

Sampling Variability and Sampling Distributions

Example: What is the average weight of women 5'1" tall between the ages of 21-45? The American Medical Association takes a sample of 1000 women between the ages of 21-45 years and with height 5'1" and finds that the mean weight is $\bar{X} = 136.2$ lbs.

1

Parameters and Statistics

- A parameter is a number that describes the population of interest. Since we usually cannot examine the entire population of interest, parameters are generally unknown.
- A statistic is a number that is computed from sample data. We often use a statistic to estimate an unknown population parameter.

sample statistic and population parameter

2

Notation

- μ = population mean
(unknown)
 - \bar{X} = sample mean
(computed from the data we have on hand from a sample of the population)
-
- σ = population standard deviation
(unknown)
 - s = sample standard deviation
(computed from the data we have on hand from a sample of the population)

3

Sampling Variability

- The basic fact that the value of a sample statistic varies in (hypothetical) repeated random sampling is called sampling variability.
- Example: If another sample of 1000 women was chosen from the same population of 5'1" women between 21-45 years old, the value of \bar{X} would almost certainly be different – something other than 136.2 lbs.

4

Q: If our goal is to estimate the mean weight of the population, how should we deal with the fact that different samples yield different estimates of the mean weight??

A: Allow a margin of error that takes sampling variability into account.

5

Confidence Intervals

- Confidence intervals are generally of the form
point estimate \pm margin of error
- Q: Why should we estimate μ with an interval of numbers? Why not just use the point estimate as our estimate of μ ?
- A: (1) Using an interval estimate (i.e., \bar{X} confidence interval) takes sampling variability into consideration, and (2) we can attach a level of confidence to an interval estimate which we cannot do with a point estimate.

6

Confidence intervals, continued

- A confidence interval for μ has two parts:
 - 1) A margin of error says how close \bar{X} lies to μ .
 - 2) A level of confidence says what percent of all possible samples satisfy the margin of error.
- Example: High school students who take the SAT math exam a second time generally score higher than on their first try. A random sample of 1000 students gains an average of $\bar{X} = 22$ points on their second try. Let $\mu =$ mean gain in score in the population of all high school students.

7

SAT Example, continued

- Suppose the change in SAT score has a sample standard deviation of $s = 50$. Here are some confidence intervals for μ .
- 90%: (19.399, 24.601)
- 95%: (18.901, 25.099)
- 99%: (17.927, 26.073)

8

- When constructing a confidence interval, you must decide on the risk you are willing to take of being wrong.
- A confidence interval is “wrong” if it doesn’t contain the true value of the population parameter.
- 99% confidence ==> 1% chance of being wrong
- 95% confidence ==> 5% chance of being wrong
- 90% confidence ==> 10% chance of being wrong
- The choice of 95% is very common because it provides a good balance between precision and reliability.

9

How confidence intervals behave

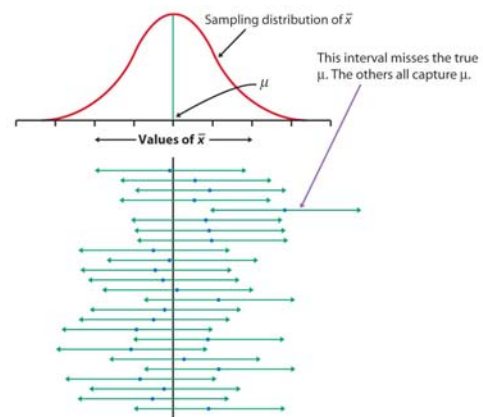
- High confidence says that our method almost always gives correct answers.
 - A small margin of error says that we have pinned down the parameter quite precisely. The margin of error determines the width of the confidence interval.
- 1) The margin of error is larger for higher confidence levels. To obtain a smaller margin of error from the same data, you must be willing to accept lower confidence.
 - 2) The margin of error is larger for smaller sample sizes.
 - 3) The margin of error is larger for populations that have lots of variability.

10

Interpreting confidence levels

- Take 95% confidence, for example.
- Practical Interpretation: We are 95% confident that the mean gain in score is between 18.9 and 25.1 points.
- Statistical Interpretation: If we repeatedly take random samples of 1000 from the population and construct 95% confidence intervals for each sample, then in the long run 95% of these confidence intervals will capture the true value of μ . Our sample is either one of the 95% for which the calculated interval captures μ , or one of the unlucky 5% that do not.

11



12

- To derive a formula for the margin of error that appropriately takes into account sampling variability, we must gain an understanding of how \bar{X} varies from one sample to the next.
- There is a pattern to the way the sample mean \bar{X} varies from sample to sample. Once this pattern of variability is understood, we can come up with the margin of error, and construct a confidence interval for the population mean that takes sampling variability into account and allows us to attach a level of confidence to our interval estimate.

13

Sampling Distribution

- The sampling distribution of a statistic is the distribution of values taken by the statistic in all possible samples of the same size from the same population.
- The sampling distribution of \bar{X} provides information about the behavior of \bar{X} in (hypothetical) repeated sampling.

14

The idea of sampling distribution

- Take many samples from the same population.
- Collect the \bar{X} 's from all the samples.
- Display the distribution of the \bar{X} 's (in a histogram, for example).
- The histogram will be bell-shaped and symmetric, centered at the population mean.
- The sampling distribution of \bar{X} is a normal distribution!

15

Facts about the sampling distribution of \bar{X}

- These facts describe how \bar{X} varies from one sample to the next:
 - 1) In repeated sampling, \bar{X} will sometimes fall above the true value of μ and sometimes below it, but there is no systematic tendency for \bar{X} to overestimate or underestimate μ . The sampling distribution of \bar{X} is centered at μ , and so \bar{X} is called an unbiased estimator of μ .
 - 2) The values of \bar{X} from larger samples are less variable than those from smaller samples. The standard deviation of the sampling distribution of \bar{X} is $\frac{\sigma}{\sqrt{n}}$.

16

Technical detail

When the population from which we sample can be modeled by a normal distribution, then

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

where μ is the mean of the population and σ is the standard deviation of the population.

More discussion later.....

17

Derivation of confidence interval for μ

$$\bar{X} \pm z_{.975} \frac{\sigma}{\sqrt{n}} \quad \bar{X} \pm t_{.975} \frac{s}{\sqrt{n}}$$

- We will derive a 95% confidence interval and then generalize the formula for any level of confidence.
- We will derive the formula under the following conditions:
 - 1) The population standard deviation σ is known, and
 - 2) The distribution of the population from which we sample can be modeled by a normal density curve.Note that condition (1) is unrealistic. We'll fix it at the end.

18