used in a particular instance as an estimate of a population parameter.

Sampling distribution A probability distribution consisting of all possible values of a sample statistic.

Unbiased A property of a point estimator that is present when the expected value of the point estimator is equal to the population parameter it estimates.

Finite population correction factor The term $\sqrt{\dfrac{N-n}{N-1}}$ that is used in the formulas for σ_x and σ_p whenever a finite population, rather than an infinite population, is being sampled. The generally accepted rule of thumb is to ignore the finite population correction factor whenever $\dfrac{n}{N} \le 0.05$.

Standard error The standard deviation of a point estimator.

Central limit theorem A theorem that enables one to use the normal probability distribution to approximate the sampling distribution of *x* whenever the sample size is large.

Relative efficiency Given two unbiased point estimators of the same population parameter, the point estimator with the smaller standard deviation is more efficient.

Consistency A property of a point estimator that is present whenever larger sample sizes tend to provide point estimates closer to the population parameter.

Stratified random sampling A probability sampling method in which the population is first divided into strata and a simple random sample is then taken from each stratum.

Cluster sampling A probability sampling method in which the population is first divided into clusters and then a simple random sample of the clusters is taken.

Systematic sampling A probability sampling method in which we randomly select one of the first *k* elements and then select every *k*-th element thereafter.

Convenience sampling A nonprobability method of sampling whereby elements are selected for the sample on the basis of convenience.

Judgment sampling A nonprobability method of sampling whereby elements are selected for the sample based on the judgment of the person doing the study.

Key formulas

Expected (mean) Value of $\bar{x}$

$$E(\bar{x}) = \mu \tag{5.1}$$

Standard Deviation of $\bar{x}$ (Standard Error)

Finite Population / Infinite Population

$$\sigma_{\bar{x}} = \sqrt{\frac{N-n}{N-1}}\left(\frac{\sigma}{\sqrt{n}}\right) \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{5.2}$$

Expected Value of $\bar{p}$

$$E(\bar{p}) = p \tag{5.3}$$

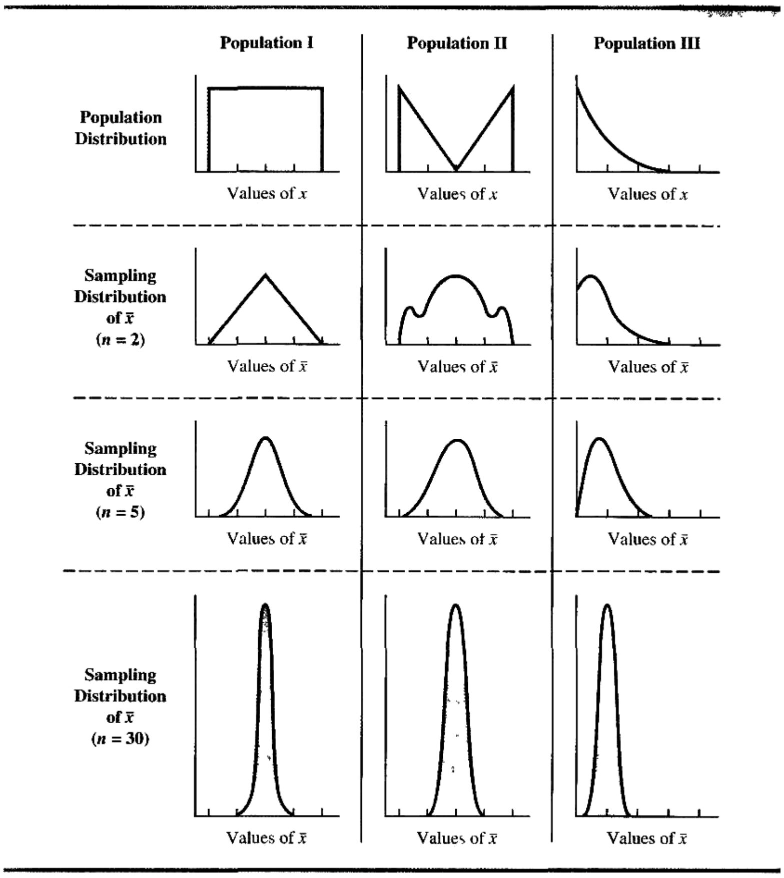
Standard Deviation of $\bar{p}$ (Standard Error)

Finite Population / Infinite Population

$$\sigma_{\bar{p}} = \sqrt{\frac{N-n}{N-1}}\sqrt{\frac{p(1-p)}{n}} \qquad \sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} \tag{5.4}$$

Lecture supplementary material

FIGURE 7.3 ILLUSTRATION OF THE CENTRAL LIMIT THEOREM FOR THREE POPULATIONS



5.2. Interval Estimation

Summary

In this lecture we considered methods for developing interval estimates of a population mean and a population proportion. A point estimator may or may not provide a good estimate of a population parameter. The use of an interval estimate provides a measure of the precision of an estimate. Both the interval estimate of the population mean and the population proportion are of the form: point estimate ± margin of error.

We presented interval estimates for a population mean for two cases. In the σ known case, historical data or other information is used to develop an estimate of σ prior to taking a sample. Analysis of new sample data then proceeds based on the assumption that σ is known. In the σ unknown case, the sample data are used to estimate both the population mean and the population standard deviation. The final choice of which interval estimation procedure to use depends upon the analyst's understanding of which method provides the best estimate of σ.

In the σ known case, the interval estimation procedure is based on the assumed value of σ and the use of the standard normal distribution. In the σ unknown case, the interval estimation procedure uses the sample standard deviation *s* and the *t* distribution. In both cases the quality of the interval estimates obtained depends on the distribution of the population and the sample size. If the population is normally distributed the interval estimates will be exact in both cases, even for small sample sizes. If the population is not normally distributed, the interval estimates obtained will be approximate. Larger sample sizes will provide better approximations, but the more highly skewed the population is, the larger the sample size needs to be to obtain a good approximation.

The general form of the interval estimate for a population proportion is $\overline{p}$ ± margin of error. In practice the sample sizes used for interval estimates of a population proportion are generally large. Thus, the interval estimation procedure is based on the standard normal distribution.

Glossary

Interval estimate An estimate of a population parameter that provides an interval believed to contain the value of the parameter. For the interval estimates in this chapter, it has the form: point estimate ± margin of error.

Margin of error The ± value added to and subtracted from a point estimate in order to develop an interval estimate of a population parameter.

σ known The condition existing when historical data or other information provides a good value for the population standard deviation prior to taking a sample. The interval estimation procedure uses this known value of *o* in computing the margin of error.

σ unknown The condition existing when no good basis exists for estimating the population standard deviation prior to taking the sample. The interval estimation procedure uses the sample standard deviation *s* in computing the margin of error.

Confidence level The confidence associated with an interval estimate. For example, if an interval estimation procedure provides intervals such that 95% of the intervals formed using the procedure will include the population parameter, the interval estimate is said to be constructed at the 95% confidence level.

Confidence coefficient The confidence level expressed as a decimal value. For example, .95 is the confidence coefficient for a 95% confidence level.

Confidence interval Another name for an interval estimate.

***t* distribution** A family of probability distributions that can be used to develop an interval estimate of a population mean whenever the population standard deviation *o* is unknown and is estimated by the sample standard deviation *s*.

Degrees of freedom A parameter of the *t* distribution. When the *t* distribution is used in the computation of an interval estimate of a population mean, the appropriate *t* distribution has *n* — 1 degrees of freedom, where *n* is the size of the simple random sample.

Key formulas

Interval Estimate of a Population Mean: σ Known

$$\overline{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \tag{5.5}$$

Interval Estimate of a Population Mean: σ Unknown

$$\overline{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}} \tag{5.6}$$

Sample Size for an Interval Estimate of a Population Mean

$$n = \frac{\left(z_{\alpha/2}\right)^2 \sigma^2}{E^2} \tag{5.7}$$

Internal Estimate of a Population Proportion

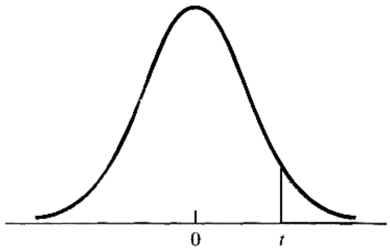
$$\bar{p} \pm z_{\alpha/2} \sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \tag{5.8}$$

Sample Size for an Interval Estimate of a Population Proportion

$$n = \frac{\left(z_{\alpha/2}\right)^2 p^*(1-p^*)}{E^2} \tag{5.9}$$

Lecture supplementary material

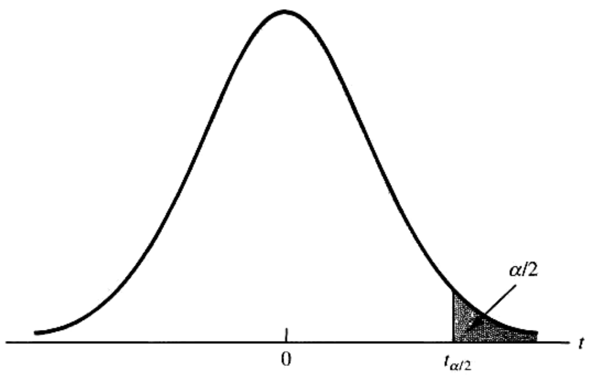
TABLE 8.2 t DISTRIBUTION TABLE FOR AREAS IN THE UPPER TAIL
EXAMPLE WITH 10 DEGREES OF FREEDOM, $t_{.025} = 2.228$



Degrees of Freedom	Area in Upper Tail					
	.20	**.10**	**.05**	**.025**	**.01**	**.005**
1	1.376	3.078	6.314	12.706	31.821	63.656
2	1.061	1.886	2.920	4.303	6.965	9.925
3	.978	1.638	2.353	3.182	4.541	5.841
4	.941	1.533	2.132	2.776	3.747	4.604
5	.920	1.476	2.015	2.571	3.365	4.032
6	.906	1.440	1.943	2.447	3.143	3.707
7	.896	1.415	1.895	2.365	2.998	3.499
8	.889	1.397	1.860	2.306	2.896	3.355
9	.883	1.383	1.833	2.262	2.821	3.250
10	.879	1.372	1.812	2.228	2.764	3.169
11	.876	1.363	1.796	2.201	2.718	3.106
12	.873	1.356	1.782	2.179	2.681	3.055
13	.870	1.350	1.771	2.160	2.650	3.012
14	.868	1.345	1.761	2.145	2.624	2.977
15	.866	1.341	1.753	2.131	2.602	2.947
16	.865	1.337	1.746	2.120	2.583	2.921
17	.863	1.333	1.740	2.110	2.567	2.898
18	.862	1.330	1.734	2.101	2.552	2.878
19	.861	1.328	1.729	2.093	2.539	2.861
20	.860	1.325	1.725	2.086	2.528	2.845
21	.859	1.323	1.721	2.080	2.518	2.831
22	.858	1.321	1.717	2.074	2.508	2.819
23	.858	1.319	1.714	2.069	2.500	2.807
24	.857	1.318	1.711	2.064	2.492	2.797
25	.856	1.316	1.708	2.060	2.485	2.787
26	.856	1.315	1.706	2.056	2.479	2.779
27	.855	1.314	1.703	2.052	2.473	2.771
28	.855	1.313	1.701	2.048	2.467	2.763
29	.854	1.311	1.699	2.045	2.462	2.756
30	.854	1.310	1.697	2.042	2.457	2.750
40	.851	1.303	1.684	2.021	2.423	2.704
50	.849	1.299	1.676	2.009	2.403	2.678
60	.848	1.296	1.671	2.000	2.390	2.660
80	.846	1.292	1.664	1.990	2.374	2.639
100	.845	1.290	1.660	1.984	2.364	2.626
∞	.842	1.282	1.645	1.960	2.326	2.576

Note: A more extensive table is provided as Table 2 of Appendix B.

FIGURE 8.5 t DISTRIBUTION WITH $\alpha/2$ AREA OR PROBABILITY IN THE UPPER TAIL



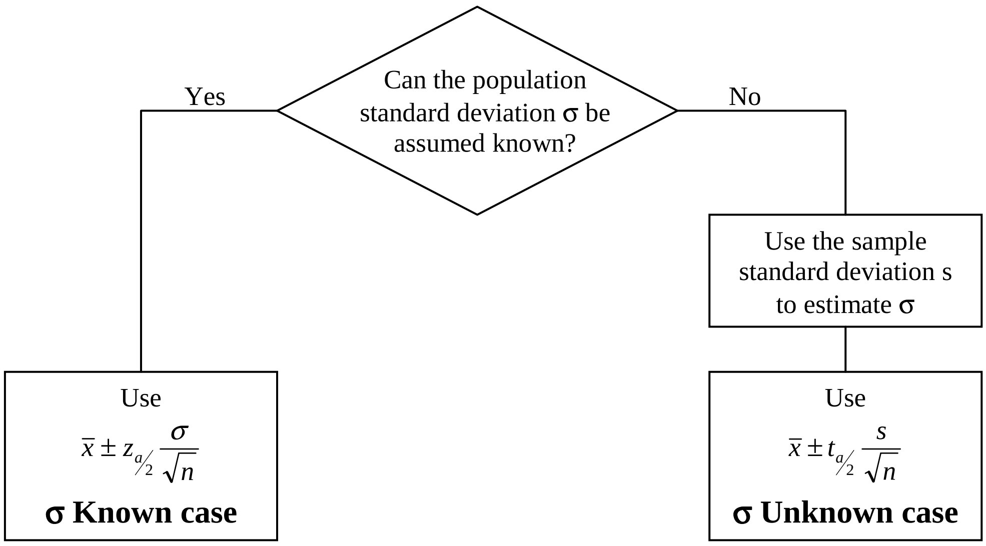


Figure. Summary of internal estimation procedures for a population mean