Sampling and Estimation

- More details on sampling, estimation, and confidence intervals will be given now
- Simple random sample: subset of a larger population created in such a way that each element of the population has an equal probability of being selected to the subset
- Sampling error: the difference between the observed and real statistics
- Sampling distribution: distribution obtained if samples are drawn from large population ("real sampling distribution"); same size samples used, population really large
- Stratified sampling: the population is divided in subpopulations (strata) based on one or more criteria; samples are then drawn from each strata according to their sizes in the overall population (Example: bond indexing
- Time-series vs. cross-sectional data

Sampling and Estimation

- Distribution of the sample mean is defined by the central limit theorem: if the population have mean ("real mean") μ and variance σ^2, then sample of the size n will have (as a) mean distribution normal distribution, with the same mean μ, and the variance $\frac{\sigma^2}{n}$ (n large).

$$\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n}, \quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

- $\sigma_{\bar{X}}$ is estimated by $s_{\bar{X}}$, $s_{\bar{X}} = \frac{s}{\sqrt{n}}$, and

$$s^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}$$

- Point estimates: example is mean value, estimated by a number
- We can also estimate intervals containing the value we are estimating with certain confidence level
- For point estimators we have formulas, e.g.

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

- There can be several estimators for the same quantity!

Sampling and Estimation

- Important properties of estimators
  - Unbiasedness: estimator expected value is the same as the mean of the sampling distribution
  - Efficiency: no other unbiased estimator has the smaller variance
  - Consistency: probability of getting estimates close to the value of the real mean increases as sample size increases
- Confidence interval $\left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$
  - $\bar{X}$ is the point estimate
  - $z_{\alpha/2}$ is the reliability factor, number based on the assumed distribution of the point estimate and the degree of confidence $(1-\alpha)$
  - Standard error $\frac{\sigma}{\sqrt{n}}$
- Usually, $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ is called precision
- For the normal distribution
  - 66% confidence $(1-\sigma)$, $z = 1$ $(1-\alpha = 0.66)$
  - 90% confidence, $z = 1.96$ $(1-\alpha = 0.9)$
  - 95% confidence $(2-\sigma)$, $z = 2$ $(1-\alpha = 0.953)$
  - 99% confidence $z = 2.58$ $(1-\alpha = 0.99)$

Sampling and Estimation

- If σ is not known, estimate for σ is used instead (s).
- If the population is large enough, reliability factors are distributed according to the t-distributions, i.e. confidence intervals are

$$\left[\bar{x} - t_{\alpha/2} \frac{s}{\sqrt{n}}, \; \bar{x} + t_{\alpha/2} \frac{s}{\sqrt{n}} \right]$$

($t_{\alpha/2}$ calculated with $n-1$ degrees of freedom!)

Distribution	Small sample	Large sample
normal, σ known	z	z
normal, σ not known	t	t (z)
non-normal, σ known	—	z
non-normal, σ not known	—	t (z)

- The sample size should be chosen appropriately, so to assure precision as desired

Sampling and Estimation

- Sources of biases:
  - Data-mining bias
  - Sample selection bias
  - Look-ahead bias
  - Time-period bias