

Chapter 9 - Estimation

Point Estimation

1. **Point Estimator** - A formula applied to a data set which results in a single value or point. In other words, a statistic. Usually, estimators are used to give plausible values of some population parameter.
 - \bar{X} , \tilde{X} , S^2 , and p are point estimators of the parameters μ , $\tilde{\mu}$, σ^2 and π respectively.
2. **Point Estimate** - The resulting value of a point estimator, when applied to a data set.
 - $\bar{x} = 27.6$, $\tilde{x} = 85$, $s^2 = 2.45$, and $p = 0.30$ are point estimates of the values of μ , $\tilde{\mu}$, σ^2 and π respectively.

Desirable Properties of a Point Estimator:

1. **Unbiased** - A point estimator is unbiased for a parameter if the mean of the estimator's sampling distribution equals the value of the parameter. Otherwise, the estimator is biased.
 - \bar{X} is an unbiased estimator of μ because $\mu_{\bar{X}} = \mu$.
 - S^2 is an unbiased estimator of σ^2 because $\mu_{S^2} = \sigma^2$. S is NOT an unbiased estimate of σ !
 - p is an unbiased estimator of π because $\mu_p = \pi$.
2. **Smallest Variability** - When choosing among unbiased estimators, the one with the smallest sampling variability (i.e. smallest standard deviation) is the best, because the point estimates will be most closely concentrated around the parameter value.
 - A Point Estimator that has properties 1 and 2 above is called a **Minimum Variance Unbiased Estimator (MVUE)**. When the data is normal, then \bar{X} is an MVUE for μ . That is, $\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}$ is smaller than the variance for any other point estimator of μ .
 - When the data is not normal, then \bar{X} is not an MVUE for μ (see page 365 of your textbook).

Interval Estimation

1. **Interval Estimator:** A formula applied to a data set which results in an interval of plausible values for some parameter. It has the form
$$\text{point estimator} \pm \text{margin of error.}$$
2. **Confidence Interval (C.I.):** An interval estimator which has a *level of confidence* attached to it.
3. **Confidence Level:** A quantity (typically stated as a percentage) describing how often a confidence interval (over all samples of a given size) captures the parameter value.
 - Commonly-used confidence levels are 90%, 95%, and 99%.
 - A 95% confidence level means that:
 - With probability .95, the formula for the confidence interval captures the value of statistic is .95
 - If confidence intervals were calculated from all possible samples, then 95% of the intervals would contain the parameter value.

Confidence Interval for μ (when σ is known) \longrightarrow $\boxed{\bar{X} \pm z_{1-\frac{\alpha}{2}} \left(\frac{\sigma}{\sqrt{n}} \right)}$

Assumptions:

- The data must be a SRS (so if you sample a finite population without replacement, then $.05N \geq n$).
- The sampling distribution of \bar{X} must be at least approximately normal (so the data is normal OR $n > 15$ for symmetric non-normal data OR $n > 30$ for skewed data).

Notation:

- the margin of error is $m = z_{1-\frac{\alpha}{2}} \left(\frac{\sigma}{\sqrt{n}} \right)$.
- $z_{1-\frac{\alpha}{2}}$ is the *critical value*, the $100 \left(1 - \frac{\alpha}{2} \right)$ percentile of the Z distribution, $Z \sim N(0, 1)$.
- α is a *significance level*
- $C = 1 - \alpha$ is the confidence level

C	$\alpha = 1 - C$	$1 - \frac{\alpha}{2}$	$z_{1-\frac{\alpha}{2}}$
0.90	0.10	0.95	
0.95	0.05	0.975	
0.99	0.01	0.995	

Find $z_{1-\frac{\alpha}{2}}$ for 90%, 95% and 99% confidence intervals using the normal table.

EXAMPLE #1: (Problem 9.37 on page 392) Seventy seven students at the University of Virginia were asked to keep a diary of conversations with their mothers, recording any lies they told during the conversations. Suppose that $\bar{x} = .5$ and $\sigma = .4$.

- Is the sample size large enough to calculate a C.I.?
- Construct a 90% C.I. for μ , the true number of lies told to mothers during each conversation.

Confidence Interval Behavior:

The margin of error m , and hence the width of the interval depends, on:

1. The confidence level C : Increasing C increases m and the interval width.
2. The variability of the response σ : Increasing σ increases m and the interval width.
3. The sample size n : Increasing n decreases m and the interval width.

- The C.I.'s with \bar{x} 's that are less than one margin of error away from μ will be the intervals that capture μ . This happens $100C\%$ of the time.
- All \bar{x} 's that are more than one margin of error away from μ will be the intervals that do not capture μ . This happens $100\alpha\%$ of the time.
- Typically, only one sample is selected from the population and therefore only one confidence interval will be calculated. The **CORRECT INTERPRETATION** of this *one* confidence interval is:

We are _____% confident that μ lies between _____ and _____.

Usually, one interprets μ in terms of the problem.

- An **INCORRECT INTERPRETATION** of this *one* confidence interval is:

There is a _____% chance (i.e. probability) that μ lies between _____ and _____.

Probability statements can only be made about confidence intervals BEFORE you calculate the specific confidence interval estimate for the data. A given interval estimate captures the parameter with probability **0** or **1**, we simply do not know which it is.

EXAMPLE #1 revisited: From our previous example, we'd interpret the C.I. like this:

We are _____% confident that, on average, University of Virginia students lie to their mothers between _____ and _____ times per conversation.

Confidence Interval for μ (σ unknown) \longrightarrow $\boxed{\bar{X} \pm t_{1-\frac{\alpha}{2}, n-1} \left(\frac{s}{\sqrt{n}} \right)}$

Assumptions:

- The data must be a SRS (so if you sample a finite population without replacement, then $.05N \geq n$).
- The sampling distribution of \bar{X} must be at least approximately normal (so the data is normal OR $n > 15$ for symmetric non-normal data OR $n > 30$ for skewed data).

t Distribution $\longrightarrow T \sim t(df)$

- The t distribution is symmetric, unimodal, bell-shaped, and centered at zero.
- The t distribution has heavier tails than the Z distribution because s (an estimate of σ) is used instead of σ .
- As the degrees of freedom (df) increases, the t distribution approaches the Z distribution.

Notation:

- the margin of error is $m = t_{1-\frac{\alpha}{2}, n-1} \left(\frac{s}{\sqrt{n}} \right)$.
- Use the t-table or R to calculate $t_{1-\frac{\alpha}{2}, n-1}$, the t *critical value*, the 100 $(1 - \frac{\alpha}{2})$ percentile of the t distribution with n - 1 degrees of freedom, $t(n - 1)$. Since $t(n - 1)$ has thicker tails than $N(0, 1)$, then $t_{1-\frac{\alpha}{2}, n-1} > z_{1-\frac{\alpha}{2}}$.
- $\frac{s}{\sqrt{n}}$ is the *standard error* of \bar{X}

EXAMPLE #3: From a SRS of 8 shipments of corn soy blend, a highly nutritious food sent for emergency relief, the mean vitamin C content (in mg/100g) is $\bar{x} = 22.5$ and the sample standard deviation is $s = 7.19$.

- Do we have a large enough sample size?
- Calculate and interpret a 99% confidence interval for μ , the true mean vitamin C content of corn soy blend.

Confidence Interval for π \longrightarrow $p \pm z_{1-\frac{\alpha}{2}} \left(\sqrt{\frac{p(1-p)}{n}} \right)$

Assumptions:

- The data must be a SRS (so if you sample a finite population without replacement, then $.05N \geq n$).
- The sampling distribution for p must be approximately normal (so $np \geq 10$ and $n(1-p) \geq 10$)

EXAMPLE #4:

In a study of heavy drinking on college campuses, 17096 students were interviewed. Of these, 3314 admitted to consuming more than 5 drinks at a time, three times a week.

- Give a point estimate of the proportion of college students who are “heavy” drinkers.
- Is the sample large enough to assume that the sampling distribution of p is approximately normal?
- Give a 99% C.I. for the proportion of all college students who are heavy drinkers.

Sample Size Calculations

Before any study, a researcher often already knows the confidence level and the *precision* (or margin of error) of a desired confidence interval. For a fixed confidence level and margin of error, the only other factor under the researcher's control is the sample size. So a researcher must collect enough data to be able to construct the desired C.I.'s.

Sample Size Calculation for Estimating μ : \longrightarrow
$$n = \left(\frac{z_{1-\frac{\alpha}{2}} \sigma}{m} \right)^2$$

- $z_{1-\frac{\alpha}{2}}$ is the critical value for the desired confidence level $C = 1 - \alpha$.
- m is the desired margin of error
- σ is the standard deviation of the population

If σ is unknown, two options for estimating σ are:

1. Use a sample standard deviation, s , from a previous study.
2. Use the anticipated range divided by 4.

EXAMPLE:

Find the sample size necessary to estimate the mean level of phosphate in the blood of dialysis patients to within 0.05 with 90% confidence. A previous study calculated a sample standard deviation of $s = 1.6$.

Sample Size Calculation for Estimating π : \longrightarrow
$$n = \pi(1 - \pi) \left(\frac{z_{1-\frac{\alpha}{2}}}{m} \right)^2$$

Two options for the value of π :

1. Use an estimate, p , from a previous study.
2. Use $\pi = \frac{1}{2}$. This is the more conservative choice because using it will result in a sample size n even larger than needed.

EXAMPLE:

Your company would like to carry out a survey of customers to determine the degree of satisfaction with your customer service. You want to estimate the proportion of customers who are satisfied. What sample size is needed to attain 95% confidence and a margin of error less than or equal to 3%, or 0.03?

R commands

```
> # EXAMPLE 1
> # Checking the z critical value
> qnorm(.95)
[1] 1.644854
> .5+c(-1,1)*qnorm(.95)*.4/sqrt(77)
[1] 0.4250206 0.5749794
>
>
>
> # EXAMPLE 2
> qnorm(.975)
[1] 1.959964
> 18.48 + c(-1,1)*qnorm(.975)*2.9/sqrt(20)
[1] 17.20904 19.75096
>
>
>
> # EXAMPLE 3
> # Comparing the z and t critical values
> qnorm(.995)
[1] 2.575829
>
> # For a t distribution, you have to tell R the degrees of freedom
> qt(.995,df=7)
[1] 3.499483
> 22.5 + c(-1,1)*qt(.995,df=7)*7.19/sqrt(8)
[1] 13.60414 31.39586
>
>
>
> # EXAMPLE 4
> n=17096
> p=3314/n
> p
[1] 0.1938465
>
> # Checking sample size
> n*p
[1] 3314
> n*(1-p)
[1] 13782
> qnorm(.995)
[1] 2.575829
> p + c(-1,1)*qnorm(.995)*sqrt(p*(1-p)/n)
[1] 0.1860588 0.2016342
```

Estimating μ after transforming data

For large sample sizes ($n > 30$), we do not need to assume that the data is normal in order to find a C.I. for μ . But when we have a small sample from a population which is clearly non-normal ($n < 15$), then a transform may be appropriate (see Chapter 7 notes). For the Box-Cox family of transforms,

$$Y_i = X_i^\lambda, \text{ if } \lambda \neq 0$$
$$Y_i = \ln(X_i), \text{ if } \lambda = 0.$$

one can not directly interpret a point estimate \bar{y} or a C.I. of μ_Y . In practice, statisticians “back-transform” point estimates and the limits of the C.I.,

$$\bar{y}^{\frac{1}{\lambda}} \quad \text{and} \quad \left(\bar{y} \pm t_{1-\frac{\alpha}{2}, n-1} \left(\frac{s_y}{\sqrt{n}} \right) \right)^{\frac{1}{\lambda}}.$$

These back-transformed values yield theoretically justifiable point estimates and C.I.’s for μ_X as long as $\frac{\sigma_X^2}{\mu_X^2}$ is “small.” In practice, you would check $\frac{s^2}{\bar{x}^2}$. The justification for $\lambda \neq 0$ is

$$\mu_Y^{\frac{1}{\lambda}} \approx \mu_X \left(1 - \frac{\lambda(\lambda-1)}{2} \frac{\sigma_X^2}{\mu_X^2} \right)^{\frac{1}{\lambda}}$$

and when $\lambda = 0$

$$\ln \mu_Y \approx \mu_X \left(e^{-\frac{\sigma^2}{2\mu^2}} \right).$$

Exercises

9.1 on p366: 1-7 odd

9.2 on p379 (C.I. for π): 11-17 odd, 21-25 odd, 27 (sample size calculation).

9.3 on p391 (C.I. for μ): 29-35 odd, 39-45 odd, 47 (sample size calculation)

Reading

Sections 9.1-9.3