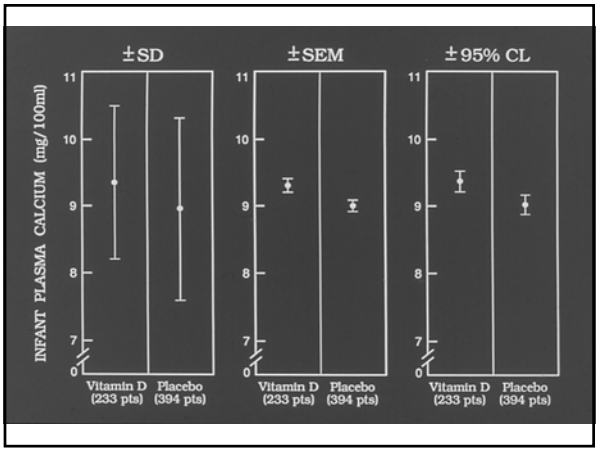
# Stats 101

***Phil Hahn***
***Queen's University***






## Objectives

- To understand and know how to calculate confidence intervals.
- Understand the relationship between confidence intervals and P values.
- Roadmap for picking the correct statistic.

## Types of Data

## Nominal Data

- Data which fits into categories
- Sex: male / female
- Pregnancy: yes / no
- Counts are compiled as a proportions

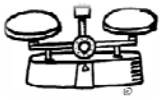
- Pregnancy Rate = 35%
- Proportion voting for Liberals = 60%

## Ordinal Data

- Categorical data with levels that can be ordered
- Pain: mild / moderate / severe
- Data often analyzed as ranks: 1, 2, 3

## Continuous Data

- Data obtained by measuring
- Height, weight, age, blood pressure
- Values are usually compiled as means



## Type of Data

- Determines where you go with your statistical analysis.

## Descriptive Statistics

## Continuous Data

## Measures of Central Tendency

- Mean - average
- Median - middle
- Mode - most

## Measures of Dispersion

- How variable is the data set?

## Range

- Lowest - highest value
- Not really that informative
- Usually the outliers
- Not used in any statistical methods

## Variance

- Measure of the spread of the data
- Average of the squared deviations of the individual values from the mean
- Sum $\frac{(\text{Mean - value})^2}{(n - 1)}$
- See: Gaddis & Gaddis. Introduction to Biostatistics: Part 2, Descriptive statistics. Annals of Emergency Medicine 1990 March;19(3):312-13.

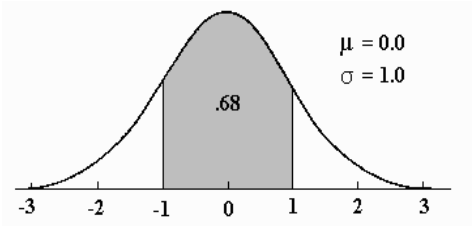
## Variance

- Sum $\frac{(\text{Mean - value})^2}{(n - 1)}$
- Squared units aren't very meaningful

## Standard Deviation (SD)

- Square root of the variance
- Measure of the spread of the sample data

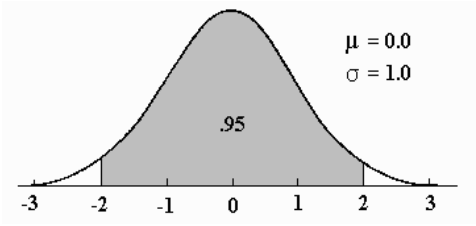
## Normal Distribution




$\overline{X}$

Mean ± 1SD contains 68% of data points

## Normal Distribution




$\overline{X}$

Mean ± 2SD contains 95% of data points

## Standard Error of the Mean (SEM)

- Indicates the degree of uncertainty in calculating a point estimate (mean or proportion)
- Somewhat abstract
- Mathematically simple: SE = SD ÷ √n
- Used to construct confidence intervals

## Confidence Interval (CI)

- Range of values within which the 'true' population parameter is believed to be found, with a given level of confidence.

## 95% Confidence Interval (CI)

- Range of possible truths.
- We can be 95% certain that the true value lies within this range.

## Confidence Interval (CI)

**Mathematically**

95% CI = sample estimate ± 1.96 x SE
(mean)
(proportion)

## Confidence Interval (CI)

**Example**

- Clinical Trial for prevention of infant hypocalcaemia

pregnant moms → Vitamin D
pregnant moms → Placebo

- Infant's plasma Ca measured after birth.

## Infant Plasma Ca (mg per 100 ml)

| Group: | N | Mean | SD |
|---|---|---|---|
| *Vitamin D* | *233* | *9.36* | *1.15* |
| Placebo | 394 | 9.01 | 1.33 |

What is the mean infant plasma Ca level if all pregnant women were given *Vitamin D*?

SE = SD ÷ √n = 1.15 ÷ √233 = 0.075

## Infant Plasma Ca (mg per 100 ml)

| Group: | N | Mean | SD |
|---|---|---|---|
| *Vitamin D* | *233* | *9.36* | *1.15* |
| Placebo | 394 | 9.01 | 1.33 |

SE = 0.075

95% CI = sample mean ± 1.96 (SE)
= 9.36 ± 1.96 (0.075)
= 9.21 to 9.51

## Infant Plasma Ca (mg per 100 ml)

| Group: | N | Mean | SD |
|---|---|---|---|
| *Vitamin D* | *233* | *9.36* | *1.15* |
| Placebo | 394 | 9.01 | 1.33 |

95% CI = 9.21 to 9.51

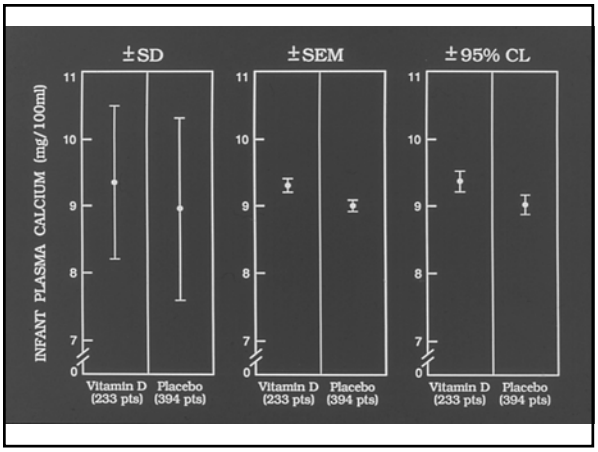
The 'true' population mean for all infants from moms on Vitamin D is between 9.21 and 9.51, with 95% certainty.

## Infant Plasma Ca (mg per 100 ml)

| Group: | N | Mean | SD |
|---|---|---|---|
| *Vitamin D* | *233* | *9.36* | *1.15* |
| Placebo | 394 | 9.01 | 1.33 |

95% CI = 9.21 to 9.51

There is a 1 in 20 chance that the 'true' mean is outside this range.




## P value

- Comes with statistical tests: t tests and chi square tests.
- Tells us the probability that the observed difference occurred by chance.

## Small P values

- Indicate that the observed difference is unlikely due to chance.

## P value < 0.05

- Considered statistically significant
- Means that the probability the result is a chance finding is less than 5%

## P value = 0.004

- Means that the probability of observing such as difference by chance is 4 in 1000.

## t test

- continuous data
- 1 comparison between 2 independent groups

## 2 groups

Lean vs Obese Women

Energy Expenditure (MJ per day)

Means 8.1 vs 10.3

*Extracted from: Douglas G. Altman DG. Practical statistics for medical research. (Chapman & Hall, London, 1991), p.192-3.*

| LEAN (n = 13) | OBESE (n = 9) |
|---|---|
| 6.1 | 8.8 |
| 7.1 | 9.2 |
| 7.5 | 92 |
| 7.5 | 9.7 |
| 7.5 | 9.7 |
| 7.6 | 10.0 |
| 7.9 | 11.5 |
| 8.1 | 11.9 |
| 8.1 | 12.8 |
| 8.1 | |
| 8.4 | |
| 10.1 | |
| 10.9 | |
| Mean = 8.1 | Mean = 10.3 |

## t test

Ho: No difference in energy expenditure

$$t = \frac{\text{mean1 - mean2}}{\text{SE (mean diff)}}$$

$$t = \frac{10.3 - 8.1}{0.56} = \frac{2.2}{0.56}$$

= 3.95

From t table or computer program

P = 0.001

| LEAN (n1 = 13) | OBESE (n2 = 9) |
|---|---|
| 6.1 | 8.8 |
| 7.1 | 9.2 |
| 7.5 | 92 |
| 7.5 | 9.7 |
| 7.5 | 9.7 |
| 7.6 | 10.0 |
| 7.9 | 11.5 |
| 8.1 | 11.9 |
| 8.1 | 12.8 |
| 8.1 | |
| 8.4 | |
| 10.1 | |
| 10.9 | |
| Mean1 = 8.1 | Mean2 = 10.3 |

## t test

P = 0.001

There is a 1 in 1000 chance of observing a difference of 2.2 when the means are really equal.

| LEAN (n = 13) | OBESE (n = 9) |
|---|---|
| 6.1 | 8.8 |
| 7.1 | 9.2 |
| 7.5 | 92 |
| 7.5 | 9.7 |
| 7.5 | 9.7 |
| 7.6 | 10.0 |
| 7.9 | 11.5 |
| 8.1 | 11.9 |
| 8.1 | 12.8 |
| 8.1 | |
| 8.4 | |
| 10.1 | |
| 10.9 | |
| 8.1 | 10.3 |

## t test

**Ho: No difference in energy expenditure**

**P = 0.001 *(less than 0.05)***

**Reject Ho:**

**Conclusion: Energy expenditure in obese women was significantly higher than lean women.**

| LEAN (n = 13) | OBESE (n = 9) |
|---|---|
| 6.1 | 8.8 |
| 7.1 | 9.2 |
| 7.5 | 92 |
| 7.5 | 9.7 |
| 7.5 | 9.7 |
| 7.6 | 10.0 |
| 7.9 | 11.5 |
| 8.1 | 11.9 |
| 8.1 | 12.8 |
| 8.1 | |
| 8.4 | |
| 10.1 | |
| 10.9 | |
| 8.1 | 10.3 |

## 95% CI
## Mean Difference

**= 1.05 to 3.41**

**Does not include 0**

**Statistically significant**

| LEAN (n1 = 13) | OBESE (n2 = 9) |
|---|---|
| 6.1 | 8.8 |
| 7.1 | 9.2 |
| 7.5 | 92 |
| 7.5 | 9.7 |
| 7.5 | 9.7 |
| 7.6 | 10.0 |
| 7.9 | 11.5 |
| 8.1 | 11.9 |
| 8.1 | 12.8 |
| 8.1 | |
| 8.4 | |
| 10.1 | |
| 10.9 | |
| 8.1 | 10.3 |

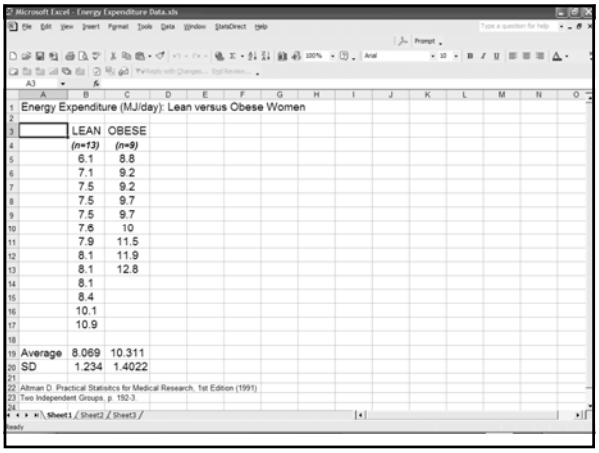
## Important Relationship

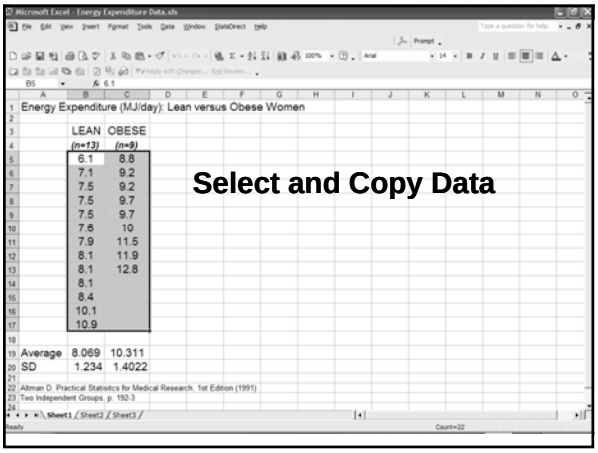
- **P values and Confidence Intervals are related.**
- **If P < 0.05, 95% confidence interval does not include the null value.**
  - **Difference = 0**
  - **Relative Risk or Odds Ratio = 1**

## Example using computer programs.

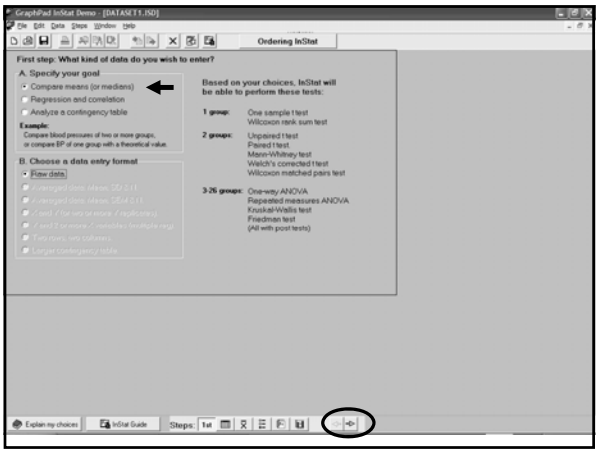
**Microsoft Excel**

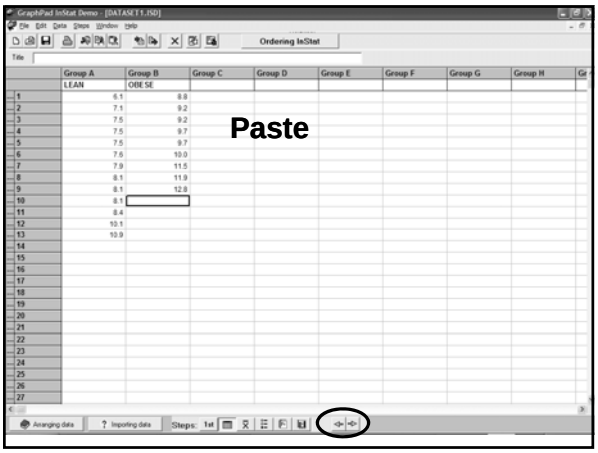
**GraphPad InStat**
***…easiest stats package in the world***

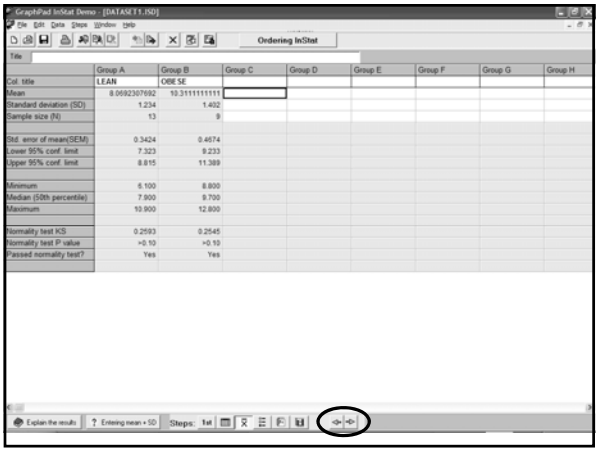


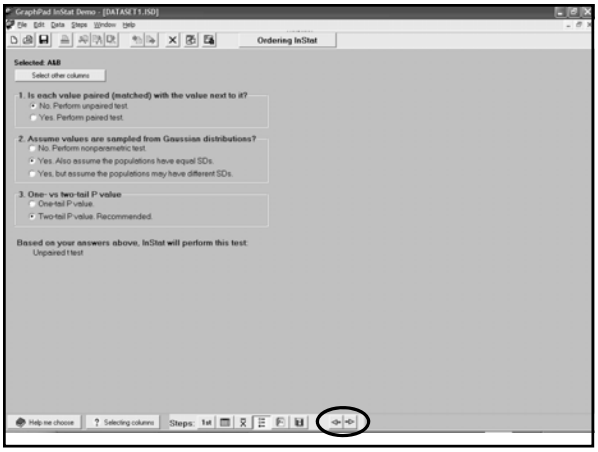



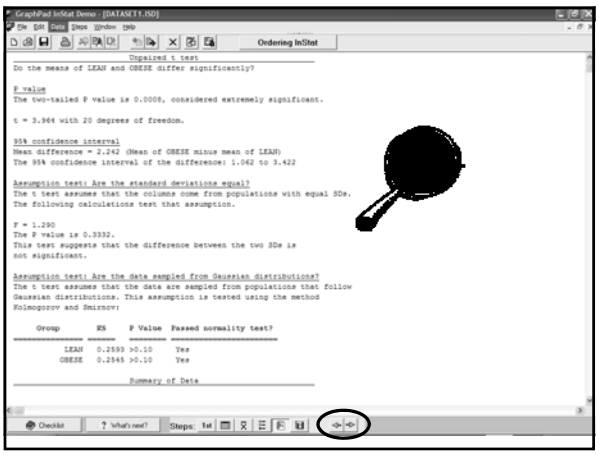

**Select and Copy Data**

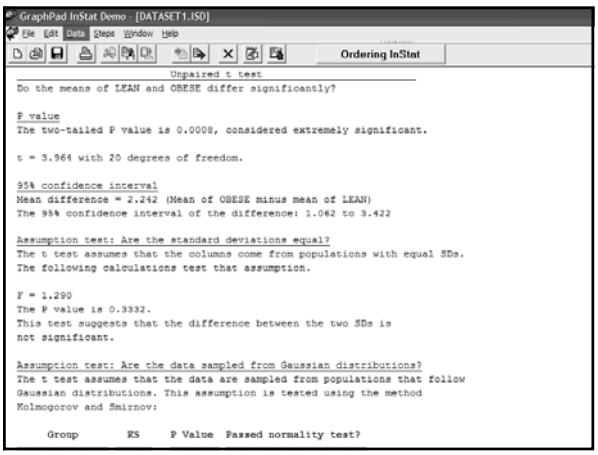

GraphPad InStat Demo - [DATASET1.ISD]
Ordering InStat
First step: What kind of data do you wish to enter?
A. Specify your goal
Compare means (or medians)
Regression and correlation
Analyze a contingency table
Example:
Compare blood pressures of two or more groups, or compare BP of one group with a theoretical value.
B. Choose a data entry format
Raw data
Based on your choices, InStat will be able to perform these tests:
1 group:
One sample t test
Wilcoxon rank sum test
2 groups:
Unpaired t test
Paired t test
Mann-Whitney test
Welch's corrected t test
Wilcoxon matched pairs test
3-26 groups:
One-way ANOVA
Repeated measures ANOVA
Kruskal-Wallis test
Friedman test
(All with post tests)
Explain my choices
InStat Guide
Steps:

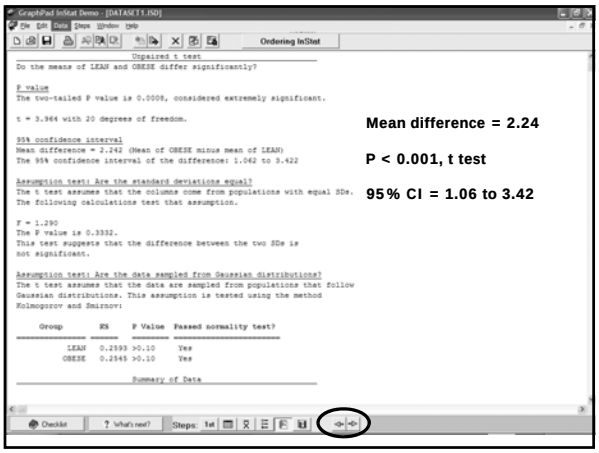

GraphPad InStat Demo - [DATASET1.ISD]
Ordering InStat
Paste
Group A
Group B
Group C
Group D
Group E
Group F
Group G
Group H
LEAN
OBESE
Arranging data
Importing data
Steps:


GraphPad InStat Demo - [DATASET1.ISD]
Ordering InStat
Group A
Group B
Group C
Group D
Group E
Group F
Group G
Group H
LEAN
OBESE
Mean
Standard deviation (SD)
Sample size (N)
Std. error of mean(SEM)
Lower 95% conf. limit
Upper 95% conf. limit
Minimum
Median (50th percentile)
Maximum
Normality test KS
Normality test P value
Passed normality test?
Explain the results
Entering mean + SD
Steps:


GraphPad InStat Demo - [DATASET1.ISD]
Ordering InStat
Selected: A&B
Select other columns
1. Is each value paired (matched) with the value next to it?
No. Perform unpaired test.
Yes. Perform paired test.
2. Assume values are sampled from Gaussian distributions?
No. Perform nonparametric test.
Yes. Also assume the populations have equal SDs.
Yes, but assume the populations may have different SDs.
3. One- vs two-tail P value
One-tail P value.
Two-tail P value. Recommended.
Based on your answers above, InStat will perform this test:
Unpaired t test
Help me choose
Selecting columns
Steps:


GraphPad InStat Demo - [DATASET1.ISD]
Ordering InStat
Unpaired t test
Do the means of LEAN and OBESE differ significantly?
P value
The two-tailed P value is 0.0008, considered extremely significant.
t = 3.964 with 20 degrees of freedom.
95% confidence interval
Mean difference = 2.242 (Mean of OBESE minus mean of LEAN)
The 95% confidence interval of the difference: 1.062 to 3.422
Assumption test: Are the standard deviations equal?
The t test assumes that the columns come from populations with equal SDs.
The following calculations test that assumption.
F = 1.290
The P value is 0.3332.
This test suggests that the difference between the two SDs is not significant.
Assumption test: Are the data sampled from Gaussian distributions?
The t test assumes that the data are sampled from populations that follow Gaussian distributions. This assumption is tested using the method Kolmogorov and Smirnov:
Group
RS
P Value
Passed normality test?
LEAN 0.2593 >0.10 Yes
OBESE 0.2545 >0.10 Yes
Summary of Data
Checklist
What's next?
Steps:

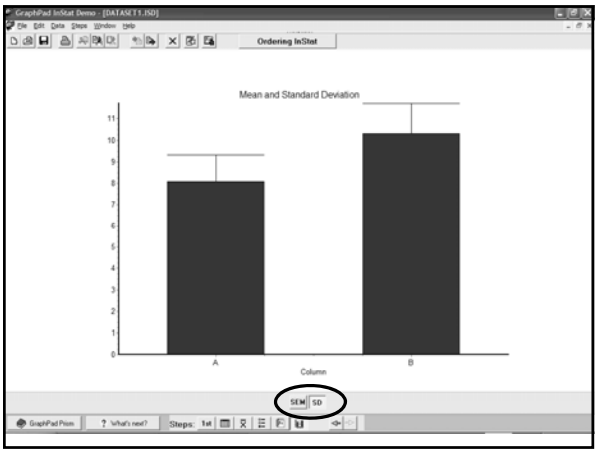

GraphPad InStat Demo - [DATASET1.ISD]
File Edit Data Steps Window Help
Ordering InStat
Unpaired t test
Do the means of LEAN and OBESE differ significantly?
P value
The two-tailed P value is 0.0008, considered extremely significant.
t = 3.964 with 20 degrees of freedom.
95% confidence interval
Mean difference = 2.242 (Mean of OBESE minus mean of LEAN)
The 95% confidence interval of the difference: 1.062 to 3.422
Assumption test: Are the standard deviations equal?
The t test assumes that the columns come from populations with equal SDs.
The following calculations test that assumption.
F = 1.290
The P value is 0.3332.
This test suggests that the difference between the two SDs is not significant.
Assumption test: Are the data sampled from Gaussian distributions?
The t test assumes that the data are sampled from populations that follow Gaussian distributions. This assumption is tested using the method Kolmogorov and Smirnov:
Group
RS
P Value
Passed normality test?

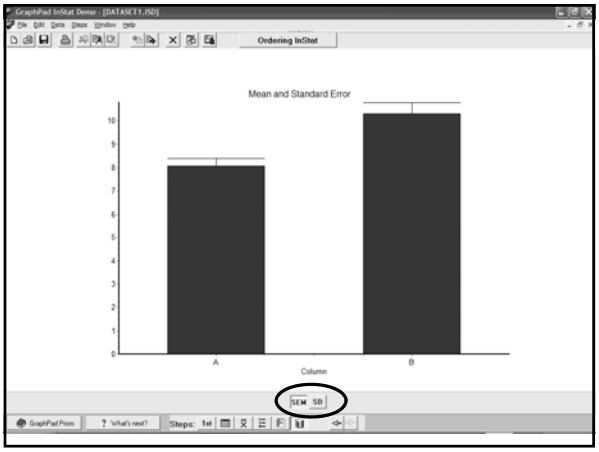


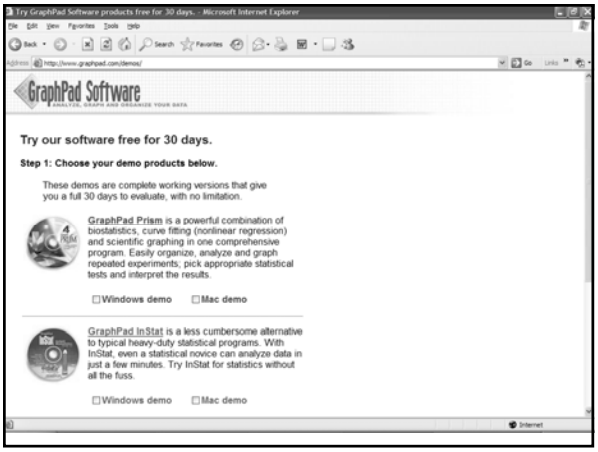

**Mean difference = 2.24**

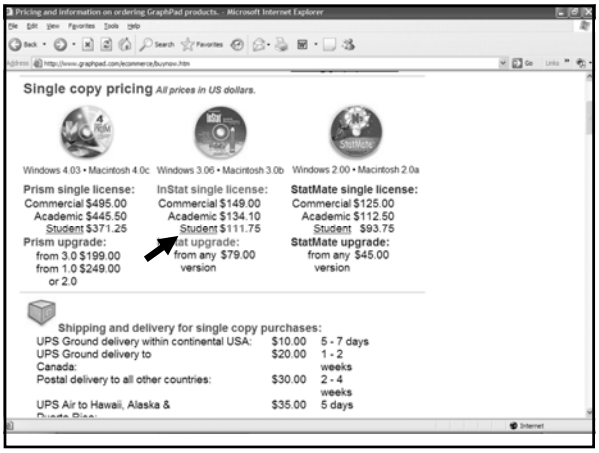
**P < 0.001, t test**

**95% CI = 1.06 to 3.42**



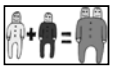




## How to Pick the Correct Statistic

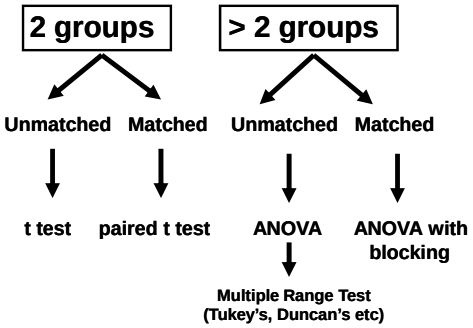
1. **Type of data**
2. **Number of comparison groups**
3. **Are the groups independent (or are they matched)**

## What is matching?

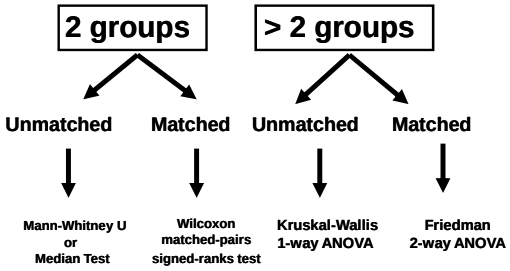


- Observations or measurements made on the same subject (or on individually matched subjects) are said to be "matched" or "paired."
- Examples: Before and after measurements in the same subject. Cases matched to controls on a confounding factor such as age.
- Most tests (t test, ANOVA, Fisher's exact test) assume independence.

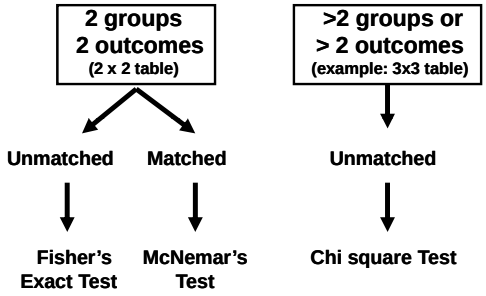
## Continuous Data




## Ordinal Data

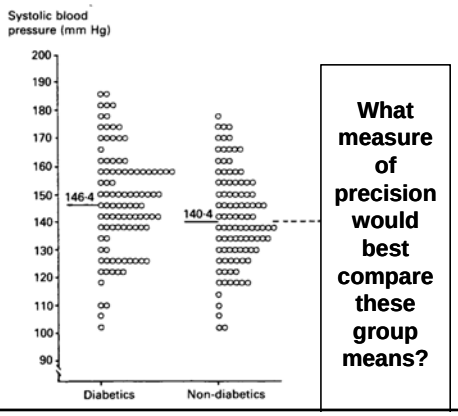


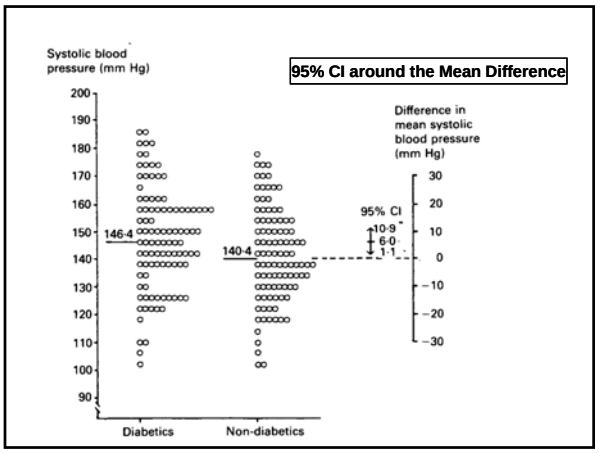

## Nominal Data



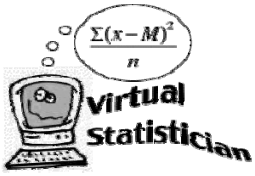

## References


- Martin J. Gardner, Douglas G. Altman. Confidence intervals rather than P values: Estimation rather than hypothesis testing. BMJ 1986;292:746-50.
- Douglas G. Altman. Why we need confidence intervals. World Journal of Surgery 2005;29:554-5.


Systolic blood pressure (mm Hg)
95% CI around the Mean Difference
Difference in mean systolic blood pressure (mm Hg)
200
190
180
170
160
150
140
130
120
110
100
90
146·4
140·4
95% CI
10·9
6·0
1·1
30
20
10
0
−10
−20
−30
Diabetics
Non-diabetics

## Questions?


$\frac{\Sigma(x-M)^2}{n}$
virtual
statistician