Quantitative Methods

Estimation and Inference

Intro and Exam Focus

- Sampling methods: probability sampling vs. nonprobability sampling
- Central limit theorem
  - Distribution of sample means
  - Standard error
  - *Confidence interval for population mean*
  - The *t*-distribution and when to use it
- Resampling: bootstrapping vs. jackknife

Sampling Methods

- **Probability sampling/(simple) random sampling:** every population member has an equal probability of being selected
  - Example of **systematic sampling:** choosing every *k*th element from population
- **Nonprobability sampling:** use judgment of researcher, or low-cost/readily available data, to select sample items
- *Care must be taken to ensure population is "stationary" (parameters do not change over the sample period), else inference is invalid*

Stratified Random Sampling

1. **Create subgroups** from population based on important characteristics (*strata*)
   - Example: create subgroups in bond index based on sector (sovereign/corporate), duration (short/long), and coupon (high/low)

2. **Select samples** from each subgroup in proportion to the size of the subgroup; matches sample distribution of characteristics of the underlying population
   - Example: bond index portfolios will have similar overall allocation to sector, duration, and coupon level as the index being tracked

Cluster Sampling

- Create subsets (clusters), each of which is representative of overall population (e.g., group residents by county)
- **One-stage cluster sampling:** take random sample of clusters and include all data from those clusters
- **Two-stage cluster sampling:** select clusters and take random samples from each

Nonprobability Sampling Methods

- **Convenience sampling:** use readily available low-cost data, not necessarily representative, perhaps for preliminary investigation
- **Judgmental sampling:** select observations from population based on analyst's judgment

Sampling Error

- **Sampling error:** difference between a sample statistic and true population parameter (e.g., $\bar{x} - \mu$)
- **Nonprobability** sampling: may lead to greater *sampling error* than probability sampling
- **Sampling distribution:** distribution of all possible values of a statistic (e.g., mean) for samples of size *n*

Central Limit Theorem: Example

Consider the distribution of a fair die (single dice):

Population parameters:

μ = (1 +2 + 3+ 4 + 5 + 6) / 6 = 3.5

σ^2 = $[(-2.5)^2 + (-1.5)^2 + (-0.5)^2 + (0.5)^2 + (1.5)^2 + (2.5)^2] / 6 = 2.92$

σ = $\sqrt{2.92} = 1.71$

Taking 5 **samples** of size 10 (n=10):

Central Limit Theorem: Example (cont.)

	Sample Number				
	1	**2**	**3**	**4**	**5**
Roll #1	2	1	2	6	2
Roll #2	1	1	6	3	5
Roll #3	6	4	4	3	6
Roll #4	3	5	3	5	5
Roll #5	1	6	1	3	2
Roll #6	4	5	2	6	2
Roll #7	5	6	3	5	5
Roll #8	5	5	5	3	5
Roll #9	1	2	2	2	1
Roll #10	2	3	5	4	6
Sample mean ($\overline{X}$)	**2.82**	**3.64**	**3.27**	**4.00**	**4.00**

Central Limit Theorem: Example (cont.)

	Sample Number				
	1	2	3	4	5
Roll #1	2	1	2	6	2
Roll #2	1	1	6	3	5
Roll #3	6	4	4	3	6
Roll #4	3	5	3	5	5
Roll #5	1	6	1	3	2
Roll #6	4	5	2	6	2
Roll #7	5	6	3	5	5
Roll #8	5	5	5	3	5
Roll #9	1	2	2	2	1
Roll #10	2	3	5	4	6
Sample mean ($\bar{X}$)	**2.82**	**3.64**	**3.27**	**4.00**	**4.00**

Average $\bar{X}$ = 3.55

Recall **μ = 3.5**

Central Limit Theorem

For **any** population with mean μ and variance σ^2, as the size of a random sample gets large, the distribution of sample means ($\bar{X}$) approaches a normal distribution with **mean μ and variance σ^2 / n.**

It allows inferences about and confidence intervals for population means, based on the distribution of sample means.

Standard Error of the Sample Mean

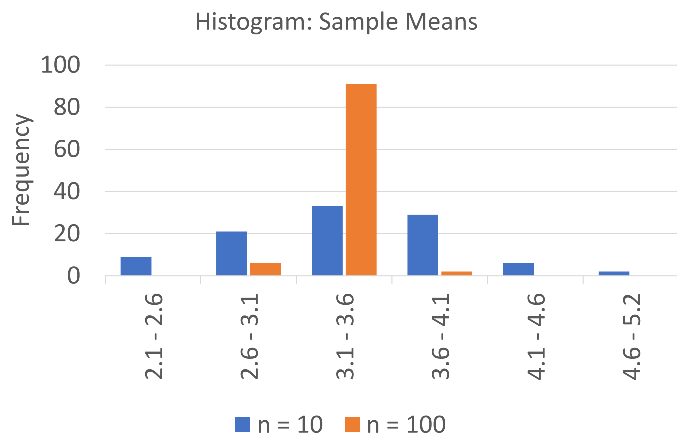
The standard deviation of the **distribution of sample means ($\bar{X}$) is** called the ***standard error*** of the sample mean.

$$SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \text{ or } \frac{s}{\sqrt{n}}$$

Use sample standard deviation, s, in calculation of SE when population standard deviation, **σ**, is not known

Standard Error of the Sample Mean (cont.)

Note: as sample size (n) increases, standard error decreases



n = 10:

SE = 1.71/√10 = 0.55

n = 100:

SE = 1.71/√100 = 0.17

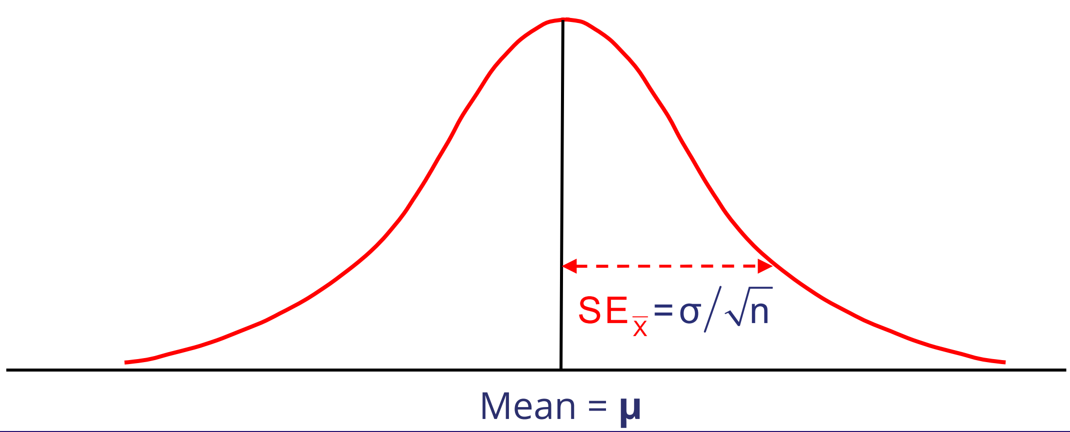
Standard Deviation vs. Standard Error

- Standard deviation, **σ**, is the dispersion of a single observation, X, from a distribution.

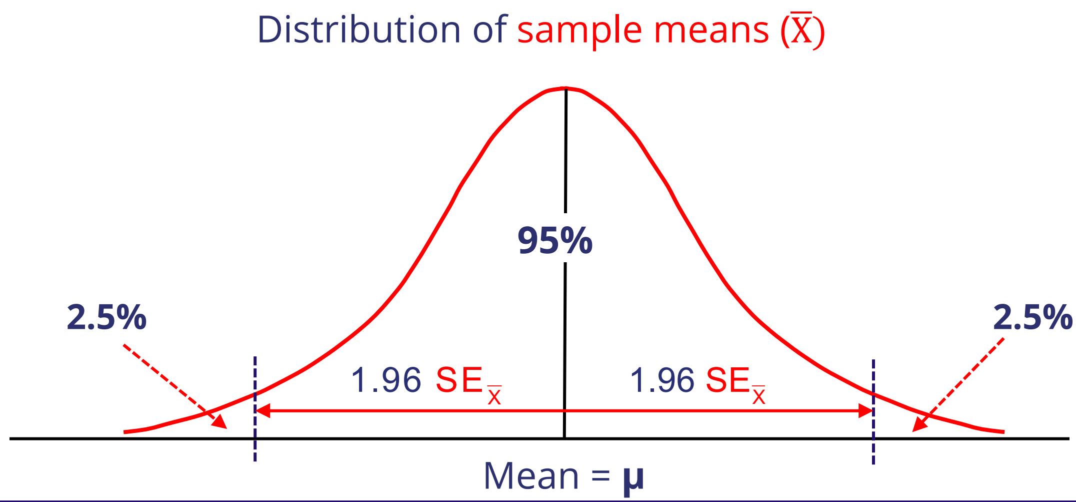
- Standard error, **$SE_{\bar{x}}$**, is the dispersion of the **sample mean** around the distribution's true population mean, **μ.**

Confidence Intervals

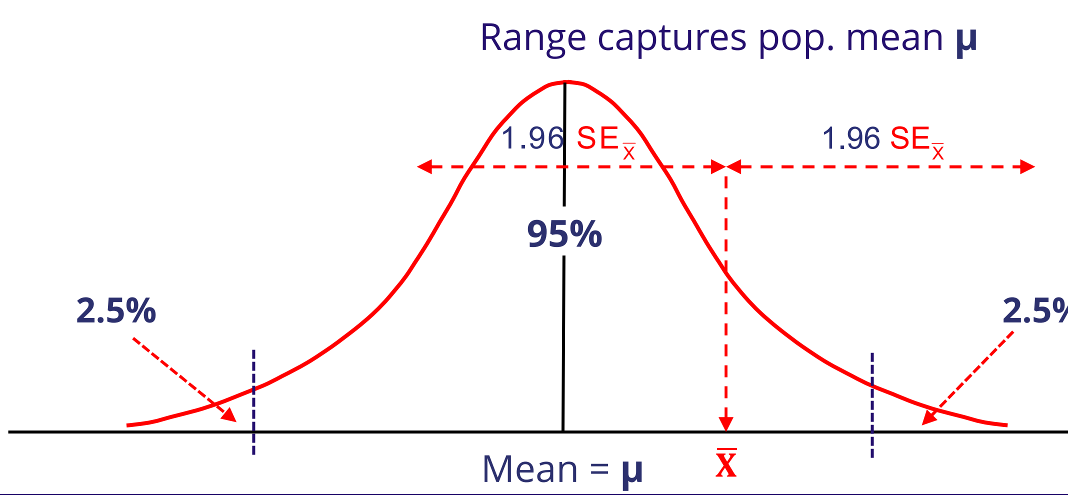
Distribution of sample means ($\bar{X}$) (assume **σ** is known)



Confidence Intervals (cont.)

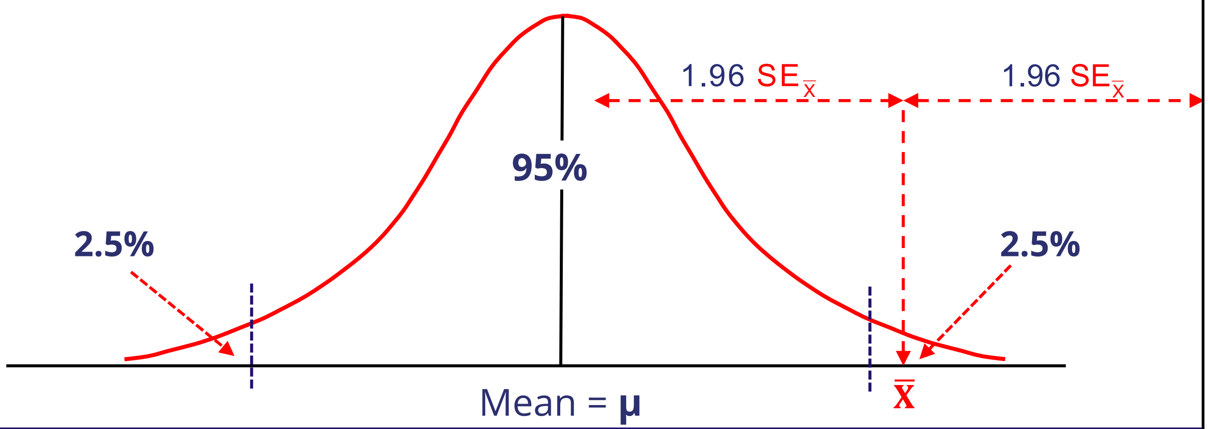


Confidence Intervals (cont.)



Confidence Intervals (cont.)

Range fails to capture pop. mean **μ**



Confidence Intervals: Example

Confidence interval for **μ** = $\bar{X}$ +/– (z_{crit} × SE $_{\bar{X}}$)

A sample of returns on 16 stocks has a sample mean of 3%, and the population standard deviation is 10%. What is the 95% confidence interval for the mean return of the sample stocks?

Confidence interval for **μ** = 3% ±

= 3% ±

=

-3

Student's *t*-Distribution

- When **σ** is not known, use $\mathbf{s_x}$ in the calculation of standard error:

$$SE_{\bar{X}} = \frac{s_X}{\sqrt{n}}$$

→ Sample mean follows the Student's *t*-distribution instead of a normal distribution (with degrees of freedom = n – 1)

→ Use *t*-distribution for critical values instead of *z*-distribution

Confidence interval for **μ** = $\bar{X}$ +/– (t_{crit} × $SE_{\bar{X}}$)

Properties of Student's *t*-Distribution

- Symmetrical (bell shaped)
- Fatter tails than a normal distribution
- Defined by single parameter, degrees of freedom (df), where df = n – 1
- As df increase, *t*-distribution approaches standard normal distribution

Student's *t*-Distribution

Level of Significance for One-Tailed Test						
df	0.100	0.050	0.025	0.01	0.005	0.0005

Level of Significance for Two-Tailed Test						
df	0.20	0.10	0.05	0.02	0.01	0.001
1	3.078	6.314	12.706	31.821	63.657	636.619
2	1.886	2.920	4.303	6.965	9.925	31.599
3	1.638	2.353	3.182	4.541	5.841	12.294
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.869
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.408
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318
13	1.350	1.771	2.160	2.650	3.012	4.221
14	1.345	1.761	2.145	2.624	2.977	4.140
15	1.341	1.753	2.131	2.602	2.947	4.073
….						
26	1.315	1.706	2.056	2.479	2.779	3.707
27	1.314	1.703	2.052	2.473	2.771	3.690
28	1.313	1.701	2.048	2.467	2.763	3.674
29	1.311	1.699	2.045	2.462	2.756	3.659
30	1.310	1.697	2.042	2.457	2.750	3.646
40	1.303	1.684	2.021	2.423	2.704	3.551
60	1.296	1.671	2.000	2.390	2.660	3.460
120	1.289	1.658	1.980	2.358	2.617	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.291

Confidence Intervals for Mean

When Sampling From A:		**Reliability Factors**	
Distribution	**Variance**	**Small Sample ($n < 30$)**	**Large Sample ($n \geq 30$)**
Normal	Known	*z*-statistic	*z*-statistic
Normal	Unknown	***t*-statistic**	*t*-statistic*
Nonnormal	Known	Not available	*z*-statistic
Nonnormal	Unknown	Not available	*t*-statistic*

* The *z*-statistic is acceptable proxy; however, *t*-statistic should technically be used

Confidence Interval Example: σ Not Known

A sample of returns on 16 stocks results in a sample mean of 3% and sample standard deviation of 10%. Assuming that the returns are approximately normal, what is the 95% confidence interval for the mean returns of the sample stocks?

A. –1.90% to +7.90%.

B. –2.33% to +8.33%.

C. –18.31% to +24.31%.

df	One-Tailed Probabilities (*p*)		
	p = 0.10	p = 0.05	p = 0.025
14	1.345	1.761	2.145
15	1.341	1.753	2.131
16	1.337	1.746	2.120

Confidence Interval Solution: σ Not Known

A. –1.90% to +7.90%.

B. –2.33% to +8.33%.

C. –18.31% to +24.31%.

df	One-Tailed Probabilities (*p*)		
	p = 0.10	p = 0.05	p = 0.025
14	1.345	1.761	2.145
15	1.341	1.753	2.131
16	1.337	1.746	2.120

Conf. int. for $\boldsymbol{\mu} = \bar{X} +/- (t_{crit} \times SE_{\bar{x}}) = 3\% \pm$

$= 3\% \pm$

$=$

-4

Resampling

- Computational methods to estimate the standard error of the sample mean include the following:
  - **Bootstrapping**: resample from original with replacement of items when drawn, calculating the sample mean each time; calculate the sample standard deviation of these sample means
  - **Jackknife:** calculate multiple sample means, each with one observation removed; calculate standard deviation of these means

Bootstrapping: CFA Institute Example

Monthly Returns of Stock	Resample 1	Resample 2	Resample 3		Resample 1000
–0.096	0.055	–0.096	–0.033	...	–0.072
–0.132	–0.033	0.055	–0.132	...	0.255
–0.191	0.255	0.055	–0.157	...	0.055
–0.096	–0.033	–0.157	0.255	...	0.296
0.055	0.255	–0.096	–0.132	...	0.055
–0.053	–0.157	–0.053	–0.191	...	–0.096
–0.033	–0.053	–0.096	0.055	...	0.296
0.296	–0.191	–0.132	0.255	...	–0.132
0.055	–0.132	–0.132	0.296	...	0.055
–0.072	–0.096	0.055	–0.096	...	–0.096
0.255	0.055	–0.072	0.055	...	–0.191
–0.157	–0.157	–0.053	–0.157	...	0.055
Sample mean	–0.019	–0.06	0.001	...	0.04

Bootstrapping: Example (cont.)

Drawing 1,000 such samples, we obtain 1,000 sample means.

The average across all resample means is −0.01367.

The sum of squares of the differences between each sample mean and the average across all resample means is 1.94143.

We calculate an estimate of the standard error of the sample mean:

$$S_{\bar{x}} = \sqrt{\frac{1}{999} \times 1.94143} = 0.04408$$

Solutions

Confidence Intervals: Example

Confidence interval for **μ** = $\bar{X}$ +/– (z_{crit} × $SE_{\bar{x}}$)

A sample of returns on 16 stocks has a sample mean of 3%, and the population standard deviation is 10%. What is the 95% confidence interval for the mean return of the sample stocks?

$$\text{Confidence interval for } \mu = 3\% \pm \left(1.96 \times \frac{10\%}{\sqrt{16}}\right)$$

$$= 3\% \pm 4.9\%$$

$$= (-1.9, 7.9\%)$$

-3

Confidence Interval Solution: σ Not Known

A. **–1.90% to +7.90%.**

B. **–2.33% to +8.33%.**

C. **–18.31% to +24.31%.**

df	**One-Tailed Probabilities (*p*)**		
	p = 0.10	**p = 0.05**	**p = 0.025**
14	1.345	1.761	2.145
15	1.341	1.753	2.131
16	1.337	1.746	2.120

$$\text{Conf. int. for } \mu = \bar{X} \text{ +/– } (t_{crit} \times SE_{\bar{x}}) = 3\% \pm \left(2.131 \times \frac{10\%}{\sqrt{16}}\right)$$

$$= 3\% \pm 5.33\%$$

$$= (-2.33\%, 8.33\%)$$

-4