Chapters 4-6: Estimation

Read sections 4.2 (except 4.2.3), 4.5.1, 5.1 (except 5.1.5), 6.1 (except 6.1.3)

Point Estimation

(4.1.1)

**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.

- $\overline{X}$, $\tilde{X}$, $S^2$, and $p$ are point estimators of the parameters $\mu$, $\tilde{\mu}$, $\sigma^2$ and $p$ respectively.

**Point Estimate** - The resulting value of a point estimator, when applied to a data set.

- $\overline{x} = 27.6$, $\tilde{x} = 85$, $s^2 = 2.45$, and $\hat{p} = 0.30$ are point estimates of the values of $\mu$, $\tilde{\mu}$, $\sigma^2$ and $p$ respectively.

Desirable Properties of a Point Estimator

(4.5 intro)

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**.

   - $\overline{X}$ is an unbiased estimator of $\mu$ because $\mu_{\overline{X}} = \mu$.
   - $S^2$ is an unbiased estimator of $\sigma^2$ because $\mu_{S^2} = \sigma^2$.
   - $S$ is NOT an unbiased estimate of $\sigma$ because $\mu_S \neq \sigma$. In fact, $\mu_S < \sigma$. However, the bias goes to zero for larger sample sizes.
   - $\hat{p}$ is an unbiased estimator of $p$ because $\mu_{\hat{p}} = p$.

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 are normal, then $\overline{X}$ is an MVUE for $\mu$. That is, $\sigma^2_{\overline{X}} = \frac{\sigma^2}{n}$ is smaller than the variance for any other point estimator of $\mu$. 
Interval Estimation

(4.2)

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 (CI)**: An interval estimator which has a *confidence level* 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 the parameter.
     - If confidence intervals were calculated from all possible samples, then 95% of the intervals would contain the parameter value.
Confidence Interval for \( \mu \) when \( \sigma \) is known (4.2.4) \[
\bar{X} \pm z_{1-\frac{\alpha}{2}} \left( \frac{\sigma}{\sqrt{n}} \right)
\]

Assumptions:
- The data must be a SRS.
- 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 are normal OR \( n \geq 15 \) for symmetric non-normal data OR \( n \geq 30 \) 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 \((1 - \frac{\alpha}{2})\) percentile of the Z distribution, \( Z \sim N(0, 1) \).
- \( \alpha \) is a significance level
- \( C = 1 - \alpha \) is the confidence level

<table>
<thead>
<tr>
<th>C</th>
<th>( \alpha = 1 - C )</th>
<th>( 1 - \frac{\alpha}{2} )</th>
<th>( z_{1-\frac{\alpha}{2}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.90</td>
<td>0.10</td>
<td>0.95</td>
<td>( z_{1-\frac{0.10}{2}} )</td>
</tr>
<tr>
<td>0.95</td>
<td>0.05</td>
<td>0.975</td>
<td>( z_{1-\frac{0.05}{2}} )</td>
</tr>
<tr>
<td>0.99</td>
<td>0.01</td>
<td>0.995</td>
<td>( z_{1-\frac{0.01}{2}} )</td>
</tr>
</tbody>
</table>

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

EXAMPLE #1: 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 CI?
- Construct a 90% CI for \( \mu \), 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 \( \sigma \): Increasing \( \sigma \) increases \( m \) and the interval width.
3. The sample size \( n \): Increasing \( n \) decreases \( m \) and the interval width.

- The CI's with \( x \)'s that are less than one margin of error away from \( \mu \) will be the intervals that capture \( \mu \). This happens 100\( C \)% of the time.
- All \( x \)'s that are more than one margin of error away from \( \mu \) will be the intervals that do not capture \( \mu \). 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 \( \mu \) lies between ______ and ______.

  Usually, one interprets \( \mu \) in terms of the problem.

- Unless you are a Bayesian statistician, an **INCORRECT INTERPRETATION** of this one confidence interval is:

  There is a _____% chance (i.e. probability) that \( \mu \) 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 CI like this:

We are _____% confident that, on average, University of Virginia students lie to their mothers between ______ and ______ times per conversation.
QUESTION: True or False?

Suppose \( \mu \) is the average weight (in pounds) of dogs in Bozeman. A 95% CI for \( \mu \) is (42, 48).

1. T / F: Ninety-five percent of the weights in the population are between 42 and 48 pounds.
2. T / F: Ninety-five percent of the weights in the sample are between 42 and 48 pounds.
3. T / F: The probability that the confidence interval (42, 48) includes \( \mu \) is 0.95.
4. T / F: The sample mean \( \bar{x} \) is in the confidence interval with probability 0.95.
5. T / F: If 100 confidence intervals were generated using this same process, approximately 5 of the confidence intervals would not include \( \mu \).

EXAMPLE #2:
Gas mileages (in MPG) of a SRS of 20 cars of a certain make and model are recorded and the sample mean is found to be 18.48 MPG. Suppose that the data are not normal but are fairly symmetric. Also suppose that the population standard deviation is known to be \( \sigma = 2.9 \) MPG.

- Do we have a large enough sample size?

- Find and interpret a 95% confidence interval for \( \mu \), the mean MPG for this vehicle.
Confidence Interval for $\mu$ when $\sigma$ unknown (5.1.4)

$$ \bar{X} \pm t_{1-\frac{\alpha}{2}, n-1} \left( \frac{s}{\sqrt{n}} \right) $$

Assumptions:

- The data must be a SRS.
- 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 are normal OR $n \geq 15$ for symmetric non-normal data OR $n \geq 30$ for skewed data).

$t$ Distribution (5.1.2) $T \sim t(df)$

- The $t$ distribution is symmetric, unimodal, bell-shaped, and centered at zero.
- The $t$ distribution has heavier tails than the normal distribution because $s$ (an estimate of $\sigma$) is used instead of $\sigma$.
- As the degrees of freedom ($df$) increases, the $t$ distribution approaches the standard normal distribution $N(0,1)$.

Notation:

- the margin of error is $m = t_{1-\frac{\alpha}{2}, n-1} \left( \frac{s}{\sqrt{n}} \right)$.
- Use the $t$-table (Table B.2 on p430 of your textbook) 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 $\bar{X}$

EXEMPLARY #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 $\mu$, the true mean vitamin C content of corn soy blend.
Confidence Interval for $p$

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

Assumptions:

- The data must be a SRS.
- If you sample a finite population without replacement, then $0.05N \geq n$.
- The sampling distribution for $\hat{p}$ must be approximately normal (so $n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$)

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 $\hat{p}$ is approximately normal?

- Give a 99% CI 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 CIs.

**Sample Size Calculation for Estimating \( \mu \):**

\[
\begin{align*}
n &= \left( \frac{z_{1 - \alpha/2} \sigma}{m} \right)^2
\end{align*}
\]

- \( z_{1 - \alpha/2} \) is the critical value for the desired confidence level \( C = 1 - \alpha \).
- \( m \) is the desired margin of error
- \( \sigma \) is the standard deviation of the population

If \( \sigma \) is unknown, two options for estimating \( \sigma \) are:

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

**EXAMPLE #5:**
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 \( p \):**

\[
\begin{align*}
n &= p(1 - p) \left( \frac{z_{1 - \alpha/2}}{m} \right)^2
\end{align*}
\]

Two options for the value of \( p \):

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

**EXAMPLE #6:**
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?
Estimating $\mu$ after transforming data

For large sample sizes ($n \geq 30$), we do not need to assume that the data $\{x_i\}$ are normal in order to find a CI for $\mu_X$. But when we have a small sample ($n < 15$) from a population which is clearly non-normal, then a “transform to normality”, such as the Box-Cox family of transforms $Y = X^\lambda$ (in Chapter 4-6, Normal Distribution notes). In many cases, one can not directly interpret the point estimate $\bar{y}$ or a CI for $\mu_Y$.

1. **Box-Cox transform $Y = X^\lambda$ with $\lambda \neq 0$**
   - The mean $\mu_X$ is always estimated by $\bar{X}$, but you may want to use medians as a measure of center due to skew or the presence of outliers.
   - $\mu_X$ is estimated by $\bar{Y}^{1/\lambda}$ if $\frac{s^2_X}{n}$ is “small”. This is because $\mu_Y^{1/\lambda} \approx \mu_X \left(1 + \frac{\lambda(\lambda - 1)}{2} \frac{s^2_Y}{\mu_X^2} \right)^{1/\lambda}$ (by the Delta Method).
   - The median $\tilde{\mu}_X$ is estimated by $\bar{Y}^{1/\lambda}$ if $Y$ is normal. This is because medians transform to medians.
   - The back-transformed interval
     \[
     \left( \left( \bar{Y} - t_{1 - \frac{\alpha}{2}, n-1} \frac{s_y}{\sqrt{n}} \right)^{1/\lambda}, \left( \bar{Y} + t_{1 - \frac{\alpha}{2}, n-1} \frac{s_y}{\sqrt{n}} \right)^{1/\lambda} \right)
     \]
     is a $100(1 - \frac{\alpha}{2})$ CI for $\mu_Y^{1/\lambda}$.
     - is a $100(1 - \frac{\alpha}{2})$ CI for the median $\tilde{\mu}_X$ if $Y$ is normal.
     - is an approximate $100(1 - \frac{\alpha}{2})$ CI for $\mu_X$ when $\frac{s^2_X}{n}$ is “small”
   - When $\lambda < 0$, you’ll need to swap the endpoints of the CI

2. **Box-Cox transform $Y = \ln(X)$ ($\lambda = 0$)**
   - The mean $\mu_X$ is always estimated by $\bar{X}$, but you may want to use medians as a measure of center due to skew or the presence of outliers.
   - $\mu_X$ is estimated by $\exp(\bar{Y})$ if $\frac{s^2_X}{n}$ is “small”. This is because $\exp(\mu_Y) \approx \mu_X \left( \exp \left( -\frac{s^2_Y}{2\mu_X^2} \right) \right)$ (by the Delta Method).
   - The median $\tilde{\mu}_X$ is estimated by $\exp(\bar{Y})$ if $Y$ is normal. This is because medians transform to medians. By the rules of logs, $\exp(\bar{Y}) = (\Pi i x_i)^{1/n}$, which is the geometric mean of the untransformed data.
   - The back-transformed interval
     \[
     \left( \exp \left( \bar{Y} - t_{1 - \frac{\alpha}{2}, n-1} \frac{s_y}{\sqrt{n}} \right), \exp \left( \bar{Y} + t_{1 - \frac{\alpha}{2}, n-1} \frac{s_y}{\sqrt{n}} \right) \right)
     \]
     - is a $100(1 - \frac{\alpha}{2})$ CI for $\exp(\mu_Y)$, the population geometric mean of $X$.
     - is a $100(1 - \frac{\alpha}{2})$ CI for the median $\tilde{\mu}_X$ if $Y$ is normal.
     - is an approximate $100(1 - \frac{\alpha}{2})$ CI for $\mu_X$ when $\frac{s^2_X}{n}$ is “small”
EXAMPLE #7:
For the timber volume data from Chapter 3 notes, we transformed volumes from \( n = 49 \) trees using the Box-Cox transform \( Y = (Volume)^{1/4} \).

- Find a point estimate for \( \mu_Y \).

- Find a point estimate for \( \mu_X \). Why is this a valid estimate?

- Find a point estimate for the median \( \tilde{\mu}_X \).

- Find a 95% CI for \( \mu_Y \).

- Find a 95% CI for \( \mu_X \). Why is this a valid CI?

- Find a 95% CI for the median \( \tilde{\mu}_X \).
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)

# 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

# EXAMPLE 7
> d=read.table("http://www.math.montana.edu/parker/courses/STAT401/data/timbervolume.txt",
header=TRUE)
> attach(d)
> Y=Volume^(1/4)  # Box-Cox transform
# Calculate statistics on original data X
> mean(Volume)
[1] 12.01837
> sd(Volume)
[1] 10.03248
> length(Volume)  # Sample size
[1] 49
> summary(Volume)
     Min.  1st Qu.   Median     Mean   3rd Qu.     Max.  
 0.7000  5.2000   8.3000  12.0200  17.1000  44.8000

# Calculate statistics on transformed data Y
> mean(Y)
[1] 1.739392
> mean(Y)^4
[1] 9.153556
> sd(Y)
[1] 0.3984408
> summary(Y)  # 5 number summary
     Min.  1st Qu.   Median     Mean   3rd Qu.     Max.  
0.91470  1.51000  1.69700  1.73900  2.03400  2.58700

# Check if mean(Y)^4 = 9.15 estimates the population mean of X
> sd(Volume)^2/mean(Volume)^2
[1] 0.6968284

# 95% CIs
# Not useful bc the Vol’s are not normal
> mean(Volume) + c(-1,1)*qt(.975,48)*sd(Volume)/sqrt(length(Volume))

# CI for mu_Y, the true mean of the Volume^(1/4)
> mean(Y) + c(-1,1)*qt(.975,48)*sd(Y)/sqrt(length(Y))
[1] 1.624946 1.853838

# CI for the true median Volume
> (mean(Y) + c(-1,1)*qt(.975,48)*sd(Y)/sqrt(length(Y)))^4

**Exercises**

Estimating means on p206: 4.7 and 4.9; p257: 5.1, 5.5
Estimating proportions on p207: 4.11 - 4.15 odd; p313: 6.5 - 6.13 odd
Sample size calculations on p259: 5.13; on p316: 6.19, 6.21