

## Estimating a population mean

- We use  $\bar{y}$  as an estimator of  $\mu$ . Is it a 'good' estimator?
- An estimator is 'good' if:
  - It is **unbiased**
  - It has **small standard error**.
- An estimator is *unbiased* if the mean of its sampling distribution equals the parameter we are trying to estimate.
  - $\bar{y}$  is unbiased for  $\mu$  because  $E(\bar{y}) = \mu_{\bar{y}} = \mu$ .
- In English: if we were to draw 100 samples of size  $n$  from some population with mean  $\mu$ , and were to compute  $\bar{y}$  in each of the 100 samples, the *average* of those 100  $\bar{y}$  would be close to  $\mu$ .

## Population mean (cont'd)

- Recall that if  $y \sim (\mu, \sigma^2)$ , then the *sampling distribution* of  $\bar{y}$  is  $N(\mu, \sigma^2/n)$ .
- As  $n$  increases,  $\sigma^2/n$  decreases: the larger the sample, the more reliable will  $\bar{y}$  be as an estimator of  $\mu$ .
- The parameter  $\frac{\sigma}{\sqrt{n}}$  is called *standard error of the mean* and is estimated by  $S/\sqrt{n}$ .
- If  $\bar{y} \sim N(\mu, \sigma^2/n)$  then

$$\text{Prob}(\bar{y} - 2\frac{\sigma}{\sqrt{n}} < \mu < \bar{y} + 2\frac{\sigma}{\sqrt{n}}) \approx 0.95$$

(exactly equal to 0.95 if we use 1.96 instead of 2).

## Confidence intervals

- Since  $\bar{y}$  will fall within  $\pm 2\sigma/\sqrt{n}$  of the population mean  $\mu$  approximately 95% of the time, then the interval

$$\bar{y} - 2\frac{\sigma}{\sqrt{n}} \text{ to } \bar{y} + 2\frac{\sigma}{\sqrt{n}}$$

*will cover  $\mu$*  about 95% of the time in repeated sampling.

- **100(1- $\alpha$ )% confidence interval for  $\mu$ :**

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

where  $z_{\alpha/2}$  is the  $z$  value with an area equal to  $\alpha/2$  to its right (see figure 1.18).

## Confidence intervals

- We can construct confidence intervals with any *confidence coefficient*  $(1 - \alpha)$ .
- For a 90% confidence interval, use 1.64 instead of 2 (or 1.96 to be precise), because:
  1.  $\alpha = 0.10$  and  $\alpha/2 = 0.05$
  2.  $z$ -value with 0.05 to its right (or 0.95 to its left) is 1.64 from standard normal table.
- For a 99% confidence interval, use 2.58 (or  $z_{0.005}$ ) instead of 2.
- Note: the wider the interval, the higher the confidence that it will cover  $\mu$ . Thus, a 99% confidence interval for  $\mu$  will always be wider than a 90% interval.

## Confidence intervals (cont'd)

- Example: Attention times given by parents to sets of twin boys during one week (Table 1.9, page 36).
- $n = 50$ ,  $\bar{y} = 20.85$  and  $S = 13.41$ .
- A 90% CI for the true mean attention time  $\mu$  is

$$\bar{y} \pm 1.64 \frac{S}{\sqrt{n}} = 20.85 \pm 1.64 \frac{13.41}{\sqrt{50}} = 20.85 \pm 3.11.$$

- 95% CI:  $\bar{y} \pm 2 \frac{S}{\sqrt{n}} = 20.85 \pm 2 \frac{13.41}{\sqrt{50}} = 20.85 \pm 3.80$ .
- 99% CI:  $\bar{y} \pm 2.57 \frac{S}{\sqrt{n}} = 20.85 \pm 2.57 \frac{13.41}{\sqrt{50}} = 20.85 \pm 4.88$ .

## Confidence intervals (cont'd)

- Note that we used the sample standard deviation  $S$  in place of the unknown population standard deviation  $\sigma$  to compute the CI.
- This is OK only if  $n$  is large enough (more than 30).
- If  $\sigma$  is unknown (as it usually is) and  $n < 30$  we compute the CI using  $t_{\alpha/2}$  instead of  $z_{\alpha/2}$  (Student's  $t$ -table instead of  $z$ -table).
- The value  $t_{\alpha/2}$  is the upper-tail  $t$ -value such that an area equal to  $\alpha/2$  lies to its right.

## Confidence intervals (cont'd)

- To get the appropriate value out of a  $t$ -table we need:
  1. The *degrees of freedom* =  $n - 1$  in this type of applications.
  2. The desired confidence coefficient  $(1 - \alpha)$ .
- For small  $n$ , a  $100(1 - \alpha)\%$   $CI$  for  $\mu$  is

$$\bar{y} \pm t_{\frac{\alpha}{2}, n-1} \frac{S}{\sqrt{n}}$$

## Confidence intervals (cont'd)

- Example 1.12, page 39: Concentrations of silica in ppm in treated saline water.
- $n = 5$  (small!),  $\bar{y} = 239.2$ ,  $S = 29.3$  and  $df = 4$ .
- For a 95% CI for the true silica concentration:  $t_{\frac{\alpha}{2},4} = t_{\frac{0.05}{2},4} = 2.776$ .
- Then, the 95% CI for  $\mu$  is

$$239.2 \pm 2.776 \frac{29.3}{\sqrt{5}} = 239.2 \pm 36.4.$$

- If we had wished to obtain a 90% or a 99% CI for the mean, then the corresponding  $t$ -values (from the table) would have been 2.132 and 4.604, respectively (see Table C.2).

## Confidence intervals (cont'd)

- Continue with same example, and now ask the following question: what sample size would we have needed if we wished to estimate the true mean silica concentration to within 10 ppm with 95% confidence?
- We wish to know what  $n$  we would need if we wished to be able to state that

$$\text{Prob}(\bar{y} - 10 < \mu < \bar{y} + 10) = 0.95.$$

- Above means that

$$t_{0.025, n-1} \frac{S}{\sqrt{n}} = 10.$$

- From expression above, we need to solve for  $n$ .

## Confidence intervals (cont'd)

- We know that the desired  $n$  *must* be larger than 5 because with  $n = 5$  we estimated  $\mu$  to within 36.4 ppm with 95% confidence.
- We need the following in order to come up with an answer:
  - Assume that  $S$  would not change with increased  $n$
  - Approximate a value of  $t_{0.025, n-1}$  to be about 2
- Then: