

Chapter 4: Sampling, Estimation, Simulation

Statistics, Data Analysis, and
Decision Modeling, Fifth Edition
James R. Evans



Need for Sampling

- Very large populations
- Destructive testing
- Continuous production process

The objective of sampling is to draw a valid inference about a population.



Sample Design

- **Sampling Plan** – a description of the approach that will be used to obtain samples from a population
 - Objectives
 - Target population
 - Population frame
 - Method of sampling
 - Operational procedures for data collection
 - Statistical tools for analysis



Sampling Methods

- Subjective
 - Judgment sampling
 - Convenience sampling
- Probabilistic
 - Simple random sampling – every subset of a given size has an equal chance of being selected



Other Sampling Methods

- Systematic sampling
- Stratified sampling
- Cluster sampling
- Sampling from a continuous process



Errors in Sampling

- Nonsampling error
 - Poor sample design
- Sampling (statistical) error
 - Depends on sample size
 - Tradeoff between cost of sampling and accuracy of estimates obtained by sampling



Random Sampling from Probability Distributions

- Random number - uniformly distributed between 0 and 1
- Excel function RAND()

TABLE 4.1 One Hundred Random Numbers

0.007120	0.215576	0.386009	0.201736	0.457990	0.127602	0.387275	0.639298	0.757161	0.285388
0.714281	0.165519	0.768911	0.687736	0.466579	0.481117	0.260391	0.508433	0.528617	0.755016
0.226987	0.454259	0.487024	0.269659	0.531411	0.197874	0.527788	0.613126	0.716988	0.747900
0.339398	0.434496	0.398474	0.622505	0.829964	0.288727	0.801157	0.373983	0.095900	0.041084
0.692488	0.137445	0.054401	0.483937	0.954835	0.643596	0.970131	0.864186	0.384474	0.134890
0.962794	0.808060	0.169243	0.347993	0.848285	0.216635	0.779147	0.216837	0.768370	0.371613
0.824428	0.919011	0.820195	0.345563	0.989111	0.269649	0.433170	0.369070	0.845632	0.158662
0.428903	0.470202	0.064646	0.100007	0.379286	0.183176	0.180715	0.008793	0.569902	0.218078
0.951334	0.258192	0.916104	0.271980	0.330697	0.989264	0.770787	0.107717	0.102653	0.366096
0.635494	0.395185	0.320618	0.003049	0.153551	0.231191	0.737850	0.633932	0.056315	0.281744

Sampling from Discrete Probability Distributions

- Rolling two dice

x	$f(x)$	$F(x)$
2	0.028	0.028
3	0.056	0.083
4	0.083	0.167
5	0.111	0.278
6	0.139	0.417
7	0.167	0.583
8	0.139	0.722
9	0.111	0.833
10	0.083	0.917
11	0.056	0.972
12	0.028	1.000

Interval			Outcome
0	to	0.028	2
0.028	to	0.083	3
0.083	to	0.167	4
0.167	to	0.278	5
0.278	to	0.417	6
0.417	to	0.583	7
0.583	to	0.722	8
0.722	to	0.833	9
0.833	to	0.917	10
0.917	to	0.972	11
0.972	to	1.000	12

Select random number, compare to interval, assign outcome



Sampling from Common Probability Distributions

- **Random variate** - a value randomly generated from a specified probability distribution
- We can transform random numbers into random variates using the cumulative distribution function or special formulas
 - Uniform random variate:

$$U = a + (b - a) * \text{RAND}()$$



Random Variates in Excel

- `NORM.INV(RAND(), mean, standard_deviation)`
- `NORM.S.INV(RAND())`
- `LOGNORM.INV(RAND(), mean, standard_deviation)`
- `BETA.INV(RAND(), alpha, beta, A, B)`
- `GAMMA.INV(RAND(), alpha, beta)`



Statistical Sampling Experiment in Finance

- In finance, one way of evaluating capital budgeting projects is to compute a profitability index (PI), which is defined as the ratio of the present value of future cash flows (PV) to the initial investment (I):

$$PI = PV/I$$

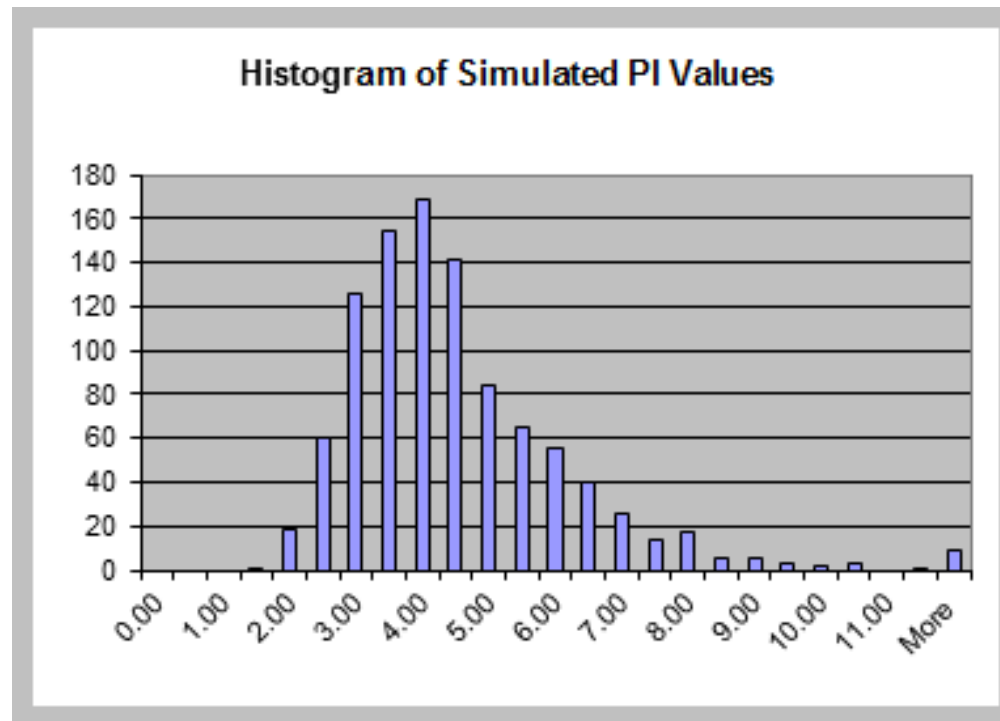
- Suppose that PV is estimated to be normally distributed with a mean of \$12 million and a standard deviation of \$2.5 million, and the initial investment is also estimated to be normal with a mean of \$3.0 million and standard deviation of \$0.8 million. What is the distribution of PI ?

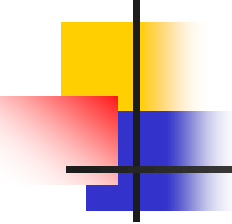


Experiment

	A	B	C	D	E
1	Profitability Index Analysis				
2					
3		Mean	Standard Deviation		
4	PV	12	2.5		
5	I	3	0.8		
6					
7	Experiment	PV	I	PI	Mean
8	1	8.396743042	3.573822001	2.349513501	4.762285
9	2	11.7446542	3.66554571	3.204067043	
10	3	11.76586862	3.554538257	3.310097619	
11	4	11.44456518	3.33708406	3.429510606	
12	5	9.373641185	3.692222659	2.538752955	
13	6	10.47906344	2.598868941	4.0321631	
14	7	14.31716958	3.203954788	4.46859289	
15	8	8.901052248	0.729081227	12.20858791	
16	9	13.99414343	3.180751244	4.399634662	
17	10	12.5758327	3.513579887	3.579207847	

Histogram of Simulated PI Values





Sampling Distributions and Sampling Error

- How good is an estimate of the mean obtained from a sample?
- Sampling distribution of the mean is the distribution of the means of all possible samples of a fixed size from some population.
- Understanding sampling distributions is the key to statistical inference.



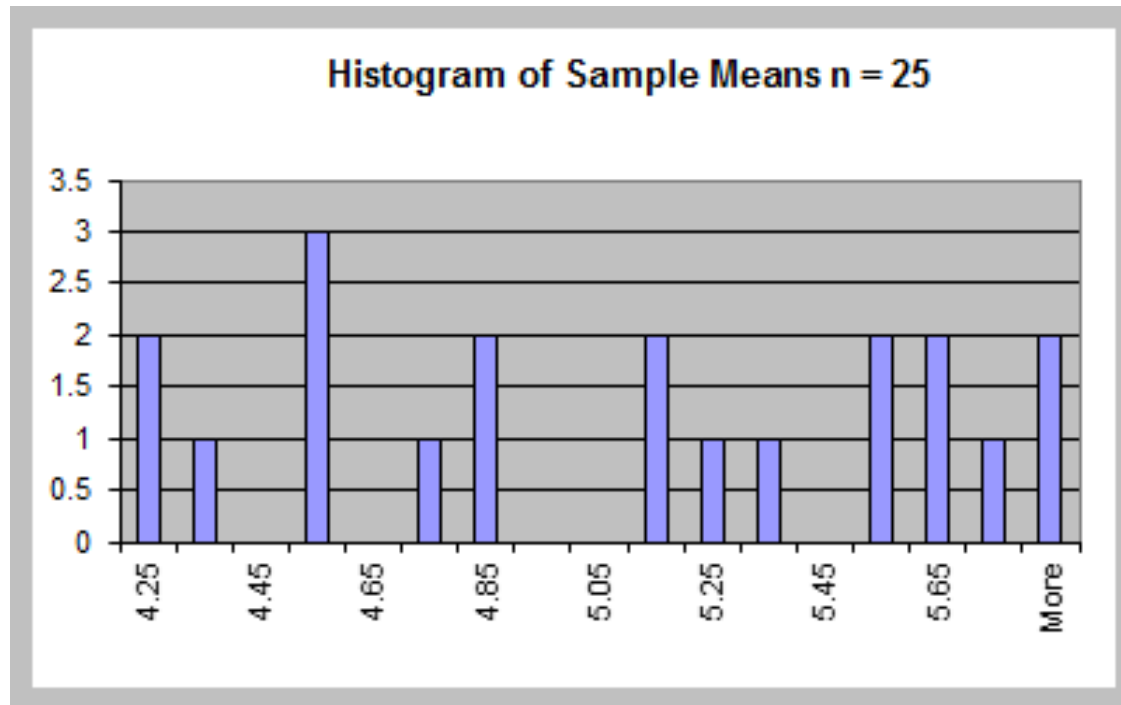
Experiment

- Assume that a random variable is uniformly distributed between 0 and 10. The expected value is $(0 + 10)/2 = 5$, the variance is $(10 - 0)^2/12 = 8.33$, and the standard deviation is 2.89.
- Generate a sample of size $n = 25$ for this random variable using the Excel function `10*RAND()` and compute the sample mean. Repeat this experiment several more times, say 20, and obtain a set of 20 sample means.

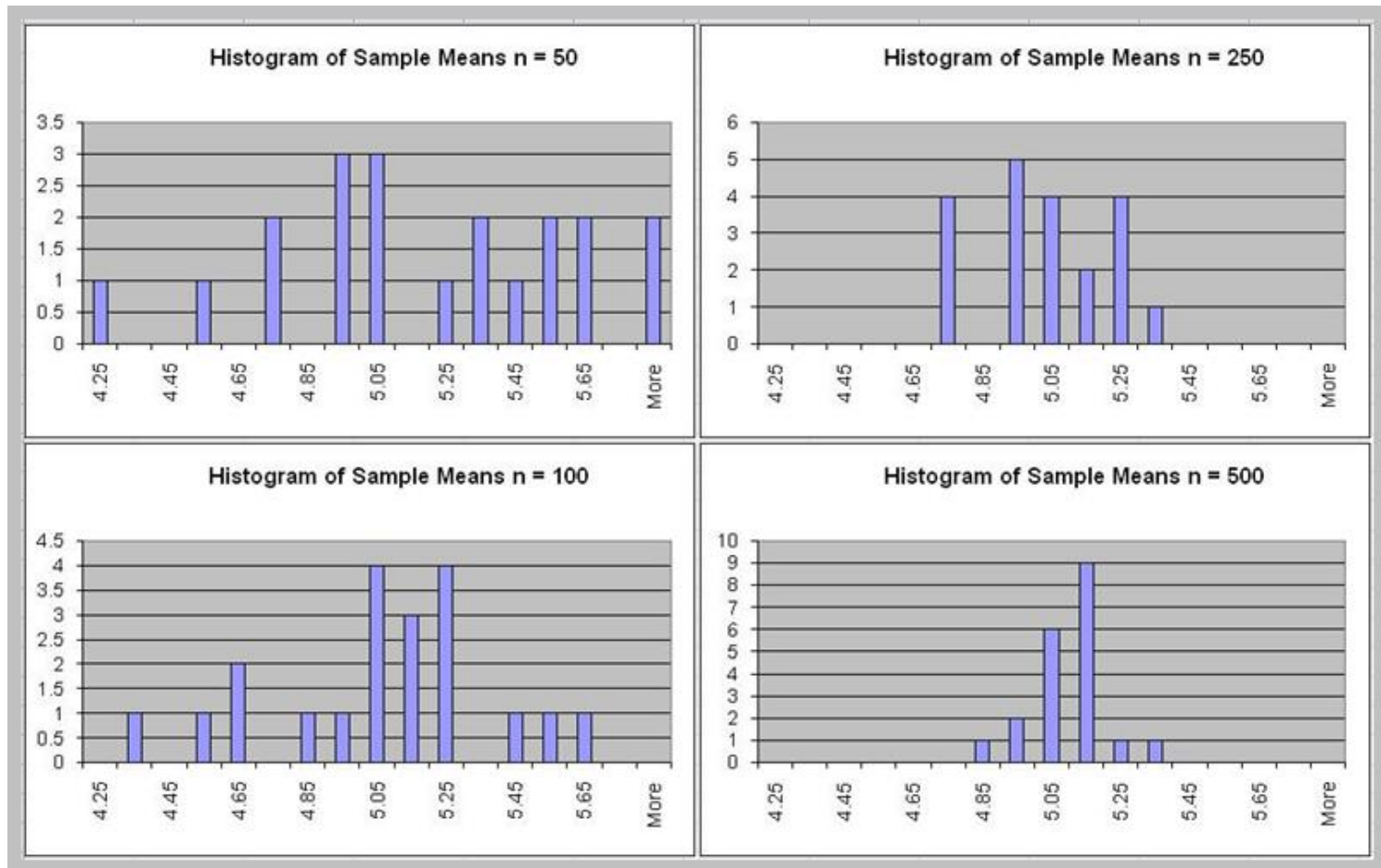
Sampling Experiment Spreadsheet

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Sampling Error Experiment																				
2	Instructions: The worksheet is designed for 20 samples with sample sizes of up to 500. To change the sample size, simply change the																				
3	range in the formulas in row 6 for computing the sample mean to include the appropriate number of observations.																				
4	Pressing the F9 key will recalculate all values.																				
5	Experiment	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
6	Sample Mean	5.011	4.774	4.983	4.808	5.130	5.030	5.118	4.958	5.155	5.128	5.031	5.014	4.821	4.838	4.853	5.070	5.072	5.020	4.953	5.109
7																					
8	Sample																				
9	1	1.528	8.061	7.232	2.881	6.074	8.949	9.507	2.190	7.463	2.487	2.878	2.504	2.927	4.982	6.996	0.999	8.756	6.911	7.891	2.572
10	2	5.299	7.018	0.975	1.298	6.561	0.731	1.612	6.376	5.861	5.613	6.556	3.836	1.231	6.969	7.960	7.285	5.870	4.508	1.678	7.679
11	3	6.174	2.333	4.188	4.832	2.992	7.205	8.786	1.440	4.328	7.471	0.459	2.217	7.020	3.243	6.356	7.152	2.631	2.845	9.526	8.879
12	4	3.609	4.155	3.552	0.188	0.619	1.766	5.173	4.625	0.339	4.067	5.250	9.360	7.693	0.239	6.761	0.449	5.816	7.534	8.348	0.074
13	5	9.000	0.636	8.796	0.916	9.754	1.690	4.458	5.460	2.032	9.930	9.107	0.522	1.899	1.672	8.205	6.523	7.115	5.267	0.558	8.224

Histogram of 20 Sample Means



Results for Other Sample Sizes





Summary

TABLE 4.2 Results from Sampling Error Experiment

Sample Size	Average of 20 Sample Means	Standard Deviation of 20 Sample Means
25	5.166	0.639449
50	4.933	0.392353
100	4.948	0.212670
250	5.028	0.167521
500	4.997	0.115949

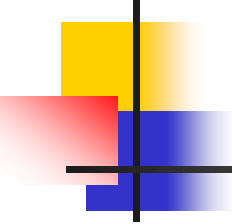
Key observations:

1. All means appear close to the expected value of 5.
2. The standard deviation gets smaller as the sample size increases.



Sampling Distribution of the Mean

- Sampling distribution of the mean is the distribution of the means of all possible samples of a fixed size from some population.
- Understanding sampling distributions is the key to statistical inference.



Properties of the Sampling Distribution of the Mean

- Expected value of the sample mean is the population mean, μ
- Variance of the sample mean is σ^2/n , where σ^2 is the variance of the population
- Standard deviation of the sample mean, called the **standard error of the mean**, is σ/\sqrt{n}



Central Limit Theorem

- If the sample size is large enough (generally at least 30, but depends on the actual distribution), the sampling distribution of the mean is approximately normal, *regardless* of the distribution of the population.
- If the population is normal, then the sampling distribution of the mean is exactly normal for any sample size.



Applying Sampling Distributions – Example 1

- Suppose that the size of individual customer orders (in dollars), X , from a major discount book publisher Web site is normally distributed with a mean of \$36 and standard deviation of \$8. What is the probability that the next individual who places an order at the Web site will purchase more than \$40?
- $P(X > 40) = 1 - \text{NORMDIST}(40, 36, 8, \text{TRUE})$
 $= 1 - 0.6915 = 0.3085$
- The calculation uses the standard deviation of the individual customer orders.



Applying Sampling Distributions – Example 2

- Suppose that a sample of 16 customers is chosen. What is the probability that the *mean purchase* for these 16 customers will exceed \$40?
- The sampling distribution of the mean will have a mean of \$36, but a standard error of $\$8/\sqrt{16} = \2 .
- $P(\bar{X} > 40) = 1 - \text{NORMDIST}(40, 36, 2, \text{TRUE})$
 $= 1 - .9772 = 0.0228$
- The calculation uses the standard error of the sampling distribution.



Sampling and Estimation

- **Estimation** – assessing the value of an unknown population parameter using sample data.
- **Point estimate** – a single number used to estimate a population parameter
- **Confidence interval estimate** – a range of values between which a population parameter is believed to be along with the probability that the interval correctly estimates the true population parameter



Common Point Estimates

TABLE 4.4 Common Point Estimates

Point Estimate	Population Parameter
Sample mean, \bar{x}	Population mean, μ
Sample variance, s^2	Population variance, σ^2
Sample standard deviation, s	Population standard deviation, σ
Sample proportion, \hat{p}	Population proportion, π



Example

	A	B	C	D	E	F
1	Cereal Name	Calories	Sodium	Fiber	Carbs	Sugars
2	Lucky Charms	110	180	0	12	12
3	Apple Jacks	110	125	1	11	14
4	Apple Cinn Cheerios	110	180	1.50	10.50	10
5	Crispy Wheat & Raisins	100	140	2	11	10
6	Honey Graham Ohs	120	220	1	12	11
7	Puffed Wheat	50	0	1	10	0
8	Honey Nut Cheerios	110	250	1.50	11.50	10
9	Trix	110	140	0	13	12
10	Post Nat. Raisin Bran	120	200	6	11	14
11	Almond Delight	110	200	1	14	8
12	Sample Mean	105	163.5	1.5	11.6	10.1
13	Sample Standard Deviation	20.1384	69.284	1.7	1.197	4.0125
14						
15	Population Mean	105.522	167.313	2.187	14.77	6.9552
16	Population Standard Deviation	18.631	79.9782	2.487	3.831	4.3758



Theoretical Issues

- **Unbiased estimator** – one for which the expected value equals the population parameter it is intended to estimate
- The sample variance is an unbiased estimator for the population variance

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

$$\sigma^2 = \frac{\sum_{i=1}^n (x_i - \mu)^2}{N}$$



Interval Estimates

- Range within which we believe the true population parameter falls
 - Example: Gallup poll – percentage of voters favoring a candidate is 56% with a 3% margin of error.
 - Interval estimate is [53%, 59%]
- A $100(1 - \alpha)\%$ probability interval is any interval $[A, B]$ such that the probability of falling between A and B is $1 - \alpha$



Confidence Intervals

- **Confidence interval (CI)** – an interval estimated that specifies the likelihood that the interval contains the true population parameter
- **Level of confidence** ($1 - \alpha$) – the likelihood that the CI contains the true population parameter, usually expressed as a percentage (90%, 95%, 99% are most common).



Confidence Intervals and *PHStat* Tools

TABLE 4.5 Common Confidence Intervals

Type of Confidence Interval	<i>PHStat</i> Tool
Mean, standard deviation known	Estimate for the mean, sigma known
Mean, standard deviation unknown	Estimate for the mean, sigma unknown
Proportion	Estimate for the proportion
Variance	Estimate for the population variance
Population total	Estimate for the population total



Confidence Interval for the Mean – σ Known

$$\bar{x} \pm z_{\alpha/2}(\sigma / \sqrt{n}) \quad (4.4)$$

$z_{\alpha/2}$ may be found from Table A.1 or using the Excel function NORM.S.INV($\alpha/2$)



Sampling From Finite Populations

- When $n > 0.05N$, use a correction factor in computing the standard error:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

Example: *Cereal* Samples (*PHStat* Tool)

	A	B
1	Calories	
2		
3	Data	
4	Population Standard Deviation	18.631
5	Sample Mean	105
6	Sample Size	10
7	Confidence Level	95%
8		
9	Intermediate Calculations	
10	Standard Error of the Mean	5.891639509
11	Z Value	-1.95996398
12	Interval Half Width	11.54740125
13		
14	Confidence Interval	
15	Interval Lower Limit	93.45259875
16	Interval Upper Limit	116.5474012
17		
18		
19	Finite Populations	
20	Population Size	67
21	FPC Factor	0.929320377
22	Interval Half Width	10.73123528
23	Interval Lower Limit	94.26876472
24	Interval Upper Limit	115.7312353

$$\bar{x} \pm z_{\alpha/2}(\sigma / \sqrt{n})$$



Key Observations

$$\bar{x} \pm z_{\alpha/2}(\sigma / \sqrt{n})$$

- As the confidence level $(1 - \alpha)$ increases, the width of the confidence interval also increases.
- As the sample size increases, the width of the confidence interval decreases.

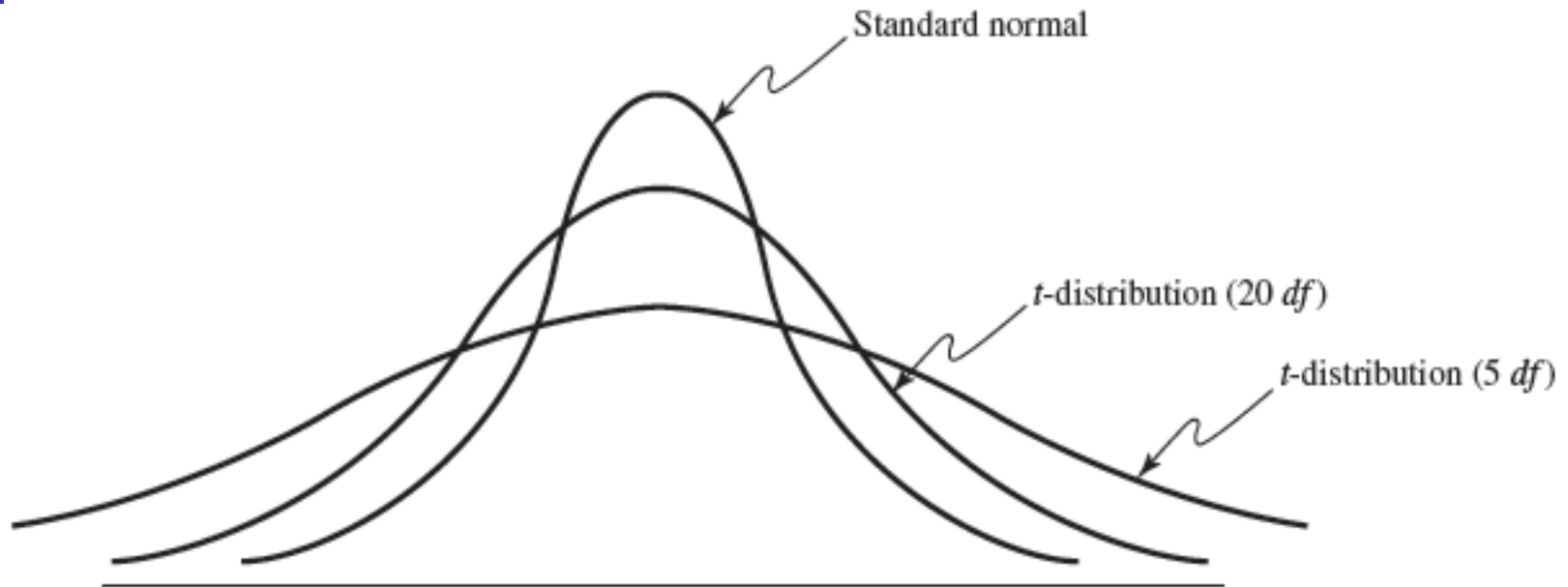


Confidence Interval for the Mean, σ Unknown

$$\bar{x} \pm t_{\alpha/2, n-1} \left(s / \sqrt{n} \right) \quad (4.6)$$

$t_{\alpha/2, n-1}$ is the value from a t-distribution with $n-1$ degrees of freedom, from Table A.2 or the Excel function T.INV($1-\alpha/2$, $n-1$)

Relationship Between Normal Distribution and t-distribution



The t-distribution yields larger confidence intervals for smaller sample sizes.

Example: *Credit Approval Decisions*

	A	B	C	D	E	F
1	Credit Approval Decisions					
2						
3	Homeowner	Credit Score	Years of Credit History	Revolving Balance	Revolving Utilization	Decision
4	Y	725	20	\$ 11,320	25%	Approve
5	Y	573	9	\$ 7,200	70%	Reject
6	Y	677	11	\$ 20,000	55%	Approve
7	N	625	15	\$ 12,800	65%	Reject
8	N	527	12	\$ 5,700	75%	Reject
9	Y	795	22	\$ 9,000	12%	Approve
10	N	733	7	\$ 35,200	20%	Approve

	A	B
1	Revolving Balance - Homeowners	
2		
3	Data	
4	Sample Standard Deviation	5393.384467
5	Sample Mean	12630.37037
6	Sample Size	27
7	Confidence Level	95%
8		
9	Intermediate Calculations	
10	Standard Error of the Mean	1037.957325
11	Degrees of Freedom	26
12	t Value	2.055529439
13	Interval Half Width	2133.551837
14		
15	Confidence Interval	
16	Interval Lower Limit	10496.82
17	Interval Upper Limit	14763.92



Sample Proportion

- **Sample proportion:** $\hat{p} = x/n$
 - x = number in sample having desired characteristic
 - n = sample size
- The sampling distribution of \hat{p} has mean π and variance $\pi(1 - \pi)/n$
- When $n\pi$ and $n(1 - \pi)$ are at least 5, the sampling distribution of \hat{p} approach a normal distribution



Confidence Interval for the Proportion

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \quad (4.7)$$

PHStat tool is available under *Confidence Intervals* option

Example: *Insurance Survey*

	A	B	C	D	E	F	G
1	Insurance Survey						
2							
3	Age	Gender	Education	Marital Status	Years Employed	Satisfaction*	Premium/Deductible**
4	36	F	Some college	Divorced	4	4	N
5	55	F	Some college	Divorced	2	1	N
6	61	M	Graduate degree	Widowed	26	3	N
7	65	F	Some college	Married	9	4	N
8	53	F	Graduate degree	Married	6	4	N
9	50	F	Graduate degree	Married	10	5	N
10	28	F	College graduate	Married	4	5	N
11	62	F	College graduate	Divorced	9	3	N
12	48	M	Graduate degree	Married	6	5	N
13	31	M	Graduate degree	Married	1	5	N
14	57	F	College graduate	Married	4	5	N
15	44	M	College graduate	Married	2	3	N
16	38	M	Some college	Married	3	2	N
17	27	M	Some college	Married	2	3	N
18	56	M	Graduate degree	Married	4	4	Y
19	43	F	College graduate	Married	5	3	Y
20	45	M	College graduate	Married	15	3	Y
21	42	F	College graduate	Married	12	3	Y
22	29	M	Graduate degree	Single	10	5	N
23	28	F	Some college	Married	3	4	Y
24	36	M	Some college	Divorced	15	4	Y
25	49	F	Graduate degree	Married	2	5	N
26	46	F	College graduate	Divorced	20	4	N
27	52	F	College graduate	Married	18	2	N
28	*Measured from 1-5 with 5 being highly satisfied.						
29	**Would you be willing to pay a lower premium for a higher deductible?						

	A	B
1	Insurance Survey	
2		
3	Data	
4	Sample Size	24
5	Number of Successes	6
6	Confidence Level	95%
7		
8	Intermediate Calculations	
9	Sample Proportion	0.25
10	Z Value	-1.95996398
11	Standard Error of the Proportion	0.088388348
12	Interval Half Width	0.173237978
13		
14	Confidence Interval	
15	Interval Lower Limit	0.076762022
16	Interval Upper Limit	0.423237978



Sampling Distribution of s

- The sample standard deviation, s , is a point estimate for the population standard deviation, σ
- The sampling distribution of s has a **chi-square (χ^2) distribution** with $n-1$ df
 - See Table A.3
 - `CHISQ.DIST(x , $deg_freedom$)` returns probability to the right of x
 - `CHISQ.INV($probability$, $deg_freedom$)` returns the value of x for a specified right-tail probability



Confidence Interval for the Variance

$$\left[\frac{(n-1)s^2}{\chi_{n-1, \alpha/2}^2}, \frac{(n-1)s^2}{\chi_{n-1, 1-\alpha/2}^2} \right] \quad (4.8)$$



Note the difference in the denominators!

Example: *Home Market Value* Data

	A	B	C	D	E
1	Home Market Value				
2					
3	Data				
4	Sample Size	42			
5	Sample Standard Deviation	10553.1			
6	Confidence Level	95%			
7					
8	Intermediate Calculations				
9	Degrees of Freedom	41			
10	Sum of Squares	4.57E+09			
11	Single Tail Area	0.025			
12	Lower Chi-Square Value	25.21452			
13	Upper Chi-Square Value	60.56057			
14					
15	Results				
16	Interval Lower Limit for Variance	7.5E+07			
17	Interval Upper Limit for Variance	1.8E+08			
18					
19	Interval Lower Limit for Standard Deviation	8683.14			
20	Interval Upper Limit for Standard Deviation	13456.9			
21					
22	Assumption:				
23	Population from which sample was drawn has an approximate normal distribution.				



Confidence Interval for a Population Total

$$N\bar{x} \pm t_{\alpha/2, n-1} N \frac{s}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \quad (4.9)$$

Example

- Sample of 20 reimbursement claims over 60 days old yielded a mean of \$185 and standard deviation of \$22.

	A	B
1	Unpaid Reimbursement Claims	
2		
3	Data	
4	Population Size	180
5	Sample Mean	185
6	Sample Size	20
7	Sample Standard Deviation	22
8	Confidence Level	95%
9		
10	Intermediate Calculations	
11	Population Total	33300.00
12	FPC Factor	0.945438918
13	Standard Error of the Total	837.1700134
14	Degrees of Freedom	19
15	t Value	2.09302405
16	Interval Half Width	1752.22
17		
18	Confidence Interval	
19	Interval Lower Limit	31547.78
20	Interval Upper Limit	35052.22



Confidence Intervals and Decision Making

- Required weight for a soap product is 64 ounces. A sample of 30 boxes found a mean of 63.82 and standard deviation of 1.05. A 95% CI is [63.43, 64.21]. What conclusion can you reach?
- What if the standard deviation was 0.46 and the CI is [64.65, 63.99]?



Confidence Intervals and Sample Size

- CI for the mean, σ known
 - Sample size needed for half-width of at most E is $n \geq (z_{\alpha/2})^2(\sigma^2)/E^2$
- CI for a proportion
 - Sample size needed for half-width of at most E is
$$n \geq \frac{(z_{\alpha/2})^2 \pi(1 - \pi)}{E^2}$$
 - Use the sample proportion as an estimate of π , or 0.5 for the most conservative estimate



Example

- For the soap filling process, the sampling error associated with the CI = [63.43, 64.21] is 0.39 (half-width). What sample size is required to reduce this to 0.15?

	A	B
1	Sample Size Determination	
2		
3	Data	
4	Population Standard Deviation	1.05
5	Sampling Error	0.15
6	Confidence Level	95%
7		
8	Intermediate Calculations	
9	Z Value	-1.95996398
10	Calculated Sample Size	188.2314822
11		
12	Result	
13	Sample Size Needed	189



Example

- How many voters must be surveyed to ensure a sampling error of at most $\pm 2\%$?

	A	B
1	Sample Size for the Proportion	
2		
3	Data	
4	Estimate of True Proportion	0.5
5	Sampling Error	0.02
6	Confidence Level	95%
7		
8	Intermediate Calculations	
9	Z Value	-1.95996398
10	Calculated Sample Size	2400.911763
11		
12	Result	
13	Sample Size Needed	2401



Prediction Intervals

- A **prediction interval** is one that provides a range for predicting the value of a new observation from the same population.
- A $100(1 - \alpha)\%$ prediction interval for a new observation is

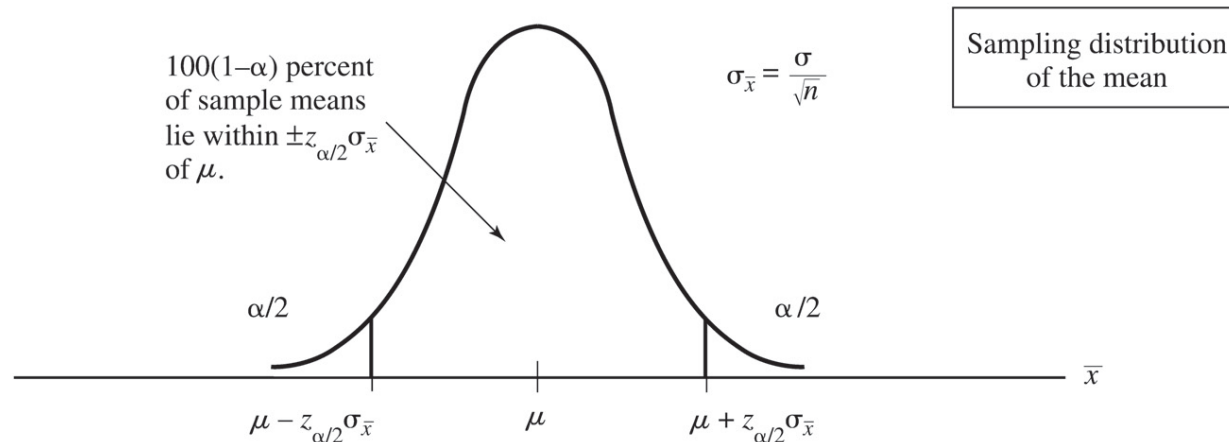
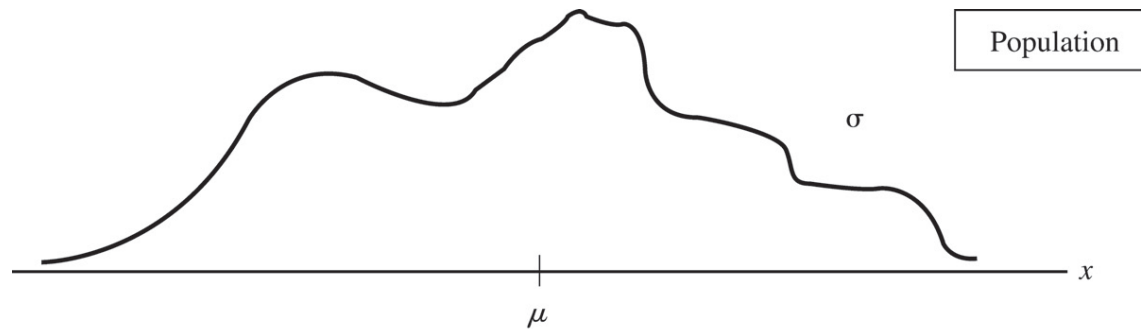
$$\bar{x} \pm t_{\alpha/2, n-1} (s \sqrt{1 + 1/n}) \quad (4.10)$$



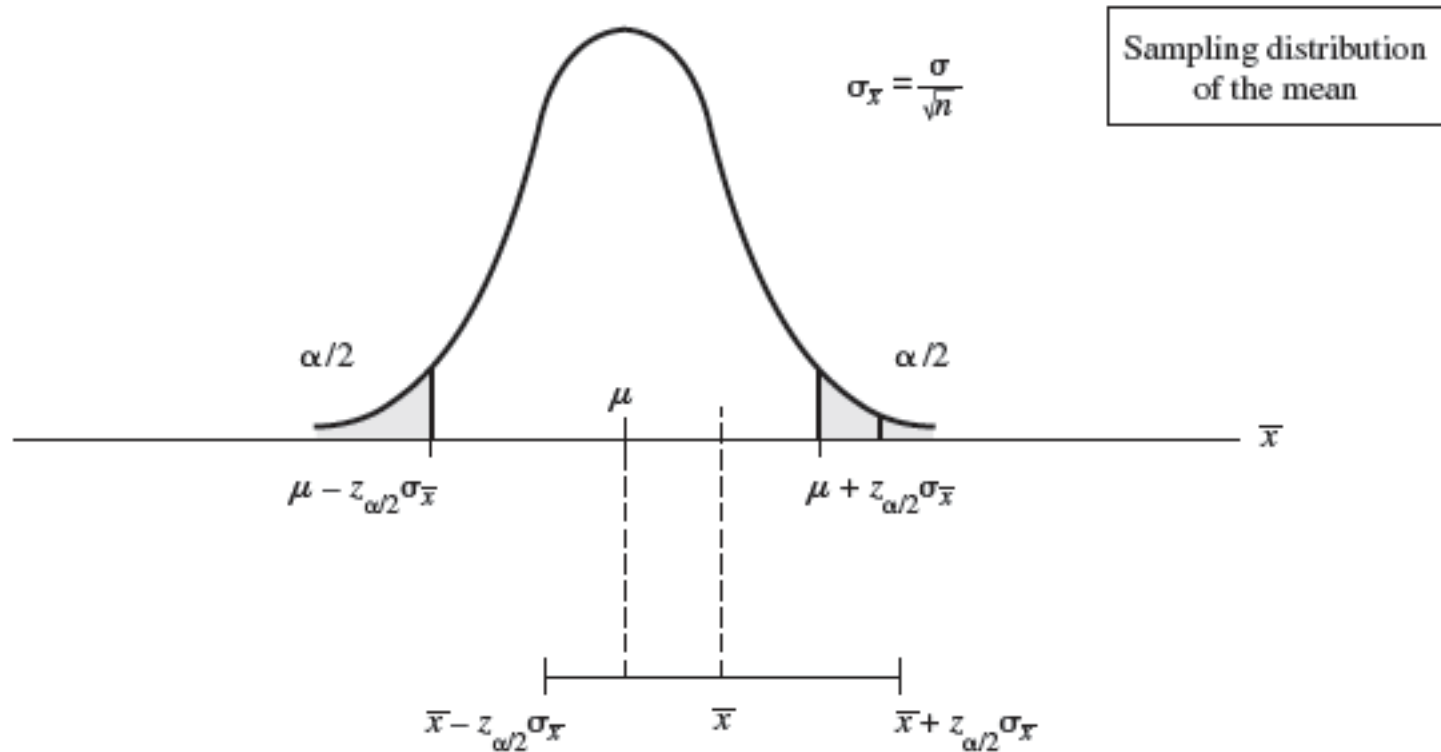
Additional Types of Confidence Intervals

- Difference in means
 - Independent samples with unequal variances
 - Independent samples with equal variances
 - Paired samples
- Difference in proportions

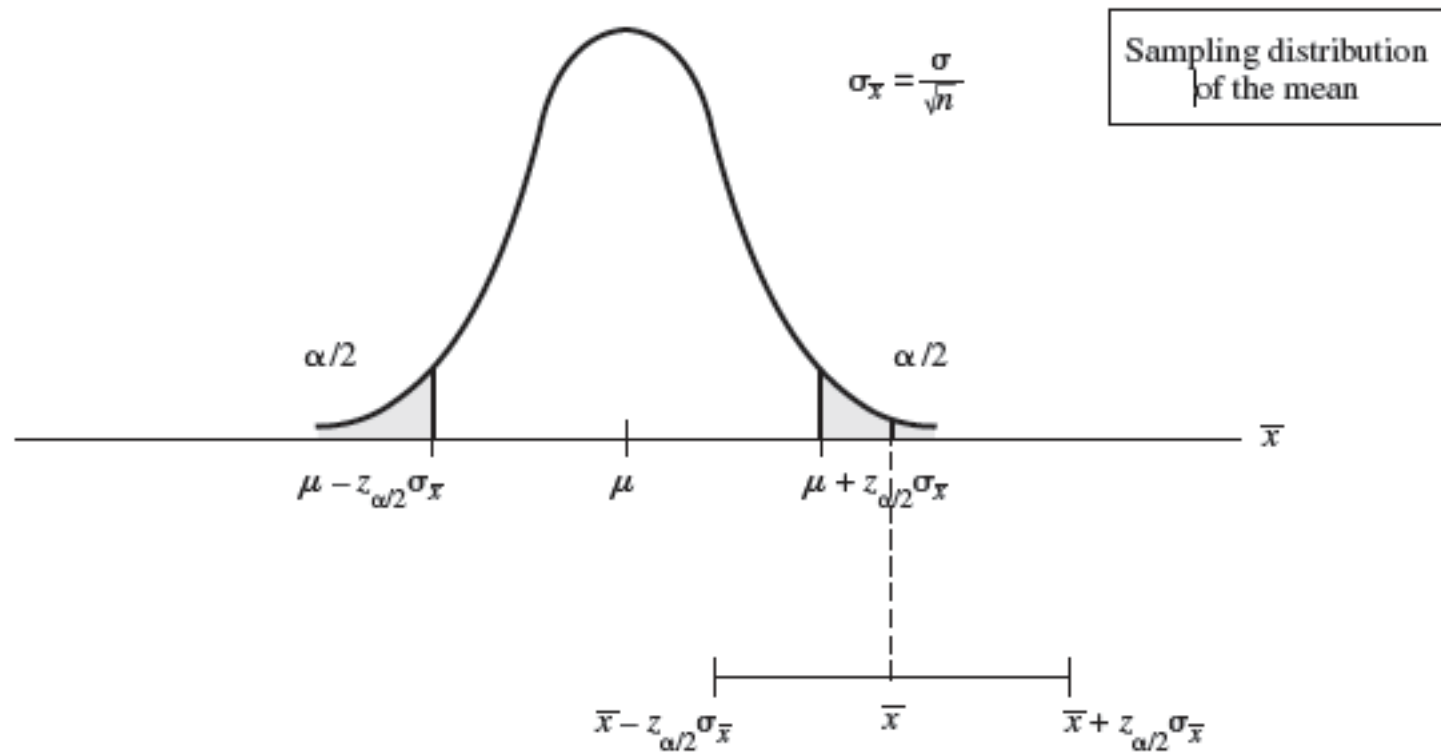
Sampling Distribution of the Mean – Theory



Interval Estimate Containing the True Population Mean



Interval Estimate Not Containing the True Population Mean





CI for Difference in Means

	Population 1	Population 2
Mean	μ_1	μ_2
Standard deviation	σ_1	σ_2
Point estimate	\bar{x}_1	\bar{x}_2
Sample size	n_1	n_2



Independent Samples with Unequal Variances

$$\bar{x}_1 - \bar{x}_2 \pm (t_{\alpha/2, df^*}) \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad (4A.6)$$

where the degrees of freedom for the t -distribution, df^* , are computed as:

$$df^* = \frac{\left[\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right]^2}{\left[\frac{(s_1^2 / n_1)^2}{n_1 - 1} \right] + \left[\frac{(s_2^2 / n_2)^2}{n_2 - 1} \right]} \quad (4A.7)$$



Independent Samples with Equal Variances

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \quad (4A.8)$$

Then the sampling distribution of $\bar{x}_1 - \bar{x}_2$ has a t -distribution with $n_1 + n_2 - 2$ degrees of freedom and standard error:

$$s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad (4A.9)$$

Therefore, a $100(1 - \alpha)\%$ confidence interval is:

$$\bar{x}_1 - \bar{x}_2 \pm \left(t_{\alpha/2, n_1 + n_2 - 2} \right) s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad (4A.10)$$



Paired Samples

For paired samples, we first compute the difference between each pair of observations, D_i , for $i = 1, \dots, n$. If we average these differences, we obtain \bar{D} , a point estimate for the mean difference between the populations. The standard deviation of the differences is similar to calculating an ordinary standard deviation:

$$s_D = \sqrt{\frac{\sum_{i=1}^n (D_i - \bar{D})^2}{n - 1}} \quad (4A.11)$$

A $100(1 - \alpha)\%$ confidence interval is:

$$\bar{D} \pm \left(t_{n-1, \alpha/2} \right) s_D / \sqrt{n} \quad (4A.12)$$



Differences Between Proportions

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1 (1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2 (1 - \hat{p}_2)}{n_2}} \quad (4A.13)$$



Summary of CI Formulas

TABLE 4A.1 Summary of Confidence Interval Formulas

Type of Confidence Interval	Formula
Mean, standard deviation known	$\bar{x} \pm z_{\alpha/2} (\sigma/\sqrt{n})$
Mean, standard deviation unknown	$\bar{x} \pm t_{\alpha/2, n-1} (s/\sqrt{n})$
Proportion	$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p} (1 - \hat{p})}{n}}$
Population total	$N\bar{x} \pm t_{\alpha/2, n-1} N \frac{s}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$
Difference between means, independent samples, equal variances	$\bar{x}_1 - \bar{x}_2 \pm (t_{\alpha/2, n_1 + n_2 - 2}) s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ $s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$
Difference between means, independent samples, unequal variances	$\bar{x}_1 - \bar{x}_2 \pm (t_{\alpha/2, df^*}) \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

Summary of CI Formulas

TABLE 4A.1 (Continued)

Type of Confidence Interval	Formula
	$df^* = \frac{\left[\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right]^2}{\left[\frac{(s_1^2/n_1)^2}{n_1 - 1} \right] + \left[\frac{(s_2^2/n_2)^2}{n_2 - 1} \right]}$
Difference between means, paired samples	$\bar{D} \pm (t_{n-1, \alpha/2})s_D / \sqrt{n}$
Differences between proportions	$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$
Variance	$\left[\frac{(n-1)s^2}{\chi_{n-1, \alpha/2}^2}, \frac{(n-1)s^2}{\chi_{n-1, 1-\alpha/2}^2} \right]$