Confidence Intervals

Ryan Miller

Grinnell College
Statistics

Introduction

Lately, we've discussed Normal probability models, and we've seen how these can be applied to two different distributions:

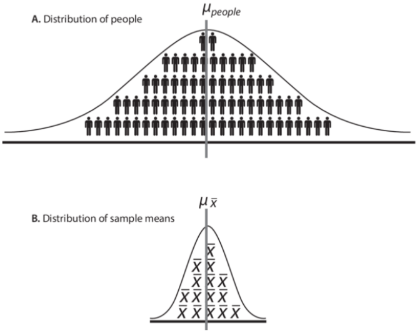


Image credit:
https://www.researchgate.net/publication/383141443_Introducing_prediction_intervals_for_sample_means

Grinnell College
Statistics

Vocabulary

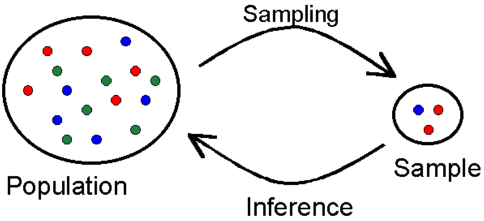
In an attempt to make their terminology less confusing/overloaded, statisticians use the following terms:

- **Standard deviation** describes the variability of *data-points* around an expected value (mean)
- **Standard error** describes the variability of *estimates* around an expected value (population parameter)
- **Sampling distribution** the distribution of *sample estimates*

Standard error is the standard deviation of the sampling, but it is given a distinct name because standard error formulas often involve the standard deviation of the cases in a sample

Statistical Inference

Statistical analyses attempt to generalize a statistic, such as the sample mean, to a broader population.



- Because sampling is a random process, any statistic calculated from a sample contains uncertainty
  - In any practical application, we never see the population, and we only get to see one sample, making Central Limit theorem a powerful tool in understanding this uncertainty

Point vs. Interval Estimates

Statisticians consider two types of estimates for an unknown population parameter:

1) **Point estimate** - a *single number* that is the *best guess* for what the population parameter is. For example, the sample mean $\overline{x}$ is a point estimate for the population's mean, μ
2) **Interval estimate** - a *range of numbers* that represent *plausible values* of the population parameter. Interval estimates usually have the form: Point Estimate $\pm$ Margin of Error

Questions

1. What are the chances that a point estimate, such as the sample mean, *exactly matches* the corresponding parameter in the population? *Hint*: think about the probability of a continuous random variable taking on any particular value.
2. For a given sample, how could you create an interval estimate that will *always* contain the population parameter of interest?

Grinnell College
Statistics

Confidence Intervals

- A **confidence interval** is an interval estimate whose margin of error is calculated using procedure with a long-run "success rate" known as a *confidence level*
  - A 95% confidence interval uses a procedure that will succeed in containing the true population parameter in 95% of different random samples (or study replications)

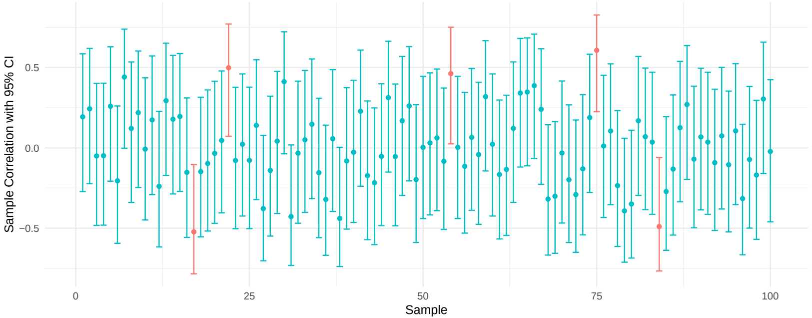
Confidence Intervals

Confidence intervals have the following properties:

- They are made with the intention of giving a plausible range of values for a population parameter
- They are the observed result of a random process (ie: sampling)
- Before that random process has been observed (ie: before we've collected a sample) the procedure used to calculate the interval has a chance of containing population parameter defined as the confidence level
    - After the random process has unfolded (ie: after we've collected our sample) the interval is no longer random, it is either correct (contains the truth) or incorrect (doesn't contain the truth)

Confidence Intervals

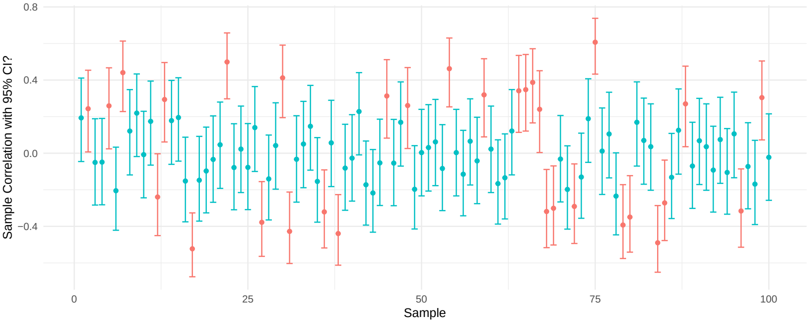
Shown below are 95% CI estimates from 100 different random samples ($n = 20$) drawn from a population with correlation of $\rho = 0$



Notice that 95 of 100 samples produced a 95% CI estimate containing the true population-level correlation. This suggests the method used to form these intervals is a *valid* 95% confidence

Not Confidence Intervals

Suppose we use a different method to calculate interval estimates for the same set of 100 randomly selected samples (of size $n = 20$):



Why isn't this a valid 95% confidence interval procedure?

Finding Confidence Interval Estimates

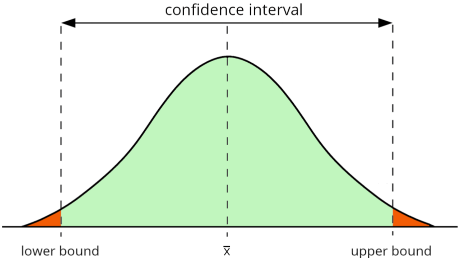
The challenge in calculating a confidence interval is properly calibrating the interval's margin of error to achieve the stated confidence level:

$$\text{Point Estimate} \pm \text{Margin of Error}$$

- If we calculate margins of error that are *too small*, the intervals will not contain the true population parameter often enough to achieve the desired confidence level
  - This increases the chances of *false positive* findings
- If we calculate margins of error that are *too large*, we are expressing an unnecessary amount of uncertainty
  - This increases the chances of *false negative* findings

Central Limit Theorem and Confidence Intervals

- We know that the sample mean follows a Normal distribution given by the Central Limit theorem
  - This distribution tells us how much sampling variability to expect when using the sample mean as an estimate
- For example, if we know that 95% of sample means are within a distance of 1.4 units from the center of the distribution, we can use 1.4 as the margin of error for a 95% confidence interval



Two Ways to use CLT

Approach #1 - Use the sample data to directly estimate the distribution of sample means and take the middle $P\%$ as the $P\%$ confidence interval estimate of the population mean:

```
## Standard error (based upon CLT)
std_error = sample_sd/sqrt(n)

## Lower endpoint
lower = qnorm(0.025, mean = sample_mean, sd = std_error)

## Upper endpoint
upper = qnorm(0.975, mean = sample_mean, sd = std_error)

## Interval
c(lower, upper)
```

Two Ways to use CLT

Approach #2 - Rather than using a specific Normal distribution, use a generic formula with an appropriate quantile from $N(0,1)$:

```
## Quantile for the middle 95% of N(0,1)
key_quantile = qnorm(0.975, mean = 0, sd = 1)

## Standard error (based upon CLT)
std_error = sample_sd/sqrt(n)

## Lower endpoint
lower = sample_mean - key_quantile*std_error

## Upper endpoint
upper = sample_mean + key_quantile*std_error

## Interval
c(lower, upper)
```

Conclusions

There are three major takeaways from this presentation:

1. Interval estimates use sample data to provide a range of plausible values for an unknown characteristic of a population.
2. Confidence intervals are interval estimates whose margin of error is determined by a procedure with a long-run success rate, known as the confidence level.
3. You can calculate a P% confidence interval by taking the middle P% of the distribution of sample estimates, or by applying a generic formula involving the point estimate, a quantile from the $N(0,1)$ distribution, and the standard error.