WILEY

# Reading 10: Sampling and Estimation

# Learning Outcome Statements

- Covered
  - 10a, 10b, 10c, 10e, 10f, 10g, 10h, 10i, 10j

- Not Covered
  - 10d, 10k

# Sampling

A **sample** is a subset of a population, often used to make inferences about population parameters. For example, a set of 500 giraffes captured and tagged for studying is a sample of the giraffe population (estimated at about 80,000).

In a **simple random sample**, every population member has an equal chance of being selected.

A **sampling distribution** is the distribution of a statistic from a number of samples of the same size. For example, if several samples of 500 giraffes were selected and the mean height of the giraffes in each sample calculated, the distribution of those means is a sampling distribution.

**Sampling error** is the difference between a sample statistic (e.g., the mean height of giraffes in a sample of size 500) and the corresponding population parameter (the mean height of all the giraffes in the world).

# Stratified Sampling

In **stratified sampling**, relevant characteristics of the population are identified, and the sample is selected to maintain the relative frequencies of those characteristics.

Example: The bonds in a given index are a mix of government bonds, investment-grade corporate bonds, and below-investment-grade corporates, with a range of maturities, embedded options, and credit ratings. If 2- to 5-year maturity, AA-rated, callable corporate bonds make up 2% of the index, then the stratified sample of those bonds chosen by an EFT designed to track the index will have 2% of its bonds being 2- to 5-year maturity, AA-rated, callable corporates. Within that category, however, the bonds can be chosen using simple sampling; i.e., any bond in that category in the index is as likely to be chosen for the ETF as any other bond in that category.

# Central Limit Theorem

Given a population with any probability distribution having mean μ and (finite) variance $\sigma^2$, if $n$ is sufficiently large, the sampling distribution of the mean $\bar{X}$ of samples of size $n$ will have an (approximately) normal distribution with mean $\mu$ and variance $\sigma^2/n$.

- The central limit theorem (CLT) applies to any distribution with a finite variance: normal, uniform, binomial, multimodal, skewed, and so on.
- The CLT allows an analyst to make inferences about the population mean given the means of random samples from that population (as long as the samples are large).
- Generally, $n \geq 30$ is considered large.

# Standard Error of the Sample Mean

The **standard error of the sample mean** is the standard deviation of the distribution of the sample means from samples of size *n*.

When we know the standard deviation ($\sigma$) of the population (rarely), the standard error of the sample mean is:

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

When we don't know the standard deviation of the population (usually), we have to use the standard deviation ($s$) of the sample, and the standard error of the sample mean is:

$$s_{\bar{X}} = \frac{s}{\sqrt{n}}$$

# Practice Question

A sample of the annual returns of  financial stocks and monthly returns of 36 financial stocks has a mean of 9% and a standard deviation of 18%. The standard error of the sample is closest to

A. 0.5

B. 1.5

C. 3.0

# Estimators: Desirable Properties – I

An **estimator** is a sample statistic used to estimate a population parameter. For example, a sample mean is an estimator of the population mean, and a sample variance is an estimator of a population variance.

Ideally, an estimator will be:

- **Unbiased**: Its expected value (the mean of its sampling distribution) will equal the population parameter it estimates.
- **Efficient**: It has the smallest variance of all unbiased estimators of the same population parameter.
- **Consistent**: As the sample size increases, the estimate tends to become more accurate (the variance of the estimator gets smaller).

# Estimators: Desirable Properties – II

Examples

- As an estimator of the population mean $\mu$, the sample mean $\mu$ has all $\bar{X}$ of the desired properties:
  - $E(\bar{X}) = \mu$
  - It has the smallest standard error of all unbiased estimators of $\mu$
  - As $n$ goes to infinity, its standard error $\sigma/\sqrt{n}$ goes to zero
- As an unbiased estimator of the population variance $\sigma^2$, the sample variance $s^2$ has all of the desired properties:
  - $E(s^2) = \sigma^2$
  - $s^2$ has the smallest standard error of all unbiased estimators of $\sigma^2$
  - As $n$ increases, the standard error of $s^2$ decreases

# Point Estimate & Confidence Interval – I

A **point estimate** is a single value estimate for a population parameter.

Example: A sample of 60 values from a population of stock betas has a mean of 1.05 and a standard error of 0.32. The sample mean—1.05—is a point estimate for the mean beta of the population.

A **confidence interval** is an interval (range) estimate for a population parameter: The parameter will be in that interval with a given probability.

Example: For the same sample, the range (0.987, 1.113) is a 95% confidence interval for the mean population beta.

# Point Estimate & Confidence Interval – II

Example: A sample of 60 values from a population of stock betas has a mean of 1.05 and a standard error of 0.32.

Because the sample is large, the CLT applies: We can base confidence intervals on normal distributions. Thus:

- A 90% confidence interval for the *mean population beta* is $\left(\mu - 1.645\sigma_{\bar{X}}, \mu - 1.645\sigma_{\bar{X}}\right)$, or (0.997, 1.103).

- A 95% confidence interval for the *mean population beta* is $\left(\mu - 1.96\sigma_{\bar{X}}, \mu - 1.96\sigma_{\bar{X}}\right)$, or (0.987, 1.113).

- A 99% confidence interval for the *mean population beta* is $\left(\mu - 2.58\sigma_{\bar{X}}, \mu - 2.58\sigma_{\bar{X}}\right)$, or (0.967, 1.133).

# Student's *t*-Distribution – I

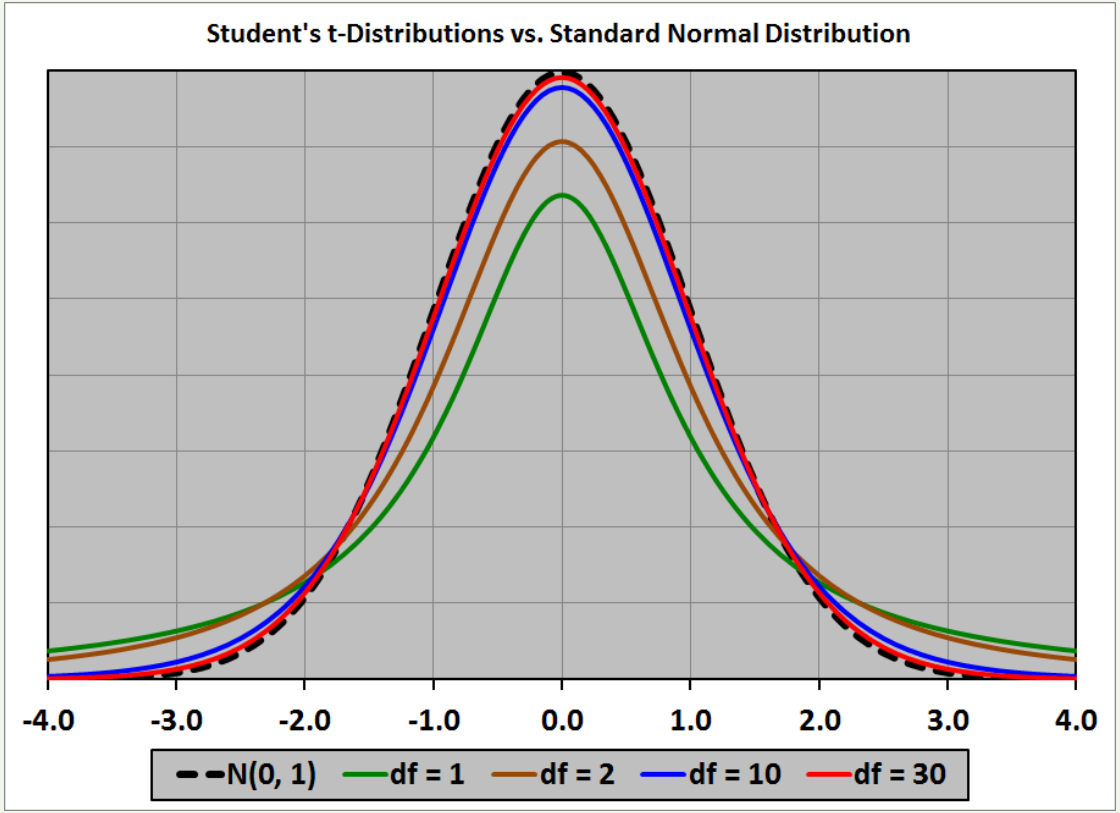
A **Student's *t*-distribution** looks very similar to a standard normal distribution. There are infinitely many *t*-distributions, one for each sample size from 2 to +∞. The characteristics of *t*-distributions are:

- They are all symmetric about zero: They have skewness of zero.
- They have less probability near the mean than the standard normal distribution.
- They have more probability in the tails than the standard normal distribution: commonly called "fat tails."
- They all have positive excess kurtosis, and the fewer the degrees of freedom, the greater the excess kurtosis.
- The degrees of freedom (*df*) is one less than the sample size:

$$df = n - 1$$

- As *df* increases, the *t*-distribution approaches a standard normal distribution.

# Student's *t*-Distribution – II



# Constructing Confidence Intervals – I

Information needed to construct a confidence interval:

1. Is it a confidence interval for the mean of a population or for an arbitrary member of the population?
2. Is the population (at least approximately) normally distributed or not?
3. Do we have a large sample or a small sample?
4. Do we know the population variance or not?
5. What degree of confidence (or level of significance) do we want?

# Constructing Confidence Intervals – II

| Which Distribution to Use | | | |
|---|---|---|---|
| Population Distribution | Population Variance | Small Sample ($n < 30$) | Large Sample ($n \geq 30$) |
| Normal | Known | $z$ | $z$ |
| Normal | Unknown | $t$ | $t^*$ |
| Nonnormal | Known | Not available | $z$ |
| Nonnormal | Unknown | Not available | $t^*$ |

*Using a $z$-distribution is also acceptable; the $t$-distribution is slightly more conservative.

The value obtained from the $z$- or $t$-distribution is called a **reliability factor**.

# Constructing Confidence Intervals – III

The **degree of confidence** is the probability that the real value will fall ***into*** the constructed confidence interval; common degrees of confidence are 90%, 95%, and 99%.

The **level of significance** ($\alpha$) is the probability that the real value will fall ***outside*** the constructed confidence interval; common levels of significance are 1%, 5%, and 10%.

The degree of confidence equals $1 - \alpha$.

The reliability factor is found from a *z*-table or *t*-table using the degree of confidence (and, for the *t*-table, the number of degrees of freedom).

# Constructing Confidence Intervals – IV

Confidence interval for the *mean* of a population:

$$\text{Point estimate} \pm \text{Reliability factor} \times \text{Standard error}$$

Confidence interval for an *arbitrary member* of a population:

$$\text{Point estimate} \pm \text{Reliability factor} \times \text{Standard deviation}$$

The point estimate will be the sample mean. Use the population standard deviation (σ) when it is known; otherwise use the sample standard deviation (*s*). The standard error is:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \quad \text{or} \quad s_{\bar{X}} = \frac{s}{\sqrt{n}}$$

# Constructing Confidence Intervals – V

Example: A sample of 60 monthly returns on an investment portfolio has a mean of 0.5% and a standard deviation of 0.3%; the standard deviation of all monthly returns is unknown. Construct a 95% confidence interval for the mean monthly return, and a 90% confidence interval for the next month's return.

As this is a large sample, we can use a *z*-distribution; the 90% reliability factor is 1.645, and the 95% reliability factor is 1.96. The 95% confidence interval for the mean is:

$$\left(0.5\% - 1.96\frac{0.3\%}{\sqrt{60}}, 0.5\% + 1.96\frac{0.3\%}{\sqrt{60}}\right) \quad \text{or} \quad (0.424\%, 0.576\%)$$

The 90% confidence interval for next month's return is:

$$\left(0.5\% - 1.645(0.3\%), 0.5\% + 1.645(0.3\%)\right) \quad \text{or} \quad (0.007\%, 0.994\%)$$

# Constructing Confidence Intervals – VI

Example: A sample of 20 P/E ratios for companies in the S&P 500 has a mean of 18.5 and a standard deviation of 3.6; the standard deviation of all monthly returns is unknown, but P/E ratios are assumed to be approximately normally distributed. Construct a 99% confidence interval for the mean P/E ratio, and a 95% confidence interval for an arbitrary firm's P/E ratio.

As this is a small sample, we must use a *t*-distribution with 19 *df*; the 95% reliability factor is 2.093, and the 99% reliability factor is 2.861. The 99% confidence interval for the mean is:

$$\left(18.5 - 2.861\frac{3.6}{\sqrt{20}}, 18.5 + 2.861\frac{3.6}{\sqrt{20}}\right) \quad \text{or} \quad (16.2, 20.8)$$

The 95% confidence interval for next month's return is:

$$\left(18.5 - 2.093(3.6), 18.5 + 2.093(3.6)\right) \quad \text{or} \quad (11.0, 26.0)$$

# Practice Question

A sample of 50 Sharpe ratios for large-cap stocks has a mean of 0.65 and a standard deviation of 0.18. A 95% confidence interval for the average large-cap Sharpe ratio is *closest to*:

A. (0.30, 1.00)

B. (0.60, 0.70)

C. (0.64, 0.66)

WILEY

# Practice Questions with Solutions

# Practice Question

A sample of the annual returns of financial stocks and monthly returns of 36 financial stocks has a mean of 9% and a standard deviation of 18%. The standard error of the sample is closest to

A. 0.5

B. 1.5

C. 3.0

Correct Answer: B

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{18}{\sqrt{36}} = \frac{18}{6} = 3.0$$

# Practice Question

A sample of 50 Sharpe ratios for large-cap stocks has a mean of 0.65 and a standard deviation of 0.18. A 95% confidence interval for the average large-cap Sharpe ratio is *closest to*:

A. (0.30, 1.00)

B. (0.60, 0.70)

C. (0.64, 0.66)

Correct answer: B. (0.60, 0.70)

$$[0.65 - 1.96(0.18/\sqrt{50}),\ 0.65 - 1.96(0.18/\sqrt{50})]$$

$$(0.600,\ 0.700)$$