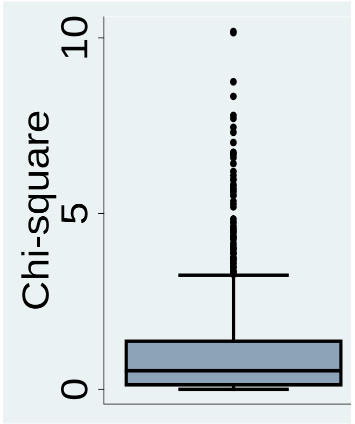
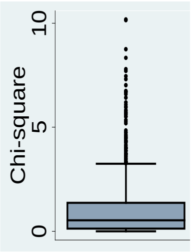
Statistics for Life: Pearls vs Sand

Charles E. McCulloch
Head
Division of Biostatistics, UCSF



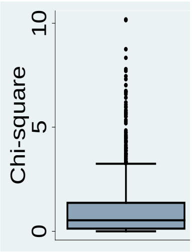
DCR, 2017

Statistics – pearls vs. sand



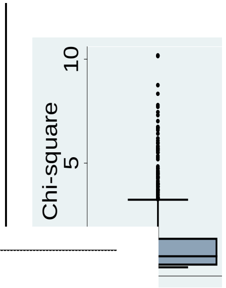
- Roles of statistics
- Example: poor sleep and dementia
- Descriptive statistics
- P-values
- Choice of statistical analysis
- Standard errors
- Confidence intervals
- Stata illustration
- Summary

Roles of statistics in clinical research



- Designing studies
  - Sample size choice
  - Comparing study designs
- Descriptive statistics
  - Describe the data in the sample
- Inferential statistics
  - What the results in the sample tell you about the population.

Example: Yaffe et al JAMA 2011



ORIGINAL CONTRIBUTION

Sleep-Disordered Breathing, Hypoxia, and Risk of Mild Cognitive Impairment and Dementia in Older Women

Kristine Yaffe, MD

Context Sleep-disordered breathing (characterized by recurrent arousals from sleep

Is sleep-disordered breathing prospectively related to mild cognitive impairment in elderly women?

About 300 women (average age 82) participated in a sleep sub-study of the Study of Osteoporotic Fractures in 2003. About 5 years later, adjudicated measures of mild cognitive impairment (or worse) were determined.

Sleep/dementia – Table 1

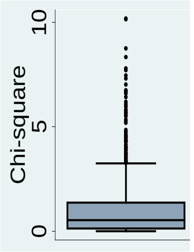
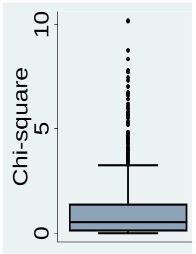


Table 1. Baseline Characteristics by Sleep-Disordered Breathing Status (N = 298)

Characteristic	Sleep-Disordered Breathing	
	No (n = 193)	Yes (n = 105)
Age, mean (SD), y	82.1 (3.2)	82.6 (3.1)
White, No. (%)	173 (90)	96 (91)
Education >high school, No. (%)	60 (31)	30 (29)
Body mass index, mean (SD)[a]	27.9 (4.8)	28.7 (5.3)
Diabetes, No. (%)	19 (10)	17 (16)
Hypertension, No. (%)	116 (60)	69 (66)
History of stroke, No. (%)	25 (13)	11 (11)
High number of depressive symptoms, No. (%)[b]	23 (12)	9 (9)
Current smoking, No. (%)	5 (3)	0
Current medication use, No. (%)		
Antidepressants	16 (8)	8 (8)
Benzodiazepines	14 (7)	9 (9)
Nonbenzodiazepine anxiolytics	5 (3)	2 (2)
Mini-Mental State Examination, mean (SD)	24.9 (1.2)	25.1 (1.1)
Modified version of Trails B, mean (SD)	130.1 (54.8)	122.9 (52.1)

[a] Body mass index was calculated as weight in kilograms divided by height in meters squared.
[b] Score of 6 or greater on Geriatric Depression Scale.
[c] Calculated using the Fisher exact test.

Typical descriptive statistics



- For numeric variables
  - Mean and standard deviation (SD)
    - +/- 2 SDs contain about 95% of the data values
  - If skewed: median and interquartile range (IQR)
    - Median: 50/50 above and below (50^{th} percentile)
    - IQR: 75^{th} percentile – 25^{th} percentile
- For categorical variables
  - Frequencies

Sleep/dementia – Table 1

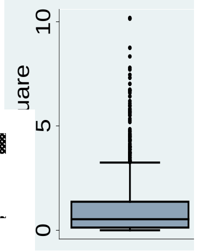
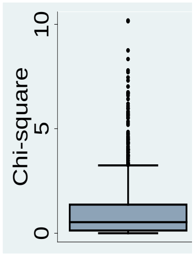


Table 1. Baseline Characteristics by Sleep-Disordered Breathing Status (N = 298)

Characteristic	Sleep-Disordered Breathing: No (n = 193)	Sleep-Disordered Breathing: Yes (n = 105)	*P* Value
Age, mean (SD), y	82.1 (3.2)	82.6 (3.1)	.24
White, No. (%)	173 (90)	96 (91)	.62
Education >high school, No. (%)	60 (31)	30 (29)	.65
Body mass index, mean (SD)[a]	27.9 (4.8)	28.7 (5.3)	.20
Diabetes, No. (%)	19 (10)	17 (16)	.11
Hypertension, No. (%)	116 (60)	69 (66)	.34
History of stroke, No. (%)	25 (13)	11 (11)	.53
High number of depressive symptoms, No. (%)[b]	23 (12)	9 (9)	.37
Current smoking, No. (%)	5 (3)	0	.17[c]
Current medication use, No. (%)			
Antidepressants	16 (8)	8 (8)	.84
Benzodiazepines	14 (7)	9 (9)	.68
Nonbenzodiazepine anxiolytics	5 (3)	2 (2)	.71
Mini-Mental State Examination, mean (SD)	24.9 (1.2)	25.1 (1.1)	.22
Modified version of Trails B, mean (SD)	130.1 (54.8)	122.9 (52.1)	.48

[a] Body mass index was calculated as weight in kilograms divided by height in meters squared.
[b] Score of 6 or greater on Geriatric Depression Scale.
[c] Calculated using the Fisher exact test.

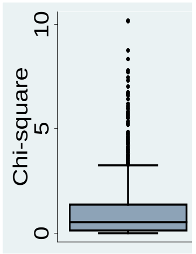
Results Table:
Poor sleep/cognition



	Total sleep time			
Cognitive Impairment	Low (N)	(<4.5hr) (%)	High (N)	(>6.5hr) (%)
Yes	41	42%	37	37%
No	57		62	
Total	98		99	

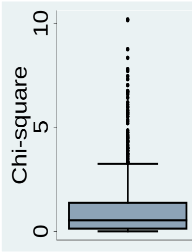
Conclusion?

How do we use this data to infer to a population?

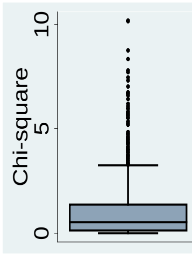


- Population of older US women.
- Want to know if the results seen in this sample of N=298 can be reliably extrapolated to the population.
- Is the data consistent with a null hypothesis of no difference in the rates of cognitive impairment in the population? Or indicative of a relationship in the population?
- *Inferential* statistics.

Chance of a result as or more extreme as observed



- Assume null hypothesis true.
- In this instance, that older US women with lower sleep time have the same chance of cognitive impairment as those with a higher time.
- Calculate the probability of a result as or more extreme as observed.
- Called the *p-value*



P-value

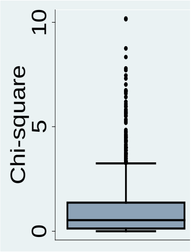
If the p-value is small, then there are two possibilities:

- The null hypothesis is true and something highly unusual occurred,

OR

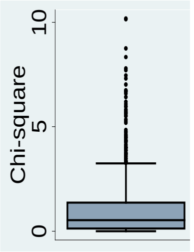
- The null hypothesis is false, and we can conclude there is a difference or relationship.

Poor sleep/cognition



	Total sleep time			
Cognitive Impairment	Low (N)	(<4.5hr)	High (N)	(>6.5hr)
Yes	41	42	37	36
No	57	56	62	63
Total	98		99	

More extreme

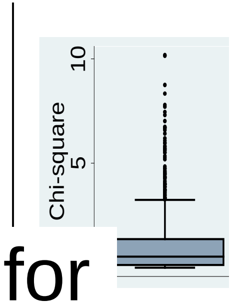


Poor sleep/cognition

	Total sleep time			
Cognitive Impairment	Low (N)	(<4.5hr) (%)	High (N)	(>6.5hr) (%)
Yes	41	42%	37	37%
No	57		62	
Total	98		99	

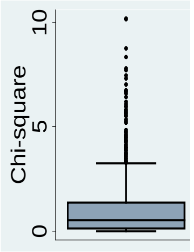
Conclusion?

Interpreting p-values – rules of thumb



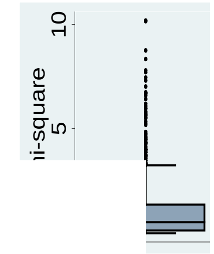
- P<0.05 is widely accepted as the cut-off point for rejecting the null hypothesis. Values less than 0.05 are commonly described as *statistically significant*.
- P>0.10: consistent with the null hypothesis.
  - But doesn't prove it.
- 0.05<P<0.10: suggestive that the alternative hypothesis is true.
- 0.01<P<0.05: supportive of the alternative hypothesis. Statistically significant.
- P<0.01: strongly supportive of the alternative hypothesis. Statistically significant.

PPPs – some sand



- P-value Pet Peeve1: Authors who describe a statistically significant result as "significant" without the "statistically" qualifier.
  - Statistically significant = detectable difference (of any magnitude)
  - Significant difference = difference of important magnitude.
- P-value Pet Peeve2: Authors who describe a lack of a statistically significant result as no effect or no association.
  - Absence of evidence is not the same as evidence of absence.
  - Need to consider confidence intervals to assert the latter.

P-value example: N Engl J Med 2010, p.363



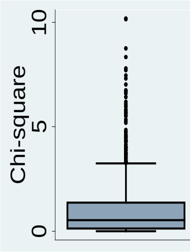
BACKGROUND

The optimal management of a torn anterior cruciate ligament (ACL) of the knee is unknown.

METHODS

We conducted a randomized, controlled trial involving 121 young, active adults with acute ACL injury in which we compared two strategies: structured rehabilitation plus early ACL reconstruction and structured rehabilitation with the option of later ACL reconstruction if needed. The primary outcome was the change from baseline to 2 years in the average score on four subscales of the Knee Injury and Osteoarthritis Outcome Score (KOOS) — pain, symptoms, function in sports and recreation, and knee-related quality of life.

P-value example: N Engl J Med 2010, p.363



Results

There were no significant differences in the change in the KOOS4 score from baseline to 2 years between the group

reconstruction and

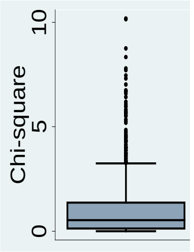
tional delayed

d 39.4,

% confidence

baseline KOOS4

P-value Practice

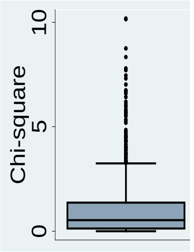


Cognitive Impairment	Sleep Disordered Breathing No (N)	(%)	Yes (N)	(%)
Yes	60	31%	47	45%
No	133		58	
Total	193		105	

Conclusion?

Probability of extreme result by chance

P-value Practice

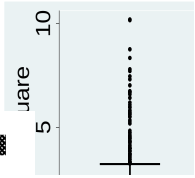


	Waking Min after Sleep Onset			
Cognitive Impairment	Low (N)	(<61min) (%)	High (N)	(>106min) (%)
Yes	31	31%	44	44%
No	68		57	
Total	99		101	

Conclusion?

Probability of extreme result by chance

Sleep/dementia – Table 1



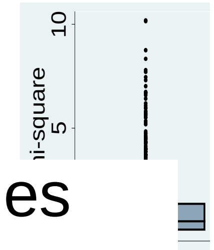
Characteristic	Sleep-Disordered Breathing: No (n = 193)	Sleep-Disordered Breathing: Yes (n = 105)	*P* Value
White, No. (%)	173 (90)	96 (91)	.62
Hypertension, No. (%)	116 (60)	69 (66)	.34
History of stroke, No. (%)	25 (13)	11 (11)	.53
High number of depressive symptoms, No. (%)[b]	23 (12)	9 (9)	.37
Current smoking, No. (%)	5 (3)	0	.17[c]
Current medication use, No. (%)			
Antidepressants	16 (8)	8 (8)	.84
Benzodiazepines	14 (7)	9 (9)	.68
Nonbenzodiazepine anxiolytics	5 (3)	2 (2)	.71
Mini-Mental State Examination, mean (SD)	24.9 (1.2)	25.1 (1.1)	.22
Modified version of Trails B, mean (SD)	130.1 (54.8)	122.9 (52.1)	.48

Current smoking, No. (%) 5 (3) 0 .17[c]

[a] Body mass index was calculated as weight in kilograms divided by height in meters squared.
[b] Score of 6 or greater on Geriatric Depression Scale.
[c] Calculated using the Fisher exact test.

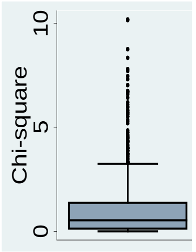
[a] Body mass index was calculated as weight in kilograms divided by height in meters squared.
[b] Score of 6 or greater on Geriatric Depression Scale.
[c] Calculated using the Fisher exact test.

Generating p-values



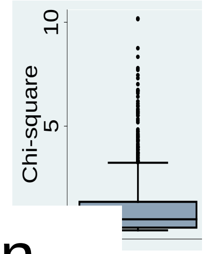
- Statistical hypothesis tests generate p-values, sometimes via arcane calculations.
- Appropriate test depends on nature of the variable studied as well as the outcome.
- Common situations
  - *Binary outcome*, compare groups: chi-square test, Fisher's exact test.
  - *Binary outcome*, compare groups, adjusting for predictor variables: logistic regression.
  - *Numerical outcome*, compare groups: t-test
  - *Numerical outcome*, compare groups, adjusting for predictor variables: linear regression
  - *Survival analysis (time to event),* compare groups: Kaplan-Meier curves, log rank test.
  - *Survival analysis*, compare groups, adjusting for predictor variables: Cox proportional hazards regression.
- They all generate a p-value, so the key is to figure out what the null hypothesis is and interpret the p-value.

Standard errors – understanding the accuracy of data from samples



- A key ingredient in statistics is the *standard error or SE*. From sample to sample, calculated statistics approximate their average value in the population, *give or take a standard error or two.*
- By knowing the SE you can delineate reasonable or unreasonable values of the unknown average values in the population.

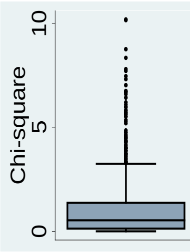
Using standard errors



- Suppose a sample of 105 participants gives a proportion of 0.45 with a SE of 0.05 (rate of cognitive impairment in the disordered breathing group). What can we say about the possibility that the true value (if we evaluated all older women in the U.S.) is as low as 0.30?
- Range of reasonable values is 0.45 plus or minus 2×(0.05) or (0.35, 0.55).
- 0.30 not reasonable for a population value.
- For a binary outcome with a sample size of n and a proportion of p, the SE is given by

$$SE = \sqrt{p(1-p)/n} = \sqrt{0.45(0.55)/105} = 0.049$$

Standard Errors (SE) vs. SD



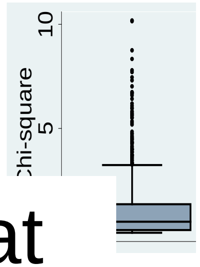
Tip:

Use standard deviations (SDs) to quantify the spread of data values in a sample or population.

Use standard errors (SEs) to quantify the precision of a sample estimate of a population quantity.

For example, if the purpose of a table or graph is to describe the variability of the data use SD. If the purpose is to suggest that the true average values in the population may differ use the SE.

Confidence Intervals

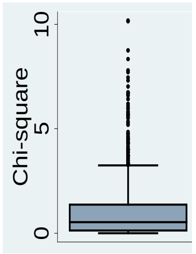


A confidence interval (CI) gives an interval that has a known probability of covering the unknown population value.

The interpretation of “95% confidence” intervals is that, on average, they cover the true population value 95% of the time

CI is often given by: *estimate* $\pm$ *2(SE)*.

Confidence Intervals and Hypothesis Tests



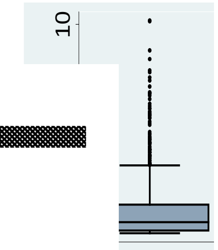
The value of 1 for a relative risk or odds ratio corresponds to the special case of equal risk.

There is a direct correspondence between confidence intervals and hypothesis tests: *A confidence interval for a relative risk or odds ratio contains the value 1 if and only if the hypothesis test cannot reject the null hypothesis* (using 95% confidence and p-value cutoff of 0.05).

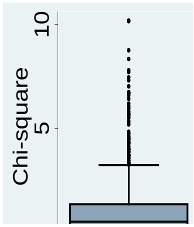
Also, a confidence interval for a difference contains 0 if and only if the hypothesis test cannot reject the null hypothesis of the difference being 0 (equal values).

Table 3. *Mild Cognitive Impairment or Dementia Among Older Women According to Hypoxia, Sleep Fragmentation, or Sleep Duration Measures*

	Mild Cognitive Impairment or Dementia, No. (%) (n = 107)	OR (95% CI)	
		Unadjusted	Adjusted[a]
Hypoxia and Disordered Breathing Measures			
Oxygen desaturation index, events/h			
<15	46 (43.0)	1 [Reference]	1 [Reference]
≥15	60 (56.1)	1.67 (1.03-2.69)	1.71 (1.04-2.83)
Oxygen saturation <90%			
<1% of sleep time	64 (59.8)	1 [Reference]	1 [Reference]
≥1% of sleep time	43 (40.2)	0.87 (0.54-1.41)	0.83 (0.51-1.38)
Sleep time in apnea or hypopnea, %			
Low (median: 0.9 [range, 0-2.2])	31 (29.0)	1 [Reference]	1 [Reference]
Mid (median: 4.4 [range, 2.3-7.0])	31 (29.0)	1.00 (0.55-1.82)	1.16 (0.61-2.20)
High (median: 16.4 [range, 7.0-66.8])	45 (42.1)	1.79 (1.01-3.20)	2.04 (1.10-3.78)

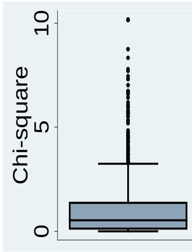


Adjustment using multivariate models



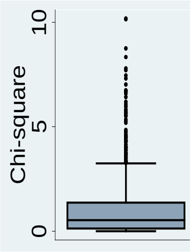
- A key reason to use multivariate models is to allow adjustment for confounders.
- The Yaffe article used logistic regression with adjustment variables that included age, race, body mass index, education level, smoking status, presence of diabetes, presence of hypertension, antidepressant use, benzodiazepine use, and use of nonbenzodiazepine anxiolytics.
- Strengthens the inference of causation.

Confidence Intervals in “Negative” Studies



When H_O is not rejected, the confidence interval gives additional information about the magnitude of the differences and can be quite useful for interpreting “negative” (not statistically significant) findings.

If the confidence interval is narrow, the possible effects will be quite limited. If it is wide, the study might be non-informative.

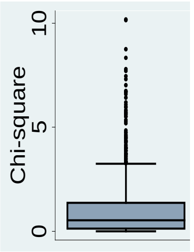


<u>Example</u>: Harris, Bebe-Donk and Alshafie (*BMC Cancer* 6: 27, 2006) study the relationship of breast cancer to use of COX-2 inhibitors in a case-control study at a single hospital. They find (their Table 2) the following odds ratios and 95% confidence intervals (CI):

COX-2	OR = 0.29,	CI (0.15, 0.59)
Acetaminophen	OR = 1.02,	CI (0.39, 2.20)
Baby aspirin	OR = 0.77,	CI (0.42, 1.41)

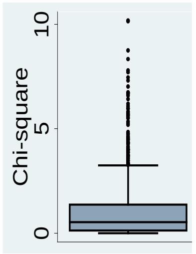
“Neither acetaminophen nor baby aspirin (81 mg) had any effect on the relative risk of breast cancer.”

Stata software



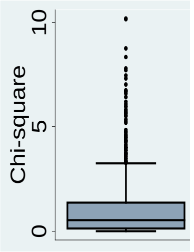
Fisher's exact test from the Yaffe article:

"Forty-seven women (44.8%) with prevalent sleep-disordered breathing developed mild cognitive impairment or dementia compared with 31.1% (60/193) of those without sleep-disordered breathing (*P*=.02)."



Stata software

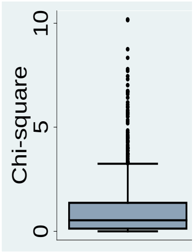
```
cogn_impair |       sleep_dis
            |         0          1 |     Total
 -----------+----------------------+----------
          0 |       133         58 |       191
            |     68.91      55.24 |     64.09
 -----------+----------------------+----------
          1 |        60         47 |       107
            |     31.09      44.76 |     35.91
 -----------+----------------------+----------
      Total |       193        105 |       298
            |    100.00     100.00 |    100.00

           Fisher's exact =                 0.023
   1-sided Fisher's exact =                 0.013
```



Stata software

Logistic regression for cognitive impairment data.

Sleep Fragmentation Measures			
Arousal index, arousals/h			
Low (median: 10.1 [range, 2.4-14.5])	44 (41.1)	1 [Reference]	1 [Reference]
Mid (median: 18.2 [range, 14.6-22.6])	30 (28.0)	0.52 (0.29-0.94)	0.54 (0.29-0.98)
High (median: 33.1 [range, 22.6-66.4])	32 (29.9)	0.59 (0.34-1.06)	0.58 (0.32-1.07)
Wake after sleep onset, min			
Low (median: 40.7 [range, 2.0-61.0])	31 (29.0)	1 [Reference]	1 [Reference]
Mid (median: 82.0 [range, 62.0-105.0])	32 (29.9)	1.06 (0.58-1.94)	1.17 (0.63-2.19)
High (median: 170.6 [range, 108.0-336.0])	44 (41.1)	1.69 (0.95-3.02)	1.79 (0.97-3.29)

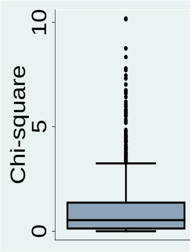


Stata software

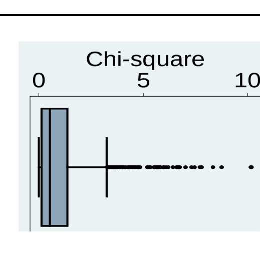
Logistic regression for cognitive impairment data.

```
.
. logistic cogn i.wake


------------------------------------------------------------------------------
 cogn_impair | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        wake |
          2  |   1.063539   .3249723     0.20   0.840     .5843353    1.935728
          3  |   1.693265   .5001159     1.78   0.075     .9491125    3.020873
             |
       _cons |   .4558824    .098795    -3.62   0.000     .2981175    .6971369
------------------------------------------------------------------------------
```

Summary

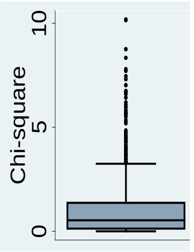


- Roles of statistics.
- Inferential: p-values, standard errors, confidence intervals.
- P<0.05. Statistically significant. Detectable difference, not necessarily important.
- Use confidence intervals to interpret “negative” studies and importance.
- Use multivariate adjustment for confounders.



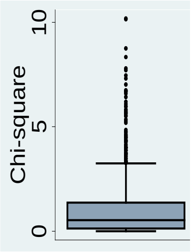
Chi-square
0
5
10

Statistical vs Practical Significance



Sample size per group	Difference between Male and Female Avg Cholesterol		
	20 mg/dl	10 mg/dl	1 mg/dl
N=10	p=0.12	p=0.43	p=0.94
N=100	$p \approx 0$	p=0.01	p=0.80
N=10,000	$p \approx 0$	$p \approx 0$	p=0.009

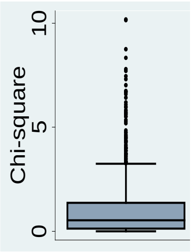
Statistical vs Practical Significance



Sample size per group	Difference between Male and Female Avg Cholesterol: 20 mg/dl	10 mg/dl	1 mg/dl
N=10	p=0.12	p=0.43	p=0.94
N=100	p≈0**	p=0.01**	p=0.80
N=10,000	p≈0**	p≈0**	p=0.009**

**Statistically significant.

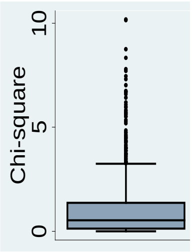
Statistical vs Practical Significance



	Difference between Male and Female Avg Cholesterol		
Sample size per group	20 mg/dl	10 mg/dl	1 mg/dl
N=10	p=0.12*	p=0.43*	p=0.94
N=100	p≈0*	p=0.01*	p=0.80
N=10,000	p≈0*	p≈0*	p=0.009

*Perhaps practically significant.

Statistical vs Practical Significance



Sample size per group	Difference between Male and Female Avg Cholesterol		
	20 mg/dl	10 mg/dl	1 mg/dl
N=10	p=0.12*	p=0.43	p=0.94
N=100	p≈0	p=0.01	p=0.80
N=10,000	p≈0	p≈0	p=0.009**

*Not statistically significant, perhaps practically significant.

**Statistically significant, but not practically significant.