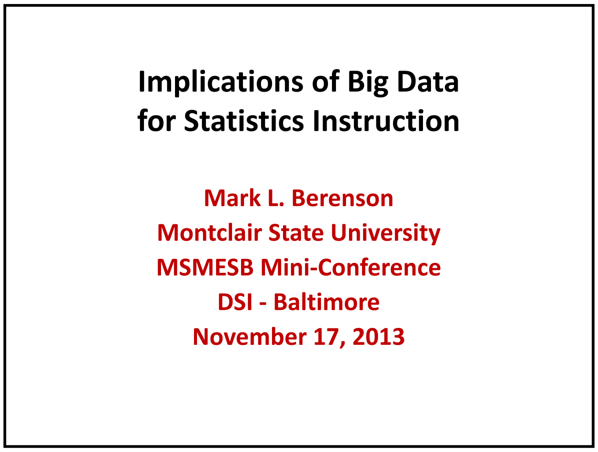


Implications of Big Data for Statistics Instruction

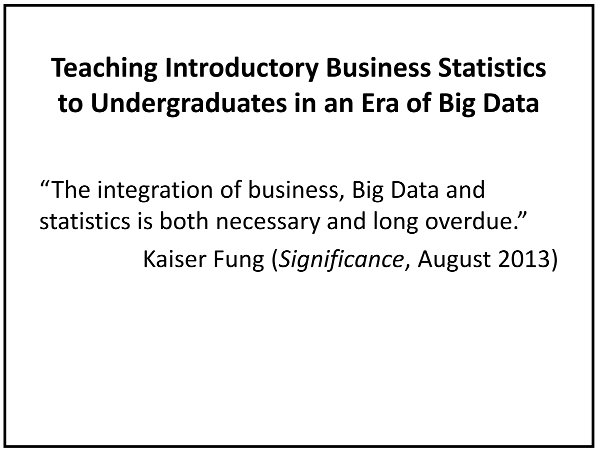
Mark L. Berenson

Montclair State University

MSMESB Mini-Conference

DSI - Baltimore

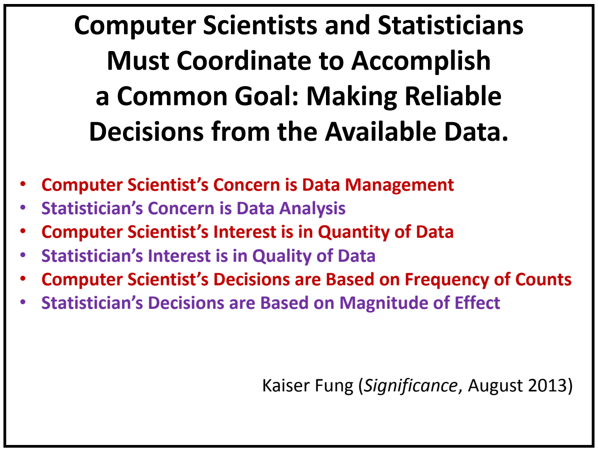
November 17, 2013



Teaching Introductory Business Statistics to Undergraduates in an Era of Big Data

"The integration of business, Big Data and statistics is both necessary and long overdue."

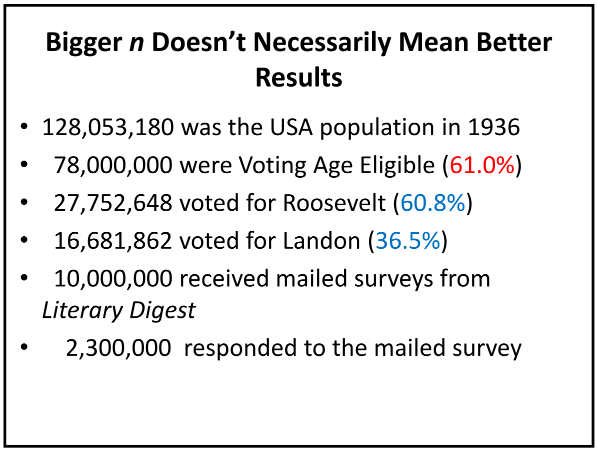
Kaiser Fung (*Significance*, August 2013)



Computer Scientists and Statisticians Must Coordinate to Accomplish a Common Goal: Making Reliable Decisions from the Available Data.

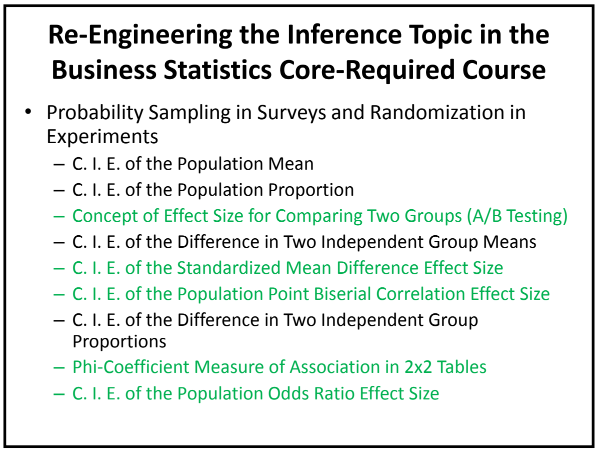
- **Computer Scientist's Concern is Data Management**
- **Statistician's Concern is Data Analysis**
- **Computer Scientist's Interest is in Quantity of Data**
- **Statistician's Interest is in Quality of Data**
- **Computer Scientist's Decisions are Based on Frequency of Counts**
- **Statistician's Decisions are Based on Magnitude of Effect**

Kaiser Fung (*Significance*, August 2013)



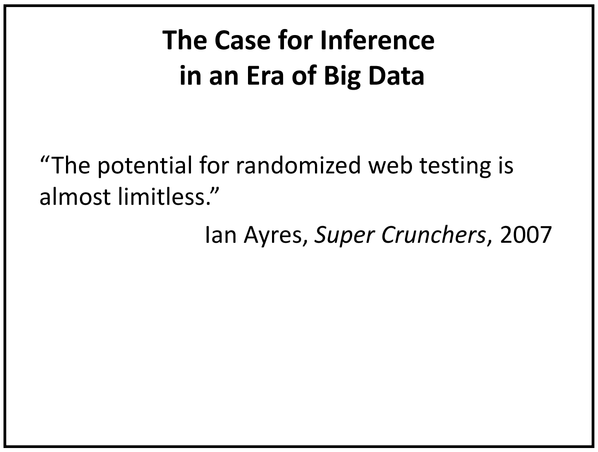
Bigger *n* Doesn't Necessarily Mean Better Results

- 128,053,180 was the USA population in 1936
- 78,000,000 were Voting Age Eligible (61.0%)
- 27,752,648 voted for Roosevelt (60.8%)
- 16,681,862 voted for Landon (36.5%)
- 10,000,000 received mailed surveys from *Literary Digest*
- 2,300,000 responded to the mailed survey



Re-Engineering the Inference Topic in the Business Statistics Core-Required Course

- Probability Sampling in Surveys and Randomization in Experiments
  - C. I. E. of the Population Mean
  - C. I. E. of the Population Proportion
  - Concept of Effect Size for Comparing Two Groups (A/B Testing)
  - C. I. E. of the Difference in Two Independent Group Means
  - C. I. E. of the Standardized Mean Difference Effect Size
  - C. I. E. of the Population Point Biserial Correlation Effect Size
  - C. I. E. of the Difference in Two Independent Group Proportions
  - Phi-Coefficient Measure of Association in 2x2 Tables
  - C. I. E. of the Population Odds Ratio Effect Size



The Case for Inference in an Era of Big Data

"The potential for randomized web testing is almost limitless."

Ian Ayres, *Super Crunchers*, 2007

The Case for Inference in an Era of Big Data

"Testing is a road that never ends. Tastes change. What worked yesterday will not work tomorrow. A system of periodic retesting with randomized trials is a way to ensure that your marketing efforts remain optimized."

Ian Ayres, *Super Crunchers*, 2007

The Case for Inference in an Era of Big Data

"Any large organization that is not exploiting both regression and randomization is presumptively missing value. Especially in mature industries, where profit margins narrow, firms ' competing on analytics' will increasingly be driven to use both tools to stay ahead. … Randomization and regression are the twin pillars of Super Crunching."

Ian Ayres, *Super Crunchers*, 2007

The Problem with Hypothesis Testing in an Era of Big Data

- H. Jeffreys (1939) and D.V. Lindley (1957) point out that any observed trivial difference will become statistically significant if the sample sizes are large enough.

The Case for Effect Size Measures to Replace Hypothesis Testing (NHST) in an Era of Big Data

The use of NHST "has caused scientific research workers to pay undue attention to the results of the tests of significance that they perform on their data and too little attention on the magnitude of the effects they are investigating."

Frank Yates (*JASA*, 1951)

The Case for Effect Size Measures to Replace Hypothesis Testing (NHST) in an Era of Big Data

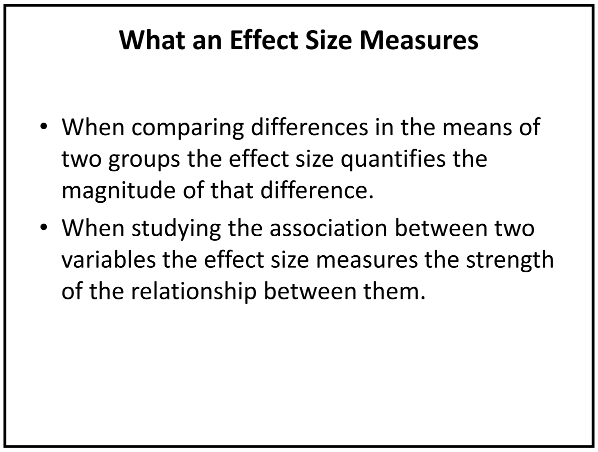
"In many experiments, it seems obvious that the different treatments must produce some difference, however small, in effect. Thus the hypothesis that there is *no* difference is unrealistic. The real problem is to obtain estimates of the size of the differences."

George W. Cochran and Gertrude M. Cox,
Experimental Design, 2nd Ed., (1957)

The Case for Effect Size Measures to Replace Hypothesis Testing (NHST) in an Era of Big Data

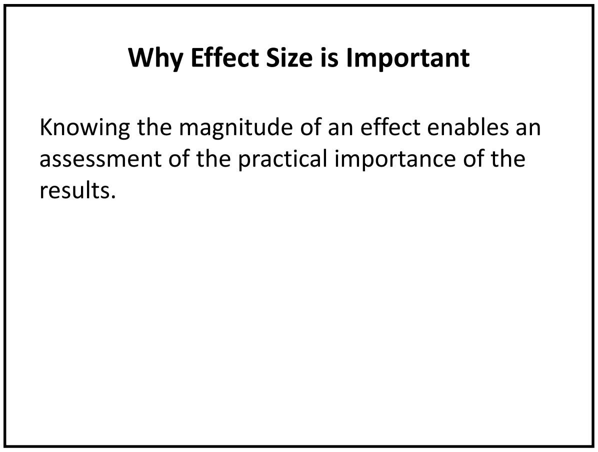
"Estimates of appropriate effect sizes and [their] confidence intervals are the minimum expectations for all APA journals."

Publication Manual of the APA (2010)



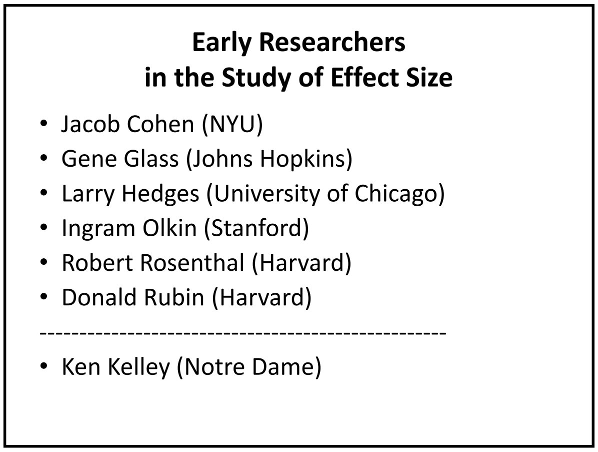
What an Effect Size Measures

- When comparing differences in the means of two groups the effect size quantifies the magnitude of that difference.
- When studying the association between two variables the effect size measures the strength of the relationship between them.



Why Effect Size is Important

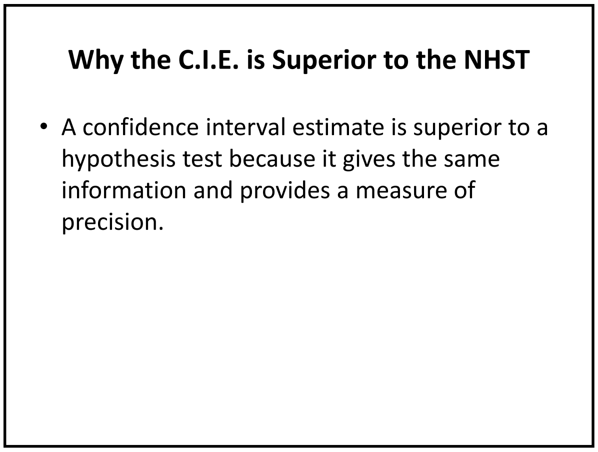
Knowing the magnitude of an effect enables an assessment of the practical importance of the results.



Early Researchers in the Study of Effect Size

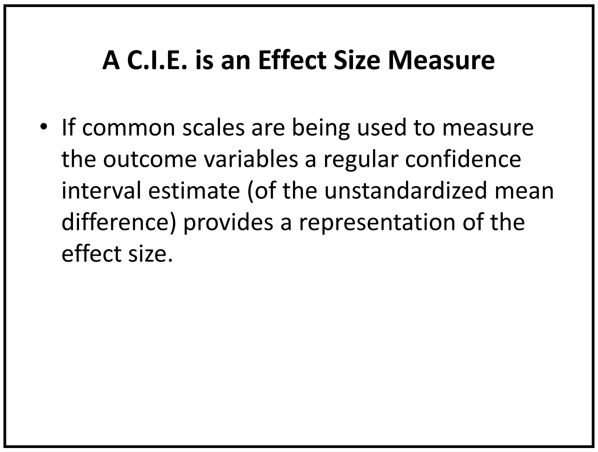
- Jacob Cohen (NYU)
- Gene Glass (Johns Hopkins)
- Larry Hedges (University of Chicago)
- Ingram Olkin (Stanford)
- Robert Rosenthal (Harvard)
- Donald Rubin (Harvard)

- Ken Kelley (Notre Dame)



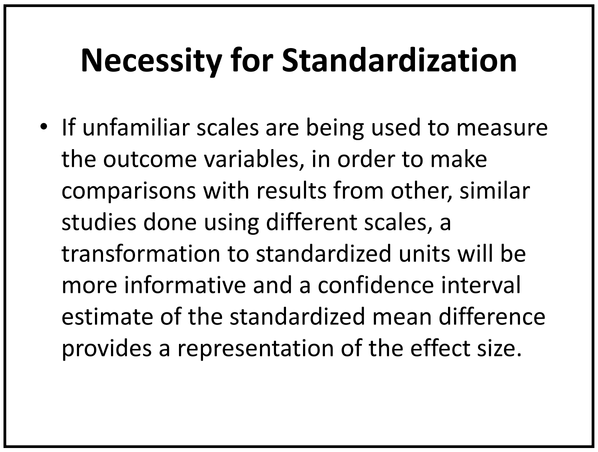
Why the C.I.E. is Superior to the NHST

- A confidence interval estimate is superior to a hypothesis test because it gives the same information and provides a measure of precision.



A C.I.E. is an Effect Size Measure

- If common scales are being used to measure the outcome variables a regular confidence interval estimate (of the unstandardized mean difference) provides a representation of the effect size.



Necessity for Standardization

- If unfamiliar scales are being used to measure the outcome variables, in order to make comparisons with results from other, similar studies done using different scales, a transformation to standardized units will be more informative and a confidence interval estimate of the standardized mean difference provides a representation of the effect size.

What Standardization Achieves

- A standardized effect size removes the sample size of the outcome variable from the effect estimate, producing a dimensionless standardized effect that can be compared across different but related outcome variables in other studies.

Cohen's Effect Size Classifications

- Cohen (1992) developed effect size cut points of .2, .5 and .8, respectively, for *small*, *medium* and *large* effects for standardized mean differences.
- Cohen classified effect size cut points of .1, .3 and .5, respectively, for *small*, *medium* and *large* effects for correlations.
- Cohen described a *medium* effect size as one in which the researcher can visually see the gains from treatment E above and beyond that of treatment C.

C. I. E. for the Difference in Means of Two Independent Groups
*Assuming Unequal Variances
*Assuming Equal Sample Sizes

As the sample sizes increase $t_{\alpha/2}$ approaches $Z_{\alpha/2}$ so that for very large n_E and n_C an approximate (1 - α) 100% confidence interval estimate of the difference in the population means $(\mu_E - \mu_C)$ is given by

$$(\bar{Y}_E - \bar{Y}_C) \pm Z_{\alpha/2}\left[\frac{S_E^2 + S_C^2}{n_*}\right]^{1/2}$$

where the equal sample sizes n_E and n_C are given by n_*.

Bonett's Effect Size C. I. E.
*Assuming Unequal Variances
*Assuming Equal Sample Sizes

Bonett's (2008) approximate (1 - α) 100% confidence interval estimate of the population standardized mean difference effect size δ is given by

$$\hat{\delta} \pm Z_{\alpha/2}\left[Var(\hat{\delta})\right]^{1/2}$$

where the estimate of δ is

$$\hat{\delta} = \frac{\bar{Y}_E - \bar{Y}_C}{\hat{\sigma}}$$

and $\hat{\sigma} = \left[\frac{(S_E^2 + S_C^2)}{2}\right]^{1/2}$ when $n_E = n_C$. The variance of the statistic $\hat{\delta}$ proposed by Bonett

reduces to

$$Var(\hat{\delta}) = \frac{\hat{\delta}^2(S_E^4 + S_C^4)}{8\hat{\sigma}^4(n_* - 1)} + \frac{2}{(n_* - 1)}$$

when the equal sample sizes n_E and n_C are represented by n_*.

Rosenthal's Effect Size Confidence Interval for the Point Biserial Correlation Coefficient

When the sample sizes are equal the point biserial correlation coefficient r_{pb} is obtained from

$$r_{pb} = \frac{\bar{Y}_E - \bar{Y}_C}{2S_Y}$$

where $\bar{\bar{Y}} = \frac{\bar{Y}_E + \bar{Y}_C}{2}$ and $S_Y = \left[\frac{\sum_{i=1}^{n_i}\sum_{j=1}^{2}(Y_{ij} - \bar{\bar{Y}})^2}{n_E + n_C}\right]^{1/2}$.

Also, for very large sample size, r_{pb} and $\hat{\delta}$ are related as follows:

$$r_{pb} = \frac{\hat{\delta}}{\left[\hat{\delta}^2 + 4\right]^{1/2}} \text{ and } \hat{\delta} = \frac{2r_{pb}}{\left[1 - r_{pb}^2\right]^{1/2}}$$

Rosenthal's Effect Size Confidence Interval for the Point Biserial Correlation Coefficient

An approximate (1 - α) 100% confidence interval estimate of the population point biserial coefficient of correlation ρ_{pb} is obtained using the Fisher Z_r transformation where

$$Z_r = 0.5\ln\left[\frac{1 + r_{pb}}{1 - r_{pb}}\right]$$

and the standard error is

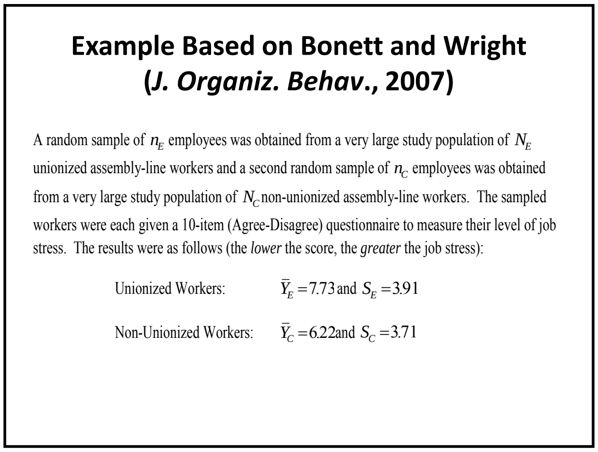
$$S_{Z_r} = \left[\frac{1}{n_E + n_C - 3}\right]^{1/2}$$

The confidence interval limits for this are obtained from $Z_r \pm Z_{\alpha/2}S_{Z_r}$ so that

$$0.5\ln\left[\frac{1 + r_{pb}}{1 - r_{pb}}\right] \pm Z_{\alpha/2}\left[\frac{1}{n_E + n_C - 3}\right]^{1/2}$$

and the approximate (1 - α) 100% confidence interval estimate of the population ρ_{pb} is obtained by taking the antilogs of the above lower and upper limits. The conversions for each limit are given by

$$\rho_{pb} = \frac{e^{2Z_r} - 1}{e^{2Z_r} + 1}$$

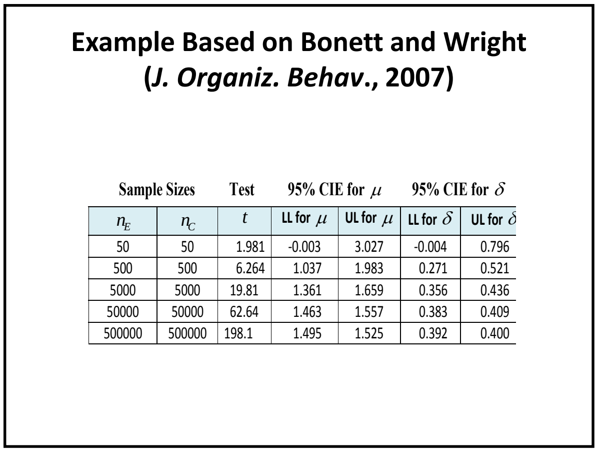


Example Based on Bonett and Wright (*J. Organiz. Behav*., 2007)

A random sample of n_E employees was obtained from a very large study population of N_E unionized assembly-line workers and a second random sample of n_C employees was obtained from a very large study population of N_C non-unionized assembly-line workers. The sampled workers were each given a 10-item (Agree-Disagree) questionnaire to measure their level of job stress. The results were as follows (the *lower* the score, the *greater* the job stress):

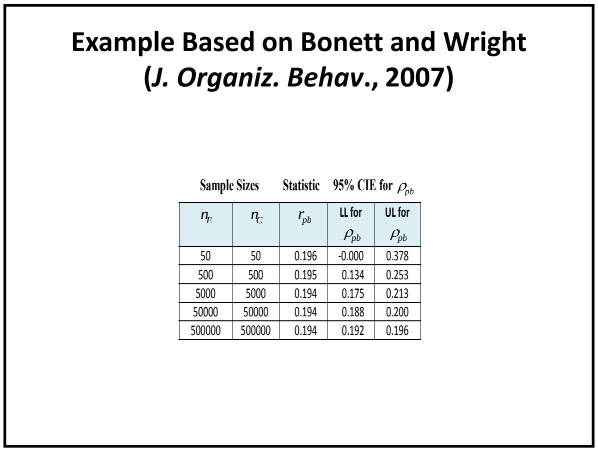
Unionized Workers: $\bar{Y}_E = 7.73$ and $S_E = 3.91$

Non-Unionized Workers: $\bar{Y}_C = 6.22$ and $S_C = 3.71$



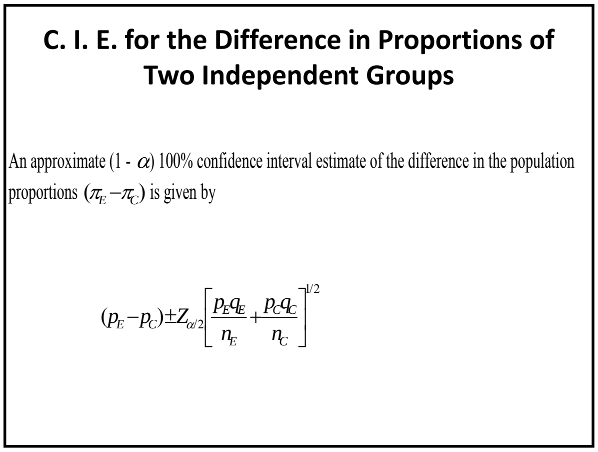
Example Based on Bonett and Wright (*J. Organiz. Behav*., 2007)

Sample Sizes		Test	95% CIE for μ		95% CIE for δ	
n_E	n_C	t	LL for μ	UL for μ	LL for δ	UL for δ
50	50	1.981	-0.003	3.027	-0.004	0.796
500	500	6.264	1.037	1.983	0.271	0.521
5000	5000	19.81	1.361	1.659	0.356	0.436
50000	50000	62.64	1.463	1.557	0.383	0.409
500000	500000	198.1	1.495	1.525	0.392	0.400



Example Based on Bonett and Wright (*J. Organiz. Behav*., 2007)

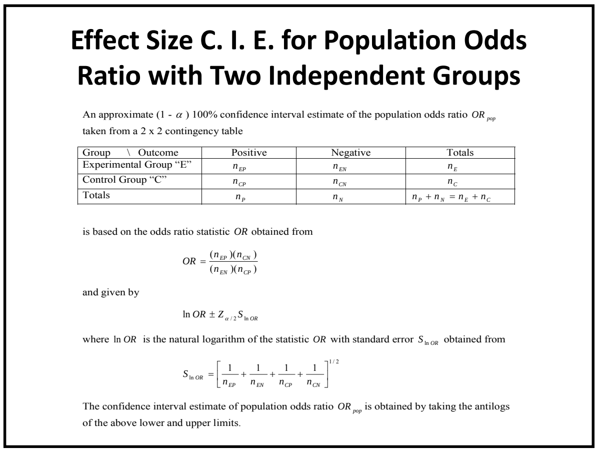
Sample Sizes		Statistic	95% CIE for ρ_{pb}	
n_E	n_C	r_{pb}	LL for ρ_{pb}	UL for ρ_{pb}
50	50	0.196	-0.000	0.378
500	500	0.195	0.134	0.253
5000	5000	0.194	0.175	0.213
50000	50000	0.194	0.188	0.200
500000	500000	0.194	0.192	0.196



C. I. E. for the Difference in Proportions of Two Independent Groups

An approximate (1 - α) 100% confidence interval estimate of the difference in the population proportions $(\pi_E - \pi_C)$ is given by

$$(p_E - p_C) \pm Z_{\alpha/2}\left[\frac{p_E q_E}{n_E} + \frac{p_C q_C}{n_C}\right]^{1/2}$$



Effect Size C. I. E. for Population Odds Ratio with Two Independent Groups

An approximate (1 - α) 100% confidence interval estimate of the population odds ratio OR_{pop} taken from a 2 x 2 contingency table

Group \ Outcome	Positive	Negative	Totals
Experimental Group "E"	n_{EP}	n_{EN}	n_E
Control Group "C"	n_{CP}	n_{CN}	n_C
Totals	n_P	n_N	$n_P + n_N = n_E + n_C$

is based on the odds ratio statistic OR obtained from

$$OR = \frac{(n_{EP})(n_{CN})}{(n_{EN})(n_{CP})}$$

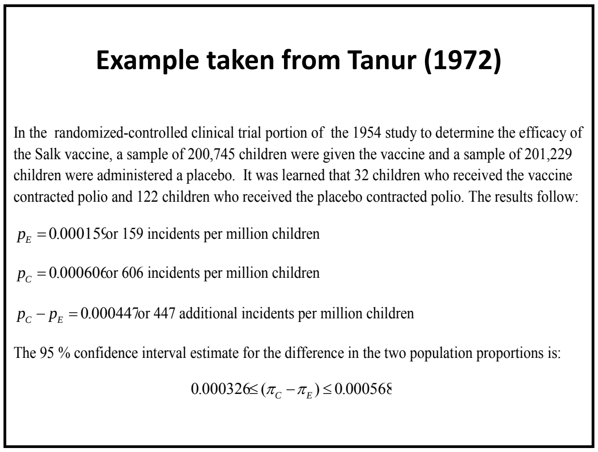
and given by

$$\ln OR \pm Z_{\alpha/2} S_{\ln OR}$$

where $\ln OR$ is the natural logarithm of the statistic OR with standard error $S_{\ln OR}$ obtained from

$$S_{\ln OR} = \left[\frac{1}{n_{EP}} + \frac{1}{n_{EN}} + \frac{1}{n_{CP}} + \frac{1}{n_{CN}}\right]^{1/2}$$

The confidence interval estimate of population odds ratio OR_{pop} is obtained by taking the antilogs of the above lower and upper limits.



Example taken from Tanur (1972)

In the randomized-controlled clinical trial portion of the 1954 study to determine the efficacy of the Salk vaccine, a sample of 200,745 children were given the vaccine and a sample of 201,229 children were administered a placebo. It was learned that 32 children who received the vaccine contracted polio and 122 children who received the placebo contracted polio. The results follow:

$p_E = 0.000159$ or 159 incidents per million children

$p_C = 0.000606$ or 606 incidents per million children

$p_C - p_E = 0.000447$ or 447 additional incidents per million children

The 95 % confidence interval estimate for the difference in the two population proportions is:

$$0.000326 \le (\pi_C - \pi_E) \le 0.000568$$

Example taken from Tanur (1972)

The χ_1^2 test statistic is 52.401

The effect size is the ϕ coefficient of correlation, $\left[\frac{\chi_1^2}{(n_E + n_C)}\right]^{1/2} = [0.0001304]^{1/2} = 0.0114$

This application shows that even an extremely small effect size can be practically important. There were 54 million children (persons 17 or under) in the USA in 1954 out of an overall population of 162 million. Of the 39000 persons who contracted polio that year, two-thirds, or 26000, were children. Therefore, in 1954 the polio incidence rate for children was 0.048% (or 481 children per million).

The odds ratio statistic OR for this study is 3.81. That is, although the incidence rate is small, the odds are 3.81 times more likely that a child given a placebo will contract polio than a child given the vaccine. Using the odds ratio as an effect size, a 95% confidence interval estimate for the population odds ratio is:

$$2.58 \le OR_{pop} \le 5.62$$

Evaluating Practical Importance via BESD Rosenthal & Rubin (1982)

Binomial Effect Size Display (BESD)

- Developed by converting various effect size measures into "correlation" effect sizes
- Obtains E group "success rate" as .5 + *r*/2
- Obtains C group "success rate" as .5 – *r*/2

BESD
Change in Success Rates for Various Values of *r*

Effect Size = Difference In Success Rate	Equivalent to Success Rate Increase	
r	From	To
0.02	0.49	0.51
0.04	0.48	0.52
0.06	0.47	0.53
0.08	0.46	0.54
0.10	0.45	0.55
0.12	0.44	0.56
0.16	0.42	0.58
0.20	0.40	0.60
0.24	0.38	0.62
0.30	0.35	0.65
0.40	0.30	0.70
0.50	0.25	0.75
0.60	0.20	0.80
0.70	0.15	0.85
0.80	0.10	0.90
0.90	0.05	0.95
1.00	0.00	1.00

BESD for Assembly-Line Stress Example

190 more workers per 1000 had lower stress if unionized.

Condition \ Result	Higher Stress	Lower Stress	Total
Unionized Workers	405	595	1000
Non-Unionized	595	405	1000
Total	1000	1000	2000

BESD for Salk Vaccine Study

114 more children per 10000 were helped by the vaccine.

Condition \ Result	Stay Healthy	Contract Polio	Total
Vaccine	5057	4943	10000
Placebo	4943	5057	10000
Total	10000	10000	20000

Summary and Conclusions

- A course in Business Statistics needs to be modified to maintain its relevance in an era of Big Data.
- Business statistics textbooks must adapt its topic coverage to introduce methodology relevant to a Big Data environment – the subject of inference must be re-engineered.
- The time has come for AACSB-accredited undergraduate programs to include a core-required course in Business Analytics as a sequel to a course in Business Statistics.

Implications of Big Data for Statistics Instruction

Mark L. Berenson

Montclair State University

MSMESB Mini-Conference

DSI - Baltimore

November 17, 2013

Teaching Introductory Business Statistics to Undergraduates in an Era of Big Data

"The integration of business, Big Data and statistics is both necessary and long overdue."

Kaiser Fung (*Significance*, August 2013)

Computer Scientists and Statisticians Must Coordinate to Accomplish a Common Goal: Making Reliable Decisions from the Available Data.

- Computer Scientist’s Concern is Data Management
- Statistician’s Concern is Data Analysis
- Computer Scientist’s Interest is in Quantity of Data
- Statistician’s Interest is in Quality of Data
- Computer Scientist’s Decisions are Based on Frequency of Counts
- Statistician’s Decisions are Based on Magnitude of Effect

Kaiser Fung (*Significance*, August 2013)

Bigger *n* Doesn't Necessarily Mean Better Results

- 128,053,180 was the USA population in 1936
- 78,000,000 were Voting Age Eligible (61.0%)
- 27,752,648 voted for Roosevelt (60.8%)
- 16,681,862 voted for Landon (36.5%)
- 10,000,000 received mailed surveys from *Literary Digest*
- 2,300,000 responded to the mailed survey

Re-Engineering the Inference Topic in the Business Statistics Core-Required Course

- Probability Sampling in Surveys and Randomization in Experiments
  - C. I. E. of the Population Mean
  - C. I. E. of the Population Proportion
  - Concept of Effect Size for Comparing Two Groups (A/B Testing)
  - C. I. E. of the Difference in Two Independent Group Means
  - C. I. E. of the Standardized Mean Difference Effect Size
  - C. I. E. of the Population Point Biserial Correlation Effect Size
  - C. I. E. of the Difference in Two Independent Group Proportions
  - Phi-Coefficient Measure of Association in 2x2 Tables
  - C. I. E. of the Population Odds Ratio Effect Size

The Case for Inference in an Era of Big Data

"The potential for randomized web testing is almost limitless."

Ian Ayres, *Super Crunchers*, 2007

The Case for Inference in an Era of Big Data

"Testing is a road that never ends. Tastes change. What worked yesterday will not work tomorrow. A system of periodic retesting with randomized trials is a way to ensure that your marketing efforts remain optimized."

Ian Ayres, *Super Crunchers*, 2007

The Case for Inference in an Era of Big Data

"Any large organization that is not exploiting both regression and randomization is presumptively missing value. Especially in mature industries, where profit margins narrow, firms ' competing on analytics' will increasingly be driven to use both tools to stay ahead. … Randomization and regression are the twin pillars of Super Crunching."

Ian Ayres, *Super Crunchers*, 2007

The Problem with Hypothesis Testing in an Era of Big Data

- H. Jeffreys (1939) and D.V. Lindley (1957) point out that any observed trivial difference will become statistically significant if the sample sizes are large enough.

The Case for Effect Size Measures to Replace Hypothesis Testing (NHST) in an Era of Big Data

The use of NHST “has caused scientific research workers to pay undue attention to the results of the tests of significance that they perform on their data and too little attention on the magnitude of the effects they are investigating.“

Frank Yates (*JASA*, 1951)

The Case for Effect Size Measures to Replace Hypothesis Testing (NHST) in an Era of Big Data

"In many experiments, it seems obvious that the different treatments must produce some difference, however small, in effect. Thus the hypothesis that there is *no* difference is unrealistic. The real problem is to obtain estimates of the size of the differences."

George W. Cochran and Gertrude M. Cox,
Experimental Design, 2nd Ed., (1957)

The Case for Effect Size Measures to Replace Hypothesis Testing (NHST) in an Era of Big Data

“Estimates of appropriate effect sizes and [their] confidence intervals are the minimum expectations for all APA journals.”

Publication Manual of the APA (2010)

What an Effect Size Measures

- When comparing differences in the means of two groups the effect size quantifies the magnitude of that difference.
- When studying the association between two variables the effect size measures the strength of the relationship between them.

Why Effect Size is Important

Knowing the magnitude of an effect enables an assessment of the practical importance of the results.

Early Researchers in the Study of Effect Size

- Jacob Cohen (NYU)
- Gene Glass (Johns Hopkins)
- Larry Hedges (University of Chicago)
- Ingram Olkin (Stanford)
- Robert Rosenthal (Harvard)
- Donald Rubin (Harvard)

- Ken Kelley (Notre Dame)

Why the C.I.E. is Superior to the NHST

- A confidence interval estimate is superior to a hypothesis test because it gives the same information and provides a measure of precision.

A C.I.E. is an Effect Size Measure

- If common scales are being used to measure the outcome variables a regular confidence interval estimate (of the unstandardized mean difference) provides a representation of the effect size.

Necessity for Standardization

- If unfamiliar scales are being used to measure the outcome variables, in order to make comparisons with results from other, similar studies done using different scales, a transformation to standardized units will be more informative and a confidence interval estimate of the standardized mean difference provides a representation of the effect size.

What Standardization Achieves

- A standardized effect size removes the sample size of the outcome variable from the effect estimate, producing a dimensionless standardized effect that can be compared across different but related outcome variables in other studies.

Cohen’s Effect Size Classifications

- Cohen (1992) developed effect size cut points of .2, .5 and .8, respectively, for *small*, *medium* and *large* effects for standardized mean differences.
- Cohen classified effect size cut points of .1, .3 and .5, respectively, for *small*, *medium* and *large* effects for correlations.
- Cohen described a *medium* effect size as one in which the researcher can visually see the gains from treatment E above and beyond that of treatment C.

C. I. E. for the Difference in Means of Two Independent Groups *Assuming Unequal Variances *Assuming Equal Sample Sizes

As the sample sizes increase $t_{\alpha/2}$ approaches $Z_{\alpha/2}$ so that for very large n_E and n_C an approximate (1 - α) 100% confidence interval estimate of the difference in the population means $(\mu_E - \mu_C)$ is given by

$$(\bar{Y}_E - \bar{Y}_C) \pm Z_{\alpha/2}\left[\frac{S_E^2 + S_C^2}{n_*}\right]^{1/2}$$

where the equal sample sizes n_E and n_C are given by n_*.

Bonett's Effect Size C. I. E.

*Assuming Unequal Variances

*Assuming Equal Sample Sizes

Bonett's (2008) approximate (1 - α) 100% confidence interval estimate of the population standardized mean difference effect size δ is given by

$$\hat{\delta} \pm Z_{\alpha/2}\left[Var(\hat{\delta})\right]^{1/2}$$

where the estimate of δ is

$$\hat{\delta} = \frac{\overline{Y}_E - \overline{Y}_C}{\hat{\sigma}}$$

and $\hat{\sigma} = \left[\frac{(S_E^2 + S_C^2)}{2}\right]^{1/2}$ when $n_E = n_C$. The variance of the statistic $\hat{\delta}$ proposed by Bonett

reduces to

$$Var(\hat{\delta}) = \frac{\hat{\delta}^2 (S_E^4 + S_C^4)}{8\hat{\sigma}^4 (n_* - 1)} + \frac{2}{(n_* - 1)}$$

when the equal sample sizes n_E and n_C are represented by n_* .

Rosenthal’s Effect Size Confidence Interval for the Point Biserial Correlation Coefficient

When the sample sizes are equal the point biserial correlation coefficient r_{pb} is obtained from

$$r_{pb} = \frac{\overline{Y}_E - \overline{Y}_C}{2S_Y}$$

where $\overline{\overline{Y}} = \frac{\overline{Y}_E + \overline{Y}_C}{2}$ and $S_Y = \left[\frac{\sum_{i=1}^{n_i} \sum_{j=1}^{2} (Y_{ij} - \overline{\overline{Y}})^2}{n_E + n_C} \right]^{1/2}$.

Also, for very large sample size, r_{pb} and $\hat{\delta}$ are related as follows:

$$r_{pb} = \frac{\hat{\delta}}{\left[\hat{\delta}^2 + 4\right]^{1/2}} \text{ and } \hat{\delta} = \frac{2r_{pb}}{\left[1 - r_{pb}^2\right]^{1/2}}$$

Rosenthal's Effect Size Confidence Interval for the Point Biserial Correlation Coefficient

An approximate (1 - α) 100% confidence interval estimate of the population point biserial coefficient of correlation ρ_{pb} is obtained using the Fisher Z_r transformation where

$$Z_r = 0.5\ln\left[\frac{1+r_{pb}}{1-r_{pb}}\right]$$

and the standard error is

$$S_{Z_r} = \left[\frac{1}{n_E + n_C - 3}\right]^{1/2}$$

The confidence interval limits for this are obtained from $Z_r \pm Z_{\alpha/2} S_{Z_r}$ so that

$$0.5\ln\left[\frac{1+r_{pb}}{1-r_{pb}}\right] \pm Z_{\alpha/2}\left[\frac{1}{n_E + n_C - 3}\right]^{1/2}$$

and the approximate (1 - α) 100% confidence interval estimate of the population ρ_{pb} is obtained by taking the antilogs of the above lower and upper limits. The conversions for each limit are given by

$$\rho_{pb} = \frac{e^{2Z_r} - 1}{e^{2Z_r} + 1}$$

Example Based on Bonett and Wright (*J. Organiz. Behav.*, 2007)

A random sample of n_E employees was obtained from a very large study population of N_E unionized assembly-line workers and a second random sample of n_C employees was obtained from a very large study population of N_C non-unionized assembly-line workers. The sampled workers were each given a 10-item (Agree-Disagree) questionnaire to measure their level of job stress. The results were as follows (the *lower* the score, the *greater* the job stress):

Unionized Workers:	$\bar{Y}_E = 7.73$ and $S_E = 3.91$
Non-Unionized Workers:	$\bar{Y}_C = 6.22$ and $S_C = 3.71$

Example Based on Bonett and Wright (*J. Organiz. Behav.*, 2007)

Sample Sizes		Test	95% CIE for μ		95% CIE for δ	
n_E	n_C	t	LL for μ	UL for μ	LL for δ	UL for δ
50	50	1.981	-0.003	3.027	-0.004	0.796
500	500	6.264	1.037	1.983	0.271	0.521
5000	5000	19.81	1.361	1.659	0.356	0.436
50000	50000	62.64	1.463	1.557	0.383	0.409
500000	500000	198.1	1.495	1.525	0.392	0.400

Example Based on Bonett and Wright (*J. Organiz. Behav.*, 2007)

Sample Sizes		Statistic	95% CIE for ρ_{pb}	
n_E	n_C	r_{pb}	LL for ρ_{pb}	UL for ρ_{pb}
50	50	0.196	-0.000	0.378
500	500	0.195	0.134	0.253
5000	5000	0.194	0.175	0.213
50000	50000	0.194	0.188	0.200
500000	500000	0.194	0.192	0.196

C. I. E. for the Difference in Proportions of Two Independent Groups

An approximate (1 - α) 100% confidence interval estimate of the difference in the population proportions $(\pi_E - \pi_C)$ is given by

$$(p_E - p_C) \pm Z_{\alpha/2}\left[\frac{p_E q_E}{n_E} + \frac{p_C q_C}{n_C}\right]^{1/2}$$

Effect Size C. I. E. for Population Odds Ratio with Two Independent Groups

An approximate (1 - α) 100% confidence interval estimate of the population odds ratio OR_{pop} taken from a 2 x 2 contingency table

Group \ Outcome	Positive	Negative	Totals
Experimental Group “E”	n_{EP}	n_{EN}	n_E
Control Group “C”	n_{CP}	n_{CN}	n_C
Totals	n_P	n_N	$n_P + n_N = n_E + n_C$

is based on the odds ratio statistic OR obtained from

$$OR = \frac{(n_{EP})(n_{CN})}{(n_{EN})(n_{CP})}$$

and given by

$$\ln OR \pm Z_{\alpha/2} S_{\ln OR}$$

where $\ln OR$ is the natural logarithm of the statistic OR with standard error $S_{\ln OR}$ obtained from

$$S_{\ln OR} = \left[\frac{1}{n_{EP}} + \frac{1}{n_{EN}} + \frac{1}{n_{CP}} + \frac{1}{n_{CN}} \right]^{1/2}$$

The confidence interval estimate of population odds ratio OR_{pop} is obtained by taking the antilogs of the above lower and upper limits.

Example taken from Tanur (1972)

In the randomized-controlled clinical trial portion of the 1954 study to determine the efficacy of the Salk vaccine, a sample of 200,745 children were given the vaccine and a sample of 201,229 children were administered a placebo. It was learned that 32 children who received the vaccine contracted polio and 122 children who received the placebo contracted polio. The results follow:

$p_E = 0.000159$ or 159 incidents per million children

$p_C = 0.000606$ or 606 incidents per million children

$p_C - p_E = 0.000447$ or 447 additional incidents per million children

The 95 % confidence interval estimate for the difference in the two population proportions is:

$$0.000326 \le (\pi_C - \pi_E) \le 0.000568$$

Example taken from Tanur (1972)

The χ_1^2 test statistic is 52.401

The effect size is the ϕ coefficient of correlation, $\left[\dfrac{\chi_1^2}{(n_E + n_C)}\right]^{1/2} = [0.0001304]^{1/2} = 0.0114$

This application shows that even an extremely small effect size can be practically important. There were 54 million children (persons 17 or under) in the USA in 1954 out of an overall population of 162 million. Of the 39000 persons who contracted polio that year, two-thirds, or 26000, were children. Therefore, in 1954 the polio incidence rate for children was 0.048% (or 481 children per million).

The odds ratio statistic OR for this study is 3.81. That is, although the incidence rate is small, the odds are 3.81 times more likely that a child given a placebo will contract polio than a child given the vaccine. Using the odds ratio as an effect size, a 95% confidence interval estimate for the population odds ratio is:

$$2.58 \le OR_{pop} \le 5.62$$

Evaluating Practical Importance via BESD
Rosenthal & Rubin (1982)

Binomial Effect Size Display (BESD)

- Developed by converting various effect size measures into “correlation” effect sizes
- Obtains E group “success rate” as .5 + *r*/2
- Obtains C group “success rate” as .5 – *r*/2

BESD

Change in Success Rates for Various Values of *r*

Effect Size = Difference In Success Rate	Equivalent to Success Rate Increase	
r	From	To
0.02	0.49	0.51
0.04	0.48	0.52
0.06	0.47	0.53
0.08	0.46	0.54
0.10	0.45	0.55
0.12	0.44	0.56
0.16	0.42	0.58
0.20	0.40	0.60
0.24	0.38	0.62
0.30	0.35	0.65
0.40	0.30	0.70
0.50	0.25	0.75
0.60	0.20	0.80
0.70	0.15	0.85
0.80	0.10	0.90
0.90	0.05	0.95
1.00	0.00	1.00

BESD for Assembly-Line Stress Example

190 more workers per 1000 had lower stress if unionized.

Condition \ Result	Higher Stress	Lower Stress	Total
Unionized Workers	405	595	1000
Non-Unionized	595	405	1000
Total	1000	1000	2000

BESD for Salk Vaccine Study

114 more children per 10000 were helped by the vaccine.

Condition \ Result	Stay Healthy	Contract Polio	Total
Vaccine	5057	4943	10000
Placebo	4943	5057	10000
Total	10000	10000	20000

Summary and Conclusions

- A course in Business Statistics needs to be modified to maintain its relevance in an era of Big Data.
- Business statistics textbooks must adapt its topic coverage to introduce methodology relevant to a Big Data environment – the subject of inference must be re-engineered.
- The time has come for AACSB-accredited undergraduate programs to include a core-required course in Business Analytics as a sequel to a course in Business Statistics.