Introduction to Sampling & Statistical Inference

BU609-4

Objectives

- Understand sampling distributions
- Understand estimation concept
- Apply sampling and estimation concepts when working with means, proportions, variances (one or two populations)

Parameters vs Statistics

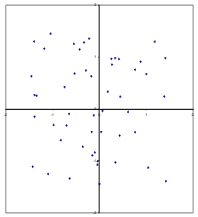
- μ
- σ
- p
- x-bar $(\bar{x})$
- s
- p-hat $(\hat{p})$

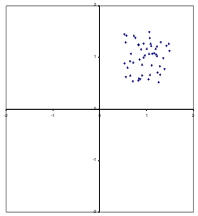
Also need to consider number of distinct populations; can work with one, two or more populations

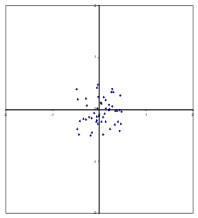
Selecting Statistic

- Selecting sample statistic to estimate parameter value depends on characteristics of statistic
- Estimator's desirable characteristics
  - ***Unbiasedness:*** unbiased estimator is one whose expected value is equal to parameter it estimates
  - ***Consistency:*** unbiased estimator is said to be consistent if difference between estimator and parameter grows smaller as sample size increases
  - ***Relative efficiency:*** For two unbiased estimators, one with smaller variance is said to be relatively efficient
- Fortunately, statisticians do this for us

Bias vs Efficiency







Upper Left: Unbiased, Inefficient

Upper Right: Biased, Efficient

Lower Left: Unbiased, Efficient

Sampling Distributions: Introduction

- In real life calculating **parameters** of populations is usually impossible
- Rather than investigating entire population, take a sample, calculate a ***statistic*** related to ***parameter*** of interest, and make an inference
- ***Sampling distribution*** of ***statistic*** is prob. dist. of statistic (mean, std dev, proportion)
  - can be Normal, or another known distribution
  - can use this to estimate how close statistic is to parameter

Opinion Polls – Example

- "Rich get Richer, Poor get Poorer, Poll Shows" (*The Record, 24-10-01)*
  - *"About 40 per cent of Canadians believe they got poorer during the last 10 years, an opinion poll suggests. The Leger Marketing survey indicated 39.6 per cent of respondents thought they were not as rich as 10 years ago, compared with 32.6 per cent who believed they were richer. Another 26.2 per cent reported no change. The survey of 1,508 Canadians has a margin of error of 2.6 percentage points, 19 times out of 20."*

Sampling Distribution of Mean

- An example
  - A die is thrown infinitely many times. Let x represent the number of spots showing on any throw.
  - Probability distribution of x is shown below
    - **What type of distribution is this?**

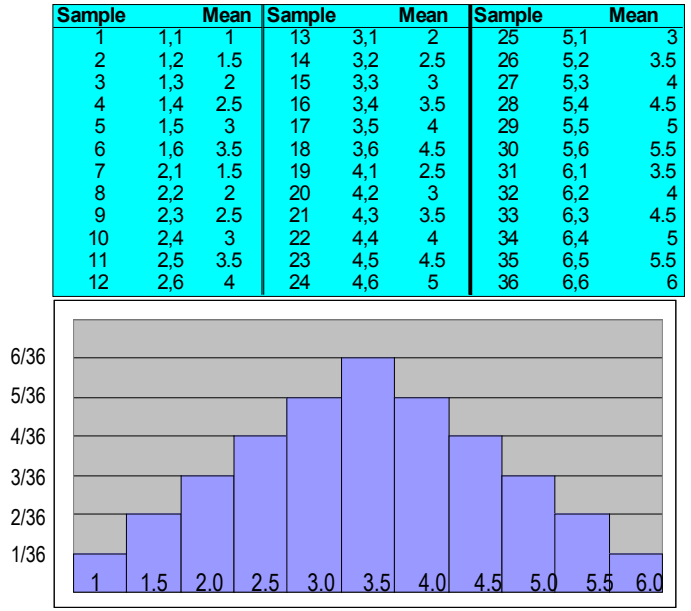
x	1	2	3	4	5	6
p(x)	1/6	1/6	1/6	1/6	1/6	1/6

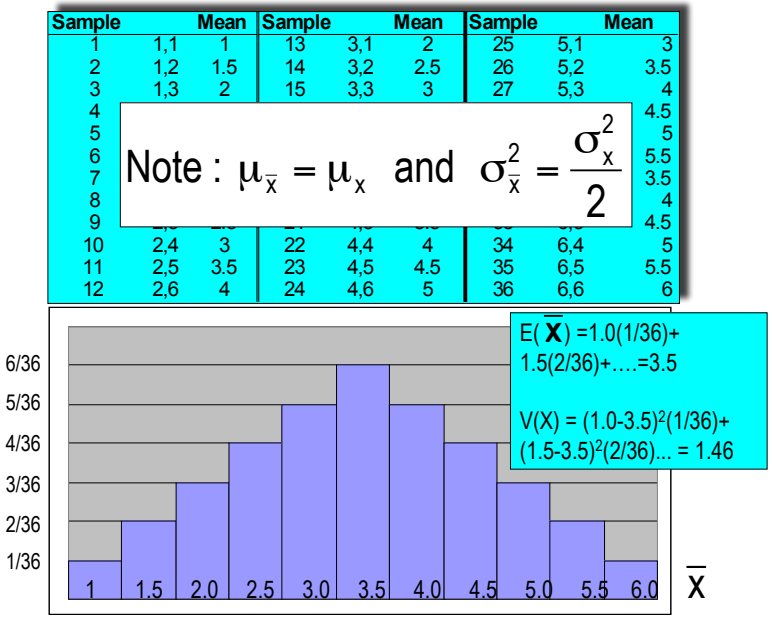
$E(X) = 1(1/6) + 2(1/6) + 3(1/6) + \ldots\ldots\ldots = 3.5$

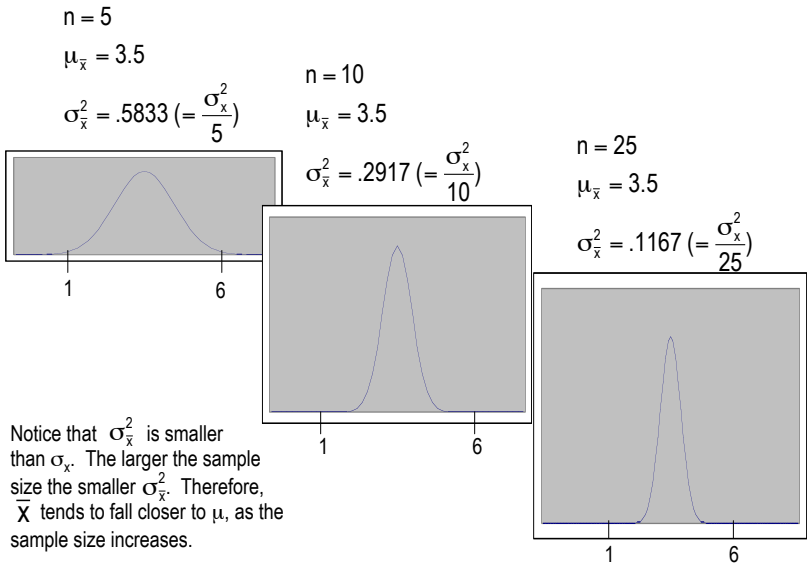
$V(X) = (1-3.5)^2(1/6) + (2-3.5)^2(1/6) + \ldots\ldots\ldots\ldots = 2.92$

- Suppose we want to estimate μ from the mean x-bar of a sample of size n = 2
- What is distribution that x-bar can follow?

Sample		Mean	Sample		Mean	Sample		Mean
1	1,1	1	13	3,1	2	25	5,1	3
2	1,2	1.5	14	3,2	2.5	26	5,2	3.5
3	1,3	2	15	3,3	3	27	5,3	4
4	1,4	2.5	16	3,4	3.5	28	5,4	4.5
5	1,5	3	17	3,5	4	29	5,5	5
6	1,6	3.5	18	3,6	4.5	30	5,6	5.5
7	2,1	1.5	19	4,1	2.5	31	6,1	3.5
8	2,2	2	20	4,2	3	32	6,2	4
9	2,3	2.5	21	4,3	3.5	33	6,3	4.5
10	2,4	3	22	4,4	4	34	6,4	5
11	2,5	3.5	23	4,5	4.5	35	6,5	5.5
12	2,6	4	24	4,6	5	36	6,6	6







Central Limit Theorem (CLT)

- If a truly random sample is independently drawn from **any** population,
- **Sampling distribution** of **sample mean** is approximately Normal (for sufficiently large sample size)
- Larger the sample size, more closely sampling distribution of sample mean will resemble Normal distribution

Guidelines for C.L.T.

- If sampling from a relatively symmetric dist. (Uniform, binomial with p=.5), n can be small (5 - 15)
- If sampling from a relatively skewed dist (binomial with p=.1, or bimodal), n should be ≥ 30
- When n ≥ 30, variability of sample mean is typically reasonable

Sampling Dist. of Sample Mean

1. $\mu_{\bar{x}} = E(\bar{x}) = \mu_x$

2. $\sigma^2_{\bar{x}} = s^2 = \dfrac{\sigma^2_x}{n} \qquad s = \sigma / \sqrt{n}$

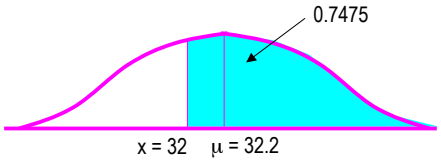
3. *If x is* Normal, $\bar{x}$ is Normal. If x is nonnormal $\bar{x}$ is approx Normal for sufficiently large sample size (n ≥ 30)

Question : How do we know σ ?

Note : SAMPLE std dev called "std error of mean"

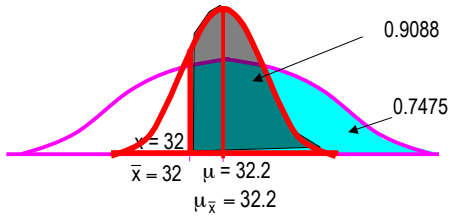
Bottling Example

- Amount of beverage in each bottle is normally distributed with mean of 32.2 ounces and standard deviation of .3 ounces.
  - Find probability that bottle bought by customer will contain more than 32 ounces.



Exercise (cont.)

- Find probability that carton of four bottles will have mean of more than 32 ounces of pop per bottle
- What about case of 8 bottles?



Finance Application of CLT

- An initial investment **I** grows over time
  - In first month grew by factor of 1.1, then shrank by factor of .97 in second month, and grew by factor of 1.2 in third month
- Value at end of 3 months = **I***1.1*.97*1.2
  - If you studied logarithms, may remember that they convert multiplications to additions
  - Logarithm of value after 3 months is log(I) + Log(1.1) + Log(.97) + Log(1.2)
- Therefore log of value of investment after large number of periods is sum of bunch of random numbers, and is hence normally distributed
  - An uncertain number whose logarithm is normally distributed is known as log normal
  - This is why investment prices are often modeled as log normal random variables

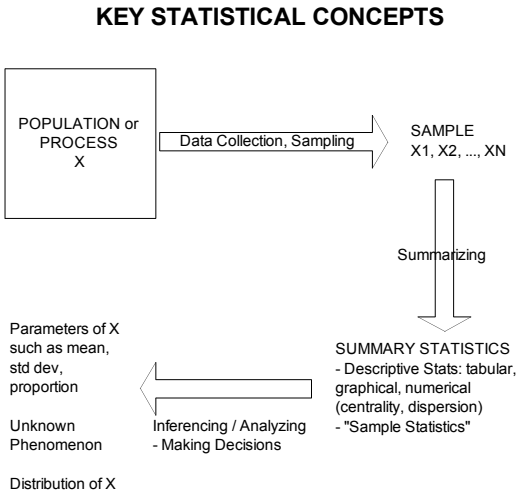
Sampling Distributions Recap

- Basis of **Statistical Inference**
- Particular sampling distribution depends on parameter (mean, variance, proportion, etc.)
  - Can take on many different forms
- Always based on **random sampling**!
- Knowing sampling distribution, and our sample statistics, we can make inferences about population parameter

Statistical Inference: Introduction

- Statistical inference is process by which we acquire information about **populations** from **samples**
- Two procedures for making inferences:
  - Estimation
  - Hypotheses testing
- Critical that sample is a **random sample**

Statistical Model

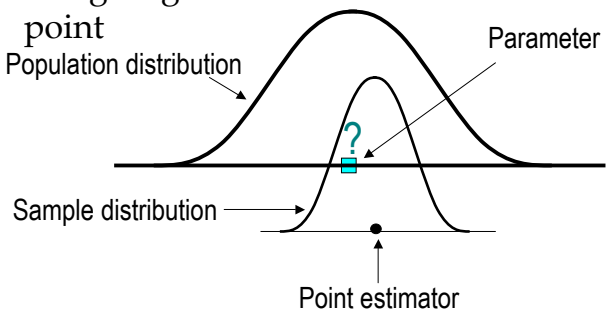


Concept of Estimation

- Objective of estimation is to determine value of a **population parameter** on the basis of a **sample statistic**
- Two types of estimators
  - Point Estimator
    - » Single value
  - Interval estimator
    - » Interval likely to contain true parameter value

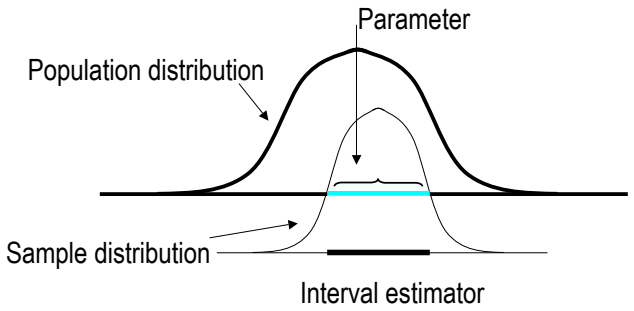
Point Estimator Example (Mean)

- Draws inference about parameter by using single value or point



Interval Estimator Example

- Interval estimator draws inferences about parameter by using an interval
- Interval estimator is affected by sample size



Confidence Statements

- In making estimations, **want to state how sure/unsure we are**
  - ***Level of significance*** = α
    - » used for hypothesis tests; .01, .05, .1 common
  - ***Confidence level*** = (1- α) x 100%
    - » used for confidence intervals; 90%, 95%, 99% common
- Random sample from pop. has (1- α) x 100% probability (or confidence) that confidence interval contains unknown pop. Parameter
  - 95% most commonly used
- In repeated samplings, (1- α) x 100% of conf. intervals will contain unknown pop. parameter

Est. Population Mean (σ Known)

- How is interval estimator produced from sampling distribution?
  - To estimate μ, sample of size n is drawn from population, and its mean (x-bar) is calculated
  - By CLT, x-bar is $N(\mu, \sigma/n^{1/2})$
  - Point estimate is sample mean (x-bar)
  - Interval is x-bar ± **so many std deviations**
    - » **#** of std dev depends on desired confidence level (how confident we want to be that interval contains population mean μ)

Commonly used Values

- Confidence level gives α (prob. of being wrong; tail area)
- Corresponds to a particular **# of standard deviations** around point estimate
  - commonly called "z-value"; obtain from Normal table or use Excel

Confidence level	α	α/2	(z) # std dev
0.90	0.10	0.05	1.645
0.95	0.05	0.025	1.96
0.98	0.02	0.01	2.33
0.99	0.01	0.005	2.575

Excel & z-values

- Based on standard Normal distribution (μ = 0, σ = 1)
- NORM**S**DIST(z)
  - Returns LHS probability for given z-value
- NORM**S**INV(α)
  - Returns z value for given LHS area
- **Drawing a diagram** helps, since tables are symmetrical and these Excel functions are not; **watch this!**

Confidence Interval for Mean (σ known)

$$\left[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

- How do we know x-bar, σ, n, $z_{\alpha/2}$?
- If we want to be 95% confident that this interval encapsulates population mean, what value of z would we use?
- What proportion of time will mean be outside this interval?

Marketing Example

- Number & types of television programs and commercials targeted at children are affected by amount of time children watch TV
  - Survey conducted among 100 North American children, in which they were asked to record # of hours they watched TV/week
  - Sample mean is 27.191 hrs
  - Population standard deviation of TV watch was known to be σ = 8.0 hrs
  - Estimate watch time with 95% confidence

Solution

- Parameter to be estimated is μ, mean time of TV watched per week per child
  - Need to compute interval estimator for μ

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 27.191 \pm z_{.025}\frac{8.0}{\sqrt{100}}$$

Since 1 - α =.95, α = .05.
Thus α/2 = .025. $Z_{.025}$ = 1.96

$$= 27.191 \pm 1.96\frac{8.0}{\sqrt{100}} = 27.191 \pm 1.57 = [25.621, 28.761]$$

LCL = 25.621, UCL = 28.761

Width of Interval Estimate

- Width of interval estimate is function of:
  - population standard deviation (σ)
  - confidence level (1 – α); z-value ($z_{\alpha/2}$)
  - sample size (n)
  - has nothing to do with N (pop. size)!
- Smaller is better (better understanding of where μ is)
  - ***How can we influence this?***

Inference About Pop. Mean (σ Unknown)

- How can we know σ when we are only working with a sample?
- When **σ** is unknown, use its point estimator **s**, and sampling distribution is **Student t** distribution
- Requires Normal (or approx Normal) population
- As sample size n increases, t-dist approaches N-dist (n > 200)

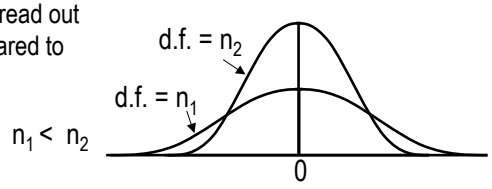
t-Statistic

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \qquad t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \qquad df = n - 1$$

When sampled population is normally distributed, statistic **t** is **Student t** distributed.

"Degrees of freedom"
- function of sample size
- determines how spread out distribution is (compared to normal distribution);
- = n - 1

t distribution is mound-shaped, and symmetrical around zero.



Confidence Interval for Mean (σ unknown, s known)

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

- This application of t-distribution is **robust**. It is an adequate approximate sampling distribution for moderately non-normal populations.
- Use Excel (or tables) for t-values

Excel & Student t-Distribution

- TDIST(x, df, tails)
  - returns "tail" probability (area under curve)
  - x = specific t-value
  - tails must be 1 or 2
- TINV(prob, df)
  - returns t-value for specified probability & degrees of freedom (**based on 2-tails**)
  - **df** = ***degrees of freedom*** (sample size minus one, n-1)
  - **prob** = α (tails area)

Practical vs Statistical Significance

- Take a large enough sample, and you can usually show statistically significant difference between two sample means (or other parameters such as variance or proportion)
- But is this difference of practical/economic significance?
  - Managerial decision re what is of practical significance

Paired Difference Test

- Use for difference of means, when samples are **not independent** (matched pairs)
  - More efficient!
  - Calculate difference for each sample pair
  - Then work with this new series of "paired differences"; same as working with a single sample
  - Symbols: μ_D and σ_D, x-bar$_D$ and s_D, n_D

Recap re Confidence Intervals

- Formula depends on **Parameter** of interest & its **Sampling Distribution**
  - Parameter defines sample statistic & sampling distribution (sometimes requirements re parent distribution)
  - We focused on μ; text also covers σ^2, p for 1 or 2 populations
- Point Estimate ± **so many std deviations** of sampling distribution
  - # of std dev depends on desired confidence level (1-α)*100%; 2-tailed, hence $\alpha/2$

Additional Topics (Level 1)

- Sampling & Estimation
  - Sampling in investment analysis
  - Data-snooping bias
  - Sample selection bias: survivorship, delisting
  - Using t-value tables
  - Conf. intervals when sampling from non-N population