VIRTUAL APPLIED DATA SCIENCE TRAINING INSTITUTE

March 24, 2022

Module 3a: Introduction to Statistics in Scientific Research

Paul Kolm, PhD
Associate Director
Center for Biostatistics, Informatics and Data Science
MedStar Health Research Institute
BERD-CTSA (Georgetown-Howard)

Outline

1. Statistics and estimation
2. Inferential statistics
3. Random sampling variability
4. Theoretical sampling distribution
5. Hypothesis testing and p values
6. Parametric statistics
7. Non-parametric statistics
8. Research design and analysis issues

Impact of Metabolic Syndrome and Diabetes on Prognosis and Outcomes With Early Percutaneous Coronary Intervention in the COURAGE (Clinical *Outcomes Utilizing Revascularization and Aggressive Drug Evaluation) Trial.* *(Maron, et. al.*, JACC, 2011)

Statistical analysis. This was a post hoc analysis. Patient characteristics across groups were assessed by analysis of variance for continuous variables and the chi-square test for discrete variables. The Cochran-Armitage test was used for trend. Estimates of the cumulative event rate were calculated by the Kaplan-Meier method. Time-to-endpoint analyses were done using Cox regression or stepwise Cox regression. Cox regression was also used for the adjusted analyses. The rate of death or MI across subsets or treatment groups was assessed by the chi-square test.

The Risk of Cardiovascular Events in Primary Care Patients Following an Episode of Severe Hypertension. (Ewen, et. al., *J. Clin. Hypertension*, 2009)

Statistical Analysis

The chi-square statistic was used to compare proportions, the Kruskal-Wallis test was conducted to compare means and medians, and the Bonferroni correction procedure was used for all pair-wise comparisons. Poisson and negative binomial regression models were used to adjust observed incidence rates for age. The cumulative incidence of first and multiple cardiovascular events for each BP category was obtained by the Kaplan-Meier method, and differences among BP categories were evaluated using the log-rank test. Time-to-event was measured in years from the date of index until the date of cardiovascular event or censoring. Hazards ratios (HRs) were estimated from the Cox proportional hazards models and adjusted for age, sex, race, diabetes status and control, hyperlipidemia, and BP category at the index date. The proportional hazard assumption was assessed for each covariate and trend tests were performed by using the median value of each BP category as a continuous variable in a proportional hazards model.

A conditional multiple event technique was applied to account for multiple cardiovascular events and the Prentice–Williams–Peterson multiplicative hazards model was used to estimate HRs. Stata version 9 (Stata Corporation, College Station, TX), SAS version 9.1 (SAS Institute, Cary, NC), and SPSS version 15.0 (SPSS Inc, Chicago, IL) were used for statistical analyses.

What are statistics?

- Summaries of individual measurements (e.g., mean)
- Individual measurements implies…..?
- Variability
- Random variable: blood pressure, heart rate, gender, race, SAQ quality of life scores
- Summaries of random variables

Estimating statistics

- What does estimation imply?
- Uncertainty
- **Think of statistics as the measurement (quantification) of uncertainty**

Diastolic BP values of 144 patients

73	42	73	67	84	75	80	78	64	73	70	92
59	77	65	69	91	71	69	62	76	89	91	79
67	76	66	80	98	70	74	71	74	59	84	73
68	74	70	72	84	65	81	54	67	83	87	63
68	93	92	71	95	69	98	74	90	100	60	82
75	91	72	82	67	71	61	94	75	76	66	76
97	58	72	57	74	65	79	73	91	79	63	80
73	92	62	82	73	90	73	83	64	70	60	79
94	73	78	99	71	90	82	74	82	77	71	64
68	73	86	89	100	77	88	72	56	85	90	88
55	68	73	85	80	69	62	73	78	73	84	70
65	81	85	90	89	64	81	77	91	95	86	73

What is the dBP mean?

$$\text{mean} = \sum_{i=1}^{144} \text{dBP}_i \,/\, 144$$

$$\text{mean} = 76.3$$

What is the dBP standard deviation?

$$\text{Variance} = \sum_{i=1}^{144} (\text{mean} - \text{dBP}_i)^2 / (144 - 1)$$

$$\text{SD} = \sqrt{\text{variance}}$$

Average variability about the mean.

Why divide by (144 – 1)?

Degrees of freedom (df)

components of a statistic that are free to vary.

$X_1 + X_2 + X_3 = 100$

50 + ?? + ?? = 100

50 + 40 + ?? = 100

Break:
Python computations of mean and standard deviation.

Inferential statistical analysis

- Comparison of two or more statistics.
- Relationship (association, correlation) of random variables.
- Prediction models of outcomes.
- **dependent variable ~ independent variable(s) + error**

Measurement Scales

- **Continuous -** infinite numerical scale where distances between scale divisions are equal - e.g., height
- **Ordinal -** numerically ordered, but distances between scale divisions are not necessarily equal - e.g., cancer stages, grades
- **Counts** – number of infections in ICU
- **Nominal** (categorical) - values are labels - e.g., gender, race
- **Time-to-event** – time continuous; event dichtomous

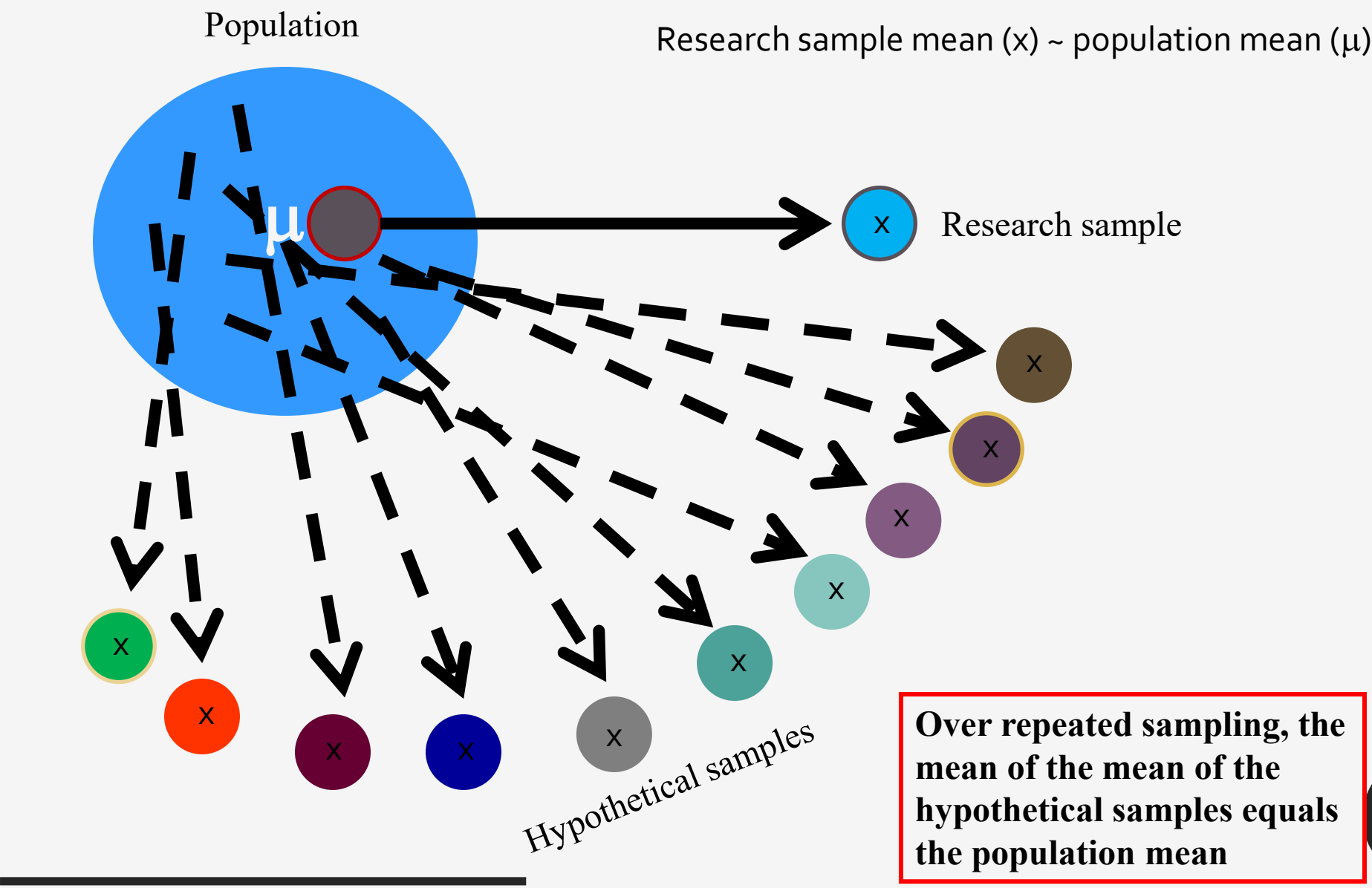
Research designs

- Retrospective, case-control
- Prospective, cohort (non-randomized)
- Experimental - randomized

*RESEARCH STUDIES DESIGNED TO ANSWER QUESTIONS ABOUT A **POPULATION***

Populations and samples

- Population is the universe of all patients (e.g., all CAD patients undergoing PCI)
- Sample is a segment of the universe typically drawn at random (more or less) from the population.
- A potential bias in estimation / analysis from the sample is **random sampling variability** (chance).



Over repeated sampling, the mean of the mean of the hypothetical samples equals the population mean

Sampling distribution *of diastolic BP*

Sample(n = 5)					Mean
63	73	61	73	66	67
90	92	56	68	60	73
69	67	85	72	73	73
90	80	55	73	77	75
69	74	91	73	66	75
64	71	80	76	94	77
62	91	79	89	62	77
81	76	91	77	63	78
95	91	73	65	72	79
72	82	73	90	81	80
83	75	69	92	79	80
89	90	77	80	82	84

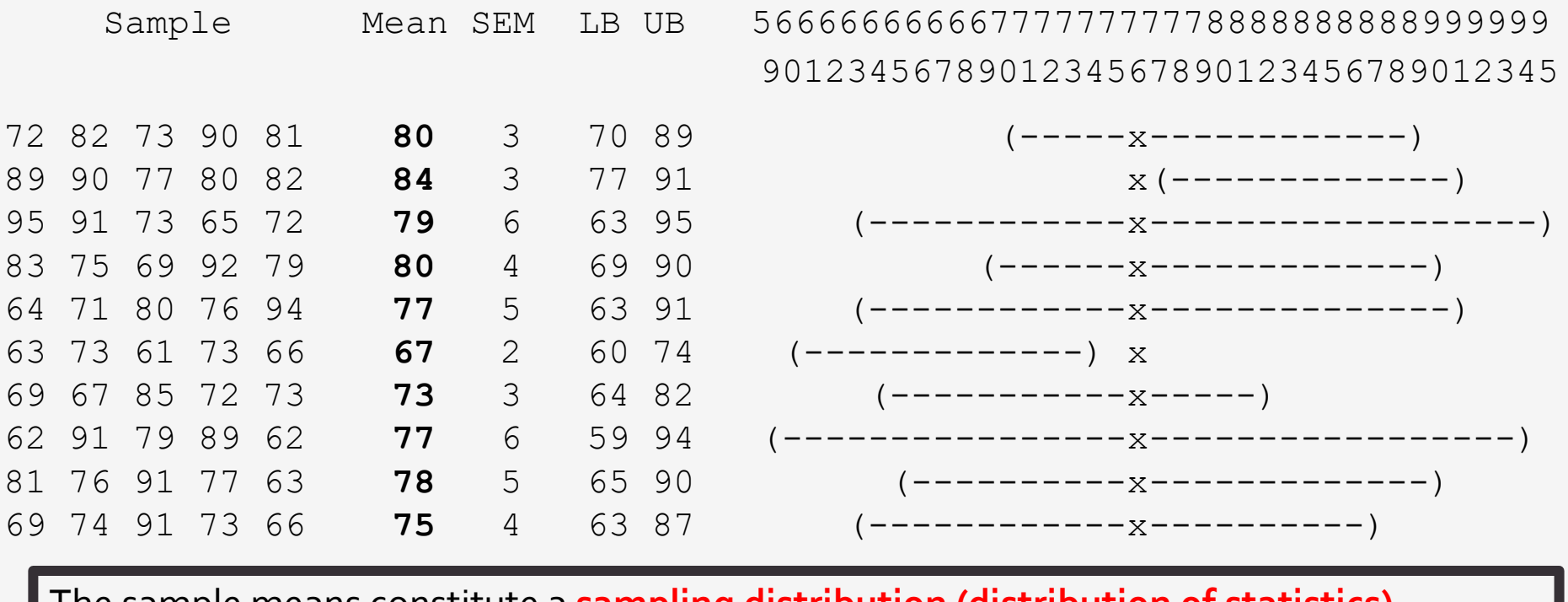
Is the sample representative of the population?

The larger the sample, the more accurate the estimation of the population parameter.

Variability

- **Standard deviation:** variability of data about its mean.
- **Standard error:** variability of a statistic (e.g. means:
- SEM = SD / √n).

95% confidence intervals for dBP from hypothetical samples (n = 5)
Population mean dBP = 76



The sample means constitute a **sampling distribution (distribution of statistics)**.
The standard error of the mean (SEM) quantifies **variability of the sampling distribution**.
95% CI is equal to the mean ± 2 x SEM (2 is standard deviation units of the normal distribution).
The mean of the sampling distribution will approximate the population mean.
The larger the sample size, the smaller the standard error.

Confidence interval

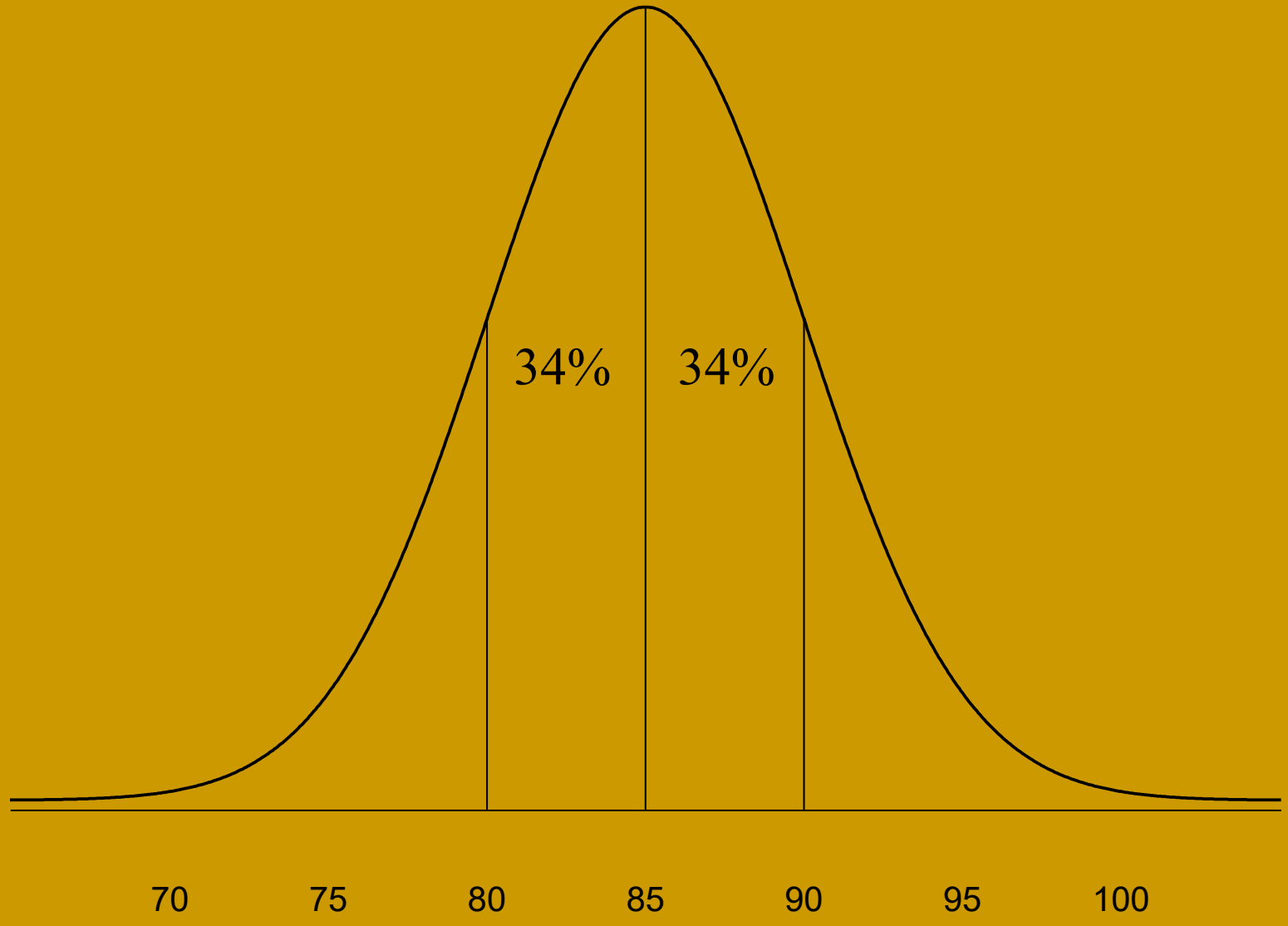
Over repeated sampling, a 95% CI will include the population mean 95 times out of every 100 random samples.

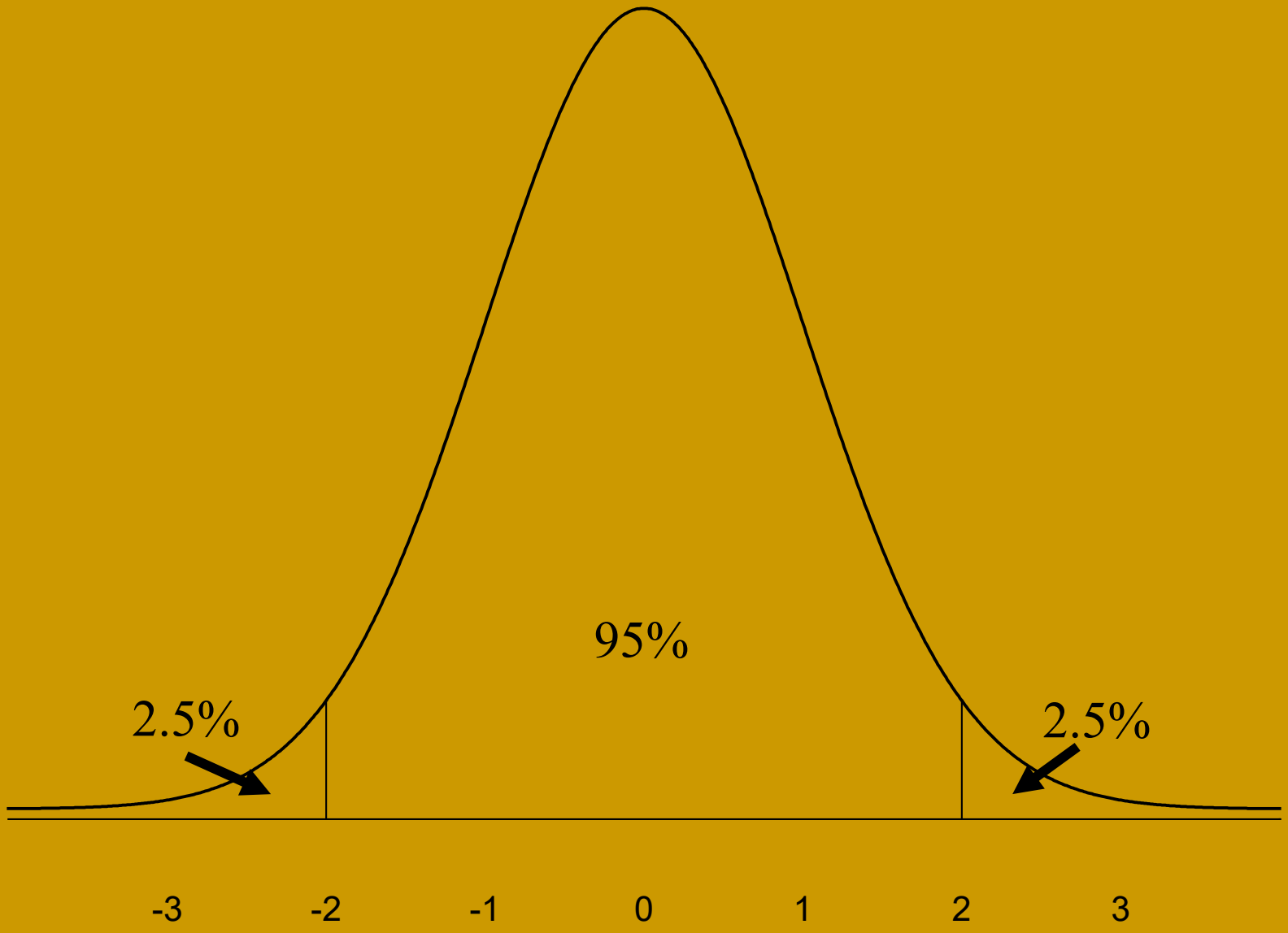
Theoretical sampling distributions

- Use theoretical distribution to model empirical distribution
- Use theoretical distribution to model distribution of sample statistics
- Theoretical distribution has a mathematical formula
- Easily calculate probabilities of random variables for given distribution

NORMAL DISTRIBUTION

A normal distribution is defined by two parameters, its mean and standard deviation. It is a symmetrical distribution in which the measures of central tendency, the mean, median and mode are equal.


34%
34%
70
75
80
85
90
95
100


95%
2.5%
2.5%
-3
-2
-1
0
1
2
3

Other theoretical distributions

- χ^2 - contingency table analysis & gof
- t - compare means & correlations
- F - analysis of variance (ANOVA)

The above distributions are sample size dependent and as sample size $\rightarrow \infty$, tend to a normal distribution

Statistical significance and p values

- All statistical analyses involve two sources of variance of the outcome variable in the data.
- **Known variance:** treatment group, age, race, etc.
- **Unknown variance:** error variance
- Statistical tests:
- T = known variance / unknown variance
- **p value is the probability of obtaining a value of T or larger, just by random sampling variability (chance)**

Hypothesis Testing

Treatment: 80 88 76 77 84 Mean = 81

Control: 84 90 86 77 93 Mean = 86

Null hypothesis: There is no difference between the means of the Treatment and Control groups (i.e., chance)

Alternative hypothesis: There is a difference between means of the Treatment and Control groups (difference unlikely to have occurred just by chance)

STATISTICAL DECISIONS

ASSUME NULL HYPOTHESIS TRUE

Statistical tests give probability of obtaining the value of the test statistic, or larger (t, F, etc.) by chance and, by inference, whether differences "statistically significant".

Conclusions from statistical test	True situation: Difference exists	True situation: No difference
Difference exists	Correct Decison Probability = 1 - Beta (Power)	Type I Error Probability = Alpha
No difference	Type II Error Probability = Beta	Correct Decision Probability = 1 - Alpha

Treatment: 80 88 76 77 84 Mean = 81 ± 5
Control: 84 90 86 77 93 Mean = 86 ± 6

t = difference in means divided by the standard error of the difference

Which t distribution? (depends on sample size)

t distribution defined by degrees of freedom (df)

df = number of components of a statistic that are free to vary

X1 + X2 + X3 = 100

50 + X2 + X3 = 100

50 + 40 + X3 = 100

Use 5 values to compute a mean, to compute variance, only 4 values needed:

$$\frac{(80-81)^2+(88-81)^2+(76-81)^2+(77-81)^2+(84-81)^2}{5-1}$$

first 4 values can be any number (free to vary); the mean is known, so once the first 4 numbers are known, there is only one possible value for the 5th number

So df = 5 -1 = 4

t-test for difference in two independent means

Difference in the two means divided by the standard error of the difference of two means

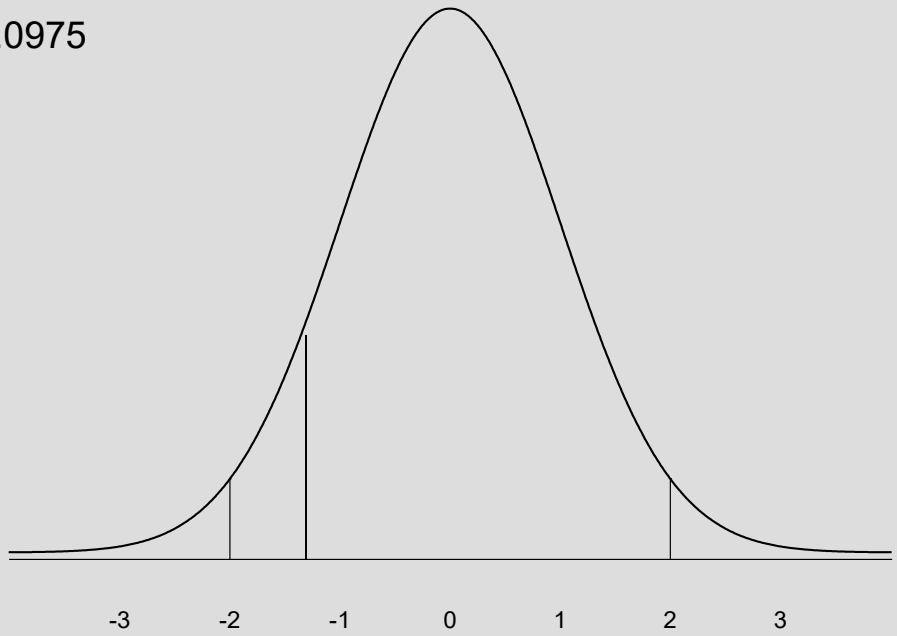
with df = $n_1 - 1 + n_2 - 1$

$$t = \frac{81 - 86}{\sqrt{\frac{5^2 (5 - 1) + 6^2 (5 - 1)}{5 - 1 + 5 - 1} \times (1/5 + 1/5)}}$$

$$= -1.4142$$

t(8) = -1.4142

p = 0.0975



p value

- one-sided = 0.0975
- two-sided = 0.0975 x 2 = 0.1950
- Fail to reject the null hypothesis at the 0.05 level

Conclusions?

- There is no difference between treatment and control, the treatment does not lower heart rate.
- NO!
- There is no evidence for a treatment effect.

Categorical outcome (2 categories)

- Positive-negative; disease-no disease dead-alive
- Treatment A: 84 of 184 patients had a positive outcome
- Treatment B: 72 of 202 patients had a positive outcome
- Proportion of positive outcomes different for A and B?

Proportion (A) = 84/184 = 0.457

standard error = $\sqrt{(0.457)*(1-0.457)/184}$

= 0.037

Proportion (B) = 72/202 = 0.356

Standard error = $\sqrt{(0.356)*(1-0.356)/202}$

= 0.034

Difference in independent proportions

z = difference in proportions divided by standard error of difference

Standard error = $\sqrt{(p*(1-p)*(1/n_1 + 1/n_2)}$ where p is overall proportion

Δ = 0.457(84/184) - 0.356 (72/202) = 0.1001

p = 156/386 = 0.4041

1 – p = 0.5959

1/184 + 1/202 = 0.0104

z = 0.1001/ $\sqrt{}$ 0.4041*0.5959*0.0104

= 2.0014, p = 0.0454 (two-sided)

Reject null hypothesis that 84/184 - 72/202 is due to sampling variability
Conclusion: Treatment A resulted in a statistically significantly greater proportion of positive outcomes than treatment B (0.457 vs. 0.356).

Chi-square

- Create row by column (r x c) table of treatment groups and outcomes
- Calculate squared difference between observed and expected frequencies, divide by expected cell frequency; sum over all cells
- Expected = (row total * column total)/grand total
- Distributed as chi-square with df = (r-1) * (c-1)

	+	-	
A	84	100	184
B	72	130	202
	156	230	386

	+	-	
A	84	100	184
B	72	130	202
	156	230	386

E1 = 184*156/386; E2 = 184*230/386; E3 = 202*156/386; E4 = 202*230/386

$$\chi^2 = \frac{(84-74.4)^2}{74.4} + \frac{(100-109.6)^2}{109.6} + \frac{(72-81.6)^2}{81.6} + \frac{(130-120.4)^2}{120.4}$$

$$= \chi^2\,(1) = 4.005,\ p = 0.0454$$

If study is prospective, calculate relative risk (RR) from 2 x 2 table:

	Disease	No Disease	
Risk factor present	84	100	184
Risk factor absent	72	130	202

RR = (84/184) / (72/202)

= 1.281

If study is retrospective, calculate odds ratio (OR) from 2 x 2 table:

	Disease	No Disease
Risk factor present	84	100
Risk factor absent	72	130
	156	230

Odds that patients with disease had risk factor = (84/156) / (72/156)
Odds that patients w/o disease had risk factor = (100/230) / (130/230)

OR = [(84/156) / (72/156)] / [(100/230) / (130/230)]
= (84*130) / (72*100)
= 1.517

RR and OR

- Prospective study estimates true incidence rate
- Retrospective study does not – samples from population of disease/no disease
- When incidence rate is small, RR and OR equal

What a p value is NOT

- The probability that the alternative hypothesis is "true".
- Probability that the null hypothesis is "false".
- Indication of importance, meaningfulness, relevance, etc. statistical significance vs. clinical significance.
- No such thing as "Highly significant" (p = 0.0001). It is or it isn't.
- No such thing as a "trend towards significance" (p = 0.06).

What a p value is

- Probability of obtaining a test statistic as large or larger than that calculated assuming the null hypothesis is true.
- Underlying assumption of repeated sampling.
- Dependent on sample size.
- **Arbitrary**

Basic concepts

Statistical Methods

- Parametric – assume some theoretical distribution for data / test statistic
- Non-parametric – no assumptions about distribution

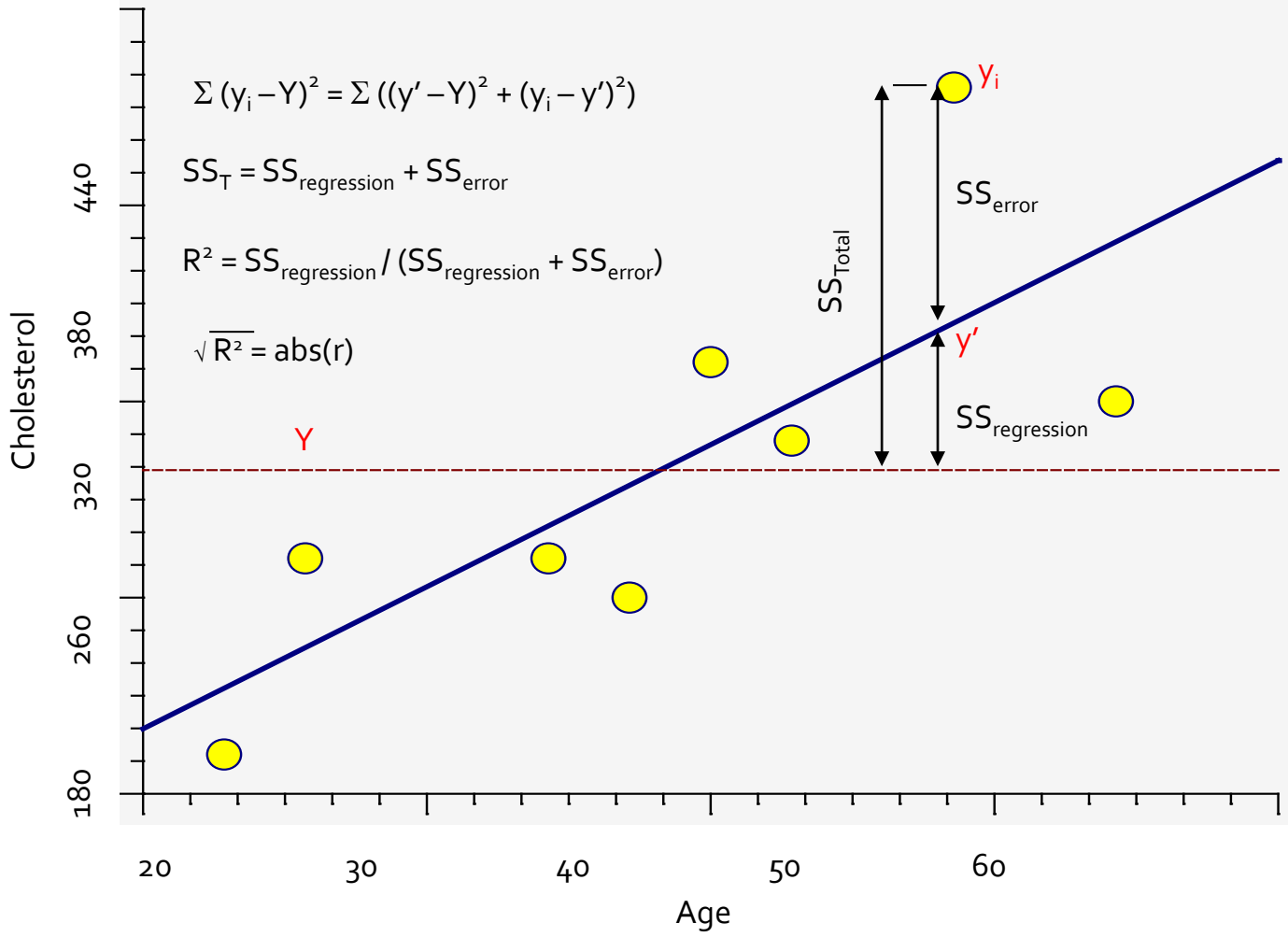
Correlation

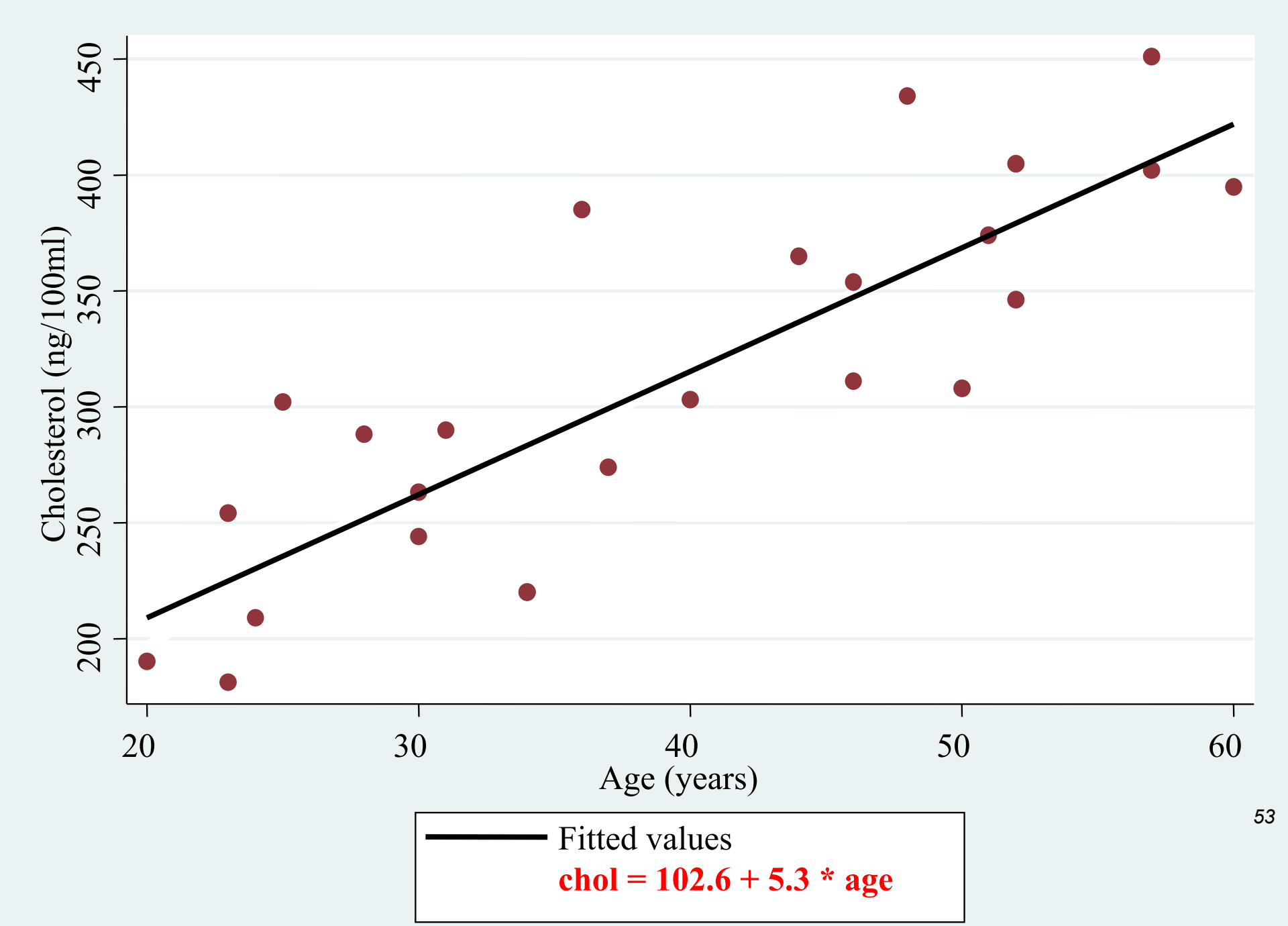
- Virtually all medical research is concerned with correlating variables – in some form or another.
- At the simplest level, correlation is an assessment of the linear rank ordering of two variables – how the values of one change in relationship to the other.
- Quantified by the correlation coefficient:

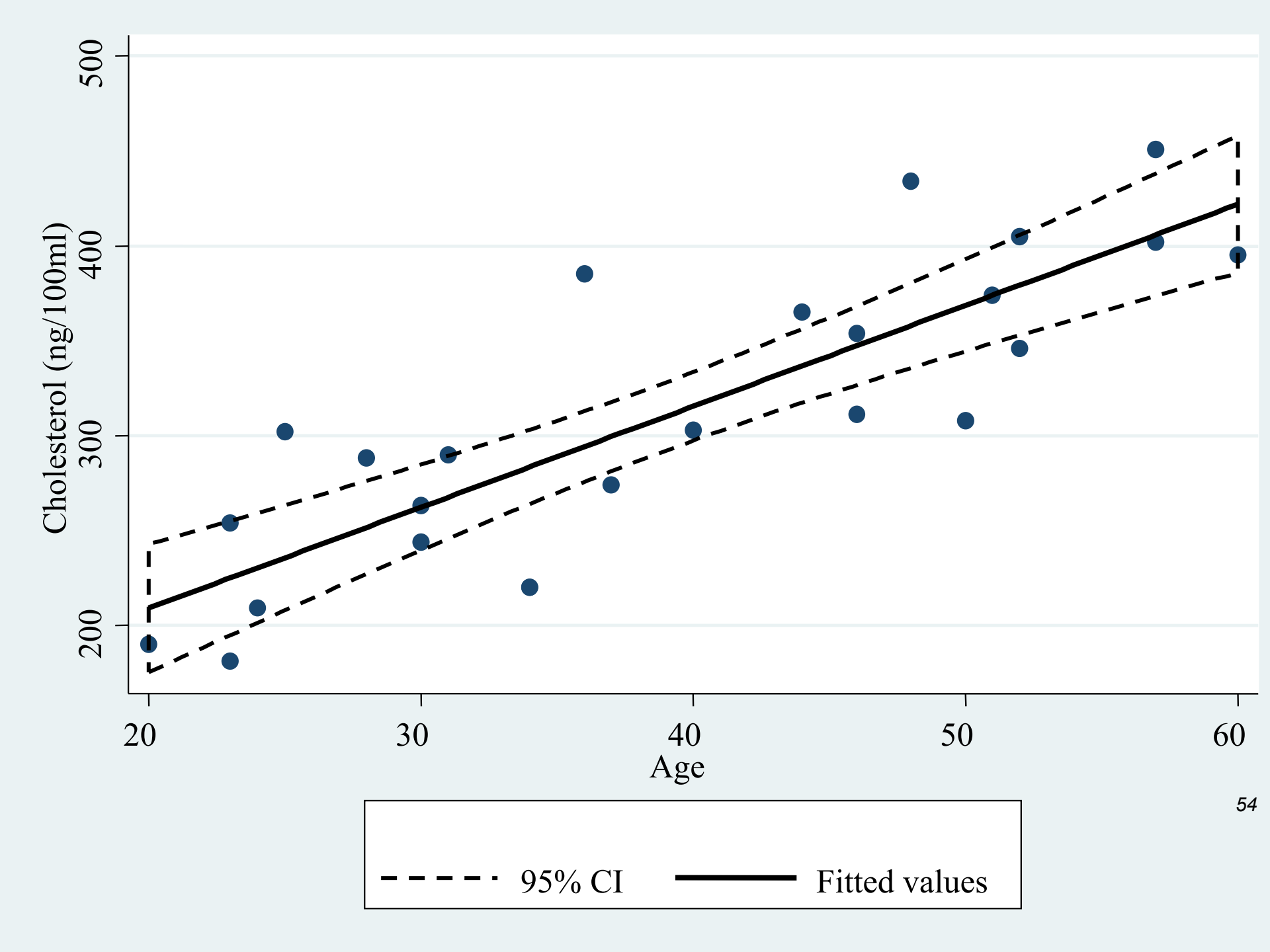
  Perfect inverse relationship (-1)
  No relationship (0)
  Perfect positive relationship (+1)

PT	Cholesterol (mg/100 ml)	Weight (kg)	Age (years)
1	354	84	46
2	190	73	20
3	405	65	52
4	263	70	30
5	451	76	57
6	302	69	25
7	288	63	28
8	385	72	36
9	402	79	57
10	365	75	44
11	209	27	24
12	290	89	31
13	346	65	52
14	254	57	23
15	395	59	60
16	434	69	48
17	220	60	34
18	374	79	51
19	308	75	50
20	220	82	34
21	311	59	46
22	181	67	23
23	274	85	37
24	303	55	40
25	244	63	30

$$r_{chol|age} = 0.84$$


$\Sigma (y_i - Y)^2 = \Sigma ((y' - Y)^2 + (y_i - y')^2)$
$SS_T = SS_{regression} + SS_{error}$
$R^2 = SS_{regression} / (SS_{regression} + SS_{error})$
$\sqrt{R^2} = abs(r)$
y_i
SS_{error}
SS_{Total}
y'
$SS_{regression}$
Y
Cholesterol
180
260
320
380
440
20
30
40
50
60
Age


Cholesterol (ng/100ml)
200
250
300
350
400
450
20
30
40
50
60
Age (years)
Fitted values
chol = 102.6 + 5.3 * age


Cholesterol (ng/100ml)
500
400
300
200
20
30
40
50
60
Age
95% CI
Fitted values

Break

General Linear Model (GLM)

dependent variable ~ independent variable(s)

$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \ldots$ where Y is a continuous outcome (dependent variable)
X's are the covariates (independents)
b's are the estimated regression coefficients.

b is the change in Y for every unit increase (decrease if b is negative) in X adjusted for all other X's in the model.

X's can be a mix of continuous, ordinal or categorical variables.

The "statistical test" is whether the regression coefficients are different from zero.

GLM special cases

$Y = b_0 + b_1X_1$ is a t-test if X is dichotomous (1/0).

$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3$ is analysis of variance (ANOVA) if X's are categorical.

If some X's continuous – analysis of covariance (ANCOVA)

Generalized Linear Model

Dichotomous outcome: success-failure; dead-alive; positive-negative
Logistic regression (logistic distribution – S curve)

Model the probability of one of the dichotomies:

$P = EXP(Y) / (1 + EXP(Y))$ where

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \ldots$$

$exp(b_i)$ = odds ratio (OR) adjusted for all other X's in the model if design is retrospective.

Risk ratio (RR) if design is prospective.

Test is whether the odds ratios are different from 1.
The 95% CI will include 1 if the odds of the event are statistically the same for values or categories of a given covariate.

Special case of logistic regression

$Y = b_0 + b_1X_1$

If X is also dichotomous, 2 x 2 contingency table.

	Outcome +	Outcome -
Treatment A	a	b
Treatment B	c	d

OR = ad/bc

Generalized Linear Model (continued)

Ordered logistic regression: **Y is ordinal**
cancer stage: I, II, III, IV

Multinomial logistic regression: **Y is nominal**
Delivery: vaginal, vaginal-assisted, cesarean

Poisson regression: **Y is a count within/over some time period.**
Number of ICU infections from 1/1/2021 – 3/31/2021
Exponentiated regression coefficients are incident rate ratios (IRR).

Time-to-event models

Event: dichotomous (death, failure, recurrence)
Time to event: from a point of origin to a terminal event (e.g., date of diagnosis of disease to date of death from disease).
The terminal event occurs after the starting event.

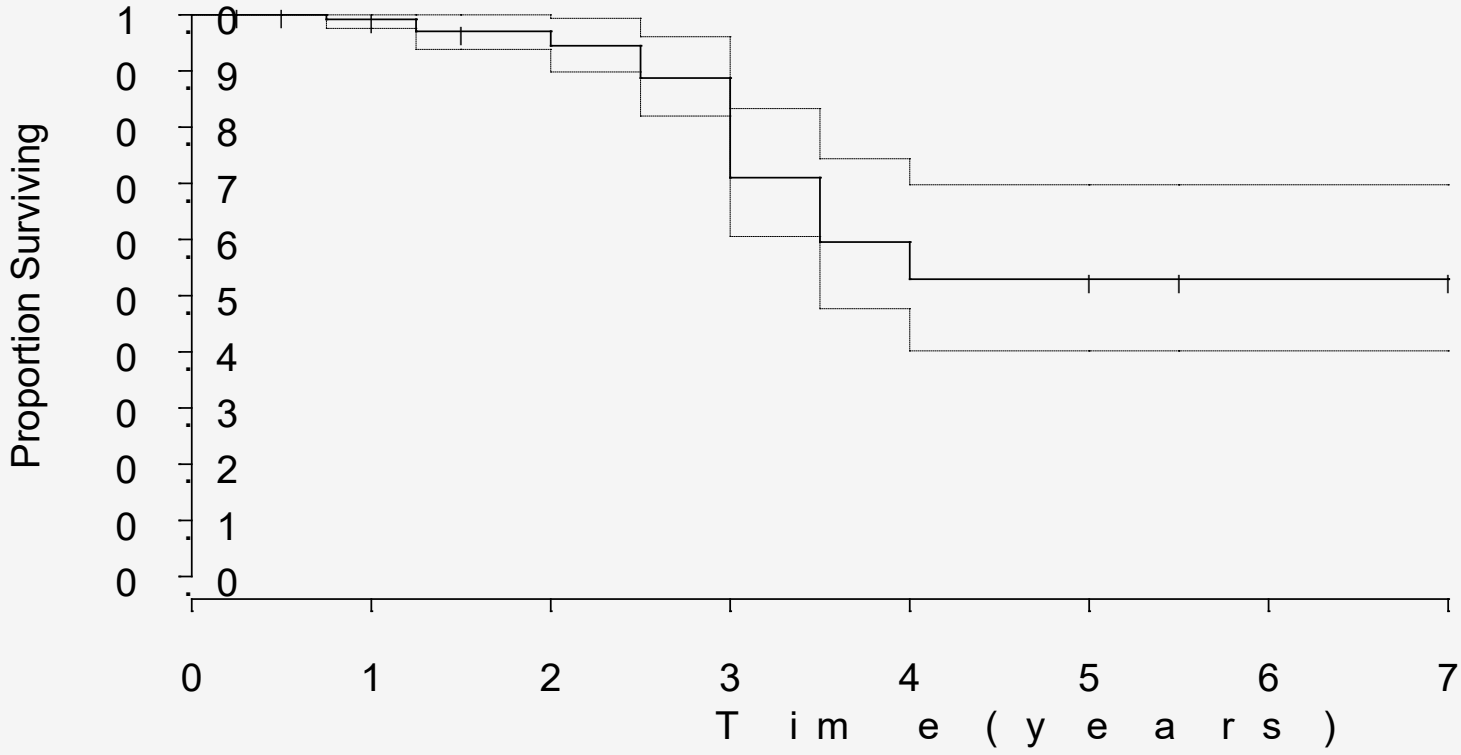
Censoring - exact survival times not known (e.g., lost to follow-up).
Survival function, S(t) - the probability that a patient survives from the time of origin to some point beyond t.
Hazard rate - the probability that a patient dies at time t conditional on having survived to time t (i.e., the instantaneous death rate for a patient that survived to time t.

Hazard function: $-H(t) = -\log(S(t))$
Survival function: $S(t) = \exp(-H(t))$

Median survival time - time at which the survival rate is 50%.

Kaplan-Meier survival functions

Survival Time(mos)	Patients	Deaths	Censored	Proportion surviving	Cumulative Survival
0	74	0	0	1.0000	1.0000
7	74	1	0	.9865	.9865
11	73	1	0	.9863	.9730
12	72	1	0	.9861	.9595
18	71	3	0	.9577	.9190
20	68	5	0	.9265	.8514
25	63	0	2	1.0000	.8514
28	61	4	1	.9344	.7956
36	56				



Kaplan-Meier survival curve with 95% CI

Kaplan-Meiers

Multivariable survival models

Cox Proportional Hazards Regression

Model - given a pattern of covariate values **z**, the survival function is:

$$S(t,z) = [S0(t)]\exp(\boldsymbol{\beta z})$$

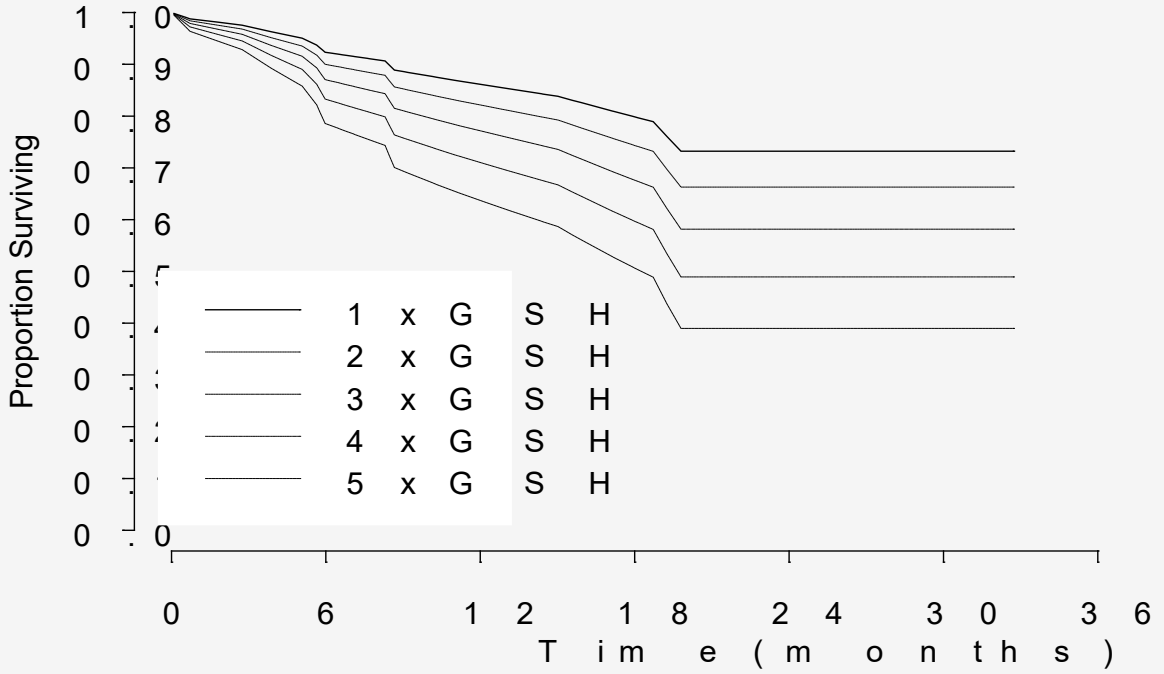
where S0(t) is the baseline survival function corresponding to **z** = 0, i.e., no effect of covariates.

β is a vector of regression coefficients, one for each covariate.

$$\mathbf{(Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \ldots)}$$

In the following example, survival of colo-rectal cancer patients is modeled as a function of tumor GSH level to normal GSH level.

Model: $S(t) = [S0(t)]\exp(\boldsymbol{\beta 1(GSH)})$



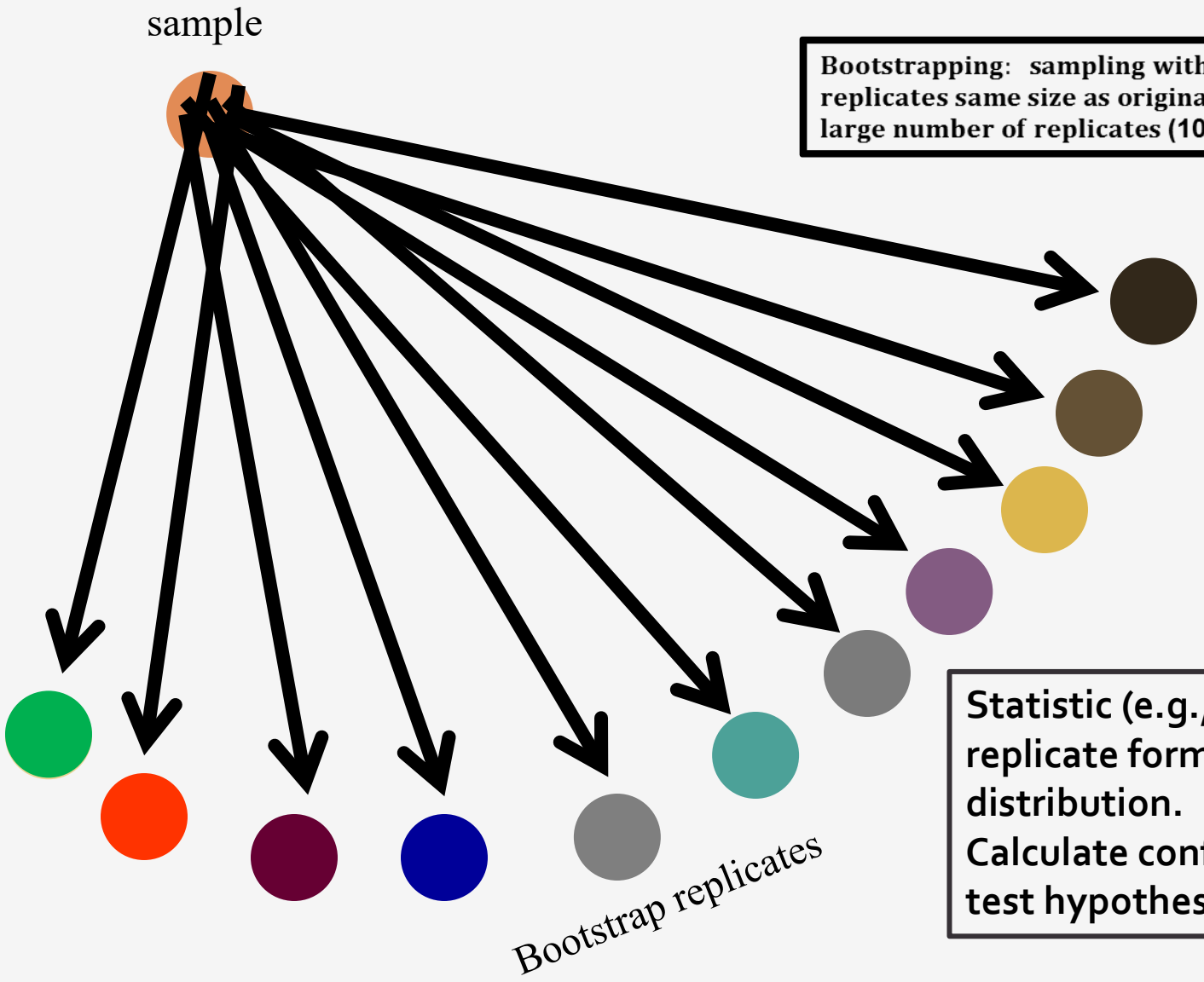
Cox regression of GSH on survival for colo-rectal cancer patients

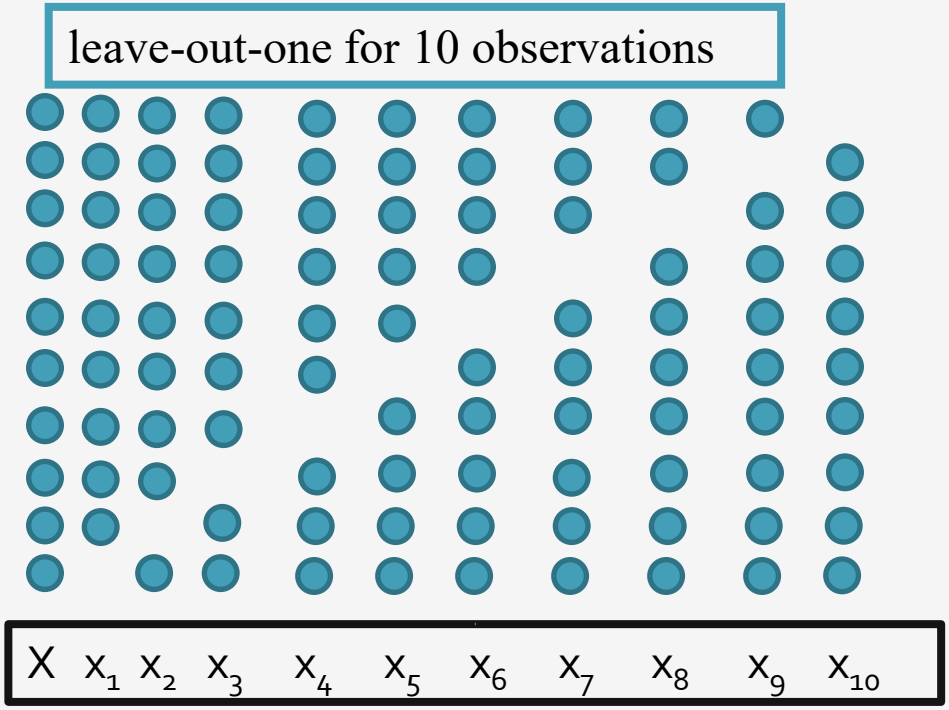
Nonparametric statistics

- Distribution-free
- No assumption is made about the distribution of the outcome variable.
- Mann-Whitney U test (t-test)
- Wilcoxon matched-pairs signed-ranks test
- Kruskal-Wallis (one-way analysis of variance)
- Spearman / Kendall correlation
- McNemar test for correlated proportions

Nonparametric statistics (continued)

- Bootstrap – randomly select an observation from sample data with replacement.
- Jackknife – leave-out-one


sample
Bootstrapping: sampling with replacement
replicates same size as original sample
large number of replicates (100-40,000)
Bootstrap replicates
Statistic (e.g., mean) of each replicate form a sampling distribution.
Calculate confidence intervals, test hypothesis



From the distribution of the x_i's, bounds of a 95% CI can be found at the lower 2.5% and the upper 2.5% of the distribution regardless of whether the distribution is symmetrical.

Considerations for statistical analysis

- Effect size
- Paired vs. unpaired data
- Clustered data
- Selection bias in non-randomized comparative effectiveness studies
- Consulting a statistician

Effect size

- How much of a difference between interventions, groups, etc. in the outcome variable is "clinically meaningful"?
- Difference in means, survival rates, proportions (size of an odds ratio if logistic regression) to detect
- Variability of the outcome: standard deviation

Data for estimating effect size

- Pilot studies
- Published studies
- Best case / worst case scenarios

Paired vs. unpaired data

- Unpaired – data from independent entities (e.g., patients in treatment groups).
- Paired – data from same patient or matched patients (e.g., repeated measures).
- Paired analysis requires fewer patients and has more statistical power.
- Within-patient variability.
- If independent groups included in design – between-group variability and within-patient variability.

The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 AUGUST 14, 2008 VOL. 359 NO. 7

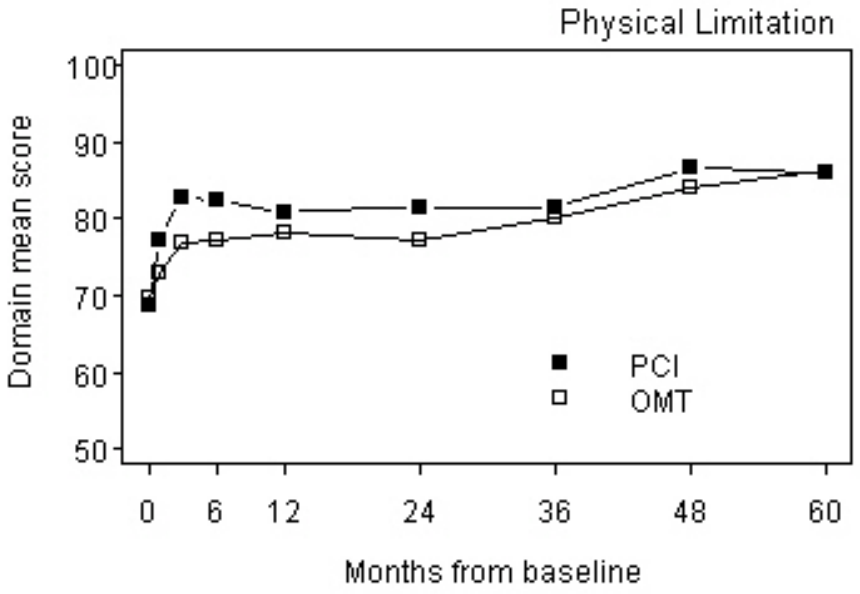
Effect of PCI on Quality of Life in Patients with Stable Coronary Disease

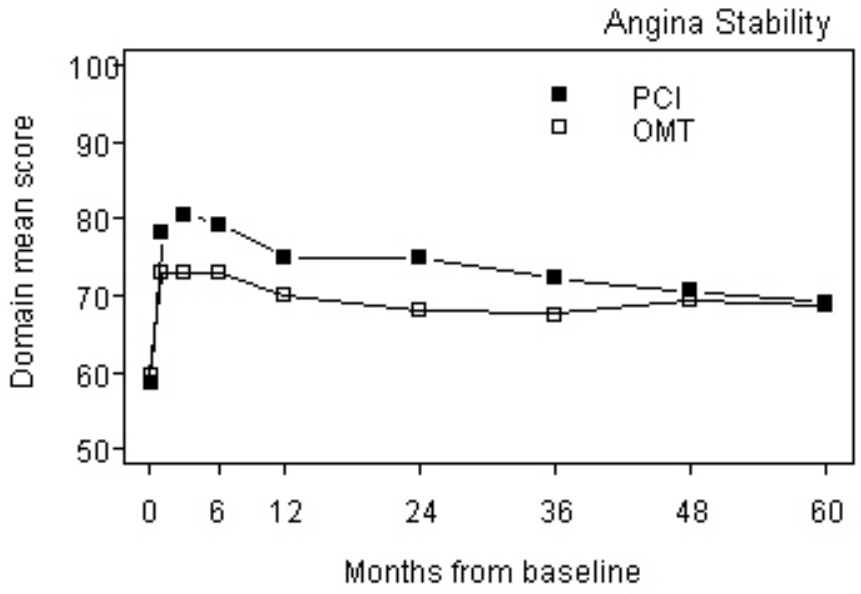
William S. Weintraub, M.D., John A. Spertus, M.D., M.P.H., Paul Kolm, Ph.D., David J. Maron, M.D., Zefeng Zhang, M.D., Ph.D., Claudine Jurkovitz, M.D., M.P.H., Wei Zhang, M.S., Pamela M. Hartigan, Ph.D., Cheryl Lewis, R.N., Emir Veledar, Ph.D., Jim Bowen, B.S., Sandra B. Dunbar, D.S.N., Christi Deaton, Ph.D., tanley Kaufman, M.D., Robert A. O'Rourke, M.D., Ron Goeree, M.S., Paul G. Barnett, Ph.D., Koon K. Teo, M.D., and William E. Boden, M.D., for the COURAGE Trial Research Group*

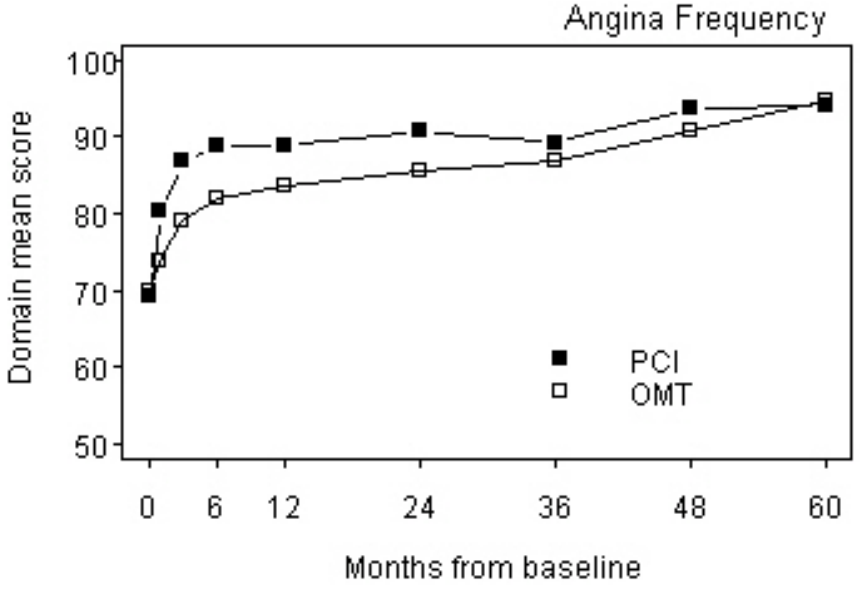
ABSTRACT

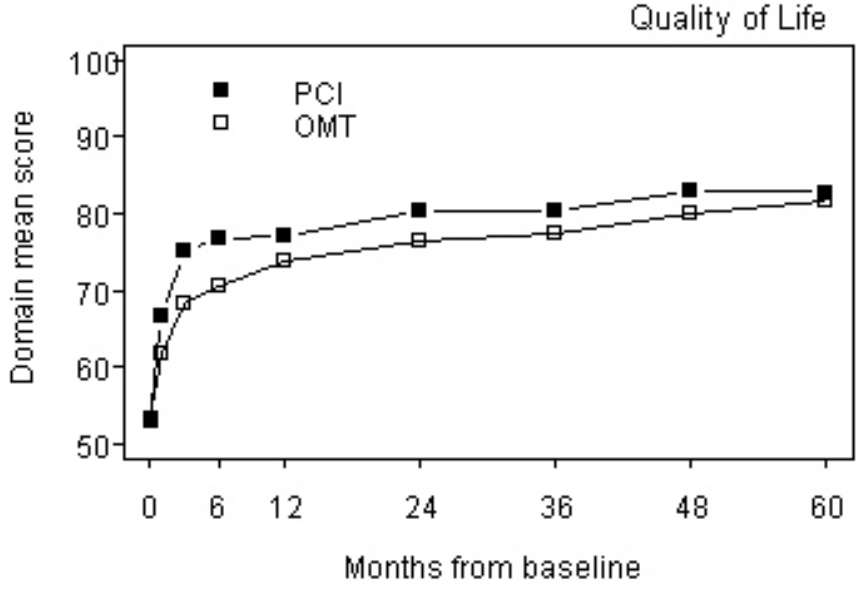
significant difference in the rate of the primary end point (death or myocardial infarction) during a median follow-up period of 4.6 years.[6]

Among patients with stable coronary artery disease, PCI is indicated for the relief of angina.[2,3,7,8] Therefore, as part of the COURAGE trial, we used the Seattle Angina Questionnaire to assess the effect of therapy on the relief of angina. In addition, we used the RAND 36-item health survey (RAND-36) to evaluate the effect of therapy on broader health status.


Physical Limitation
Domain mean score
100
90
80
70
60
50
0
6
12
24
36
48
60
Months from baseline
PCI
OMT


Angina Stability
Domain mean score
100
90
80
70
60
50
0
6
12
24
36
48
60
Months from baseline
PCI
OMT


Angina Frequency
Domain mean score
100
90
80
70
60
50
0
6
12
24
36
48
60
Months from baseline
PCI
OMT

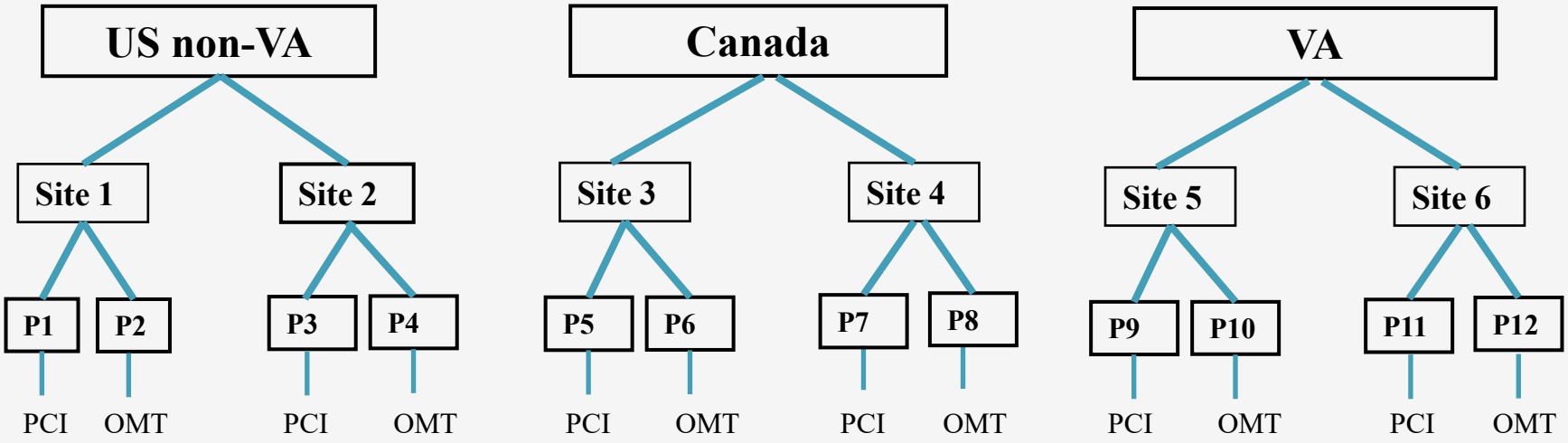

Quality of Life
Domain mean score
100
90
80
70
60
50
0
6
12
24
36
48
60
Months from baseline
PCI
OMT

Clustered or Multilevel

- Hierarchical
- Clustered
- Levels considered random effects
- Variance component – variance of a random effect

- **Fixed effects** – effects attributable to a finite set of levels of a factor (variable) that are of specific interest, e.g., treatment, sex, race, age.
- **Random effects** – effects attributable to an infinite (or at least large) set of levels of a factor of which only a random (more or less) sample appear in the data, e.g., clinical trial sites, geographic regions, patients/subjects are always a random effect.

COURAGE TRIAL



Random effects issues

- Data from a given cluster may be correlated, i.e., not independent.
- Correlation may affect model parameters and standard errors if not accounted for.
- Estimates / conclusions may be misleading.

Comparative Effectiveness Analysis in Non-randomized Studies

- ***Pros***

  _more data

  _data may already exist (EMR)

  _less expensive

  _realistic clinical practice environment

- ***Cons***

  _potential for selection bias and confounding

Solutions to reducing selection bias

- Case-control matching (age, race, sex, etc.)
- Covariate adjustment
- Stratification
- Propensity score analysis
- Inverse probability weighting (IPW)
- Instrumental variables

Thank you!!

paul.kolm@medstar.net