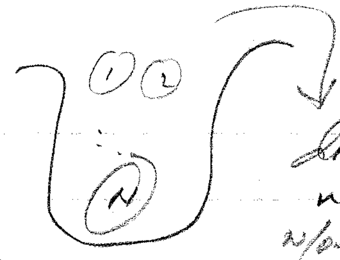
Ch. 7 Survey Sampling

Def: Simple Random Sampling.



drawn n w/out rep.

$\mu = \frac{1}{N}\sum^{N} x_i$

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

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.

Remark: Any statistic computed from a random sample is a random variable, and has an associated sampling distribution.

The sampling distribution of $\bar{X}$ determines how accurately $\bar{X}$ estimates μ; roughly speaking, the more tightly the sampling distribution is centered around μ, the better the estimate.

C.L.T. $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$.

As a measure of the center of the sampling distribution, we use $E[\bar{X}] = \mu$ and as a measure of the dispersion of the sampling distribution, we will use $SD(\bar{X}) = \frac{\sigma}{\sqrt{n}}$.

<u>Thm A</u>: With SRS $E[\bar{X}] = \mu$.

$\bar{X}$ is an unbiased estimator of μ

<u>Lemma B</u>: With SRS, without rep

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

<u>Thm B</u>:
$$Var(\bar{X}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$$
$$= \frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right).$$
$$= \frac{\sigma^2}{n}\left(f.p.c\right)$$

finite population correction

If N is large $Var(\bar{X}) \approx \frac{\sigma^2}{n}$

Estimation of the Population Variance

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \qquad \text{biased.}$$

<u>Thm:</u> $E(\hat{\sigma}^2) = \sigma^2\left(\frac{n-1}{n}\right)\left(\frac{N}{N-1}\right)$.

<u>Cor:</u> An unbiased estimate of $Var(\bar{X})$ is

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

$$= \frac{S^2}{n}\left(1 - \frac{n}{N}\right)$$

If N is large $S_{\bar{X}}^2 \doteq \frac{S^2}{n}$.

In practice we ignore the finiteness of the population and assume $n \ll N$. This implies independence in the sampling and

$$Cov(X_i, X_j) \doteq 0.$$

C.L.T. Normal Approximation to the Sampling Distribution of $\bar{X}$

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

The spread of the sampling distribution and therefore the precision of $\bar{X}$ are determined by the sample size n and not by the population size N.

Confidence Intervals

$100(1-\alpha)\%$ CI for μ, large $n \geq 30$

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$100(1-\alpha)\%$ CI for μ, normal population, small $n < 30$

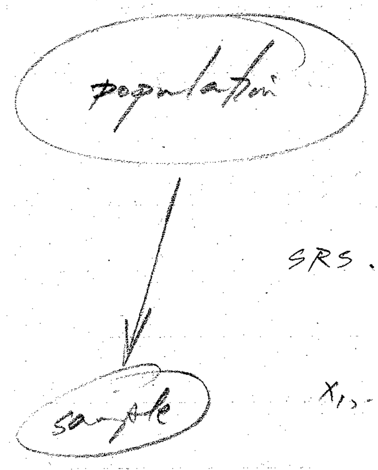
$$\bar{x} \pm t_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$100(1-\alpha)\%$ CI for π, large n

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

Inference Procedures

① Point Estimation



μ

$X_1, \dots, X_n \qquad \bar{X}$

Estimate μ with $\bar{x}$, $\bar{x}$ is unbiased.

② Interval Estimation:

Give a SRS, $X_1, \dots, X_n$ from a population with unknown μ and known σ^2. a $100(1-\alpha)\%$ confidence is a random interval s.t.

$$P\left(-z_{\alpha/2} \le Z \le z_{\alpha/2}\right) = 1-\alpha$$

$$P\left(-z_{\alpha/2} \le \underbrace{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}_{N(0,1)} \le z_{\alpha/2}\right) = 1-\alpha$$

$$P\left(\mu - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \bar{X} \le \mu + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

here $\bar{X}$ is a r.v.

$$P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

here the endpoints are random.

Before we collect our data the confidence interval

$$\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

has a 95% probability of including μ. After we collect our data, the CI

$$\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

can be interpreted as follows:

"We are $100(1-\alpha)\%$ confidence that our interval includes μ."

Note: We don't not use the word probability when discussing a single confidence interval. A single interval either includes μ or it doesn't, we don't know.

see p. 204.

Given a SRS $X_1, \dots, X_n$ from a population with unknown μ and unknown σ^2 a $100(1-\alpha)\%$ confidence interval is

$$P\left(-t_{\alpha/2} < T < t_{\alpha/2}\right) = 1-\alpha$$

$$P\left(-t_{\alpha/2} < \frac{\bar{X}-\mu}{s/\sqrt{n}} < t_{\alpha/2}\right) = 1-\alpha$$

$$= P\left(\mu - t_{\alpha/2}\frac{s}{\sqrt{n}} \le \bar{X} \le \mu + t_{\alpha/2}\frac{s}{\sqrt{n}}\right) = 1-\alpha$$

here $\bar{X}$ is a r.v.

$$P\left(\bar{X} - t_{\alpha/2}\frac{s}{\sqrt{n}} \le \bar{X} \le \mu + t_{\alpha/2}\frac{s}{\sqrt{n}}\right) = 1-\alpha$$

so $\bar{X} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}$ is a $100(1-\alpha)\%$ CI for μ.

Aside: sample size calculations

"Forethought in Statistics"

find the sample size n need to have a margin-of-error of .03 at 95% confidence when estimate the population proportion π.

$$100(1-\alpha)\%\ CI \quad \hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = .03$$

conservative, let $\hat{p} = .5$

$$1.96\sqrt{\frac{(.5)^2}{n}} = .03$$

$$\left[\frac{(1.96)(.5)}{.03}\right]^2 = n$$

$$1067 = n$$

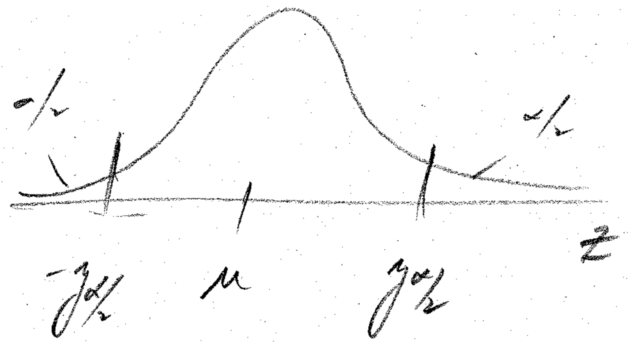
Hypothesis Testing

$H_0: \mu = \mu_0$

$H_A: \mu \neq \mu_0$

$X_1, \ldots, X_n$ a r.s.

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1^2).$$



Reject H_0 if Z falls in a tail.

Figure 7-6
Uniform distribution

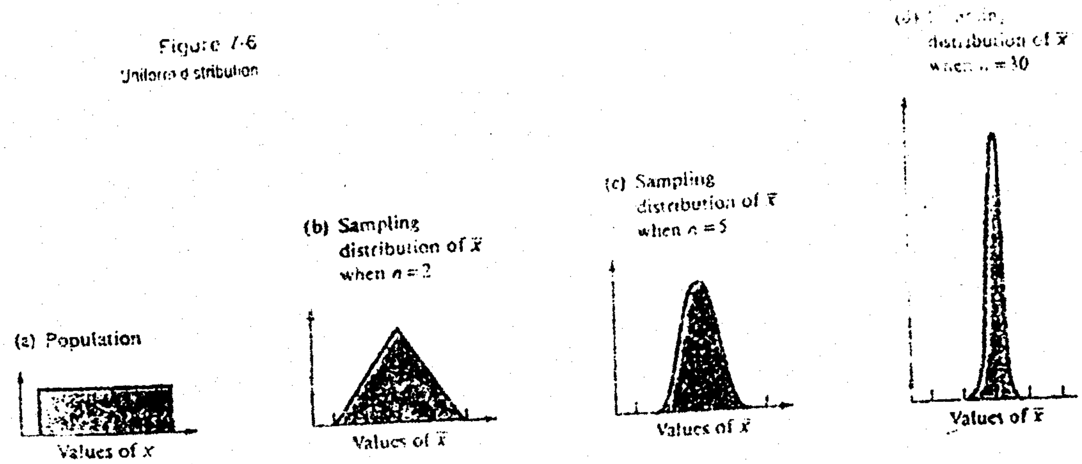


Figure 7-7
U-shaped distribution

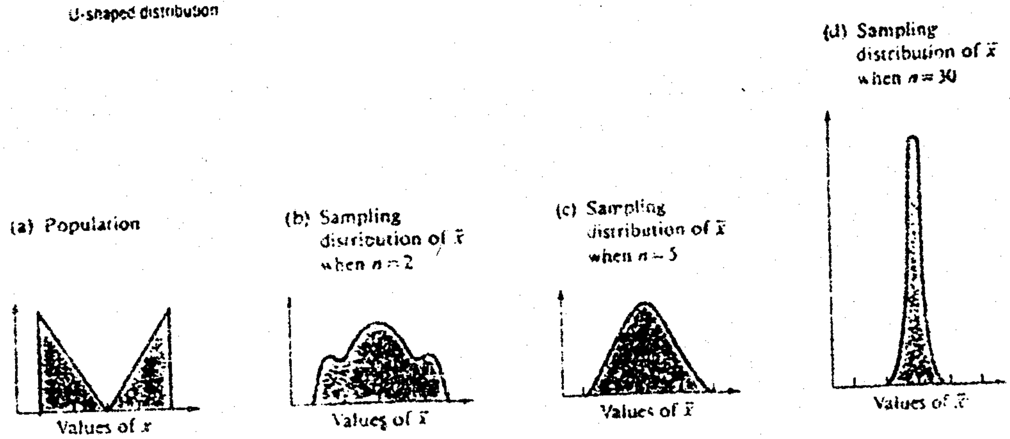


Figure 7-8
J-shaped distribution

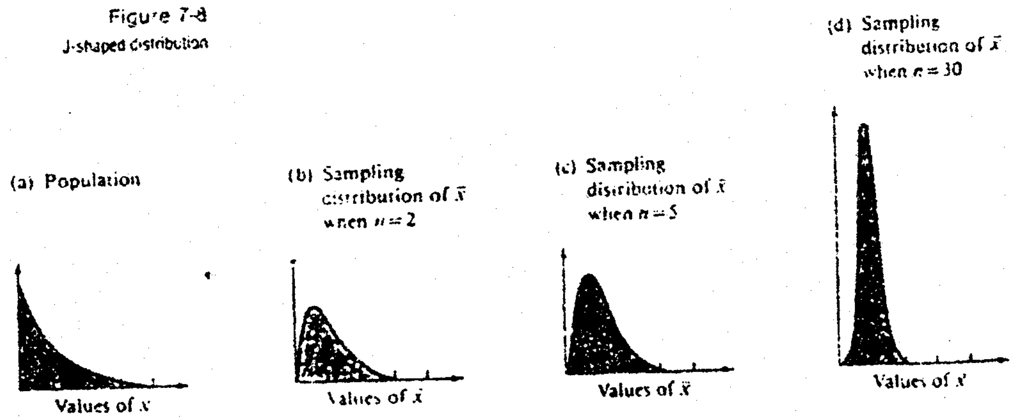


Figure 7-9
Normal distribution