Data Analytics Study Guide

Why Sample?

Most of Chapter 7 involves finding a statistical measure of a big data set like the mean, and then taking a small sample of the data and finding the mean again, and then comparing the two. Standard deviations, confidence intervals, all are about comparing the numbers we get from a small sample to the numbers we get from the entire data set. Why are we even doing this? If we have all the data, why sample at all?

1. In many cases 'all the data' is changing. Even if you have all the minute-to-minute stock prices for AAPL in your spreadsheet – that dataset is growing every minute. So is the weather data. It's never all in your sheet. You are always going to be sampling, like it or not. These are the statistical tools to do it accurately.

2. From sampling we learn how much **variability** exists between samples. If we took many random samples from the same data, how alike would they be? If you are trying to predict the future – the future data is going to be a 'sample' of the total data. How much will it vary from past samples?

Ch7_PracticeData.xlsm

The 'Actual Data' is the Employee | Salary | Training data from Chapter 7 of the book, expanded to 100 rows. I left my numbers in the sheet for clarity – make a copy, clear out the tables and overwrite them with your own experiments.

1. Start with Actual Data
2. Assign a random number to each row
3. Sort the rows
4. Take the top 30 rows → that's your Random Sample
5. Find the mean of that sample → xbar
6. Find the proportion who said Yes → pbar
7. Compare to the Actual Data

This gives you xbar and pbar for one sample – but it doesn't tell you about variance. Imagine there were 10,000 rows in Actual Data instead of 100. One 30-row sample doesn't tell us enough.

What is a Point Estimator (xbar and pbar) ?

Goal: find the population mean (x) and proportion (p)
Problem: We don't have all the data (N) so we can't
Solution: Choose a random sample of size n and find the sample mean (xbar) and the sample proportion (pbar). These are called **point estimators.**

Population Parameter	***Point Estimator***	***Example***
x (mean)	xbar (sample mean)	$40,000 avg salary
p (proportion)	pbar (sample proportion)	60% of respondents said 'Yes'

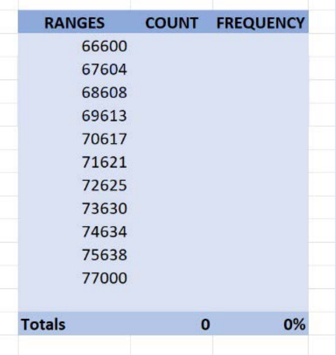
Ch7_PracticeData.xlsm

What if you **repeated** the steps above 100 times, found 100 xbar (or pbars), and averaged them all together. An average of an average. The **RunSamples** macro generates 100 samples of 30 each and computes xbar and pbar for each sample.

1. Run the 'RandomSamples' Macro
2. Creates a new 'Results' tab with a random number, like 'Results_911'
3. Runs 100 times
4. Each iteration selects 30 row random sample from Actual Data
5. Calculates xbar and pbar for that sample

Range = width of data = largest value – smallest value
Buckets = how many 'ranges' we want to divide the data into
Interval = how 'big' each bucket is

Make a results table like this:



RANGES	COUNT	FREQUENCY
66600		
67604		
68608		
69613		
70617		
71621		
72625		
73630		
74634		
75638		
77000		
Totals	0	0%

In the **Count** column use Excel Function to make a Frequency Distribution
=FREQUENCY(<xbar or pbar column> , <ranges column>)

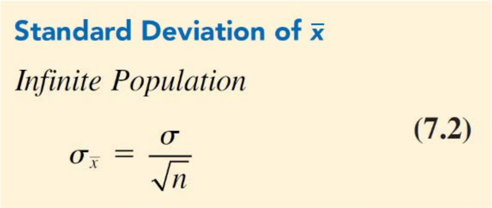
This is a **Relative Frequency Distribution** for the 100 values of xbar (or pbar). It tells us what percentage of the means fell within each range during the trial runs – and for any 'trial run' in the future which buckets the mean will *probably* fall into.

RandomSamples runs 100 times. The problem is, there are *many* more than 100 different combinations:

N = 100 rows
r = 30 sample size
Total possible samples = n! / r! (n – r)! = **29,000,000,000,000,000,000,000,......**

Because there are so many possible samples, any one sample we generate will be arbitrary. Any value for xbar and pbar will be a **random number** which may or may not match the actual population parameter. So in addition to a value for xbar and pbar we have:

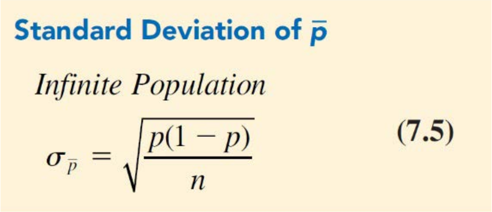
1. **Standard deviation of the point estimator**
   How much do xbar / pbar vary from the true population x / p across all the random samples



Standard Deviation of $\bar{x}$

Infinite Population

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \qquad \textbf{(7.2)}$$



Standard Deviation of $\bar{p}$

Infinite Population

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} \qquad \textbf{(7.5)}$$

If: $\sigma_{xbar} = 2$

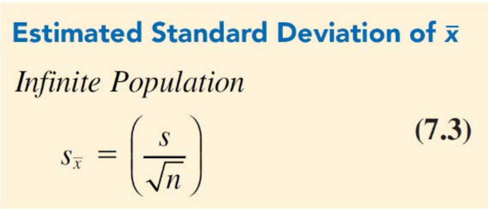
Then: across all possible random samples of size 30 from our population of 100, the sample mean xbar is expected to vary by +/- 2 units from the population mean x.

But: The right sides of the equations require the population standard deviation (σ) and the actual population proportion (p)

So: We use…..

2. ***Estimated* standard deviation**

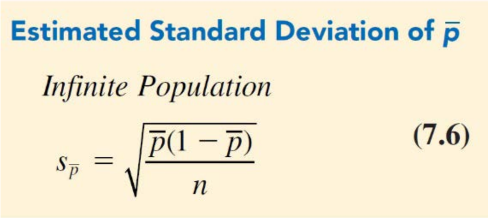
Because we often **don't know** the population σ and p, we calculate the **estimated** standard deviation for xbar and pbar using the sample standard deviation (s) and sample proportion (pbar).



Estimated Standard Deviation of $\bar{x}$

Infinite Population

$$s_{\bar{x}} = \left(\frac{s}{\sqrt{n}}\right) \qquad \textbf{(7.3)}$$



Estimated Standard Deviation of $\bar{p}$

Infinite Population

$$s_{\bar{p}} = \sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \qquad \textbf{(7.6)}$$

s = sample standard deviation
pbar = sample proportion

Standard Deviation of xbar / pbar = Standard Error (SE)
xbar / pbar is expected to vary by +/- SE from the true population x / p.

What's a Sampling Distribution?

Problem: We still don't have the actual population mean (x) and proportion (p)

So we imagine a huge set containing all the possible random samples – and all the possible sample means. And we create a chart showing how likely each mean is by how common it is in the set (its frequency). Most of the means cluster in the middle, and therefore have the highest frequency. The outlier means are distributed along the edges with very low frequency because they occur very seldomly. This is the **sampling distribution.**

The SE gives us how spread apart the values are from the middle. A small SE means the xbars are clustered near the center. A large SE means they are spread out across the distribution.

If the sample size is small relative to the population the SE will be large, to account for outlier sample means. As the sample size increases to match the population, SE decreases, the spread becomes narrower, and the xbars/pbars vary less and less from actual x/p.

Sampling Distribution
Ingredients:

1. Expected value population param (if available)
   sample param (if not)

2. Standard Error standard deviation *of* xbar or pbar

3. Probability distribution (Normal if n >= 30)

What is a Normal Distribution?
pg 339

If the distribution is 'normal' we know it's a bell curve and we know the intervals. For any normally distributed random variable, 90% of the values lie within 1.645 standard deviations of the mean, 95% of the values lie within 1.960 standard deviations of the mean, and 99% of the values lie within 2.576 standard deviations of the mean.

sample mean = population mean
sample std dev (Standard Error) = <u>population std dev</u> / sqrt(sample size)

←----------------------------- 90% Confidence Interval ---------------------------------→
- (1.645 * SE) sample mean + (1.645 * SE)

we call 1.645 in this case the **z-value.**

Problem: Usually we don't know the population std dev (b/c we don't have all the data) so we need another statistical tool to estimate the Standard Error.

What is the T-value?
Estimate the Standard Error using the *sample* standard deviation (s).

SE’ = s / sqrt(n)

s = sample standard deviation
n = sample size

AND THEN get the **T-Value**

Excel: t = T.INV.2T(1 - confidence , degrees)

Ingredients:

1 - confidence = .05 (100 - .05 = 95% probability)
n = sample size
degrees of freedom = n – 1

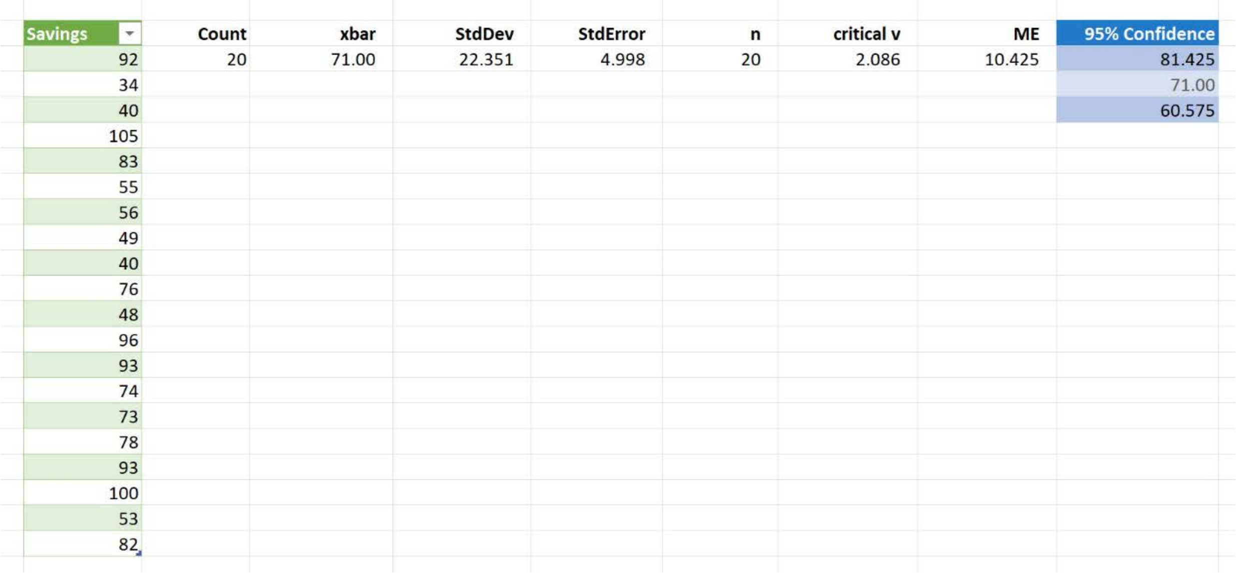
What is Margin of Error?

← ---------------------------- Range of xbar / pbar ------------------------------- →

- margin of error	**population param**	**+ margin of error**

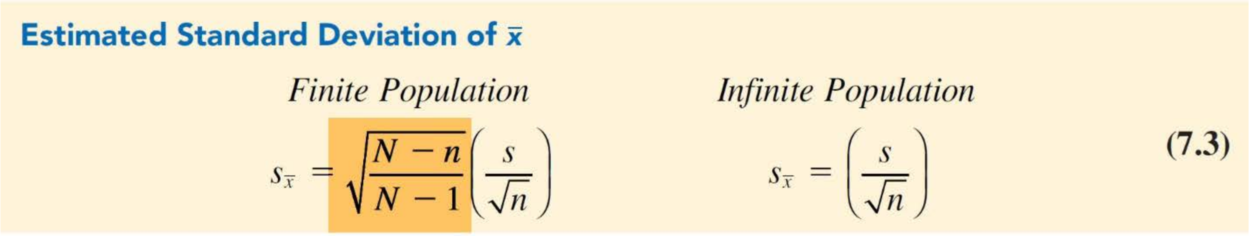
ME = t * SE’
t = T-value (for 95%)
SE’ = estimated Standard Error

95% Confidence interval = xbar +/- ME



Savings	Count	xbar	StdDev	StdError	n	critical v	ME	95% Confidence
92	20	71.00	22.351	4.998	20	2.086	10.425	81.425
34								71.00
40								60.575
105								
83								
55								
56								
49								
40								
76								
48								
96								
93								
74								
73								
78								
93								
100								
53								
82								

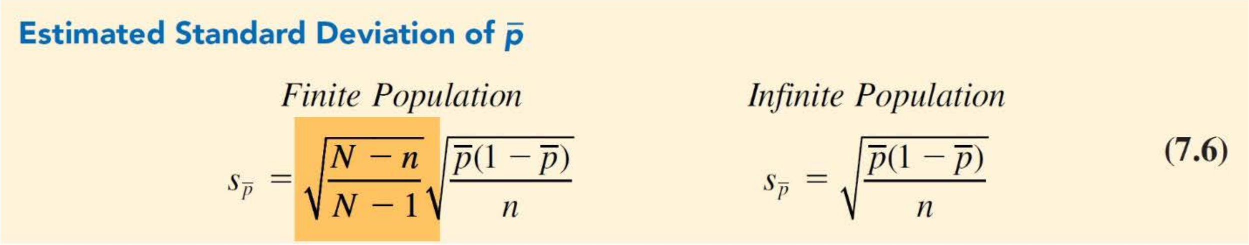
When To Use: The Finite Population Correction Factor
pgs 332



Estimated Standard Deviation of $\bar{x}$

Finite Population *Infinite Population*

$$s_{\bar{x}} = \sqrt{\frac{N-n}{N-1}}\left(\frac{s}{\sqrt{n}}\right) \qquad s_{\bar{x}} = \left(\frac{s}{\sqrt{n}}\right) \tag{7.3}$$



Estimated Standard Deviation of $\bar{p}$

Finite Population *Infinite Population*

$$s_{\bar{p}} = \sqrt{\frac{N-n}{N-1}}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \qquad s_{\bar{p}} = \sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \tag{7.6}$$

Compare the sample size (n) to the population size (N).

n/N > 0.05 → sample is a large % of the total
use the correction factor

n/N <= 0.05 → sample is a small % of the total
don't use

total might as well be infinite by comparison