The Central Limit Theorem and Confidence Intervals

STAT-UB.0001 Statistics for Business Control

Ningshan Zhang

IOMS Department
nzhang@stern.nyu.edu

Jul 26, 2018

* Office Hours: Wed & Fri 10:00 - 11:30 AM, KMC 8-174

Review

Continuous random variables

- Probability density function (pdf)
- Area under the curve

Normal distribution

- Normal distribution's pdf
- Standard normal distribution, z-tables
- Convert normal to standard normal

Populations and Samples

There are many times that we want to know something about a population, but it is unrealistic to gather all the information to obtain an exact answer.

Example:

- What is the average amount of time that NYU undergraduates spend text messaging per day?
- What is the average income of all Stern graduates?

Populations and Samples

As an alternative, we can take a random subset of items or individuals from the population, and use this random sample to draw conclusions about the population.

- A random subset is an unbiased sample of the population.
- Will we get a precise answer to our original question?

Population Parameters and Sample Statistics

Population parameters

- Descriptive measures of a *population*
- Example: population mean, population variance

Sample statistics

- Descriptive measures of a *sample*
- Example: sample mean, sample variance

Goal: use sample statistics to make inferences about the parameters of a population.

Population Parameters and Sample Statistics

Often, sample statistics are estimators of population parameters.

Population parameter	Sample statistic
μ	$\bar{x}$
σ^2	s^2
σ	s
p	$\hat{p}$

However, sample statistics are subject to randomness.

- Different samples will give different estimates.
- Estimates are (almost) never exactly correct.

* Link to Animations of Sampling Variation

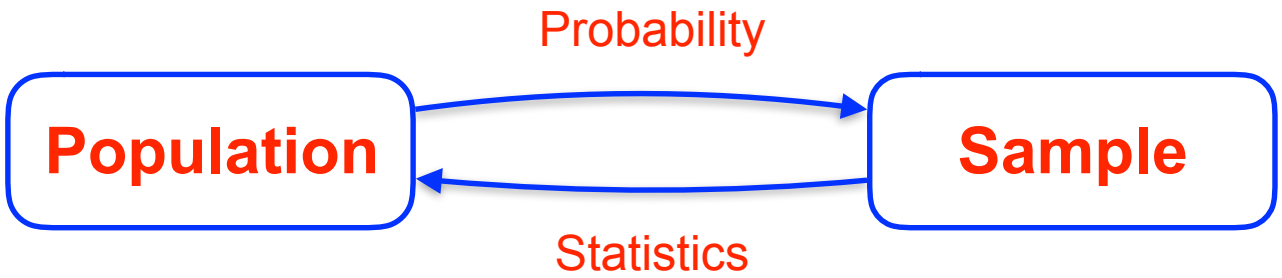
Sampling Distribution

Sample statistic is a random variable:

- Random experiment: randomly draw n observations from the population.
- Outcome of the random experiment: sample statistic calculated from the sample of n observations.

Since sample statistic is a random variable it follows some distribution. We refer to this distribution as a *sampling distribution*.

Populations and Samples



- Probability: how would the sample statistic fluctuate around the population parameter?
- Statistics: given the sample, what are possible values of the population parameter?

Sampling Distribution of Sample Mean $\bar{X}$

Given a sample of n observations, $X_1, \cdots, X_n$, the sample mean is computed as

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

We will study the sampling distribution of $\bar{X}$:

- Expected value of $\bar{X}$, denoted by $\mathbb{E}(\bar{X})$ or $\mu_{\bar{X}}$.
- Standard deviation of $\bar{X}$, denoted by $\text{sd}(\bar{X})$ or $\sigma_{\bar{X}}$.
- Histogram of $\bar{X}$.

The Central Limit Theorem (CLT)

Suppose $X_1, X_2, \ldots, X_n$ are sampled independently from a population with mean μ and standard deviation σ. Let $\bar{X}$ be the sample mean,

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Then,

- $\mu_{\bar{X}} = \mathbb{E}(\bar{X}) = \mu$,
- $\sigma_{\bar{X}} = \mathrm{sd}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$,
- If n is sufficiently large ($n \geq 30$), then $\bar{X}$ is approximately normal.

Expected Value of $\bar{X}$

The expected value of $\bar{X}$ is μ:

$$\mu_{\bar{X}} = \mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(X_i) = \mu.$$

- We will refer to $\mu_{\bar{X}}$ as the "mean of sample mean".

Standard Deviation of $\bar{X}$

The standard deviation of $\bar{X}$ is $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$.

- We will refer to $\sigma_{\bar{X}}$ as the "standard deviation of sample mean".
- The variance of $\bar{X}$ is $\frac{\sigma^2}{n}$.

Distribution of $\bar{X}$

When n is sufficiently large, $\bar{X}$ is approximately normally distributed, with mean $\mu_{\bar{X}} = \mu$ and standard deviation $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$.

- In general, $n \geq 30$ is sufficiently large.
- $\bar{X}$ is approximately normal, even if the population is not.
- $\bar{X}$ fluctuates around population mean μ:

$$\mathbb{P}(\mu - \frac{2\sigma}{\sqrt{n}} < \bar{X} < \mu + \frac{2\sigma}{\sqrt{n}}) = 0.95.$$

Distribution of $\bar{X}$

What if $n < 30$?

- We need additional assumptions about the population distribution, in order to understand the distribution of $\bar{X}$.

Table: Relationship between population and sample mean.

Population	**Sample size** n	**Sample mean** $\bar{X}$
Normal	any $n \geq 1$	Normal
Any distribution	$n \geq 30$	Approximately normal

CLT Demo

This website has a nice demo of the Central Limit Theorem:
`http://onlinestatbook.com/stat_sim/sampling_dist`

Try the following settings:

- Set the population to be skewed, set $n = 25$.
- Set a custom population, set $n = 25$.
- Set a custom population, set $n = 5$.
- Set the population to be normal, set $n = 5$.

Estimation

In practice we rarely know the true value of the population parameters, but we can use sample data to estimate them.

- Point estimation: use a single number to estimate the population parameter of interest.
- Interval estimation: construct an interval that contains the true parameter value with a certain probability.

Point Estimation

Example:

- $\bar{X}$ is a point estimator of μ.
- s^2 is a point estimator of σ^2.

Point estimates are problematic:

- (Almost) never exactly correct.
- Tell us nothing about their own variability.

Interval Estimation

Confidence Intervals:

- A *confidence interval* (CI) is a formula that tells us how to construct an interval using sample data, such that the interval contains the true parameter value with a certain probability.
- The formula varies depending on the the sample size, and the information available about the population.

Interval Estimation

We will study the following cases:

- Confidence intervals for population mean μ
  - Known population variance vs. unknown population variance
  - Large sample vs. small sample
- Confidence intervals for population proportion p

CI for the Mean: Known Variance

Setup: assume the population variance σ^2 is known, build a CI for the population mean μ using a sample of n observations.

- By CLT, when $n \geq 30$, $\bar{X}$ is (roughly) normally distributed with mean μ and standard deviation $\sigma/\sqrt{n}$.
- Let $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$, then Z is a standard normal (roughly).
- In particular,

$$\mathbb{P}(-1.96 < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < 1.96) = 0.95.$$

CI for the Mean: Known Variance

Rewrite the expression:

$$\mathbb{P}(-1.96 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1.96) = 0.95$$

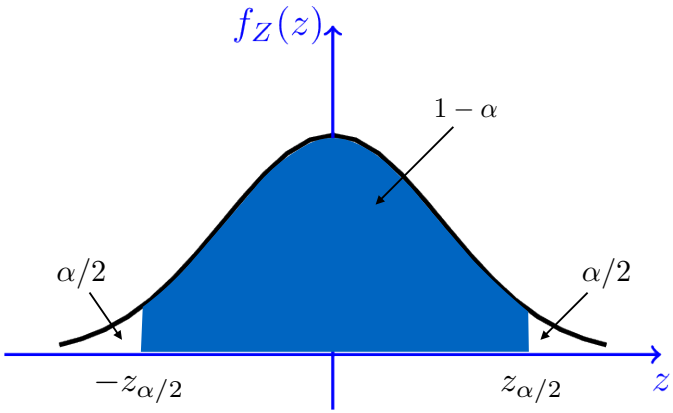
$$\iff \mathbb{P}(\bar{X} - \frac{1.96\sigma}{\sqrt{n}} < \mu < \bar{X} + \frac{1.96\sigma}{\sqrt{n}}) = 0.95$$

Thus, the interval $(\bar{X} - \frac{1.96\sigma}{\sqrt{n}}, \bar{X} + \frac{1.96\sigma}{\sqrt{n}})$ is a CI for μ with confidence level 0.95.

CI for the Mean: Known Variance

In general, let $z_{\alpha/2}$ be the value such that

$$\mathbb{P}(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha.$$



Example:

- $\alpha = 0.05$, $z_{0.025} = 1.96$.
- $\alpha = 0.1$, $z_{0.05} = 1.64$.

Then the interval $(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}})$ is a CI for μ with confidence level 1α.

Example

In an attempt to estimate the average number of sick days an employee uses in a year in a large firm, a HR manager examined a sample of 200 employees. Sample average was $\bar{X} = 3.47$. Assuming a population standard deviation of $\sigma = 2.8$ days, build a 90% CI for μ, the true average among all employees.

Interpretations of Confidence Intervals

In the example, we found that $(3.14, 3.80)$ is a 90% CI for μ.

Which of the following statements is true?

(a) There is a 0.90 probability that μ is between 3.14 and 3.80.

(b) μ will be between 3.14 and 3.80 90% of the time.

(c) In 90% of all future samples, $\bar{X}$ will be between 3.14 and 3.80.

(d) μ is between 3.14 and 3.80.

(e) None of the above.

Interpretations of Confidence Intervals

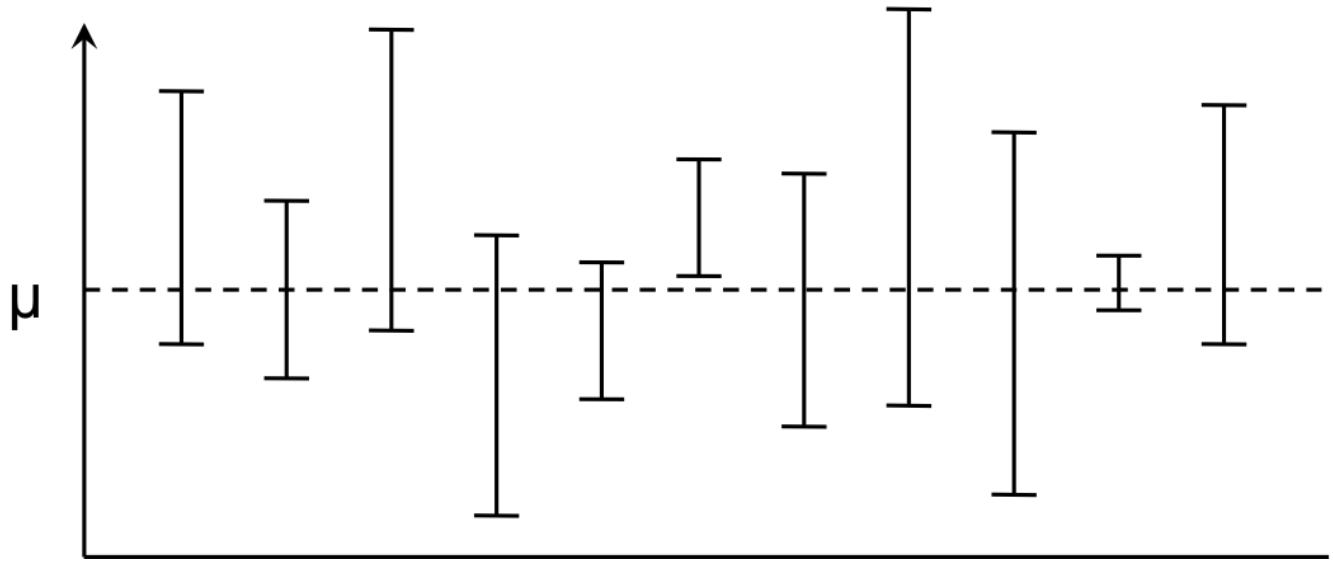
Warning: The practical interpretation of a CI is tricky!

- μ is nonrandom, it makes no sense to talk about the probability of μ.
- The "$1-\alpha$ confidence level" refers to the process of constructing confidence intervals, not to the particular CI estimate obtained from the given sample.

Interpretations of Confidence Intervals

The proportion of these intervals (in the long run) that contain μ is equal to $1 - \alpha$.

Figure: Different samples give different CIs.



Interpretations of Confidence Intervals

In practice, we only have one sample, so why should we care about confidence intervals?

- $1 - \alpha$ represents the proportion of times that we successfully obtain a CI that covers μ.
- Generally, we make α small so that most of the confidence intervals that we compute will contain μ.

Why not construct a 100% confidence interval?

Interpretations of Confidence Intervals

Compare: the price a customer will be willing to pay for our new, revolutionary cell phone is between

- \$0 and \$250 M, with 100% certainty.
- \$355 and \$420 M, with 90% certainty.

The price of higher certainty is lack of accuracy.