

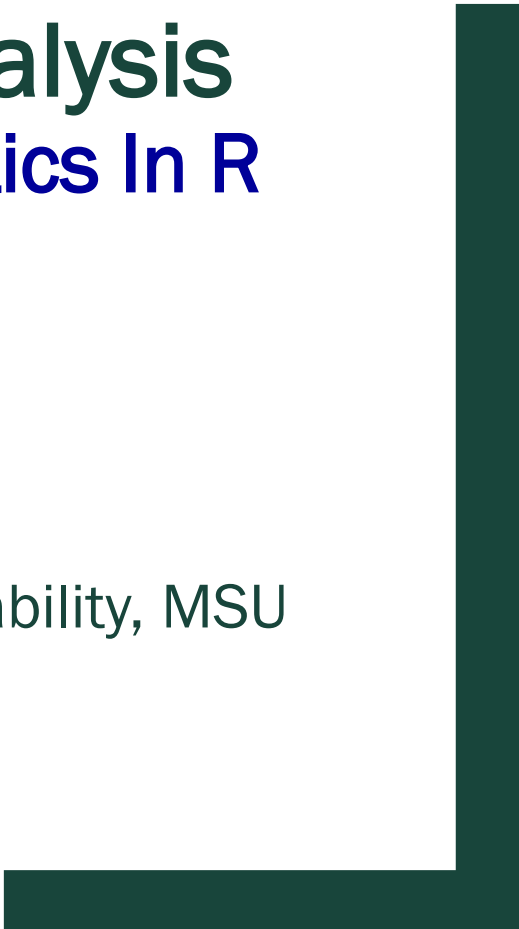


Basic Statistical Analysis

Translational Bioinformatics In R

Workshop

Yuehua Cui
Department of Statistics and Probability, MSU
(cuiy@msu.edu)



Topics to cover

- Correlation Analysis
- Linear Regression and Logistic Regression
- Confounding Factor
- P-value and FDR
- Survival Analysis

1. Correlation Analysis

1.1 Continuous data

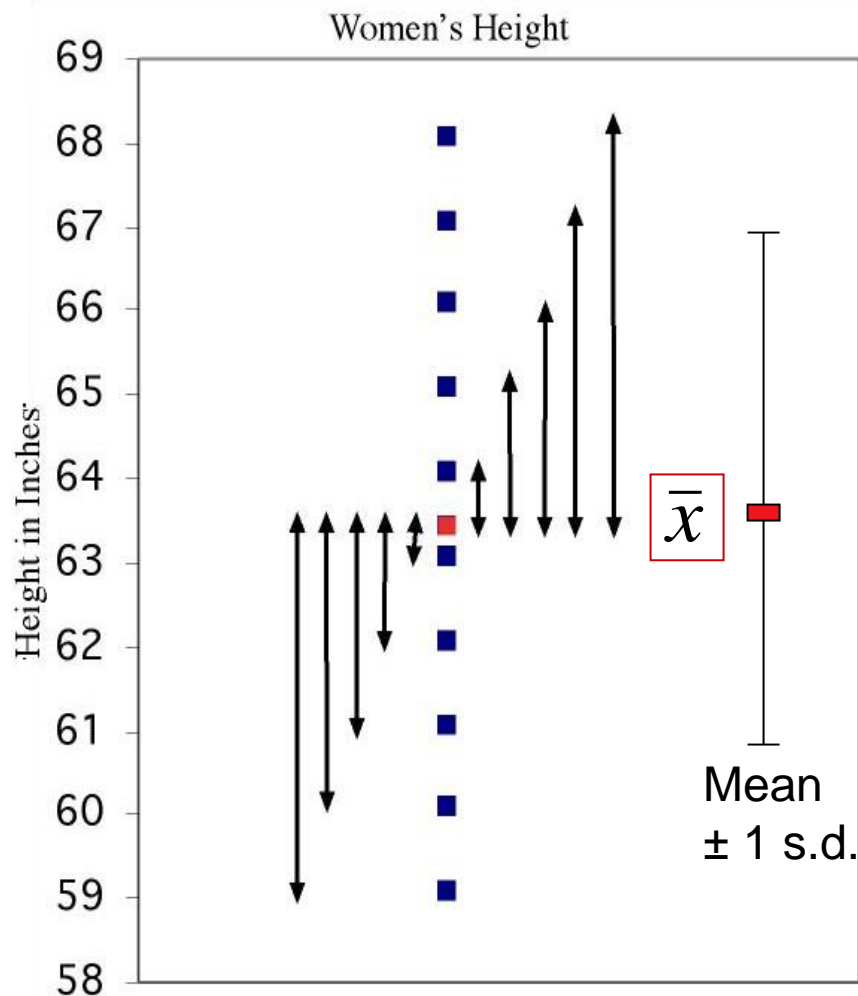
- Pearson correlation
- Spearman correlation

1.2 Categorical data

- Fisher test
- Chi-square test

Measure of spread: the standard deviation

The standard deviation “s” is used to describe the variation around the mean. Like the mean, it is not resistant to skew or outliers.



1. First calculate the **variance s^2** .

$$s^2 = \frac{1}{n-1} \sum_1^n (x_i - \bar{x})^2$$

2. Then take the square root to get the **standard deviation s** .

$$s = \sqrt{\frac{1}{n-1} \sum_1^n (x_i - \bar{x})^2}$$

1.1 Continuous data: Pearson correlation

- This was introduced by Karl Pearson(1867-1936)
- The correlation coefficient is a measure of the direction and strength of a linear relationship.
- It is calculated using the mean and the standard deviation of both the x and y variables.

$$r = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

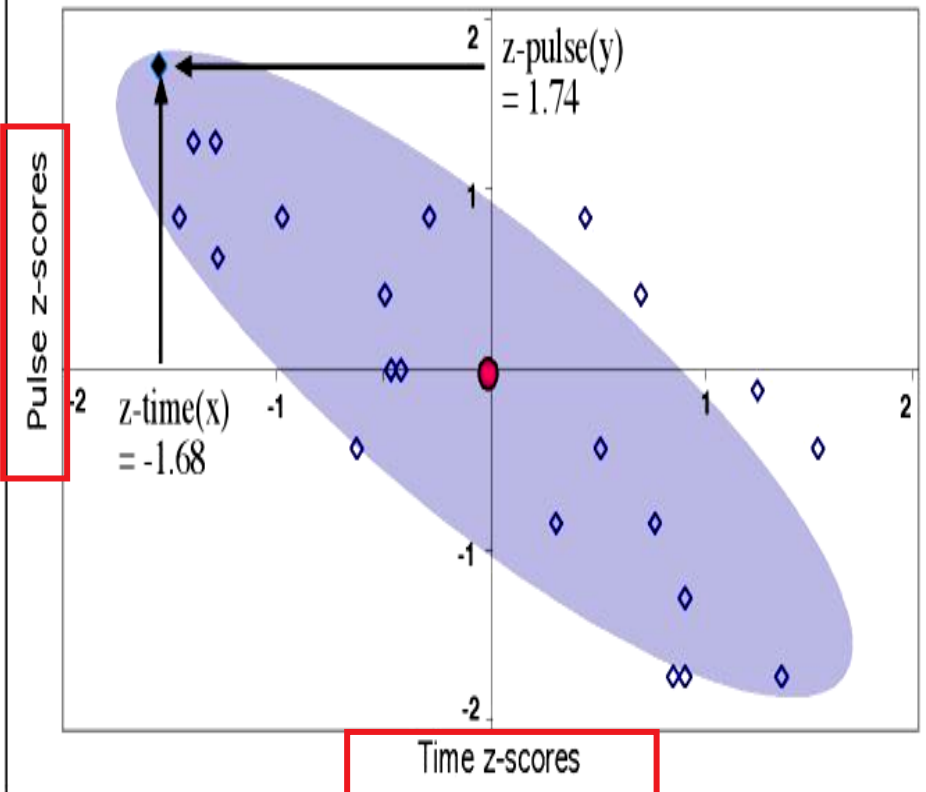
- Correlation can only be used to describe quantitative variables. Categorical variables don't have means and standard deviations.



$$r = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

z for time
z for pulse

Product of z-scores for this point =
 (z-pulse)(z-time) = (-1.74)(1.68) = -2.92

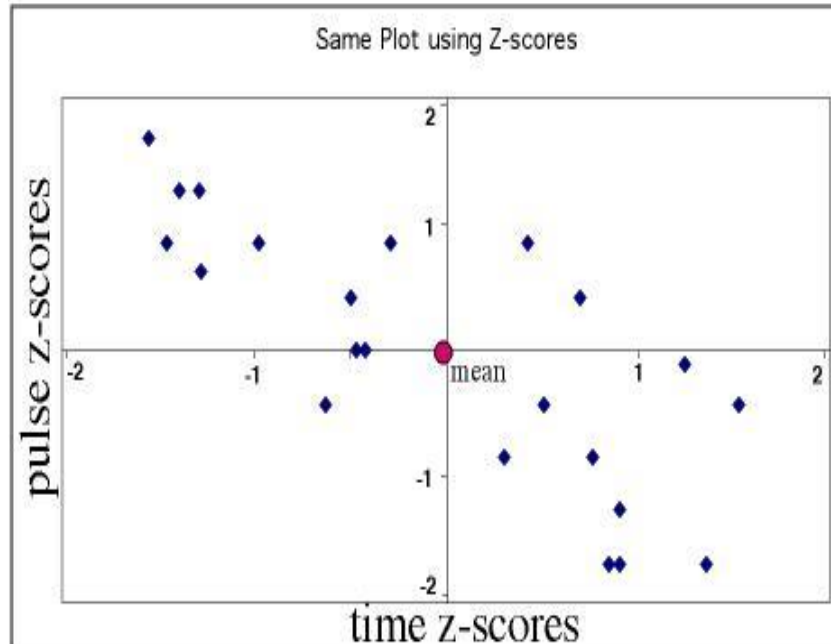
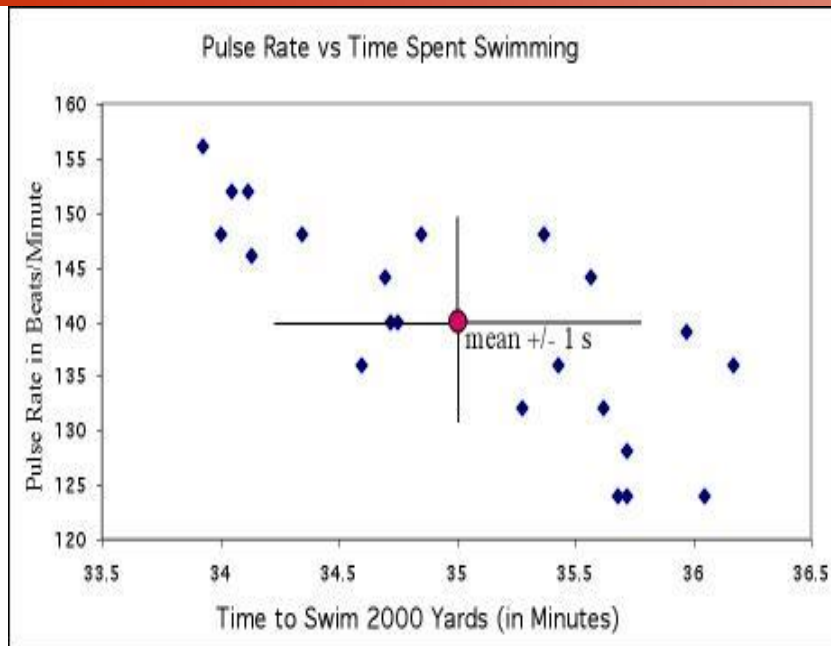


Part of the calculation involves finding z, the standardized score we used when working with the normal distribution.

Standardization:

Allows us to compare correlations between data sets where variables are measured in different units or when variables are different.

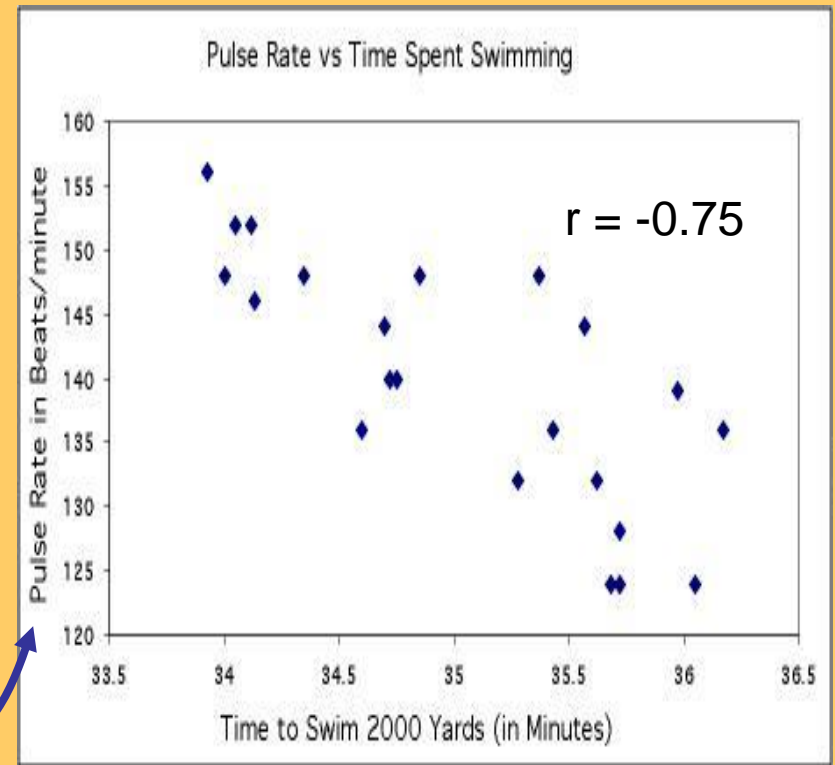
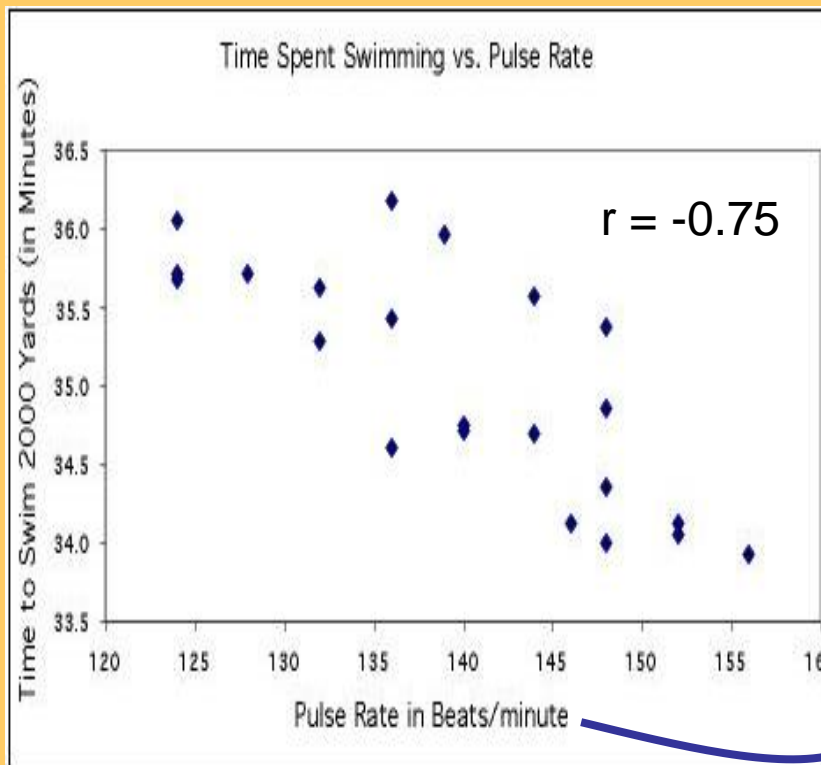
For instance, we might want to compare the correlation between [swim time and pulse], with the correlation between [swim time and breathing rate].



“r” does not distinguish x & y

The correlation coefficient, r, treats x and y symmetrically.

$$r = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$



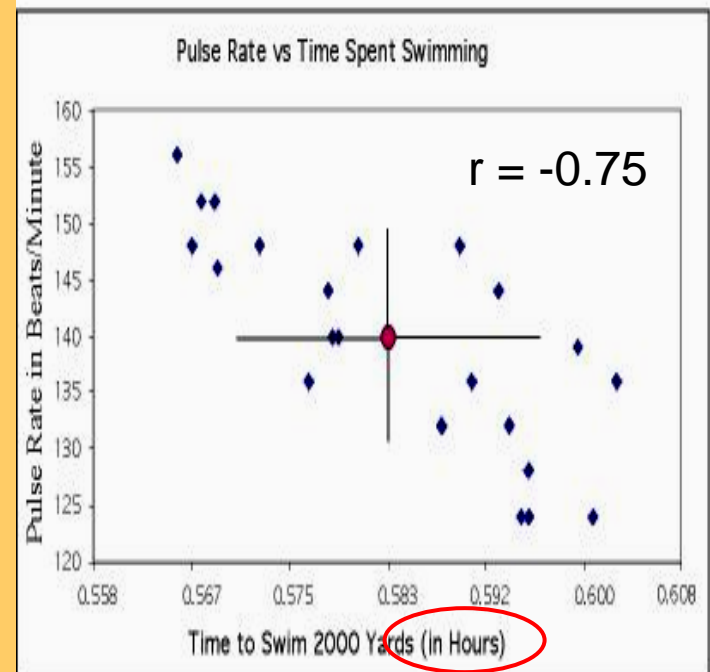
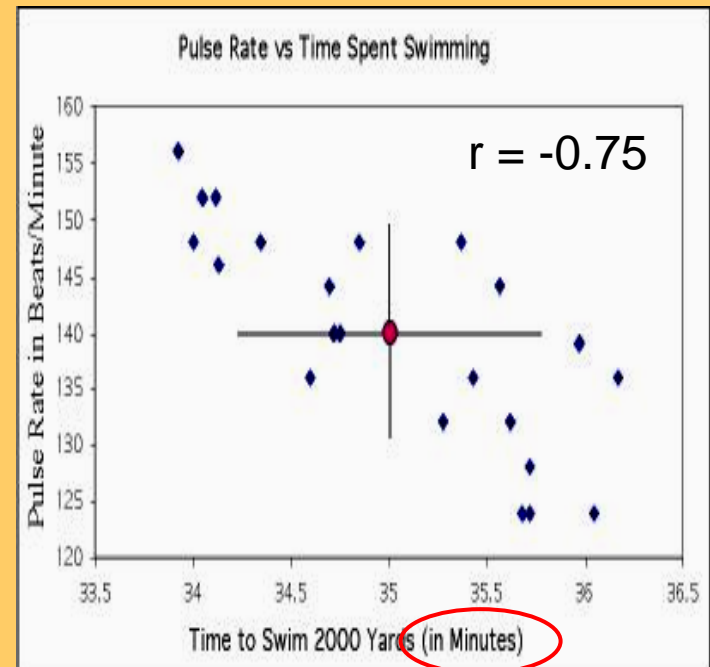
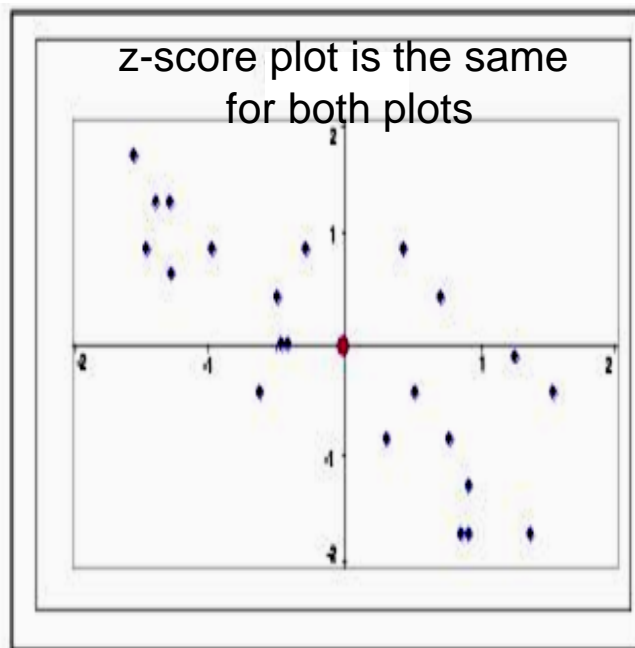
"Time to swim" is the explanatory variable here, and belongs on the x axis. However, in either plot r is the same ($r = -0.75$).

"r" has no unit

Changing the units of variables does not change the correlation coefficient "r", because we get rid of all our units when we standardize (get z-scores).

$$r = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

z for time z for pulse

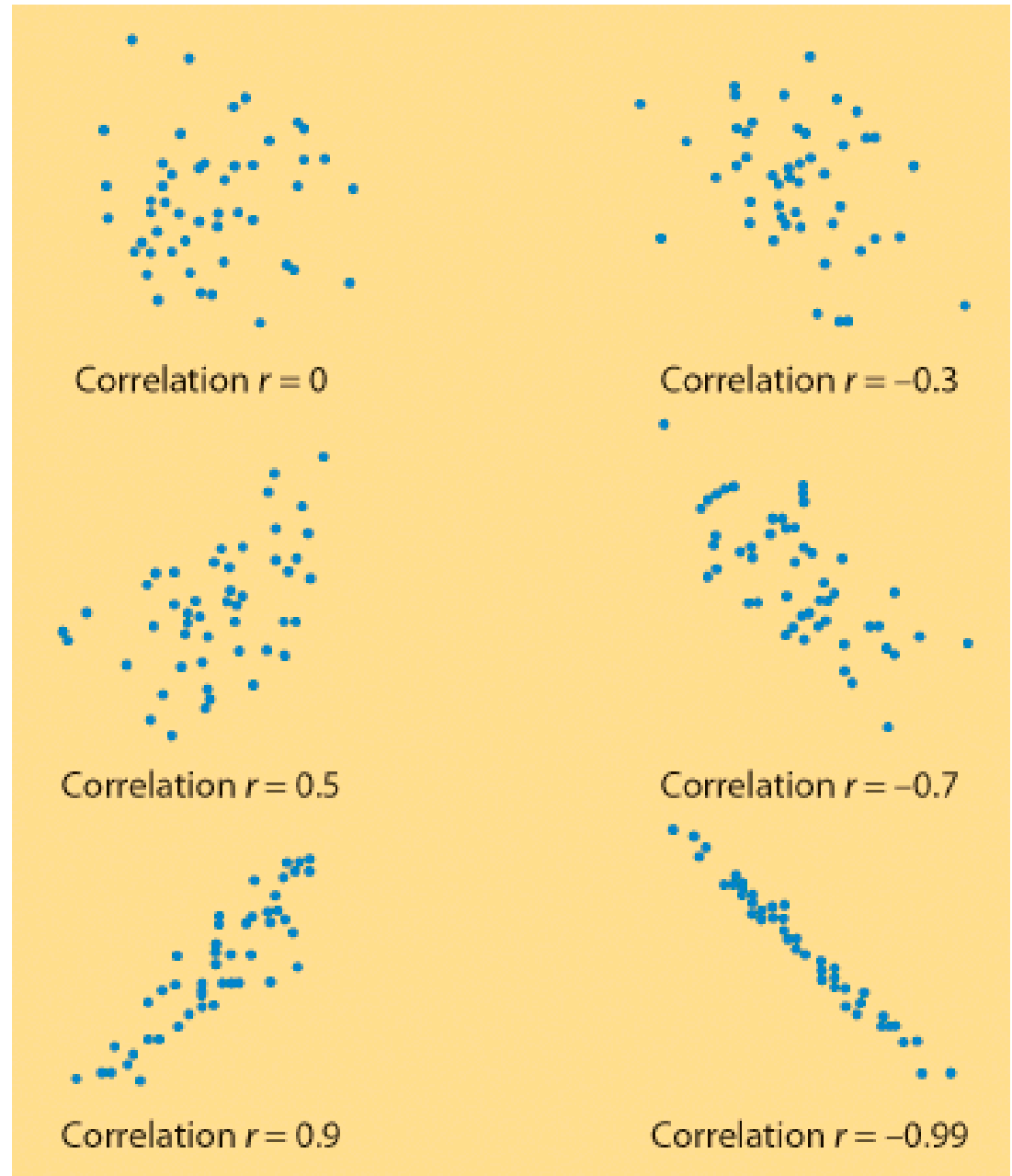


"r" ranges from -1 to +1

"r" quantifies the **strength** and **direction** of a **linear relationship** between 2 quantitative variables.

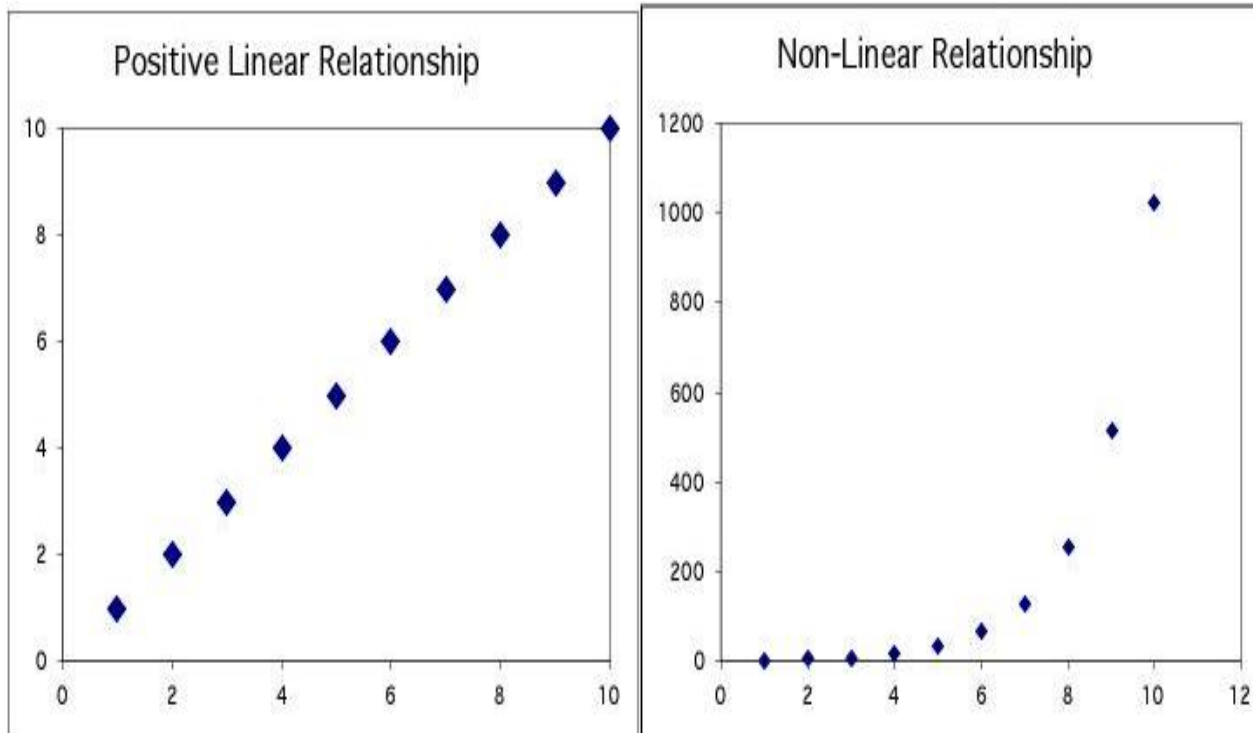
Strength: how closely the points follow a straight line.

Direction: is positive when individuals with higher X values tend to have higher values of Y .



Correlation only describes linear relationships

No matter how strong the association, r does not describe curved relationships.



Note: You can sometimes transform a non-linear association to a linear form, for instance by taking the logarithm. You can then calculate a correlation using the transformed data.

1.1 Continuous data: Spearman correlation

- This was introduced by Charles Edward Spearman(1863-1945)
- The Spearman rank correlation coefficient is the nonparametric version of the Pearson correlation coefficient and is defined as the Pearson correlation coefficient between the **rank variables**.

1.1 Continuous data: Pearson correlation

The formula of r_s for Spearman correlation is

$$r_s = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n^3 - n}$$

where

d_i is the difference between the two ranking
 n is the number of observations.

1.1 Continuous data: Pearson correlation

Assumptions:

- Data must be at least ordinal
- Scores on one variable must be monotonically related to the other variable
- Note: Spearman rank correlation test does not make any assumptions about the distribution

1.1 Continuous data: Example

■ Sample question

The scores for nine students in history and algebra are as follows:

History: 35, 23, 47, 17, 10, 43, 9, 6, 28

Algebra: 30, 33, 45, 23, 8, 49, 12, 4, 31

Compute the Spearman rank correlation.

1.1 Continuous data: Example

■ Answer

Step 1: rank each student

History	Rank	Algebra	Rank
35	3	30	5
23	5	33	3
47	1	45	2
17	6	23	6
10	7	8	8
43	2	49	1
9	8	12	7
6	9	4	9
28	4	31	4

1.1 Continuous data: Example

■ Answer

Step 2: calculate difference between the ranks (d) and square (d^2)

History	Rank	Algebra	Rank	d	d^2
35	3	30	5	2	4
23	5	33	3	2	4
47	1	45	2	1	1
17	6	23	6	0	0
10	7	8	8	1	1
43	2	49	1	1	1
9	8	12	7	1	1
6	9	4	9	0	0
28	4	31	4	0	0

1.1 Continuous data: Example

■ Answer

Step 3: sum (add up) all the d^2 scores

$$\Sigma d^2 = 4 + 4 + 1 + 0 + 1 + 1 + 1 + 0 + 0 = 12$$

Step 4: insert the values in the formula.

$$\begin{aligned}\rho &= 1 - (6 \cdot 12) / (9(81-1)) = 1 - 72/720 = 1-0.1 \\ &= 0.9\end{aligned}$$

1.1 Continuous data: Comparison

- The fundamental difference between the two correlation coefficients is that the Pearson coefficient works with a linear relationship between the two variables whereas the Spearman Coefficient works with monotonic relationships as well.
- One more difference is that Pearson works with raw data values of the variables whereas Spearman works with rank-ordered variables.

1.2 Categorical data: 2x2 Contingency Table

- The table shows the data from a study of 91 patients who had a myocardial infarction (Snow 1965). One variable is treatment (propranolol versus a placebo), and the other is outcome (survival for at least 28 days versus death within 28 days).

		<u>OUTCOME</u>		
		Survival for at least 28 days	Death	Total
<u>Treatment</u>	Propranolol	38	7	45
	Placebo	29	17	46
Total		67	24	91

1.2 Categorical data: 2x2 Contingency Table

Hypotheses for Two-way Tables:

- The null hypothesis H_0 is simply that ***there is no association*** between the row and column variable.
- The alternative hypothesis H_a is that ***there is an association*** between the two variables. It doesn't specify a particular direction and can't really be described as one-sided or two-sided.

1.2 Categorical data: 2x2 Contingency Table

Hypothesis statement in Our Example

- Null hypothesis: the method of treating the myocardial infarction patients did not influence the proportion of patients who survived for at least 28 days.
- The alternative hypothesis is that the outcome (survival or death) depended on the treatment, meaning that the outcomes was the dependent variable and the treatment was the independent variable.

A 2×2 table

- There are two **categorical variables**, each has two categories shown in the following table.

		Variable A		
		A1	A2	marginal
Variable B	B1	$n_{11} (\pi_{11})$	$n_{21} (\pi_{21})$	$n_{.1} (\pi_{.1})$
	B2	$n_{12} (\pi_{12})$	$n_{22} (\pi_{22})$	$n_{.2} (\pi_{.2})$
	marginal	$n_{1.} (\pi_{1.})$	$n_{2.} (\pi_{2.})$	$n (\pi_{..})$

- $\pi_{ij} = n_{ij}/n$;
- $\pi_{i.} = n_{i.}/n$
- $\pi_{.j} = \frac{n_{.j}}{n}$
- $\pi_{..} = \frac{n}{n} = 1$
- Question: Are variable A and B independent?

The Chi-Square Test of independence

The chi-square statistic (χ^2) is a measure of how much the observed cell counts in a two-way table diverge from the expected cell counts.

The formula for the χ^2 statistic is:

(summed over all $r * c$ cells in the table)

$$\chi^2 = \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$$

$$\text{Degrees of freedom (df)} = (r - 1) \times (c - 1)$$

Tip: First, calculate the χ^2 components, (observed-expected)²/expected, for each cell of the table, and then sum them up to arrive at the χ^2 statistic.

■ $\pi_{i.}$ and $\pi_{.j}$ are the marginal probability, and \rightarrow

$$\pi_{i.} = \sum_{j=1}^J \pi_{ij}$$

■ Under the null of independence:

$$\pi_{.j} = \sum_{i=1}^I \pi_{ij}$$

$$\pi_{ij} = \pi_{i.}\pi_{.j}$$

We thus consider testing the following null hypothesis:

$$H_0: \pi_{ij} = \pi_{i.}\pi_{.j}, \quad i = 1, \dots, I, \quad j = 1, \dots, J$$

versus the alternative that the π_{ij} are free. Under H_0 , the mle of π_{ij} is

$$\begin{aligned} \hat{\pi}_{ij} &= \hat{\pi}_{i.}\hat{\pi}_{.j} \\ &= \frac{n_{i.}}{n} \times \frac{n_{.j}}{n} \end{aligned}$$

simply

Under the alternative, the mle of π_{ij} is

$$\tilde{\pi}_{ij} = \frac{n_{ij}}{n}$$

1.2 Categorical data: 2x2 Contingency Table

- Expected cell counts

		<u>OUTCOME</u>		
		Survival for at least 28 days	Death	Total
<u>Treatment</u>	Propranolol	33.13	11.87	45
	Placebo	33.87	12.13	46
	Total	67	24	91

1.2 Categorical data: Chi-square test

When to Use Chi-Square Test for Independence

- The sampling method is simple random sampling.
- The variables under study are each categorical.
- If sample data are displayed in a contingency table, the expected frequency count for each cell of the table is **at least 5**.

1.2 Categorical data: Chi-square test

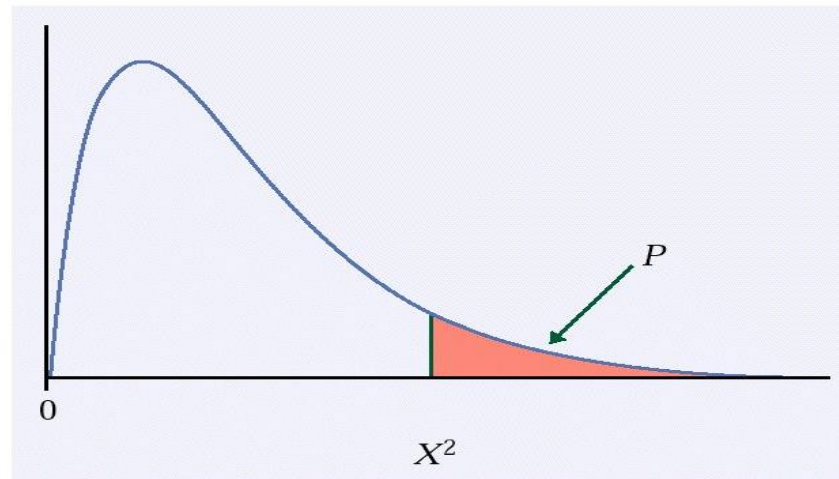
CHI-SQUARE TEST FOR TWO-WAY TABLES

The null hypothesis H_0 is that there is no association between the row and column variables in a two-way table. The alternative is that these variables are related.

If H_0 is true, the chi-square statistic X^2 has approximately a χ^2 distribution with $(r - 1)(c - 1)$ degrees of freedom.

The P -value for the chi-square test is

$$P(\chi^2 \geq X^2)$$



where χ^2 is a random variable having the $\chi^2(\text{df})$ distribution with $\text{df} = (r - 1)(c - 1)$.

1.2 Categorical data: Chi-square test

Calculation of the chi-square (χ^2) value

$$\begin{aligned}\chi^2 &= \sum \left[\frac{(O - E)^2}{E} \right] \\&= \frac{(38 - 33.13)^2}{33.13} + \frac{(7 - 11.87)^2}{11.87} + \frac{(29 - 33.87)^2}{33.87} + \frac{(17 - 12.13)^2}{12.13} \\&= \frac{(4.87)^2}{33.13} + \frac{(-4.87)^2}{11.87} + \frac{(-4.87)^2}{33.87} + \frac{(-4.87)^2}{12.13} \\&= \frac{23.72}{33.13} + \frac{23.72}{11.87} + \frac{23.72}{33.87} + \frac{23.72}{12.13} \\&= 0.72 + 2 + 0.7 + 1.96 = 5.38\end{aligned}$$

$$df = (R - 1)(C - 1) = 1$$

$$p - \text{value} < 0.025$$

Interpretation :

The results noted in this 2×2 table are statistically significant.

That is, it is highly probable (only 1 chance in about 50 of being wrong) that the investigator can reject the null hypothesis of independence and accept the alternative hypothesis that propranolol does affect the outcome of myocardial infarction.

Music and wine purchase decision

H_0 : No relationship between music and wine

H_a : Music and wine are related

Observed counts

Wine	Music		
	None	French	Italian
French	30	39	30
Italian	11	1	19
Other	43	35	35

Expected counts

Wine	Music		
	None	French	Italian
French	34.222	30.556	34.222
Italian	10.716	9.568	10.716
Other	39.062	34.877	39.062

We calculate nine χ^2 components and sum them to produce the χ^2 statistic:

$$\begin{aligned}\chi^2 &= \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \\&= \frac{(30 - 34.222)^2}{34.222} + \frac{(39 - 30.556)^2}{30.556} + \frac{(30 - 34.222)^2}{34.222} \\&\quad + \frac{(11 - 10.716)^2}{10.716} + \frac{(1 - 9.568)^2}{9.568} + \frac{(19 - 10.716)^2}{10.716} \\&\quad + \frac{(43 - 39.062)^2}{39.062} + \frac{(35 - 34.877)^2}{34.877} + \frac{(35 - 39.062)^2}{39.062} \\&= 0.5209 + 2.3337 + 0.5209 + 0.0075 + 7.6724 \\&\quad + 6.4038 + 0.3971 + 0.0004 + 0.4223 \\&= 18.28\end{aligned}$$



1.2 Categorical data: Chi-square test

■ Summary: Computations for Two-way Tables

1. create the table, including observed cell counts, column and row totals.
2. Find the expected cell counts.
 - Determine if a χ^2 test is appropriate
 - Calculate the χ^2 statistic and number of degrees of freedom
3. Find the approximate P -value
4. Draw conclusions about the association between the row and column variables.

1.2 Categorical data

- Question

What do we do if the expected values in any of the cells in a 2x2 table is below 5?

1.2 Categorical data: Fisher test

Before we proceed with the Fisher test, we first introduce some notation. We represent the cells by the letters a , b , c and d , call the totals across rows and columns *marginal totals*, and represent the grand total by n . So the table now looks like this:

	men	women	total
dieting	a	b	$a + b$
not dieting	c	d	$c + d$
totals	$a + c$	$b + d$	n

1.2 Categorical data: Fisher test

- Fisher showed that the probability of obtaining any such set of values was given by the hypergeometric distribution:

$$p = \binom{a+b}{a} \binom{c+d}{c} / \binom{n}{a+c}$$
$$= \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!}$$

1.2 Categorical data: Fisher test

For example, a sample of teenagers might be divided into male and female on the one hand, and those that are and are not currently dieting on the other. We hypothesize, perhaps, that the proportion of dieting individuals is higher among the women than among the men, and we want to test whether any difference of proportions that we observe is significant. The data might look like this:

	men	women	total
dieting	1	9	10
not dieting	11	3	14
totals	12	12	24

1.2 Categorical data: Fisher test

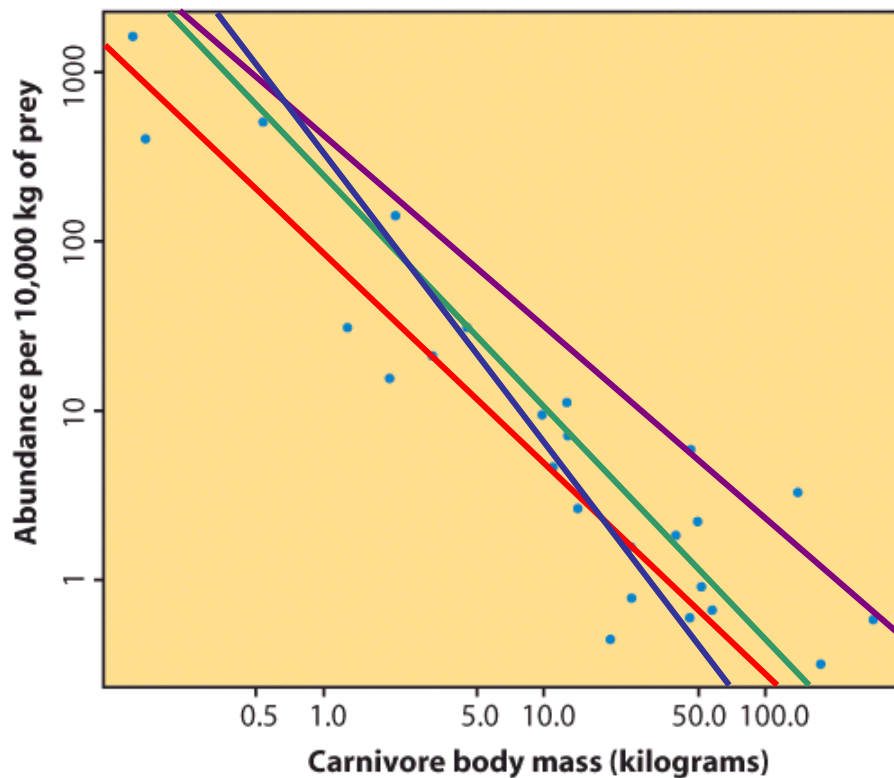
	men	women	total
dieting	1	9	10
not dieting	11	3	14
totals	12	12	24

$$F = \frac{10! 14! 12! 12!}{24! 1! 9! 11! 3!} = 0.00134$$

Recall that p-value is the probability of observing data as extreme or more extreme if the null hypothesis is true. So the p-value is this problem is 0.00137.

In R, `fisher.test(.)`

2. Linear Regression and Logistic Regression



Correlation tells us about *strength* (scatter) and *direction* of the linear relationship between two quantitative variables.

In addition, we would like to have a numerical description of how both variables vary together. For instance, is one variable increasing faster than the other one? And we would like to make predictions based on that numerical description.

But which line best describes our data?

2. Introduction to Regression Analysis

- Regression Analysis is used to:

1. Predict the value of a dependent variable based on the value of at least one variable.

2. Explain the impact of changes in an independent variable on the dependent variable

- Dependent Variable: the variable we wish to predict or explain

- Independent Variable: the variable used to explain the dependent variable

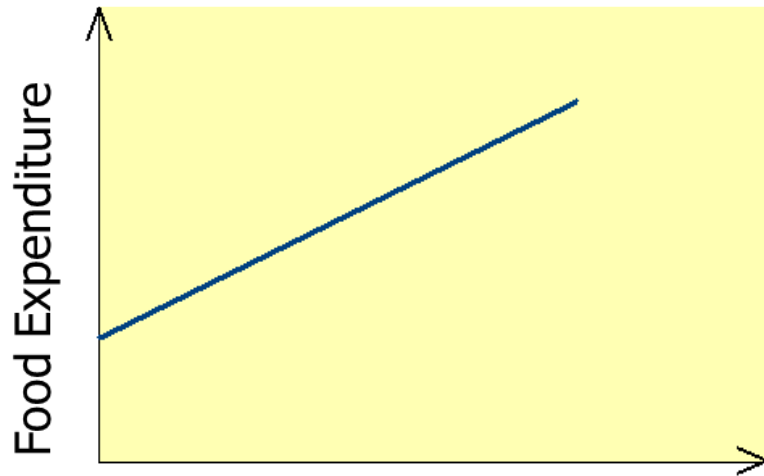
2.1 Linear Regression model

■ Definition

A (simple) regression model that gives a straight-line relationship between two variables is called a linear regression model.

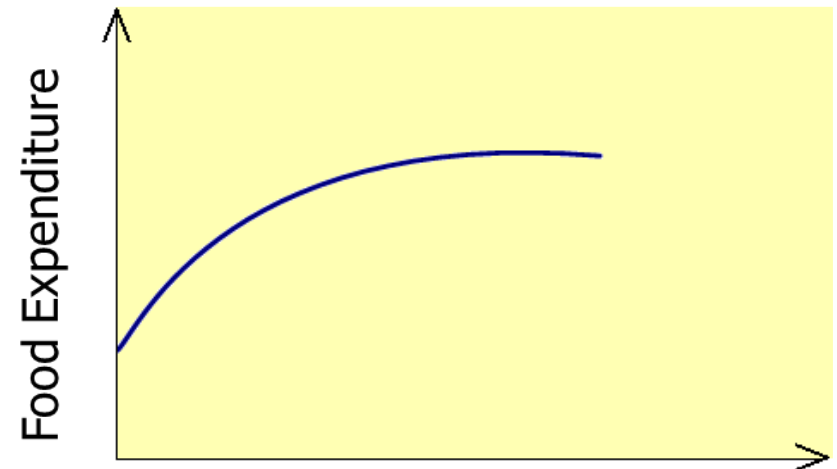
In the regression model $y = a + bx + e$, where a is called the y-intercept or constant term, b is the slope, and e is the random error term. The dependent and independent variables are y and x , respectively.

2.1 Linear Regression model



Income

(a)



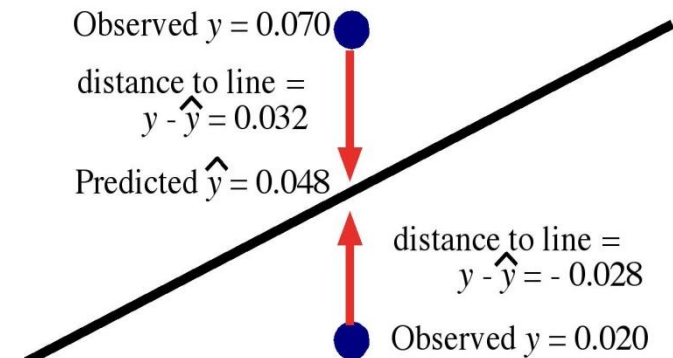
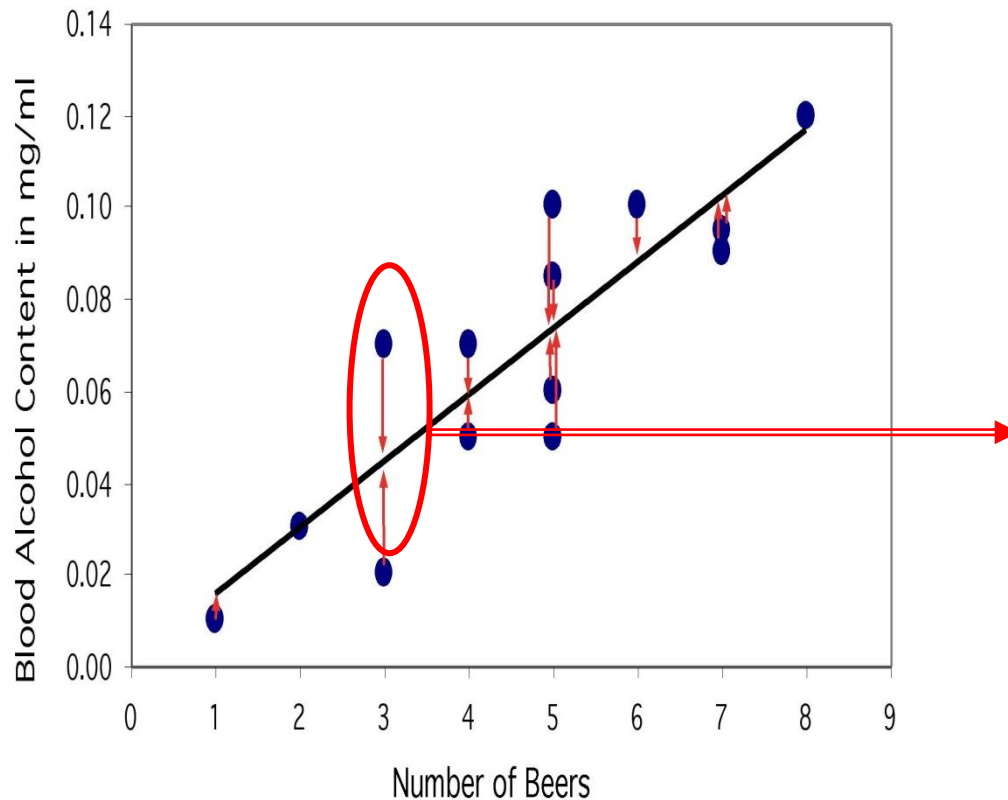
Income

(b)

- Relationship between food expenditure and income.
(a) Linear relationship. (b) Nonlinear relationship.

The regression line

The **least-squares regression line** is the unique line such that the sum of the squared vertical (y) distances between the data points and the line is as small as possible.



Distances between the points and line are squared so all are positive values. This is done so that distances can be properly added (Pythagoras).

The **least-square line** of y on x is the line that makes the sum of the squares of the vertical distances of the data points from the line as small as possible. More formally, the least-square line minimize the quantity, $\sum_{i=1}^n [y_i - (a + bx_i)]^2$

Computing the least-square line

Mathematically, given n pairs of numbers $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, find the number a and b that minimize the function

$$f(a, b) = \sum_{i=1}^n [y_i - (a + bx_i)]^2$$

Let \bar{x} and \bar{y} represent the mean of x_1, x_2, \dots, x_n and y_1, y_2, \dots, y_n respectively. The least-square line is the line

$$\hat{y} = \hat{a} + \hat{b}x$$

with **slope**

$$\hat{b} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

and **intercept**

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

\hat{y} (read “y hat”) in the equation means that the line gives a *predicted* response \hat{y} for any x . It is also called the *fitted* value. The predicted response will usually not be exactly the same as the actually *observed* response y . The “hat” over the letters a and b is used to remind us that these are the values that minimize the sum of squared deviations.

2.1 Linear Regression model

- In the model $\hat{y} = a + bx$, a and b , which are calculated using sample data, are called the estimates of A and B .

$$b = \frac{SS_{xy}}{SS_{xx}} \quad \text{and} \quad a = \bar{y} - b\bar{x}$$

$$SS_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y})$$

$$SS_{xx} = \sum (x_i - \bar{x})^2$$

2.1 Linear Regression model

- The error sum of squares, denoted SSE, is

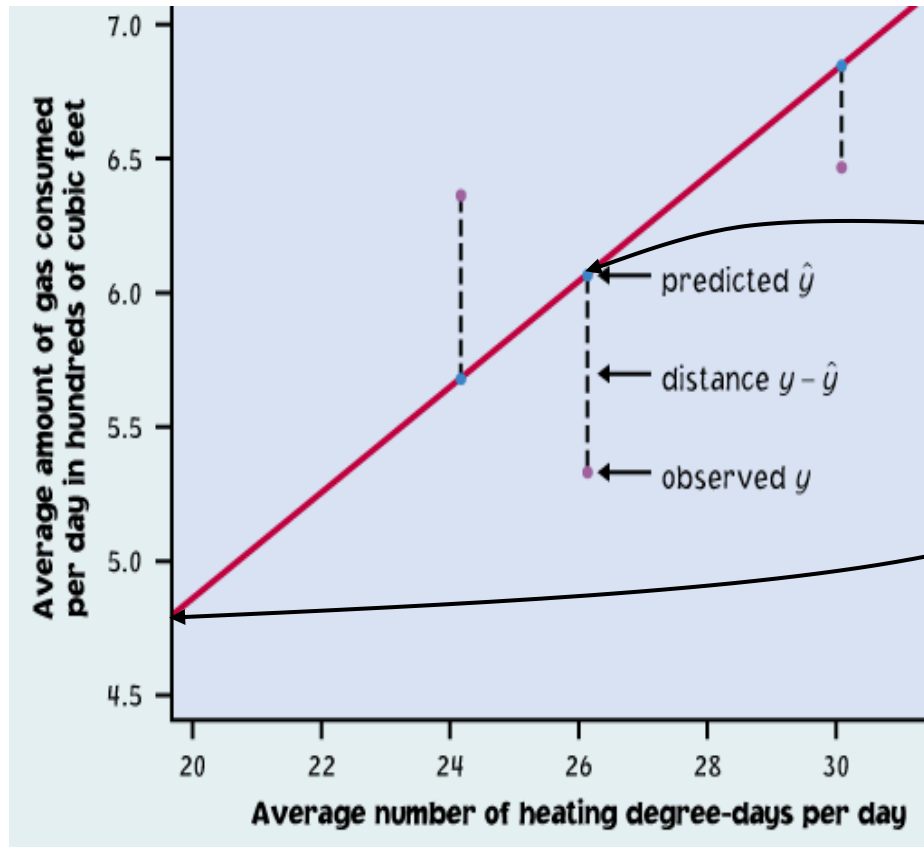
$$SSE = \sum e^2 = \sum (y - \hat{y})^2$$

- The values of a and b that give the minimum SSE are called the least square estimates of A and B, and the regression line obtained with these estimates is called the least square line.

Properties

The least-squares regression line can be shown to have this equation:

$$\hat{y} = \hat{a} + \hat{b}x$$



\hat{y} is the predicted y value (y hat)

\hat{b} is the **slope**

\hat{a} is the **y-intercept**

Example

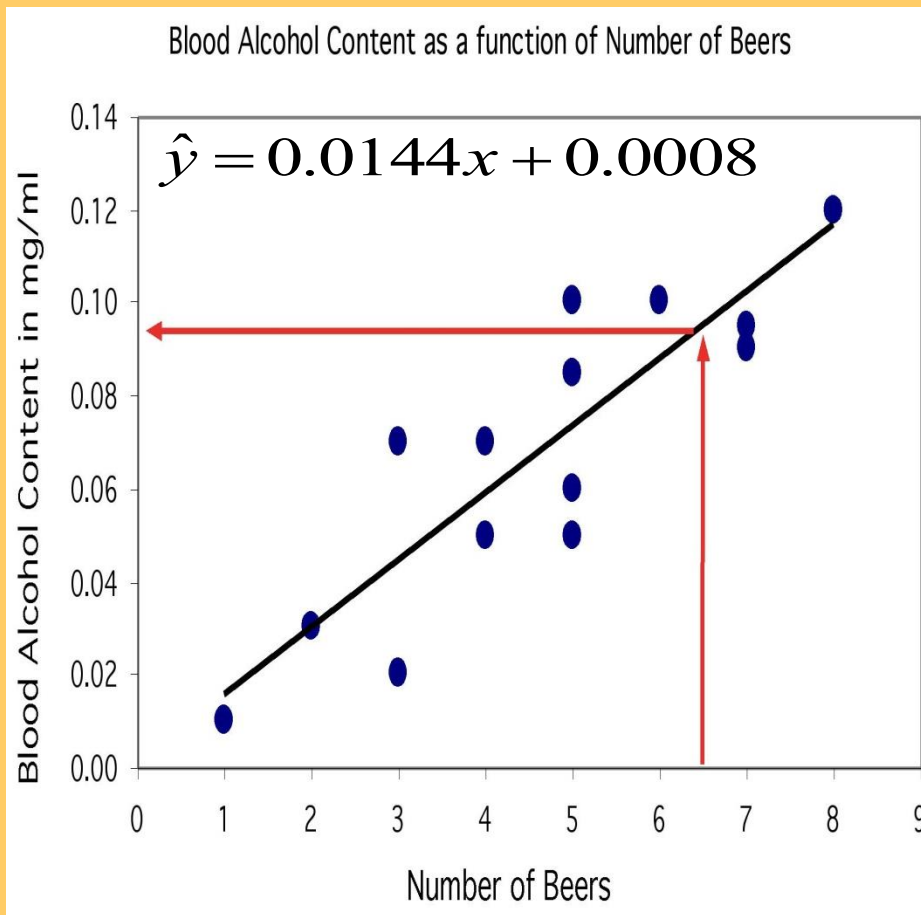
i	x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	1	4	-3	-2.85714	8.57143	9
2	2	6	-2	-0.85714	1.71429	4
3	3	7	-1	0.14286	-0.14286	1
4	4	5	0	-1.85714	0	0
5	5	8	1	1.14286	1.14286	1
6	6	8	2	1.14286	2.28571	4
7	7	10	3	3.14286	9.42857	9
mean	$\bar{x} = 4$	$\bar{y} = 6.86$				
Total	28	48	0	0	23	28

Based on the calculation on the above table, it's easy to get $\hat{b} = 23 / 28 = 0.8214$ and $\hat{a} = 48 / 7 - 0.8214 \times (28 / 7) = 3.5714$. Therefore, the fitted **least square line** is given by,

$$\hat{y} = 3.5714 + 0.8214x$$

Making predictions

The equation of the least-squares regression allows you to predict y for any x within the range studied.



Nobody in the study drank 6.5 beers, but by finding the value of \hat{y} from the regression line for $x = 6.5$ we would expect a blood alcohol content of 0.094 mg/ml.

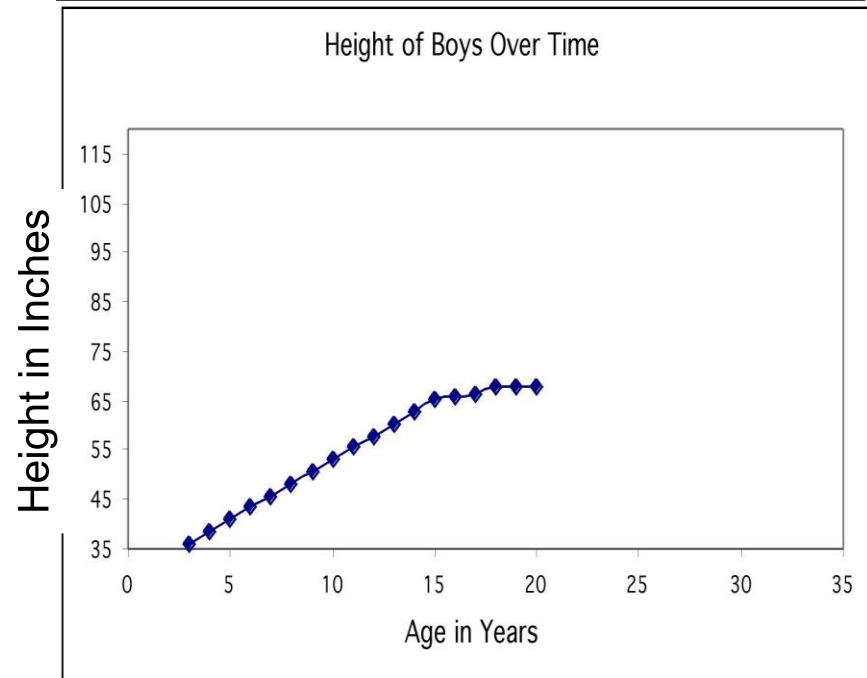
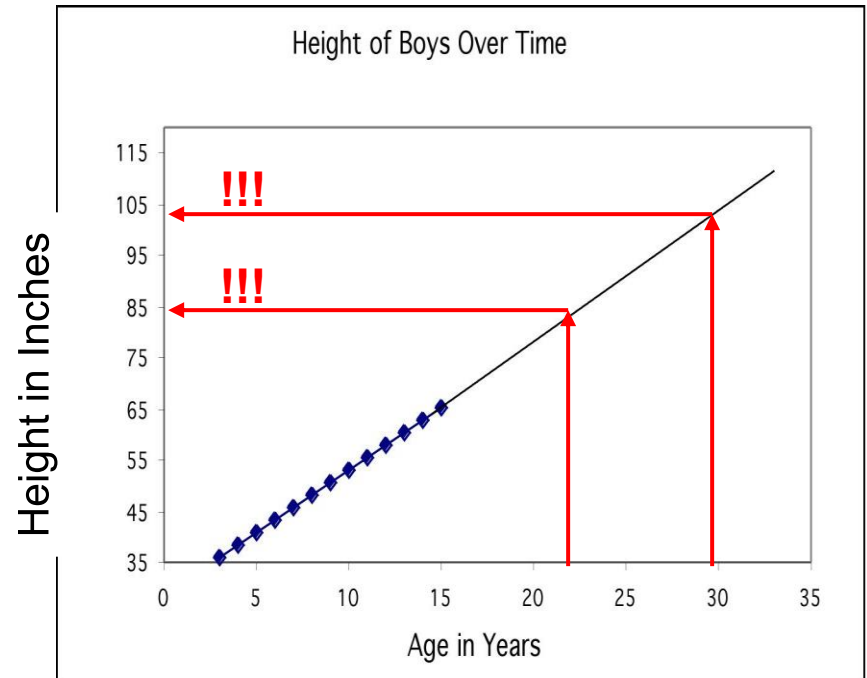
$$\hat{y} = 0.0144 * 6.5 + 0.0008$$

$$\hat{y} = 0.0936 + 0.0008 = 0.0944 \text{ mg/ml}$$

Extrapolation

Extrapolation is the use of a regression line for predictions *outside the range of x values* used to obtain the line.

This can be a very stupid thing to do, as seen here.



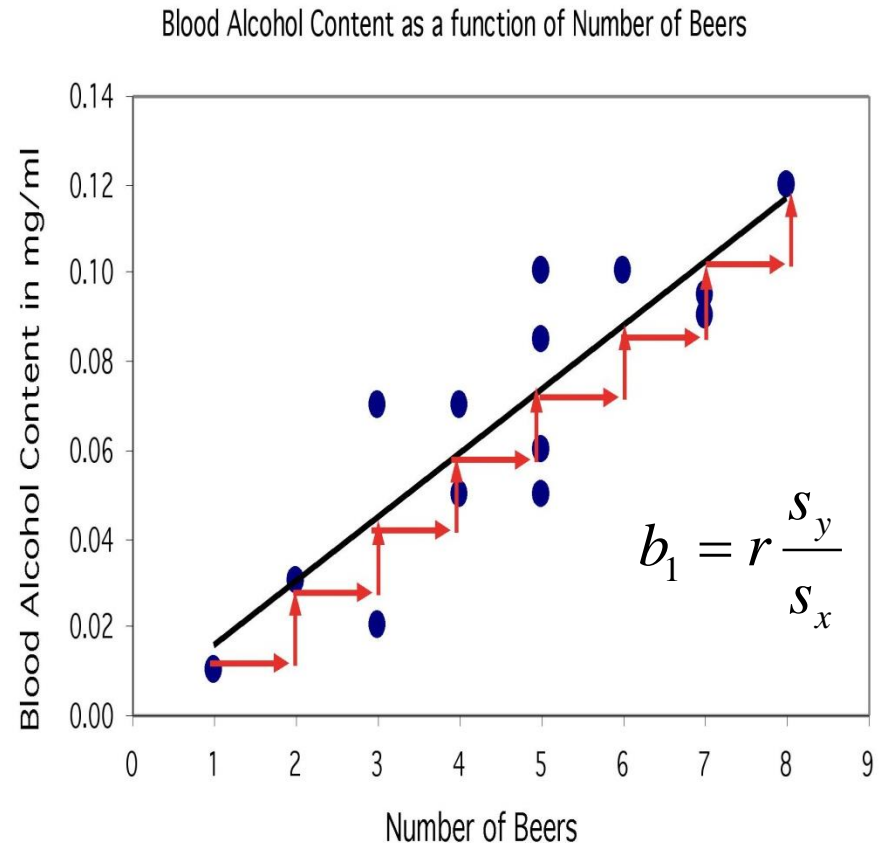
Coefficient of determination, r^2

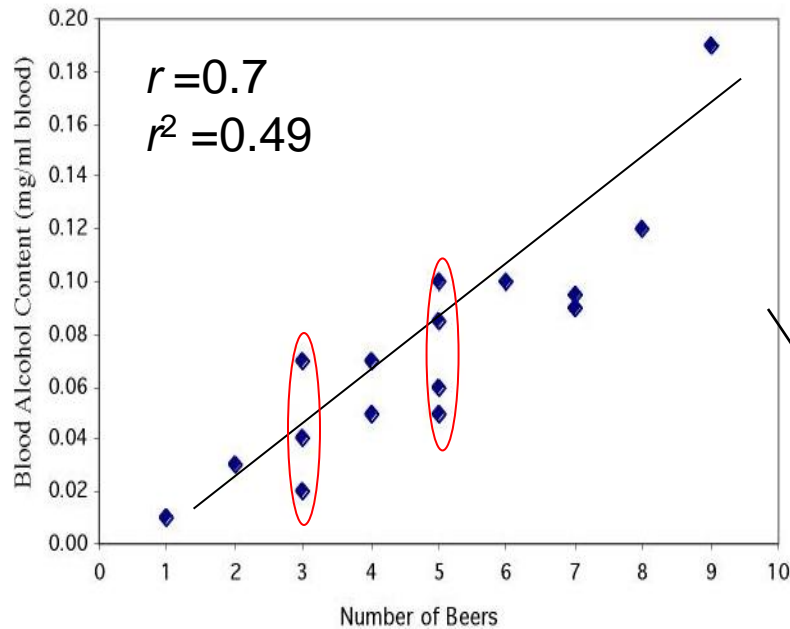
□ *Two important questions:*

- *Does data fit the linear model adequately?*
- *Will model predict response well enough to be useful?*

r^2 , the coefficient of determination, is the square of the correlation coefficient.

r^2 represents **the percentage of the variance in y** (vertical scatter from the regression line) **that can be explained by changes in x .**

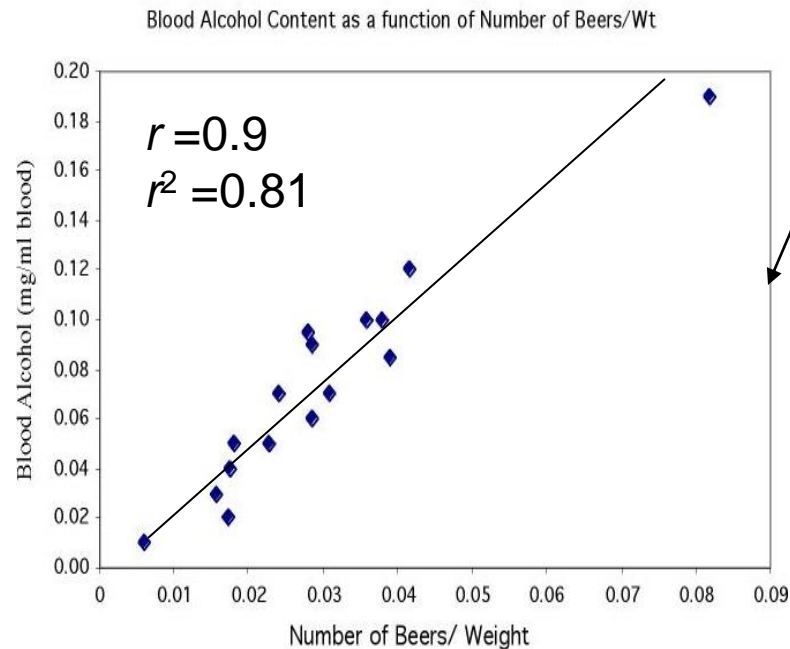




There is quite some variation in BAC for the same number of beers drank. A person's blood volume is a factor in the equation that was overlooked here.



We changed number of beers to number of beers/weight of person in lb.



- In the first plot, number of beers only explains 49% of the variation in blood alcohol content.
- But number of beers / weight explains 81% of the variation in blood alcohol content.
- Additional factors contribute to variations in BAC among individuals (like maybe some genetic ability to process alcohol).

Transforming relationships

A scatterplot might show a clear relationship between two quantitative variables, but issues of influential points or nonlinearity prevent us from using correlation and regression tools.

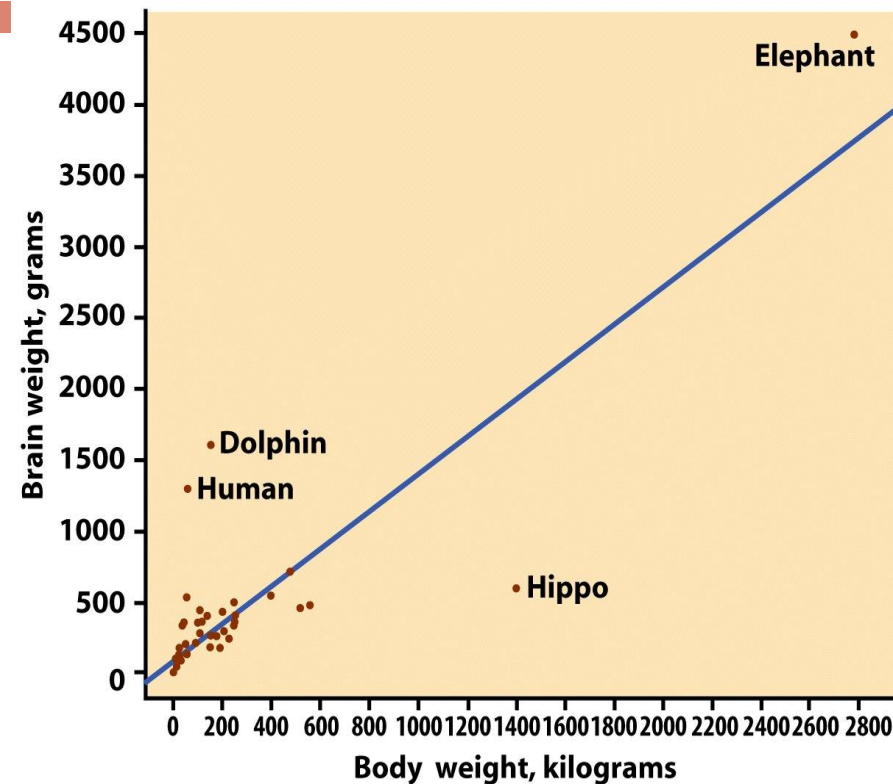
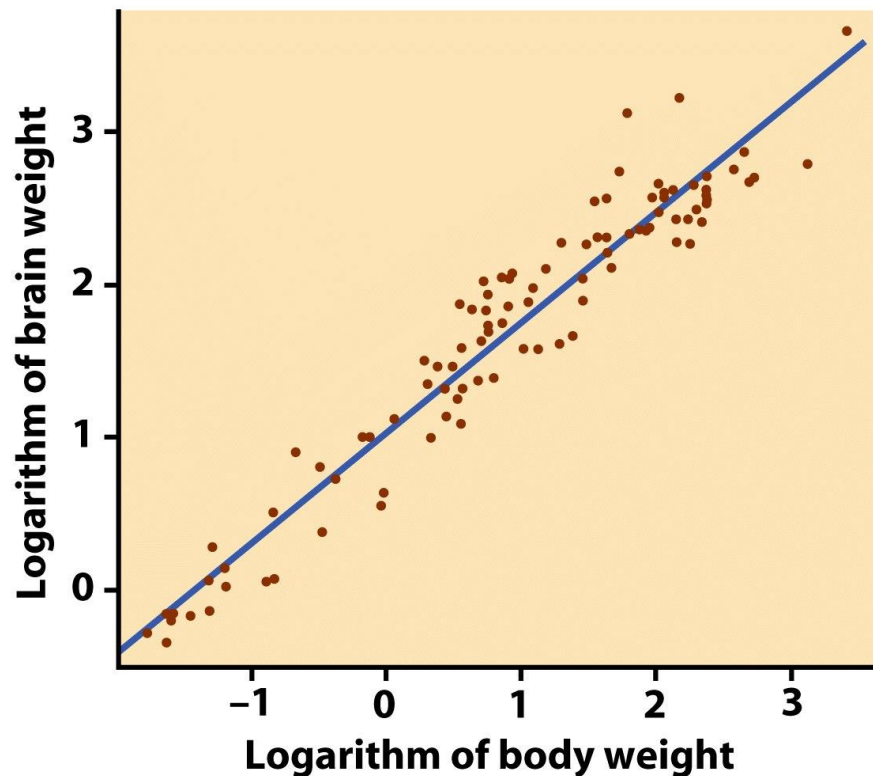
Transforming the data – changing the scale in which one or both of the variables are expressed – can make the shape of the relationship linear in some cases.

Example: Patterns of growth are often exponential, at least in their initial phase. Changing the response variable y into $\log(y)$ or $\ln(y)$ will transform the pattern from an upward-curved exponential to a straight line.

Body weight and brain weight in 96 mammal species

$r = 0.86$, but this is misleading.

The elephant is an influential point. Most mammals are very small in comparison. Without this point, $r = 0.50$ only.



Now we plot the log of brain weight against the log of body weight.

The pattern is linear, with $r = 0.96$.
The vertical scatter is homogenous
→ good for predictions of brain weight from body weight (in the log scale).

Always plot your data!

The correlations all give $r \approx 0.816$, and the regression lines are all approximately $\hat{y} = 3 + 0.5x$. For all four sets, we would predict $\hat{y} = 8$ when $x = 10$.

Table 2.8 Four data sets for exploring correlation and regression

Data Set A

x	10	8	13	9	11	14	6	4	12	7	5
y	8.04	6.95	7.58	8.81	8.33	9.96	7.24	4.26	10.84	4.82	5.68

Data Set B

x	10	8	13	9	11	14	6	4	12	7	5
y	9.14	8.14	8.74	8.77	9.26	8.10	6.13	3.10	9.13	7.26	4.74

Data Set C

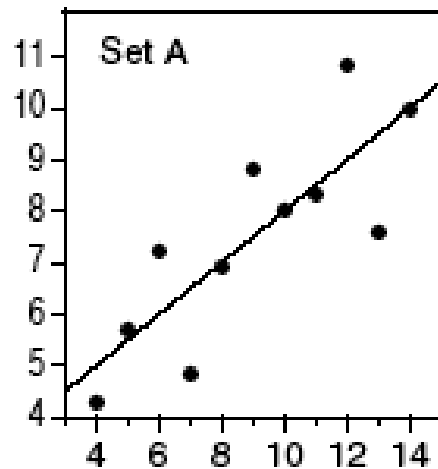
x	10	8	13	9	11	14	6	4	12	7	5
y	7.46	6.77	12.74	7.11	7.81	8.84	6.08	5.39	8.15	6.42	5.73

Data Set D

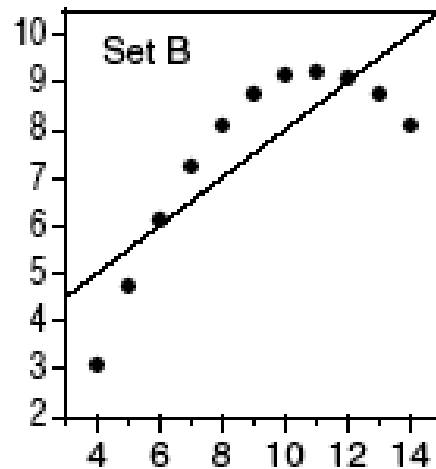
x	8	8	8	8	8	8	8	8	8	8	19
y	6.58	5.76	7.71	8.84	8.47	7.04	5.25	5.56	7.91	6.89	12.50

Source: Frank J. Anscombe, "Graphs in statistical analysis," *The American Statistician*, 27 (1973), pp. 17–21.

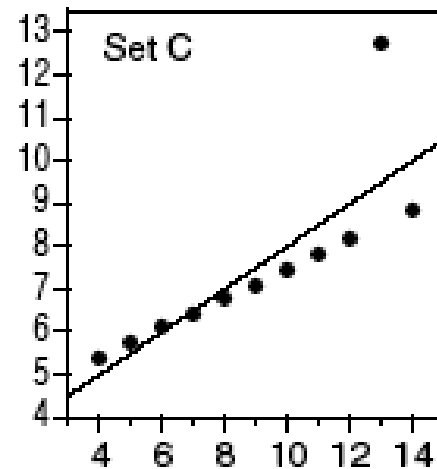
However, making the scatterplots shows us that the correlation/
regression analysis is not appropriate for all data sets.



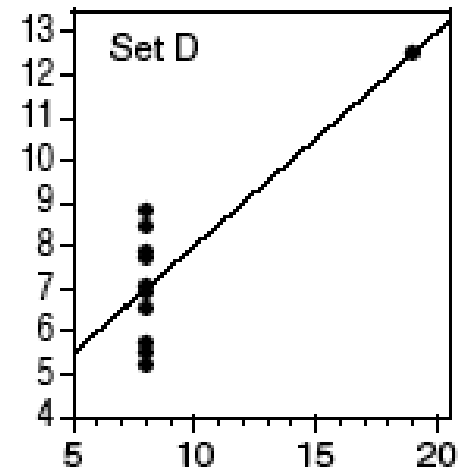
Moderate linear
association;
regression OK.



Obvious
nonlinear
relationship;
regression
not OK.



One point deviates
from the highly
linear pattern; this
outlier must be
examined closely
before proceeding.



Just one very
influential point; all
other points have
the same x value;
a redesign is due
here.

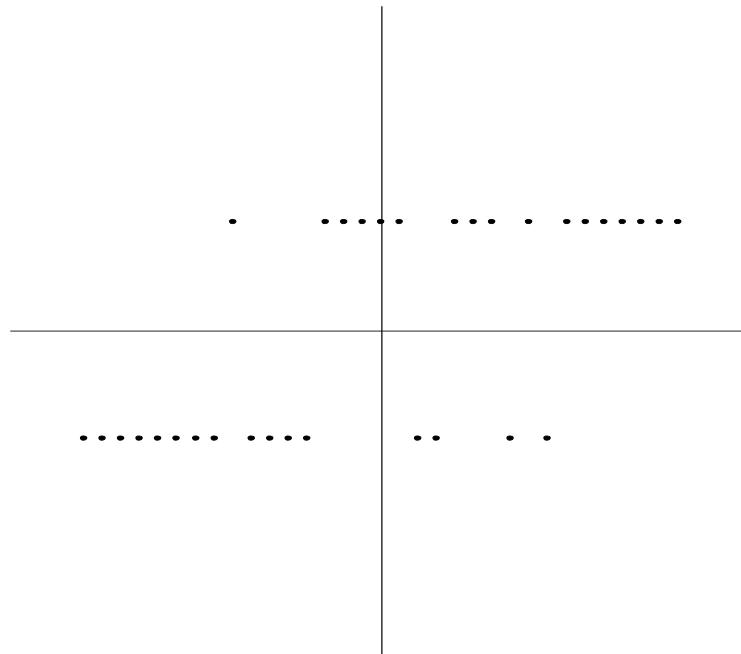
Explaining association: causation

- ❑ Association is not causation.
- ❑ Even if an association is very strong, this is not by itself good evidence that a change in x will cause a change in y , so does NOT imply causation.
- ❑ Example 1: Daughter's body mass index depends on mother's body mass index. This is an example of direct causation.
- ❑ Example 2: Married men earn more than single men. Can a man raise his income by getting married?
- ❑ Only careful experimentation can show causation.

2.2 Logistic Regression model

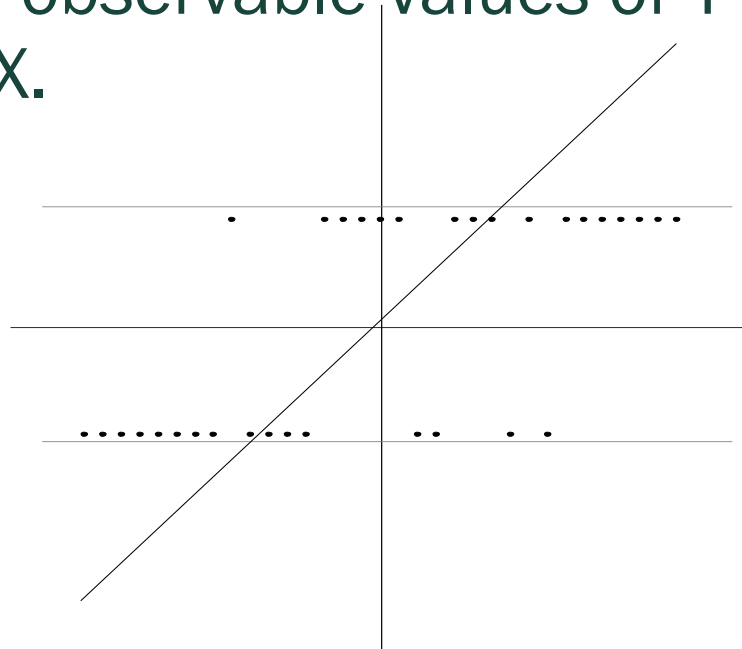
■ A Problem with Linear Regression

What if you have a binary outcome variable?



2.2 Logistic Regression model

- We could severely simplify the plot by drawing a line between the means for the two dependent variable levels, but this is problematic in two ways: (a) the line seems to oversimplify the relationship and (b) it gives predictions that cannot be observable values of Y for extreme values of X .



2.2 Logistic Regression model

- The reason this doesn't work is because the approach is analogous to fitting a linear model to the probability of the event. As you know, probabilities can only take values between 0 and 1. Hence, we need a different approach to ensure that our model is appropriate for the data.

2.2 Logistic Regression model

- Odds Ratios (OR) can be useful for comparisons.
- Suppose we have a trial to see if an intervention T reduces mortality, compared to a placebo, in patients with high cholesterol. The odds ratio is

$$OR = \frac{odds(death|intervention\ T)}{odds(death|placebo)}$$

- The OR describes the benefits of intervention T:
- $OR < 1$: the intervention is better than the placebo
- $OR = 1$: there is no difference between the intervention and the placebo
- $OR > 1$: the intervention is worse than the placebo

Odds

Definition:

$$\frac{\pi}{1 - \pi} = \frac{P(Yes)}{P(No)}$$

is the odds of Yes.

$$odds = \frac{\pi}{1 - \pi} \Leftrightarrow \pi = \frac{odds}{1 + odds}$$

Binary Logistic Regression Model


Y = Binary response

X = Quantitative predictor

π = proportion of 1's (yes, success) at any X

Equivalent forms of the logistic regression model:

Logit form

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X$$


Probability form

$$p = \frac{e^{b_0 + b_1 X}}{1 + e^{b_0 + b_1 X}}$$

What does this look like?

N.B.: This is natural log (aka “ln”)

$$\pi = \text{Prob}(Y = 1 | X = x)$$

2.2 Logistic Regression model

- The logistic distribution constrains the estimated probabilities to lie between 0 and 1.
- The estimated probability is:

$$\pi = 1/[1 + e^{(\beta_0 + \beta_1 X)}]$$

- if you let $\beta_0 + \beta_1 X = 0$, then $p = .50$
 - as $\beta_0 + \beta_1 X$ gets really big, p approaches 1
 - as $\beta_0 + \beta_1 X$ gets really small, p approaches 0
- ```
> fit=glm(Y~X, family=binomial, data=data)
> summary(fit)
```

## 2.2 Logistic Regression model

### ■ Interpretation of $\beta_0$

$\beta_0$  is the log of the odds of success at zero values for all covariates

### ■ Interpretation of $\beta_1$

$\beta_1$  is the increase in the log odds ratio associated with a one-unit increase in  $X$

- If  $\beta_1 = 0$ , there is no association between changes in  $X$  and changes in success probability ( $OR = 1$ ).
- If  $\beta_1 > 0$ , there is a positive association between  $X$  and  $p$  ( $OR > 1$ ).
- If  $\beta_1 < 0$ , there is a negative association between  $X$  and  $p$  ( $OR < 1$ )



$X$  is replaced by  $X + 1$ :

$$odds = e^{b_0 + b_1 X}$$

is replaced by

$$odds = e^{b_0 + b_1 (X+1)}$$

So the ratio is

$$\frac{e^{b_0 + b_1 (X+1)}}{e^{b_0 + b_1 X}} = e^{b_0 + b_1 (X+1) - (b_0 + b_1 X)} = e^{b_1}$$

## Interpreting “Slope” using Odds Ratio

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X$$

$$\Rightarrow \text{odds} = e^{\beta_0 + \beta_1 X}$$

When we increase  $X$  by 1, the ratio of the new odds to the old odds is  $e^{\beta_1}$ .

i.e. odds are multiplied by  $e^{\beta_1}$ .

## Example: Golf Putts

| Length | 3   | 4   | 5   | 6   | 7   |
|--------|-----|-----|-----|-----|-----|
| Made   | 84  | 88  | 61  | 61  | 44  |
| Missed | 17  | 31  | 47  | 64  | 90  |
| Total  | 101 | 119 | 108 | 125 | 134 |

Build a model to predict the proportion of putts made (success) based on length (in feet).

# Logistic Regression for Putting

Call:

```
glm(formula = Made ~ Length, family = binomial, data =
Putts1)
```

Deviance Residuals:

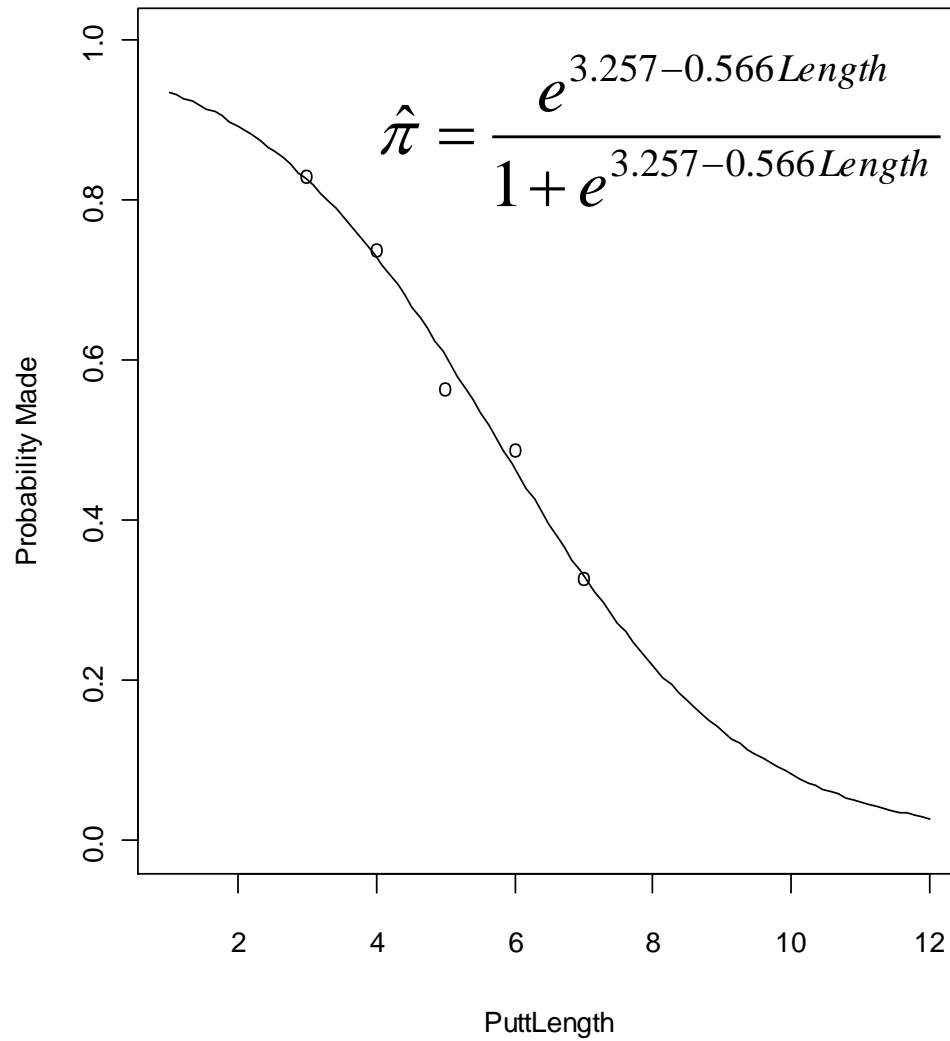
| Min     | 1Q      | Median | 3Q     | Max    |
|---------|---------|--------|--------|--------|
| -1.8705 | -1.1186 | 0.6181 | 1.0026 | 1.4882 |

Coefficients:

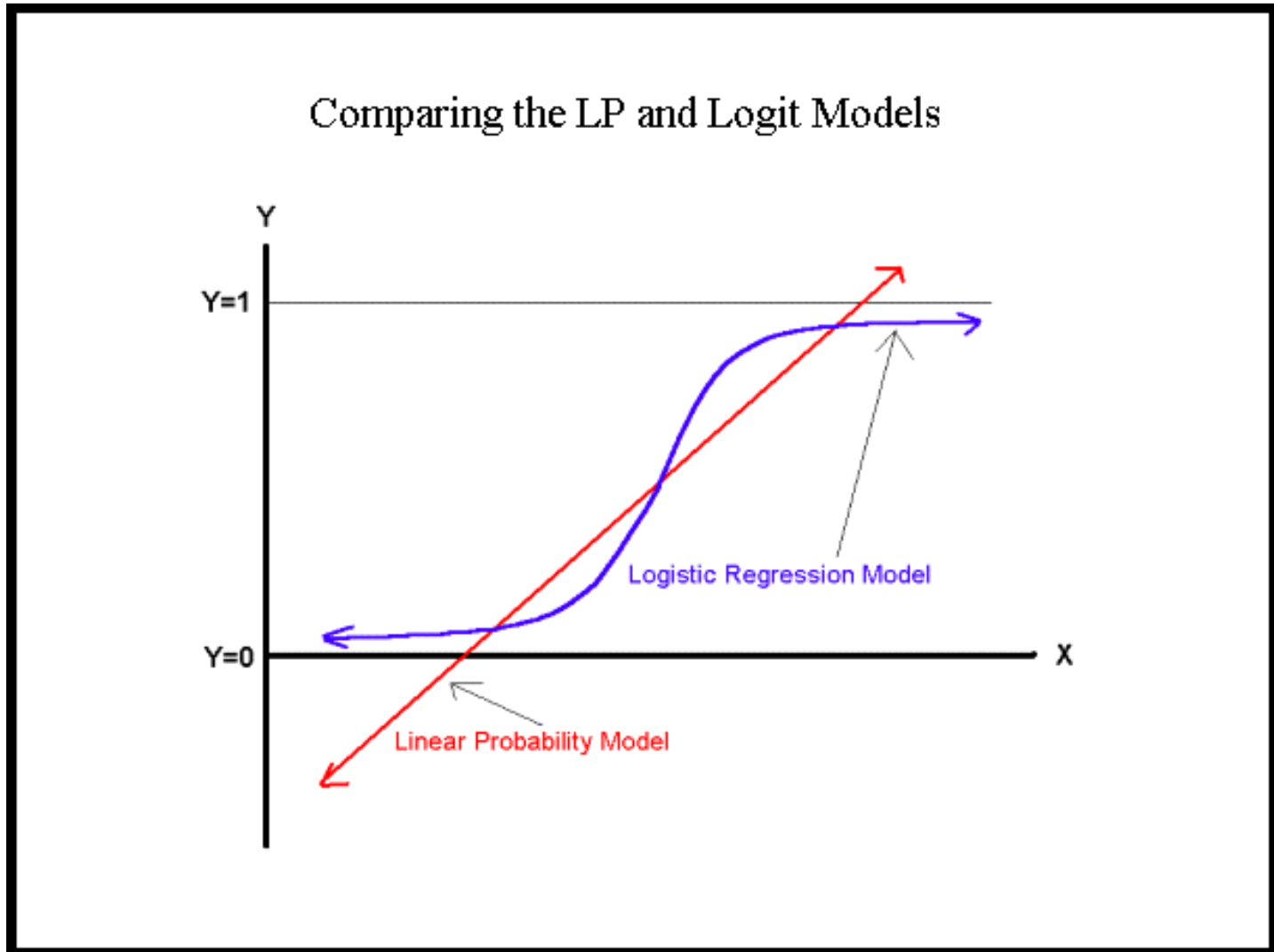
|             | Estimate | Std. Error | z value | Pr(> z ) |     |
|-------------|----------|------------|---------|----------|-----|
| (Intercept) | 3.25684  | 0.36893    | 8.828   | <2e-16   | *** |
| Length      | -0.56614 | 0.06747    | -8.391  | <2e-16   | *** |

---

# Probability Form of Putting Model



## 2.2 Logistic Regression model



# Binary Logistic Regression Model

$Y = \text{Binary}$

$X = \text{Single predictor}$

$\pi = \text{proportion of 1's (yes, success) at any } x$

Equivalent forms of the logistic regression model:

Logit form  $\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X$

Probability form 
$$\pi = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

# Binary Logistic Regression Model

$Y = \text{Binary}$

$X_1, X_2, \dots, X_k = \text{Multiple}$

$\pi = \text{proportion of 1's at any } x_1, x_2, \dots, x_k$

Equivalent forms of the logistic regression model:

Logit form  $\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$

Probability form  $\pi = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k}}$

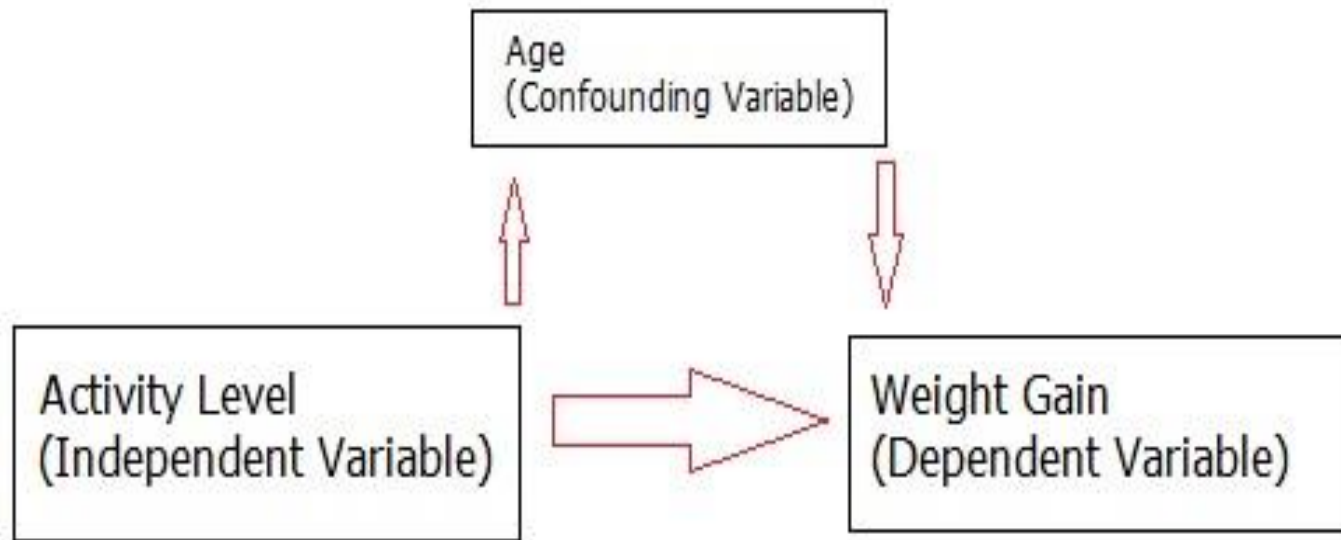


### 3. Confounding Factors

### 3. Confounding Factors

- A third variable correlated with both the dependent variable and the independent variable.
- Distortion of true effect of exposure on a disease by third factor/variable.
- The factor can cause over/under estimation of true effect. In other words it biases our study.

### 3. Confounding Factors



- A confounding variable can have a hidden effect on your experiment's outcome.

## Example 2: Wood dust, respiratory disease and smoking

- **Hypothesis and design:** Suppose that we conduct a fixed cohort study to estimate the effect of exposure to wood dust on the occurrence of chronic respiratory disease (CRD) in middle-aged, male furniture workers.
- **Potential confounder:** Since cigarette smoking is a known cause of the disease, we will control for smoking as a confounder, using stratified analysis.

## Example 3: Physical activity, coronary heart disease (CHD), and age and gender

- **Hypothesis and design:** Suppose that we conduct a cohort study to estimate the effect of physical activity level on the occurrence of CHD in a population of adults, aged 50-69.
- **Potential confounders:** Since age and sex are known risk factors for CHD, we will control for these variables as confounders, using stratified analysis. The different strata are formed from the cross-classification of both variables (covariates)–i.e., younger men, older men, younger women, and older women.

### 3. Confounding Factors: Bias

**Confounding bias** is the result of having confounding variables in your model. It has a direction, depending on if it over- or underestimates the effects of your model:

- Positive confounding is when the observed association is biased away from the null. In other words, it **overestimates** the effect.
- Negative confounding is when the observed association is biased toward the null. In other words, it **underestimates** the effect.

# 3. Confounding Factors: Adjustment

## Experimental design:

- **Randomization** : Randomly assigning experimental units to treatment groups. It balances known and unknown confounders. Limited applicability.
- **Restriction** : Using only one category for confoundings. Eg. Taking only smokers. Study results can not be generalized.
- **Matching** : Making a match of an experimental unit for different treatment groups.

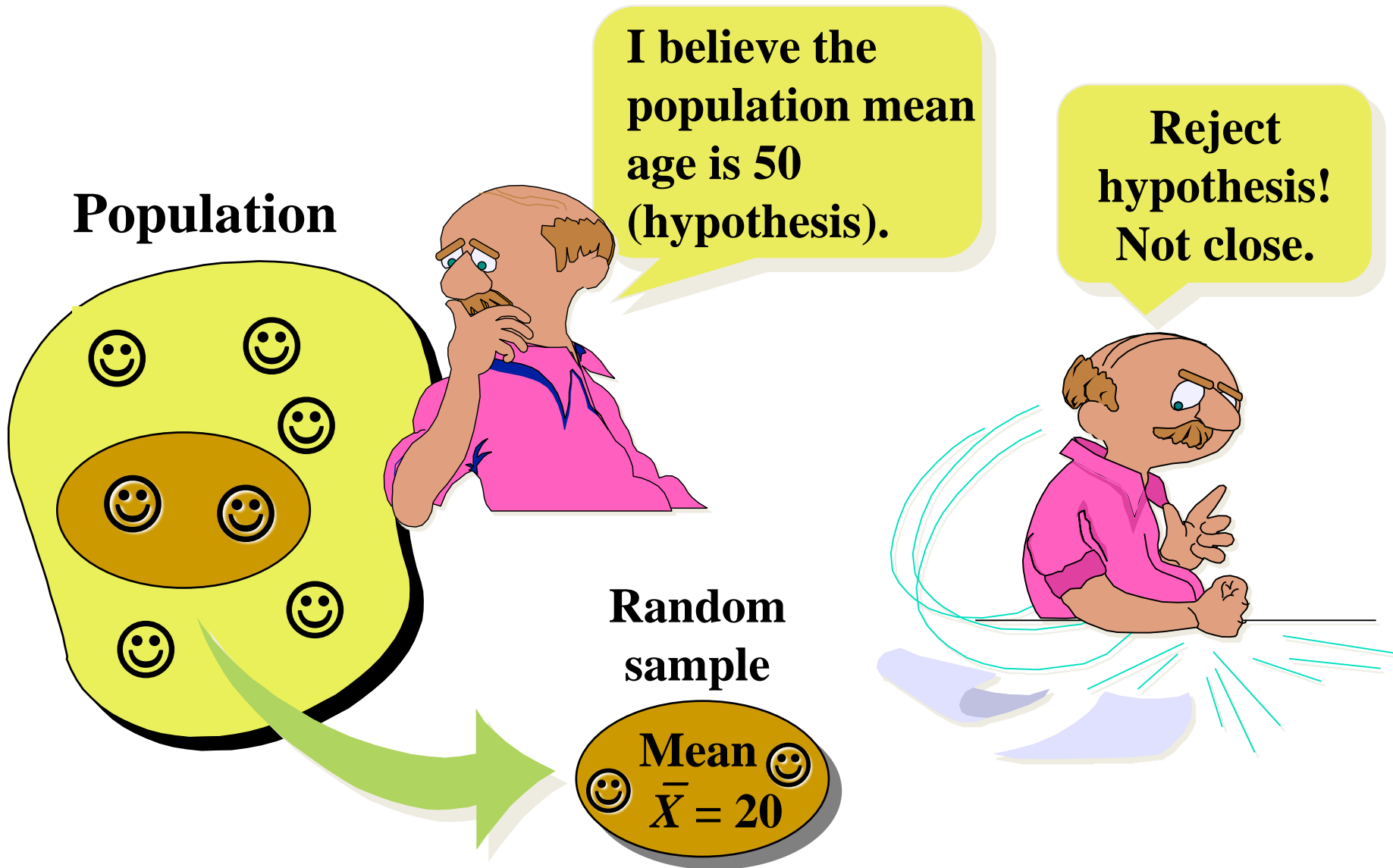
### 3. Confounding Factors: Adjustment

- If confounders are observable, then you can include them as covariates in your regression model to adjust their effects.
- Subgroup analysis. For example, gender might have some effect on your outcome, then do analysis focusing on the two groups separately.
- If unobservable, then need careful experimental design.



## 4. P-value and FDR

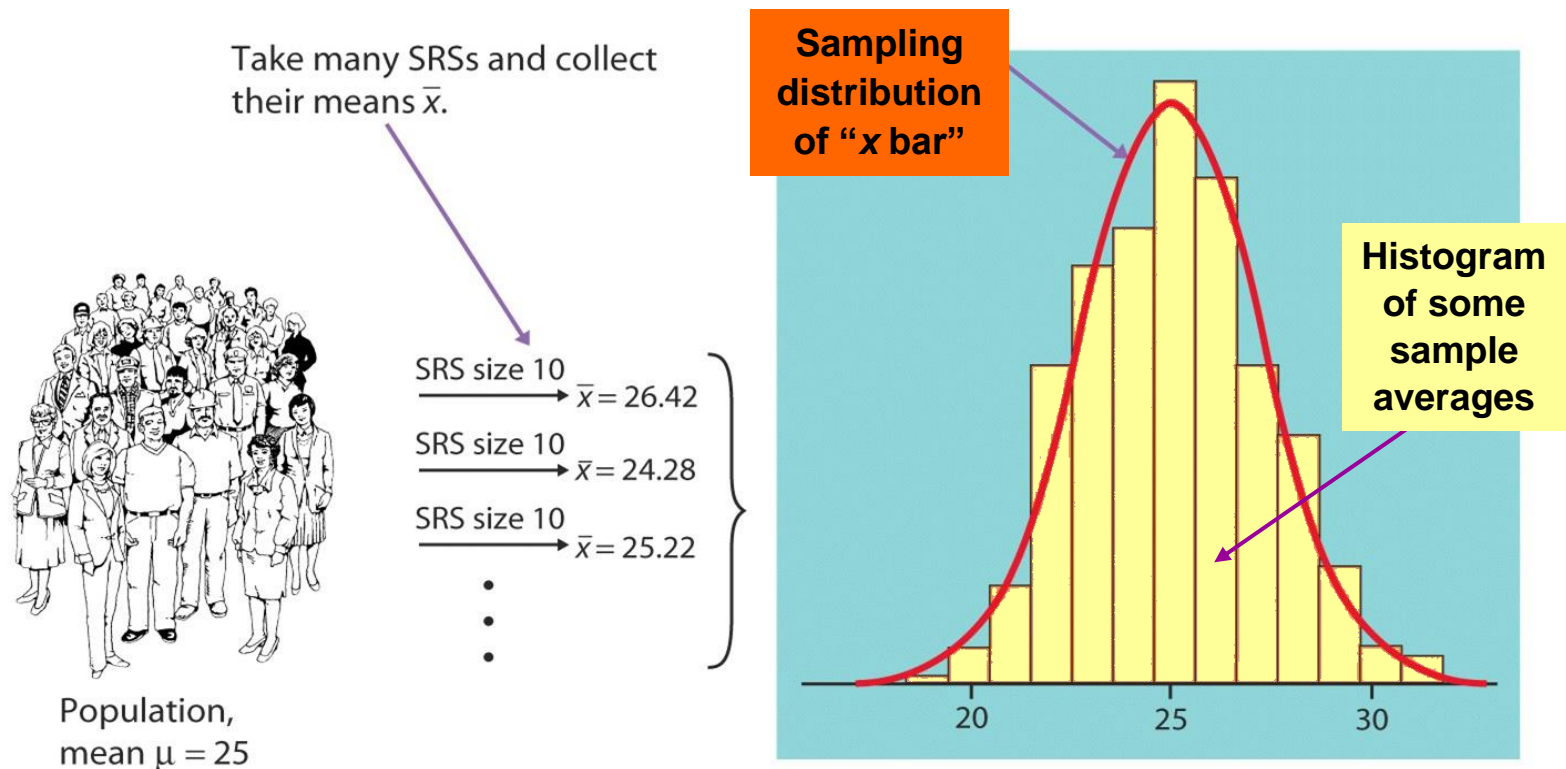
# Hypothesis Testing



# Sampling distribution of the sample mean

We take many random samples of a given size  $n$  from a population with mean  $\mu$  and standard deviation  $\sigma$ .

Some sample means will be above the population mean  $\mu$  and some will be below, making up the sampling distribution.



# Reasoning of Significance Tests

We can rely on the properties of the sample distribution to test hypotheses.

Example: You are in charge of quality control in your food company. You sample randomly four packs of cherry tomatoes, each labeled 1/2 lb. (227 g).

The average weight from your four boxes is 222 g. Obviously, we cannot expect boxes filled with whole tomatoes to all weigh exactly half a pound. Thus,

- Is the somewhat smaller weight simply due to chance variation?
- Is it evidence that the calibrating machine that sorts cherry tomatoes into packs needs revision?



# Elements of a Test of Hypothesis

1. Null hypothesis ( $H_0$ ): A theory about the specific values of one or more population parameters. The theory generally represents the status quo, which we adopt until it is proven false.
2. *Alternative (research) hypothesis* ( $H_a$ ): A theory that contradicts the null hypothesis. The theory generally represents that which we will adopt only when sufficient evidence exists to establish its truth.

# Elements of a Test of Hypothesis

3. *Test statistic*: A sample statistic used to decide whether to reject the null hypothesis.
4. *Rejection region*: The numerical values of the test statistic for which the null hypothesis will be rejected. The rejection region is chosen so that the probability is  $\alpha$  that it will contain the test statistic when the null hypothesis is true, thereby leading to a Type I error. The value of  $\alpha$  is usually chosen to be small (e.g., .01, .05, or .10) and is referred to as the **level of significance** of the test.

# Elements of a Test of Hypothesis

5. *Assumptions:* Clear statement(s) of any assumptions made about the population(s) being sampled.
6. *Experiment and calculation of test statistic:* Performance of the sampling experiment and determination of the numerical value of the test statistic.

# Elements of a Test of Hypothesis

## 7. *Conclusion:*

- a. If the numerical value of the test statistic falls in the rejection region, we reject the null hypothesis and conclude that the alternative hypothesis is true.
- b. We know that the hypothesis-testing process will lead to this conclusion incorrectly (Type I error) only  $100\alpha\%$  of the time when  $H_0$  is true.
- c. If the test statistic does not fall in the rejection region, we do not reject  $H_0$ . Thus, we reserve judgment about which hypothesis is true.
- d. We do not conclude that the null hypothesis is true because we do not (in general) know the probability  $\beta$  that our test procedure will lead to an incorrect acceptance of  $H_0$  (Type II error).



# *p*-value

- Probability of obtaining a test statistic as extreme or more extreme ( $\leq$  or  $\geq$ ) than actual sample value, **given  $H_0$  is true**
- Can be thought of as a measure of the “credibility” of the null hypothesis  $H_0$ . So the smaller the *p*-value, the stronger the evidence to against the null.
- *$\alpha$  is the nominal level of significance. This value is assumed by an analyst.*
- *p*-value is called “**observed level of significance**”. It is the smallest value of  $\alpha$  for which  $H_0$  can be rejected.
- It is used to make rejection decision
  - If *p*-value  $\geq \alpha$ , do not reject  $H_0$
  - If *p*-value  $< \alpha$ , reject  $H_0$



## Does the packaging machine need revision?

- $H_0: \mu = 227$  g versus  $H_a: \mu \neq 227$  g
- What is the probability of drawing a random sample such as yours if  $H_0$  is true?

$$\bar{x} = 222\text{g} \quad \sigma = 5\text{g} \quad n = 4$$

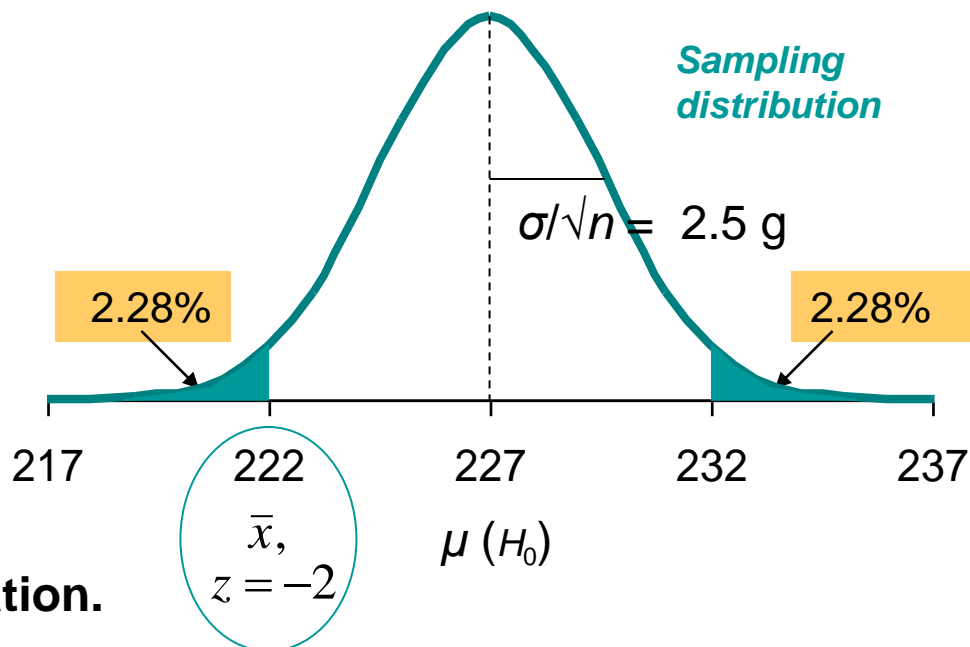
$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{222 - 227}{5/\sqrt{4}} = -2$$

From table A, the area under the standard normal curve to the left of  $z$  is 0.0228.

Thus, P-value =  $2 \times 0.0228 = 4.56\%$ .

The probability of getting a random sample average so different from  $\mu$  is so low that we reject  $H_0$ .

➔ **The machine does need recalibration.**



## 4. P-value and FDR

---

### ▣ Importance of Multiple Testing

Genomics = Lots of Data = Lots of Hypothesis Tests

- ▣ A typical microarray experiment might result in performing 10000 separate hypothesis tests. If we use a standard p-value cut-off of 0.05, we'd expect 500 genes to be deemed "significant" by chance.

False positive rate increase as the number of tests increases.

---

| m      | 1    | 5    | 10   | 50   |
|--------|------|------|------|------|
| P(F>0) | 0.05 | 0.23 | 0.40 | 0.92 |

$$\begin{aligned}P(F > 0) &= 1 - P(F = 0) \\&= 1 - \prod_{i=1}^m P(p_i > 0.05 | H_0) \\&= 1 - (0.95)^m\end{aligned}$$

For  $m=1$ ,  $P(F>0)=0.05$ . As  $m$  increases, this probability approaches to 1.

## 4. False Discovery Rate (FDR)

---


- Controlling FWER is extremely conservative We might be willing to accept A FEW false positives
- FDR = Fraction of “false significant results” among the significant results you found


□  $FDR =$

|             | Declared non-sign. | Declared sign. | Total |
|-------------|--------------------|----------------|-------|
| True $H_0$  | U                  | V              | $M_0$ |
| False $H_0$ | T                  | S              | $M_1$ |
| Total       | M-R                | R              | M     |

# False Discovery Rate Illustrated

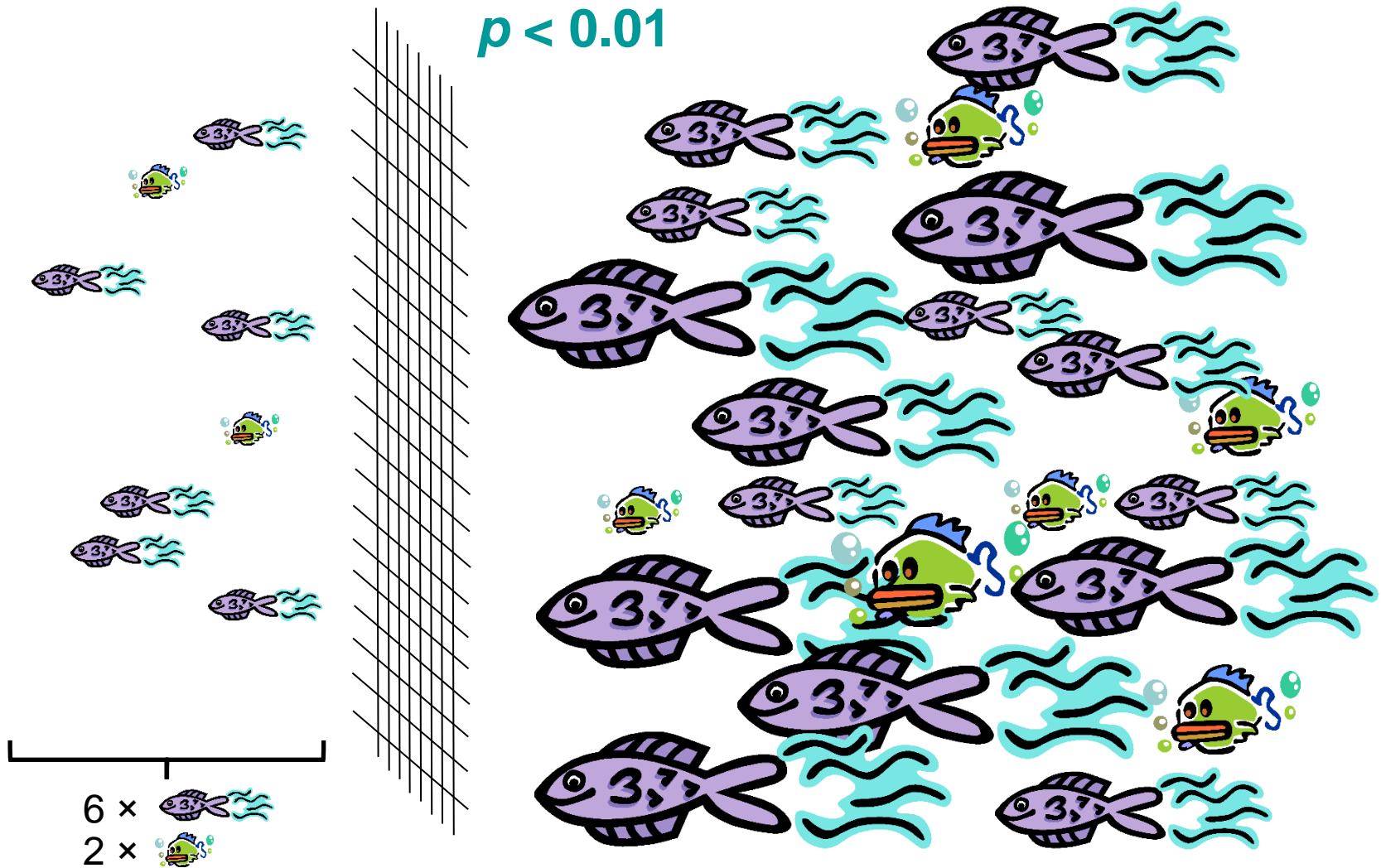
- We can, instead, attempt to minimize the *false discovery rate* (FDR)—the probability of  $H_0$  being true given a significant test.

Tests for which  $H_0$  true = 

Tests for which  $H_A$  true = 

Then, we are interested to control the FDR, i.e., for those tests called significant, what is the probability that it is actually coming from  $H_0$ .

# False Discovery Rate Illustrated



$$\Pr(H_0 \mid p < 0.01) = 0.75$$

$$\Pr(p < 0.01 \mid H_0) = 0.01$$

## 4. Benjamini and Hochberg FDR

To control FDR at level  $\delta$ :

- Order the unadjusted p-values:  $p_1 \leq p_2 \leq \dots \leq p_m$
- Then find the test with the highest rank,  $j$ , for which the p value,  $p_j$ , is less than or equal to  $(j/m) \times \delta$
- Declare the tests of rank 1, 2, ...,  $j$  as significant

$$p(j) \leq \delta \frac{j}{m}$$



## 4. B&H FDR Example

- Controlling the FDR at  $\delta = 0.05$

| Rank (j) | P-value | $(j/m) \times \delta$ | Reject $H_0$ ? |
|----------|---------|-----------------------|----------------|
| 1        | 0.0008  | 0.005                 | 1              |
| 2        | 0.009   | 0.010                 | 1              |
| 3        | 0.165   | 0.015                 | 0              |
| 4        | 0.205   | 0.020                 | 0              |
| 5        | 0.396   | 0.025                 | 0              |
| 6        | 0.450   | 0.030                 | 0              |
| 7        | 0.641   | 0.035                 | 0              |
| 8        | 0.781   | 0.040                 | 0              |
| 9        | 0.900   | 0.045                 | 0              |
| 10       | 0.993   | 0.050                 | 0              |

# Another example

The results of comparing 17 exploratory behavior measures between eight C57 and eight BALB mice

| Measure                               | Observed<br><i>P</i> -values | Rank (i) | Bonferroni<br>threshold | FDR (BH) thresholds      |
|---------------------------------------|------------------------------|----------|-------------------------|--------------------------|
| Lingering time (prop.)                | 0.000001                     | 1        | 0.0029                  | 0.0029 <i>q/m</i>        |
| Lingering speed (cm/s)                | 0.000013                     | 2        | 0.0029                  | 0.0058                   |
| Early activity in move segments (m)   | 0.000065                     | 3        | 0.0029                  | 0.0088                   |
| Early activity (m)                    | 0.00063                      | 4        | 0.0029                  | 0.0117                   |
| Spread of lingering (cm)              | 0.0008                       | 5        | 0.0029                  | 0.0147                   |
| Dynamics of activity                  | 0.0017                       | 6        | <b>0.0029</b> Bon       | 0.0176                   |
| Dynamics of diversity                 | 0.0032                       | 7        | 0.0029                  | 0.0205                   |
| Number of excursions                  | 0.0065                       | 8        | 0.0029                  | 0.0235                   |
| Movement speed (cm/s)                 | 0.0148                       | 9        | 0.0029                  | <b>0.0264</b>            |
| Spread of move segments               | 0.049                        | 10       | 0.0029                  | 0.0294                   |
| Stops per excursions (upper quartile) | 0.094                        | 11       | 0.0029                  | 0.0323                   |
| Center activity (prop.)               | 0.11                         | 12       | 0.0029                  | 0.0352                   |
| Center rest (prop.)                   | 0.15                         | 13       | 0.0029                  | 0.0382                   |
| Activity (m)                          | 0.24                         | 14       | 0.0029                  | 0.0411                   |
| Lingering activity (prop.)            | 0.45                         | 15       | 0.0029                  | 0.0441                   |
| Diversity                             | 0.56                         | 16       | 0.0029                  | 0.047                    |
| Lingering at home base (prop.)        | 0.87                         | 17       | 0.0029                  | 0.05 Start here <i>q</i> |

↑  
Reject  
 $H_0$

***m* = 17**

Compare

## 4. q-value

- q-value is defined as the minimum FDR that can be attained when calling that “feature” significant (i.e., expected proportion of false positives incurred when calling that feature significant)
- The estimated q-value is a function of the p-value for that test and the distribution of the entire set of p-values from the family of tests being considered (Storey and Tibshiriani 2003)

## 4. q-value

- Thus, in an array study testing for differential expression, if gene X has a q-value of 0.013 it means that 1.3% of genes that show p-values at least as small as gene X are false positives

## 5. Survival Analysis

# Survival Analysis

- In many medical studies, the primary endpoint is time until an event occurs (e.g. death, remission)
- Data are typically subject to **censoring** when a study ends before the event occurs
- Survival Function - A function describing the proportion of individuals surviving to or beyond a given time. Notation:
  - $T \equiv$  survival time of a randomly selected individual
  - $t \equiv$  a specific point in time.
  - $S(t) = P(T > t) \equiv$  Survival Function
  - $\lambda(t) \equiv$  instantaneous failure rate at time  $t$  aka hazard function

# Kaplan-Meier Estimate of Survival Function

- Case with no censoring during the study
  - Identify the observed failure times:  $t_{(1)} < \dots < t_{(k)}$
  - Number of individuals at risk before  $t_{(i)} \equiv n_i$
  - Number of individuals with failure time  $t_{(i)} \equiv d_i$
  - Estimated hazard function at  $t_{(i)}$ :

$$\hat{\lambda}_i = \frac{d_i}{n_i}$$

- Estimated Survival Function at time  $t$

$$\hat{S}(t) = \prod_{i|t_{(i)} \leq t} (1 - \hat{\lambda}_i) = \frac{\# \text{ with } T > t}{n}$$

(when no censoring)

# Example - Navelbine/Taxol vs Leukemia

- Mice given P388 murine leukemia assigned at random to one of two regimens of therapy
  - Regimen A - Navelbine + Taxol Concurrently
  - Regimen B - Navelbine + Taxol 1-hour later
- Under regimen A, 9 of  $n_A=49$  mice died on days: 6,8,22,32,32,35,41,46, and 54. Remainder  $> 60$  days
- Under regimen B, 9 of  $n_B=15$  mice died on days: 8,10,27,31,34,35,39,47, and 57. Remainder  $> 60$  days



# Example - Navelbine/Taxol vs Leukemia

Regimen A

Regimen B

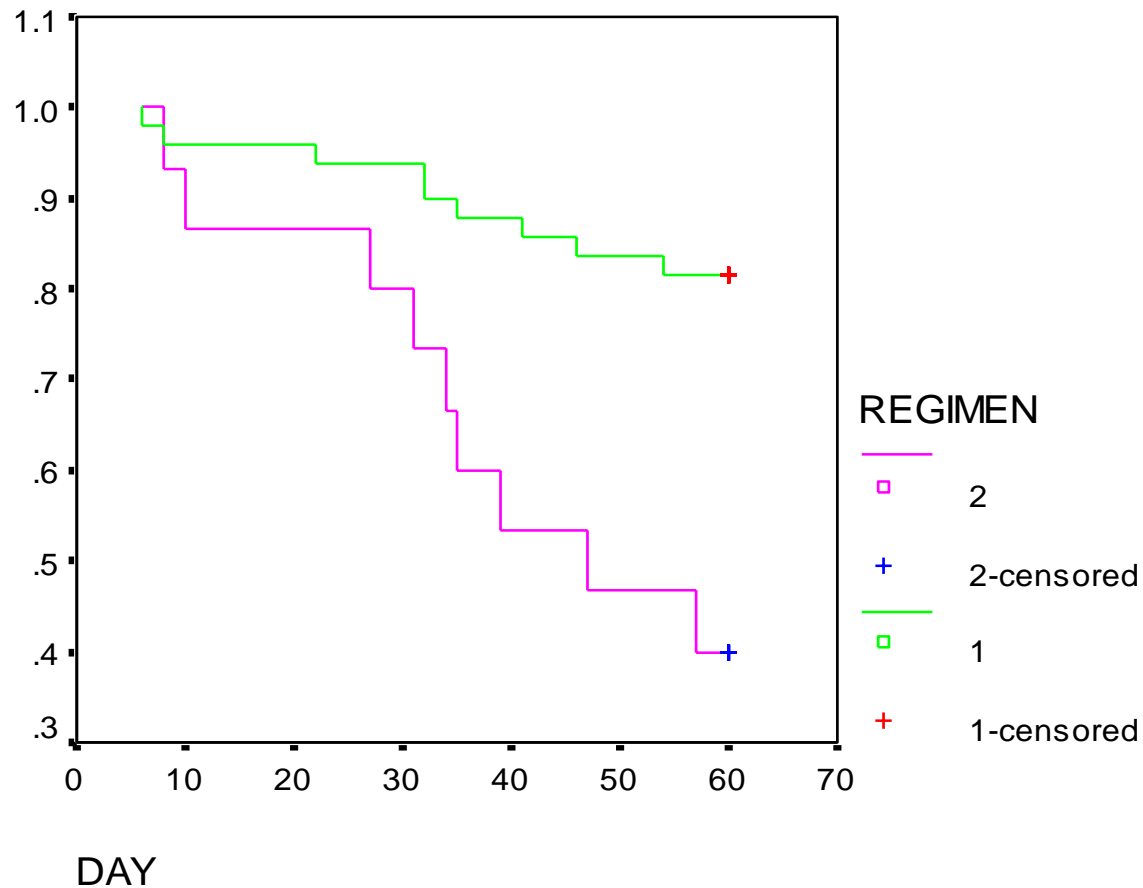
| $i$ | $t_{(i)}$ | $n_i$ | $d_i$ | $\lambda_i$ | $S(t_{(i)})$ | $i$ | $t_{(i)}$ | $n_i$ | $d_i$ | $\lambda_i$ | $S(t_{(i)})$ |
|-----|-----------|-------|-------|-------------|--------------|-----|-----------|-------|-------|-------------|--------------|
| 1   | 6         | 49    | 1     | .020        | .980         | 1   | 8         | 15    | 1     | .067        | .933         |
| 2   | 8         | 48    | 1     | .021        | .959         | 2   | 10        | 14    | 1     | .071        | .867         |
| 3   | 22        | 47    | 1     | .021        | .939         | 3   | 27        | 13    | 1     | .077        | .800         |
| 4   | 32        | 46    | 2     | .043        | .899         | 4   | 31        | 12    | 1     | .083        | .733         |
| 5   | 35        | 44    | 1     | .023        | .878         | 5   | 34        | 11    | 1     | .091        | .667         |
| 6   | 41        | 43    | 1     | .023        | .858         | 6   | 35        | 10    | 1     | .100        | .600         |
| 7   | 46        | 42    | 1     | .024        | .837         | 7   | 39        | 9     | 1     | .111        | .533         |
| 8   | 54        | 41    | 1     | .024        | .817         | 8   | 47        | 8     | 1     | .125        | .467         |
|     |           |       |       |             |              | 9   | 57        | 7     | 1     | .143        | .400         |

$$\hat{\lambda}_1^A = \frac{1}{49} = .020 \quad \hat{S}^A(6) = 1 - .020 = .980$$

$$\hat{\lambda}_2^A = \frac{1}{48} = .021 \quad \hat{S}^A(8) = .980(1 - .021) = .959$$

# Example - Navelbine/Taxol vs Leukemia

Survival Functions



## 5. KM Plot: example 2

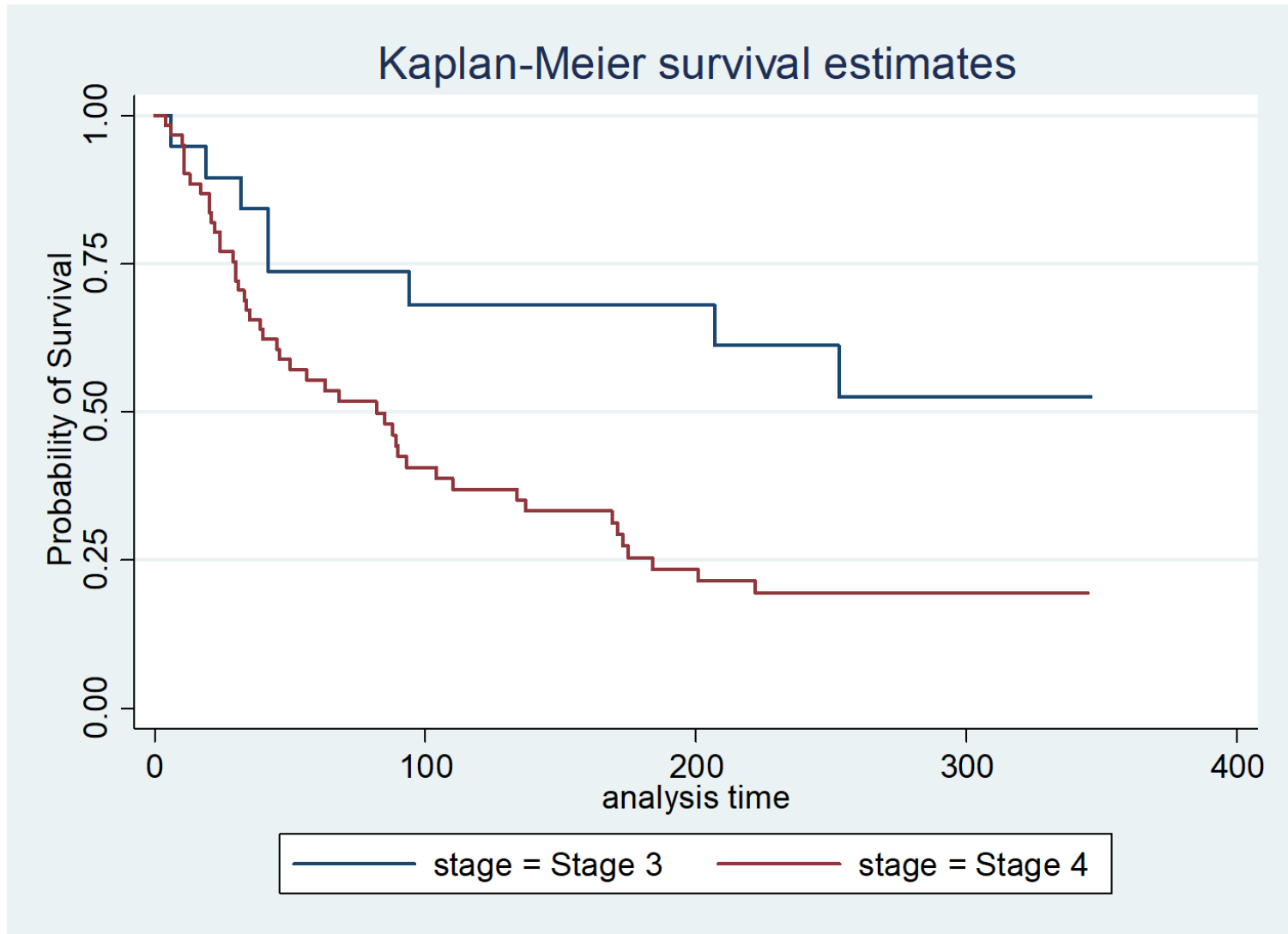
### Survival in lymphoma patients

- Armitage et al. (2002: p. 579) discuss the following data on patient survival after recruitment into a clinical of patients with diffuse histiocytic lymphoma (KcKelvey et al. Cancer 1976; 38: 1484 – 93).

# 5. Survival Analysis: KM Plot

| Follow-up (days) |                          |     |     |     |                           |     |     |     |
|------------------|--------------------------|-----|-----|-----|---------------------------|-----|-----|-----|
|                  | Dead at end of follow-up |     |     |     | Alive at end of follow-up |     |     |     |
| Stage 3          |                          |     |     |     |                           |     |     |     |
|                  | 6                        | 19  | 32  | 42  | 43                        | 126 | 169 | 211 |
|                  | 42                       | 94  | 207 | 253 | 227                       | 255 | 270 | 310 |
|                  |                          |     |     |     | 316                       | 335 | 346 |     |
| Stage 4          |                          |     |     |     |                           |     |     |     |
|                  | 4                        | 6   | 10  | 11  | 41                        | 43  | 61  | 61  |
|                  | 11                       | 11  | 13  | 17  | 160                       | 235 | 247 | 260 |
|                  | 20                       | 20  | 21  | 22  | 284                       | 290 | 291 | 302 |
|                  | 24                       | 24  | 29  | 30  | 304                       | 341 | 345 |     |
|                  | 30                       | 31  | 33  | 34  |                           |     |     |     |
|                  | 35                       | 39  | 40  | 45  |                           |     |     |     |
|                  | 46                       | 50  | 56  | 63  |                           |     |     |     |
|                  | 68                       | 82  | 85  | 88  |                           |     |     |     |
|                  | 89                       | 90  | 93  | 104 |                           |     |     |     |
|                  | 110                      | 134 | 137 | 169 |                           |     |     |     |
|                  | 171                      | 173 | 175 | 184 |                           |     |     |     |
|                  | 201                      | 222 |     |     |                           |     |     |     |

# 5. Survival Analysis: KM Plot



# 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)

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

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

Can take on any form

## 5. Survival Analysis: Hazard Ratio

- Hazard ratio (HR): the ratio of the hazard rates corresponding to the conditions described by two levels of an explanatory variable.

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

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

## 5. Survival Analysis: Hazard Ratio

- $HR(T \text{ vs. } C) > 1$ : The treatment group experiences a higher hazard over the control group(control group is favored)
- $HR(T \text{ vs. } C) = 1$ : No difference between the treatment and the control group
- $HR(T \text{ vs. } C) < 1$ : The control group experiences a higher hazard over the treatment group(treatment group is favored)



## 5. Survival Analysis: Hazard Ratio

### Example

- $HR(T \text{ vs. } C) = 1.3$ , the hazard of the treatment group is increased by 30% compared to the control group and for  $HR(T \text{ vs. } C) = 0.7$  the hazard of the treatment group is by 30% decreased compared to the control group