Linear Regression and Correlation

- Explanatory and Response Variables are Numeric
- Relationship between the mean of the response variable and the level of the explanatory variable assumed to be approximately linear (straight line)
- Model:

$$Y = \beta_0 + \beta_1 x + \varepsilon \qquad \varepsilon \sim N(0, \sigma)$$

  - $\beta_1 > 0 \Rightarrow$ Positive Association
  - $\beta_1 < 0 \Rightarrow$ Negative Association
  - $\beta_1 = 0 \Rightarrow$ No Association

Least Squares Estimation of β_0, β_1

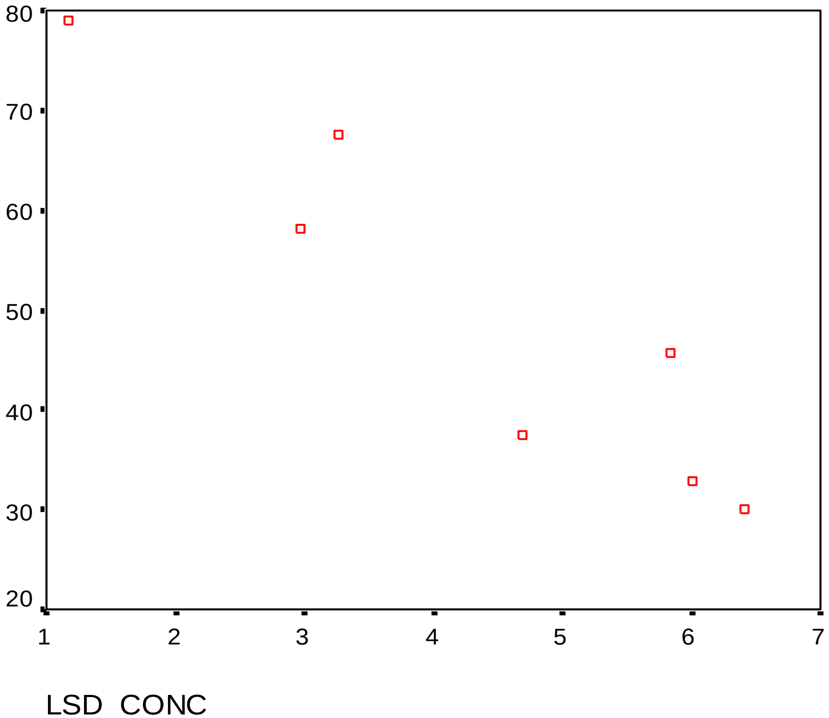
- $\beta_0 \equiv$ Mean response when x=0 (y-intercept)
- $\beta_1 \equiv$ Change in mean response when x increases by 1 unit (slope)
- β_0, β_1 are unknown parameters (like μ)
- $\beta_0+\beta_1 x \equiv$ Mean response when explanatory variable takes on the value x
- Goal: Choose values (estimates) that minimize the sum of squared errors (*SSE*) of observed values to the straight-line:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \qquad SSE = \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 = \sum_{i=1}^{n}\left(y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)\right)^2$$

Example - Pharmacodynamics of LSD

- Response (*y*) - Math score (mean among 5 volunteers)
- Predictor (*x*) - LSD tissue concentration (mean of 5 volunteers)
- Raw Data and scatterplot of Score vs LSD concentration:

Score (y)	LSD Conc (x)
78.93	1.17
58.20	2.97
67.47	3.26
37.47	4.69
45.65	5.83
32.92	6.00
29.97	6.41



Source: Wagner, et al (1968)

Least Squares Computations

$$S_{xx} = \sum \left(x - \overline{x}\right)^2$$

$$S_{xy} = \sum \left(x - \overline{x}\right)\left(y - \overline{y}\right)$$

$$S_{yy} = \sum \left(y - \overline{y}\right)^2$$

$$\hat{\beta}_1 = \frac{\sum \left(x - \overline{x}\right)\left(y - \overline{y}\right)}{\sum \left(x - \overline{x}\right)^2} = \frac{S_{xy}}{S_{xx}}$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

$$s^2 = \frac{\sum \left(y - \hat{y}\right)^2}{n-2} = \frac{SSE}{n-2}$$

Example - Pharmacodynamics of LSD

Score (y)	LSD Conc (x)	x-xbar	y-ybar	Sxx	Sxy	Syy
78.93	1.17	-3.163	28.843	10.004569	-91.230409	831.918649
58.20	2.97	-1.363	8.113	1.857769	-11.058019	65.820769
67.47	3.26	-1.073	17.383	1.151329	-18.651959	302.168689
37.47	4.69	0.357	-12.617	0.127449	-4.504269	159.188689
45.65	5.83	1.497	-4.437	2.241009	-6.642189	19.686969
32.92	6.00	1.667	-17.167	2.778889	-28.617389	294.705889
29.97	6.41	2.077	-20.117	4.313929	-41.783009	404.693689
350.61	**30.33**	**-0.001**	**0.001**	**22.474943**	**-202.487243**	**2078.183343**

(Column totals given in bottom row of table)

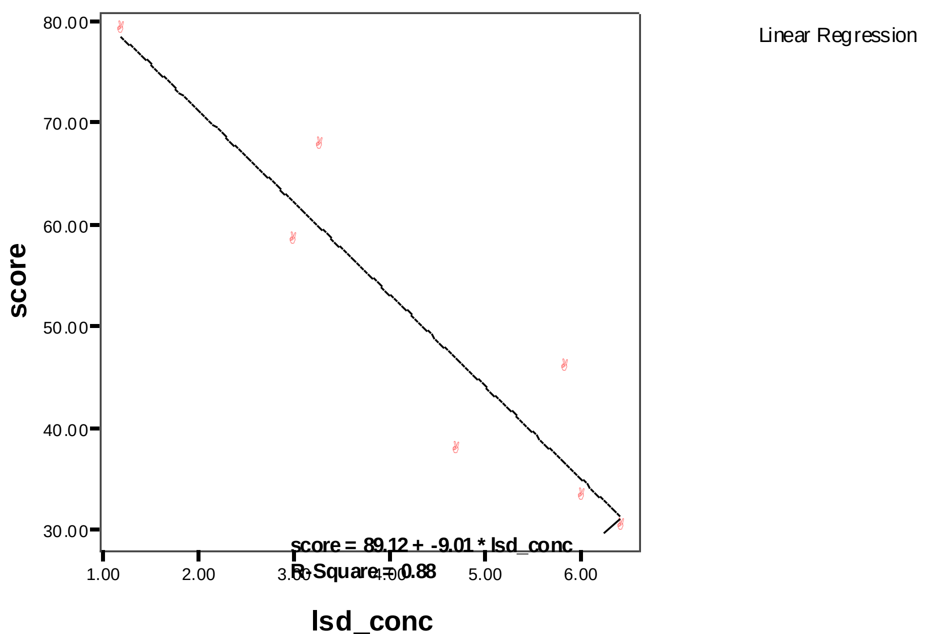
$$\bar{y} = \frac{350.61}{7} = 50.087 \qquad \bar{x} = \frac{30.33}{7} = 4.333$$

$$\hat{\beta}_1 = \frac{-202.4872}{22.4749} = -9.01 \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 50.09 - (-9.01)(4.33) = 89.10$$

$$\hat{y} = 89.10 - 9.01x \qquad s^2 = 50.72$$

SPSS Output and Plot of Equation

Math Score vs LSD Concentration (SPSS)



Inference Concerning the Slope (β_1)

- Parameter: Slope in the population model (β_1)
- Estimator: Least squares estimate: $\hat{\beta}_1$
- Estimated standard error: $\hat{\sigma}_{\hat{\beta}_1} = s / \sqrt{S_{xx}}$

- Methods of making inference regarding population:
  - Hypothesis tests (2-sided or 1-sided)
  - Confidence Intervals

Hypothesis Test for β_1

- 2-Sided Test
  - H_0: $\beta_1 = 0$
  - H_A: $\beta_1 \neq 0$

$$T.S.: t_{obs} = \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}}$$

$$R.R.: |t_{obs}| \geq t_{\alpha/2, n-2}$$

$$P-val: 2P(t \geq |t_{obs}|)$$

- 1-sided Test
  - H_0: $\beta_1 = 0$
  - H_A^+: $\beta_1 > 0$ or
  - H_A^-: $\beta_1 < 0$

$$T.S.: t_{obs} = \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}}$$

$$R.R.^+: t_{obs} \geq t_{\alpha, n-2} \quad R.R.^-: t_{obs} \leq -t_{\alpha, n-2}$$

$$P-val^+: P(t \geq t_{obs}) \quad P-val^-: P(t \leq t_{obs})$$

(1-α)100% Confidence Interval for β_1

$$\hat{\beta}_1 \pm t_{\alpha/2}\,\hat{\sigma}_{\hat{\beta}_1} \equiv \hat{\beta}_1 \pm t_{\alpha/2}\,\frac{s}{\sqrt{S_{xx}}}$$

- Conclude positive association if entire interval above 0
- Conclude negative association if entire interval below 0
- Cannot conclude an association if interval contains 0
- Conclusion based on interval is same as 2-sided hypothesis test

Example - Pharmacodynamics of LSD

$$n = 7 \quad \hat{\beta}_1 = -9.01 \quad s = \sqrt{50.72} = 7.12 \quad S_{xx} = 22.475$$

$$\hat{\sigma}_{\hat{\beta}_1} = \frac{7.12}{\sqrt{22.475}} = 1.50$$

- Testing H_0: $\beta_1 = 0$ vs H_A: $\beta_1 \neq 0$

$$T.S.: t_{obs} = \frac{-9.01}{1.50} = -6.01 \qquad R.R.: |t_{obs}| \geq t_{.025,5} = 2.571$$

- 95% Confidence Interval for β_1 :

$$-9.01 \pm 2.571(1.50) \equiv -9.01 \pm 3.86 \equiv (-12.87, -5.15)$$

Correlation Coefficient

- Measures the strength of the linear association between two variables
- Takes on the same sign as the slope estimate from the linear regression
- Not effected by linear transformations of *y* or *x*
- Does not distinguish between dependent and independent variable (e.g. height and weight)
- Population Parameter - ρ
- Pearson's Correlation Coefficient:

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \qquad -1 \leq r \leq 1$$

Correlation Coefficient

- Values close to 1 in absolute value $\Rightarrow$ strong linear association, positive or negative from sign
- Values close to 0 imply little or no association
- If data contain outliers (are non-normal), Spearman's coefficient of correlation can be computed based on the ranks of the *x* and *y* values
- Test of $H_0{:}\rho = 0$ is equivalent to test of $H_0{:}\beta_1=0$
- Coefficient of Determination (r^2) - Proportion of variation in *y* "explained" by the regression on *x*:

$$r^2 = (r)^2 = \frac{S_{yy} - SSE}{S_{yy}} \qquad 0 \le r^2 \le 1$$

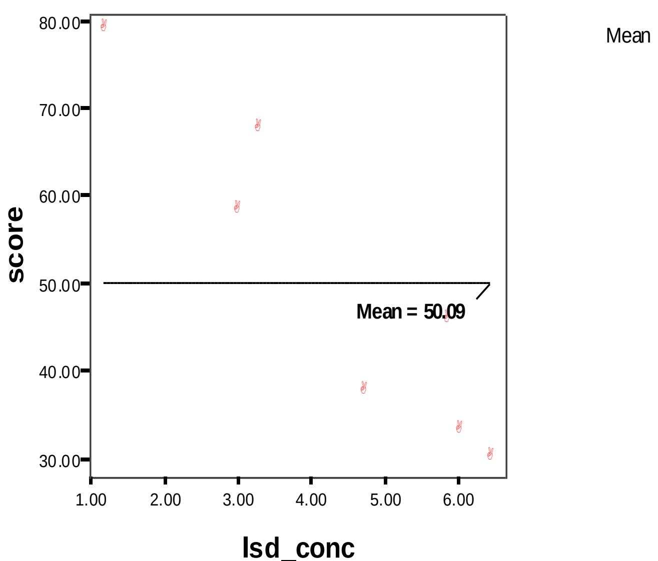
Example - Pharmacodynamics of LSD

$$S_{xx} = 22.475 \quad S_{xy} = -202.487 \quad S_{yy} = 2078.183 \quad SSE = 253.89$$

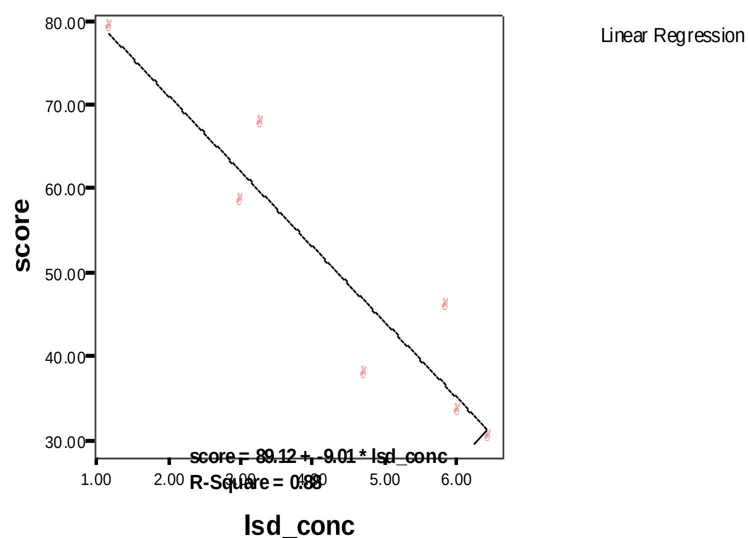
$$r = \frac{-202.487}{\sqrt{(22.475)(2078.183)}} = -0.94$$

$$r^2 = \frac{2078.183 - 253.89}{2078.183} = 0.88 = (-0.94)^2$$

S_{yy}



SSE



Example - SPSS Output
Pearson’s and Spearman’s Measures

Analysis of Variance in Regression

- Goal: Partition the total variation in *y* into variation “explained” by *x* and random variation

$$(y_i - \bar{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

$$\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2$$

- These three sums of squares and degrees of freedom are:
  - **Total** (S_{yy}) $df_{\text{Total}} = n$-1
  - **Error** (*SSE*) $df_{\text{Error}} = n$-2
  - **Model** (*SSR*) $df_{\text{Model}} = 1$

Analysis of Variance in Regression

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
Model	SSR	1	$MSR = SSR/1$	$F = MSR/MSE$
Error	SSE	n-2	$MSE = SSE/(n-2)$	
Total	S_{yy}	n-1		

- Analysis of Variance - F-test
- $H_0: \beta_1 = 0$ $H_A: \beta_1 \neq 0$

$$T.S.: F_{obs} = \frac{MSR}{MSE}$$

$$R.R.: F_{obs} \geq F_{\alpha,1,n-2}$$

$$P-val: P(F \geq F_{obs})$$

Example - Pharmacodynamics of LSD

- Total Sum of squares:

$$S_{yy} = \sum (y_i - \overline{y})^2 = 2078.183 \qquad df_{Total} = 7 - 1 = 6$$

- Error Sum of squares:

$$SSE = \sum (y_i - \hat{y}_i)^2 = 253.890 \qquad df_{Error} = 7 - 2 = 5$$

- Model Sum of Squares:

$$SSR = \sum (\hat{y}_i - \overline{y})^2 = 2078.183 - 253.890 = 1824.293 \qquad df_{Model} = 1$$

Example - Pharmacodynamics of LSD

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	*F*
Model	1824.293	1	1824.293	35.93
Error	253.890	5	50.778	
Total	2078.183	6		

- Analysis of Variance - *F*-test
- $H_0: \beta_1 = 0 \qquad H_A: \beta_1 \neq 0$

$$T.S.: F_{obs} = \frac{MSR}{MSE} = 35.93$$

$$R.R.: F_{obs} \geq F_{.05,1,5} = 6.61$$

$$P-val: P(F \geq 35.93)$$

Example - SPSS Output

Multiple Regression

- Numeric Response variable (Y)
- p Numeric predictor variables
- Model:

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$$

- Partial Regression Coefficients: $\beta_i \equiv$ effect (on the mean response) of increasing the i^{th} predictor variable by 1 unit, **holding all other predictors constant**

Example - Effect of Birth weight on Body Size in Early Adolescence

- Response: Height at Early adolescence (n =250 cases)
- Predictors (p=6 explanatory variables)
  - Adolescent Age (x_1, in years -- 11-14)
  - Tanner stage (x_2, units not given)
  - Gender (x_3=1 if male, 0 if female)
  - Gestational age (x_4, in weeks at birth)
  - Birth length (x_5, units not given)
  - Birthweight Group (x_6=1,...,6 <1500g (1), 1500-1999g(2), 2000-2499g(3), 2500-2999g(4), 3000-3499g(5), >3500g(6))

Source: Falkner, et al (2004)

Least Squares Estimation

- Population Model for mean response:

$$E(Y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

- Least Squares Fitted (predicted) equation, minimizing *SSE*:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p \qquad SSE = \sum \left(Y - \hat{Y} \right)^2$$

- All statistical software packages/spreadsheets can compute least squares estimates and their standard errors

Analysis of Variance

- Direct extension to ANOVA based on simple linear regression
- Only adjustments are to degrees of freedom:
  - $df_{\text{Model}} = p$ $df_{\text{Error}} = n\text{-}p\text{-}1$

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
Model	SSR	p	$MSR = SSR/p$	$F = MSR/MSE$
Error	SSE	$n\text{-}p\text{-}1$	$MSE = SSE/(n\text{-}p\text{-}1)$	
Total	S_{yy}	$n\text{-}1$		

$$R^2 = \frac{S_{yy} - SSE}{S_{yy}} = \frac{SSR}{S_{yy}}$$

Testing for the Overall Model - F-test

- Tests whether **any** of the explanatory variables are associated with the response
- H_0: $\beta_1 = \cdots = \beta_p = 0$ (None of the x^s associated with y)
- H_A: Not all $\beta_i = 0$

$$T.S.: F_{obs} = \frac{MSR}{MSE} = \frac{R^2 / p}{(1 - R^2)/(n - p - 1)}$$

$$R.R.: F_{obs} \geq F_{\alpha, p, n-p-1}$$

$$P-val: P(F \geq F_{obs})$$

Example - Effect of Birth weight on Body Size in Early Adolescence

- Authors did not print ANOVA, but did provide following:
  - n=250 p=6 R^2=0.26
- H_0: $\beta_1 = \cdots = \beta_6 = 0$
- H_A: Not all $\beta_i = 0$

$$T.S.: F_{obs} = \frac{MSR}{MSE} = \frac{R^2 / p}{(1-R^2)/(n-p-1)} =$$

$$= \frac{0.26/6}{(1-0.26)/(250-6-1)} = \frac{.0433}{.0030} = 14.2$$

$$R.R.: F_{obs} \geq F_{\alpha,6,243} = 2.13$$

$$P-val: P(F \geq 14.2)$$

Testing Individual Partial Coefficients - *t*-tests

- Wish to determine whether the response is associated with a single explanatory variable, after controlling for the others

- $H_0: \beta_i = 0$ $H_A: \beta_i \neq 0$ (2-sided alternative)

$$T.S.: t_{obs} = \frac{\hat{\beta}_i}{\hat{\sigma}_{\hat{\beta}_i}}$$

$$R.R.: |t_{obs}| \geq t_{\alpha/2, n-p-1}$$

$$P-val: 2P(t \geq |t_{obs}|)$$

Example - Effect of Birth weight on Body Size in Early Adolescence

Variable	b	s_b	$t=b/s_b$	P-val (z)
Adolescent Age	2.86	0.99	2.89	.0038
Tanner Stage	3.41	0.89	3.83	<.001
Male	0.08	1.26	0.06	.9522
Gestational Age	-0.11	0.21	-0.52	.6030
Birth Length	0.44	0.19	2.32	.0204
Birth Wt Grp	-0.78	0.64	-1.22	.2224

Controlling for all other predictors, adolescent age, Tanner stage, and Birth length are associated with adolescent height measurement

Models with Dummy Variables

- Some models have both numeric and categorical explanatory variables (Recall **gender** in example)
- If a categorical variable has *k* levels, need to create *k*-1 dummy variables that take on the values 1 if the level of interest is present, 0 otherwise.
- The baseline level of the categorical variable for which all *k*-1 dummy variables are set to 0
- The regression coefficient corresponding to a dummy variable is the difference between the mean for that level and the mean for baseline group, controlling for all numeric predictors

Example - Deep Cervical Infections

- Subjects - Patients with deep neck infections
- Response (*Y*) - Length of Stay in hospital
- Predictors: (One numeric, 11 Dichotomous)
  - Age (x_1)
  - Gender (x_2=1 if female, 0 if male)
  - Fever (x_3=1 if Body Temp > 38C, 0 if not)
  - Neck swelling (x_4=1 if Present, 0 if absent)
  - Neck Pain (x_5=1 if Present, 0 if absent)
  - Trismus (x_6=1 if Present, 0 if absent)
  - Underlying Disease (x_7=1 if Present, 0 if absent)
  - Respiration Difficulty (x_8=1 if Present, 0 if absent)
  - Complication (x_9=1 if Present, 0 if absent)
  - WBC > 15000/mm^3 (x_{10}=1 if Present, 0 if absent)
  - CRP > 100μg/ml (x_{11}=1 if Present, 0 if absent)

Source: Wang, et al (2003)

Example - Weather and Spinal Patients

- Subjects - Visitors to National Spinal Network in 23 cities Completing SF-36 Form
- Response - Physical Function subscale (1 of 10 reported)
- Predictors:
  - Patient's age (x_1)
  - Gender (x_2=1 if female, 0 if male)
  - High temperature on day of visit (x_3)
  - Low temperature on day of visit (x_4)
  - Dew point (x_5)
  - Wet bulb (x_6)
  - Total precipitation (x_7)
  - Barometric Pressure (x_7)
  - Length of sunlight (x_8)
  - Moon Phase (new, wax crescent, 1st Qtr, wax gibbous, full moon, wan gibbous, last Qtr, wan crescent, presumably had 8-1=7 dummy variables)

Source: Glaser, et al (2004)

Analysis of Covariance

- Combination of 1-Way ANOVA and Linear Regression
- Goal: Comparing numeric responses among *k* groups, adjusting for numeric concomitant variable(s), referred to as **Covariate(s)**
- Clinical trial applications: Response is Post-Trt score, covariate is Pre-Trt score
- Epidemiological applications: Outcomes compared across exposure conditions, adjusted for other risk factors (age, smoking status, sex,...)

Multivariate Linear Regression

Dr. Kourosh Sayehmiri

Multivariate Analysis

- Every program has three major elements that might affect cost:
  - Size
    - Weight, Volume, Quantity, etc...
  - Performance
    - Speed, Horsepower, Power Output, etc...
  - Technology $\mathbf{Y_i = b_0 + b_1X + \varepsilon_\iota}$
    - Gas turbine, Stealth, Composites, etc…

- So far we've tried to select cost drivers that

Multivariate Analysis

- What if one variable is not enough?
- What if we believe there are other significant cost drivers?
- In Multivariate Linear Regression we will be working with the following model:

$$\mathbf{Y_i = b_0 + b_1X_1 + b_2X_2 + \ldots + b_kX_k + \varepsilon_i}$$

- What do we hope to accomplish by bringing in additional independent variables?

Multiple Regression

$$y = a + b_1x_1 + b_2x_2 + \ldots + b_kx_k + \varepsilon$$

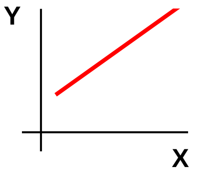
- In general the underlying math is similar to the simple model, but matrices are used to represent the coefficients and variables
  - Understanding the math requires background in Linear Algebra
  - Demonstration is beyond the scope of the module, but can be obtained from the references
- Some key points to remember for multiple regression include:
  - Perform residual analysis between each X variable and Y
  - Avoid high correlation between X variables
  - Use the “Goodness of Fit” metrics and statistics to guide you toward a good model

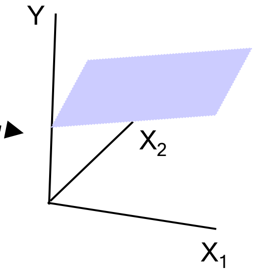
Multiple Regression

- If there is more than one independent variable in linear regression we call it *multiple regression*
- The general equation is as follows:

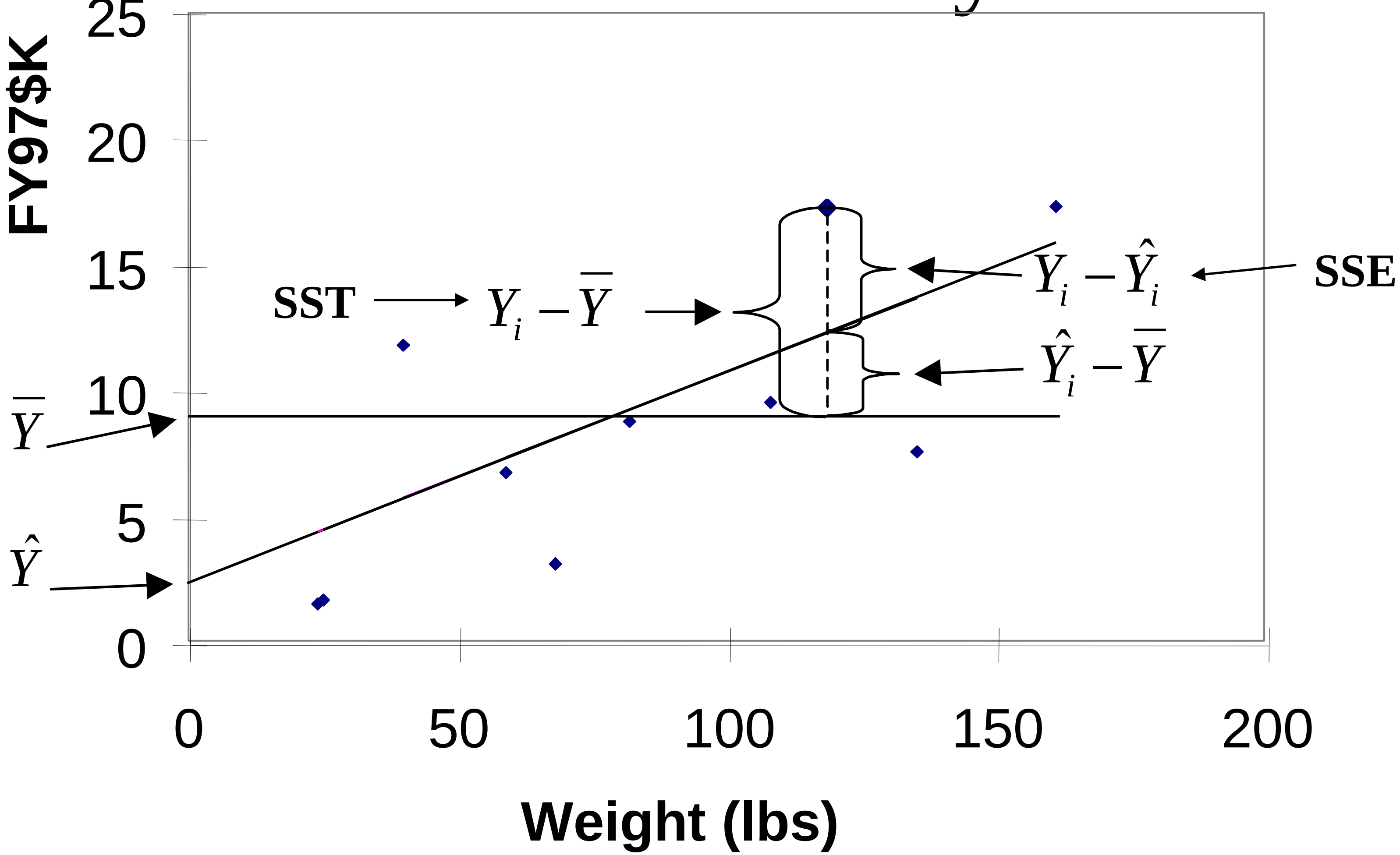
$$y = a + b_1x_1 + b_2x_2 + \ldots + b_kx_k + \varepsilon$$

  - So far, we have seen that for one independent variable, the equation forms a line in 2-dimensions
  - For two independent variables, the equation forms a plane in 3-dimensions
  - For three or more variables, we are working in higher dimensions and cannot picture the equation
- The math is more complicated, but the results can be easily obtained from a regression tool like the one in Excel





Multivariate Analysis



Multivariate Analysis

- Regardless of how many independent variables we bring into the model, we cannot change the total variation:

$$SST = \sum (y_i - \bar{y})^2$$

$$SSE = \sum (y_i - \hat{y}_X)^2$$

- We can only attempt to minimize the unexplained variation:

Multivariate Analysis

- The same regression assumptions still apply:
  - Values of the independent variables are known.
  - The e_i are normally distributed random variables with mean equal to zero and constant variance.
  - The error terms are uncorrelated

- We will introduce Multicollinearity and talk further about the t-statistic.

Multivariate Analysis

- What do the coefficients, $(b_1, b_2, \ldots, b_k)$ represent?
- In a simple linear model with one X, we would say b_1 represents the change in Y given a one unit change in X.
- In the multivariate model, there is more of a conditional relationship.
  - Y is determined by the combined effects of all the X's.

Multicollinearity

- One factor in the ability of the regression coefficient to accurately reflect the marginal contribution of an independent variable is the amount of independence between the independent variables.
- If X_i and X_j are statistically independent, then a change in X_i has no correlation to a change in X_j.
- Usually, however, there is some amount of correlation between variables.

Multicollinearity

- One of the ways we can detect multicollinearity is by observing the regression coefficients.
- If the value of b_1 changes significantly from an equation with X_1 only to an equation with X_1 and X_2, then there is a significant amount of correlation between X_1 and X_2.
- A better way of detecting this is by looking at a pairwise correlation matrix.

Multicollinearity

- In general, multicollinearity does not necessarily affect our ability to get a good fit, nor does it affect our ability to obtain a good prediction, *provided that we maintain the multicollinear relationship between variables*.
- How do we determine that relationship?
- Run simple linear regression between the two correlated variables.

Effects of Multicollinearity

- Creates variability in the regression coefficients
  - First, when X_1 and X_2 are highly correlated, the coefficients of each may change significantly from the one-variable models to the multivariable models.
  - Consider the following equations from the missile data set:

Cost = (-24.486) + 7.7899 * Weight

Cost = 59.575 + 0.3096 * Range

Cost = (-21.878) + 8.3175 * Weight + (-0.0311) * Range

Effects of Multicollinearity

- Example

Cost	Thrust	Weight
10	7	18
20	8	44
30	17	57
30	13	67
50	22	112
60	34	112
70	39	128
80	39	165

Effects of Multicollinearity

Regression Statistics	
Multiple R	0.9781
R Square	0.9568
Adjusted R Square	0.9496
Standard Error	5.6223
Observations	8

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	4197.838	4197.838	132.799	0.000
Residual	6	189.662	31.610		
Total	7	4387.500			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	2.712	4.078	0.665	0.531	-7.268	12.691
Thrust	1.834	0.159	11.524	0.000	1.445	2.224

$$Cost = 2.712 + 1.834 \times (Thrust)$$

Effects of Multicollinearity

Regression Statistics	
Multiple R	0.9870
R Square	0.9742
Adjusted R Square	0.9699
Standard Error	4.3465
Observations	8

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	4274.147	4274.147	226.240	0.000
Residual	6	113.353	18.892		
Total	7	4387.500			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-0.4177	3.3142	-0.1260	0.9038	-8.5273	7.6920
Weight	0.5026	0.0334	15.0413	0.0000	0.4209	0.5844

$$Cost = (-0.418) + 0.503 \times (Weight)$$

Effects of Multicollinearity

Regression Statistics	
Multiple R	0.9997
R Square	0.9995
Adjusted R Square	0.9992
Standard Error	0.6916
Observations	8

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	4385.108	2192.554	4583.300	0.000
Residual	5	2.392	0.478		
Total	7	4387.500			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-0.5062	0.5274	-0.9598	0.3813	-1.8620	0.8496
Thrust	0.8291	0.0544	15.2300	0.0000	0.6892	0.9690
Weight	0.2925	0.0148	19.7856	0.0000	0.2545	0.3305

$$Cost = (-0.506) + 0.829 \times (Thrust) + 0.293 \times (Weight)$$

Effects of Multicollinearity

$$Cost = 2.712 + 1.834 \times (Thrust)$$

$$Cost = (-0.418) + 0.503 \times (Weight)$$

$$Cost = (-0.506) + 0.829 \times (Thrust) + 0.293 \times (Weight)$$

- Notice how the coefficients have changed by using a two variable model.
- This is an indication that Thrust and Weight are correlated.
- We now regress Weight on Thrust to see what the relationship is between the two variables.

Effects of Multicollinearity

Regression Statistics	
Multiple R	0.9331
R Square	0.8706
Adjusted R Square	0.8491
Standard Error	5.1869
Observations	8

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1086.454	1086.454	40.383	0.001
Residual	6	161.421	26.903		
Total	7	1247.875			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	0.107	3.955	0.027	0.979	-9.571	9.784
Weight	0.253	0.040	6.355	0.001	0.156	0.351

$$Thrust \approx 0.25 \times Weight$$

Effects of Multicollinearity

- **System 1 holds the required relationship between Weight and Thrust (approximately), while System 2 does not.**
- **Notice the variation in the cost estimates for System 2 using the three CERs.**
- **However, System 1, since Weight and Thrust follow the required relationship, is estimated fairly precisely by all three CERs.**

	System 1	System 2
Weight	95	25
Thrust	25	12
Cost (Weight)	47.33	12.15
Cost (Thrust)	48.56	24.72
Cost (Weight, Thrust)	48.01	16.76

Effects of Multicollinearity

- When multicollinearity is present we can no longer make the statement that b_1 is the change in Y for a unit change in X_1 while holding X_2 constant.
  - The two variables may be related in such a way that precludes varying one while the other is held constant.
  - For example, perhaps the only way to increase the range of a missile is to increase the amount of the propellant, thus increasing the missile

Remedies for Multicollinearity?

- Drop a variable and ignore an otherwise good cost driver?
  - Not if we don't have to.
- Involve technical experts.
  - Determine if the model is correctly specified.
- Combine the variables by multiplying or dividing them.

- Rule of Thumb for determining if you have

More on the t-statistic

- Lightweight Cruise Missile Database:

Missile	Unit Cost (CY95$K)	Empty Weight	Max Speed	Range
A	290	39	0.7	600
B	420	54	0.66	925
C	90	16	0.84	450
D	95	15	0.59	420
E	420	57	0.37	1000
F	380	52	0.52	800
G	370	52	0.63	790
H	450	63	0.44	1600

More on the t-statistic

I. Model Form and Equation

Model Form: **Linear Model**

Number of Observations: 8

Equation in Unit Space: Cost = -29.668 + 8.342 * Weight + 9.293 * Speed + -0.03 * Range

II. Fit Measures (in Unit Space)

Coefficient Statistics Summary

Variable	Coefficient	Std Dev of Coefficient	t-statistic (coeff/sd)	Significance
Intercept	-29.668	45.699	-0.649	0.5517
Weight	8.342	0.561	14.858	0.0001
Speed	9.293	51.791	0.179	0.8666
Range	-0.03	0.028	-1.055	0.3509

Goodness of Fit Statistics

Std Error (SE)	R-Squared	R-Squared (adj)	CV (Coeff of Variation)
14.747	0.994	0.99	0.047

Analysis of Variance

Due to	Degrees of Freedom	Sum of Squares (SS)	Mean Squares (SS/DF)	F-statistic	Significance
Regression (SSR)	3	146302.033	48767.344	224.258	0
Residuals (Errors) (SSE)	4	869.842	217.46		
Total (SST)	7	147171.875			

More on the t-statistic

I. Model Form and Equation

Model Form: **Linear Model**
Number of Observations: 8
Equation in Unit Space: Cost = -21.878 + 8.318 * Weight + -0.031 * Range

II. Fit Measures (in Unit Space)

Coefficient Statistics Summary

Variable	Coefficient	Std Dev of Coefficient	t-statistic (coeff/sd)	Significance
Intercept	-21.878	12.803	-1.709	0.1481
Weight	8.318	0.49	16.991	0
Range	-0.031	0.024	-1.292	0.2528

Goodness of Fit Statistics

Std Error (SE)	R-Squared	R-Squared (adj)	CV (Coeff of Variation)
13.243	0.994	0.992	0.042

Analysis of Variance

Due to	Degrees of Freedom	Sum of Squares (SS)	Mean Squares (SS/DF)	F-statistic	Significance
Regression (SSR)	2	146295.032	73147.516	417.107	0
Residuals (Errors) (SSE)	5	876.843	175.369		
Total (SST)	7	147171.875			

Selecting the Best Model

Choosing a Model

- We have seen what the linear model is, and explored it in depth
- We have looked briefly at how to generalize the approach to non-linear models
- You may, at this point, have several significant models from regressions
  - One or more linear models, with one or more significant variables
  - One or more non-linear models
- Now we will learn how to choose the “best model”

Steps for Selecting the “Best Model”

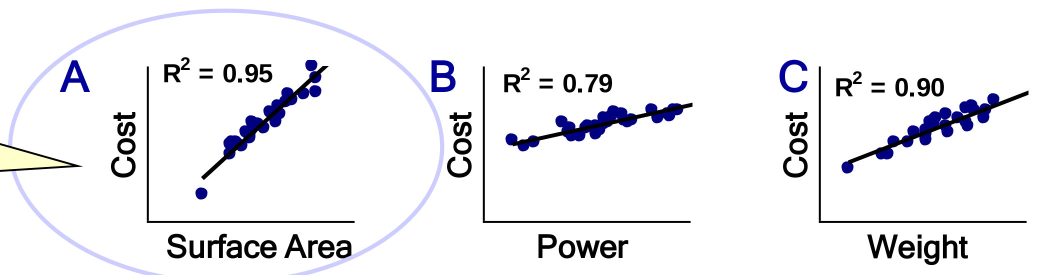
- You should already have rejected all non-significant models first
  - If the F statistic is not significant
- You should already have stripped out all non-significant variables and made the model “minimal”
  - Variables with non-significant t statistics were already removed
- Select “withi
- Select “acros

We will examine each in more detail…

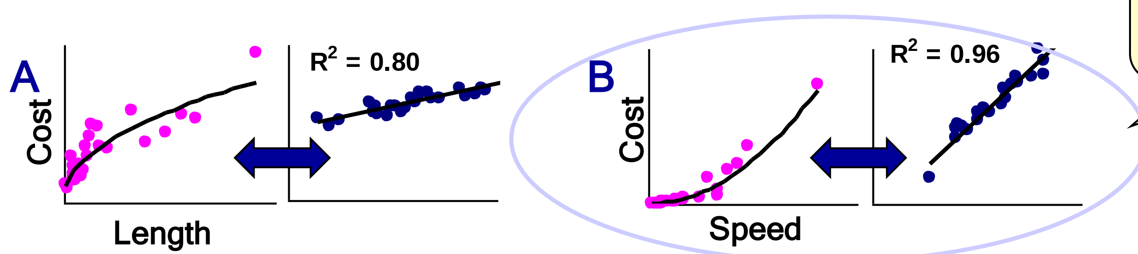
Selecting "Within Type"

- Start with only significant, "minimal" models
- In choosing among "models of a similar form", R^2 is the criterion
- "Models of a similar form" means that you will compare
  - e.g., linear models with other linear models
  - e.g., power models with other power models

Select the model with the highest R^2



Select the model with the highest R^2



Tip: If a model has a lower R^2, but has variables that are more useful for decision makers, retain these, and consider using them for CAIV trades and the like

Selecting "Across Type"

- Start with only significant, "minimal" models
- In choosing among "models of a different form", the SSE in unit space is the criterion
- "Models of a different form" means that you will compare:
  - e.g., linear models with non-linear models
  - e.g., power models with logarithmic models
- We must compute the SSE by:
  - Computing $\hat{Y}$ *in unit space* for each data point
  - Subtracting each $\hat{Y}$ from its corresponding actual Y value
  - Sum the squared values, this is the SSE
- An example follows…

Warning: We cannot use R^2 to compare models of different forms because the R^2 from the regression is computed on the transformed data, and thus is distorted by the transformation

Introduction to Survival Analysis

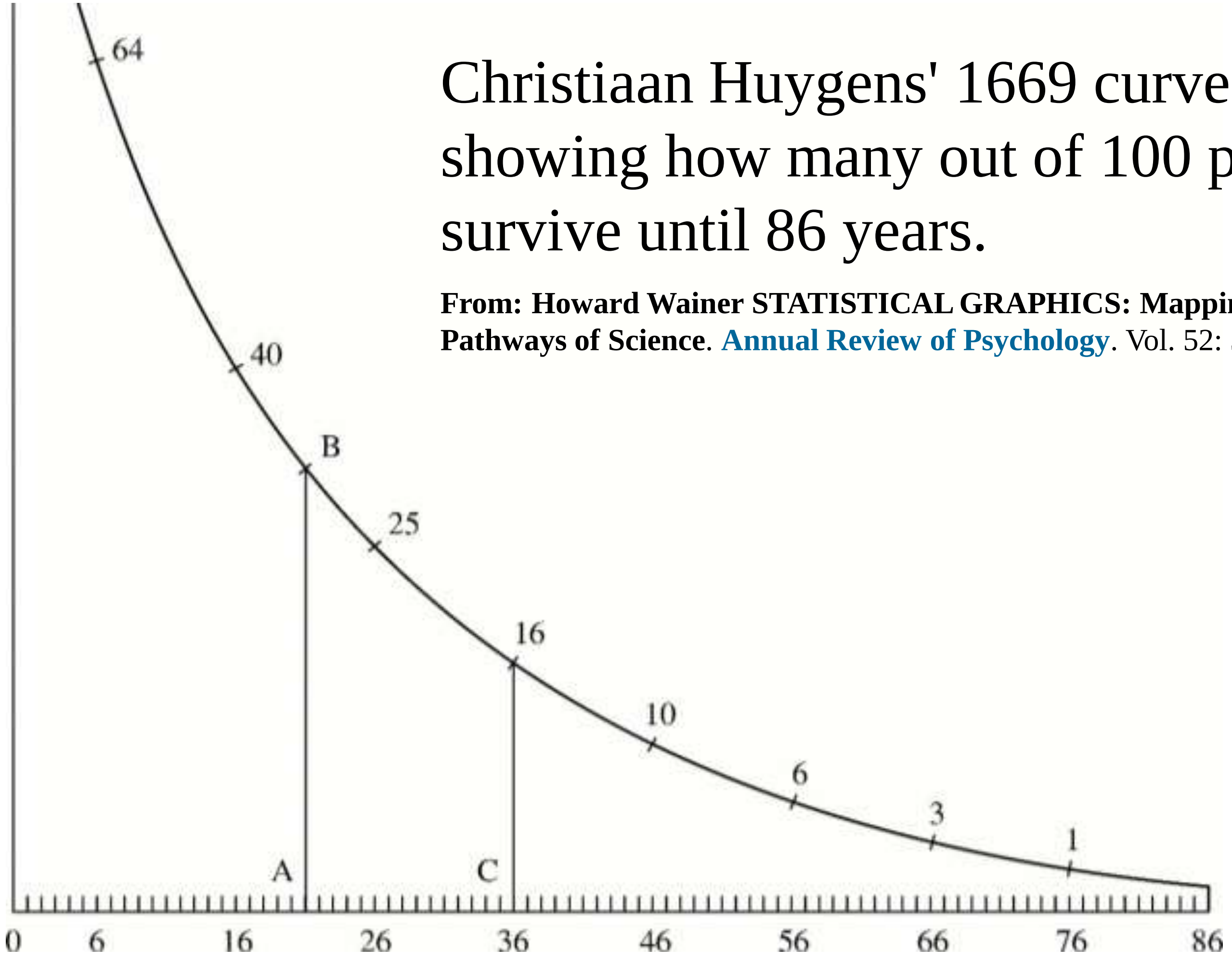
Dr. Kourosh sayehmiri   Ph.D.

In Biostatistics

Overview

- What is survival analysis?
- Terminology and data structure.
- Survival/hazard functions.
- Parametric versus semi-parametric regression techniques.
- Introduction to Kaplan-Meier methods (non-parametric).

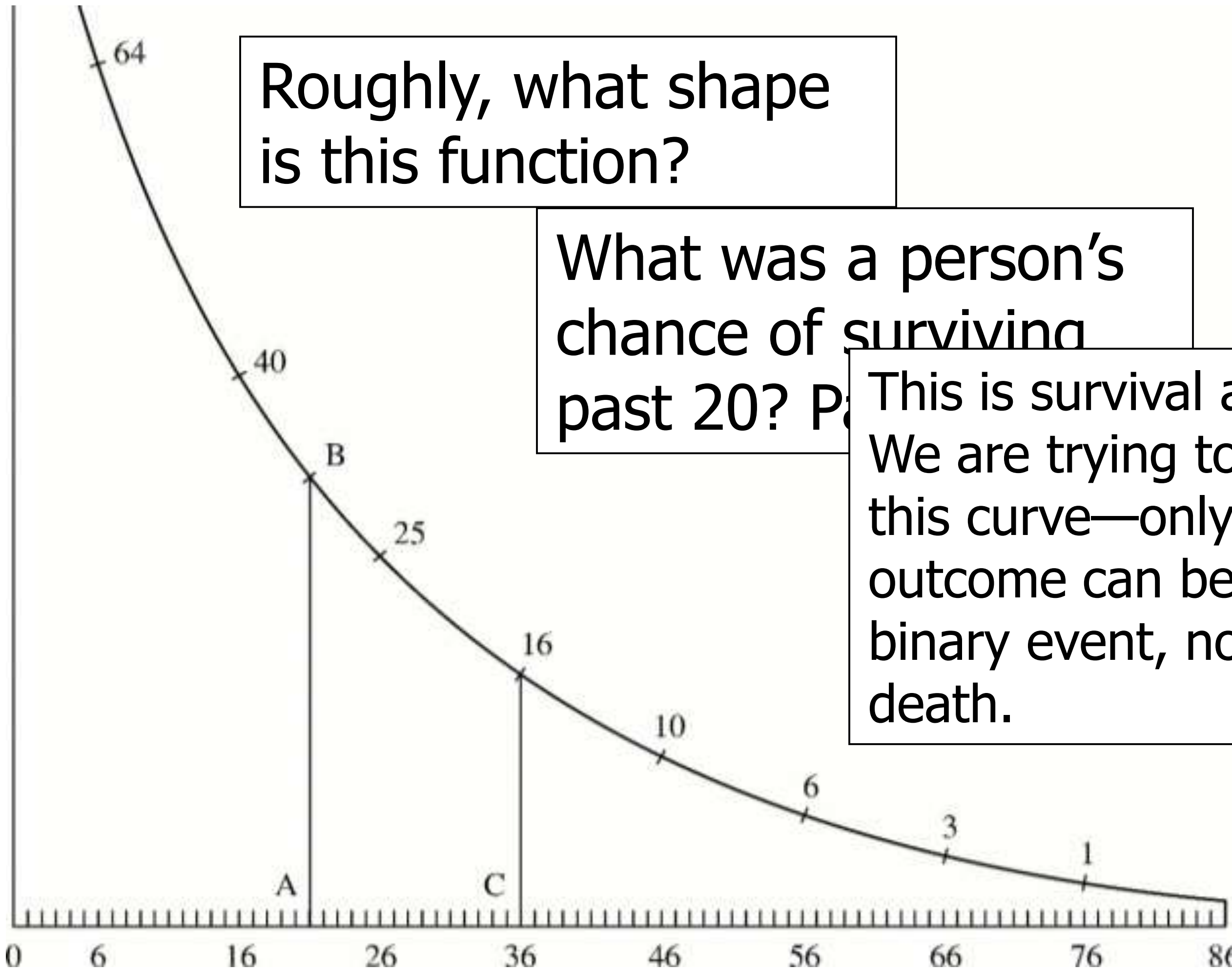
Early example of survival analysis, 1669



Christiaan Huygens' 1669 curve showing how many out of 100 people survive until 86 years.

From: Howard Wainer STATISTICAL GRAPHICS: Mapping the Pathways of Science. **Annual Review of Psychology**. Vol. 52: 305-335.

Early example of survival analysis



Roughly, what shape is this function?

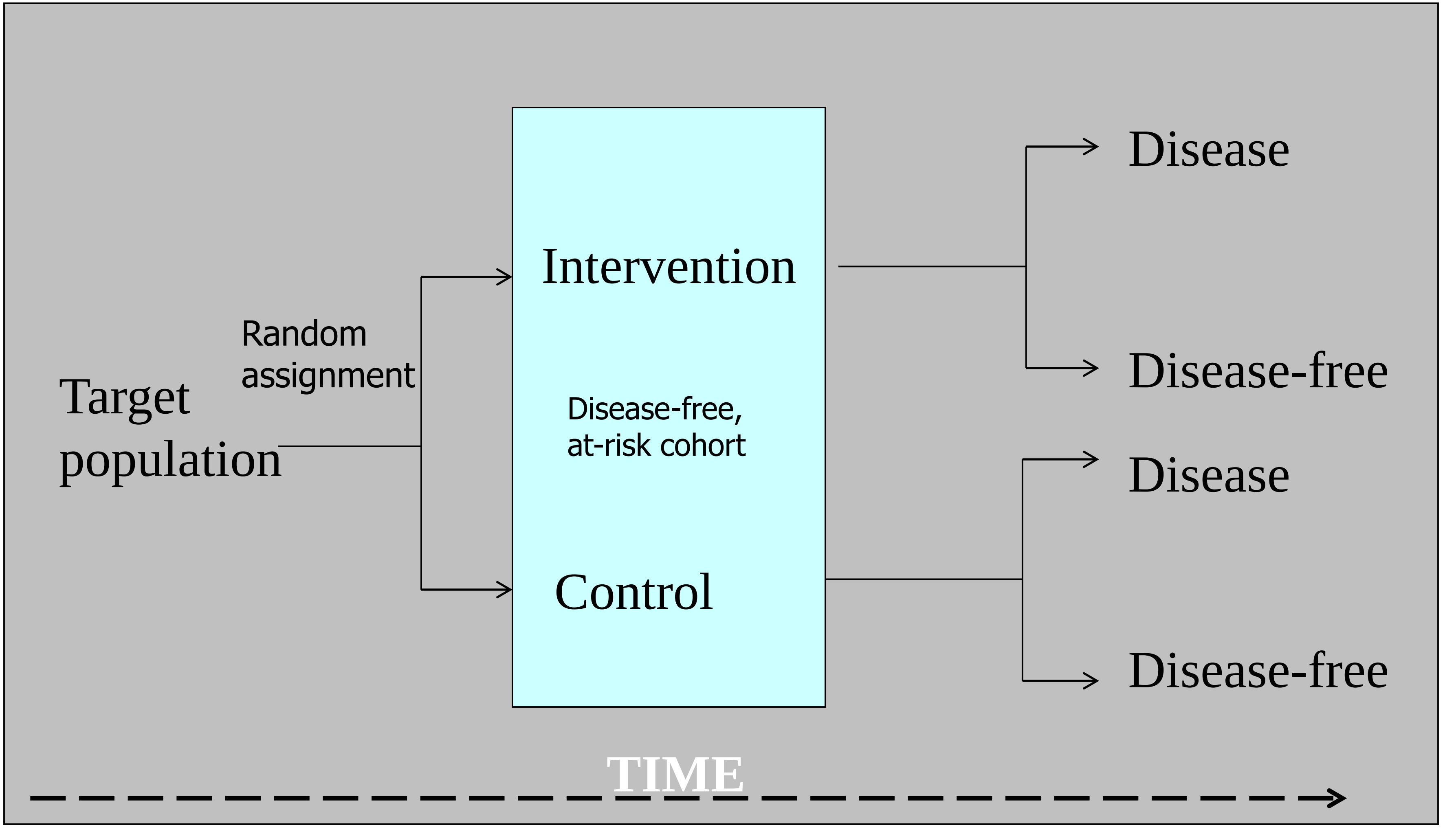
What was a person's chance of surviving past 20? Pa

This is survival analysis! We are trying to estimate this curve—only the outcome can be any binary event, not just death.

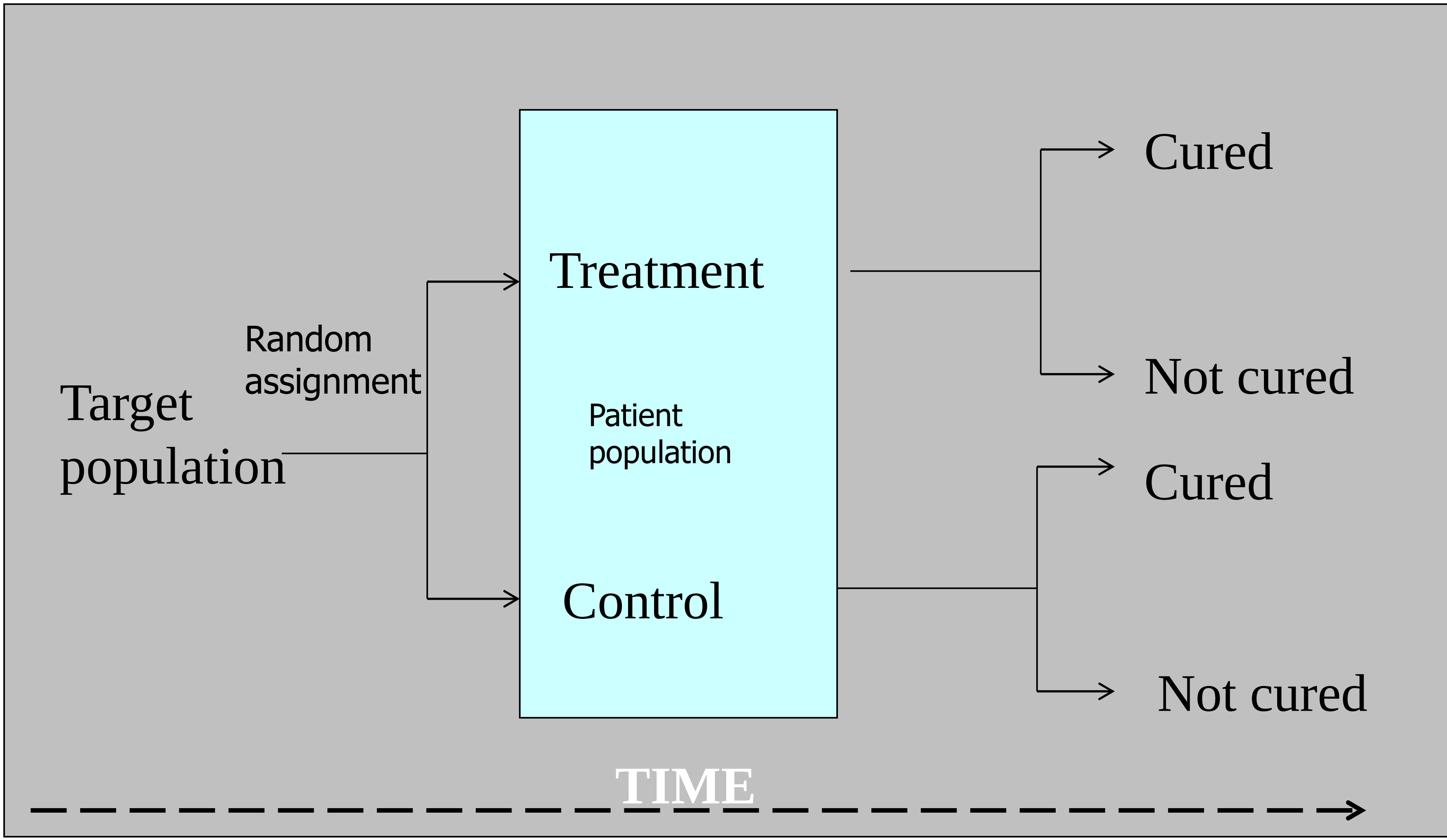
What is survival analysis?

- Statistical methods for analyzing longitudinal data on the occurrence of events.
- Events may include death, injury, onset of illness, recovery from illness (binary variables) or transition above or below the clinical threshold of a meaningful continuous variable (e.g. CD4 counts).
- Accommodates data from randomized clinical trial or cohort study design.

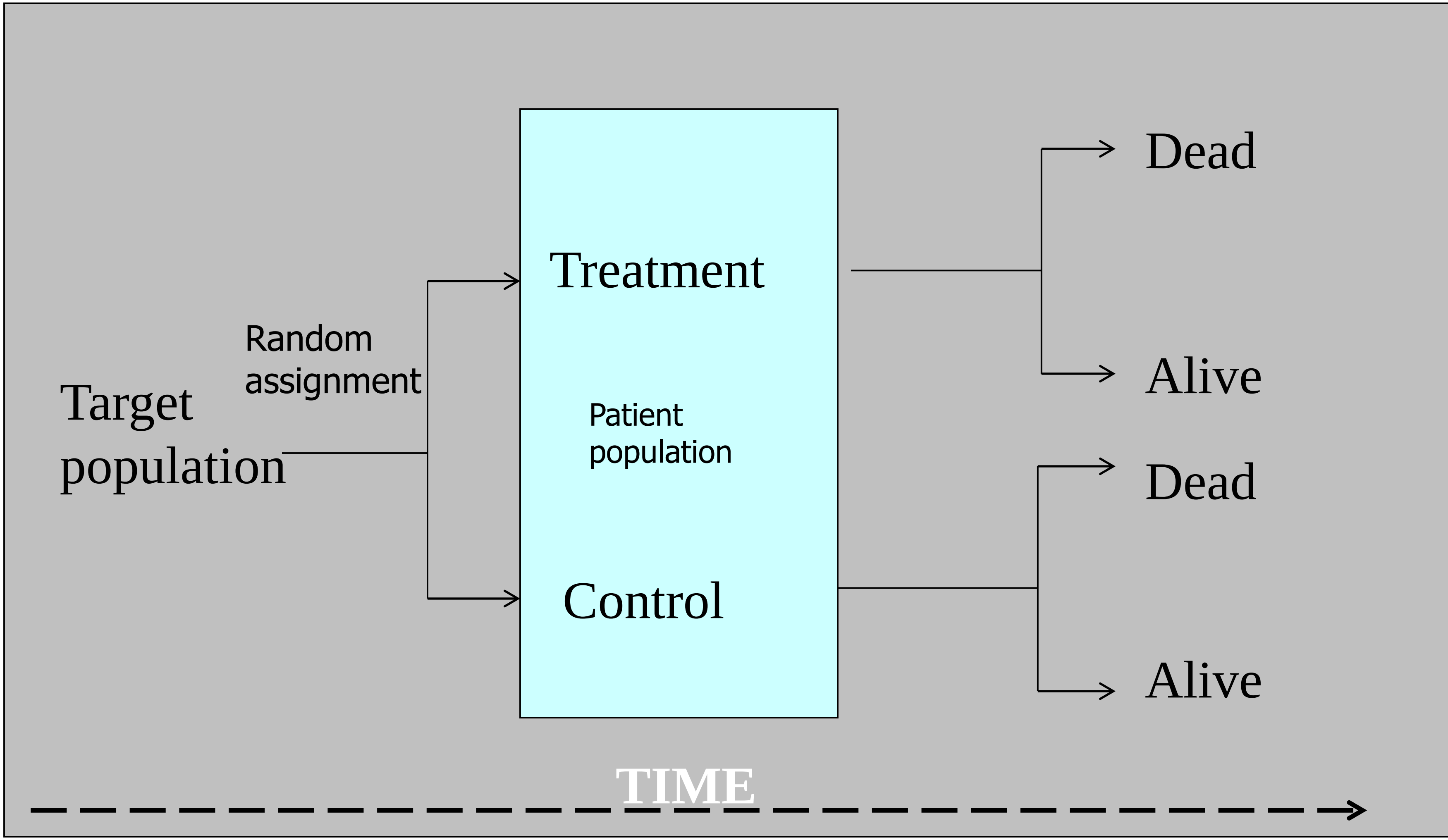
Randomized Clinical Trial (RCT)



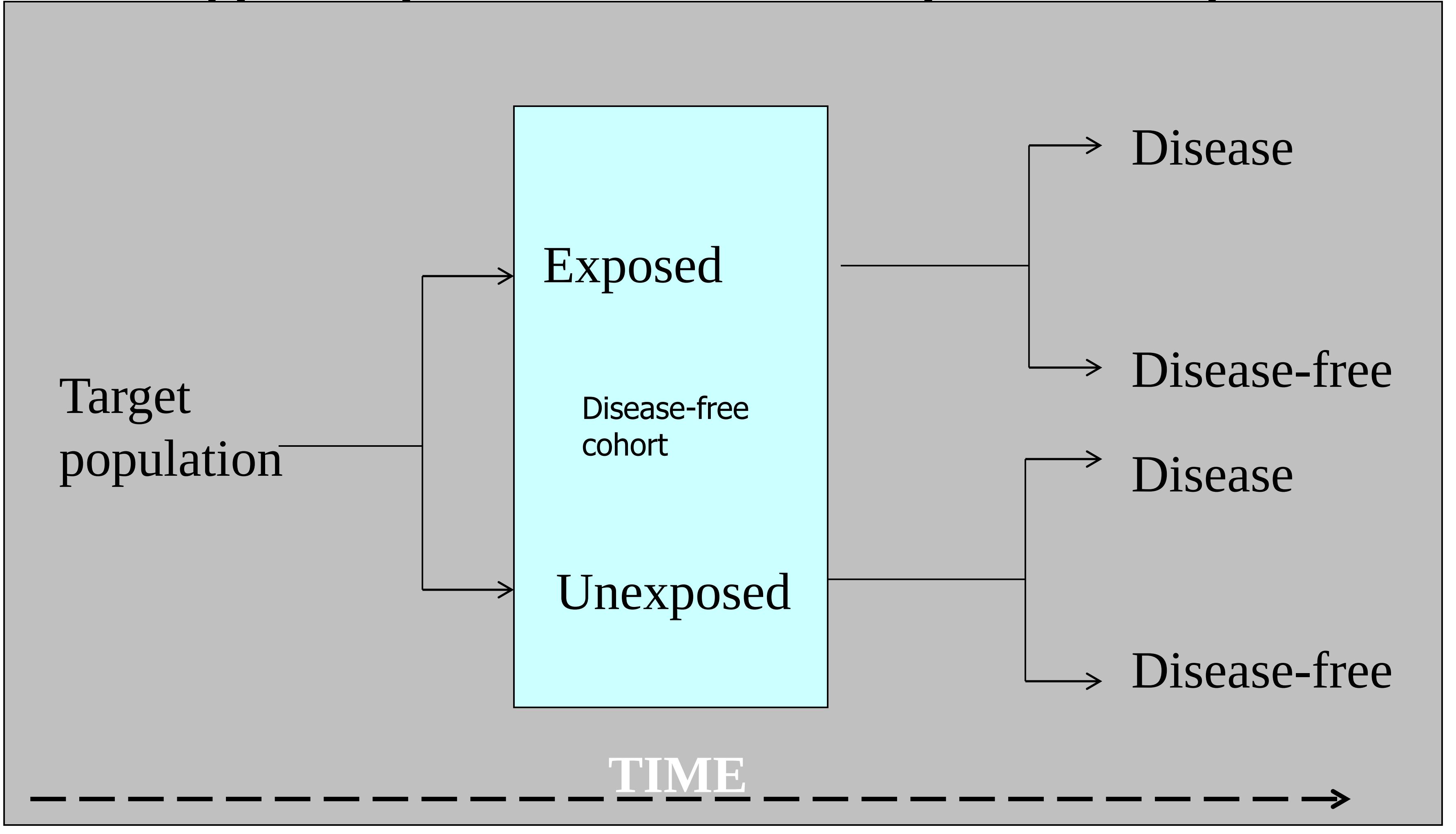
Randomized Clinical Trial (RCT)



Randomized Clinical Trial (RCT)

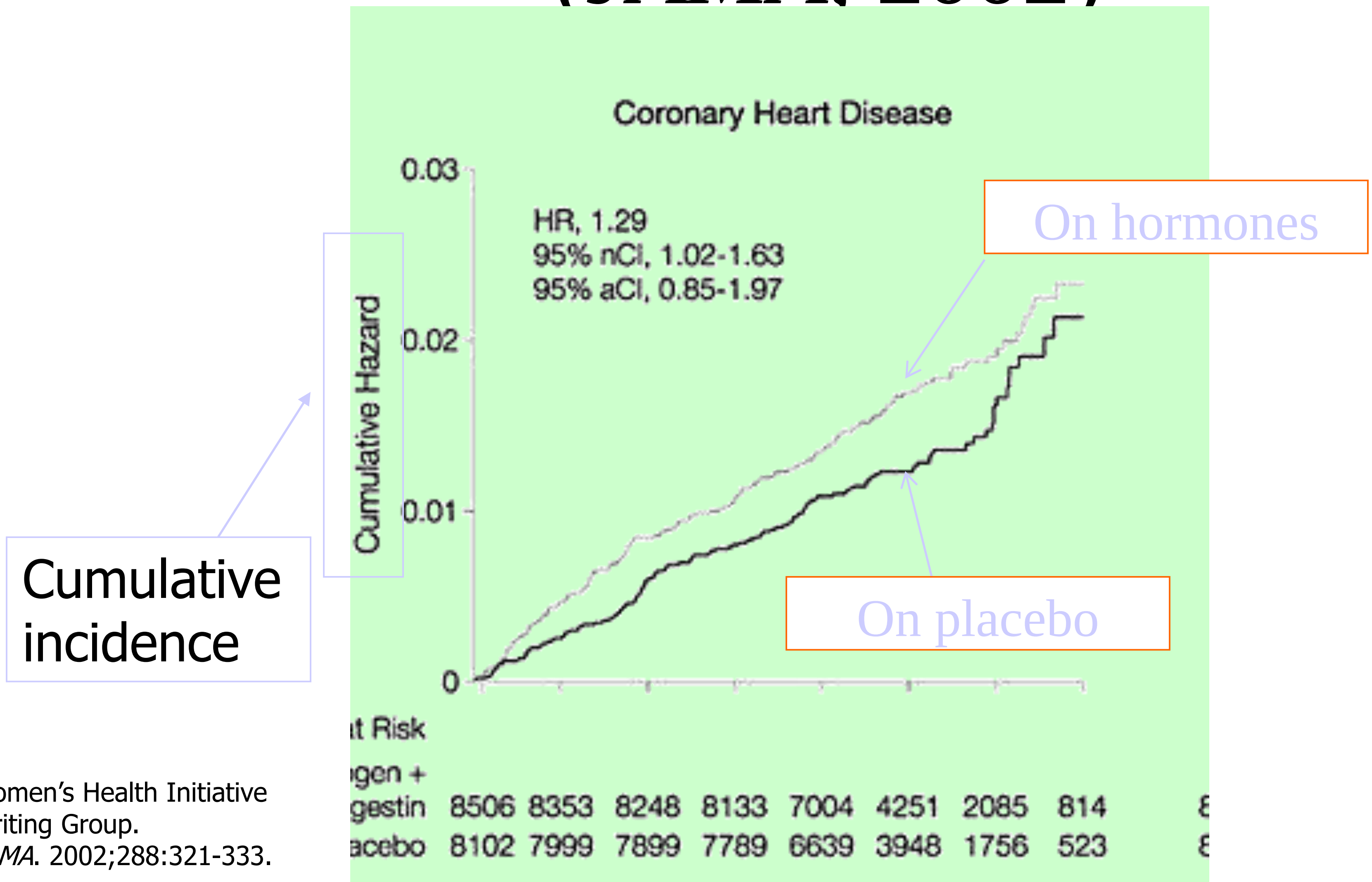


Cohort study (prospective/retrospective)



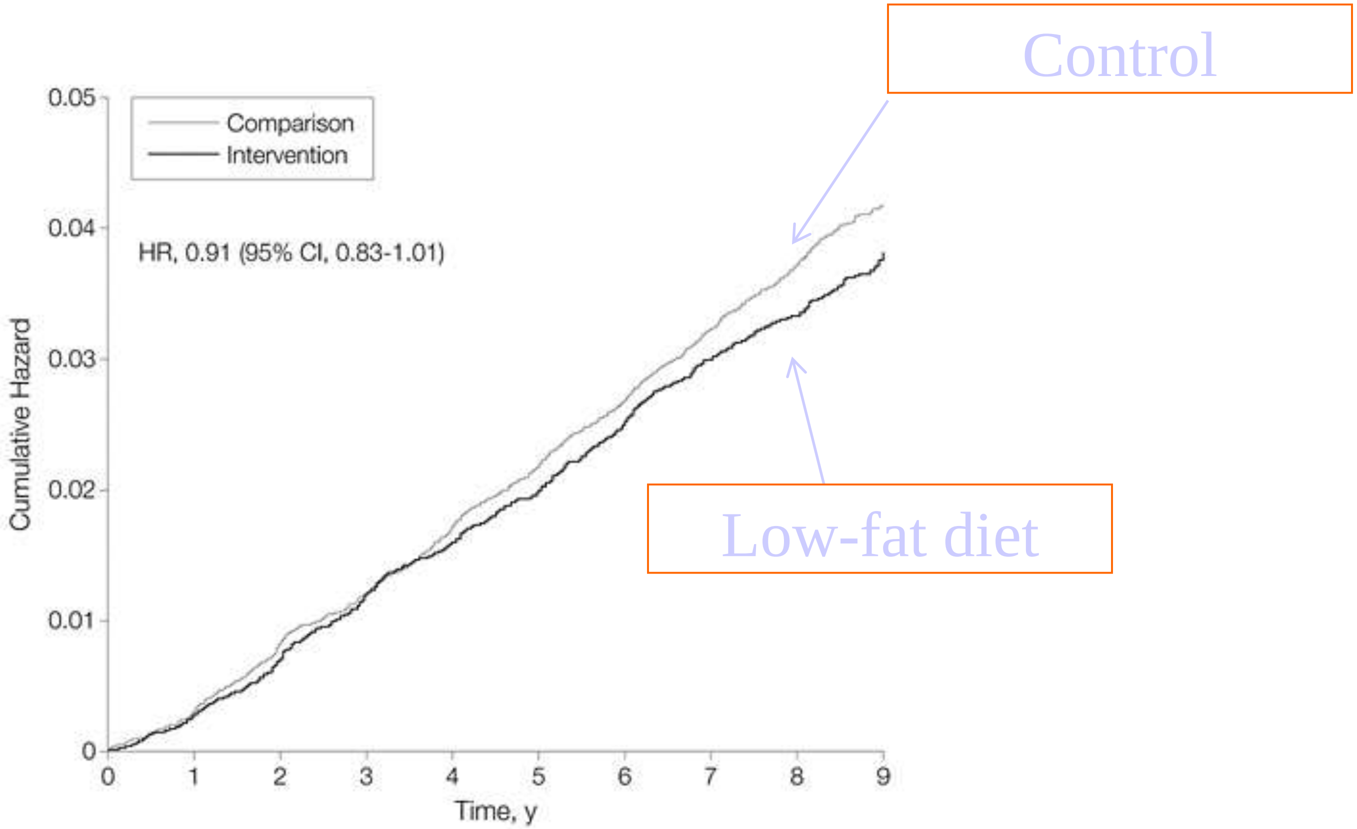
Examples of survival analysis in medicine

RCT: Women's Health Initiative (*JAMA*, 2002)



Women's Health Initiative Writing Group. *JAMA*. 2002;288:321-333.

WHI and low-fat diet…



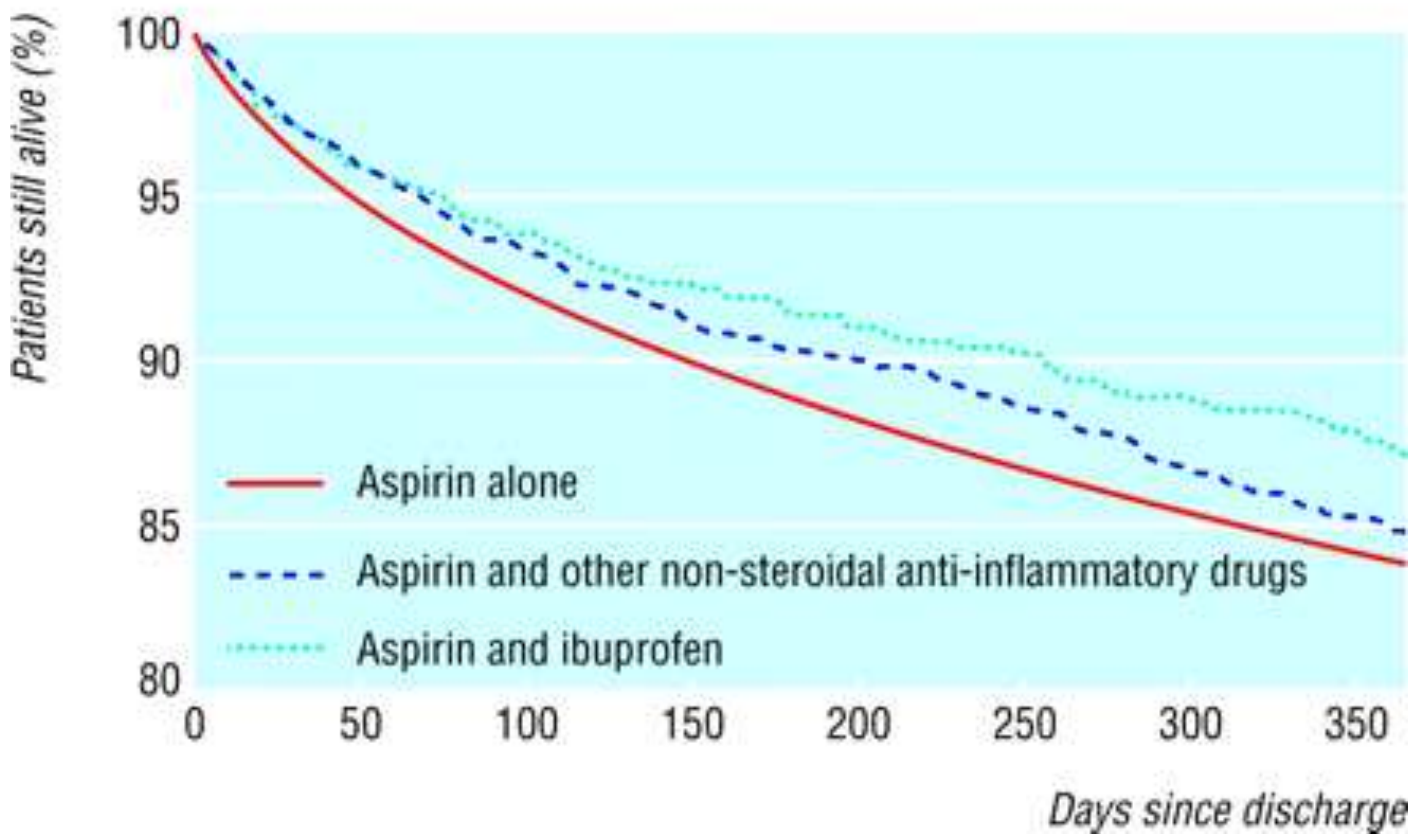
	0	1	2	3	4	5	6	7	8	9
Events										
Intervention		47	79	92	80	72	94	89	46	33
Comparison		74	140	123	137	136	137	145	97	58
No. at Risk										
Intervention	19541	19328	19084	18798	18520	18263	17900	15507	10245	5075
Comparison	29294	28908	28536	28195	27806	27372	26977	23337	15373	7580

Prentice et al. *JAMA*, February 8, 2006; 295: 629 - 642.

Retrospective cohort study:

From December 2003 *BMJ*:

Aspirin, ibuprofen, and mortality after myocardial infarction: retrospective cohort study



Curits et al. *BMJ* 2003;327:1322-1323.

Objectives of survival analysis

- **Estimate time-to-event for a group of individuals**, such as time until second heart-attack for a group of MI patients.
- **To compare time-to-event between two or more groups**, such as treated vs. placebo MI patients in a randomized controlled trial.
- **To assess the relationship of co-variables to time-to-event**, such as: does weight, insulin resistance, or cholesterol influence survival time of MI patients?

Note: expected time-to-event = 1/incidence rate

Why use survival analysis?

1. Why not compare mean time-to-event between your groups using a t-test or linear regression?

-- ignores censoring

2. Why not compare proportion of events in your groups using risk/odds ratios or logistic regression?

--ignores time

Survival Analysis: Terms

- Time-to-event: The time from entry into a study until a subject has a particular outcome
- Censoring: Subjects are said to be censored if they are lost to follow up or drop out of the study, or if the study ends before they die or have an outcome of interest. They are counted as alive or disease-free for the time they were enrolled in the study.
  - If dropout is related to both outcome and treatment, dropouts may bias the results

Data Structure: survival analysis

Two-variable outcome :

- Time variable: t_i = time at last disease-free observation or time at event
- Censoring variable: c_i =1 if had the event; c_i =0 no event by time t_i

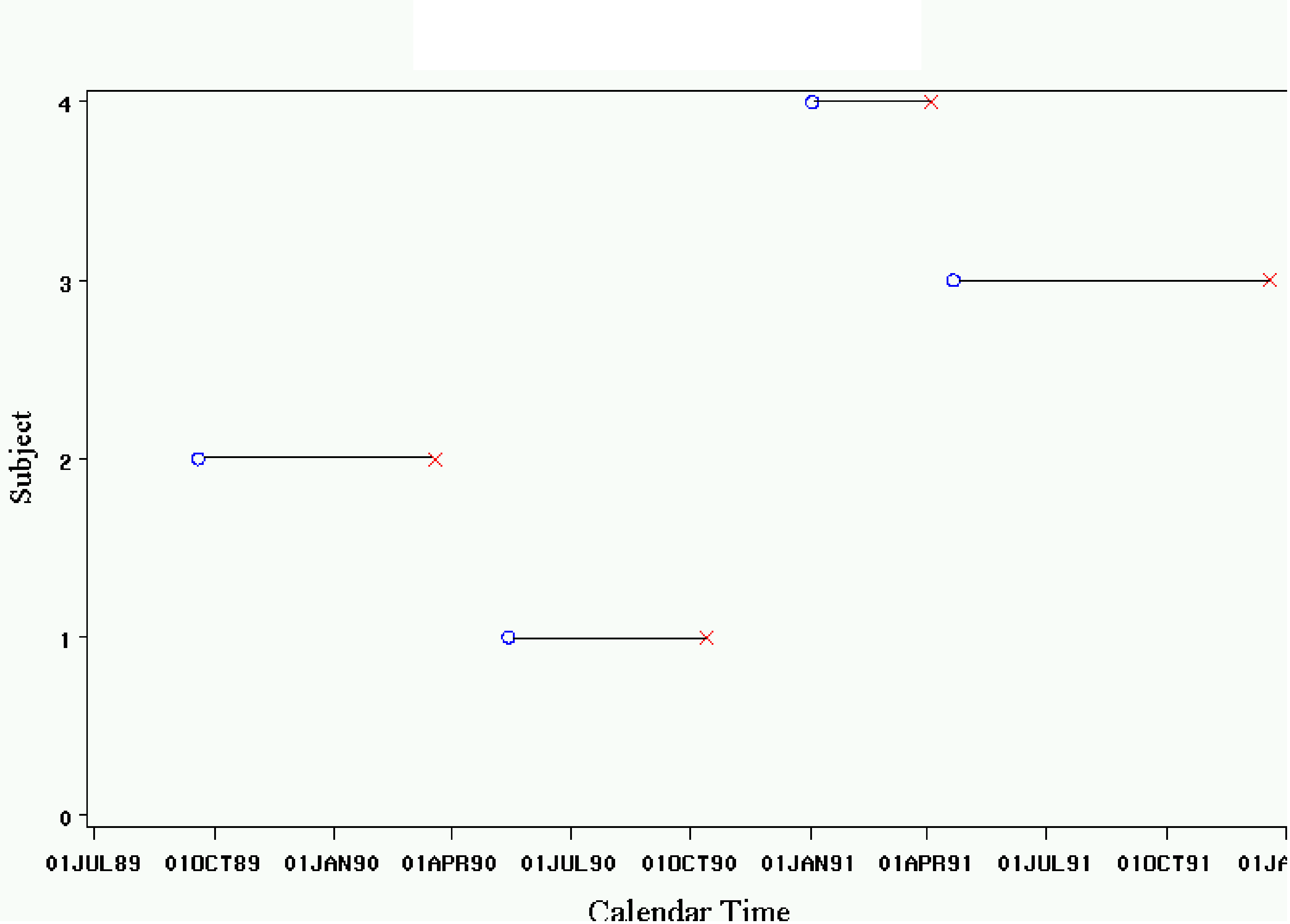
Right Censoring (T>t)

<u>Common examples</u>

- Termination of the study
- Death due to a cause that is not the event of interest
- Loss to follow-up

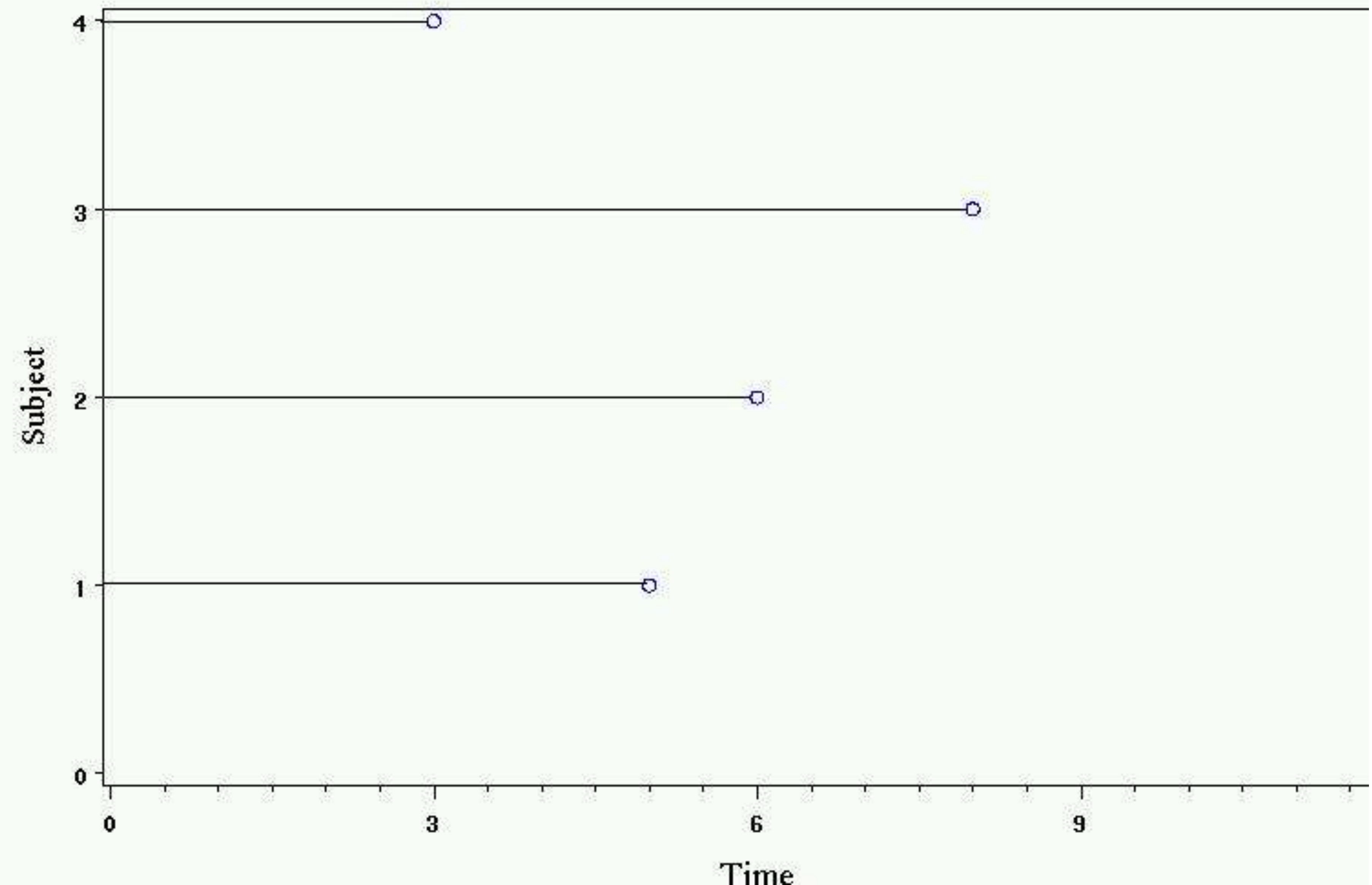
We know that subject survived at least to time *t*.

Choice of time of origin. Note varying start times.



Count every subject's time since their baseline data collection.

Right-censoring!



Introduction to survival distributions

- T_i the event time for an individual, is a random variable having a probability distribution.
- Different models for survival data are distinguished by different choice of distribution for T_i.

Describing Survival Distributions

Parametric survival analysis is based on so-called “Waiting Time” distributions (ex: exponential probability distribution).

<u>The idea is this:</u>

Assume that times-to-event for individuals in your dataset follow a continuous probability distribution (which we may or may not be able to pin down mathematically).

For all possible times T_i after baseline, there is a certain probability that an individual will have an event at exactly time T_i. For example, human beings have a certain probability of dying at ages 3, 25, 80, and 140: P(T=3), P(T=25), P(T=80), P(T=140). These probabilities are obviously vastly different.

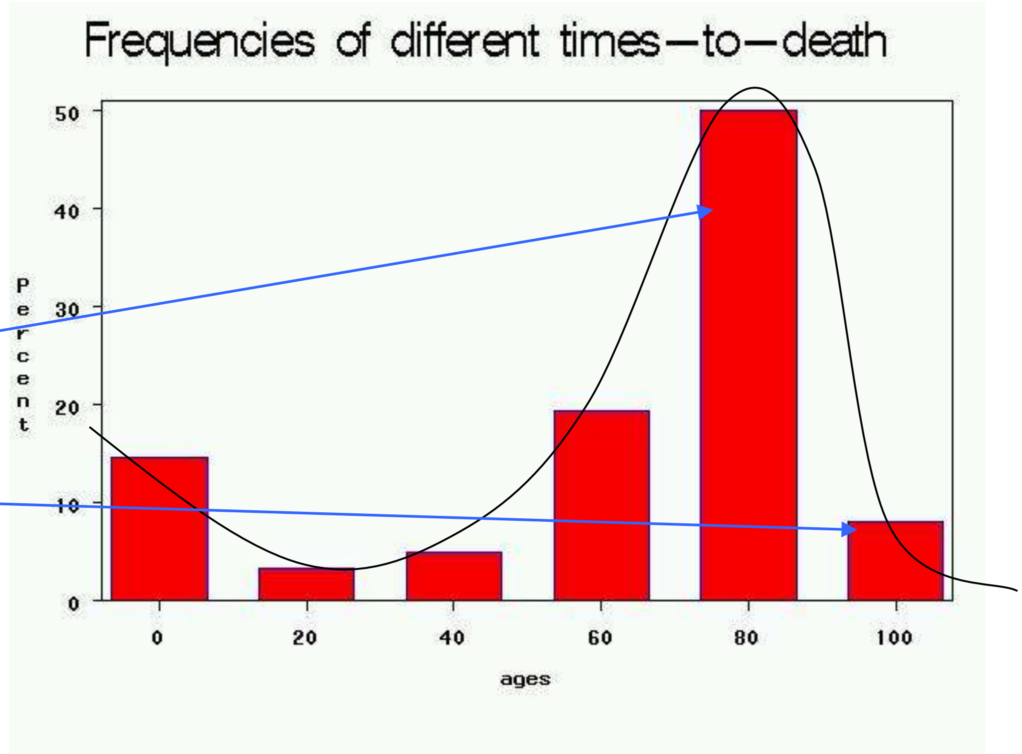
Probability density function: f(t)

In the case of human longevity, T_i is unlikely to follow a normal distribution, because the probability of death is not highest in the middle ages, but at the beginning and end of life.

Hypothetical data:

People have a high chance of dying in their 70's and 80's;

BUT they have a smaller chance of dying in their 90's and 100's, because few people make it long enough to die at these ages.



Probability density function: f(t)

The probability of the failure time occurring at exactly time t (out of the whole range of possible t's).

$$f(t) = \lim_{\Delta t \longrightarrow 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t}$$

Survival function: 1-F(t)

The goal of survival analysis is to estimate and compare survival experiences of different groups.

Survival experience is described by the cumulative survival function:

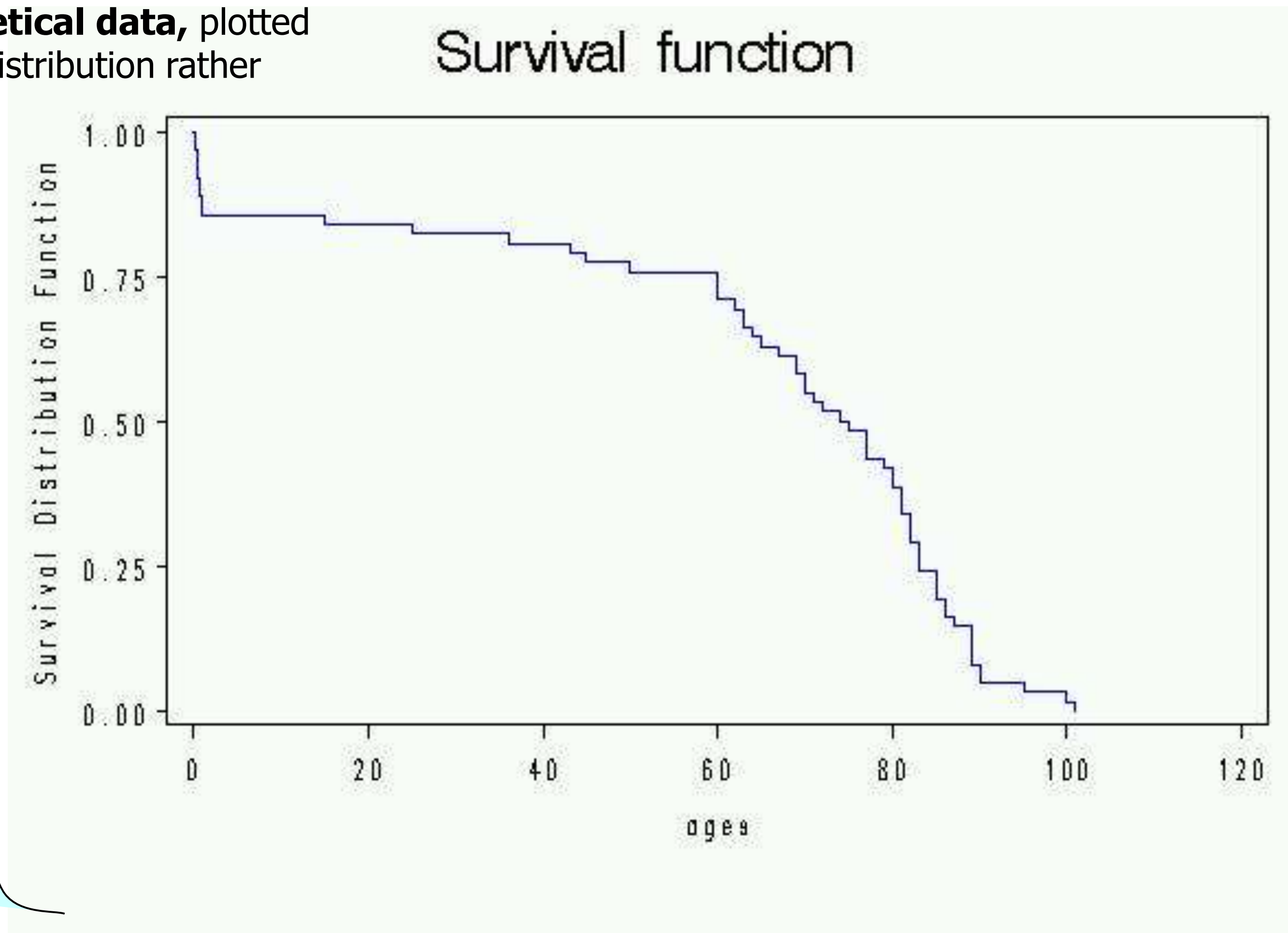
$$S(t) = 1 - P(T \leq t) = 1 - F(t)$$

F(t) is the CDF of f(t), and is "more interesting" than f(t).

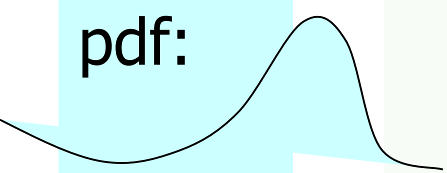
Example: If t=100 years, S(t=100) = probability of surviving beyond 100 years.

Cumulative survival

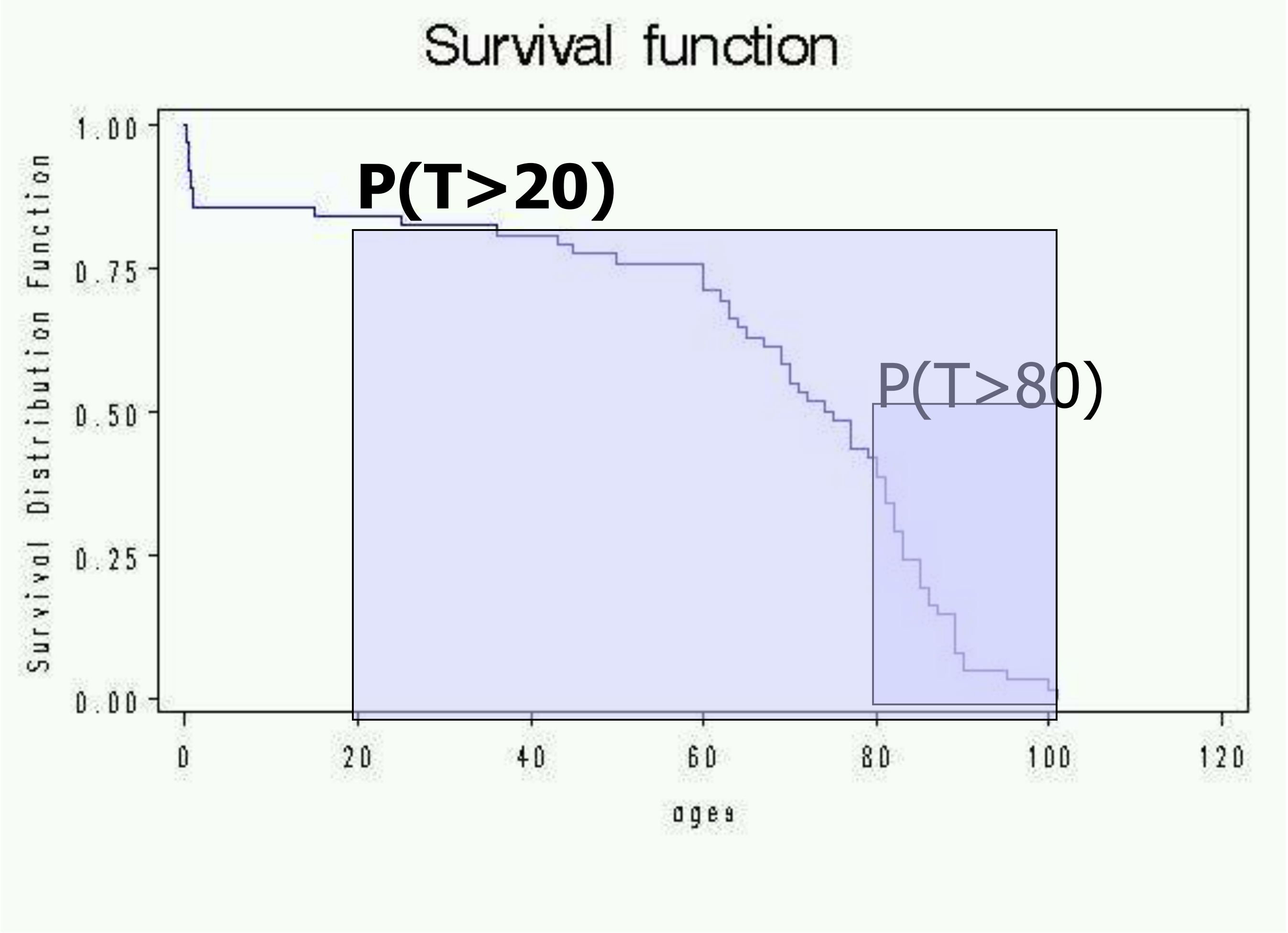
Same hypothetical data, plotted as cumulative distribution rather than density:



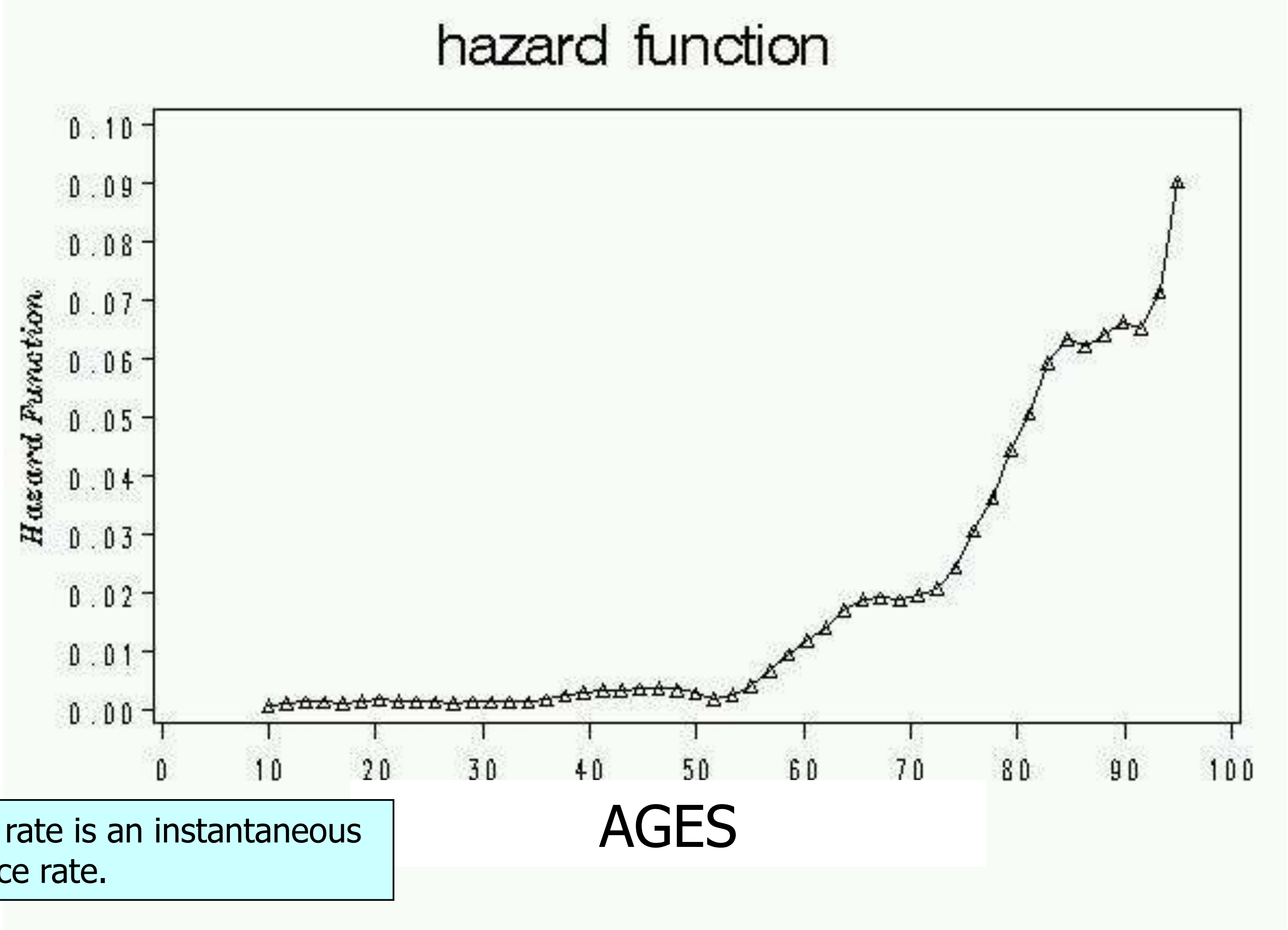
Recall pdf:



Cumulative survival



Hazard Function: new concept



Hazard rate is an instantaneous incidence rate.

Hazard function

$$h(t) = \lim_{\Delta t \longrightarrow 0} \frac{P(t \leq T < t + \Delta t / T \geq t)}{\Delta t}$$

<u>In words:</u> the probability that ***if you survive to t***, you will succumb to the event in the next instant.

$$\text{Hazard from density and survival: } h(t) = \frac{f(t)}{S(t)}$$

Derivation (Bayes' rule):

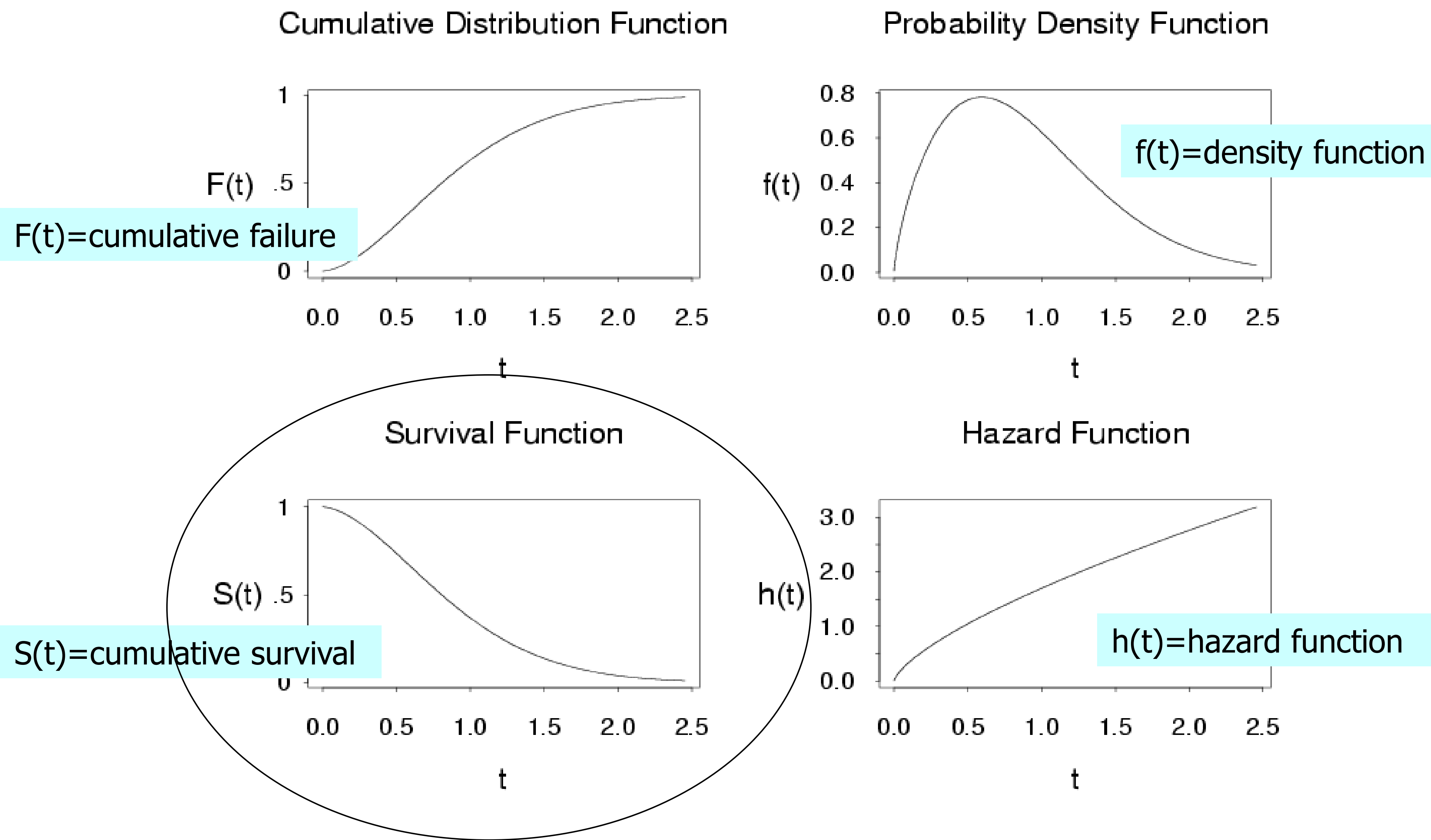
$$h(t)dt = P(t \leq T < t + dt / T \geq t) = \frac{P(t \leq T < t + dt \;\&\; T \geq t)}{P(T \geq t)} = \frac{P(t \leq T < t + dt)}{P(T \geq t)} = \frac{f(t)dt}{S(t)}$$

Hazard vs. density

This is subtle, but the idea is:

- When you are born, you have a certain probability of dying at any age; that's the probability density (think: marginal probability)
  - Example: a woman born today has, say, a 1% chance of dying at 80 years.
- However, as you survive for awhile, your probabilities keep changing (think: conditional probability)
  - Example, a woman who is 79 today has, say, a 5% chance of dying at 80 years.

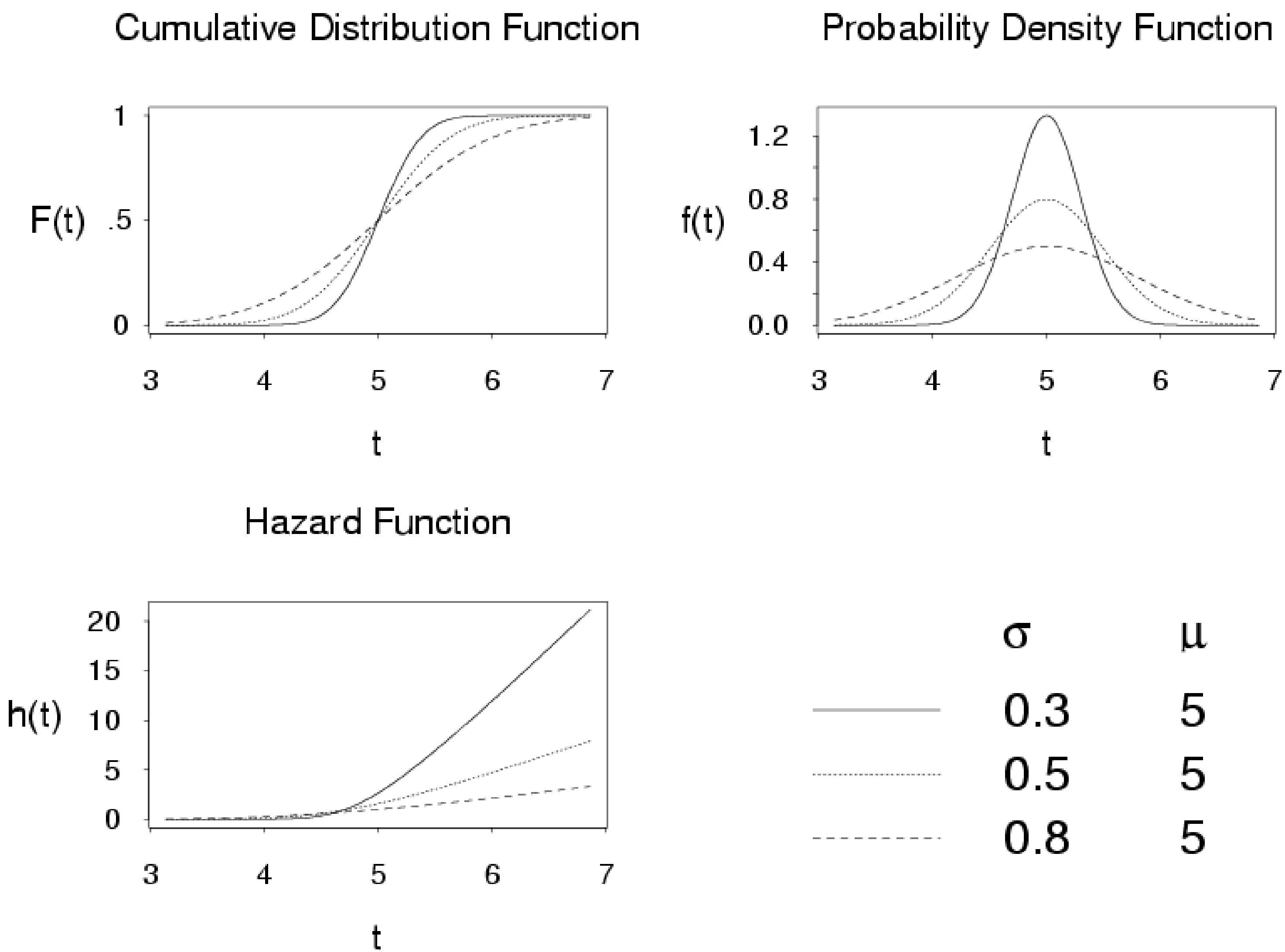
A possible set of probability density, failure, survival, and hazard functions.



A probability density we all know: the normal distribution

- What do you think the hazard looks like for a normal distribution?
- Think of a concrete example. Suppose that times to complete the midterm exam follow a normal curve.
- What's your probability of finishing at any given time given that you're still working on it?

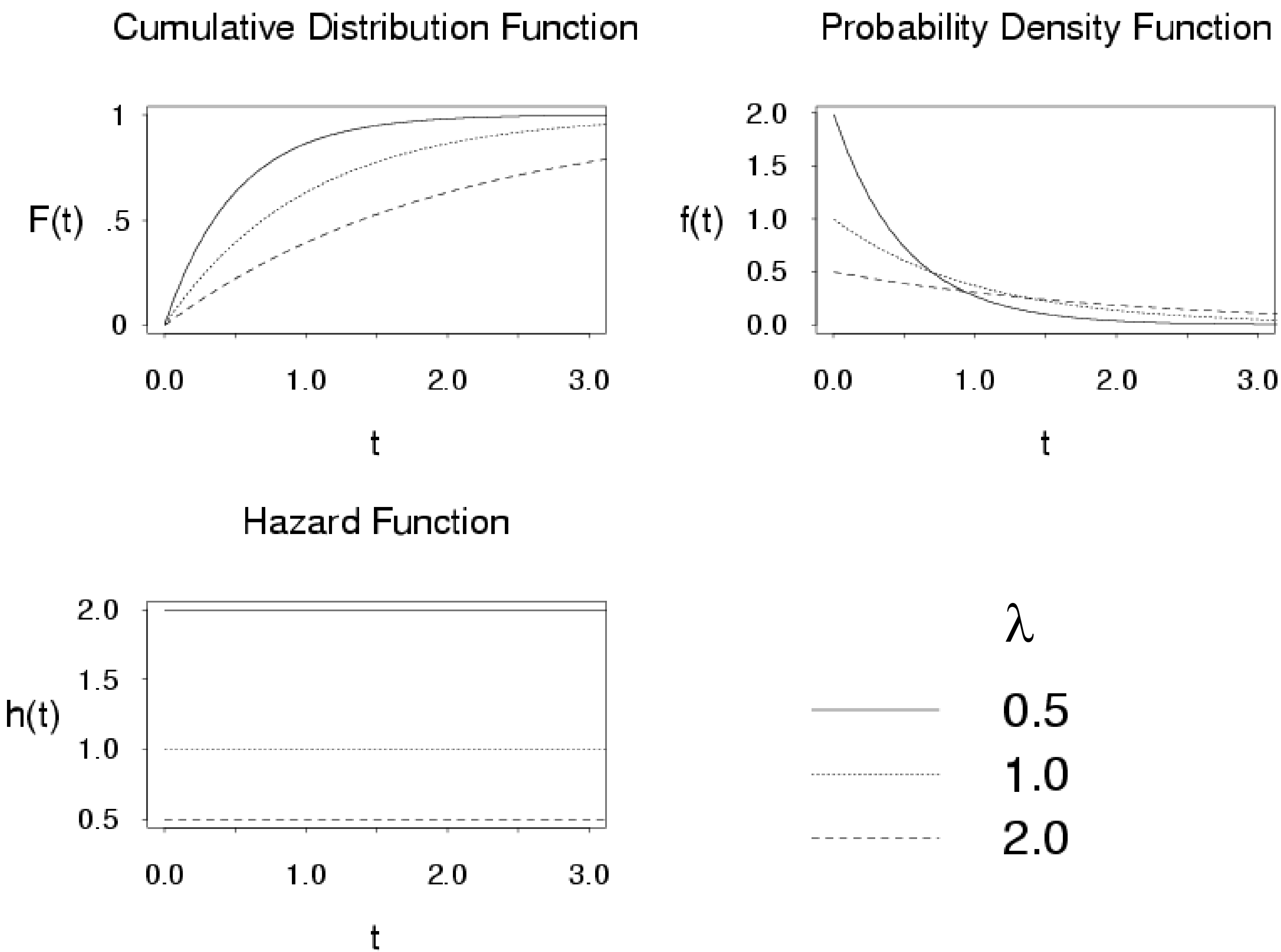
f(t), F(t), S(t), and h(t) for different normal distributions:



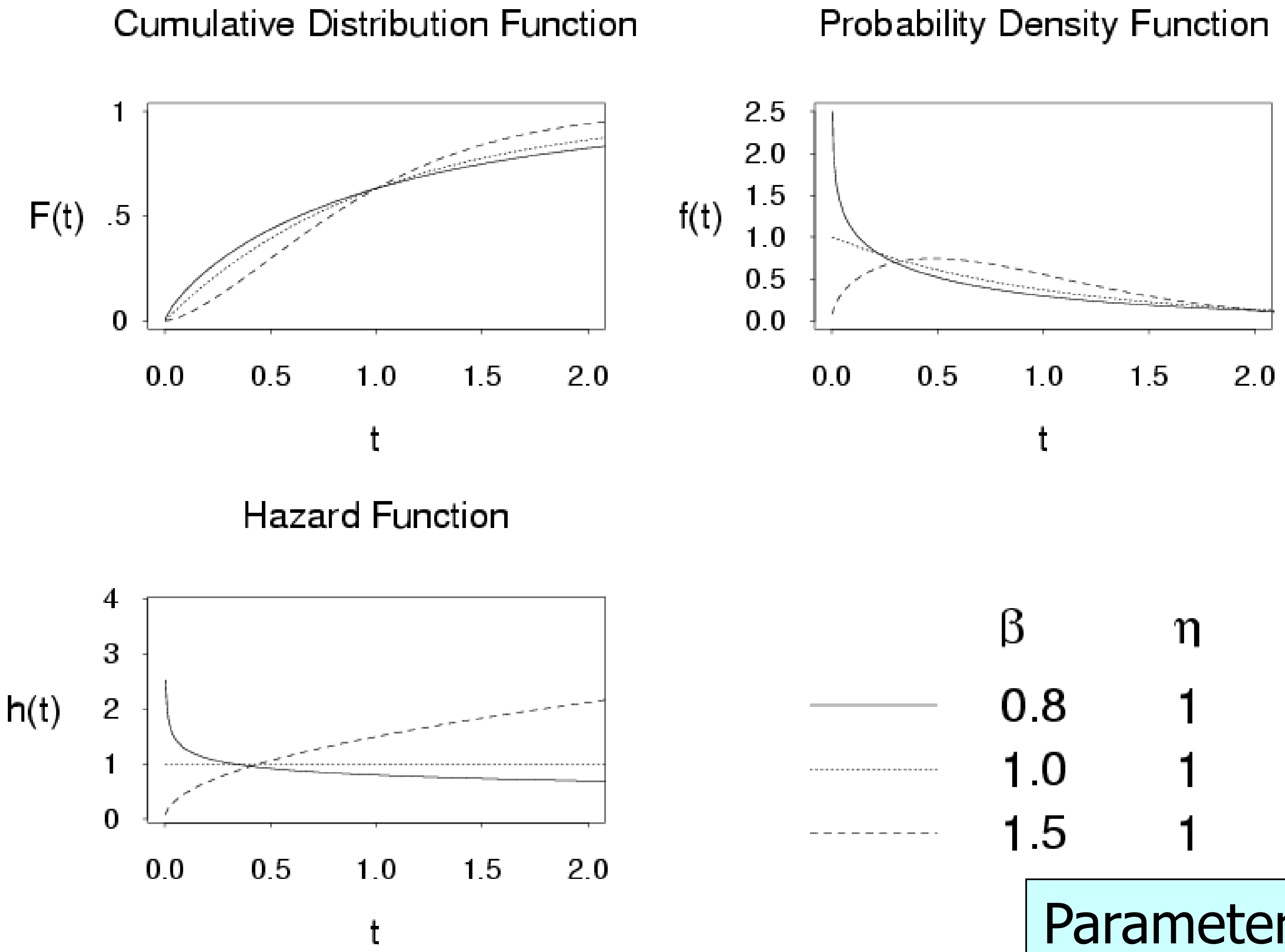
Examples: common functions to describe survival

- Exponential (hazard is constant over time, simplest!)
- Weibull (hazard function is increasing or decreasing over time)

f(t), F(t), S(t), and h(t) for different exponential distributions:



f(t), F(t), S(t), and h(t) for different Weibull distributions:



Parameters of the Weibull distribution

Exponential

Constant **hazard function**: $h(t) = h$

Exponential **density function:** $P(T = t) = f(t) = he^{-ht}$

Survival function:

$$P(T > t) = S(t) = \int_t^{\infty} he^{-hu} du = -e^{-hu}\Big|_t^{\infty} = 0 - -e^{-ht} = e^{-ht}$$

With num

Why isn't the cumulative probability of survival just 90% (rate of .01 for 10 years = 10% loss)?

$$h(t) = .01 \text{ cases/person} - \text{year}$$

Incidence rate (constant).

$$P(t = 10) = .01e^{-.01(10)} = .01e^{-.1} = 0.009$$

Probability of developing disease <u>at year 10.</u>

$$S(t) = e^{-.01t} = 90.5\%$$

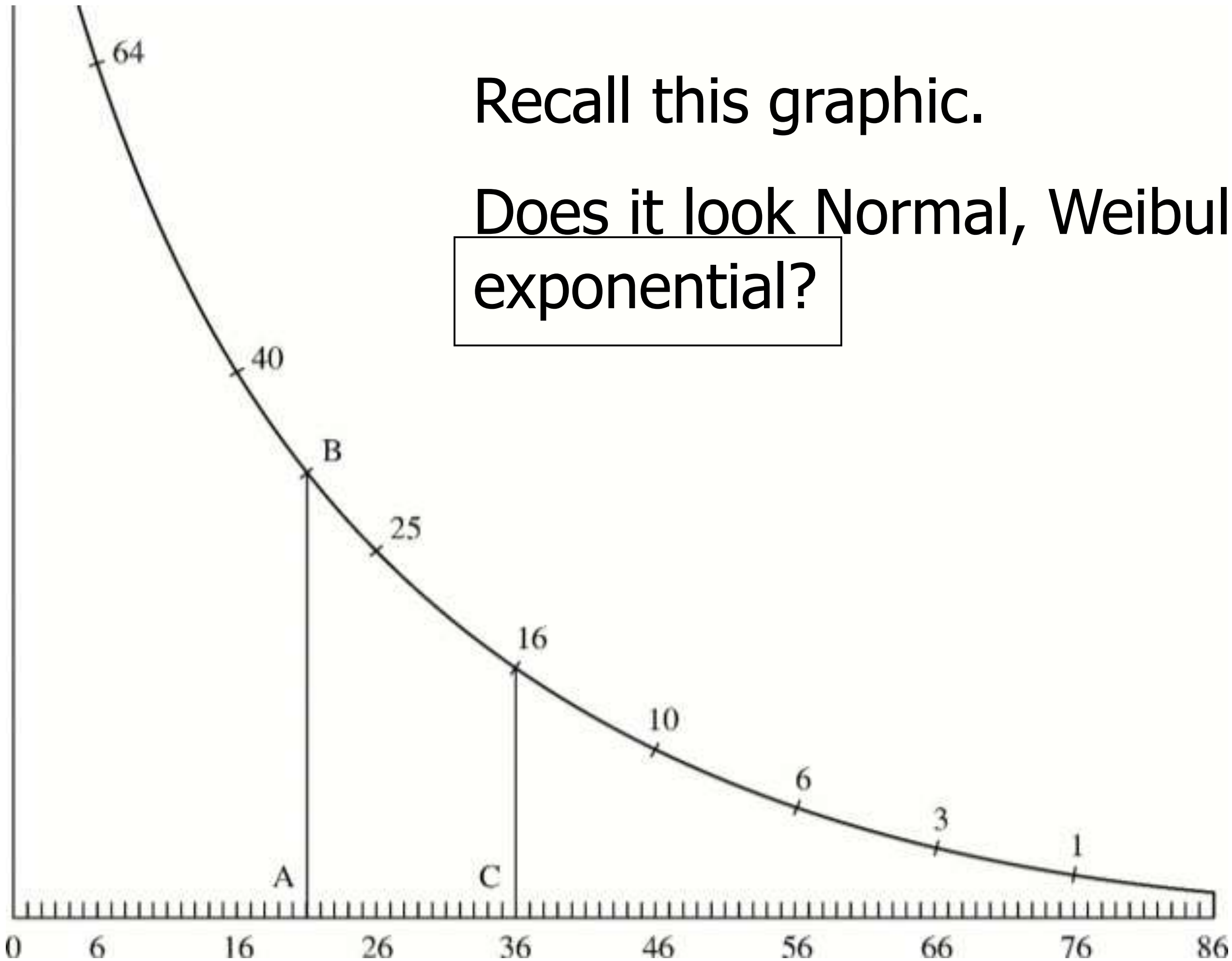
Probability of surviving <u>past year 10.</u>

(cumulative risk through year 10 is 9.5%)

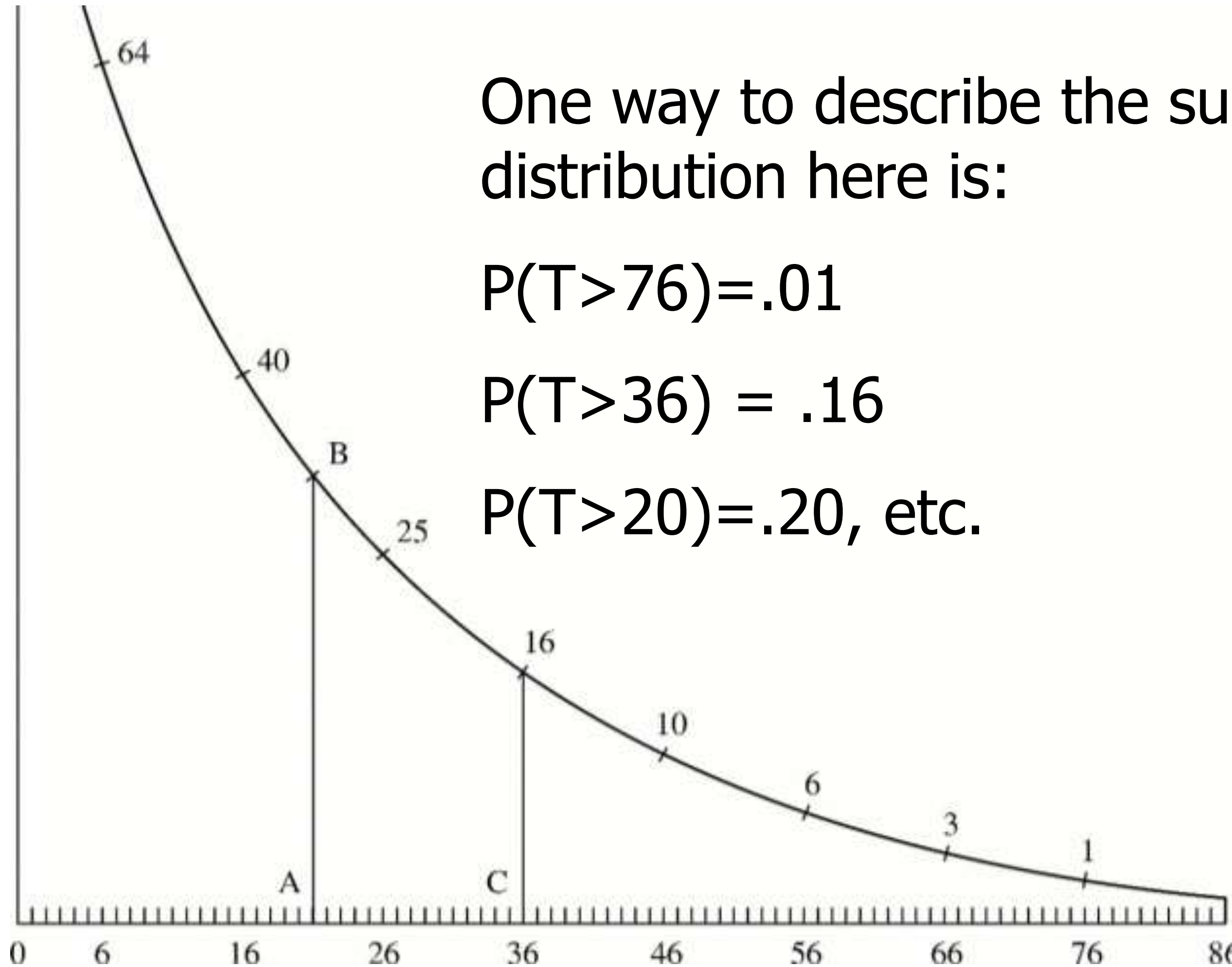
Example…

Recall this graphic.

Does it look Normal, Weibull, exponential?



Example…



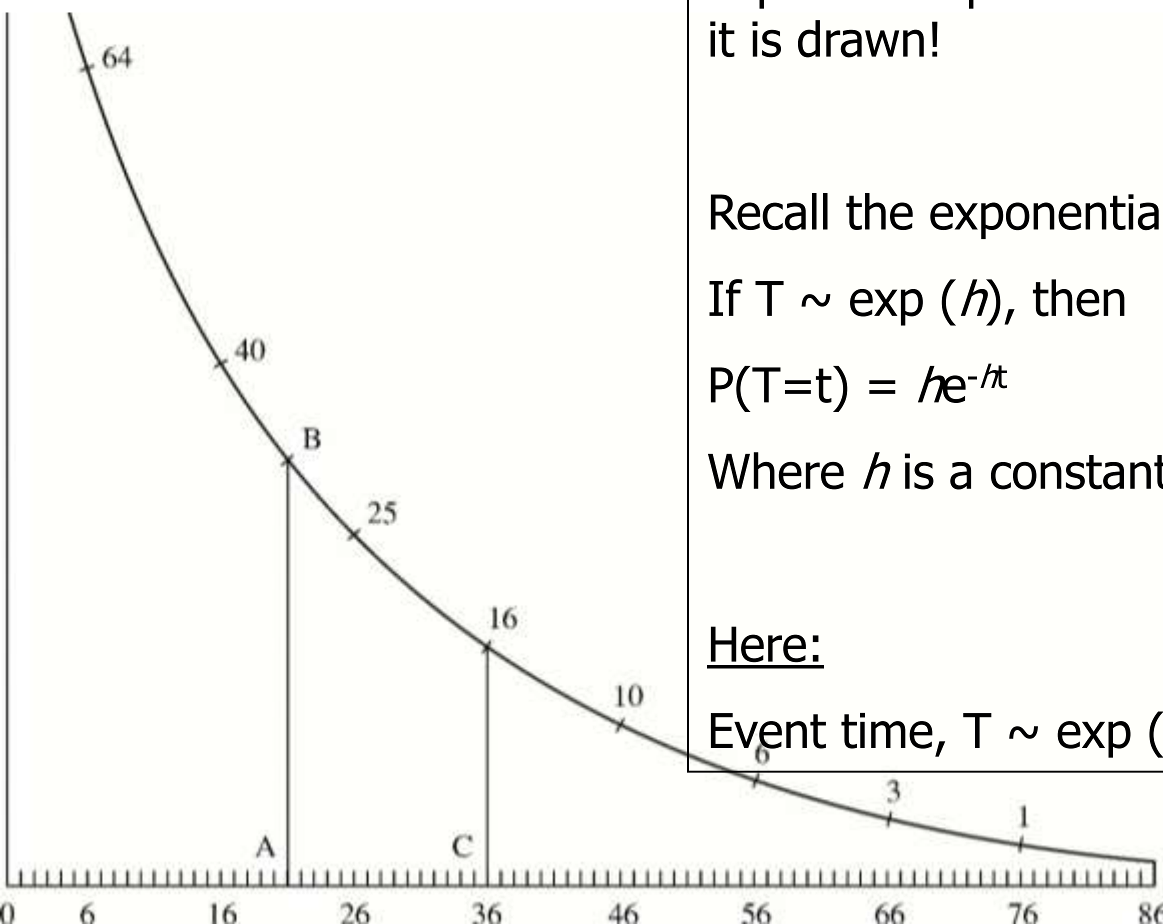
One way to describe the survival distribution here is:

P(T>76)=.01

P(T>36) = .16

P(T>20)=.20, etc.

Example…



Or, more compactly, try to describe this as an exponential probability function—since that is how it is drawn!

Recall the exponential probability distribution:

If T ~ exp (h), then

$P(T=t) = he^{-ht}$

Where h is a constant rate.

<u>Here:</u>

Event time, T ~ exp (Rate)

Example…

To get from the instantaneous probability (density), $P(T=t) = he^{-ht}$, to a cumulative probability of death, integrate:

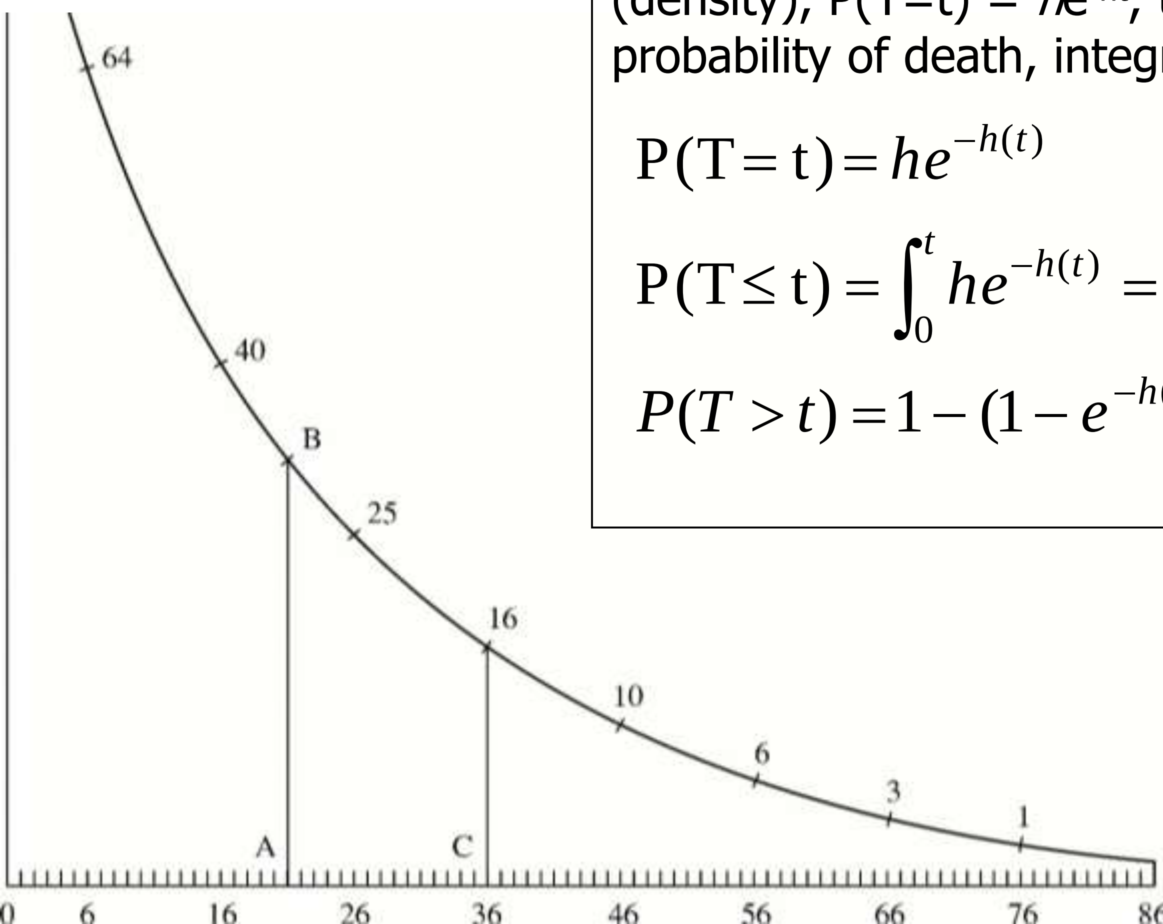
$$P(T=t) = he^{-h(t)}$$

$$P(T \leq t) = \int_0^t he^{-h(t)} = 1 - e^{-h(t)}$$

Area to the left

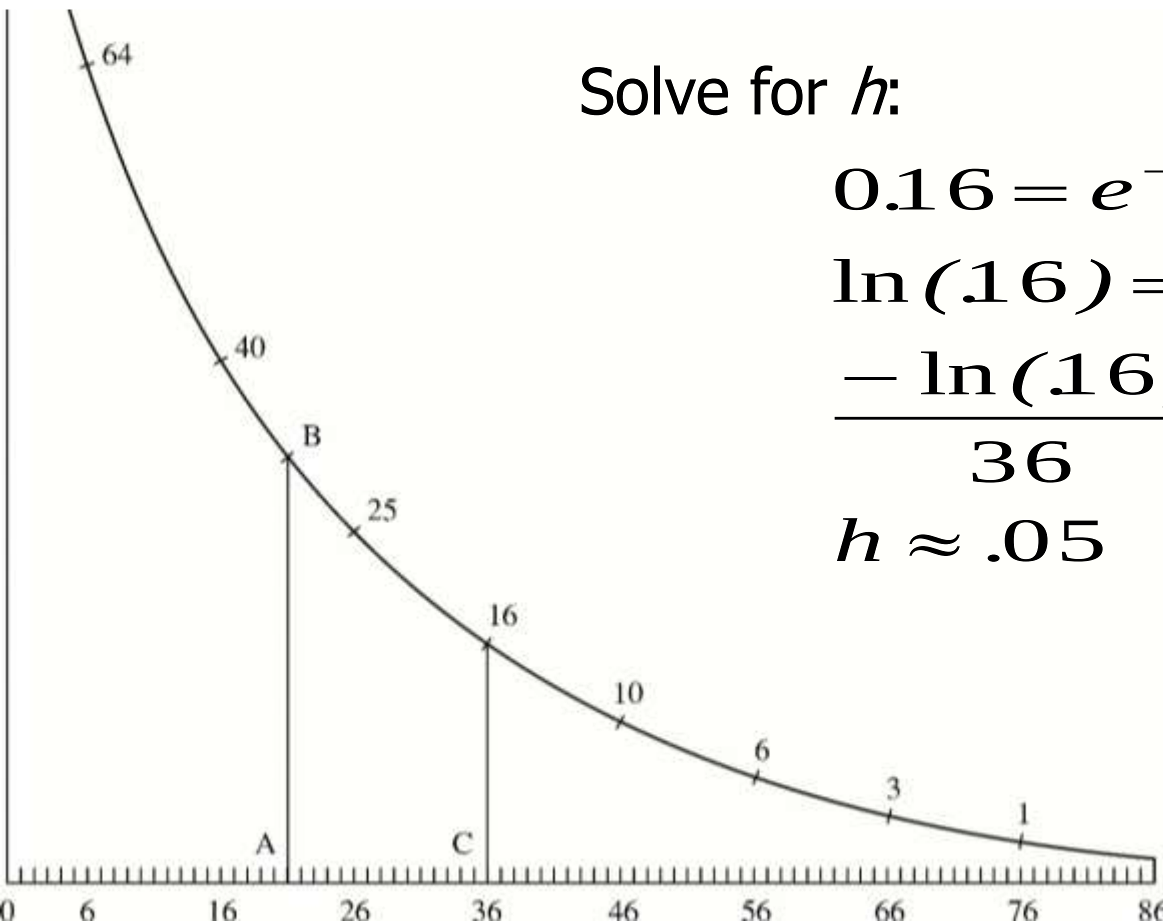
$$P(T > t) = 1 - (1 - e^{-h(t)}) = e^{-h(t)}$$

Area to the right



Example…

$$P(T > age) = e^{-h(age)}$$



Solve for *h*:

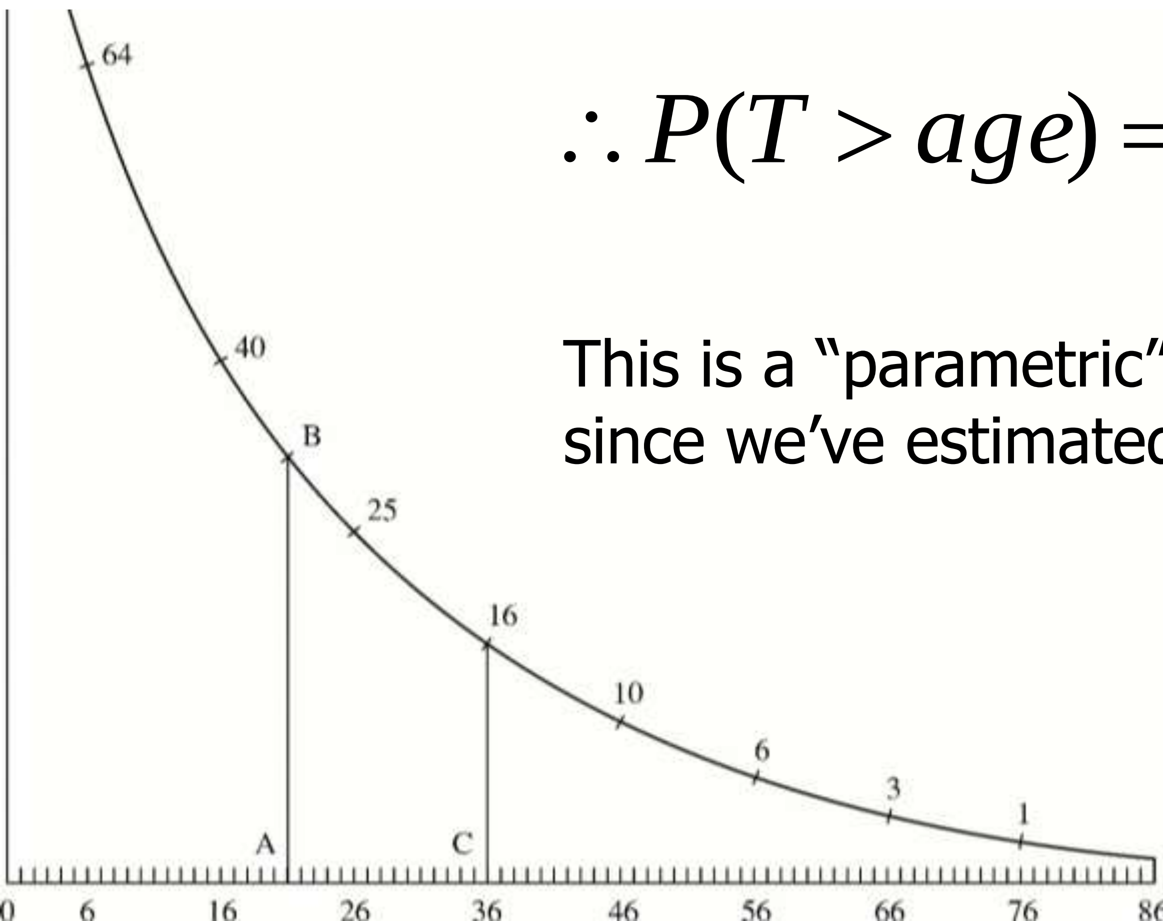
$$0.16 = e^{-h(36)}$$

$$\ln(.16) = -h36$$

$$\frac{-\ln(.16)}{36} = h$$

$$h \approx .05$$

Example…



$$\therefore P(T > age) = e^{-.05(age)}$$

This is a "parametric" survivor function, since we've estimated the parameter *h*.

Hazard rates could also change over time…

$$h(t) = .01 * \mathrm{t}$$

$$\mathrm{h}(5) = .05$$

$$\mathrm{h}(10) = .1$$

Example: Hazard rate increases linearly with time.

Relating these functions (a little calculus just for fun…):

Hazard from density and survival: $h(\mathrm{t}) = \dfrac{f(t)}{S(t)}$

Survival from density: $\mathrm{S}(\mathrm{t}) = \int_t^{\infty} f(u)du$

Density from survival: $\mathrm{f}(t) = -\dfrac{dS(t)}{dt}$

Density from hazard: $\mathrm{f}(t) = h(t)e^{(-\int_0^t h(u)du)}$

Survival from hazard: $\mathrm{S}(\mathrm{t}) = e^{(-\int_0^t h(u)du)}$

Hazard from survival: $h(\mathrm{t}) = -\dfrac{\mathrm{d}}{\mathrm{dt}} \ln S(t)$

Getting density from hazard…

$$h(t) = .01 * \mathrm{t}$$

$$\mathrm{h}(5) = .05$$

$$\mathrm{h}(10) = .1$$

Example: Hazard rate increases linearly with time.

$$\text{Density from hazard: } \mathrm{f}(t) = h(t)e^{(-\int_0^t h(u)du)}$$

$$\mathrm{f}(t) = .01 * te^{(-\int_0^t .01tdu)} = .01(t)e^{-\int_0^t .01udu} = .01(t)e^{-.005t^2}$$

$$f(t=5) = .01(5)e^{-.005(25)} = .05e^{-.125} = .044$$

$$f(t=10) = .1(10)e^{-.005(100)} = .1e^{-.5} = .06$$

Getting survival from hazard…

$$h(t) = .01 * \mathrm{t}$$

$$\mathrm{h}(10) = .1$$

$$\mathrm{h}(5) = .05$$

$$\text{Survival from hazard: } \mathrm{S}(\mathrm{t}) = e^{(-\int_0^t h(u)du)}$$

$$\mathrm{S}(\mathrm{t}) = e^{(-\int_0^t .01u\,du)} = e^{-.005t^2}$$

$$S(10) = e^{-.005(100)} = .60$$

$$S(5) = e^{-.005(25)} = .88$$

Parametric regression techniques

- Parametric multivariate regression techniques:
  - Model the underlying hazard/survival function
  - Assume that the dependent variable (time-to-event) takes on some known distribution, such as Weibull, exponential, or lognormal.
  - Estimates *parameters* of these distributions (e.g., baseline hazard function)
  - Estimates covariate-adjusted hazard ratios.
    - A hazard ratio is a ratio of hazard rates

Many times we care more about comparing groups than about estimating absolute survival.

The model: parametric reg.

Components:

- A baseline hazard function (which may change over time).
- A linear function of a set of k fixed covariates that when exponentiated gives the relative risk.

Exponential model assumes fixed baseline hazard that we can estimate.

$$\log h_i(t) = \mu + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$$

Weibull model models the baseline hazard as a function of time. Two parameters (shape and scale) must be estimated to describe the underlying hazard function over time.

$$\log h_i(t) = \mu + \alpha \log t + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$$

The model

When exponentiated, risk factor coefficients from both models give hazard ratios (relative risk).

Components:

- A baseline hazard functio[n]
- A linear function of a set of k fixed covariates that when exponentiated gives the relative risk.

$$\log h_i(t) = \mu + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$$

$$\log h_i(t) = \mu + \alpha \log t + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$$

Cox Regression

- Semi-parametric
- Cox models the effect of predictors and covariates on the hazard rate but leaves the baseline hazard rate unspecified.
- Also called proportional hazards regression
- Does NOT assume knowledge of absolute risk.
- Estimates *relative* rather than *absolute* risk.

The model: Cox regression

Components:

•A baseline hazard function that is left unspecified but must be positive (=the hazard when all covariates are 0)

•A linear function of a set of k fixed covariates that is exponentiated. (=the relative risk)

$$\log h_i(t) = \boxed{\log h_0(t)} + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$$

Can take on any form

$$h_i(t) = \boxed{h_0(t)} e^{\beta_1 x_{i1} + \ldots + \beta_k x_{ik}}$$

The model

The point is to compare the hazard rates of individuals who have different covariates:

Hence, called *Proportional* hazards:

$$HR = \frac{h_1(t)}{h_2(t)} = \frac{\cancel{h_0(t)}e^{\beta x_1}}{\cancel{h_0(t)}e^{\beta x_2}} = e^{\beta(x_1 - x_2)}$$

Hazard functions should be strictly parallel.

Introduction to Kaplan-Meier

Non-parametric estimate of the survival function:

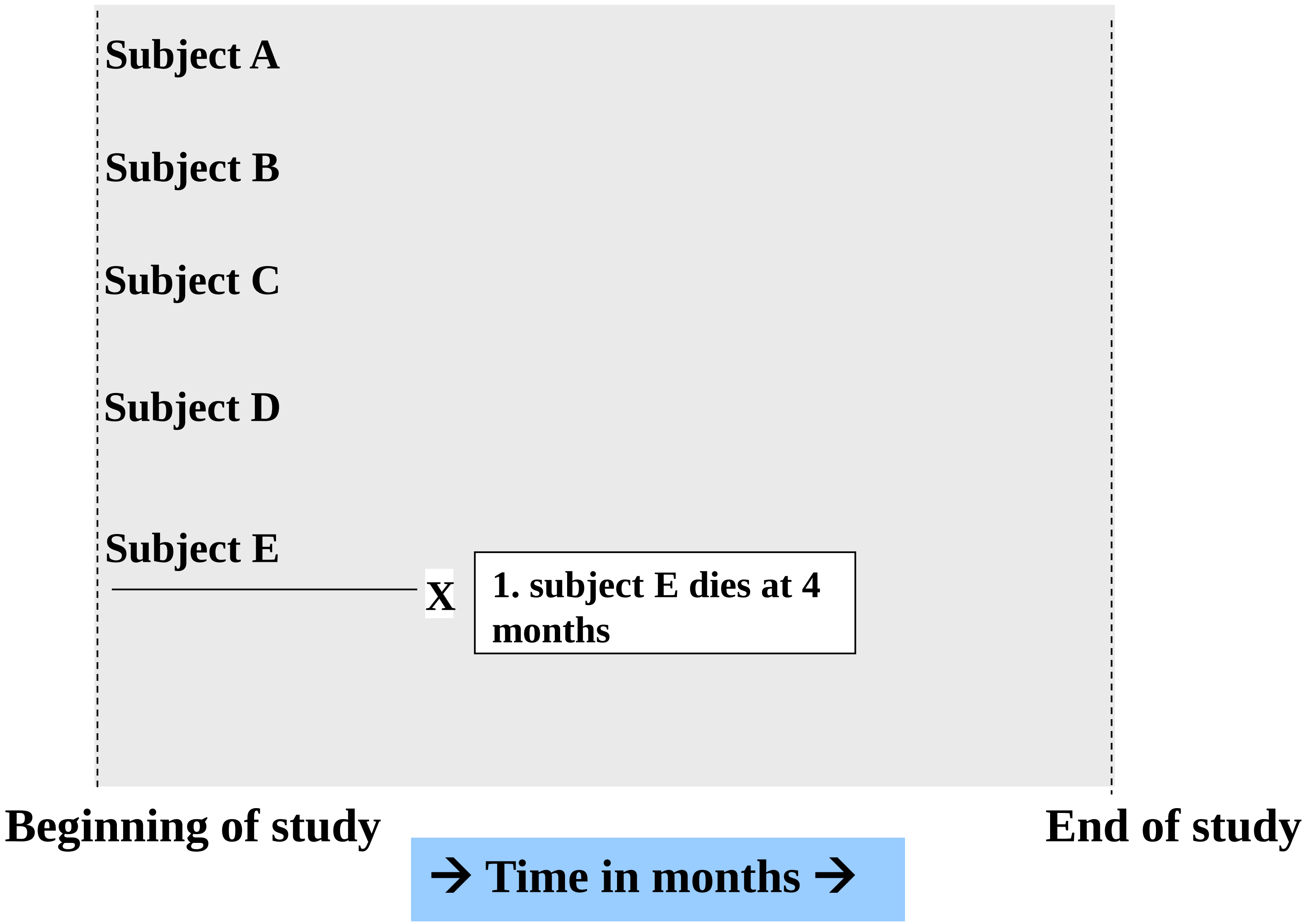
No math assumptions! (either about the underlying hazard function or about proportional hazards).

Simply, the empirical probability of surviving past certain times in the sample (taking into account censoring).

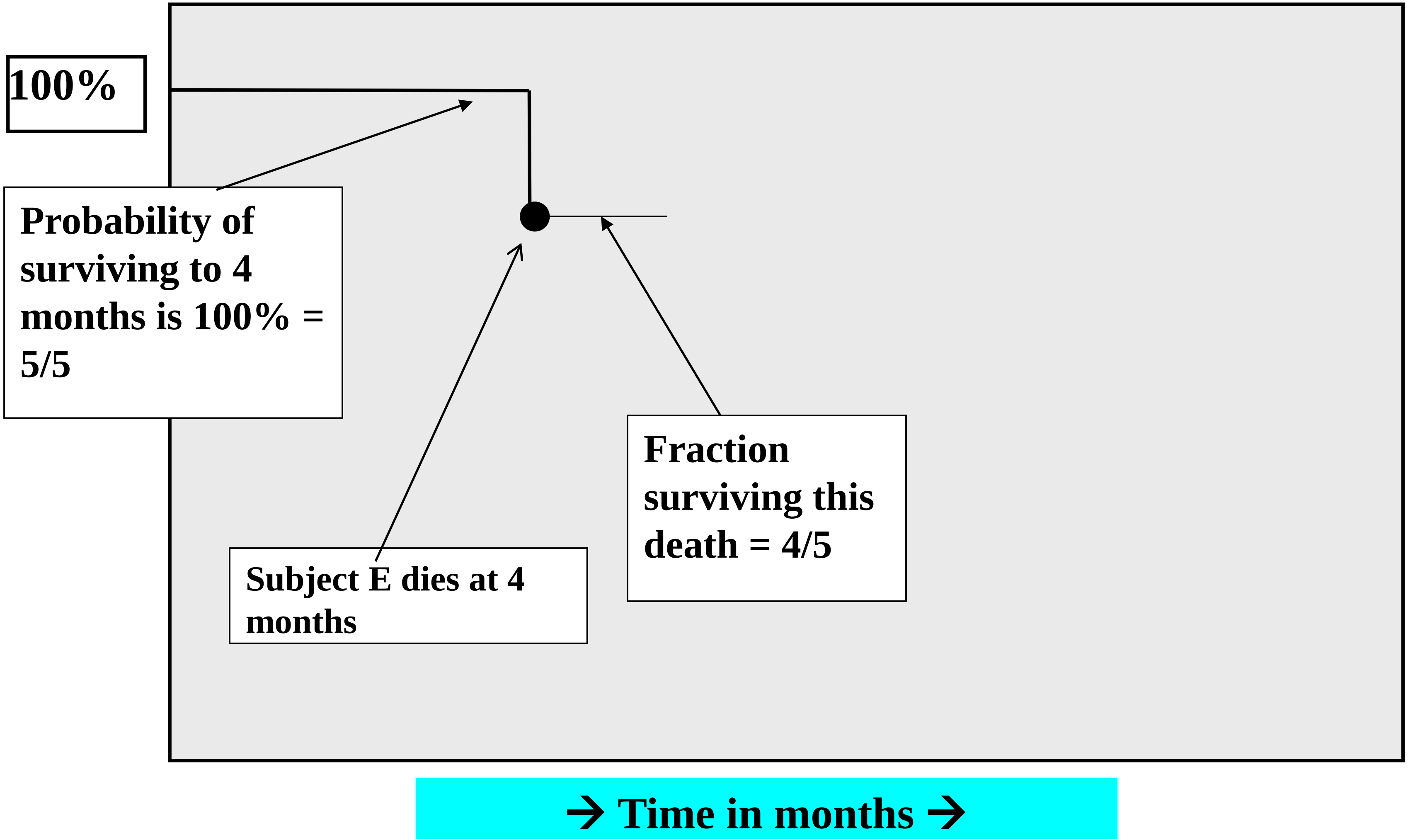
Introduction to Kaplan-Meier

- Non-parametric estimate of the survival function.
- Commonly used to describe survivorship of study population/s.
- Commonly used to compare two study populations.
- Intuitive graphical presentation.

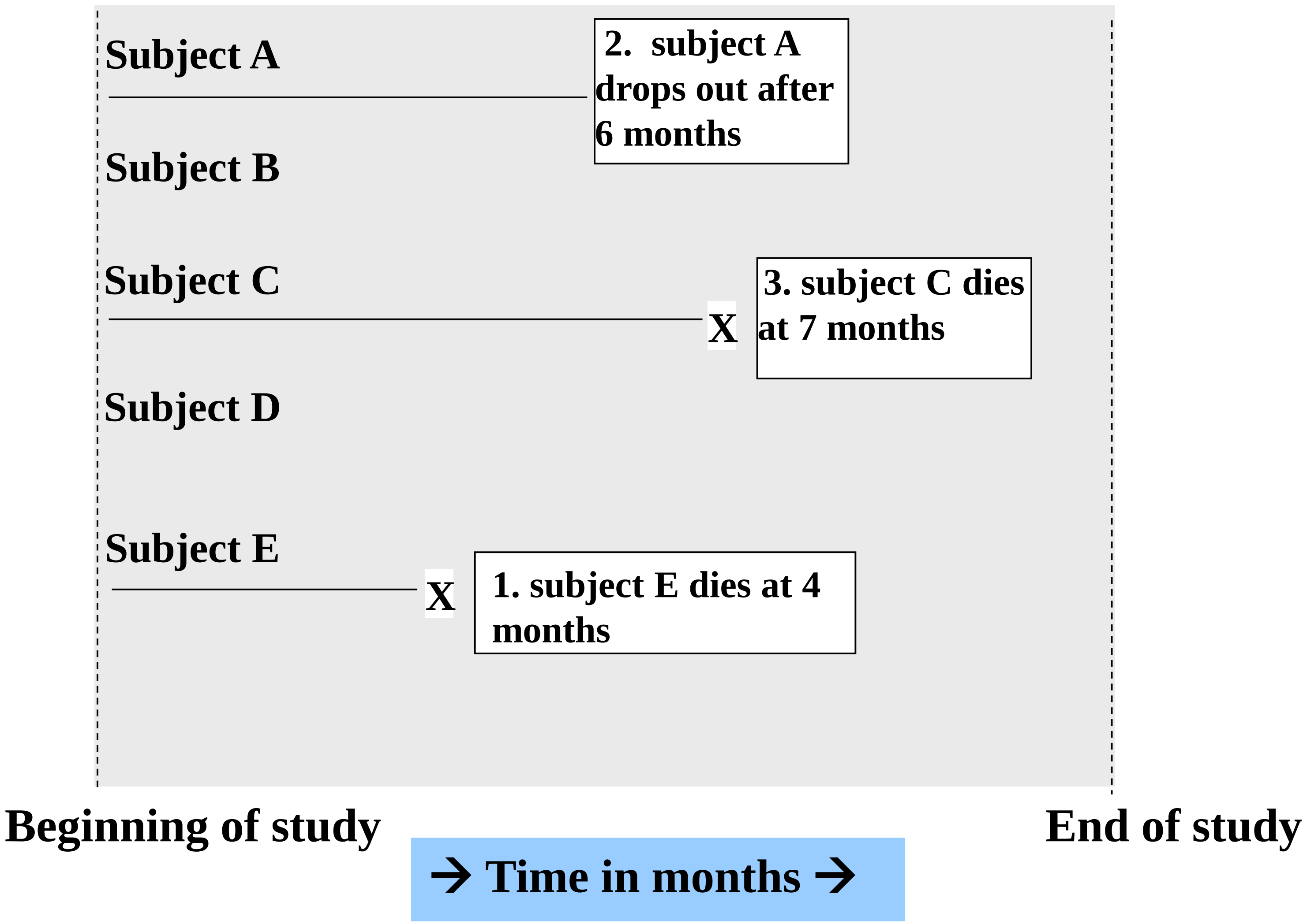
Survival Data (right-censored)



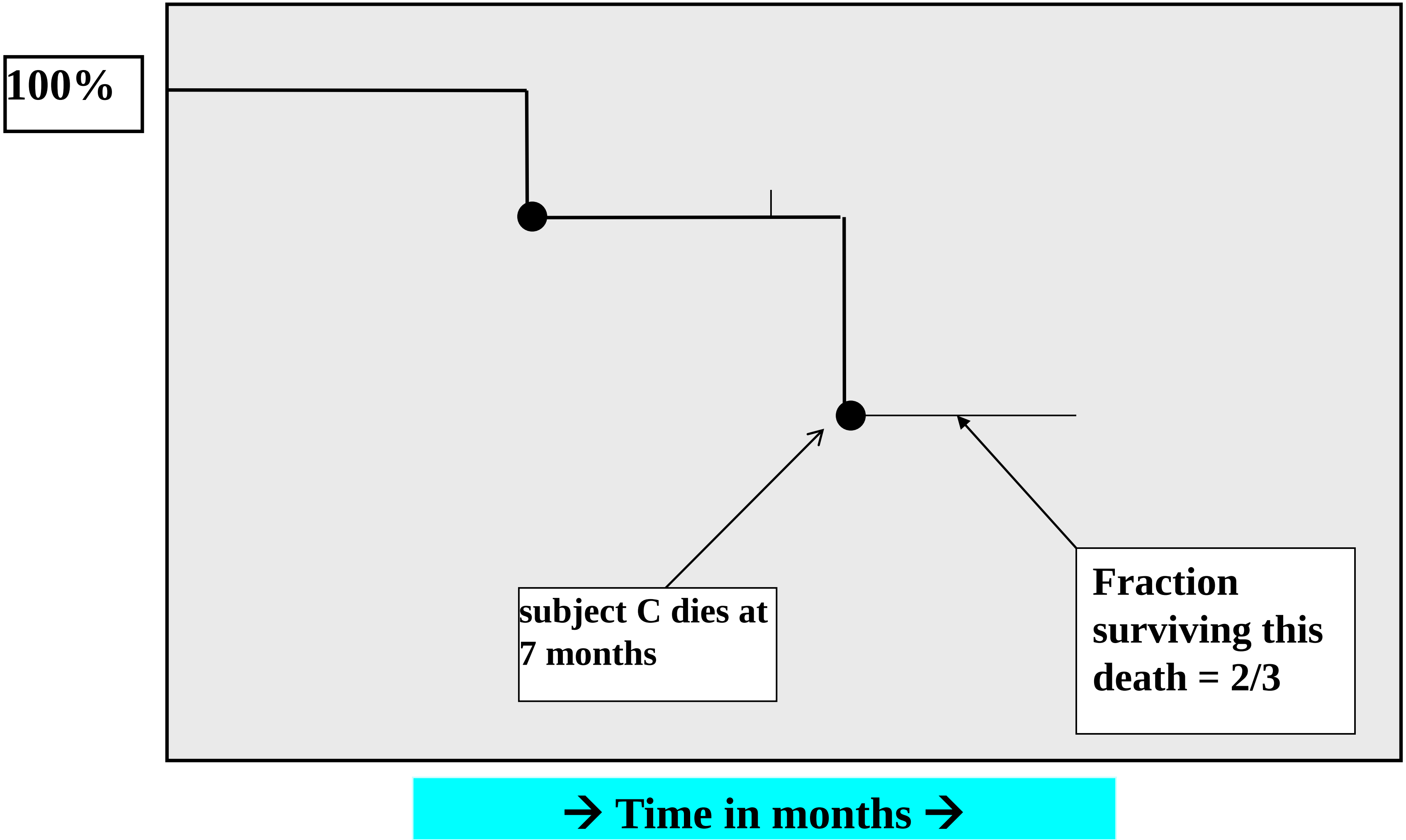
Corresponding Kaplan-Meier Curve



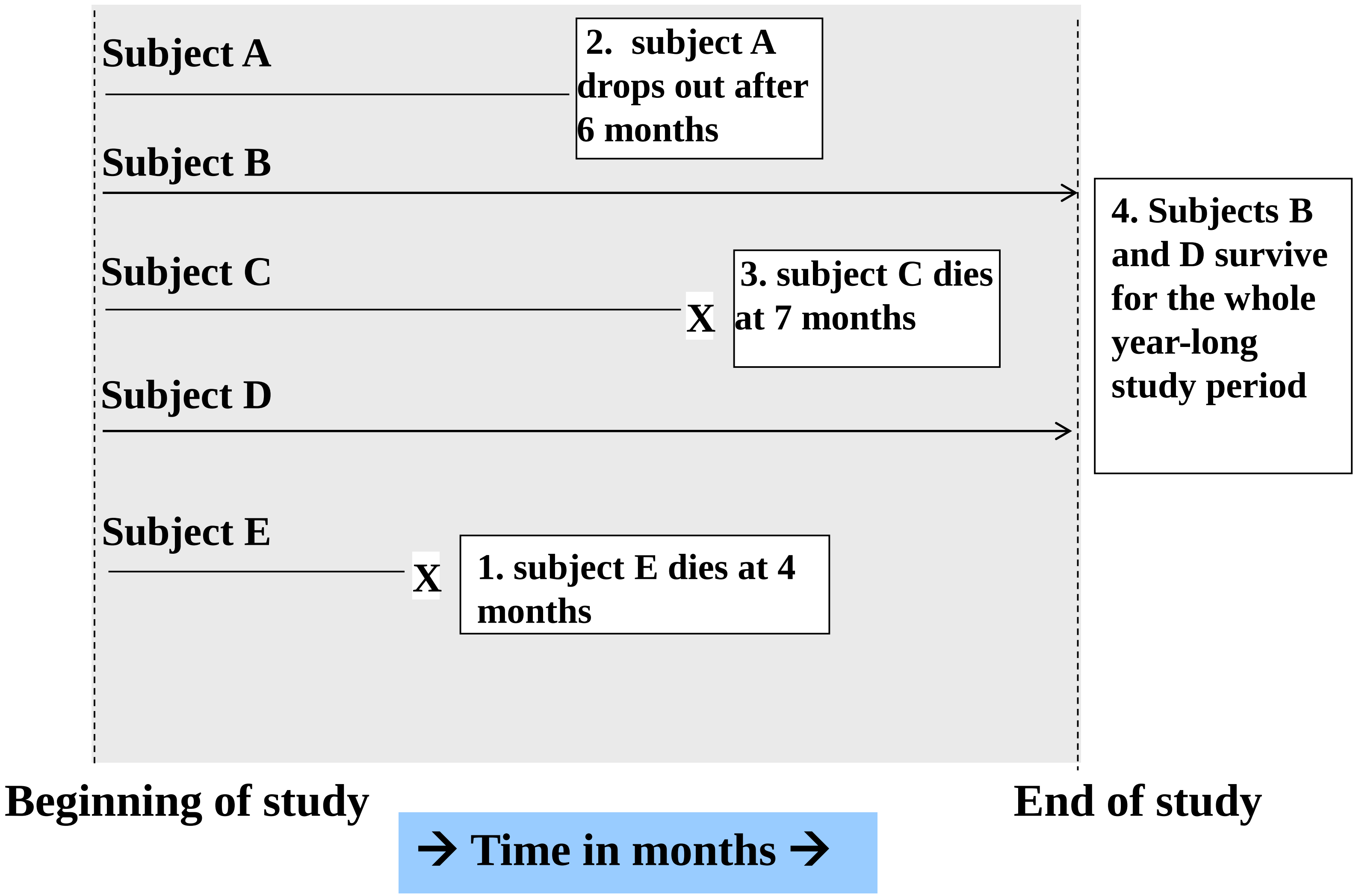
Survival Data



Corresponding Kaplan-Meier Curve

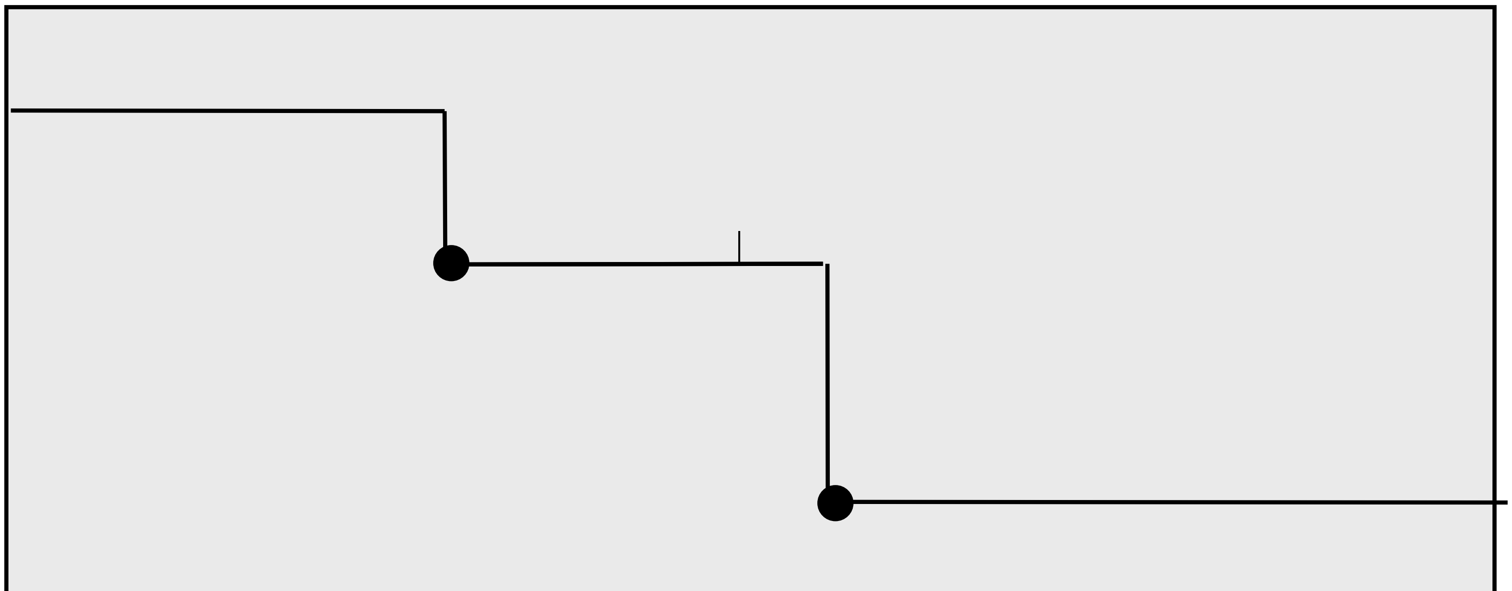


Survival Data



Corresponding Kaplan-Meier Curve

100%



Rule from prob
P(A&B)=P(A)*
In survival ana
P(surviving int

∴Product limit estimate of survival =
P(surviving interval 1/at-risk up to failure 1) *
P(surviving interval 2/at-risk up to failure 2)
= 4/5 * 2/3= .5333

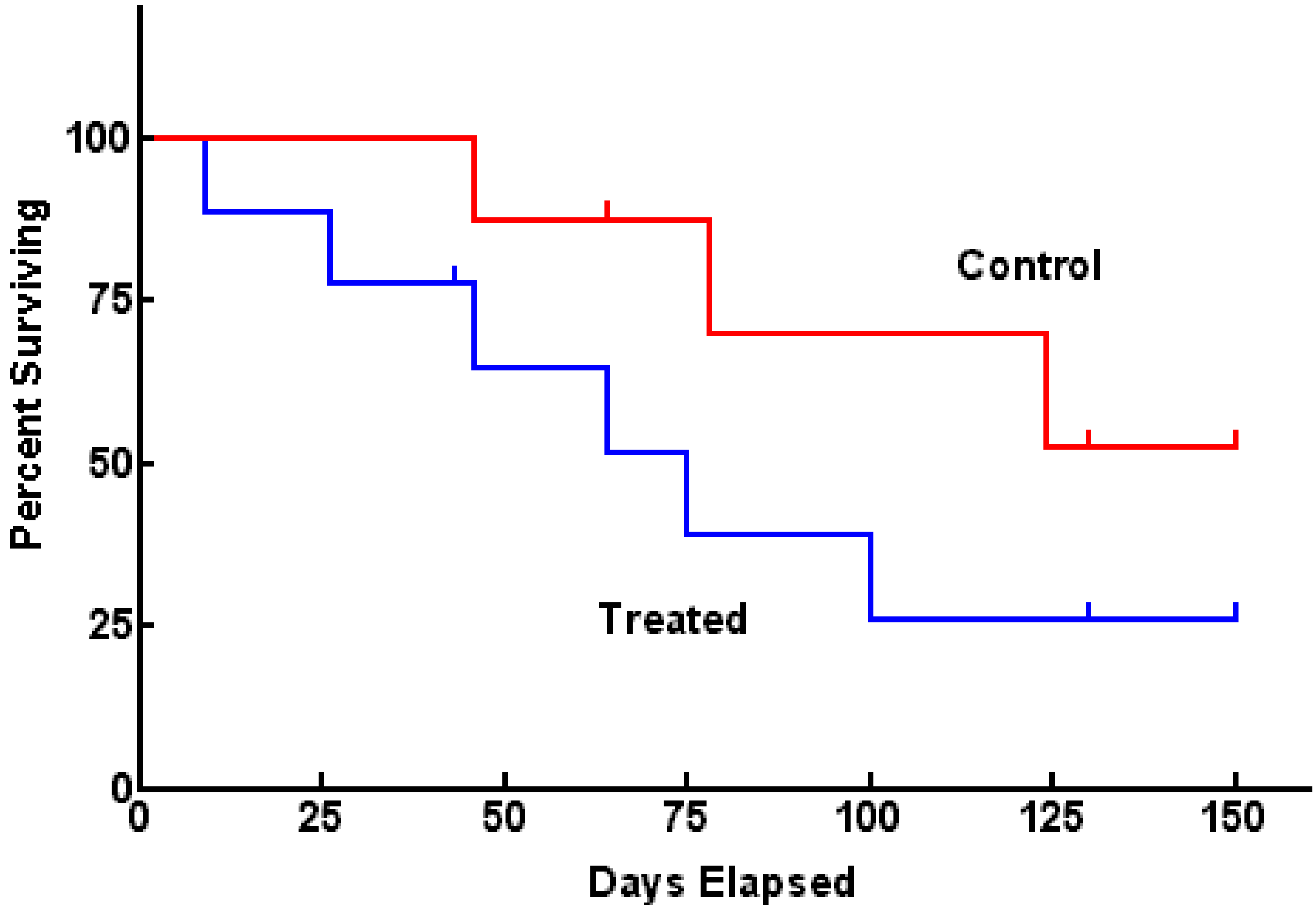
→ Time in months →

The product limit estimate

- The probability of surviving in the entire year, taking into account censoring
- = (4/5) (2/3) = 53%

- NOTE: > 40% (2/5) because the one drop-out survived at least a portion of the year.
- AND <60% (3/5) because we don't know if the one drop-out would have survived until the end of the year.

Comparing 2 groups

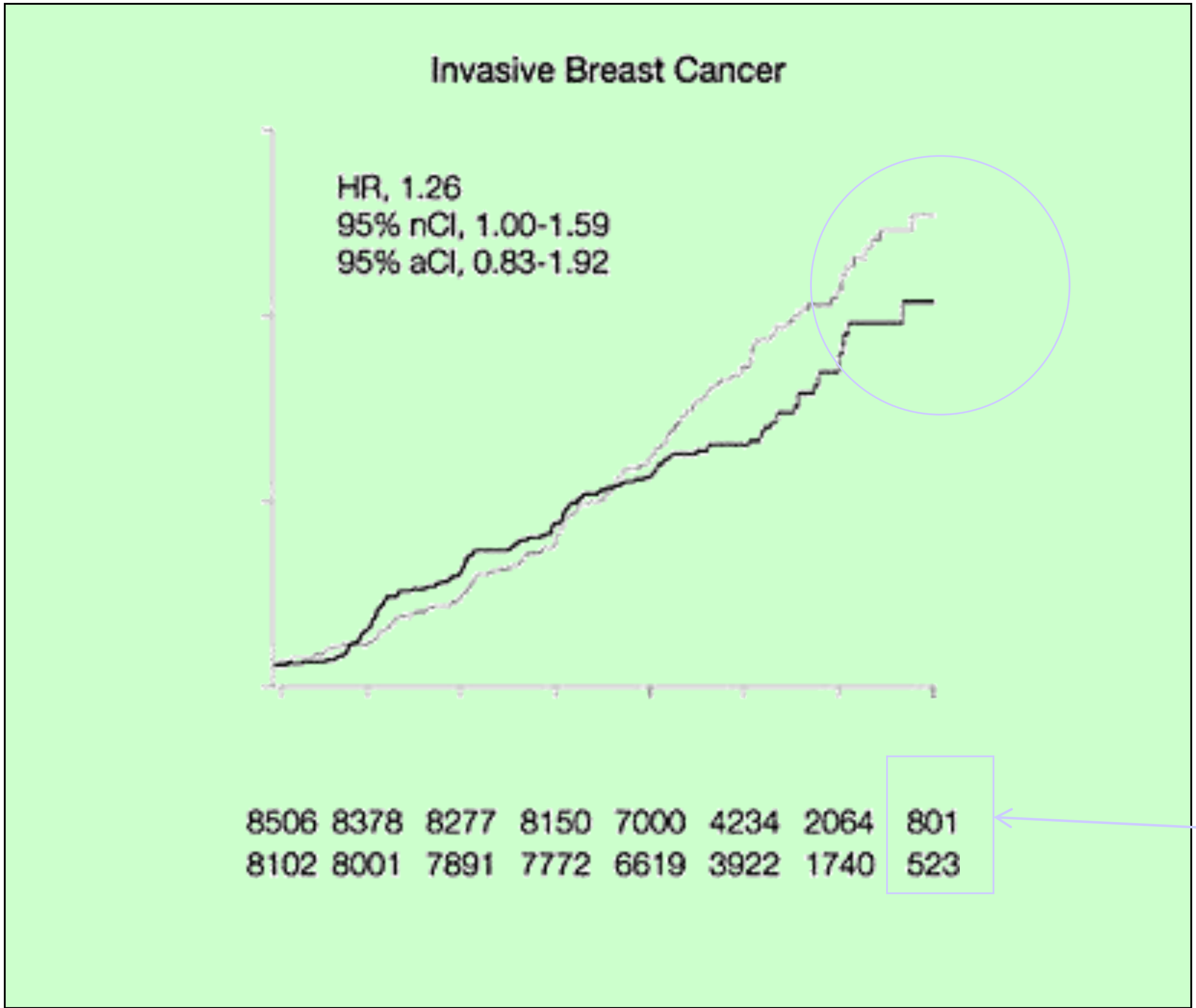


Use log-rank test to test the null hypothesis of no difference between survival functions of the two groups (more on this next time)

Caveats

- Survival estimates can be unreliable toward the end of a study when there are small numbers of subjects at risk of having an event.

WHI and breast cancer



Small numbers left

Women's Health Initiative Writing Group. *JAMA*. 2002;288:321-333.

Limitations of Kaplan-Meier

- Mainly descriptive
- Doesn’t control for covariates
- Requires categorical predictors
- Can’t accommodate time-dependent variables

References

Paul Allison. *Survival Analysis Using SAS*. SAS Institute Inc., Cary, NC: 2003.