TELFER

VOTRE LIEN AVEC CE QUI COMPTE — CONNECTS YOU TO WHAT MATTERS

Estimating and Testing a Mean

ADM 2304

Winter 2016

Review of Sampling Distributions

- If a population has mean μ and stdev σ, then the sample mean (X-bar) from a “large” sample will have a normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$.
- Provided the original population is not extremely skewed, a sample size of $n > 30$ is “large” enough.

Probability Intervals

- The probability that **X-bar** falls inside

  $\boldsymbol{\mu \pm z_{\alpha/2} * \sigma/\sqrt{n}}$, is $1 - \alpha$.

  (meaning $100(1 - \alpha)\%$ of all sample means will fall inside this interval)
- *(μ - $\mathbf{z_{\alpha/2}}$ * $\boldsymbol{\sigma/\sqrt{n}}$, μ + $\mathbf{z_{\alpha/2}}$ * $\boldsymbol{\sigma/\sqrt{n}}$) is a fixed **probability interval**, because there is a fixed probability that the sample mean falls inside it, for given fixed values of μ **and** σ.*

Confidence Interval for μ

- **X-bar** $\pm$ **$z_{\alpha/2}$** * **$\sigma/\sqrt{n}$** is a $100(1 - \alpha)$ % **confidence interval**. It covers the mean μ if and only if **X-bar** falls inside the probability interval: $\mu \pm z_{\alpha/2} * \sigma/\sqrt{n}$.
- This confidence interval can be calculated only when sample data are available to calculate the sample mean. However, it assumes that σ is known, even though μ may be unknown.

Confidence Interval for a (population) mean (σ *known*)

The 100(1-α)% confidence interval is:

$$\bar{X} \pm z_{\alpha/2} * \sigma/\sqrt{n},$$

where $z_{\alpha/2}$ is the upper α/2 point of the standard normal distribution.

(For $n \le 30$, the original population must be approximately normally distributed as the CLT does not apply;

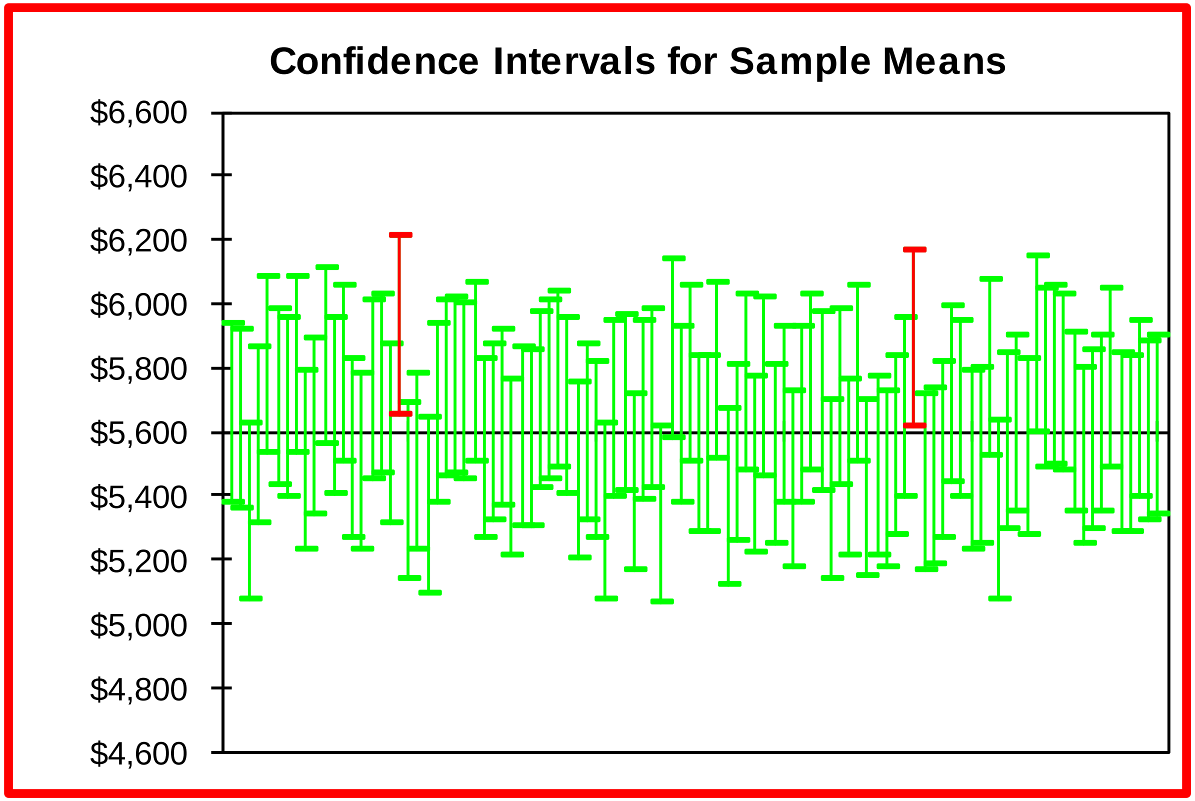
If the original population is not extremely skewed, then $n > 30$ is large enough for the CLT to apply.)

If n/N > .05, the Finite Population Correction (FPC) = $\sqrt{[(N-n)/(N-1)]}$ *should be used to adjust the standard deviation.*

Sampling Interpretation

- For example, if we define an interval within 2 standard deviations on either side of **X-bar**, we are 95% “confident” the interval will cover **μ** because 95% of **all such intervals** will cover the true population mean **μ**.
- Demo using *conf.xlt* to show the “sampling interpretation of confidence intervals”.

Simulation of Confidence Intervals



True Mean	$5,600
Sigma	1000
# of Samples	100
Sample Size	50

Confidence Interval
95.0 %

Size of Interval
xbar (+ or -) $277.18

Alpha	$Z_{(1-alpha/2)}$
5.0%	1.960

Generate Random Sample Means

% of Intervals That Include the True Mean **98%**

What if the variance is not known?

- We replace the unknown $\boldsymbol{\sigma}$ by its estimate s, the sample standard deviation;
- For small samples, the critical value $\mathbf{z_{\alpha/2}}$ must be replaced by a similar value from the t-distribution;
- This is symmetric like the normal distribution but with more probability in the tails; however, it approaches the normal shape as n gets large (meaning $n > 30$).

t-distribution

- For different sample sizes n, the value of $\mathbf{t}_{\alpha/2}$ allows for more uncertainty than the value of $z_{\alpha/2}$ does. For example, where the confidence coefficient is 95%,

n=5 uses d.f.=4	$\mathbf{t}_{\alpha/2}$ =2.776,
n=10 uses d.f.=9	$\mathbf{t}_{\alpha/2}$ =2.262,
n=20 uses d.f.=19	$\mathbf{t}_{\alpha/2}$ =2.093,
n=30 uses d.f.=29	$\mathbf{t}_{\alpha/2}$ =2.045,
n infinite	$\mathbf{t}_{\alpha/2} = \mathbf{z}_{\alpha/2} = 1.96$.

- n-1 represents the degrees of freedom for the sample standard deviation **s** (that's why we divide by n-1 in the formula for **s**);
- When the number of degrees of freedom is large, the distribution of **t** approaches the standard normal **z** distribution (see bottom row of Table T on page A-58 in Appendix B or the t-table tab of “ProbabilityTables.xls”.

Confidence Interval for a (population) mean (σ *unknown*)

The 100(1-α)% confidence interval is:

$$\bar{X} \pm t_{\alpha/2} * (s/\sqrt{n}) * FPC,$$

where $t_{\alpha/2}$ is the upper $\alpha/2$ point of the t- distribution with n-1 degrees of freedom.

For $n \leq 30$, the original population must be approximately normally distributed;

For $n > 30$, the original population must not be extremely skewed and $t_{\alpha/2} \cong z_{\alpha/2}$

Example

- Take a sample of male student heights and estimate the average male height using a 95% confidence interval.

Using Minitab

- Click on the **Stat** Menu, and select **Basic Statistics.**
- Select the **1-sample t** choice (only select **1-sample z** if the population standard deviation “sigma” is known).
- You can specify a column that contains the individual data values or specify the sample size **n**, sample mean **X-bar** and sample standard deviation **s**. Note that the interval assumes an infinite population.
- **Options** -- specify confidence level (default is 95%). For a 2-sided confidence interval, leave the default “**not equal**” “alternative”; for 1-sided confidence interval, specify the “less than” or “greater than” alternative, depending on which alternative you believe.

Sample Size Determination for Estimating a Population Mean

- To determine the sample size requirement for a survey to estimate a population mean, we need to know or to guess at the variability of the data (as measured by the standard deviation σ), as well as to determine the confidence level $(1 - \alpha)$ and the margin of error ($\pm$ M).
- Since $\pm \mathbf{z}_{\alpha/2} * \sigma / \sqrt{n} = \pm M$, solve for $n = (\mathbf{z}_{\alpha/2} * \sigma / M)^2$; we use $\mathbf{z}_{\alpha/2}$ instead of $t_{\alpha/2}$, in expectation of a sample size requirement greater than 30. If σ is not known, then we need to estimate it using the standard deviation s from a small sample.
- The more variable the data, the larger the sample size required to achieve a given margin of error for estimating the mean.

Example

- We wish to estimate the mean male student height using a 95% confidence interval with a margin of error of $\pm .5$ in.
- What sample size would be required?

Testing a Mean

We have the weight loss in pounds (negative values denote weight gains) for a random sample of 15 individuals who were selected to go on a special 4-week diet.

Setting the alternative

If we think (before seeing the data) that the diet is effective, then we should do a 1-sided test:

H_0: $\mu = \mu_0$ or $\mu \leq \mu_0$

H_a: $\mu > \mu_0$

For this example, the hypothesized value μ_0 is zero, and we have:

H_0: $\mu = 0$ or $\mu \leq 0$

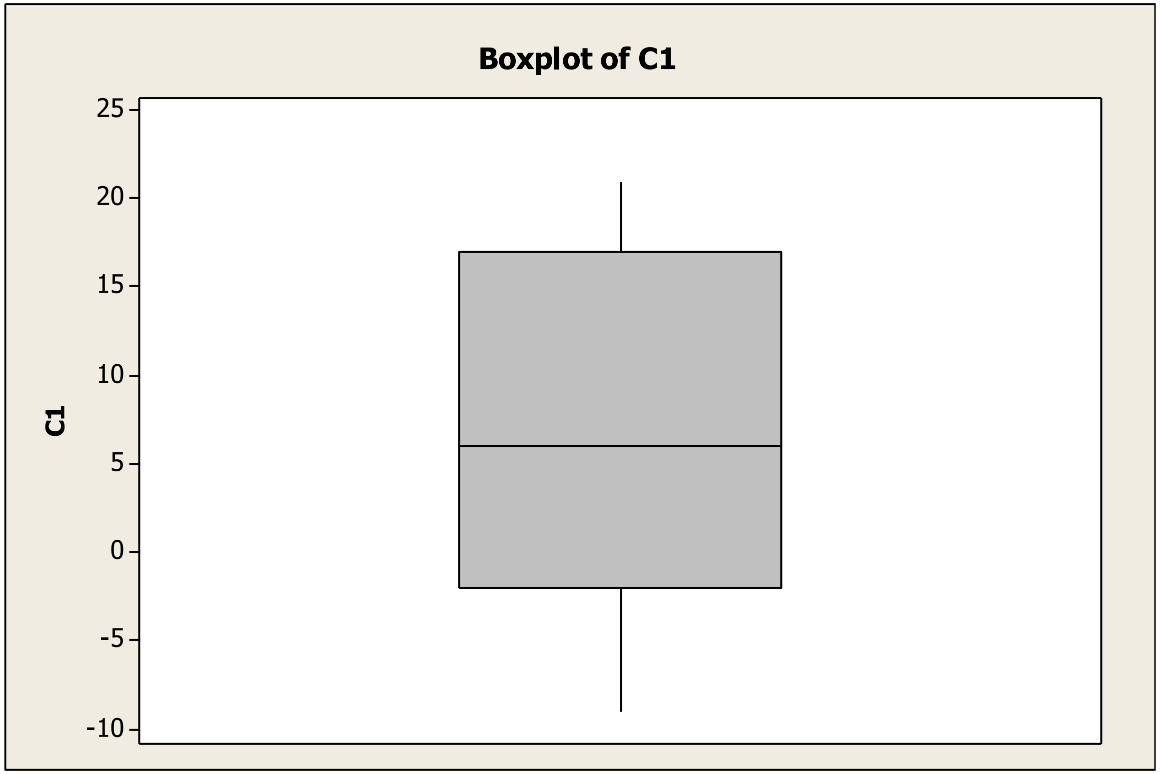
H_a: $\mu > 0$

Data

Suppose the fifteen weight losses are:

20	-8	6	20	15	-3	0	21
3	-2	7	9	-9	-2	17	

The boxplot shows the data can be assumed to come from an approximately normal distribution (since the data are symmetric and there are no outliers):



How do we proceed?

- To estimate the mean, we use the sample mean **x-bar**, based on a sample of n.
- If we assume the null hypothesis is true, then the sample mean has mean μ_0 and standard deviation $\sigma/\sqrt{n}$. Since the data come from a normal distribution, the sample mean is normally distributed. (If n > 30, then the sampling distribution is normal provided the population is not extremely skewed.)
- If the value of the population standard deviation σ is known, we can standardize the sample mean by calculating the z-statistic:

$$z_{stat} = (\text{x-bar} - \mu_0) / (\sigma/\sqrt{n}).$$

Decision Rule

- The rejection region is $z_{stat} > z_{\alpha}$, because only large positive values of z_{stat} argue for the alternative hypothesis and against the null hypothesis.
- Similarly the p-value is the probability that we would obtain a Z-value *larger* than the one we obtained in the sample, given that the null hypothesis is true, i.e. $P(Z > z_{stat})$.

What if σ is not known?

- If the value of the population standard deviation σ is not known, then we calculate the t-statistic:

$$t_{stat} = (\text{x-bar} - \mu_0) / (s/\sqrt{n}).$$

- The rejection region becomes $t_{stat} > t_\alpha$ and we look up the t-distribution with n-1 degrees of freedom. The p-value is $P(\ t > t_{stat}\ )$.

2-sided test

A 2-sided test would be reasonable if, prior to seeing the data, we think the diet does something, but we are not sure whether it results in a weight loss or a weight gain:

H_0: $\mu = 0$

H_a: $\mu \neq 0$ ($\mu < 0$ or $\mu > 0$)

2-sided test (cont'd)

- Here we use a 2-tailed rejection region because both very large positive values and very large negative values (the sign does not matter) of the t-statistic argue against the null hypothesis.
- For a test restricting the probability of the Type I error to alpha, we divide alpha by 2 and define a "2-tailed" or "2-sided" rejection region, each with probability $\alpha/2$ under the null hypothesis:

We reject the null hypothesis if

$$t_{stat} < -t_{\alpha/2} \quad \text{or} \quad t_{stat} > t_{\alpha/2} \,,$$

$$\text{or } |\, t_{stat} \,| > t_{\alpha/2} \,.$$

p-value

- The p-value for a 2-sided test is

  $P(t < -|t_{stat}| \text{ or } t > |t_{stat}|)$

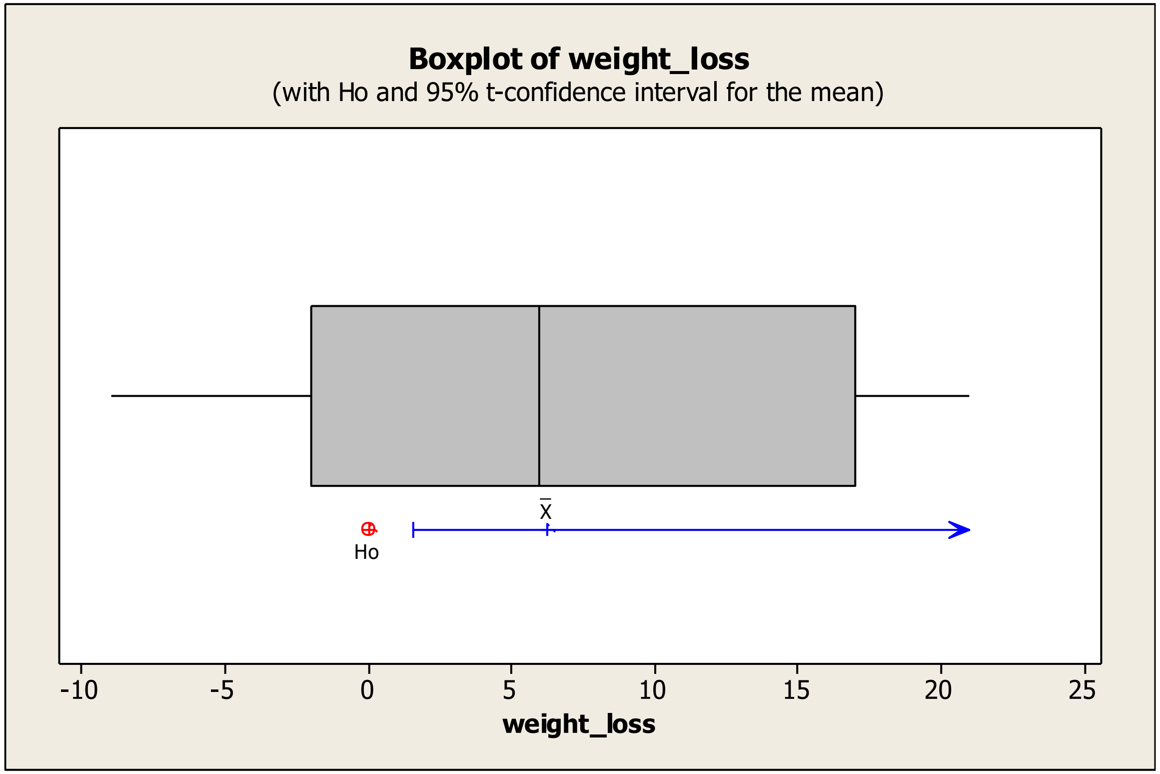
  $= 2\,P(t > |t_{stat}|) \text{ or } 2\,P(t < -|t_{stat}|)$

  since the sign of the t-statistic does not matter.
- Only the absolute value counts as a measure of how far the sample result is from 0 and we calculate the probability that the t distribution is more extreme than $|t_{stat}|$ in both the positive and the negative direction.

Using Minitab

- Go to the **Stat** Menu, select **Basic Statistics** and **1-sample t**
- (select **1-sample z** if the population standard deviation is known)

The 1-sided confidence interval
“at least 1.55477” is shown in blue; it does not cover the hypothesized mean of 0, as indicated in red.



One-Sample T: weight_loss

Test of mu = 0 vs > 0

Variable	N	Mean	StDev	SE Mean	95% Lower Bound	T	P
weight_loss	15	6.26667	10.36110	2.67522	1.55477	2.34	**0.017**

One-Sample T: weight_loss

Test of mu = 0 vs < 0

Variable	N	Mean	StDev	SE Mean	95% Upper Bound	T	P
weight_loss	15	6.26667	10.36110	2.67522	10.97857	2.34	**0.983**

One-Sample T: weight_loss

Test of mu = 0 vs not = 0

Variable	N	Mean	StDev	SE Mean	95% CI	T	P
weight_loss	15	6.26667	10.36110	2.67522	(0.52888, 12.00445)	2.34	**0.034**

If we take the perspective of the weight loss clinic, then the appropriate test is the first one since we would expect a positive weight loss if the diet is effective. The second test is not necessary since the positive z-statistic is clearly not inconsistent with the null hypothesis (mu ≥ 0). If we did not know what to expect, then the third test might be appropriate.