Linear Regression Analysis

Lecture 1- Prelude

周在莹

清华大学统计与数据科学系

zyzhou@mail.tsinghua.edu.cn



本节揭晓：

- 为什么叫“回归”？
- 学习这门课程需要知道哪些基本“术语”？



Outline

- 1. History
- 2. Before You Started
  - 2.1 EDA
  - 2.2 Terminologies
  - 2.3 Possible Objectives
  - 2.4 General Steps

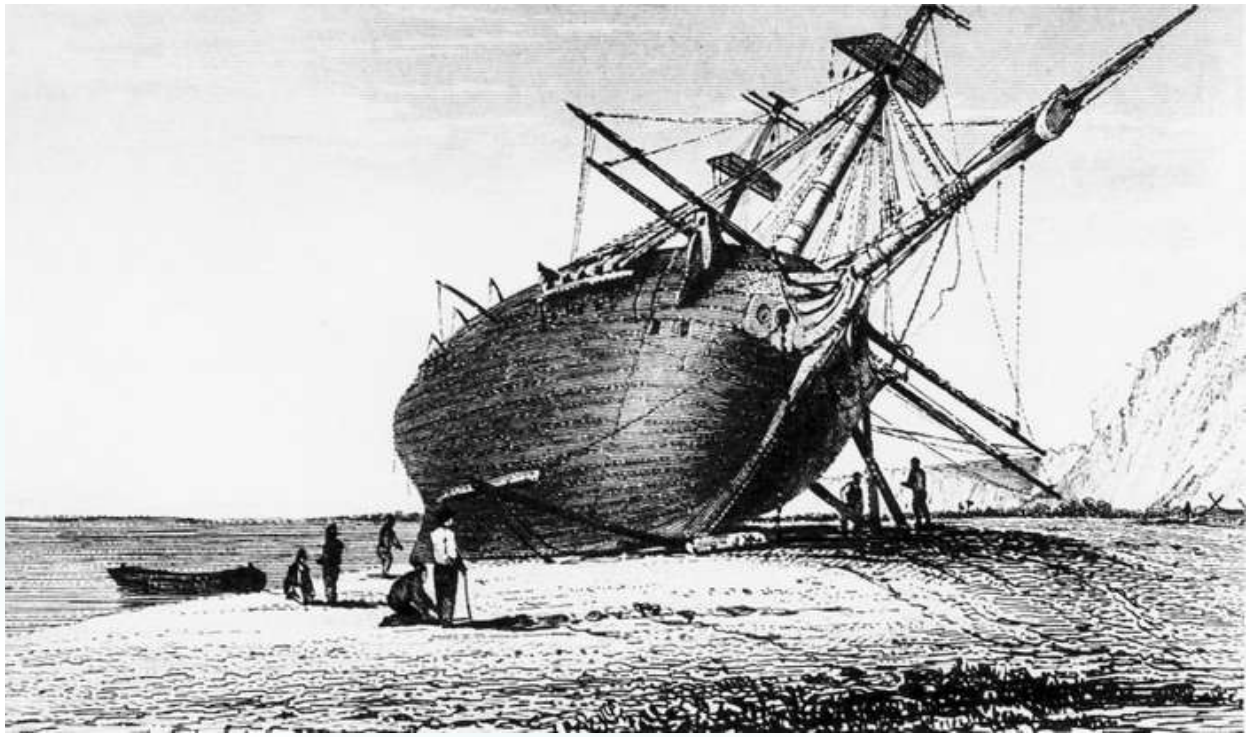


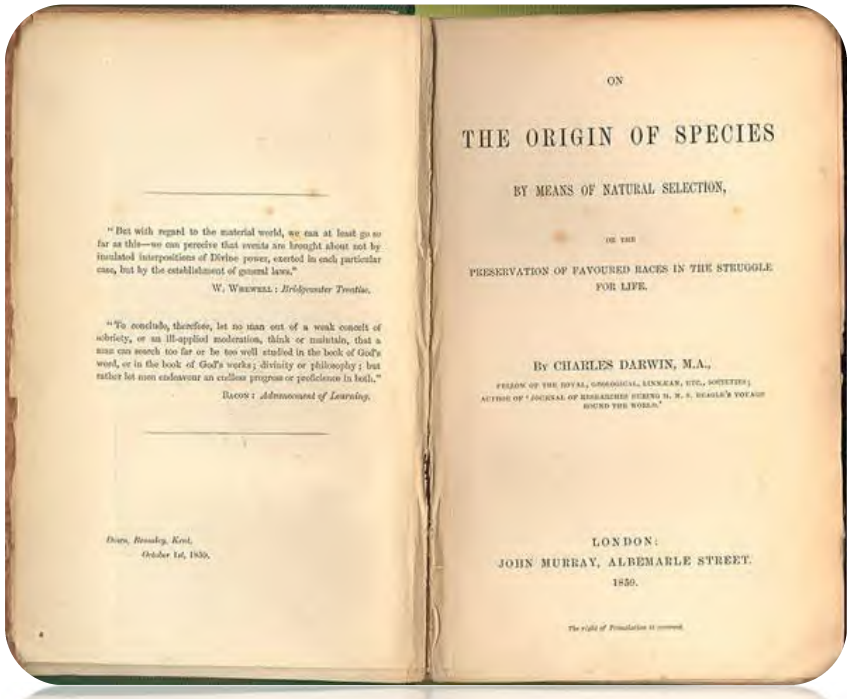
Background Reading

- Regression-type problems first considered in the 18th century to aid navigation with the use of astronomy
- Legendre developed the method of least squares in 1805
- Gauss claimed to have developed the method a few years earlier and in 1809 showed that least squares is the optimal solution when the errors are normally distributed
- Used almost exclusively in the physical sciences until later in the 19th century



The Big Bang





ON

THE ORIGIN OF SPECIES

BY MEANS OF NATURAL SELECTION,

OR THE

PRESERVATION OF FAVOURED RACES IN THE STRUGGLE FOR LIFE.

By CHARLES DARWIN, M.A.,

LONDON:
JOHN MURRAY, ALBEMARLE STREET.
1859.

"But with regard to the material world, we can at least go so far as this—we can perceive that events are brought about not by insulated interpositions of Divine power, exerted in each particular case, but by the establishment of general laws."

W. WHEWELL: *Bridgewater Treatise.*

"To conclude, therefore, let no man out of a weak conceit of sobriety, or an ill-applied moderation, think or maintain, that a man can search too far or be too well studied in the book of God's word, or in the book of God's works; divinity or philosophy; but rather let men endeavour an endless progress or proficience in both."

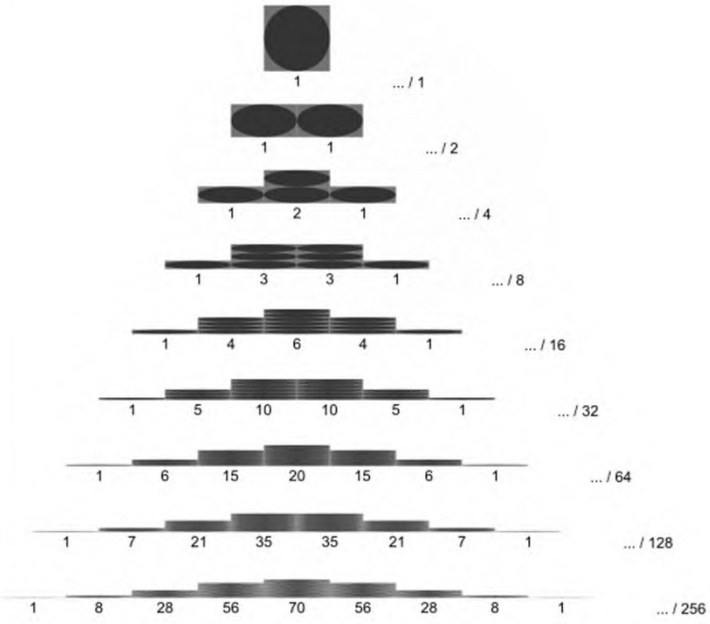
BACON: *Advancement of Learning.*

Down, Bromley, Kent,
October 1st, 1859.

Galton's Quincunx





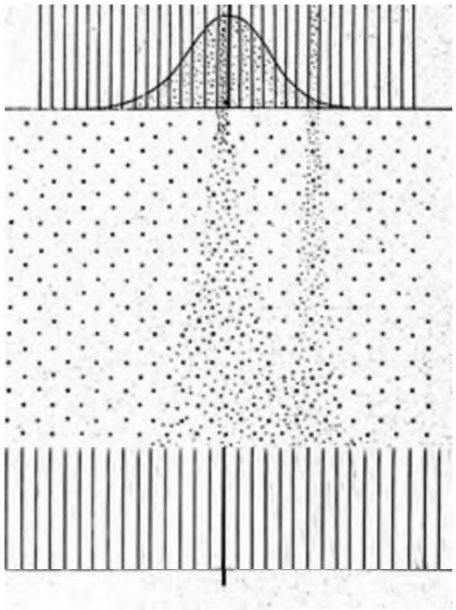


单选题 1分

设置

7

假设某个种群，初始父代某个特征的分布为一个正态分布，则依据该模型，落入槽中所形成的子代分布与父代相同吗?



是，种群特征稳定，因此分布至少相差不大

不是，钉子的层数越多，子代的正态分布越分散

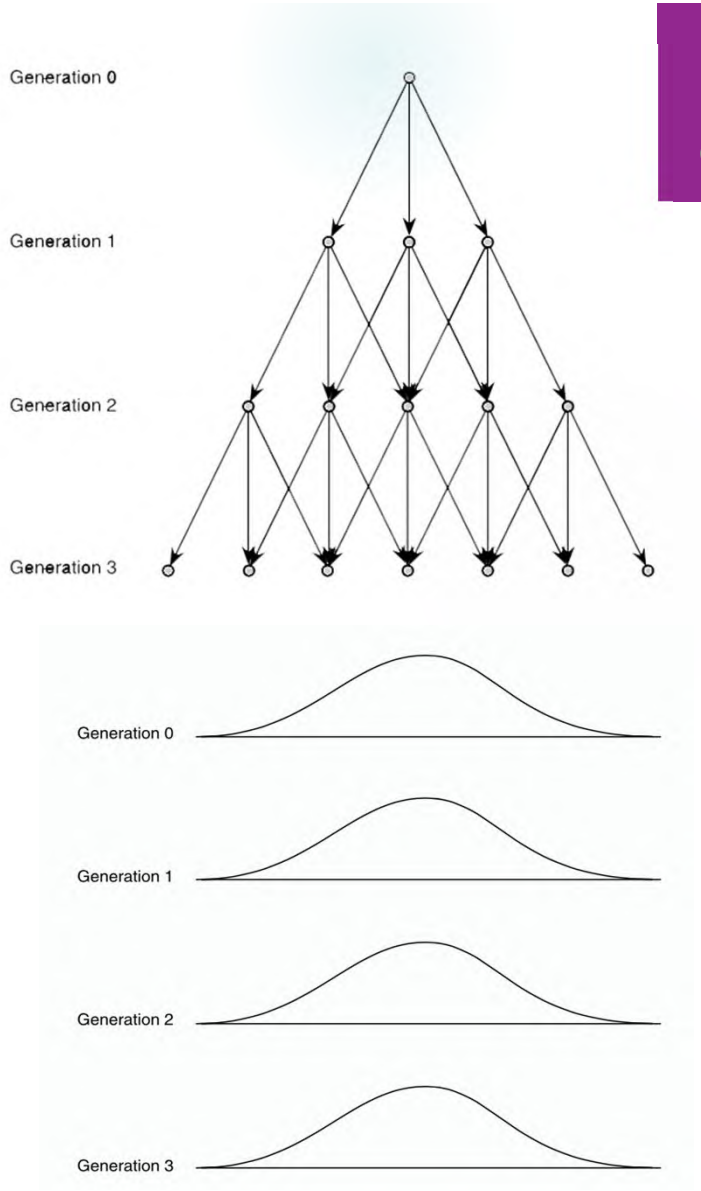
不是，钉子的层数越多，子代的正态分布越集中

不是，子代压根不一定是正态分布

提交

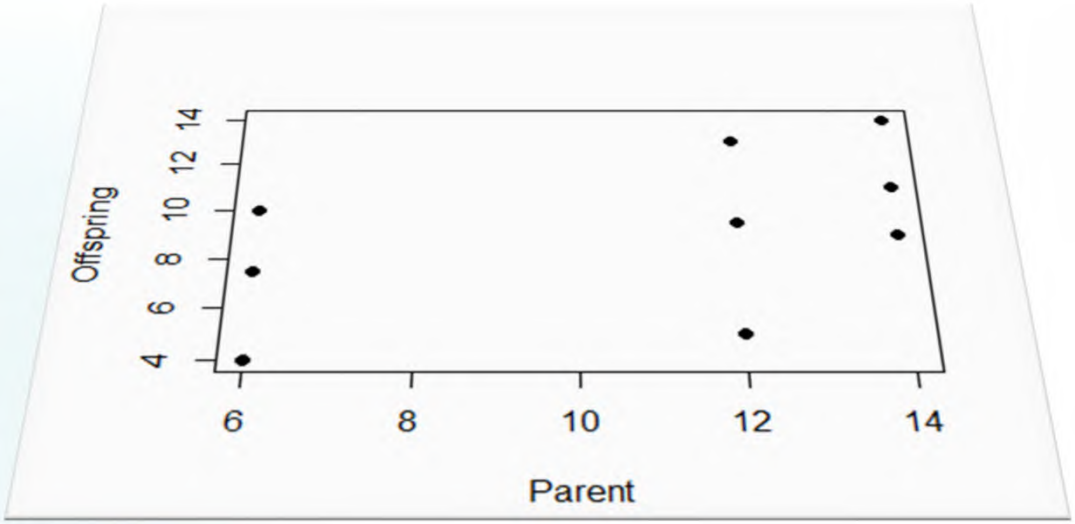
The Lurking Problem

- Intergenerational transfers passed heritable variation to offspring
  - The same parent would produce offspring with different, heritable traits
- Population dispersion is stable over the short term
  - Indeed this stability is essential for the very definition of a species



The Sweet Peas Experiment from 1875

- Weights of the daughter seeds (Offspring) against the weights of the mother seeds (Parent)



- Least-squares method of estimation by Gauss (1809) and Legendre (1806)



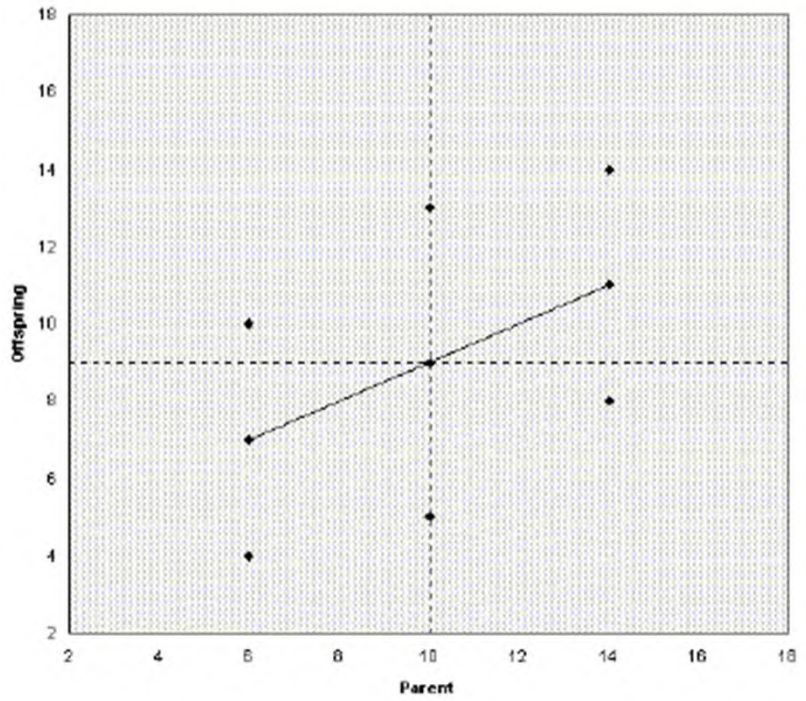
填空题 1分

设置

10

取每个母代对应后代的中位数，大致连成如图一条直线。

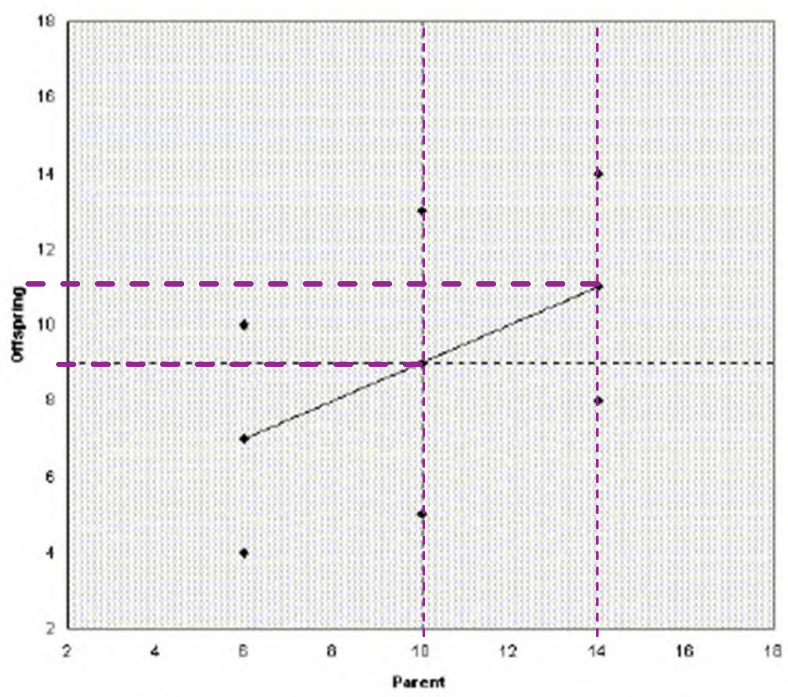
如果该直线斜率是 [填空1]，则不支持 Galton 的“品质”可遗传假说。



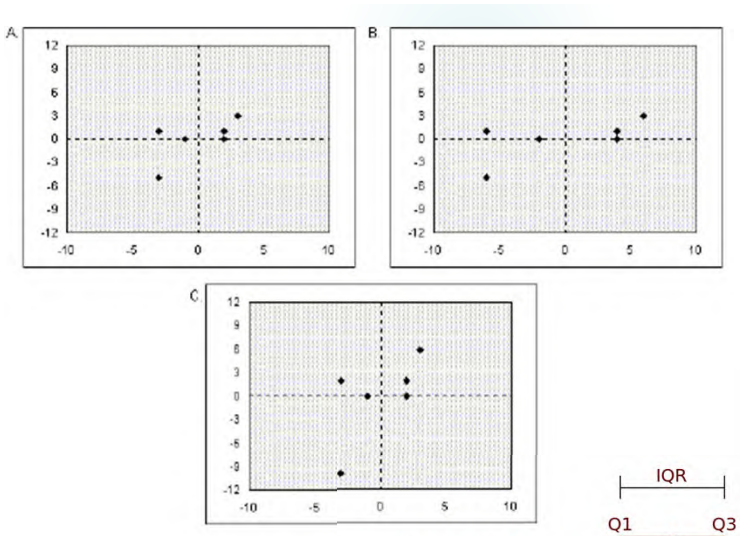
作答

Regression & Slope

- Galton's first documented study of this type suggested a slope of 0.33
  - Affirmed his basic assumptions concerning the heritability of "characters"
- This finding is prototypical of regression to the mean
  - He designated the slope '*r*' for regression



Generality of Regression Slope



- Galton believed he had estimated a single, universal heredity constant

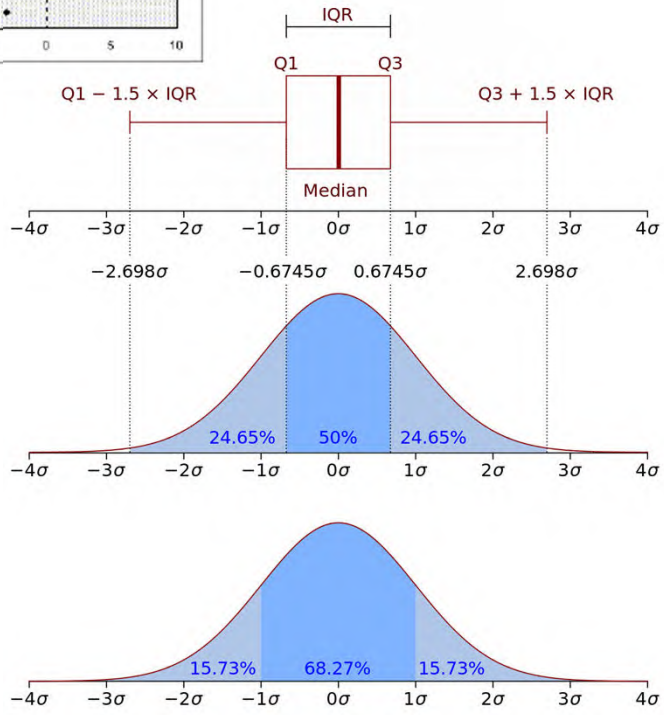
$$y - \bar{y} = r(x - \bar{x})$$

- Different sets of measurements!

$$y - \bar{y} = r\,(S_y/S_x)(x - \bar{x})$$

  - Used an estimated value of r
  - (S_y/S_x) allowed him to adjust the observed slope

Quantile Deviation：semi-interquartile range

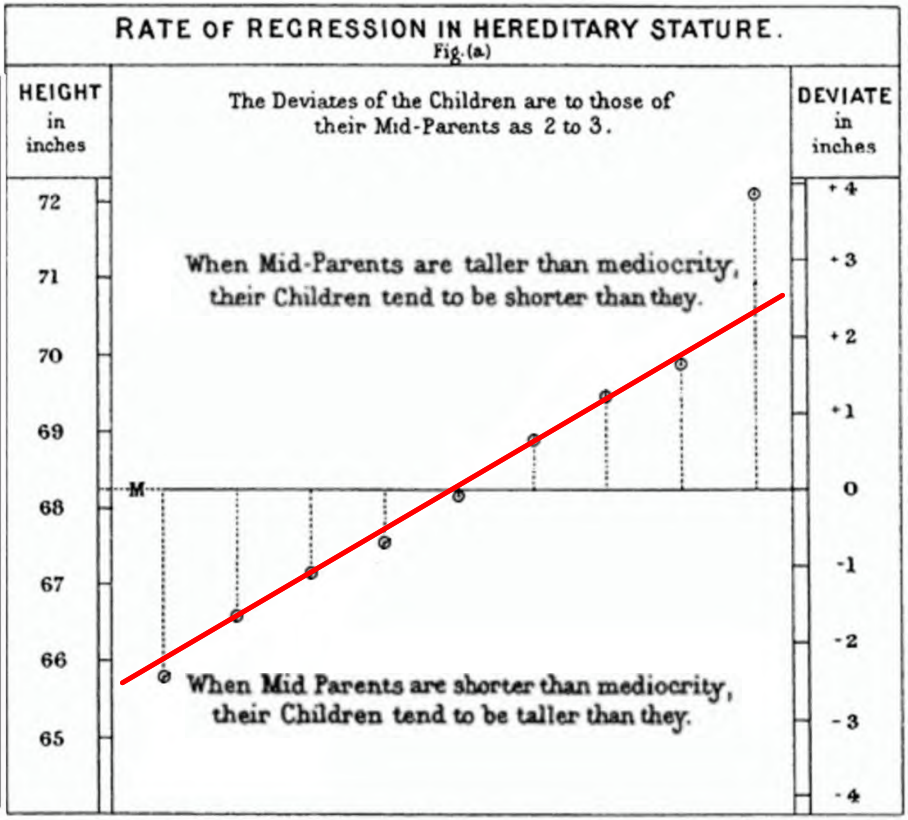


Data on Family Heights

NUMBER OF ADULT CHILDREN OF VARIOUS STATURES BORN OF 205 MID-PARENTS OF VARIOUS STATURES.
(All Female heights have been multiplied by 1·08).

Heights of the Mid-parents in inches.	Heights of the Adult Children.														Total Number of		Medians.
	Below	62·2	63·2	64·2	65·2	66·2	67·2	68·2	69·2	70·2	71·2	72·2	73·2	Above	Adult Children.	Mid-parents.	
Above ..	..	..	..	..	..	..	..	..	..	..	..	1	3	..	4	5	..
72·5	..	..	..	..	..	..	..	1	2	1	2	7	2	4	19	6	72·2
71·5	..	..	..	..	1	3	4	3	5	10	4	9	2	2	43	11	69·9
70·5	1	..	1	..	1	1	3	12	18	14	7	4	3	3	68	22	69·5
69·5	..	..	1	16	4	17	27	20	33	25	20	11	4	5	183	41	68·9
68·5	1	..	7	11	16	25	31	34	48	21	18	4	3	..	219	49	68·2
67·5	..	3	5	14	15	36	38	28	38	19	11	4	..	..	211	33	67·6
66·5	..	3	3	5	2	17	17	14	13	4	..	..	..	..	78	20	67·2
65·5	1	..	9	5	7	11	11	7	7	5	2	1	..	..	66	12	66·7
64·5	1	1	4	4	1	5	5	..	2	..	..	..	..	..	23	5	65·8
Below ..	1	..	2	4	1	2	2	1	1	..	..	..	..	..	14	1	..
Totals ..	5	7	32	59	48	117	138	120	167	99	64	41	17	14	928	205	..
Medians ..	..	..	66·3	67·8	67·9	67·7	67·9	68·3	68·5	69·0	69·0	70·0	..	..	..	..	..

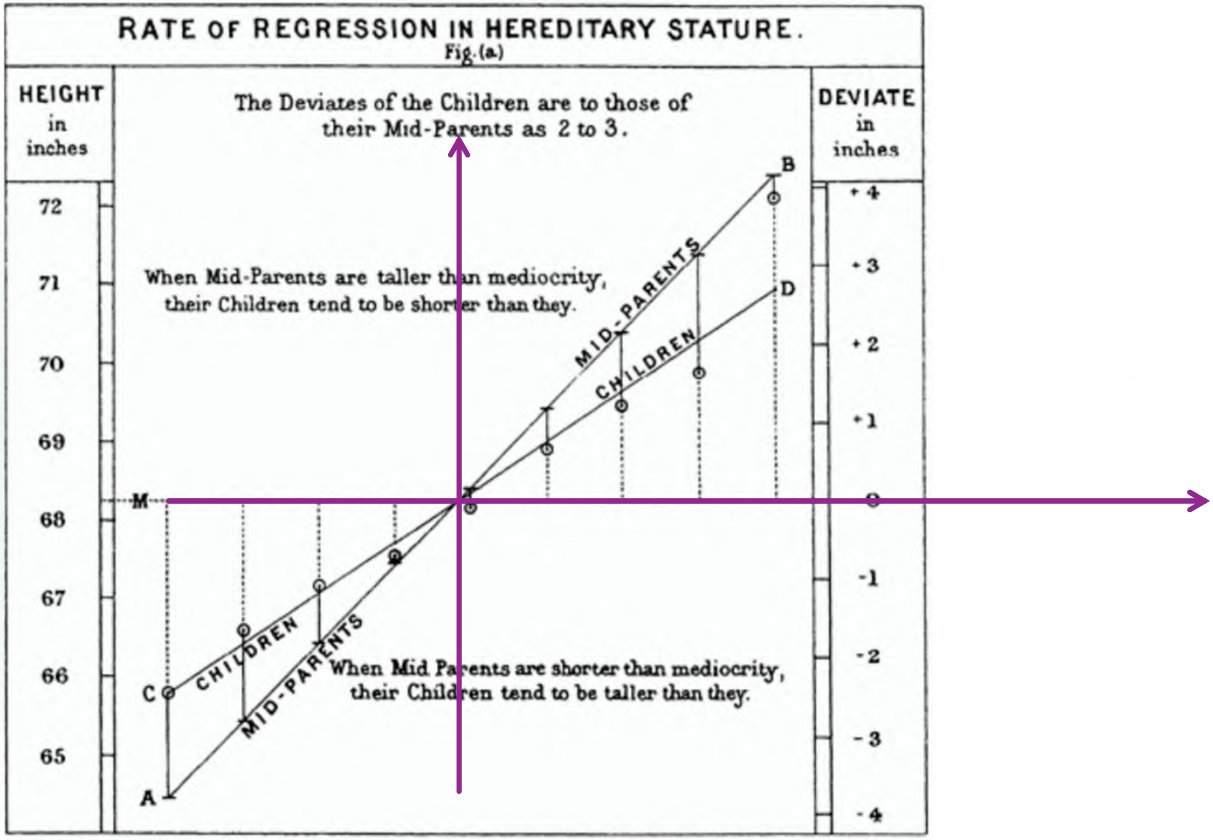
NOTE.—In calculating the Medians, the entries have been taken as referring to the middle of the squares in which they stand. The reason why the headings run 62·2, 63·2, &c., instead of 62·5, 63·5, &c., is that the observations are unequally distributed between 62 and 63, 63 and 64, &c., there being a strong bias in favour of integral inches. After careful consideration, I concluded that the headings, as adopted, best satisfied the conditions. This inequality was not apparent in the case of the Mid-parents.



- Regression effect

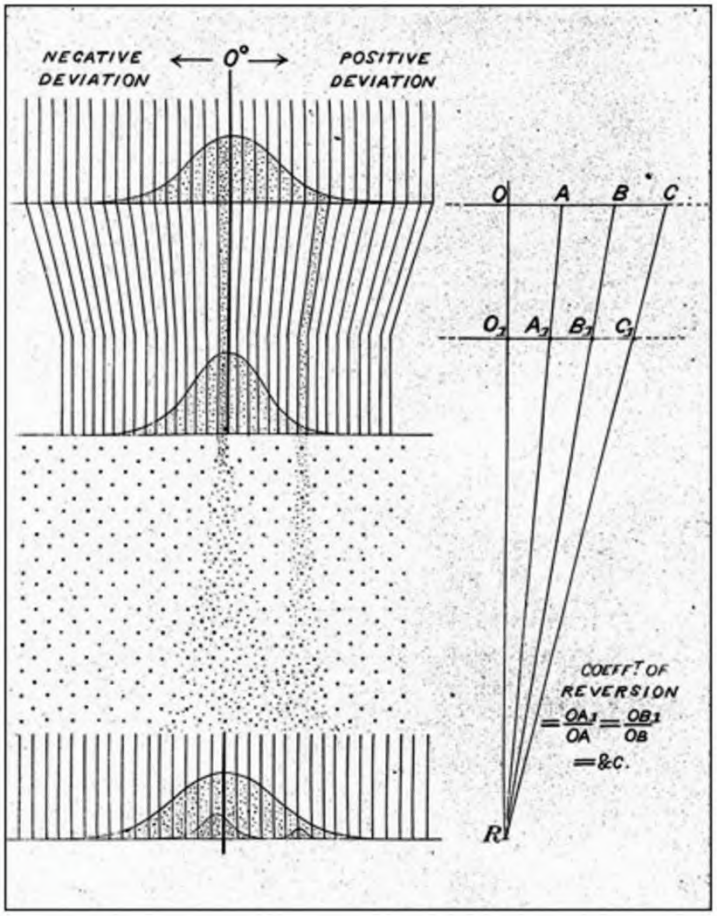
The Regression Line

- ‘Regression towards mediocrity’



Introducing the Inclined Chutes

- Galton called this reversion
  - Not coincidental but necessary for population stability
- His 1877 version of the quincunx, showing:
  - how the inclined chutes compensate for the increase in dispersion below to maintain constant population dispersion
  - how the offspring of two of the upper-level stature groups can be traced through the process to the lowest level

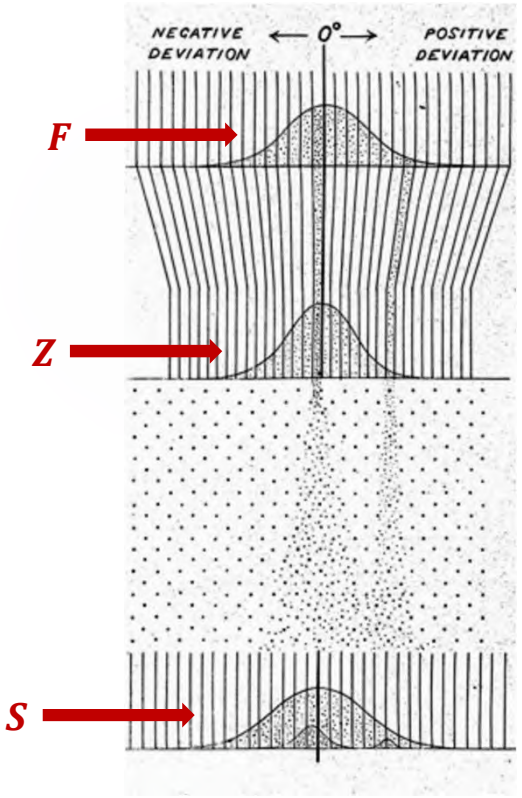


Math behind the Scene

- $F \sim N(0, \sigma_1^2)$
- What do the inclined chutes do?
  - $Z = cF \sim N(0, c^2\sigma_1^2)$, with $c < 1$
- What do the rows of pegs do?
  - X_i: change caused by row i

$$\{X_i\} \text{ are i.i.d.} \Rightarrow X = \sum_{i=1}^{n} X_i \xrightarrow{d} N(0, \sigma^2)$$

  - $Var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} Var(X_i) \uparrow$
- $S = Z + X = F$



单选题 1分

设置

$$y - \bar{y} = r\,(S_y/S_x)(x - \bar{x})$$

17

假设 Galton的助手不小心弄反了 ‘mid-parents’ 和 ‘children’ 的标签，则 Galton 会得出回归线的斜率是多少呢？还会发现“回归”现象吗？

NUMBER OF ADULT CHILDREN OF VARIOUS STATURES BORN OF 205 MID-PARENTS OF VARIOUS STATURES.
(All Female heights have been multiplied by 1·08).

Heights of the Mid-parents in inches.	Heights of the Adult Children.													Total Number of		Medians.
	Below	62·2	63·2	64·2	65·2	66·2	67·2	68·2	69·2	70·2	71·2	72·2	73·2	Above	Adult Children.	Mid-parents.	
Above ..	..	..	..	..	..	..	..	..	..	..	..	1	3	..	4	5	..
72·5	..	..	..	..	..	..	..	1	2	1	2	7	2	4	19	6	72·2
71·5	..	..	..	..	1	3	4	3	5	10	4	9	2	2	43	11	69·9
70·5	1	..	1	..	1	1	3	12	18	14	7	4	3	3	68	22	69·5
69·5	..	..	1	16	4	17	27	20	33	25	20	11	4	5	183	41	68·9
68·5	1	..	7	11	16	25	31	34	48	21	18	4	3	..	219	49	68·2
67·5	..	3	5	14	15	36	38	28	38	19	11	4	..	..	211	33	67·6
66·5	..	3	3	5	2	17	17	14	13	4	..	..	..	..	78	20	67·2
65·5	1	..	9	5	7	11	11	7	7	5	2	1	..	..	66	12	66·7
64·5	1	1	4	4	1	5	5	..	2	..	..	..	..	..	23	5	65·8
Below ..	1	..	2	4	1	2	2	1	1	..	..	..	..	..	14	1	..
Totals ..	5	7	32	59	48	117	138	120	167	99	64	41	17	14	928	205	..
Medians ..	..	..	66·3	67·8	67·9	67·7	67·9	68·3	68·5	69·0	69·0	70·0	..	..	..	..	..

NOTE.—In calculating the Medians, the entries have been taken as referring to the middle of the squares in which they stand. The reason why the headings run 62·2, 63·2, &c., instead of 62·5, 63·5, &c., is that the observations are unequally distributed between 62 and 63, 63 and 64, &c., there being a strong bias in favour of integral inches. After careful consideration, I concluded that the headings, as adopted, best satisfied the conditions. This inequality was not apparent in the case of the Mid-parents.

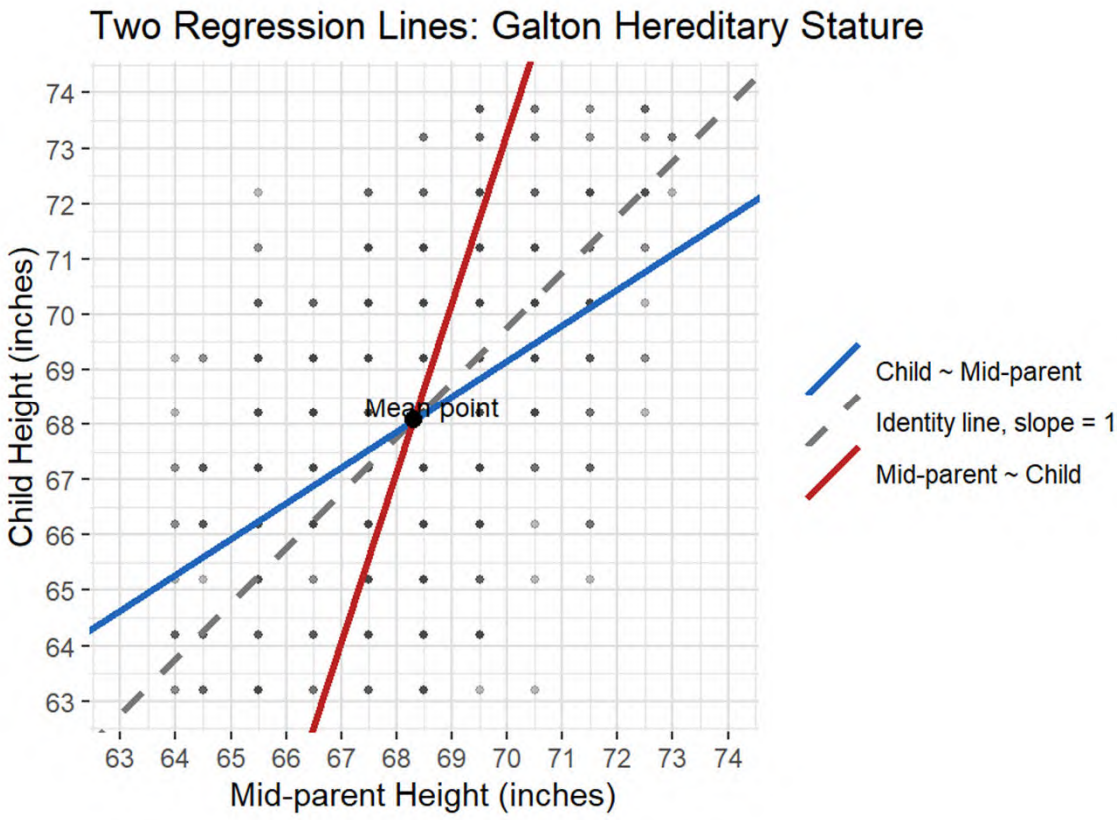
会

不会

提交

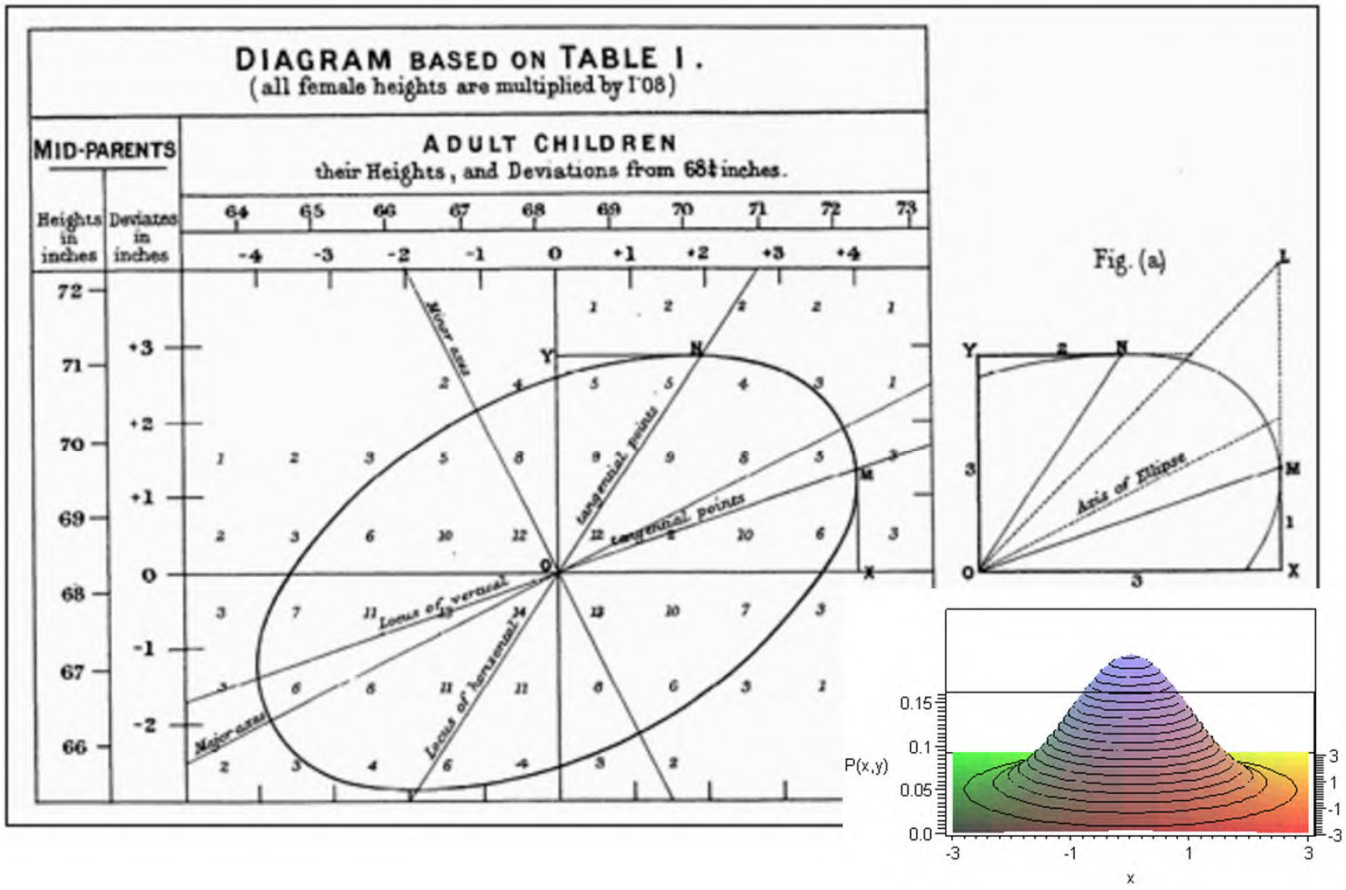
Two Regression Lines Together

- 'Regression towards mediocrity'
- Similar results for brothers data
- The explanation must be statistical, not biological



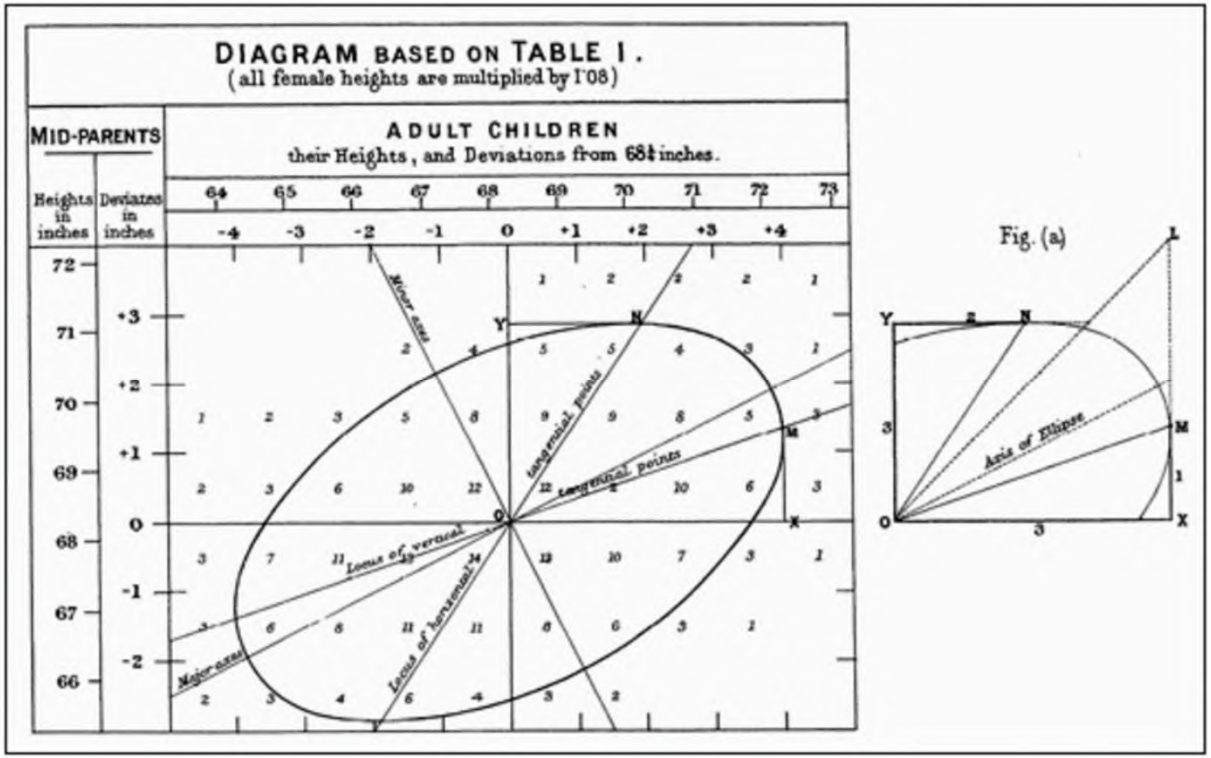
Geometric Relationship of the Data

- Galton smoothed the counts to make the pattern stand out better
- Could see a rough elliptical contour emerging around the densest portion



Bivariate Normal Density & Regression Lines

- Galton found a theoretical representation for the table- the bivariate normal density
- Two "regression lines":
  - OM- the expected height of the mid-parent as a function of the child's height
  - ON- the expected height of the offspring as a function of the mid-parent's height
- $E(Y|X = x)$



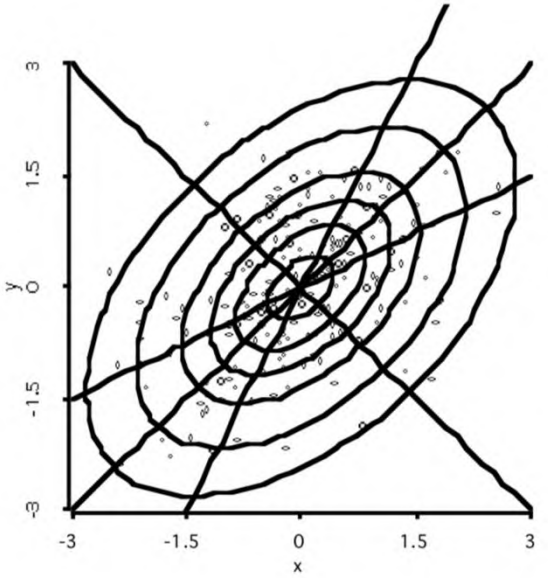
主观题 10分

设置

21

- X and Y have a standard bivariate normal distribution with $\rho = 0.5$
- Here is a picture of a sample of 200 from a standard bivariate normal distribution with $\rho = 0.5$. Also shown in the picture are:
  - Some level curves for the pdf
  - The axes of the ellipse
  - The line showing $E(Y|X = x)$ as a function of x
  - The line showing $E(X|Y = y)$ as a function of y

L2

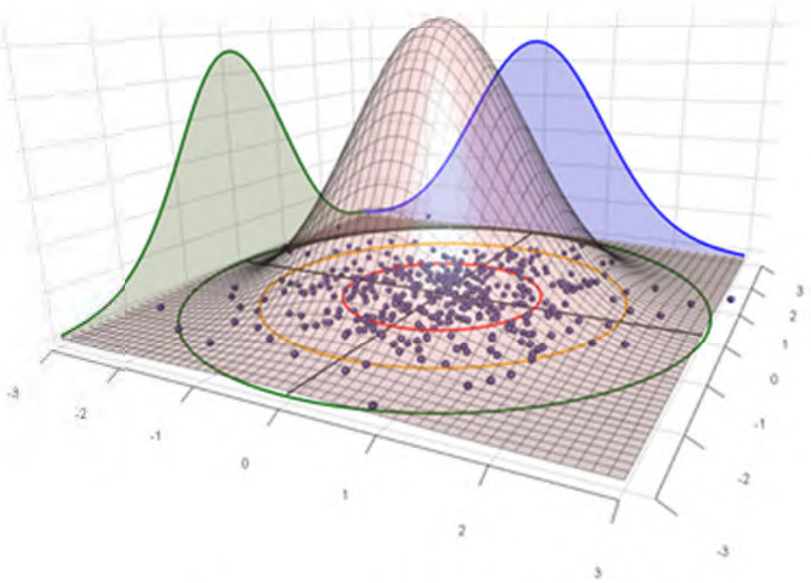


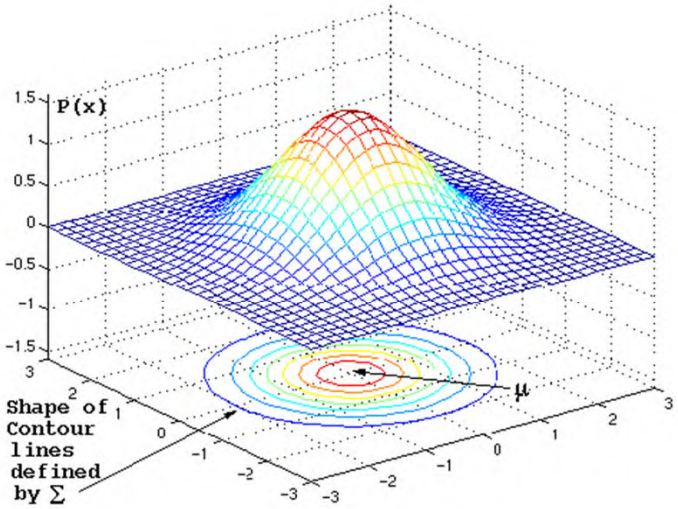
L1

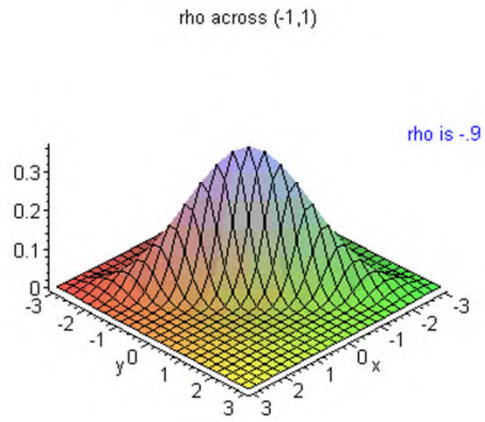
- Q1. Which line is $E(Y|X = x)$?
- Q2. Which line is $E(X|Y = y)$?
- Q3. What happens if $\rho = 0$?
- Q4. What happens if $\rho = 1$?

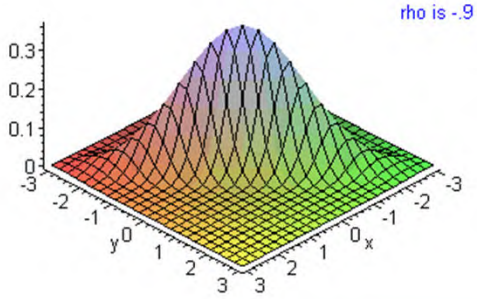
作答

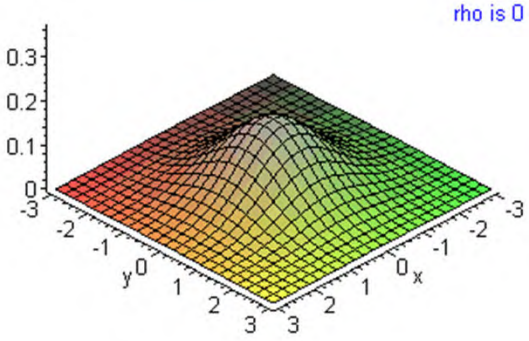
Quick Review of Bivariate Normal

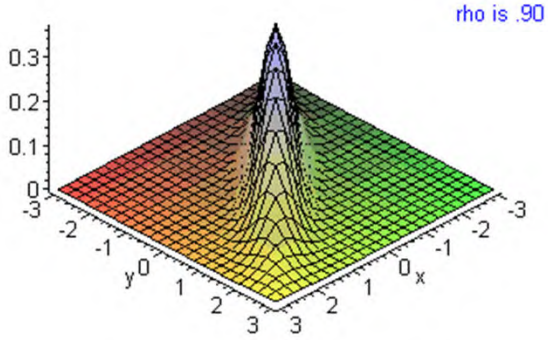






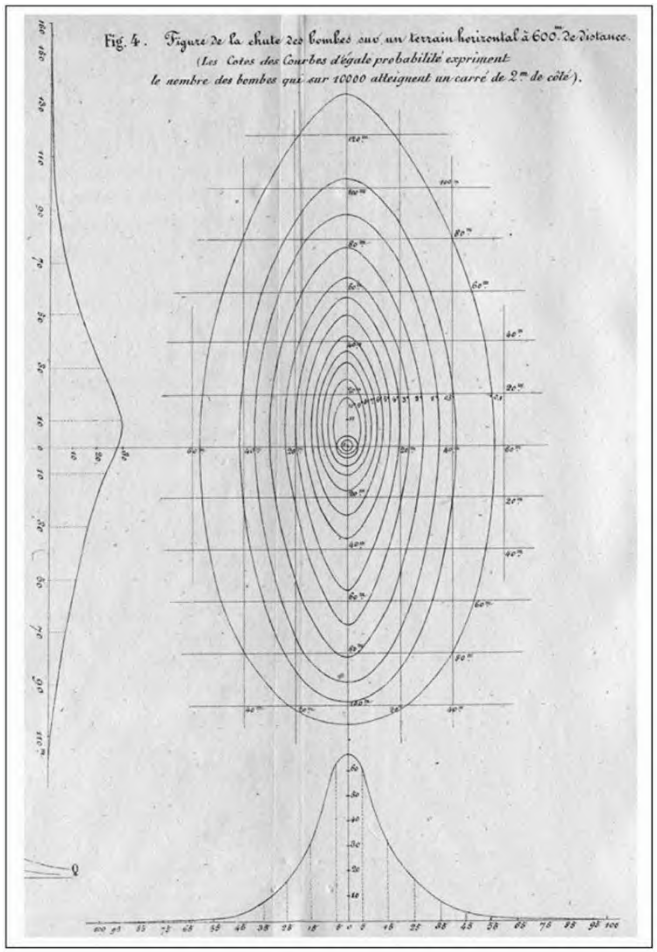






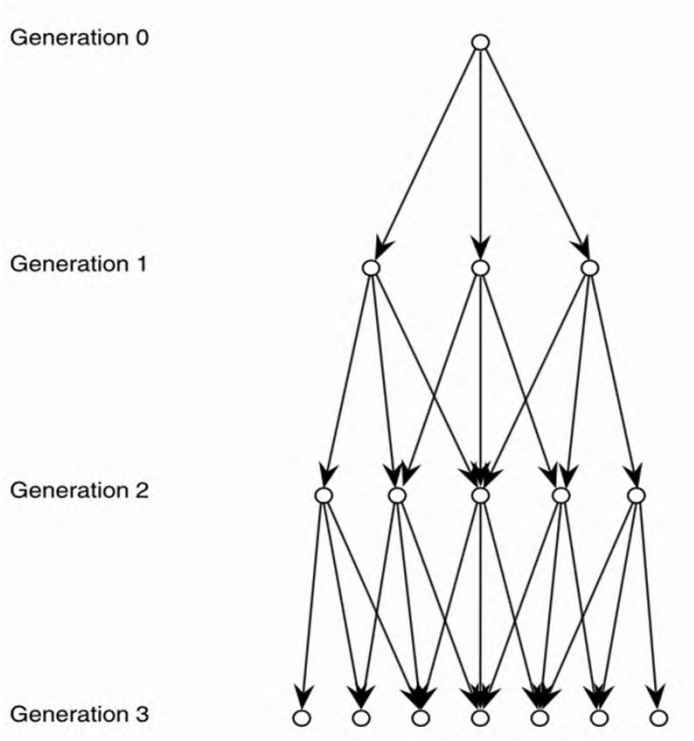
Far-reaching Impact

- A measure of association: the correlation coefficient was born
- *Francis Edgeworth*, *G. Udny Yule*, and *Karl Pearson* took the idea to higher dimensions
  - partial correlations
  - multidimensional least squares
  - principal components of variance
- "*Statistics had jumped off the two-dimensional page of tables of numbers, emerging as a technology for handling problems of great complexity*"



Galton's Solution to Darwin's Problem

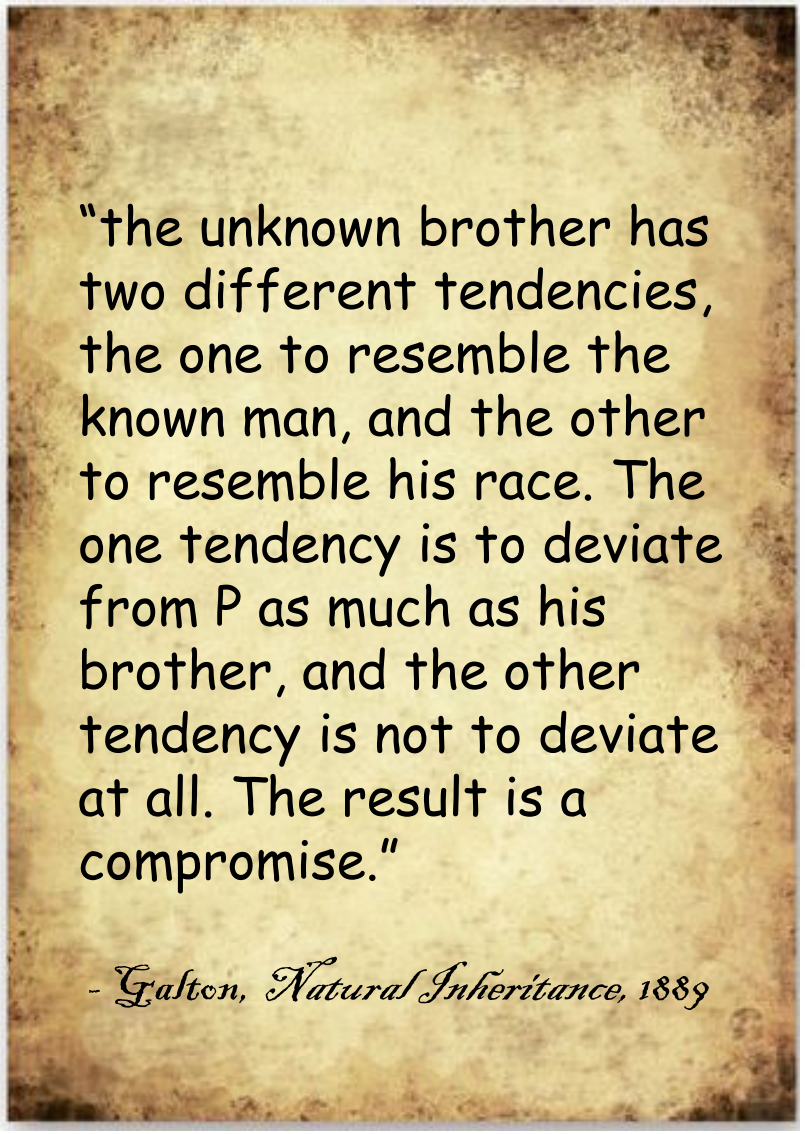
- Regression toward the mean was
  - a necessary requirement for natural selection
  - a simple consequence of the imperfect correlation between parents and offspring
- The population dispersion is in an approximate evolutionary equilibrium
  - The movement from the population center toward the extremes was balanced by the movement back





Galton's Interpretation

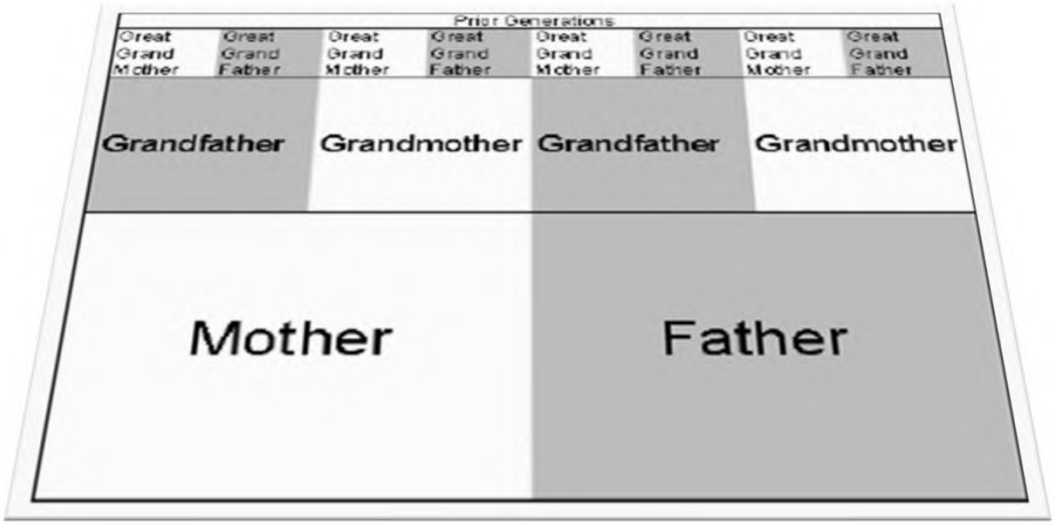
- $F = G + V_1, S = G + V_2,$
  - G: an unobserved persistent (genetic) component that blood relatives hold in common with each other
  - V_i: unobserved transitory (random) components, uncorrelated with G and with each other
  - $E(G) = p$
  - $E(V_1) = E(V_2) = 0$





One Step Further: Multiple Regression

- Galton noticed that
  - prior generations could also influence individual characteristics
  - effects are diminishing
- He had hit upon the kernel of the idea of multiple regression:
  - A variable may be influenced by a multitude of causes of greater and lesser importance
  - Some of these causes may even overlap with one another



Now get back to our time:
Are you ready for the journey?



Before You Started

- Statistics
  - starts with a problem
  - proceeds with the collection of data
  - continues with the data analysis
  - finishes with conclusions
- A common mistake is to plunge into a complex analysis without paying attention to what the objectives are or even whether the data are appropriate for the proposed analysis



Formulate the Problem Correctly

- Understand the physical background
- Understand the objective
- Make sure you know what the client wants
- Put the problem into statistical terms

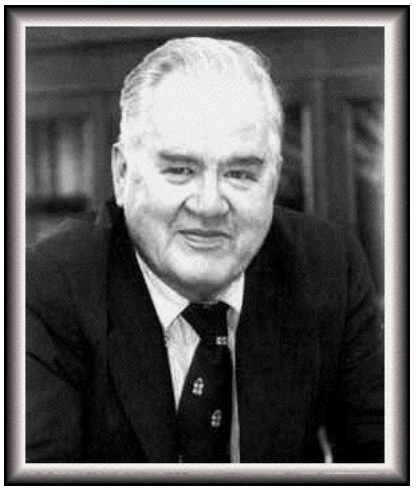


How the Data Were Collected?

- Were they obtained via a designed sample survey. **How** the data were collected has a crucial impact on what conclusions can be made
- Is there nonresponse? The data you do not see may be just as important as the data you do see
- Are there missing values? This is a common problem that is troublesome and time consuming to handle
- How are the data coded? In particular, how are the qualitative variables represented?
- What are the units of measurement?
- Beware of data entry errors and other corruption of the data. Perform some data sanity checks

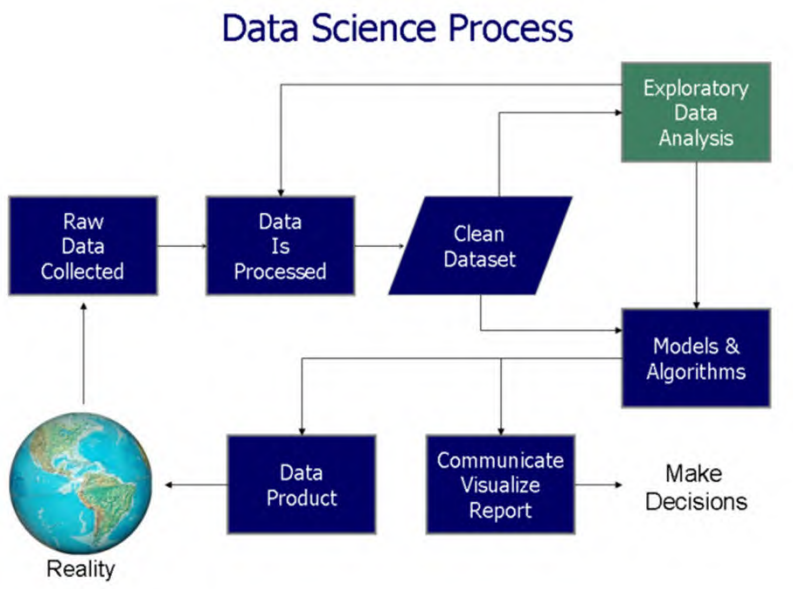


Exploratory Data Analysis (EDA)



- Promoted by John Tukey to encourage statisticians to explore the data, and possibly formulate hypotheses that could lead to new data collection and experiments

- **EDA** is an approach
  - to analyzing data sets
  - to summarize their main characteristics
  - often with visual methods



EDA in Practice

- Numerical summaries such as means, standard deviations, maximum and minimum, correlations and whatever else is appropriate to the specific dataset
- Graphical summaries
  - For one variable, boxplots, histograms, density plots
  - For two/more variables, scatterplots are standard
  - interactive and dynamic graphics
- Look for outliers, data-entry errors, skewed or unusual distributions and structures
- Simple but vital!



An Example: Female Pima Indians

- The National Institute of Diabetes and Digestive and Kidney Diseases conducted a study on 768 adult female Pima Indians living near Phoenix
- The following variables were recorded:
  - number of times pregnant
  - plasma glucose concentration at 2 hours in an oral glucose tolerance test
  - diastolic blood pressure(mmHg)
  - triceps skin fold thickness (mm)
  - 2-hour serum insulin (mu U/ml)
  - body mass index (weight in kg/(height in m^2))
  - diabetes pedigree function
  - age (years)
  - a test whether the patient showed signs of diabetes



- The purpose of the study
- How the data were collected
- Look at the data

> library(faraway)

> pima

	pregnant	glucose	diastolic	triceps	insulin	bmi	diabetes	age
1	6	148	72	35	0	33.6	0.627	50
2	1	85	66	29	0	26.6	0.351	31
3	8	183	64	0	0	23.3	0.672	32
…much deleted…								
768	1	93	70	31	0	30.4	0.315	23



```
> summary(pima)
```

pregnant	glucose	diastolic	triceps
Min. : 0.000	Min. : 0.0	Min. : 0.00	Min. : 0.00
1st Qu.: 1.000	1st Qu.: 99.0	1st Qu.: 62.00	1st Qu.: 0.00
Median : 3.000	Median :117.0	Median : 72.00	Median :23.00
Mean : 3.845	Mean :120.9	Mean : 69.11	Mean :20.54
3rd Qu.: 6.000	3rd Qu.:140.2	3rd Qu.: 80.00	3rd Qu.:32.00
Max. :17.000	Max. :199.0	Max. :122.00	Max. :99.00

insulin	bmi	diabetes	age	test
Min. : 0.0	Min. : 0.00	Min. :0.0780	Min. :21.00	Min. :0.000
1st Qu.: 0.0	1st Qu.:27.30	1st Qu.:0.2437	1st Qu.:24.00	1st Qu.:0.000
Median : 30.5	Median :32.00	Median :0.3725	Median :29.00	Median :0.000
Mean : 79.8	Mean :31.99	Mean :0.4719	Mean :33.24	Mean :0.349
3rd Qu.:127.2	3rd Qu.:36.60	3rd Qu.:0.6262	3rd Qu.:41.00	3rd Qu.:1.000
Max. :846.0	Max. :67.10	Max. :2.4200	Max. :81.00	Max. :1.000



Sanity Checks

```
> sort(pima$diastolic)
 [1]  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
[18]  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
[35]  0 24 30 30 38 40 44 44 44 44 46 46 48 48 48 48 48
...etc...
> pima$diastolic [pima$diastolic == 0] <- NA
> pima$glucose [pima$glucose == 0] <- NA
> pima$triceps [pima$triceps == 0]  <- NA
> pima$insulin [pima$insulin == 0] <- NA
> pima$bmi [pima$bmi == 0] <- NA
```



Factors- Categorical Variables

```
> pima$test <- factor (pima$test)
> summary (pima$test)
   0    1
 500  268
> levels(pima$test) <- c('negative', 'positive')
> summary (pima$test)
negative  positive
     500       268
```



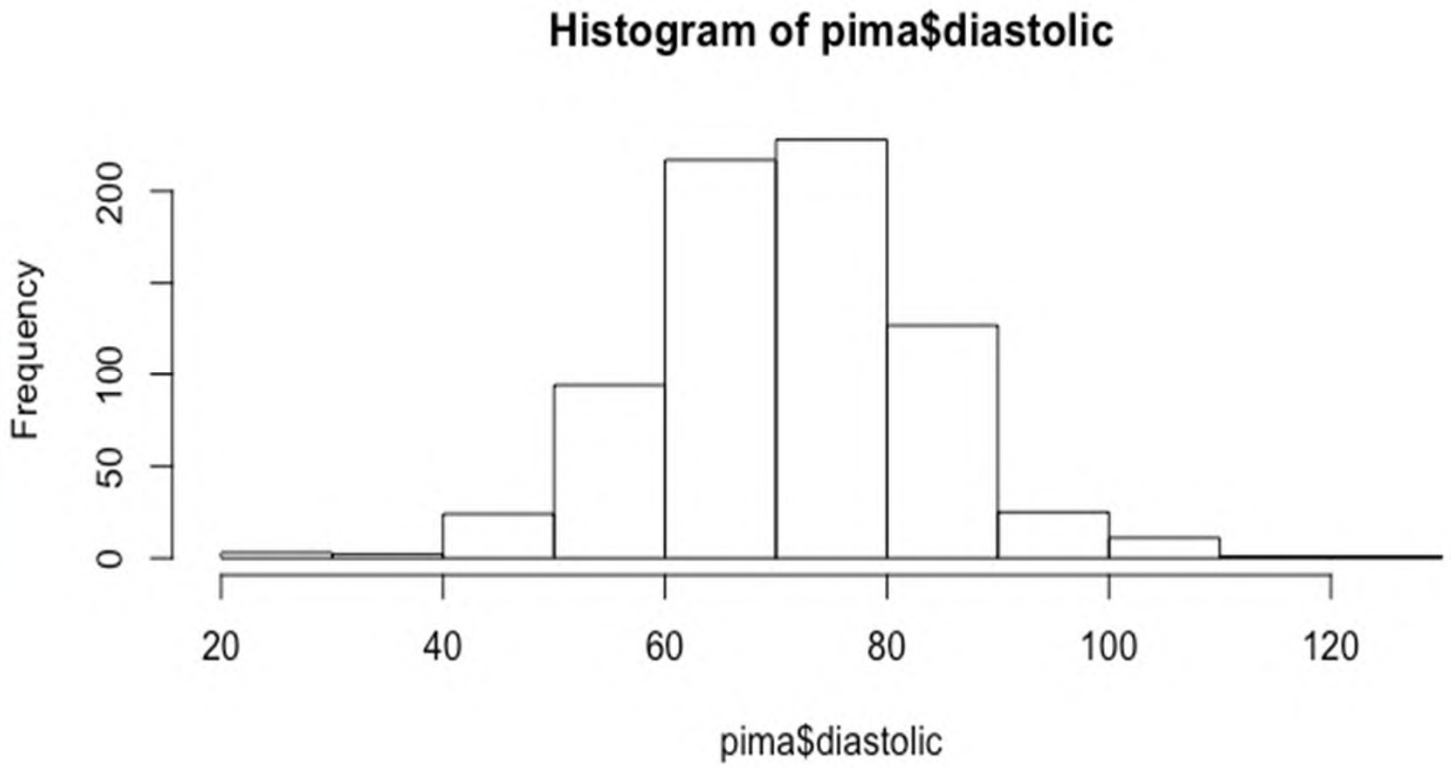
> summary (pima)

```
    pregnant         glucose       diastolic        triceps         insulin     
 Min.   : 0.000   Min.   : 44.0   Min.   : 24.00   Min.   : 7.00   Min.   : 14.00  
 1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 64.00   1st Qu.:22.00   1st Qu.: 76.25  
 Median : 3.000   Median :117.0   Median : 72.00   Median :29.00   Median :125.00  
 Mean   : 3.845   Mean   :121.7   Mean   : 72.41   Mean   :29.15   Mean   :155.55  
 3rd Qu.: 6.000   3rd Qu.:141.0   3rd Qu.: 80.00   3rd Qu.:36.00   3rd Qu.:190.00  
 Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00   Max.   :846.00  
                  NA's   :5       NA's   :35       NA's   :227     NA's   :374     
      bmi           diabetes          age            test       
 Min.   :18.20   Min.   :0.0780   Min.   :21.00   negative:500  
 1st Qu.:27.50   1st Qu.:0.2437   1st Qu.:24.00   positive:268  
 Median :32.30   Median :0.3725   Median :29.00                 
 Mean   :32.46   Mean   :0.4719   Mean   :33.24                 
 3rd Qu.:36.60   3rd Qu.:0.6262   3rd Qu.:41.00                 
 Max.   :67.10   Max.   :2.4200   Max.   :81.00                 
```



Histogram

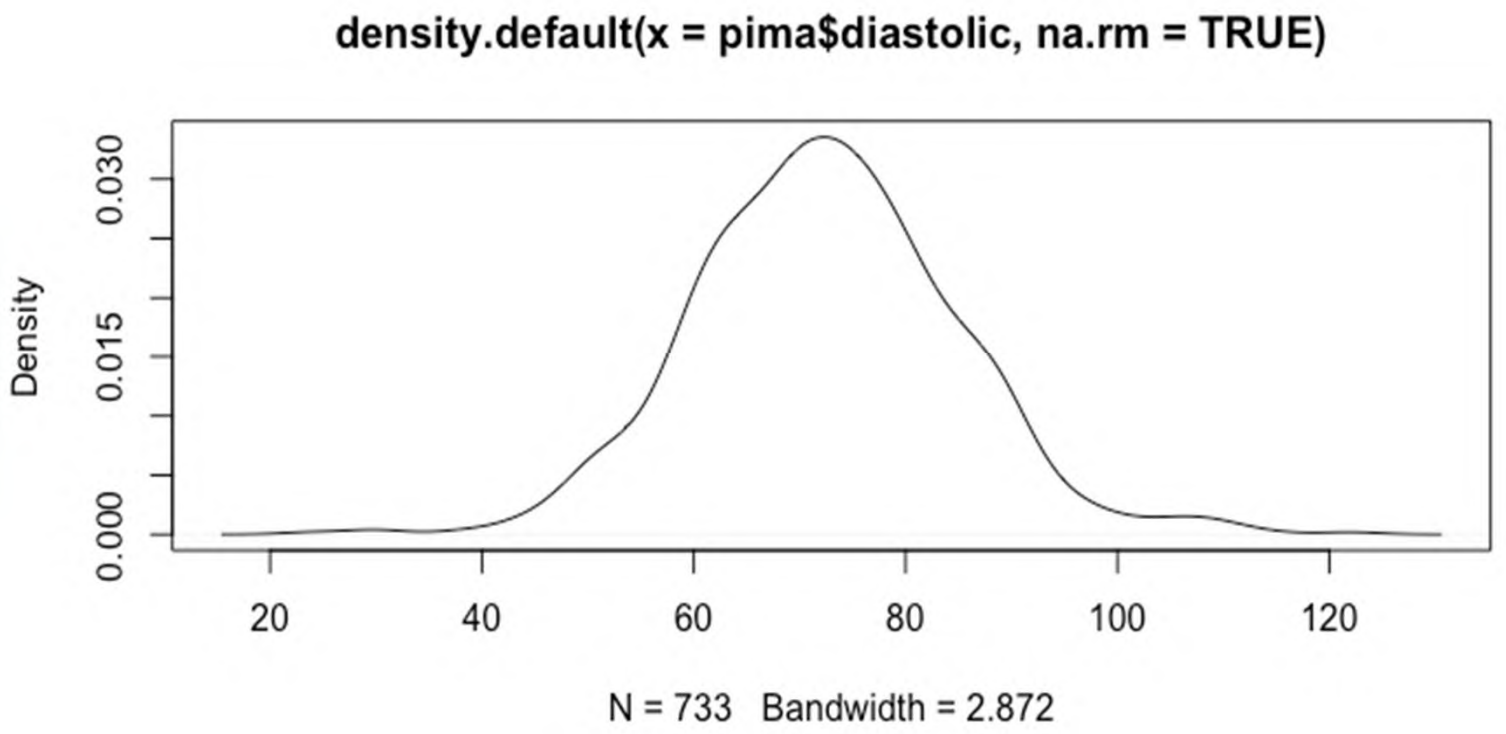
> hist (pima$diastolic)





Kernel Density Estimate

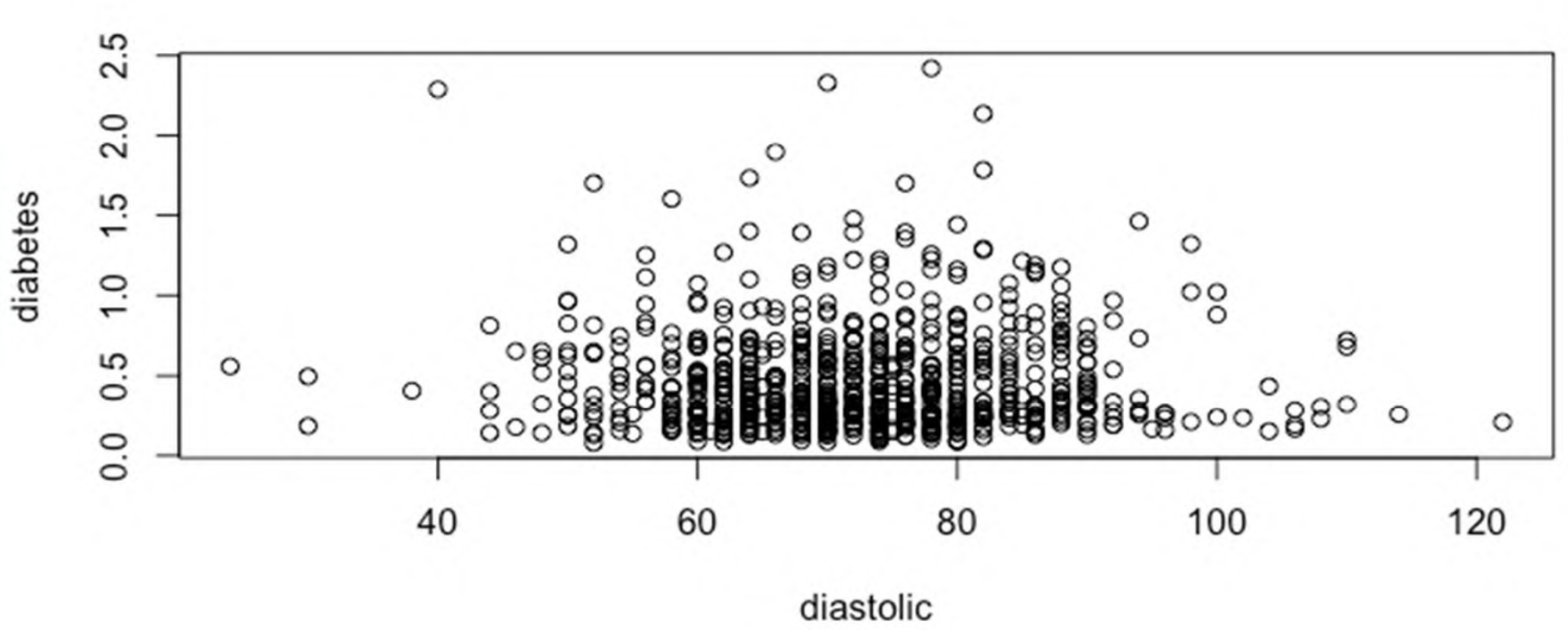
> plot (density (pima$diastolic, na.rm=TRUE))





Scatterplot

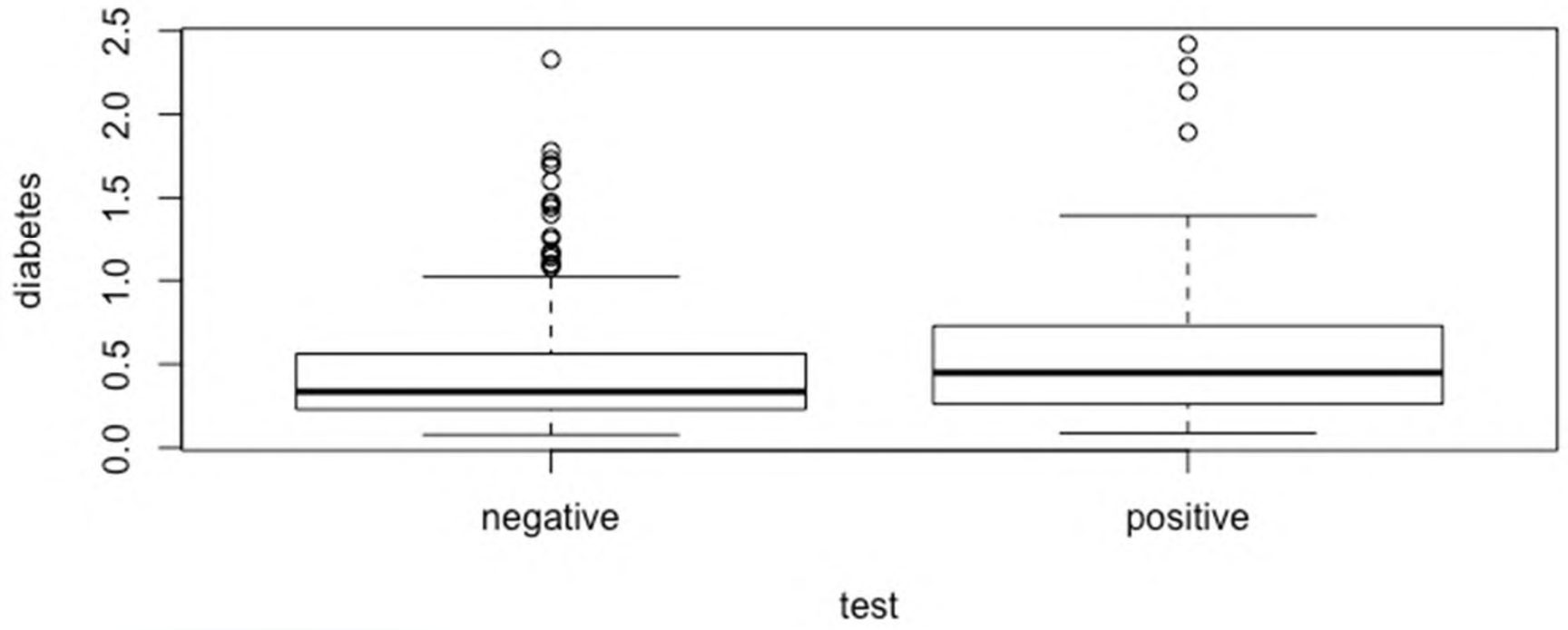
> plot (diabetes ~ diastolic, pima)





Side-by-side Boxplot

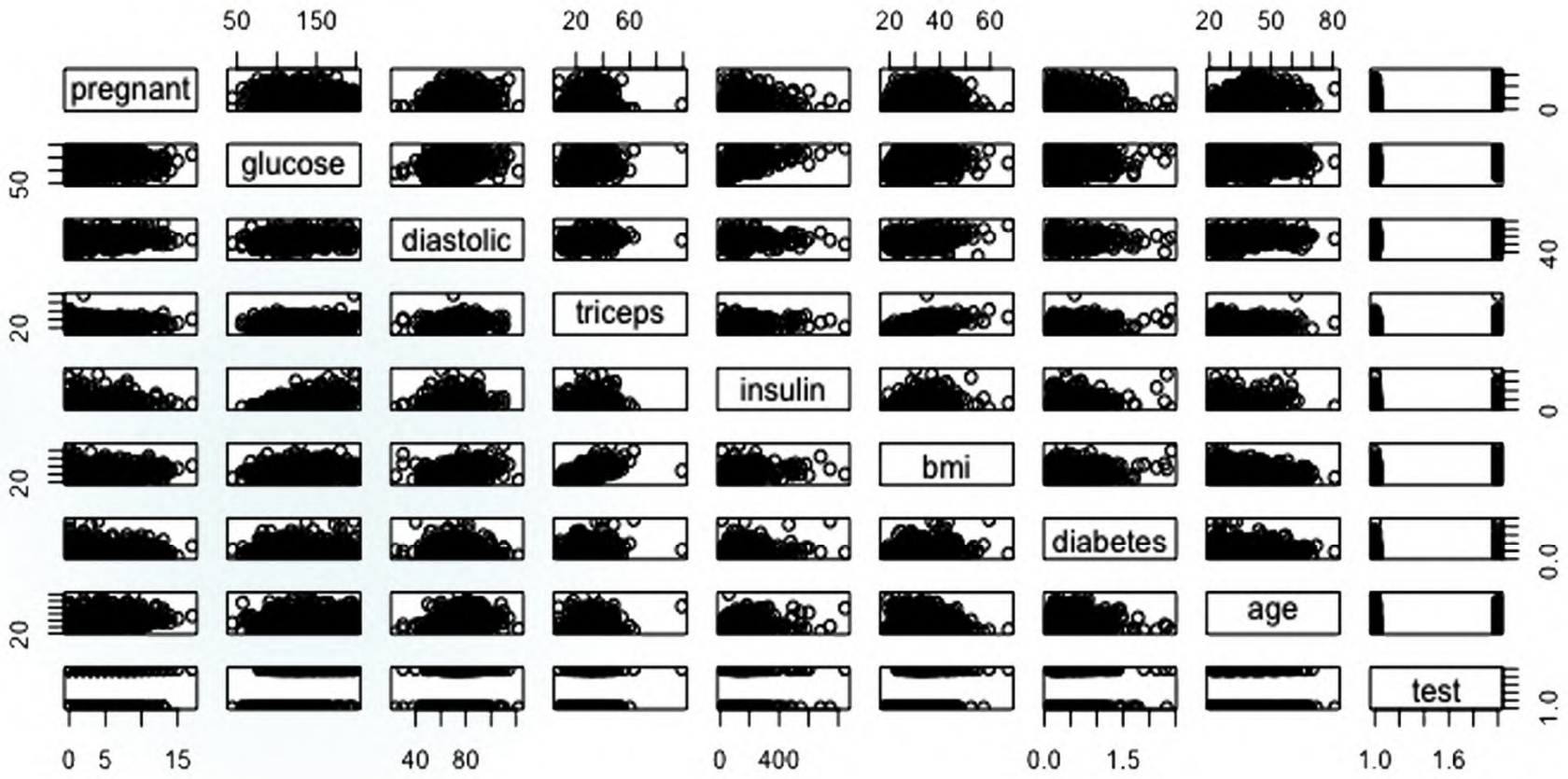
> plot (diabetes ~ test, pima)





Scatterplot Matrix

> pairs (pima)



Analysis: Terminologies

- X: predictor/ input/ independent or explanatory variable
- Y: response/ output/ dependent variable
- $Y \sim X$: simple regression
- $Y \sim X_1, \ldots, X_p$: multiple/multivariate regression
  - A multiple regression has more than one X in one formula
  - A multivariate regression has more than one Y, but in different formulae
- A **multivariate multiple regression** has **multiple** X's to predict **multiple** Y's with each Y in a different formula, usually based on the **same** data



Various Forms of Regressions

- *Y* continuous; *X* continuous, discrete or categorical
  - diabetes ~ diastolic + bmi : a multiple regression, only quantitative variables, will be tackled shortly
  - bmi ~ diastolic + test: both quantitative and qualitative predictors, will be considered on analysis of covariance
  - diastolic ~ test : only qualitative predictors, called analysis of variance (ANOVA)
- test ~ diastolic + bmi : *Y* qualitative (categorical, here binary), needs a logistic regression
- pregnant ~ diastolic + bmi : *Y* discrete (count), needs a Poisson regression



Possible Objectives

- Prediction of future observations
- Assessment of the effect of, or relationship between, explanatory variables and the response
- A general description of data structure



Steps in a Regression Analysis

- A regression analysis includes the following steps:
  - Statement of the problem
  - Selection of potentially relevant variables
  - Data collection
  - Model specification
  - Choice of fitting method
  - Model fitting
  - Model validation and criticism
  - Using the chosen model(s) for the solution of the posed problem



Now we're ready to try some regression therapy…

See you next week!



第1讲

- 1.1 回归概念的历史
- 1.2 预备知识
  - 1.2.1 探索性数据分析（EDA）
  - 1.2.2 术语
  - 1.2.3 研究目标
  - 1.2.4 一般步骤



请回答：

- 为什么叫“回归”？
- 学习这门课程需要知道哪些基本“术语”？



Objectives （仅复习自测用，不上交）

- Know how Galton discovered ‘regression’.
- Understand what is the regression phenomenon and especially, the relationship of slope and correlation coefficient.
- Understand the mathematical formulation used to describe the actions of the quincunx.
- Understand the connection between regression and the bivariate normal distribution.
- Know the concept of EDA.
- Know the terminologies used in regression.
- Distinguish various forms of regression.
- Understand the modelling procedure for a regression analysis.



Homework

- Prob-1. Use the brothers data collected by Galton. Plot the points and explain the regression effect there.

Relative number of Brothers of various Heights to Men of various Heights, Families of Five Brothers and upwards being excluded.

Heights of the men in inches.	Heights of their brothers in inches.													Total cases.	Medians.
	Below 63	63·5	64·5	65·5	66·5	67·5	68·5	69·5	70·5	71·5	72·5	73·5	Above 74		
74 and above	1	1	..	..	..	..	..	1	1	..	5	3	12	24	
73·5.....	..	..	..	..	..	1	3	4	8	3	3	2	3	27	
72·5.....	..	..	..	..	1	1	6	5	9	9	8	3	5	47	71·1
71·5.....	..	1	..	1	2	8	11	18	14	20	9	4	..	88	70·2
70·5.....	..	..	1	1	7	19	30	45	36	14	9	8	1	171	69·6
69·5.....	..	1	2	1	11	20	36	55	44	17	5	4	2	198	69·5
68·5.....	..	1	5	9	18	38	46	36	30	11	6	3	..	203	68·7
67·5.....	2	4	8	26	35	38	38	20	18	8	1	1	..	199	67·7
66·5.....	4	3	10	33	28	35	20	12	7	2	1	..	..	155	67·0
65·5.....	3	3	15	18	33	36	8	2	1	1	..	..	..	110	66·5
64·5.....	3	8	12	15	10	8	5	2	1	..	..	..	..	64	65·6
63·5.....	5	2	8	3	3	4	1	1	..	1	..	..	1	20	
Below 63	5	5	3	3	4	2	..	..	..	..	..	..	1	23	
Totals	23	29	64	110	152	200	204	201	169	86	47	28	25	1329	



填空题 10分

设置

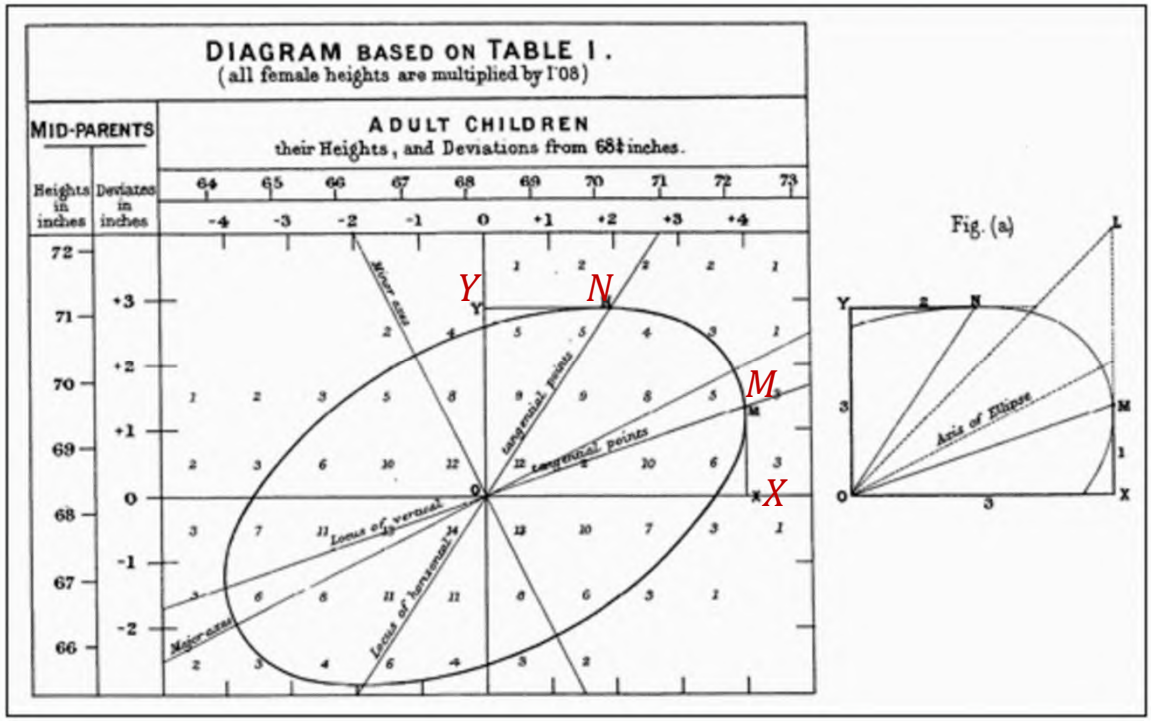
53

Prob-2 (A). 请用 X，Y 等分别写出两条回归线的数学表达式：

OM, (the expected height of the mid-parent as a function of the child's height): [填空1]

ON, (the expected height of the offspring as a function of the mid-parent's height): [填空2]

（提示：请注意字母的大小写。）



作答

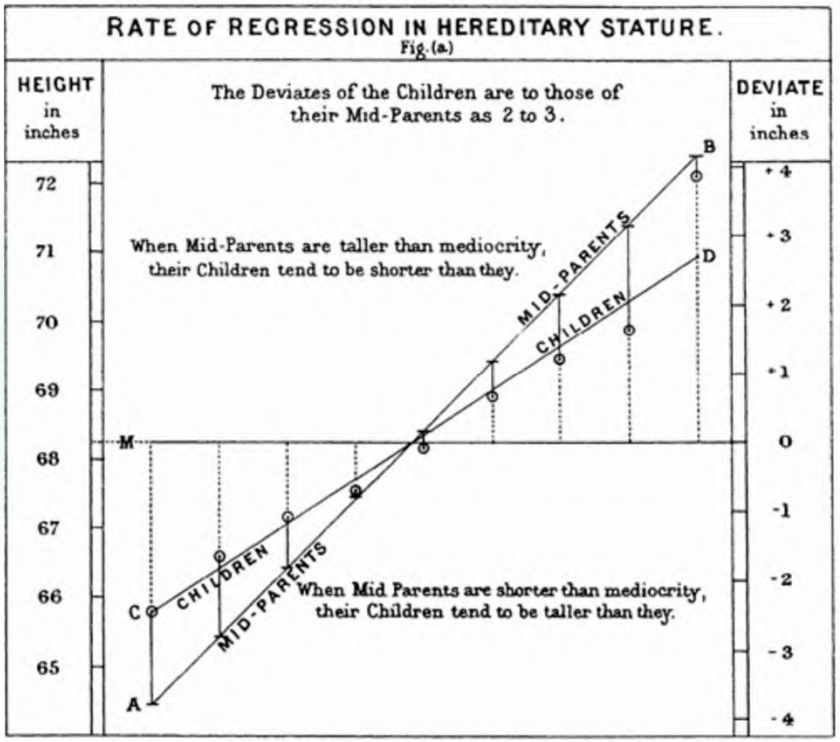
填空题 15分

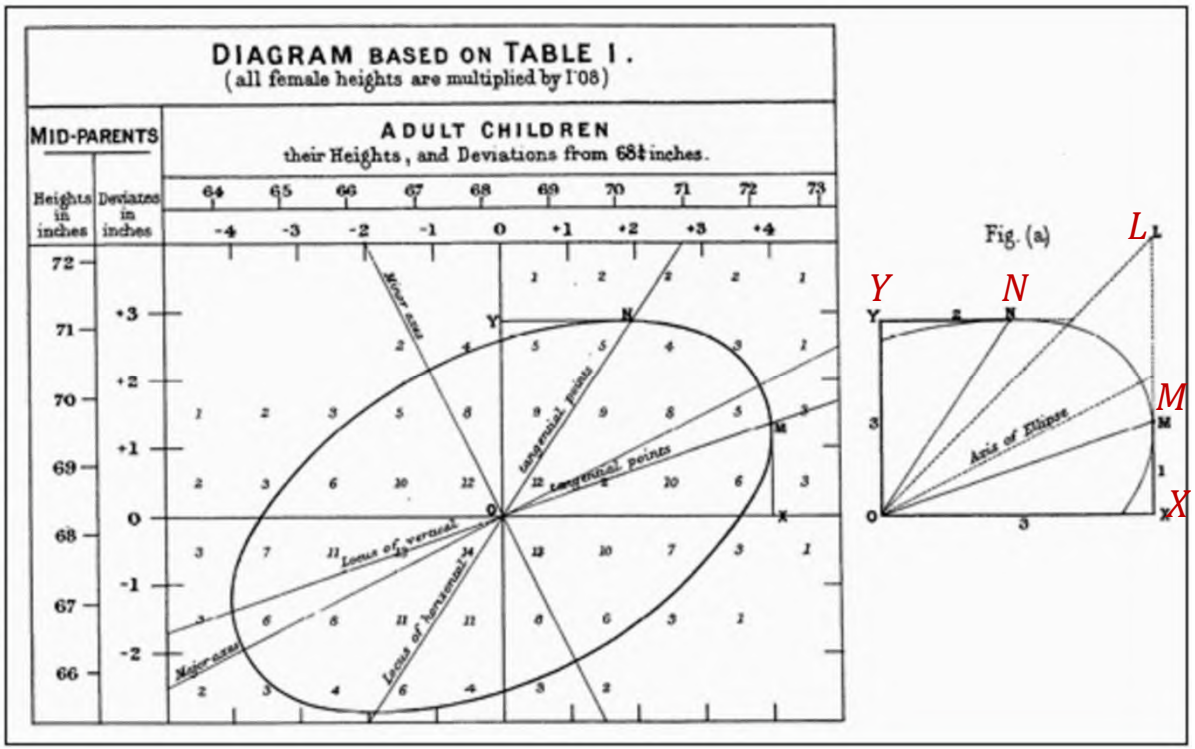
设置

54

Prob-2 (B). 比较：左图中AB线对应右图的 [填空1]，

CD线对应右图的 [填空2].





作答

填空题 2分

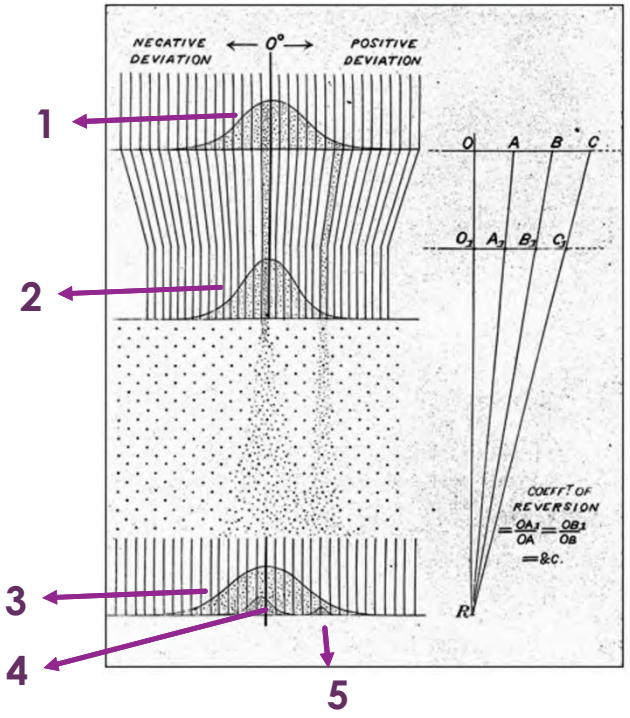
设置

55

Prob-3 (A). 此处我们看到若干个正态分布的形状(记为1~5)，其中：

mid-parent population，对应 [填空1]；

each subgroup’s offspring as a normal distribution with the same dispersion，指的是 [填空2]。



作答

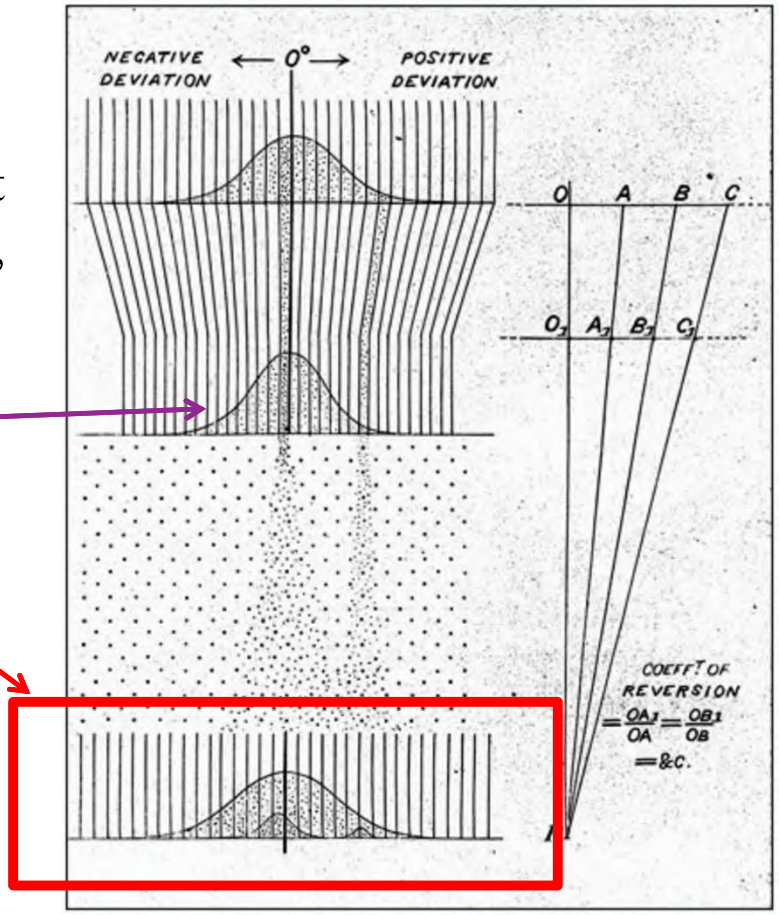
Prob-3 (B).

Galton demonstrated that there was variability among each of the families, yet the families combined to produce a stable, normally distributed population.

经过倾斜槽后仍为正态分布吗?

子代整体的分布为什么是正态分布?

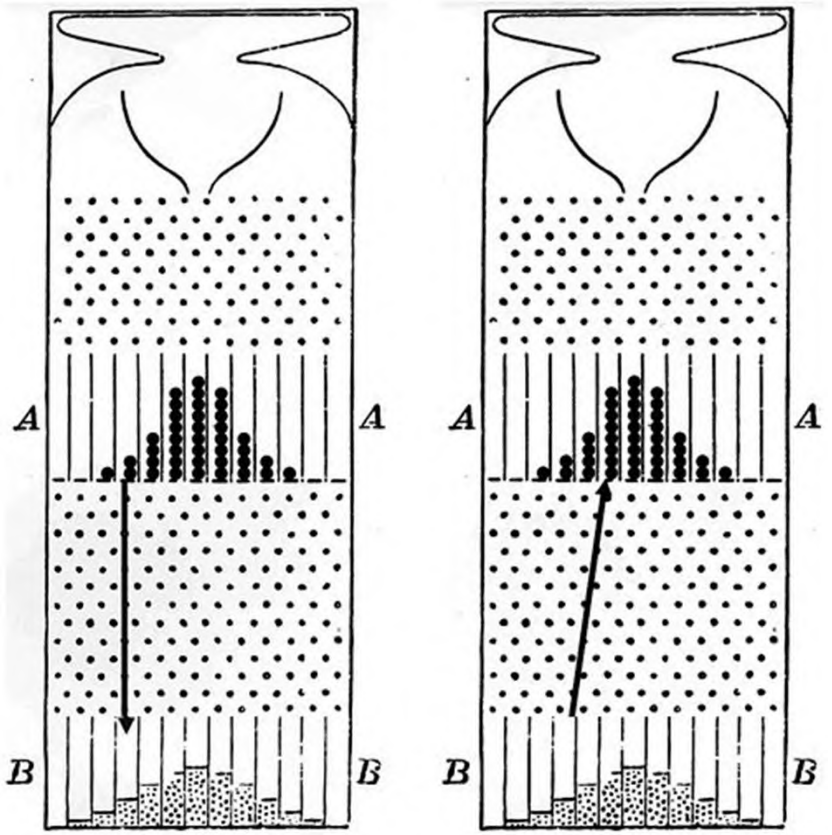
注意底层红框中的三个正态分布，是否是小（局部）正态分布相加组成了大（整体）正态分布?



作答

Homework Cont'd

- Prob-4. The left panel shows the average final position of shot released from an upper compartment; the right panel shows the average initial position of shot that arrived in a lower compartment. *(Galton 1889).* That is, if you were to release the shot in a single midlevel compartment, on average they would fall directly below. But if we look at a lower compartment on the left, and ask where the residents of that lower compartment were likely to have descended from, the answer is closer to the middle. Explain this paradox both mathematically and intuitively.



Homework Cont'd

- Prob-5. Suppose that the heights, in inches, of the women in a certain population follow the normal distribution with mean 65 inches and standard deviation 1 inch. The heights of men in this population follow the normal distribution with mean 68 inches and standard deviation 2 inches. One woman is selected at random and, independently, one man is selected at random. Find the probability that the woman will be taller than the man.



Background Reading

- Read Appendix A.
  - A.3 : random variables
  - A.4 : probability distributions
- Chapter 1
  - 1.3 : simple linear regression
  - 1.6 : estimation of regression function
  - 1.7 : estimation of error variance
  - 1.8 : normal error regression model



Further Reading

- Galton did not play an important role in the acceptance of Darwin's theory, but the methods he developed were extremely important in early twentieth-century biology
- He introduced the correlation coefficient and simple variance component models, in effect finding by statistical methods alone some of the results that would flow from the rediscovery of Mendel's work in 1900
- In 1918, Ronald A. Fisher extended the variance calculations to correlations and partial correlations of relatives under Mendelian assortment to all manner of relationships, and much of modern quantitative genetics was born
- The variance components idea became key to much quantitative and educational psychology. Galton's idea of separating permanent and transient effects was at the heart of the model that economist Milton Friedman proposed in 1957 in his *Theory of the Consumption Function*, for which he won the 1976 Nobel Prize



下节预告：

- 简单线性模型 SLR 是如何建立起来的?
  - 是怎么书写的?
  - 前提假设有哪些?
  - 参数的意义是什么?
  - 怎么进行参数估计?
  - 怎么在 R 上实现?
- OLS 和 MLE 是什么？怎么用在 SLR 这里的?
- …
- 我们将在下节课揭晓答案！一定要来哦~~~

