Introduction to Biostatistics

Amylou Dueck, PhD
Rare Disease Scholars Program
Sept 7, 2016

Types of Data

- **Quantitative variable:** a variable that can only be recorded using a number (a variable in which taking a mean makes sense)
- **Examples:** BMI, height, blood pressure, number of pregnancies, number of hospitalizations, days of survival
- **Descriptive statistics:** mean, standard deviation, median, inter-quartile range, range, histogram, box plot (special care needed for censored variables)

Types of Data

- **Quantitative variable:** a variable that can only be recorded using a number (a variable in which taking a mean makes sense)

Continuous variables

- **Examples:** BMI, height, blood pressure, number of pregnancies, number of hospitalizations, days of survival
- **Descriptive statistics:** mean, standard deviation, median, inter-quartile range, range, histogram, box plot (special care needed for censored variables)

Types of Data

- **Quantitative variable:** a variable that can only be recorded using a number (a variable in which taking a mean makes sense)
- **Examples:** BMI, height, blood pressure, <u>number of pregnancies, number of hospitalizations,</u> days of survival

  Discrete/count variables
- **Descriptive statistics:** mean, standard deviation, median, inter-quartile range, range, histogram, box plot (special care needed for censored variables)

Types of Data

- **Quantitative variable:** a variable that can only be recorded using a number (a variable in which taking a mean makes sense)
- **Examples:** BMI, height, blood pressure, number of pregnancies, number of hospitalizations, <u>days of survival</u> ← Censored variable
- **Descriptive statistics:** mean, standard deviation, median, inter-quartile range, range, histogram, box plot (special care needed for censored variables)

Types of Data

- **Qualitative (or categorical) variable:** a variable that describes a quality or attribute of the individual
- **Example:** Histologic stage, gender, occupation
- **Descriptive statistics:** frequencies, relative frequencies, bar chart, pie chart

Types of Data

- **Qualitative (or categorical) variable:** a variable that describes a quality or attribute of the individual

Ordinal variable

- **Example:** Histologic stage, gender, occupation
- **Descriptive statistics:** frequencies, relative frequencies, bar chart, pie chart

Types of Data

- **Qualitative (or categorical) variable:** a variable that describes a quality or attribute of the individual

Binary variable

- **Example:** Histologic stage, gender, occupation
- **Descriptive statistics:** frequencies, relative frequencies, bar chart, pie chart

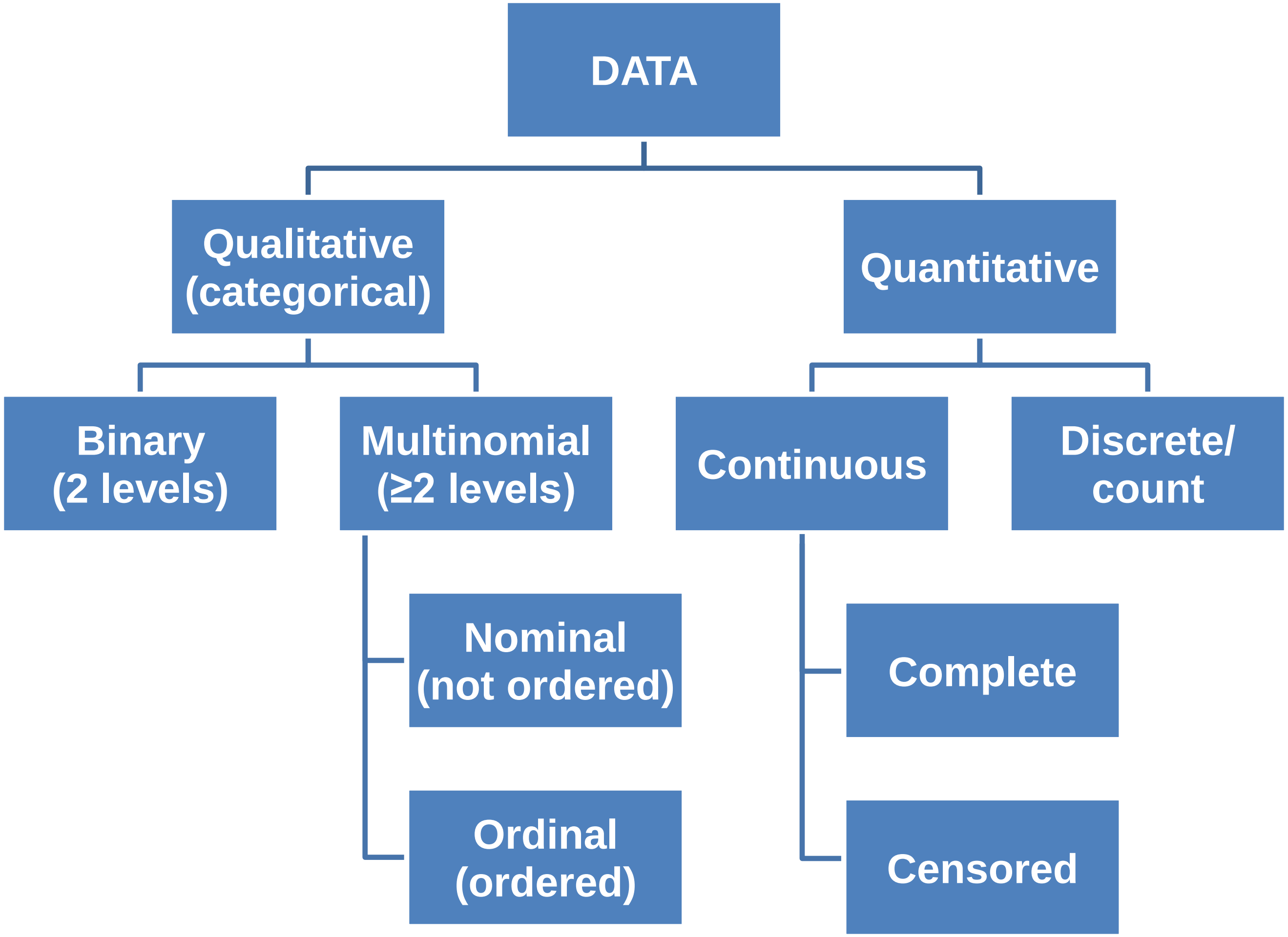
Types of Data

- **Qualitative (or categorical) variable:** a variable that describes a quality or attribute of the individual

Nominal variable

- **Example:** Histologic stage, gender, <u>occupation</u>
- **Descriptive statistics:** frequencies, relative frequencies, bar chart, pie chart

Taxonomy


DATA
Qualitative (categorical)
Quantitative
Binary (2 levels)
Multinomial (≥2 levels)
Continuous
Discrete/ count
Nominal (not ordered)
Ordinal (ordered)
Complete
Censored

Summarizing Quantitative Data
(complete continuous or discrete data)

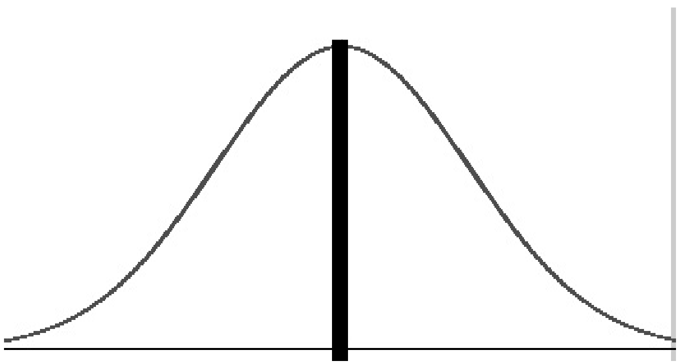
- **Mean:** Simple average
- **Standard deviation:** Measure of how spread out the data are (square root of the "almost average" [use *n*-1 instead of *n*] squared deviation from the mean)
- **Median:** Middle of the ordered values
- **Range:** Largest and smallest value

Example

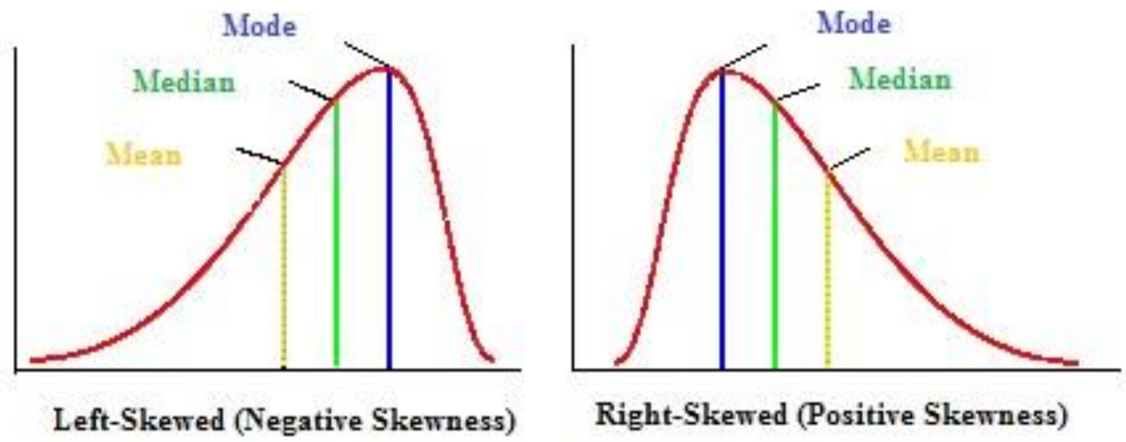
- Donations to ASH Scholars in the amounts of
  - \$20, \$20, \$10, \$30, \$40
  - Sum of these 5 donations is \$120
- Mean is \$120/5 = \$24
- Sum of squared deviations from mean = $(\$20-\$24)^2 + (\$20-\$24)^2 + (\$10-\$24)^2 + (\$30-\$24)^2 + (\$40-\$24)^2 =$ \$520
- SD = sqrt (\$520 / [5-1]) = \$11.40
- Ordered values: \$10, \$20, \$20, \$30, \$40
- Median = \$20
- Range = \$10 to \$40

Mean or Median?

- If data are **symmetric**, mean & median in same place
- If data are **skewed**, mean & median in different places



Mean,
Median,
Mode



- Also, mean is sensitive to outlying values!

Example

- Donations to ASH Scholars in the amounts of
  - $20, $20, $10, $30, $40, **$1,100**
- Mean is **$203.33**
- SD is $439.39
- Median is **$25** (average of the two middle values)
- Range = $10 to $1,100
- **Mean or median?**

Point Estimates

- **Point estimate:** a single value given as an estimate of a parameter of a population (thanks, Google!)
- **Example:** Mean of a sample for a complete continuous variable is a point estimate of the true population mean (sample mean estimates the “location” or “center” for the given variable)

N=**<u>10</u>** women:
Mean weight = 141.5 lb

N=**<u>100</u>** women:
Mean weight = 141.5 lb

N=**<u>1000</u>** women:
Mean weight = 141.5 lb

Confidence Intervals

- We need more information than just a point estimate!
- **Confidence interval:** provides a range of plausible values for the population mean
- We characterize confidence intervals by the confidence level (%)
  - Traditionally, 90%, 95%, or 99% confidence intervals are constructed

N=10 subjects:
Mean weight = 141.5 lb
SD = 30 lb
95% CI: **122.9 - 160.1 lb**

N=100 subjects:
Mean weight = 141.5 lb
SD = 30 lb
95% CI: **135.6 - 147.4 lb**

N=1000 subjects:
Mean weight = 141.5 lb
SD = 30 lb
95% CI: **139.6 - 143.4 lb**

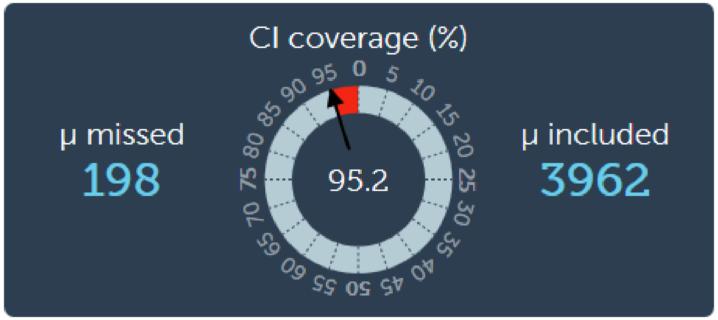
https://www.mccallum-layton.co.uk/tools/statistic-calculators/confidence-interval-for-mean-calculator/#confidence-interval-for-mean-calculator

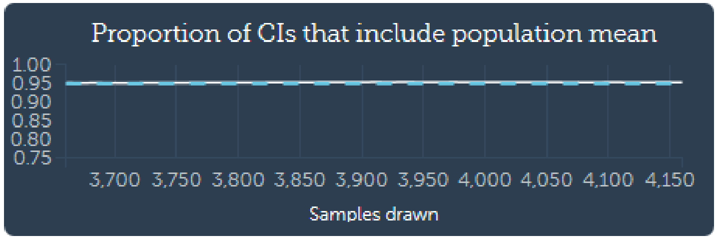
Confidence Intervals

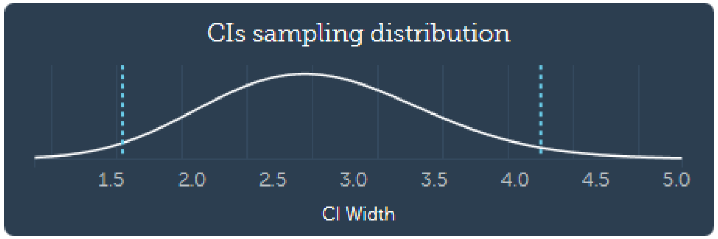
- What does the confidence level mean?
  - (1) Take a sample of size *n* from the population
  - (2) Compute the sample mean
  - (3) Construct a 95% confidence interval
  - (4) Repeat (1)-(3) a thousand times
  - 95% of the confidence intervals will contain the true population mean
  - 5% of the confidence intervals will not
  - Nice visualization: http://rpsychologist.com/d3/CI/

Slide me

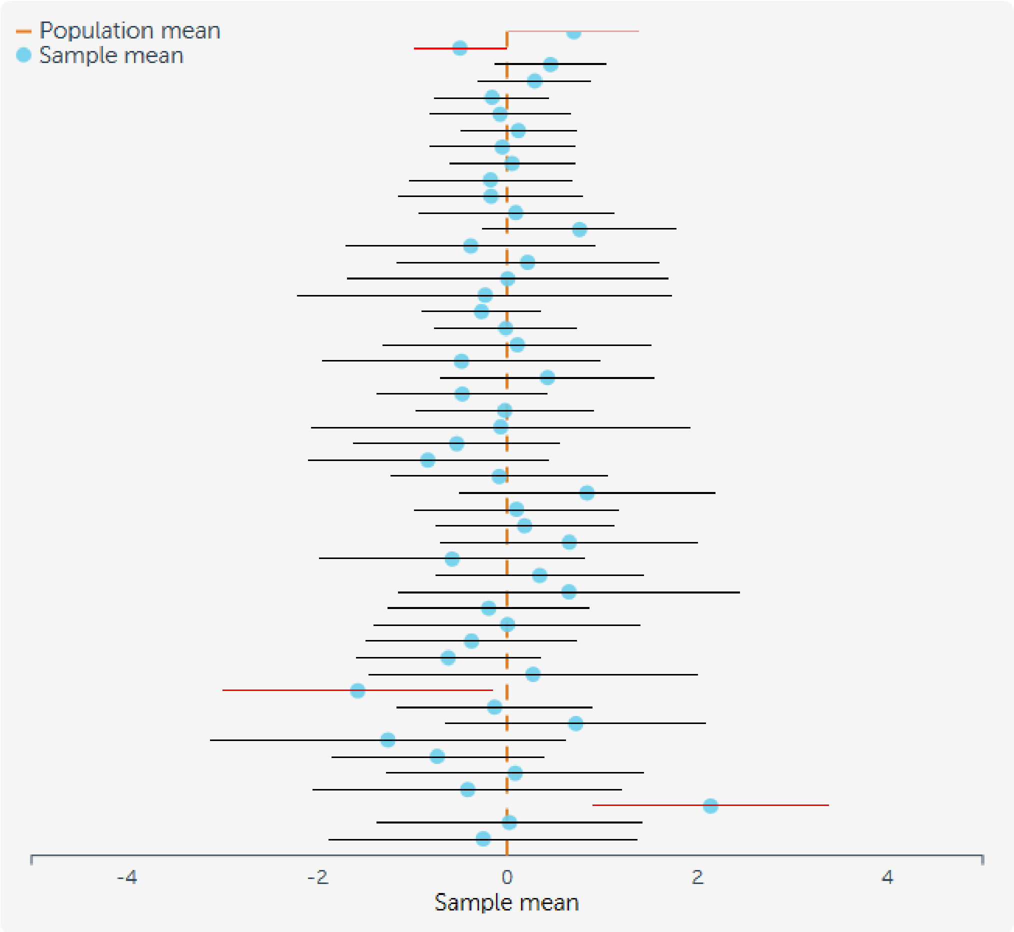
Simulation statistics







95% confidence intervals



Sample size - 10 +

Speed - resume +

About the visualization

Some say that a shift from hypothesis testing to confidence intervals and estimation will lead to fewer statistical misinterpretations.

General Properties of Confidence Intervals

- As sample size gets bigger, …

  … confidence interval gets narrower
  - Narrowing in on your target!
- As confidence level gets larger, …

  … confidence interval gets wider
  - To be more confident, you have to include more real estate in your interval!

Hypothesis Testing and P-values

- Ho = null hypothesis
  - Usually “no difference between groups”
- Ha = alternative hypothesis
  - Usually “difference between groups” (this is what you are trying to show)
- (1) Assume Ho is true
- (2) Collect a sample of data
- (3) Figure out the likelihood that the data you observed occurred under Ho (this is the p-value)
- (4) If small p-value (≤0.05), data are **unlikely** to occur under Ho, so **conclude Ha must be true**
- (5) If large p-value (>0.05), date are **reasonably likely** to occur under Ho, so **do not reject Ho (unable to conclude Ha is true)**

Alpha and Power

- **Alpha = Type I error =** probability of declaring Ha to be true when Ho is really true
- **Beta = Type II error =** probability of not rejecting Ho (not concluding Ha) when Ha is really true
- **Power =** 1-beta = probability of concluding Ha when Ha is really true

		Reality	
		<u>Ho Innocent</u> (Ho True)	**<u>Ho Guilty</u> (Ha True)**
Decision	**Reject Ho (conclude Ha true)** **<u>GUILTY!</u>**	Type I Error (α)	Correct Decision
	Fail to Reject Ho (unable to conclude Ha true) **<u>NOT GUILTY!</u>**	Correct Decision	Type II Error (β)

Sample Size Calculation

- Which is worse – Type I or Type II error?
  - Type I error!
- We construct our tests to ensure a low probability of this error
  - Traditionally 0.05 (maybe 0.10 or 0.20 for ph II)
  - Two-sided (maybe one-sided for ph II)
- Then choose a sample size (“evidence”) to ensure acceptable power (probability of “conviction”)

Sample Size Calculation

- We expect the mean QOL of two equally sized groups to be **6** and **7**
- We also expect the standard deviation of the QOL scores to be **1.67**
- Set alpha=0.05 (two-sided)

Power	Sample Size
80%	44 per group (88 total)*
85%	51 per group (102 total)*
90%	59 per group (118 total)*

http://www.stat.ubc.ca/~rollin/stats/ssize/n2.html

*Consider adding patients to allow for dropouts!

Confidence Intervals and Hypothesis Testing

- If the 95% confidence interval does NOT contain the null hypothesis (Ho), then you can reject Ho!
- **Example:** Cox proportional hazards model (of survival) with two groups, Ho is hazard ratio = 1 (ie, the risk of death is the same in both groups).
  - 95% CI = [1.4, 2.6], then we can reject Ho and conclude that there is a difference between groups!
  - 95% CI = [0.8 1.9], then we are not able to reject Ho and we conclude that there is not enough evidence to conclude that there is a difference between groups.

What Type of Analysis Do I Use?

- Depends on the type of data you have
- Depends on how may groups you have

Which method should I use?

	One Group	Binary Predictor: 2 Independent Samples	Binary Predictor: 2 Paired/Matched Samples	Continuous Predictor & Multiple Predictors
Continuous Outcome	t-test & CI Wicoxon's signed rank test	2-sample t-test & CI Wilcoxon's rank sum test	Paired t-test & CI Wilcoxon's signed rank test	Linear regression & ANOVA
Binary or Categorical Outcome	Z test & CI Exact Binomial test	Z test & Chi-squared test Relative risk & CI Odds ratio & CI Fisher's exact test	McNemar's test McNemar's odds ratio & CI Sign test	Logistic regression
Time-to-Event Outcome	Kaplan-Meier curve	Kaplan-Meier curve & logrank test	Not covered within this site: see Other Regression Topics	Cox regression

	Assessing Agreement Without a Gold Standard	Assessing Agreement With a Gold Standard
Assessing Agreement	Kappa Bland-Altman	Sensitivity, specificity, positive & negative predictive values, ROC Curves McNemar's test & the sign test

Parametric vs Nonparametric

- **Parametric:** Rely on data following a particular distribution (eg, that data are normally distributed like a bell-shaped curve)
  – Thankfully, due to the central limit theorem, with a large enough sample, these assumptions are reasonable in most circumstances!
- **Non-parametric:** Make no or few assumptions about the underlying distribution (usually rely on rank order of data values)
  – Non-parametric statistics are generally used with small sample sizes and generally have less power than the corresponding parametric procedure (double whammy!)

Parametric vs Nonparametric

Parametric	**Non-parametric**
Two-sample t-test	Wilcoxon Rank Sum Test (aka, Mann-Whitney U Test)
ANOVA	Kruskal-Wallis
Pearson Correlation	Spearman Correlation

Correlation

- A measure of association between two continuous variables
- **Example:** Years of education and annual income are positively correlated (ie, as years of education increases, annual income increases)
- **Example:** Years of education and years in jail are negatively correlated (ie, as the years of education increases, years in jail decreases)

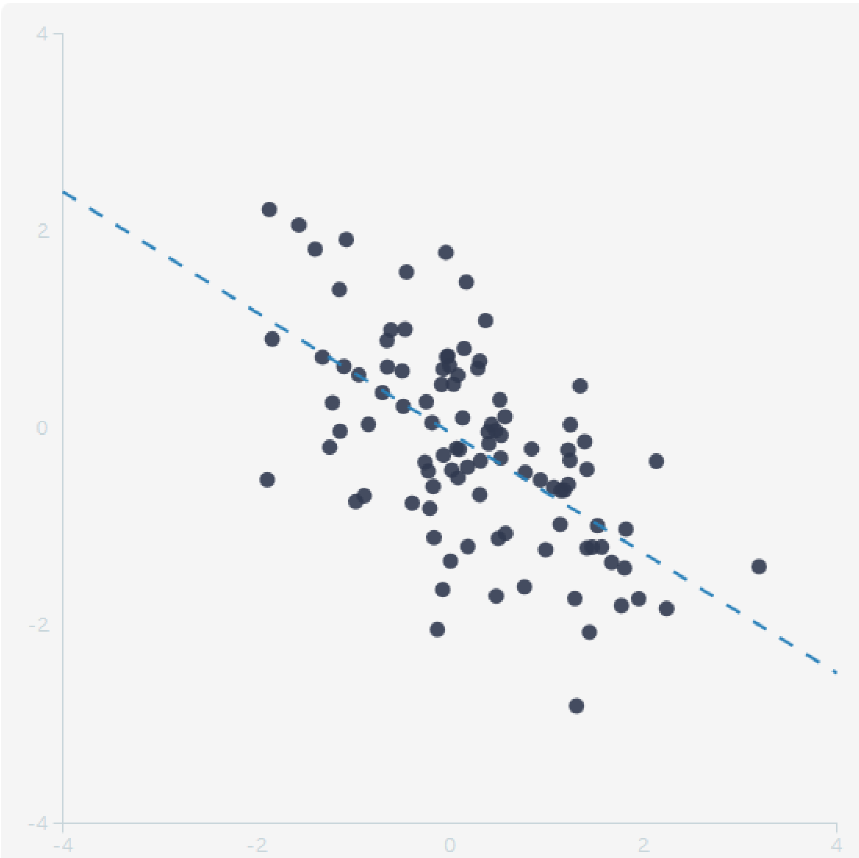
Correlation

- **Parametric =** Pearson correlation
  - Measures how closely pairs of values fall on a straight line (graphically)
  - Range: -1 to 1
  - Negative correlation = variables moving in opposite directions
  - Positive correlation = variables are moving in the same direction
  - The closer the value gets to -1 or 1, the stronger the correlation
  - Software typically produces p-values for Pearson correlations
    - Ho: Correlation = 0 (ie, testing for no correlation)
    - P-value ≤0.05 means that there is significant non-zero correlation (not that there is strong correlation!)
- **Non-parametric =** Spearman correlation

outliers.

http://rpsychologist.com/d3/correlation/

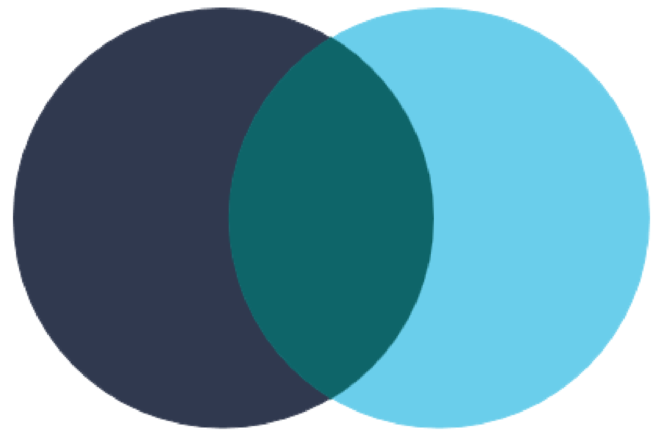
Slide me



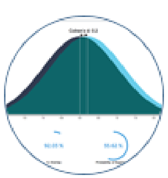
Correlation: -0.61

Sample size 100 New sample

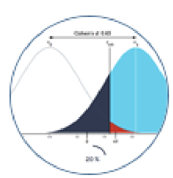
Shared variance: 37.2%



More visualizations



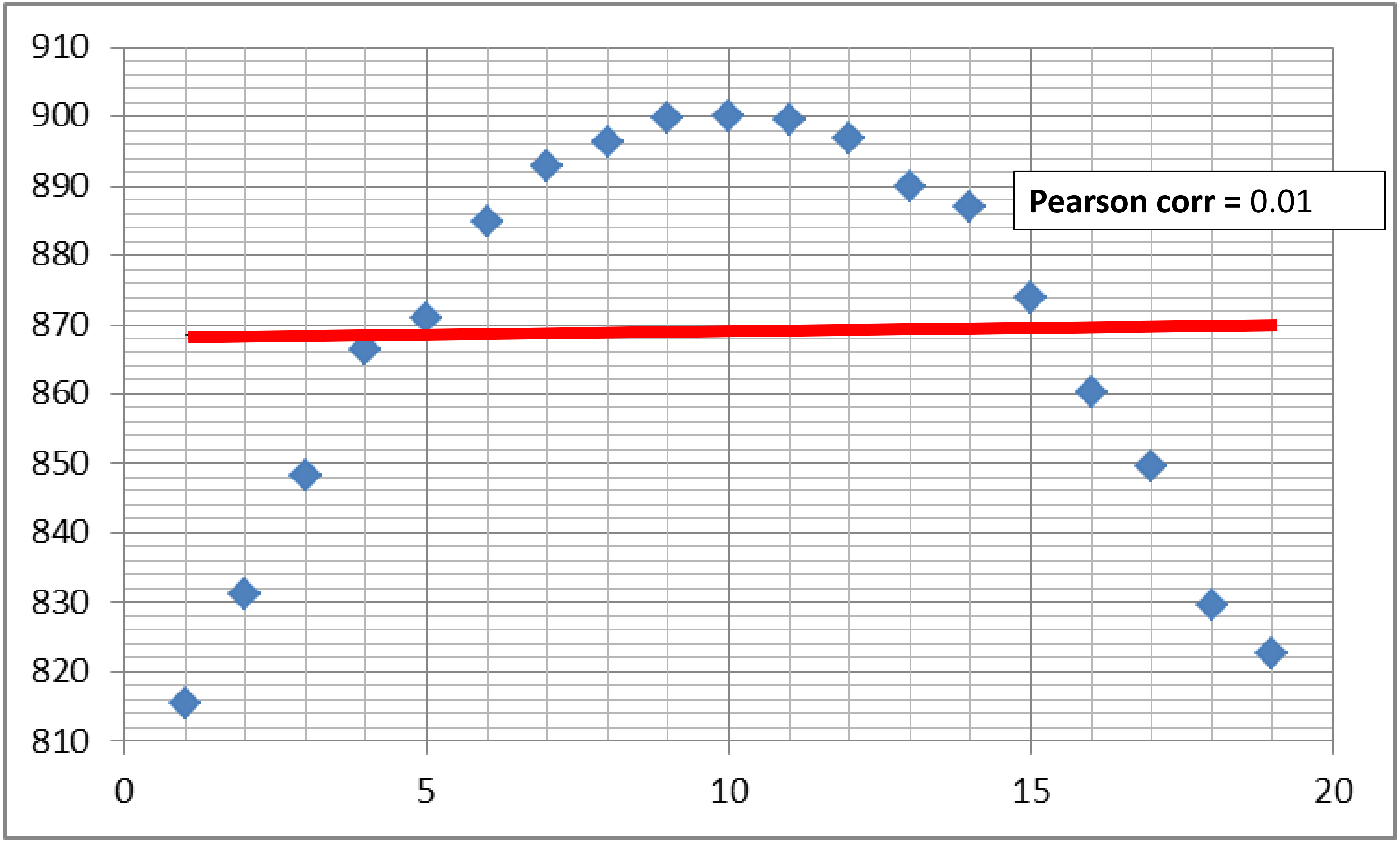
Cohen's d



NHST

Suggestions

Have any suggestion? Send them to me, my contact info can be found here.



Pearson corr = 0.01
910
900
890
880
870
860
850
840
830
820
810
0
5
10
15
20

CORRELATION DOES NOT IMPLY CAUSATION!!!

Fig. 1
IS FACEBOOK DRIVING THE GREEK DEBT CRISIS?

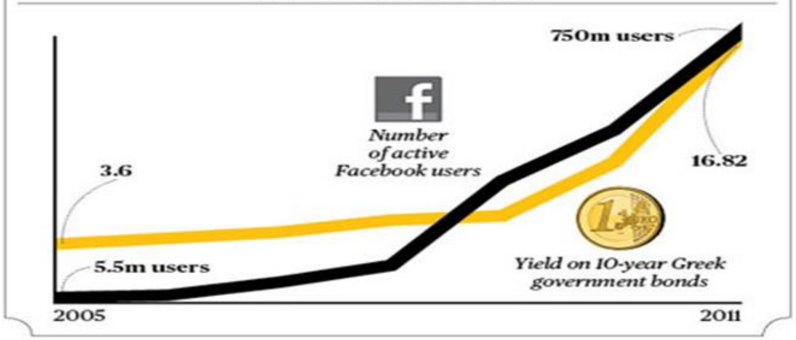


Fig. 2
IS GLOBAL WARMING A HOAX PROPAGATED BY SCIENTISTS?

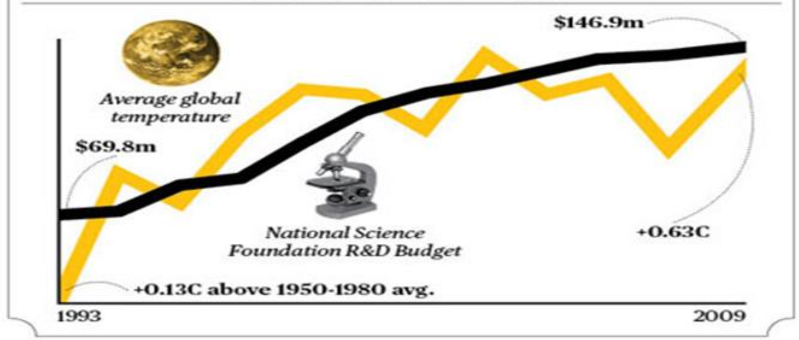


Fig. 3
DID AVAS CAUSE THE U.S. HOUSING BUBBLE?

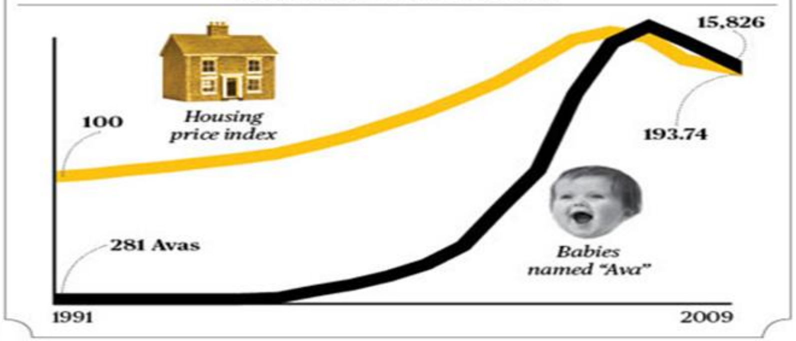


Fig. 4
WOULD M. NIGHT SHYAMALAN START MAKING GOOD MOVIES AGAIN IF PEOPLE BOUGHT MORE NEWSPAPERS?

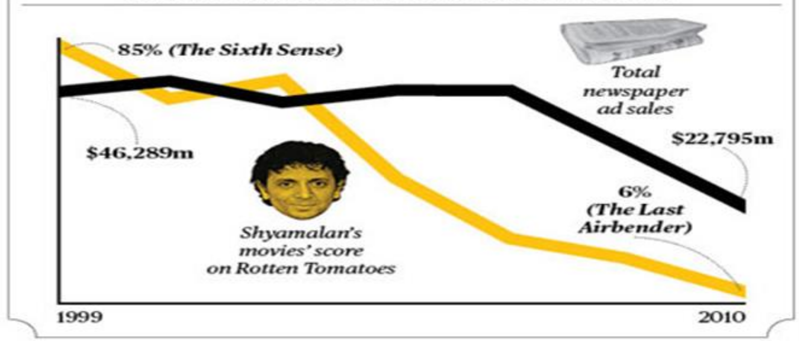


Fig. 5
DID WE TV SABOTAGE MICHELE BACHMANN'S CANDIDACY BY TAKING *STATEN ISLAND CAKES* OFF THE AIR?

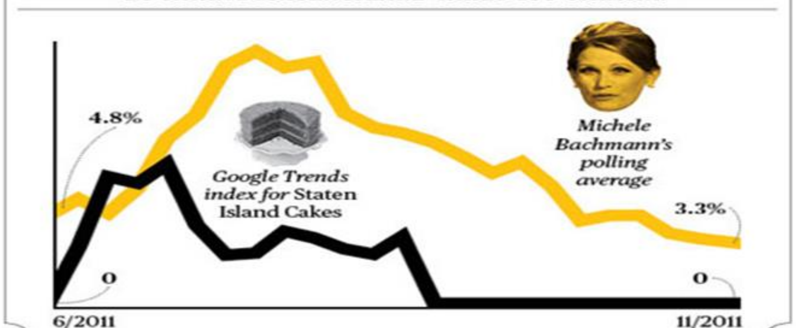


Fig 6
IS THIS MOUNTAIN RANGE AFFECTING THE MURDER RATE?

ALAMY(3); BLOOMBERG(1); GETTY IMAGES(7). DATA: FIG 1: FACEBOOK, BLOOMBERG; FIG 2: NASA, NATIONAL SCIENCE FOUNDATION; FIG 3: U.S. SOCIAL SECURITY ADMINISTRATION, FEDERAL HOUSING FINANCE AGENCY; FIG 4: ROTTEN TOMATOES, NEWSPAPER ASSOCIATION OF AMERICA; FIG 5: GOOGLE, REAL CLEAR POLITICS; FIG 6: NEW YORK LAW ENFORCEMENT AGENCY

Statistical vs Clinical Significance

P-value is not the whole story!

- Does it make sense to adopt a therapeutic agent because the p-value is **0.048** and at the same time ignore another therapeutic agent because the p-value is **0.052**?
  - These two results are entirely consistent!
- A very **large study** may result in a **very small p-value** but a small magnitude of effect
- P-value gives no indication about the clinical importance of the observed association

It's Official:

Bacon and Sausage Cause Cancer

(and are as Dangerous as Cigarettes)

EatDrinkBetter.com

Is bacon really as bad as smoking?

- In terms of statistical evidence, YES!
- In terms of clinical significance, NO!
- Smoking increases your **<u>relative risk</u>** of lung cancer by 2500%
- Eating two slices of bacon per day increases your **<u>relative risk</u>** of colorectal cancer by 18%
- Lifetime **<u>(absolute) risk</u>** of colorectal cancer increases from about 5% to 6%

Top 3 Perils of Statistical Analysis

- Multiplicity
  - Problem behind multiple endpoints, subgroup analysis, interim analysis, cutpoint determination,…
- Missing data
- Multivariate regression

Multiplicity

- **Problem:** "Torture numbers, and they'll confess to anything."
  - Each test has a type I error rate of 5%, but when you perform multiple tests, the type I error rate overall can be much greater than 5%!

Subgroup Example (Lancet 1988, 2[8607]:349-360)

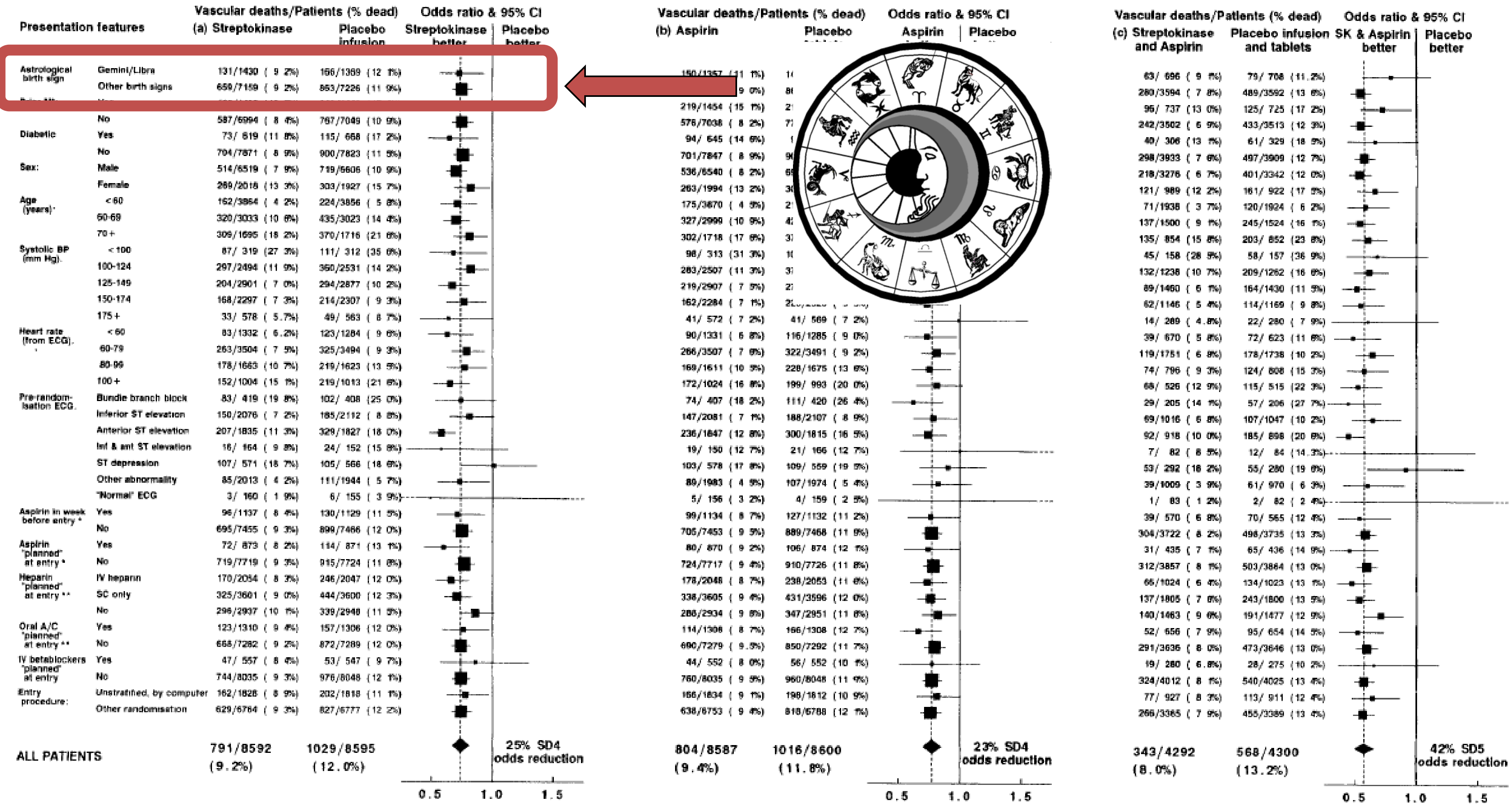
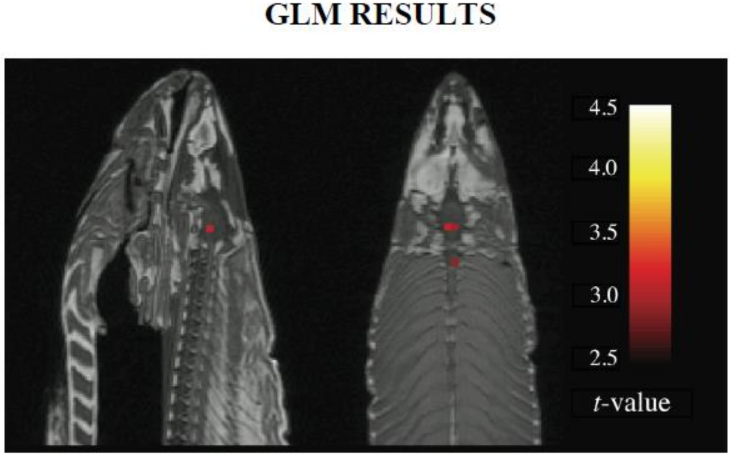


Fig 5—Subgroup analyses of the odds of vascular death in days 0–35.

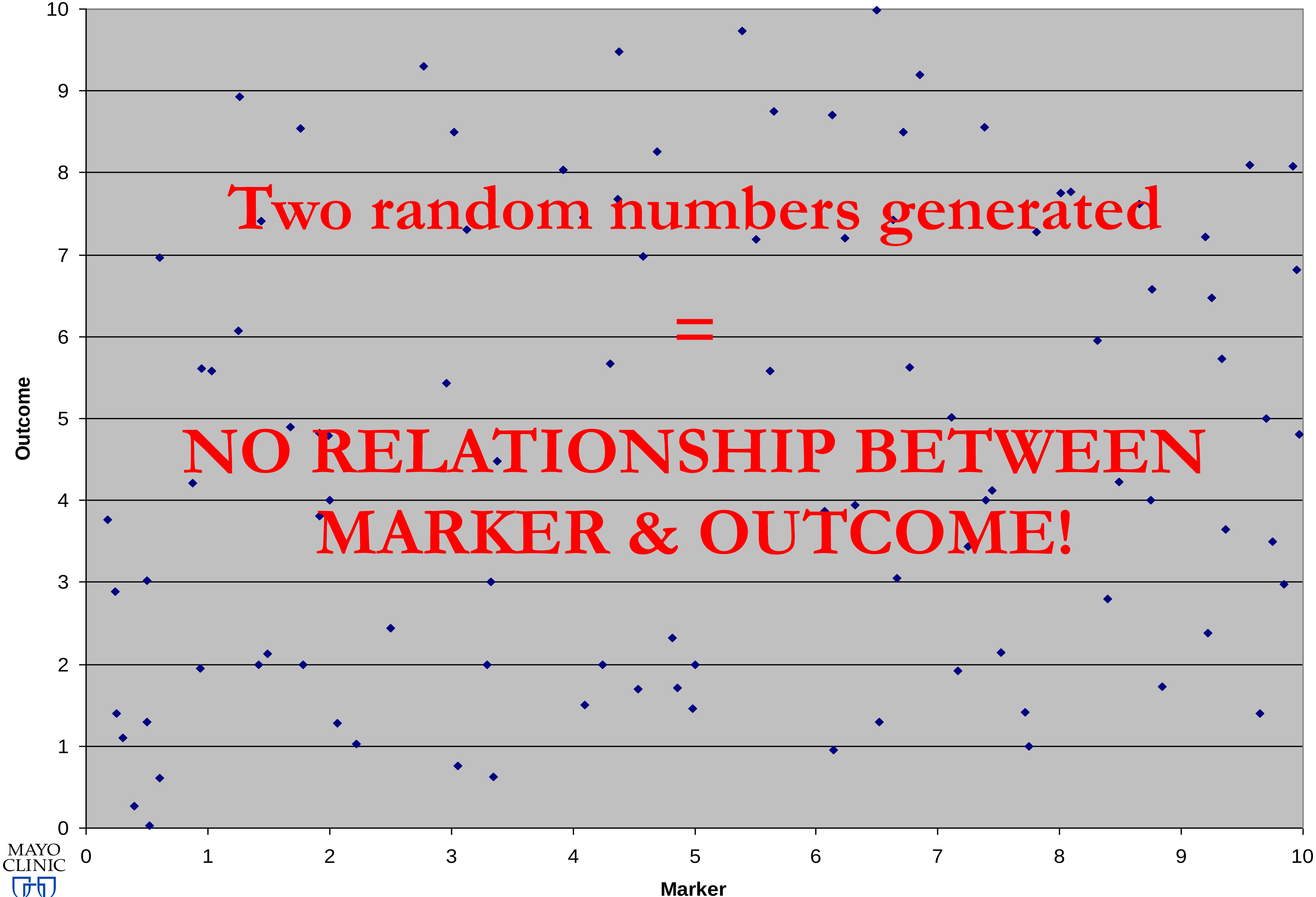
Square sizes[5] and 95% confidence intervals are as in fig 3. Asterisks denote subsidiary analyses that were prespecified in the protocol for aspirin (* and **) or for streptokinase (**). (The sum of the 26 χ-squared test statistics for heterogeneity in the 26 different non-astrological subgroup analyses in fig 5(a) and 5(b) was 58·5 on 50 degrees of freedom, NS. If no real heterogeneity of effect existed then about 1 or 2 of these 26 heterogeneity tests would be expected to yield a $p<0{\cdot}05$ result by chance alone, and in fact only the 1 for aspirin and previous MI did so: all other heterogeneity tests, including that for streptokinase and ECG, were $p>0{\cdot}05$.)

Multiple Testing Example (Bennett, Neuroimage 2009, 47[Suppl 1]:S125)

- Statistically significant difference in emotional responses when presenting different human faces to a subject.
- **The subject being a salmon.**
- **A dead salmon to be particular.**

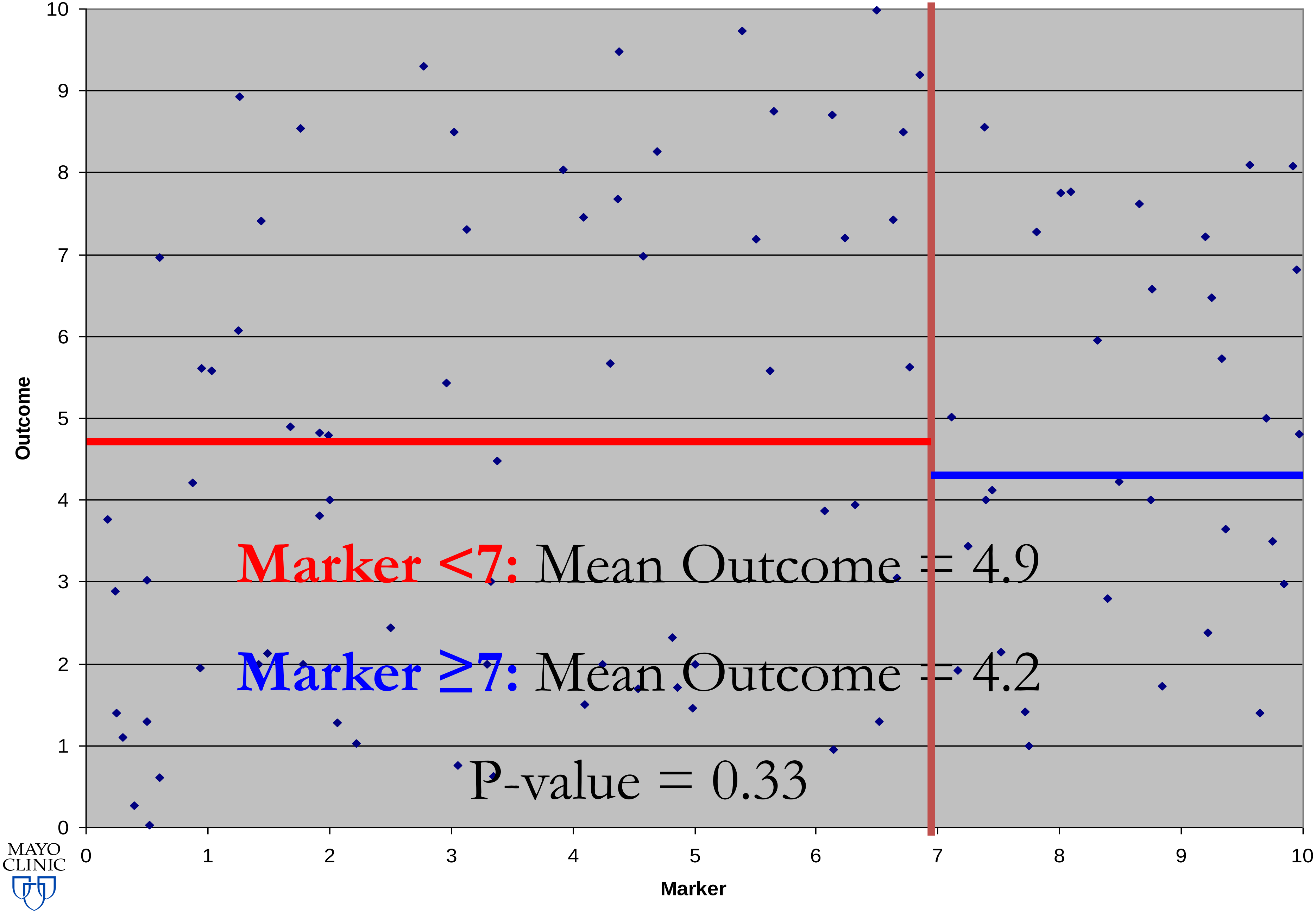


Cut Point Determination - Example



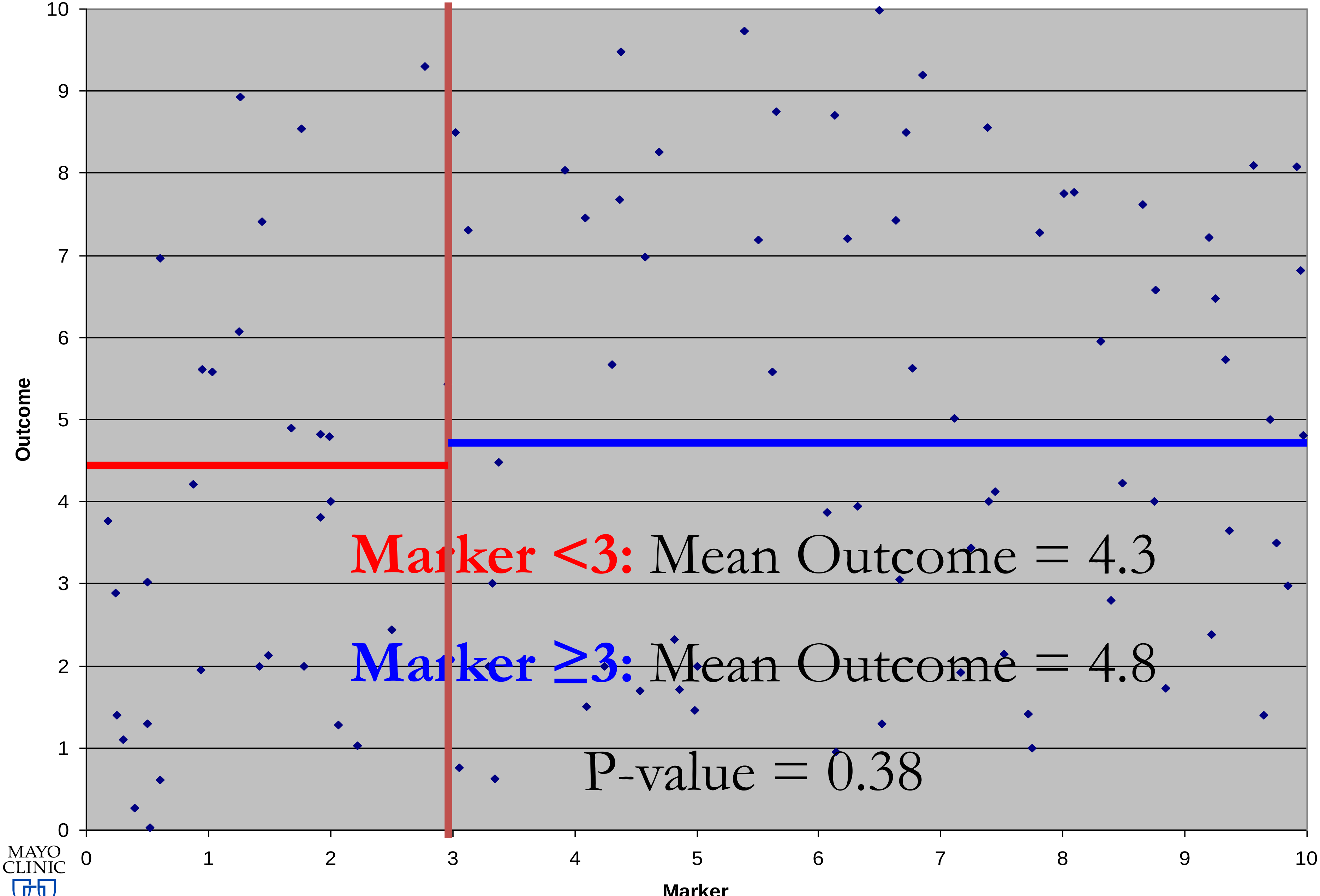
MAYO CLINIC

Cut Point Determination - Example



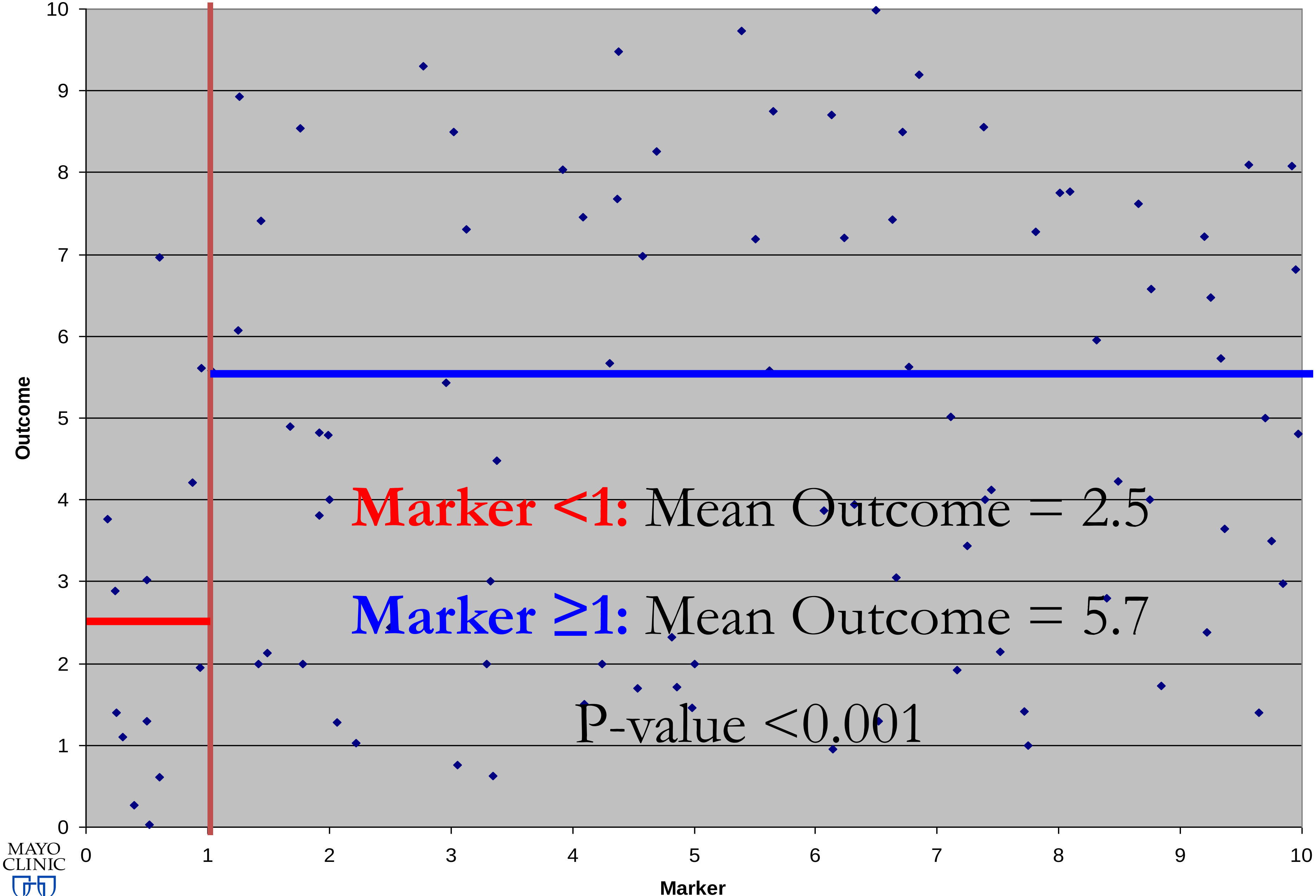
MAYO CLINIC

Cut Point Determination - Example



MAYO CLINIC

Cut Point Determination - Example



MAYO CLINIC

PLOT YOUR DATA!!!!

Missing Data

- **Problem:** Less data = less power, and missingness can be "meaningful"!
  - What does the missing data tell us?
- **Example:** Fatigued patients may be the least likely to fill out a fatigue survey
  - Complete data: 1,4,6,5,2,3,3,4,6,7 (mean = 4.1)
  - Actual data (missing is random): 1,X,6,5,2,X,3,4,X,7 (mean = 4.0)
  - Actual data (three worst fatigue scores missing): 1,4,X,5,2,3,3,4,X,X (mean = 3.1)
- Not just a problem with survey data – can be an issue with clinical data as well!
- **Rigorous data collection a must!** (statistical analysis can only "fix" so many problems)

Multivariate Regression

- **Problem:** Art more than science
- Need to understand relationships among ALL your variables to properly interpret a multivariate model
- Prediction models require validation (minimum internal validation; better external validation; best external validation on multiple datasets)
  - Model selection picks the model which is best for YOUR data and may not extend to another dataset
- **Example**
  - Dependent variable: oxygen saturation
  - Possible independent variables: age, weight, run time, resting pulse, running pulse, maximum pulse
- MANY POSSIBLE "MODELS"

Number in Model	R-Square	Variables					
1	0.7434	runtime					
1	0.1595	rstpulse					
1	0.1584	runpulse					
1	0.0928	age					
1	0.056	maxpulse					
1	0.0265	weight					
2	0.7642	runtime	age				
2	0.7614	runtime	runpulse				
2	0.7452	runtime	maxpulse				
2	0.7449	runtime	weight				
2	0.7435	runtime	rstpulse				
3	0.8111	runtime	age	runpulse			
3	0.81	runtime	runpulse	maxpulse			
3	0.7817	runtime	age	maxpulse			
3	0.7708	runtime	age	weight			
3	0.7673	runtime	age	rstpulse			
3	0.7619	runtime	runpulse	rstpulse			
3	0.7618	runtime	weight	runpulse			
3	0.7462	runtime	weight	maxpulse			
3	0.7452	runtime	maxpulse	rstpulse			
3	0.7451	runtime	weight	rstpulse			
4	0.8368	runtime	age	runpulse	maxpulse		
4	0.8165	runtime	age	weight	runpulse		
4	0.8158	runtime	weight	runpulse	maxpulse		
4	0.8117	runtime	age	runpulse	rstpulse		
4	0.8104	runtime	runpulse	maxpulse	rstpulse		
4	0.7862	runtime	age	weight	maxpulse		
4	0.7834	runtime	age	maxpulse	rstpulse		
4	0.775	runtime	age	weight	rstpulse		
4	0.7623	runtime	weight	runpulse	rstpulse		
4	0.7462	runtime	weight	maxpulse	rstpulse		
5	0.848	runtime	age	weight	runpulse	maxpulse	
5	0.837	runtime	age	runpulse	maxpulse	rstpulse	
5	0.8176	runtime	age	weight	runpulse	rstpulse	
5	0.8161	runtime	weight	runpulse	maxpulse	rstpulse	
5	0.7887	runtime	age	weight	maxpulse	rstpulse	
5	0.5541	age	weight	runpulse	maxpulse	rstpulse	
6	0.8487	runtime	age	weight	runpulse	maxpulse	rstpulse

Which model should you pick? Here are 5 models (with 4 variables each) which explain >80% of the variability in oxygen saturation!

MAYO CLINIC

Multivariate Regression

- Univariate:
  - Running pulse: p=0.02
  - Running time: p<0.001
  - Age: p=0.10
  - Weight: p=0.38
- Multivariate Model #1:
  - Running pulse: p=0.15
  - Running time: p<0.001
- Multivariate Model #2:
  - Running pulse: p=0.002
  - Age: p=0.003
  - Weight: p=0.23

Running pulse not related to oxygen saturation?

Not according to this model!

Results of the model are dependent on what's in the model!

I have no money for a statistician. ~~$$$~~ What do I do?

- Core funding to support statistical analysis
  – Cancer center support grant, CTSA grant
- Institutional/departmental small grants
- Ask your department chair (discretionary dollars)
- Contact your statistician – he/she might be aware of other internal funding opportunities

DOs and **DON'Ts** of working with a statistician

DON'T:

- Start any request with: This should only take a few minutes…
- End any request with: …and my abstract/submission deadline is tomorrow.
- Ask for just a p-value
- Collect data in Excel
- Send updated data without consulting your statistician
- Ask for analysis and then wait a year before getting around to writing the manuscript (and then ask for updated analysis)
- Do the analysis yourself and ask your statistician to fill in the statistical methods section
- Omit your statistician from the author list and then ask your statistician for help on addressing reviewer comments

There is no such thing as a statistical emergency, only poor planning!

DOs and **DON'Ts** of working with a statistician

DO:

- Contact your statistician as early as possible in the research process (**BEFORE COLLECTING DATA**) – if you wait until after, it's often too late!
- Give your statistician as much time as possible on all requests (min 4 weeks)
- Communicate with your statistician about timelines & budgets
- Communicate with your statistician about any changes to data or the project
- Send the IRB number for all requests
- Prioritize analyses (outline your abstract, create mock tables before you request analyses)
- Include reasonable effort on all grants for your statistician
- Include your statistician as a coauthor (preferably 2nd author) on all abstracts and papers (acknowledgement section is not enough)

Treat your statistician as an integral member of your research team!

QUESTIONS?

THANKS!

dueck@mayo.edu

@BiostatGirl