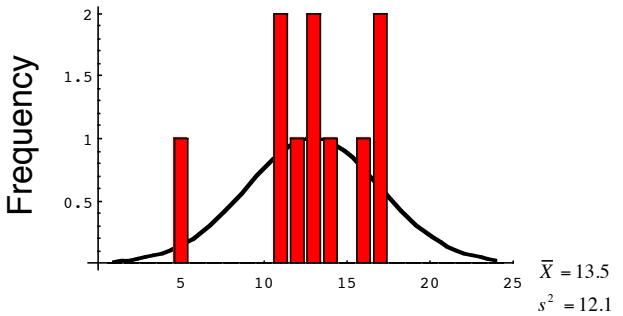
Estimating with uncertainty

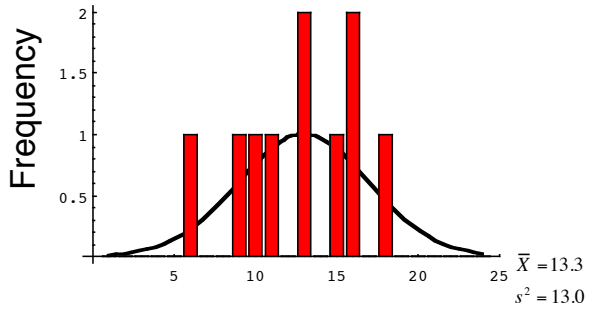
Chapter 4

Sample size 10 from Normal distribution with μ=13 and σ^2=16



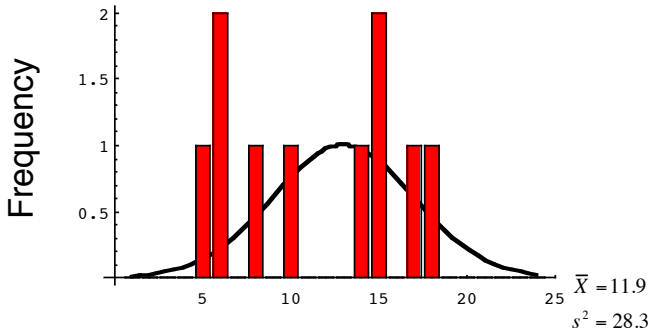
$\bar{x}$

Another sample of 10 from same distribution



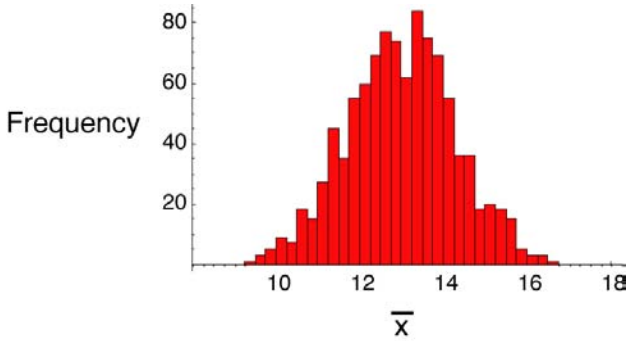
$\bar{x}$

A third sample of 10 from the same distribution

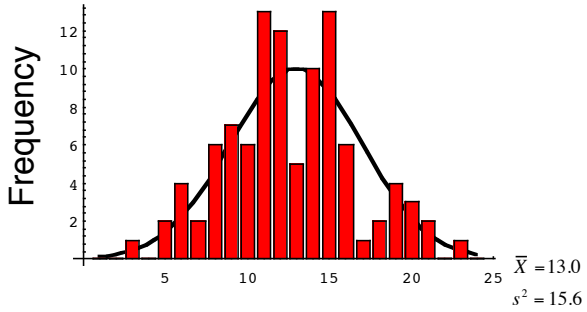


$\bar{x}$

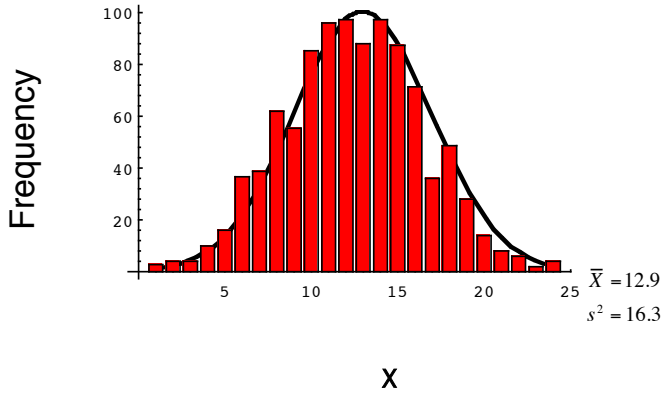
Distribution of the means of 1000 samples, each of sample size 10



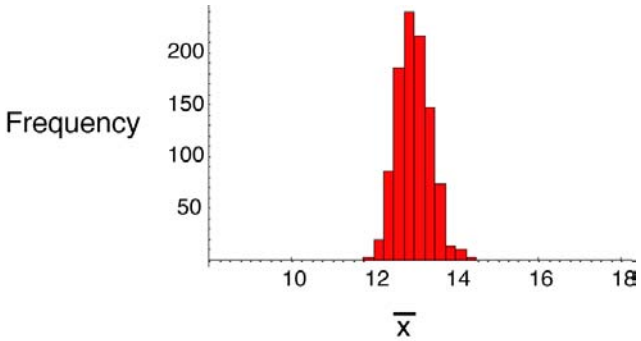
A sample of 100 from the same population distribution



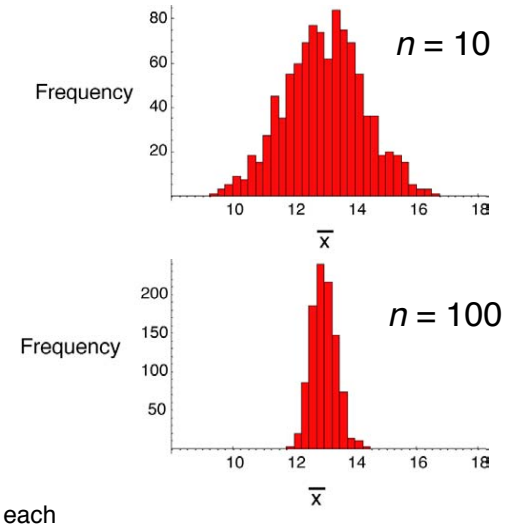
A sample of 1000 from the same population distribution



Distribution of the means of 1000 samples, each of sample size 100

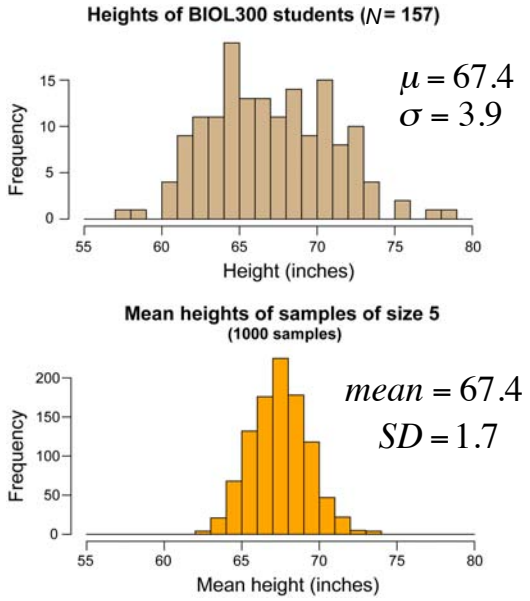


Variation in sample means decreases with sample size



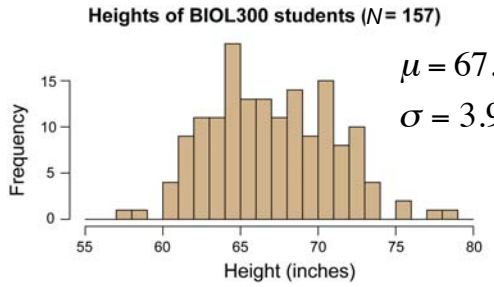
1000 samples each

The *standard error* of an estimate is the standard deviation of its sampling distribution. The standard error predicts the sampling error of the estimate.



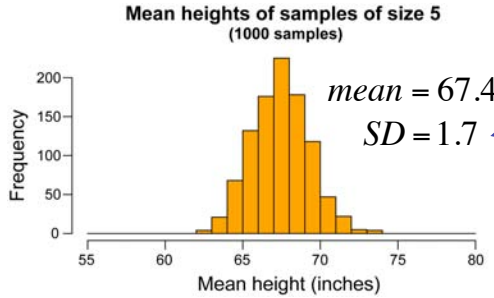
Standard error of the mean

$$\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}}$$



$$\mu_{\bar{Y}} = \mu = 67.4$$

$$\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}} = \frac{3.9}{\sqrt{5}} = 1.7$$



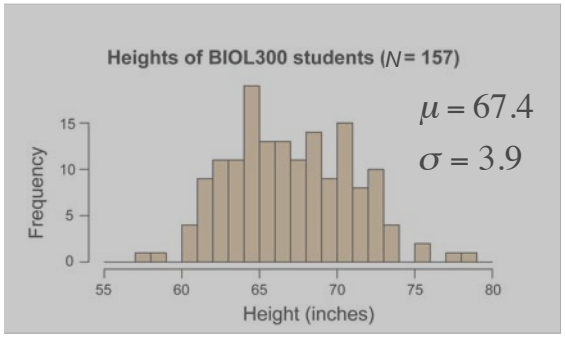
The math works!

The problem is, we rarely know σ.

Estimate of the standard error of the mean

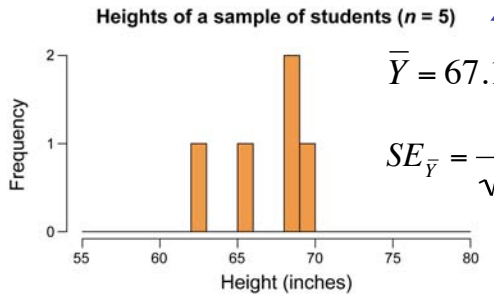
$$SE_{\bar{Y}} = \frac{s}{\sqrt{n}}$$

This gives us some knowledge of the likely difference between our sample mean and the true population mean.



In most cases, we don't know the real population distribution.

We only have a sample.



$$\bar{Y} = 67.1 \qquad s = 3.1$$

$$SE_{\bar{Y}} = \frac{s}{\sqrt{n}} = \frac{3.1}{\sqrt{5}} = 1.4$$

We use this as an estimate of $\sigma_{\bar{Y}}$

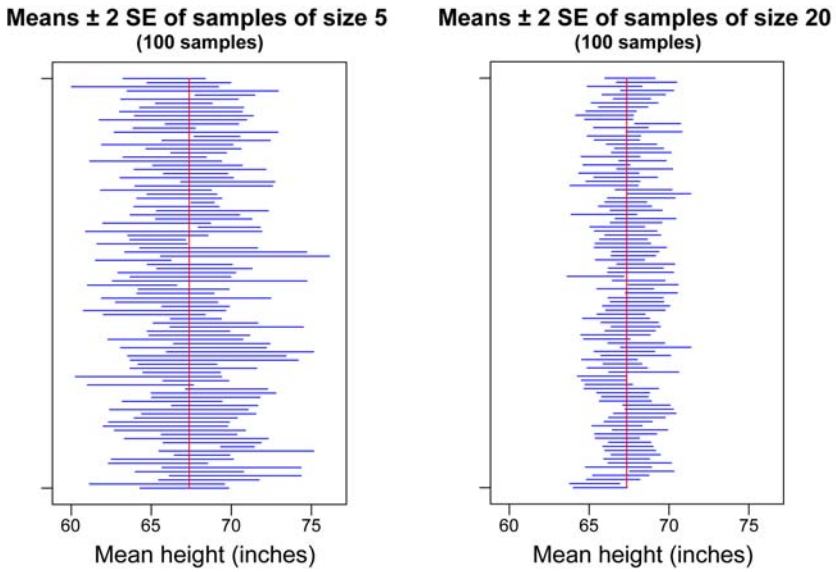
Confidence interval

The 95% confidence interval provides a plausible range for a parameter. All values for the parameter lying within the interval are plausible, given the data, whereas those outside are unlikely.

The 2SE rule-of-thumb

The interval from $\overline{Y} - 2\,SE_{\overline{Y}}$ to $\overline{Y} + 2\,SE_{\overline{Y}}$ provides a rough estimate of the 95% confidence interval for the mean.

(*Assuming normally distributed population and/or sufficiently large sample size.*)



Use correct language when talking about confidence intervals

Not correct:

"There is a 95% probability that the population mean is within a particular 95% confidence interval"

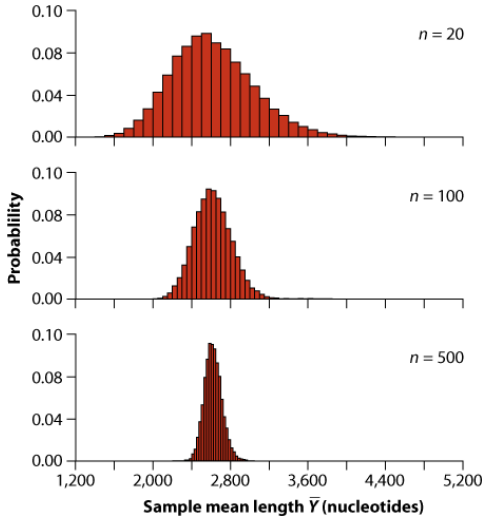
Correct:

"95% of all 95% confidence intervals calculated from samples include the population mean."

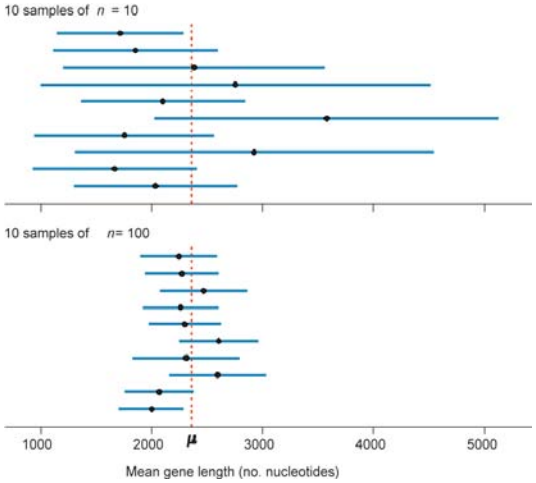
or:

"We are 95% confident that the population mean lies within the 95% confidence interval."

Sample means of gene sizes



Confidence interval



US counties with high kidney cancer death

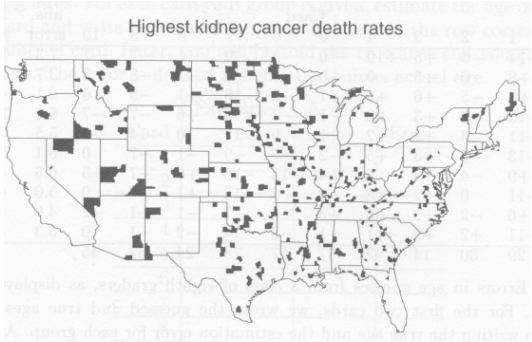


Fig. 2.3 The counties of the United States with the highest 10% age-standardized death rates for cancer of kidney/ureter for U.S. white males, 1980–1989.

US counties with low kidney cancer death

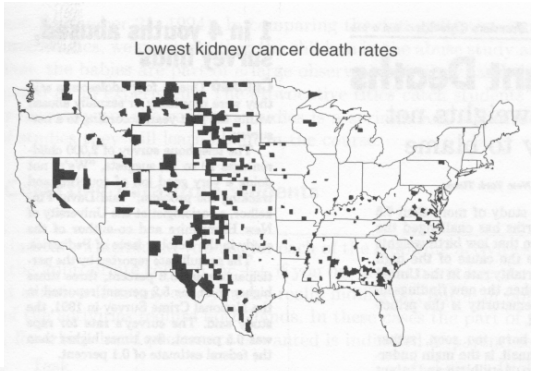


Fig. 2.4 The counties of the United States with the *lowest* 10% age-standardized death rates for cancer of kidney/ureter for U.S. white males, 1980–1989.

Variation in cancer rates decreases with population size of counties

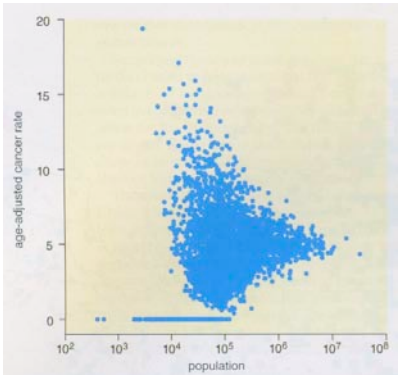


Figure 3. When age-adjusted kidney-cancer rates in U.S. counties are plotted against the log of county population, the reduction of variation with population becomes obvious. This is the typical triangle-shaped bivariate distribution.

Wainer (2007) The most dangerous equation. *American Scientist* 95: 249-256.

Pseudoreplication

The error that occurs when samples are not independent, but they are treated as though they are.

Example: Pseudoreplication

You are interested in average pulse rate of mountain climbers. Since they are hard to find, you decide to take 10 measurements from each climber. You study 6 climbers, so you have 60 measurements.

What is your sample size (*n*)*?*

Avoiding pseudoreplication

You are interested in average pulse rate of mountain climbers. Since they are hard to find, you decide to take 10 measurements from each climber. You study 6 climbers, so you have 60 measurements.

Take the mean blood pressure for each climber, so that you have 6 pulse rates, one for each climber ($n = 6$).