

## 22S:105

## Statistical Methods and Computing

Sample size for confidence intervals  
with  $\sigma$  known  
 $t$  Intervals

Lecture 13  
Mar. 7, 2011

Kate Cowles  
374 SH, 335-0727  
kcowles@stat.uiowa.edu

Sample size for a study involving a confidence interval

- Suppose a group of obstetricians wish to carry out a study to estimate  $\mu$ , the mean birth-weight in the population of infants born at UIHC.
- Suppose the obstetricians believe that the population standard deviation of birthweights of infants born at UIHC is the same as that of infants overall in the US.

$$\sigma = 15 \text{ oz}$$

- The obstetricians would like a 95% confidence interval for  $\mu$  that is no wider than 4 ounces. That is, they want a margin of error  $\leq 2$  ounce.
- How many infants do they need in their study?

## The margin of error

- The **margin of error** is the value that we add onto  $\bar{x}$  and subtract from  $\bar{x}$  to get the endpoints of a confidence interval.
- For confidence intervals for the mean of a normal population with  $\sigma$  known, this is

$$m = z^* \frac{\sigma}{\sqrt{n}}$$

- Equivalently, the margin of error is one half the width of the c.i.
- The margin of error depends on
  - the level of confidence desired
  - the population standard deviation (which we can't control!)
  - the sample size (*not* the population size)

- Let  $m$  denote the margin of error. Then

$$\begin{aligned} m &= z^* \frac{\sigma}{\sqrt{n}} \\ \sqrt{n} &= z^* \frac{\sigma}{m} \\ n &= \left( z^* \frac{\sigma}{m} \right)^2 \\ n &= \left( 1.96 * \frac{15}{2} \right)^2 \\ &= 216.09 \end{aligned}$$

- A sample size must always be rounded *up*, so they need 217 infants in their study.

## Sample size continued

What makes a sample size large?

$$n = \left( z^* \frac{\sigma}{m} \right)^2$$

get help from a statistician on computing measures of center and intervals that are not sensitive to outliers.

- Check your data for skewness and other signs that the population they came from may not be normal. If the sample size is large (i.e.  $n \geq 30$ ) the central limit theorem says the approach is valid. If the sample size is small, the confidence level will not be correct.
- The formula  $\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$  requires that we know the exact value of the population standard deviation  $\sigma$ , which we never do.

\* Moore, David S. (2000) *The Basic Practice of Statistics*, 2nd ed., W.H. Freeman and Co.

## Caveats regarding our formula for computing confidence intervals for population means

- The data must be a *simple random sample* from the population.
  - We are not in too big trouble if the data can plausibly be thought of as observations taken at random from the population.
- “There is no correct method for inference from data haphazardly collected with bias of unknown size. Fancy formulas cannot rescue badly produced data.”\*
- Watch out for outliers in your dataset, because they can have a large effect on both the point estimate of  $\mu$  and the confidence interval.

If outliers are not data errors, and if there is no subject-matter reason for deleting them,

## What to do when we believe the population is normal but we don't know $\sigma$

Assumptions behind this method

- The data are a *simple random sample* from the population of interest.
- Values in the population follow a *normal distribution* with mean  $\mu$  and standard deviation  $\sigma$ . Both  $\mu$  and  $\sigma$  are unknown.

The sample mean  $\bar{x}$  is still our *point estimate* of the unknown population mean  $\mu$ .

$\bar{x}$  still comes from a normal distribution with mean  $\mu$  and standard deviation  $\frac{\sigma}{\sqrt{n}}$ .

- We will *estimate*  $\sigma$  by the sample standard deviation  $s$ .
- Then we estimate the standard deviation of  $\bar{x}$  by  $\frac{s}{\sqrt{n}}$

## Standard errors

When we use the data to estimate the standard deviation of a *statistic*, the result is called the *standard error* of the statistic.

The standard error of the sample mean  $\bar{x}$  is  $\frac{s}{\sqrt{n}}$ .

When we are *estimating*  $\sigma$  with  $s$ , we need to make our confidence interval *wider* to account for the uncertainty in estimation.

- (What if we had gotten a sample that happened to give a sample standard deviation  $s$  that was much smaller than the population standard deviation  $\sigma$ ?)
- We do this by multiplying  $\frac{s}{\sqrt{n}}$  by something *bigger* than  $z^*$ .

## $t$ intervals

When we claimed to know  $\sigma$ , we computed confidence intervals for  $\mu$  as

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$$

where  $z^*$  was the appropriate cutoff value from a standard normal distribution.

When we don't know  $\sigma$ , we will compute confidence intervals for  $\mu$  as

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$$

## The $t$ distribution

- There is a different  $t$  distribution for every sample size.
  - We identify different  $t$  distributions by their *degrees of freedom*,  $n - 1$ .
- The density curve for  $t$  distributions is
  - symmetric around 0
  - bell-shaped (and has only one mode)
- The spread of  $t$  distributions is greater than the spread of the standard normal distribution.
  - The smaller the degrees of freedom, the more spread out the  $t$  distribution is.
  - The larger the degrees of freedom, the closer the density curve for a  $t$  distribution is to a standard normal curve.

- \* This makes sense because the larger the sample size, the better an estimate  $s$  is likely to be for  $\sigma$  (i.e., the less extra uncertainty is introduced by estimating  $\sigma$  instead of knowing its value)

### More on the $t$ distribution

If  $\bar{x}$  is the sample mean of a simple random sample of size  $n$  value from a normal population with mean  $\mu$  and standard deviation  $\sigma$ , then the random quantity

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

follows a  $t$  distribution

### Example

- We have data on a simple random sample of 10 birthweights of infants born at Boston City Hospital.
- We wish to estimate the mean  $\mu$  of birthweights in the population served by this hospital.
- This population may be different from the population of all US birthweights, so we cannot assume that we know either  $\mu$  or  $\sigma$ .

### Constructing confidence intervals for $\mu$ when $\sigma$ is unknown

To construct a level  $C$  confidence interval for  $\mu$

- Draw a simple random sample of size  $n$  from the population. The population is assumed to be normal.
- Compute the sample mean  $\bar{x}$  and the sample standard deviation  $s$ .
- Then the level  $C$  confidence interval for  $\mu$  is

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$$

where  $t^*$  cuts off the upper  $\frac{1-C}{2}$  area under the density curve for a  $t$  distribution with  $n - 1$  degrees of freedom.

- Use Table A.2 at the back of your textbook to find  $t^*$ .

- Our data values are:

Infant	Birthweight in ounces
1	97
2	117
3	140
4	78
5	99
6	148
7	108
8	135
9	126
10	121

First calculate

$$\bar{x} = 116.90 \quad s = 21.70$$

The degrees of freedom are  $10 - 1 = 9$ . For a 95% confidence interval, we need the value of  $t^*$  that cuts off an area of .025 in the upper tail.

From Table C, we find  $t^* = 2.262$ .

Our confidence interval is

$$\begin{aligned} \bar{x} \pm t^* \frac{s}{\sqrt{n}} &= \\ 116.90 \pm 2.262 \frac{21.70}{\sqrt{10}} &= \\ 116.90 \pm 15.22 &= (101.38, 132.42) \end{aligned}$$

The interval is so wide because of

- the relatively small sample size
- the relatively large variation between birth-weights (large  $s$ )