Probability and Statistics

Introduction to Statistical Inference

谢润烁 Nanjing University, 2023 Fall

Common Questions

- What is Statistics?
- Is Statistics a part of Mathematics?
- What’s the relation between Statistics and Probability?
- What’s the relation between Statistics and Machine Learning?

What is Statistics?

人工智能 统计学 应用统计学

关注者 3,055 | 被浏览 782,064

如何评价诺奖得主 Thomas J. Sargent「人工智能其实就是统计学」的观点?

原文：诺奖得主Thomas J. Sargent：人工智能其实就是统计学 只不过用了一个很华丽的辞藻

「人工智能其实就是统计学，只不过用了一个很华丽的辞藻，其实就是统计学。好多的公式都非常老，但是所有的人工智能利用的都是统计学来解决问题。」。8 月 11 日，2011 年诺贝尔经济学奖获得者 Thomas J. Sargent 在由厚益控股和《财经》杂志联合主办主题为「共享全球智慧引领未来科技」的世界科技创新论坛上如此表示。

他还提出，「有好多应用科学像工程学、物理学、经济学，我们会建立一些模型模拟世界运营......我们的目的是希望解释我们所观察到的世界上的现象，而我们关键的工具是使用模型，然后放到电脑里模拟。」

Machine Learning是一个立于统计和CS之间的领域，CMU里ML系的统计教授和CS教授大概四六开。

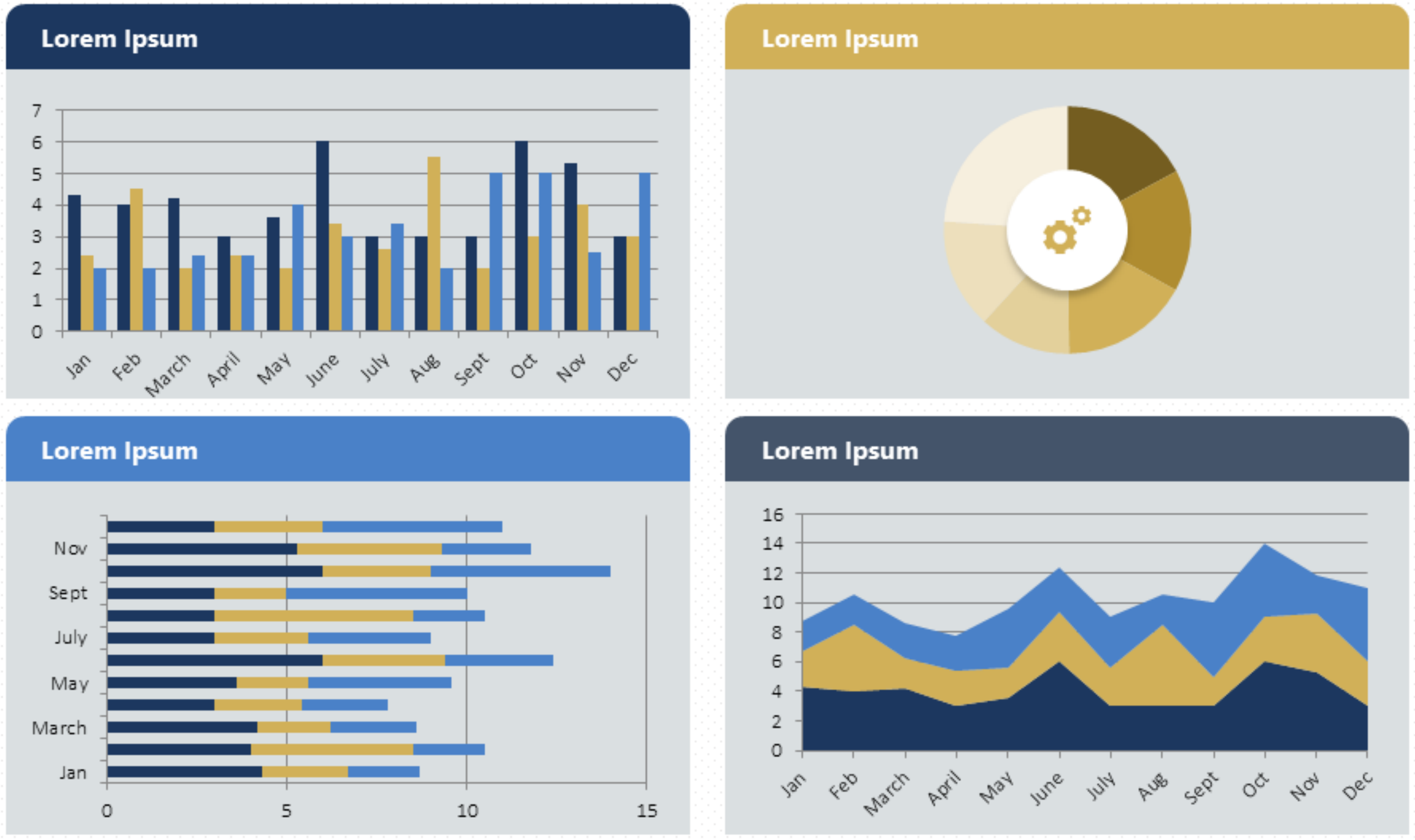
CS的ML和统计的ML有很多值得互相学习的地方，因而它们**正在变得越来越接近**。

Reference: https://www.zhihu.com/question/289782814
https://www.zhihu.com/question/48706077/answer/152022118

What is Statistics?

- A discipline that deals with data
  - Collection
  - Organization
  - Presentation

→ Descriptive



In high school

Sampling(抽样/取样)

一般地，设一个总体含有 N（N 为正整数）个个体，从中逐个抽取 n（$1 \leqslant n < N$）个个体作为样本，如果抽取是放回的，且每次抽取时总体内的各个个体被抽到的概率都相等，我们把这样的抽样方法叫做放回简单随机抽样；如果抽取是不放回的，且每次抽取时总体内未进入样本的各个个体被抽到的概率都相等，我们把这样的抽样方法叫做不放回简单随机抽样．放回简单随机抽样和不放回简单随机抽样统称为**简单随机抽样**（simple random sampling）．通过简单随机抽样获得的样本称为简单随机样本．

> 从总体中，逐个不放回地随机抽取 n 个个体作为样本，一次性批量随机抽取 n 个个体作为样本，两种方法是等价的．

与放回简单随机抽样比较，不放回简单随机抽样的效率更高，因此实践中人们更多采用不放回简单随机抽样．除非特殊声明，本章所称的简单随机抽样指不放回简单随机抽样．

上面我们按性别变量，把高一学生划分为男生、女生两个身高差异较小的子总体分别进行抽样，进而得到总体的估计．一般地，按一个或多个变量把总体划分成若干个子总体，每个个体属于且仅属于一个子总体，在每个子总体中独立地进行简单随机抽样，再把所有子总体中抽取的样本合在一起作为总样本，这样的抽样方法称为**分层随机抽样**（stratified random sampling），每一个子总体称为**层**．在分层随机抽样中，如果每层样本量都与层的大小成比例，那么称这种样本量的分配方式为**比例分配**．

In high school

Correlation coefficient(相关系数) & Scatter plot(散点图)

为了消除度量单位的影响，需要对数据作进一步的“标准化”处理. 我们用

$$s_x=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2},\quad s_y=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2}$$

分别除 $x_i-\overline{x}$ 和 $y_i-\overline{y}(i=1,2,\cdots,n)$，得

$$\left(\frac{x_1-\overline{x}}{s_x},\ \frac{y_1-\overline{y}}{s_y}\right),\ \left(\frac{x_2-\overline{x}}{s_x},\ \frac{y_2-\overline{y}}{s_y}\right),\ \cdots,\ \left(\frac{x_n-\overline{x}}{s_x},\ \frac{y_n-\overline{y}}{s_y}\right).$$

为简单起见，把上述“标准化”处理后的成对数据分别记为

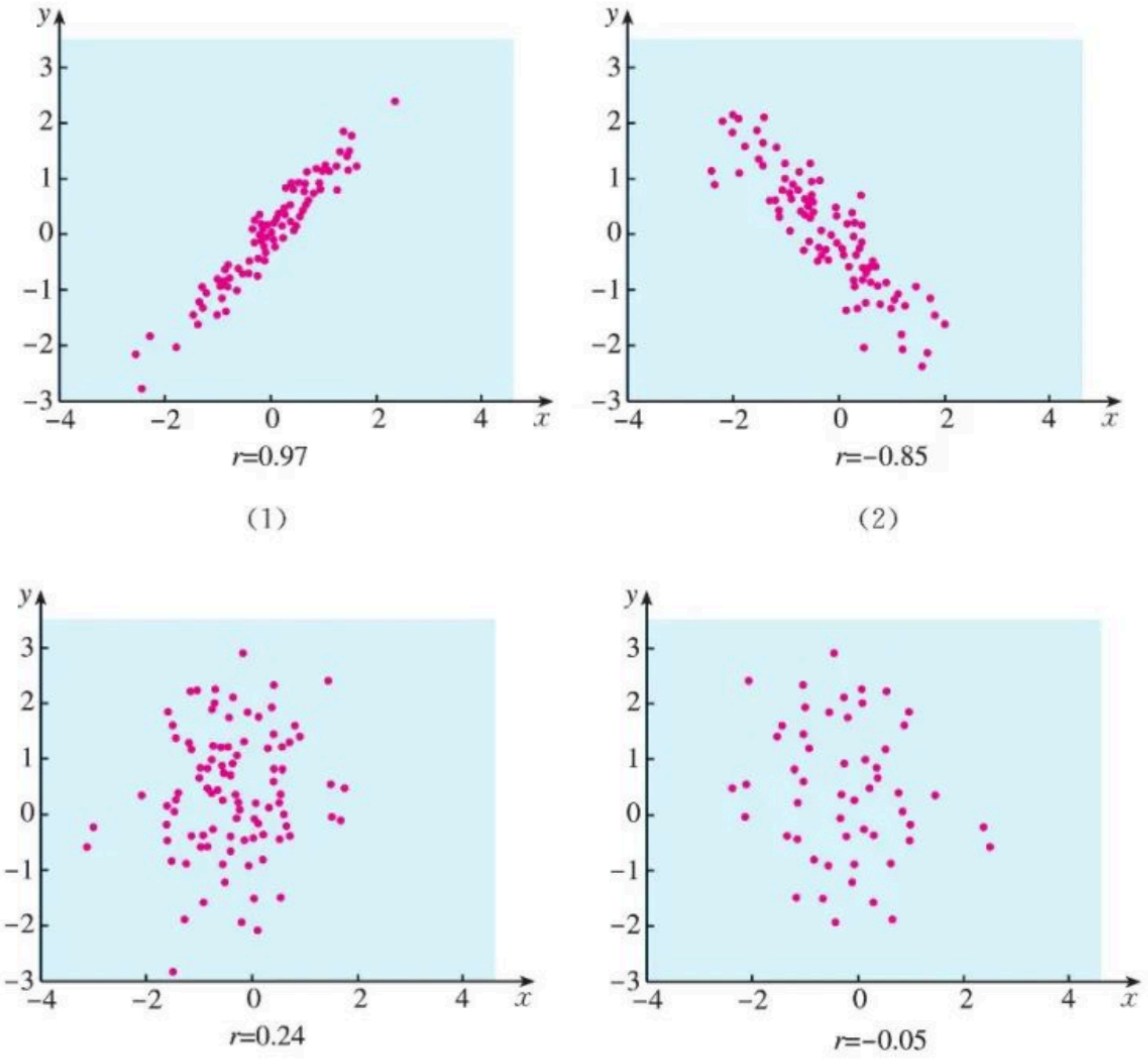
$$(x'_1,\ y'_1),\ (x'_2,\ y'_2),\ \cdots,\ (x'_n,\ y'_n),$$

仿照 L_{xy} 的构造，可以得到

$$r=\frac{1}{n}(x'_1y'_1+x'_2y'_2+\cdots+x'_ny'_n)$$

$$=\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}}.$$

我们称 r 为变量 x 和变量 y 的**样本相关系数** (sample correlation coefficient).



What is Statistics?

- A discipline that deals with data
  - Collection
  - Organization → Descriptive
  - Presentation

人工智能 统计学 应用统计学

如何评价诺奖得主 Thomas J. Sargent「人工智能其实就是统计学」的观点?

但是这句话在大众媒体上的意思则是：

人工智能就是，柱状图，饼图，折线图，散点图。

最广大民众能理解，并且愿意理解的统计学的极限，就是散点图，最多再加上一条拟合曲线。

而在极限之下，更大多数人能理解的，就是：

今年7月，合肥市区二手房价格同比增长，5.8%，环比增长，4.0%。

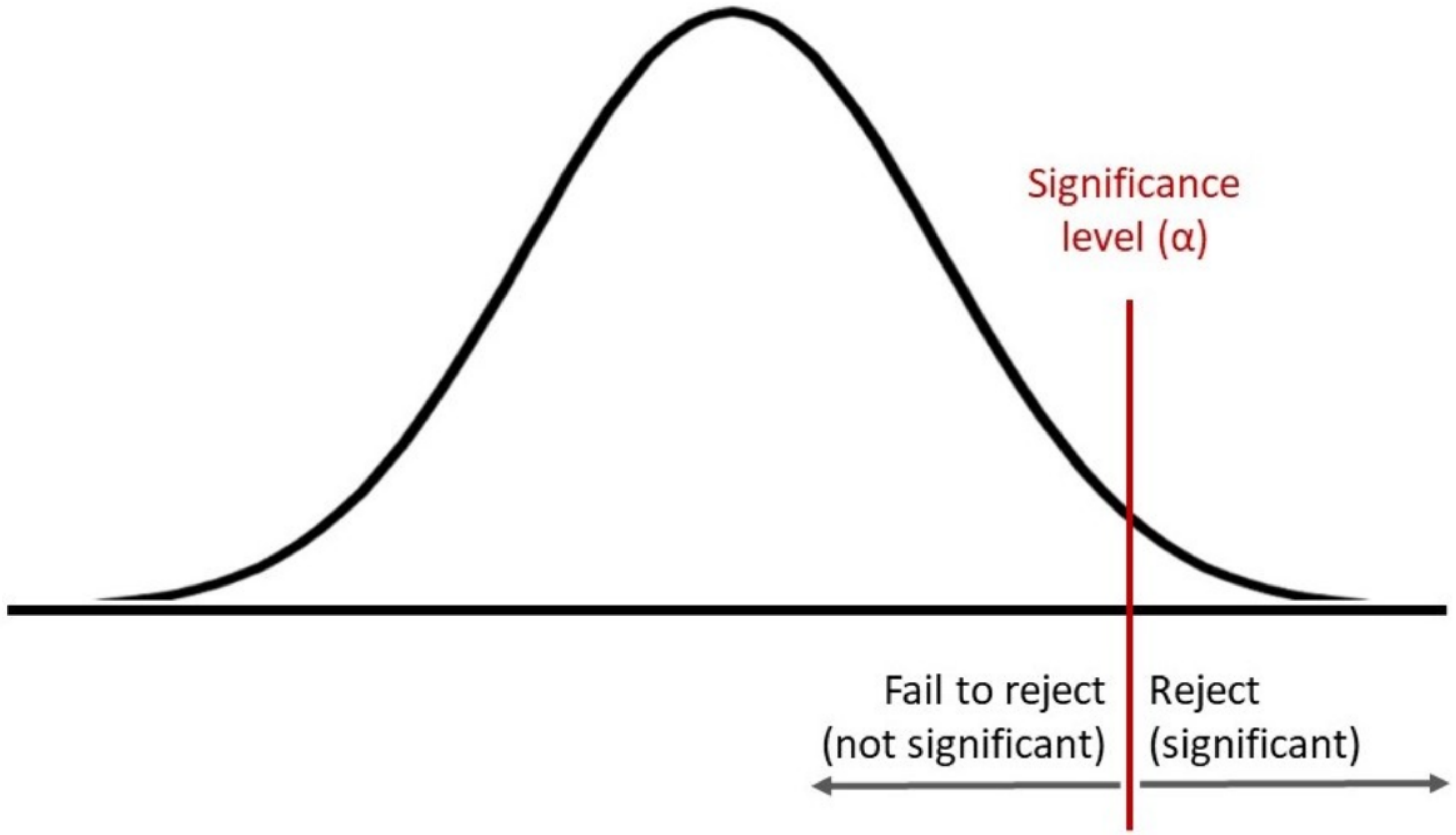
这才是媒体受众理解到的统计学。

What is Statistics?

- A discipline that deals with data
  - Collection
  - Organization
  - Presentation
  - Analysis
  - Interpretation

Descriptive

Inferential



In high school

Scatter plot(散点图)

- Can we do more about that?

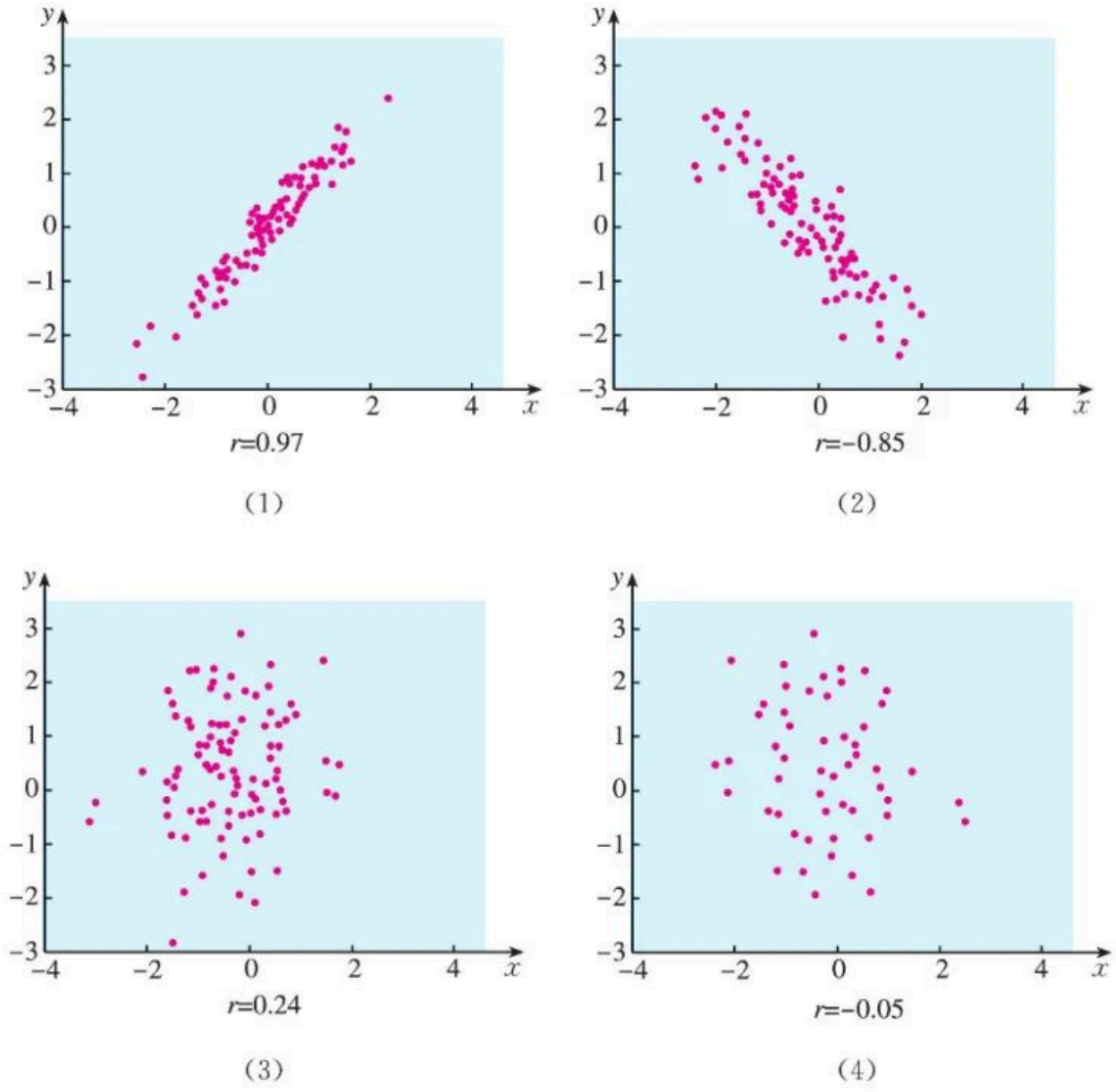


图 8. 1-5

In high school

Linear Regression

用 x 表示父亲身高，Y 表示儿子身高，e 表示随机误差. 假定随机误差 e 的均值为 0，方差为与父亲身高无关的定值 σ^2，则它们之间的关系可以表示为

$$\begin{cases} Y=bx+a+e, \\ E(e)=0,\ D(e)=\sigma^2. \end{cases} \quad (1)$$

> 为什么假设 $E(e)=0$，而不假设其为某个不为 0 的常数?

我们称（1）式为 Y 关于 x 的**一元线性回归模型**（simple linear regression model）. 其中，Y 称为**因变量**或**响应变量**，x 称为**自变量**或**解释变量**；a 和 b 为模型的未知参数，a 称为截距参数，b 称为斜率参数；e 是 Y 与 $bx+a$ 之间的随机误差. 模型中的 Y 也是随机变量，其值虽然不能由变量 x 的值确定，但是却能表示为 $bx+a$ 与 e 的和（叠加），前一部分由 x 所确定，后一部分是随机的. 如果 $e=0$，那么 Y 与 x 之间的关系就可用一元线性函数模型来描述.

In high school

Least squares method(最小二乘法)

上式是关于 b 的二次函数，因此要使 Q 取得最小值，当且仅当 b 的取值为

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

综上，当 a，b 的取值为

$$\begin{cases}\hat{b}=\dfrac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2},\\ \hat{a}=\overline{y}-\hat{b}\overline{x}\end{cases}\tag{2}$$

时，Q 达到最小.

In high school

Test of Independence(独立性检验)

表 8.3-2 单位：人

学校	数学成绩		合计
	不优秀（$Y=0$）	优秀（$Y=1$）	
甲校（$X=0$）	33	10	43
乙校（$X=1$）	38	7	45
合计	71	17	88

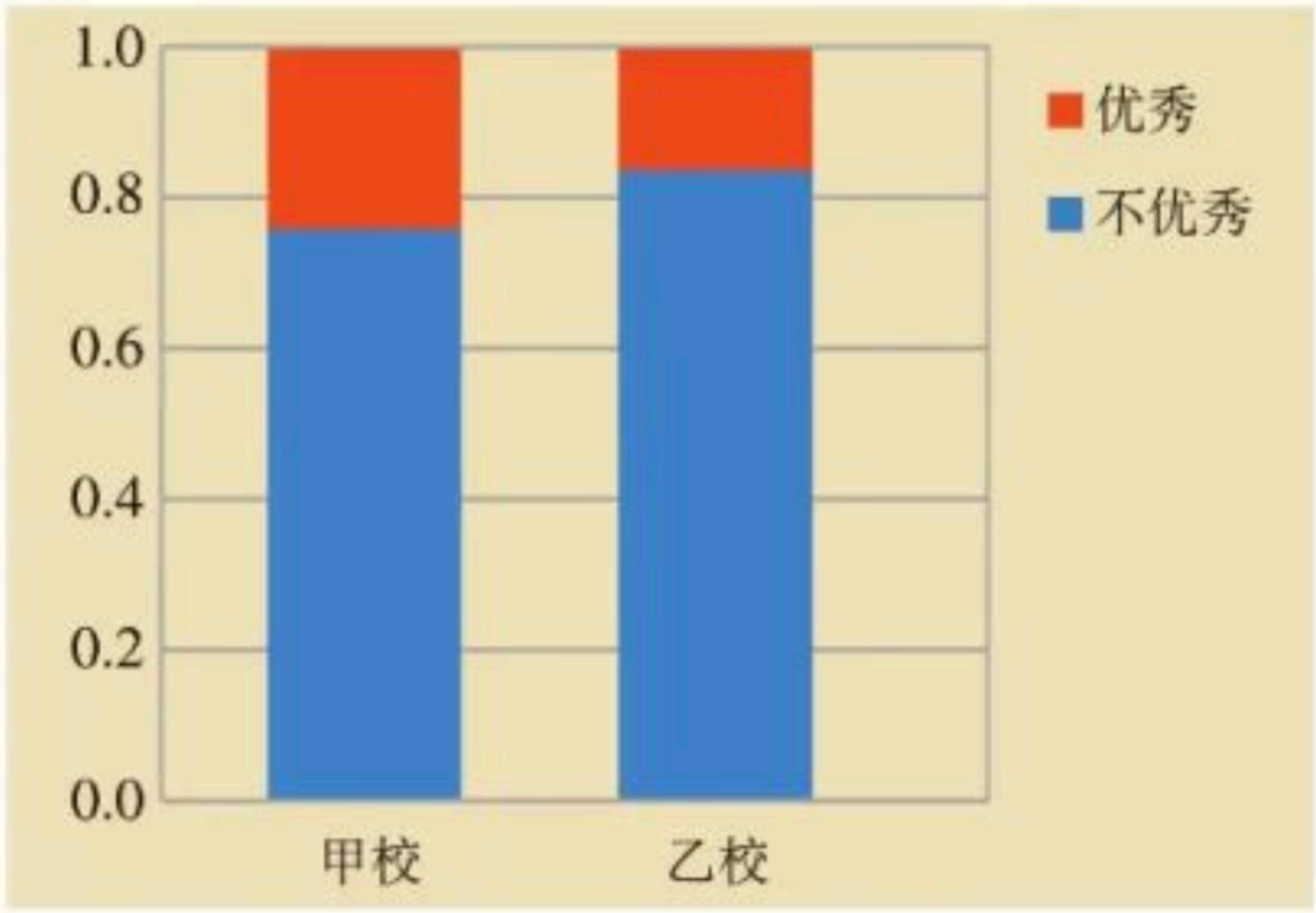


图 8.3-1

解：零假设为

H_0：分类变量 X 与 Y 相互独立，即两校学生的数学成绩优秀率无差异.

根据表 8.3-2 中的数据，计算得到

$$\chi^2=\frac{88\times(33\times7-10\times38)^2}{43\times45\times71\times17}\approx0.837<2.706=x_{0.1}.$$

根据小概率值 $\alpha=0.1$ 的 χ^2 独立性检验，没有充分证据推断 H_0 不成立，因此可以认为 H_0 成立，即认为两校的数学成绩优秀率没有差异.

Statistical Inference

Reference: https://statisticsbyjim.com/hypothesis-testing/statistical-inference/

Problems

- Classification
  - Parameter Estimation
  - Hypothesis Testing
  - Regression Analysis and Correlation Analysis
  - ……
- Our mission:
  - sample $\rightarrow$ population

Probability vs. Statistics

- Probability
  - Previous studies showed that the drug was 80% effective. Then for a study on 100 patients,
    - in average 80 will be cured
    - at least 65 will be cured with 99.99% chances
- Statistics
  - Observe that 78/100 patients were cured
  - We (will be able to) conclude that we are 95% confident that for other studies the drug will be effective on between 69.88% and 86.11% of patients

Basic Concepts

- Population (总体): the set of similar items or events which is of interest for some question or experiment
  - Similar to the sample space Ω in probability theory
- Sample (样本): a subset of population, which are viewed as random variables
  - e.g. $X_1, X_2, \dots, X_n$, i.i.d.
- Statistic (统计量): any quantity computed from values in a sample
  - e.g. $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, $S^2 = \frac{1}{n-1}\sum_{i=0}^{n} (X_i - \bar{X})^2$

Reading: https://statisticsbyjim.com/basics/population-vs-sample/

Parameter Estimation

All models are wrong😨 But some are useful😋

- We assume that the population follows a distribution F from a set of distribution $F(\Theta)$
- Parameter: Θ
- E.g.
  - $N(\mu, \sigma^2)$
  - $\text{Poisson}(\lambda)$
  - ……

Hypothesis Testing

Study shows that …

Example: Lady Tasting Tea

- The *lady tasting tea* is a randomized experiment devised by *Ronald Fisher*
  - reported in his book *The Design of Experiments* (1935)
- The lady claimed that
  - she could tell whether the tea or the milk was added first to a cup
- How to validate it?

Reference: https://en.wikipedia.org/wiki/Lady_tasting_tea

Example: The 1970 draft lottery

Table 3.3. *Draft numbers assigned by lottery.*

day	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sep.	Oct.	Nov.	Dec.
1	305	086	108	032	330	249	093	111	225	359	019	129
2	159	144	029	271	298	228	350	045	161	125	034	328
3	251	297	267	083	040	301	115	261	049	244	348	157
4	215	210	275	081	276	020	279	145	232	202	266	165
5	101	214	293	269	364	028	188	054	082	024	310	056
6	224	347	139	253	155	110	327	114	006	087	076	010
7	306	091	122	147	035	085	050	168	008	234	051	012
8	199	181	213	312	321	366	013	048	184	283	097	105
9	194	338	317	219	197	335	277	106	263	342	080	043
10	325	216	323	218	065	206	284	021	071	220	282	041
11	329	150	136	014	037	134	248	324	158	237	046	039
12	221	068	300	346	133	272	015	142	242	072	066	314
13	318	152	259	124	295	069	042	307	175	138	126	163
14	238	004	354	231	178	356	331	198	001	294	127	026
15	017	089	169	273	130	180	322	102	113	171	131	320
16	121	212	166	148	055	274	120	044	207	254	107	096
17	235	189	033	260	112	073	098	154	255	288	143	304
18	140	292	332	090	278	341	190	141	246	005	146	128
19	058	025	200	336	075	104	227	311	177	241	203	240
20	280	302	239	345	183	360	187	344	063	192	185	135
21	186	363	334	062	250	060	027	291	204	243	156	070
22	337	290	265	316	326	247	153	339	160	117	009	053
23	118	057	256	252	319	109	172	116	119	201	182	162
24	059	236	258	002	031	358	023	036	195	196	230	095
25	052	179	343	351	361	137	067	286	149	176	132	084
26	092	365	170	340	357	022	303	245	018	007	309	173
27	355	205	268	074	296	064	289	352	233	264	047	078
28	077	299	223	262	308	222	088	167	257	094	281	123
29	349	285	362	191	226	353	270	061	151	229	099	016
30	164		217	208	103	209	287	333	315	038	174	003
31	211		030		313		193	011		079		100

Table 3.4. *Average draft number per month.*

January	201.2	July	181.5
February	203.0	August	173.5
March	225.8	September	157.3
April	203.7	October	182.5
May	208.0	November	148.7
June	195.7	December	121.5

Is it a Coincidence?

Parameter Estimation

Parameter Estimation

- Assume a set of models with unknown parameters
  - e.g. normal distribution with expectation μ and variance σ^2
- Some common methods
  - Method of Moments (矩估计法)
  - MLE (最大似然估计)
  - MAP (最大后验估计)

Method of Moments

- Distribution: $p(x;\theta_1,\ldots,\theta_k)$ with mth moment $\alpha_m(\theta_1,\ldots,\theta_k)$, $m=1,2,\ldots,k$
- Sample $X_1,\ldots,X_n$ with mth moment α_m, $m=1,2,\ldots,k$
- Solve Equation

$$\alpha_m = \alpha_m(\theta_1,\ldots,\theta_k), m=1,2,\ldots,k$$

- Get

$$\hat{\theta}_m = g_m(X_1,X_2,\ldots,X_n)$$

Maximum Likelihood Estimate(MLE)

What is likelihood?

- Distribution: $p(x; \Theta)$
- Sample: $X_1, \ldots, X_n$
- Likelihood function:

$$L(X_1, \ldots, X_n; \Theta) = \prod_{i=1}^{n} p(X_i; \Theta)$$

- Maximum likelihood:

$$\max_{\Theta} L(X_1, \ldots, X_n; \Theta)$$

Maximum Likelihood Estimate(MLE)

- Maximize $p(\text{Data} \mid \Theta)$
- Sometimes it is too difficult or impossible to calculate
  - ill-posed problem (不适定问题)
  - Solutions: Approximation
    - e.g. EM Algorithm when there are latent variable (隐变量)

Bayesian vs. Frequentist

MLE vs. MAP

- MLE: $\max_{\Theta} p(\text{Data} \mid \Theta)$
- MAP: $\max_{\Theta} p(\Theta \mid \text{Data})$

Maximum A Posteriori (MAP)

- $p(\Theta \mid \text{Data}) = \dfrac{p(\text{Data} \mid \Theta)p(\Theta)}{p(\text{Data})}$
- Difference between MLE and MAP:
  - The *a priori* part $p(\Theta)$
  - A certain $p(\Theta)$ makes MLE and MAP with same result

Hypothesis Testing

Hypothesis Testing

- Null Hypothesis H_0
- If strong enough evidence again H_0 is found:
  - reject H_0
- Else: accept H_0
- A little bit like Presumption of innocence (无罪推定原则)
- How to measure or quantify “strong enough”?

Two kinds of errors

- Type I error (false positive)
  - H_0 is true but reject
- Type II error (false negative)
  - H_0 is false but accept

	Rejects H_0	Accept H_0
Null is True	Type I Error	Correct
Null is False	Correct	Type II error

Reference: https://statisticsbyjim.com/hypothesis-testing/types-errors-hypothesis-testing/

Significance level

- The significance level α is the probability of rejecting the null hypothesis when it is true
    - i.e. the probability of making type I error
- The researcher determines the significance level before conducting the experiment
    - Normally, α takes 0.1, 0.05, 0.01, …, *etc.*
- How to set α?

p-value

- The probability that you would **obtain the effect observed in your sample**, if the null hypothesis is true for the populations
  - Lower p-values indicate greater evidence against the null hypothesis
- If the p-value is less than the significance level α
  - reject the null hypothesis
  - conclude that the effect is statistically significant

Reference: https://statisticsbyjim.com/glossary/p-value/

Example: Lady Tasting Tea

- The lady claimed that
  - she could tell whether the tea or the milk was added first to a cup
- The experiment provides a subject with 8 randomly ordered cups of tea
  - 4 prepared by pouring milk and then tea ($\mathcal{MT}$)
  - 4 by pouring tea and then milk ($\mathcal{TM}$)
  - The lady is asked to pick 4 cups of $\mathcal{TM}$ from these 8 cups
  - The lady did it successfully

Reference: https://en.wikipedia.org/wiki/Lady_tasting_tea

Example: The 1970 draft lottery

Table 3.3. *Draft numbers assigned by lottery.*

day	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sep.	Oct.	Nov.	Dec.
1	305	086	108	032	330	249	093	111	225	359	019	129
2	159	144	029	271	298	228	350	045	161	125	034	328
3	251	297	267	083	040	301	115	261	049	244	348	157
4	215	210	275	081	276	020	279	145	232	202	266	165
5	101	214	293	269	364	028	188	054	082	024	310	056
6	224	347	139	253	155	110	327	114	006	087	076	010
7	306	091	122	147	035	085	050	168	008	234	051	012
8	199	181	213	312	321	366	013	048	184	283	097	105
9	194	338	317	219	197	335	277	106	263	342	080	043
10	325	216	323	218	065	206	284	021	071	220	282	041
11	329	150	136	014	037	134	248	324	158	237	046	039
12	221	068	300	346	133	272	015	142	242	072	066	314
13	318	152	259	124	295	069	042	307	175	138	126	163
14	238	004	354	231	178	356	331	198	001	294	127	026
15	017	089	169	273	130	180	322	102	113	171	131	320
16	121	212	166	148	055	274	120	044	207	254	107	096
17	235	189	033	260	112	073	098	154	255	288	143	304
18	140	292	332	090	278	341	190	141	246	005	146	128
19	058	025	200	336	075	104	227	311	177	241	203	240
20	280	302	239	345	183	360	187	344	063	192	185	135
21	186	363	334	062	250	060	027	291	204	243	156	070
22	337	290	265	316	326	247	153	339	160	117	009	053
23	118	057	256	252	319	109	172	116	119	201	182	162
24	059	236	258	002	031	358	023	036	195	196	230	095
25	052	179	343	351	361	137	067	286	149	176	132	084
26	092	365	170	340	357	022	303	245	018	007	309	173
27	355	205	268	074	296	064	289	352	233	264	047	078
28	077	299	223	262	308	222	088	167	257	094	281	123
29	349	285	362	191	226	353	270	061	151	229	099	016
30	164		217	208	103	209	287	333	315	038	174	003
31	211		030		313		193	011		079		100

Table 3.4. *Average draft number per month.*

January	201.2	July	181.5
February	203.0	August	173.5
March	225.8	September	157.3
April	203.7	October	182.5
May	208.0	November	148.7
June	195.7	December	121.5

Is it a Coincidence?

Example: The 1970 draft lottery

Table 3.5. *Index numbers for the 1970 draft lottery.*

month	1	2	3	4	5	6	7	8	9	10	11	12
index	5	4	1	3	2	6	8	9	10	7	11	12

- H_0: the lottery is fair
  - Then this permutation would have to be a “random” permutation
- For a random permutation $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_{12})$ of the numbers $1,\ldots,12$,
  - We define the distance measure $d(\sigma)$ by $d(\sigma) = \sum_{i=1}^{12} |\sigma_i - i|$
- $d(\sigma^*) = 18$, for σ^* from Table 3.5
- A Monte Carlo study shows that $\Pr(d(\sigma) \leq 18) \approx 0.009$
- Reject H_0 if our significance level $\alpha \gg 0.009$

Let’s recap those common questions

Common Questions

- What is Statistics?
- **Is Statistics a part of Mathematics?**
- **What's the relation between Statistics and Probability Theory?**
- What's the relation between Statistics and Machine Learning?

Statistics vs. Mathematics

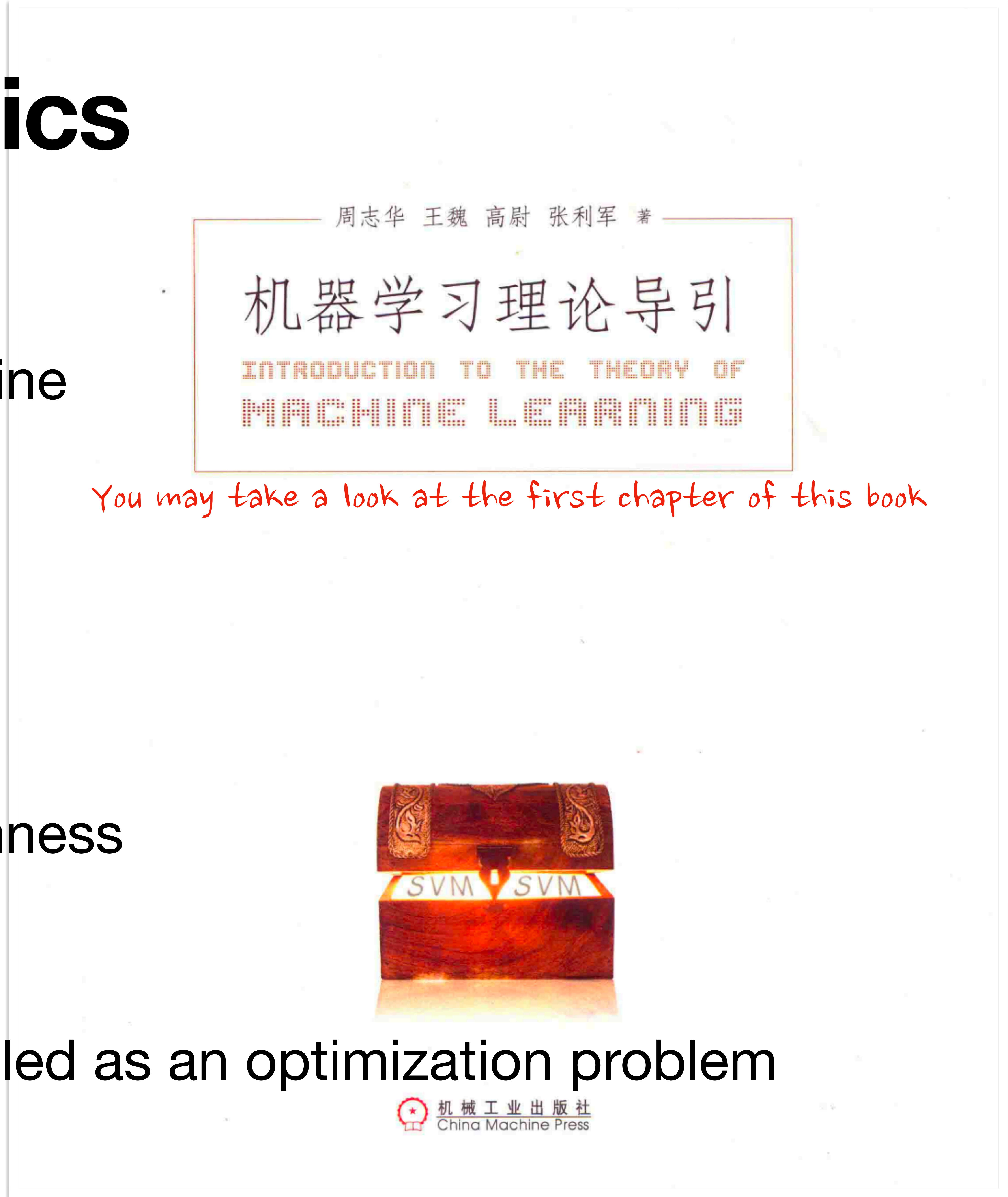
Reference: 学数学专业的人是怎么看待统计学的？ - 学弱猹的回答 - 知乎
https://www.zhihu.com/question/424487147/answer/1517834128

如果要从high-level的角度来看，数学和统计其实是完全不一样的，看教育部把统计学划为一级学科就明白了，这个“完全不一样”重点是突出在**二者的思维方式有本质的不同**。数学追求的是绝对的对错，但是统计学追求的是对现实的刻画和预测，从这一点上来说**统计学更像是物理，甚至于可以说它像是哲学，而不是数学。**

理学	0701 数学 0702 物理学 0703 化学 0704 天文学 0705 地理学 0706 大气科学 0707 海洋科学 0708 地球物理学 0709 地质学 0710 生物学 0711 系统科学 0712 科学技术史（分学科，可授理学、工学、农学、医学学位） 0713 生态学 0714 统计学（可授理学、经济学学位）

Statistics vs. Mathematics

- Statistics is a rather independent discipline
  - which uses many mathematical tools
- For example
  - Probability Theory
    - Statistics is about data with randomness
  - Optimization
    - Parameter Estimation is often modeled as an optimization problem



You may take a look at the first chapter of this book

Common Questions

- What is Statistics?
- Is Statistics a part of Mathematics?
- What’s the relation between Statistics and Probability?
- **What’s the relation between Statistics and Machine Learning?**

Statistics vs. Machine Learning

- No essential difference
- Many ML theories and algorithms are developed by statisticians
- Key elements of ML (new scenarios)
  - Computer: powerful computing
  - Data: big data
    - More samples
    - More dimensions

Statistics vs. Machine Learning

- Therefore, ML is different from *traditional statistics*
  - Which focus on models but not data
- And there are differences between different communities (CS and Stat)
  - Instead of asking “differences between Statistics and ML”
  - It is more interesting to ask
    - “Differences between CS community and Stat community in terms of ML research”

Reference: https://www.zhihu.com/question/48706077

ML: Statistics vs. CS

数据挖掘 统计学 机器学习 计算机科学 统计学习

关注者 2,476 被浏览 506,535

统计 PhD 和 CS PhD 学的 machine learning 有何不同?

相似的问题：机器学习专家与统计学家观点上有哪些不同？

他们也关注了该问题

已关注 写回答 邀请回答 好问题 77 3 条评论 分享 ··· 收起

关于理论保证（Theoretical Guarantees）

统计的算法往往一开始就有理论保证，即使这些保证可能和现实差比较远。比如大O上可能有一个很大的常数。

而因为CS不愿意做假设，所以CS的算法很少给理论保证，给的时候也会给得非常保守。

一个擅长 burning brain

一个擅长 BURNING NVIDIA

发布于 2022-04-11 10:20

统计：CS这帮人为了出结果什么破结构都敢往模型里堆

CS：统计这帮人为了证bound什么破assumption都敢用

编辑于 2017-03-21 11:54

ML: Statistics vs. CS

From *Statistical Modeling: The Two Cultures* by Leo Breiman

- "There are two cultures in the use of statistical modeling to reach conclusions from data."
  - "One assumes that the data are generated by a **given stochastic data model**."
  - "The other uses **algorithmic models** and treats the data mechanism as unknown."
- "The statistical community has been committed to the almost **exclusive** use of data models.
  - "This commitment has led to **irrelevant theory**, **questionable conclusions**"
  - "and has kept statisticians from working on a large range of interesting current problems."

ML: Statistics vs. CS

From *Statistical Modeling: The Two Cultures* by Leo Breiman

- “If our goal as a field is to use data to solve problems, then”
  - “we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.”

Finally

Further Readings:
https://statisticsbyjim.com/hypothesis-testing/interpreting-p-values/

Interesting Topics

That we do not cover here

- Misuse and Misinterpretation of Statistics
- History of Mathematical Statistics, Statistical Learning and Machine Learning
- The relation between ML Theory and TCS
- ……

Reference: Boosting学习理论的探索——一个跨越30年的故事 by 周志华

2013年的工作引起了很多反响，如在2014年国际人工智能大会(AAAI)上，国际人工智能学会主席、卡内基梅隆大学机器学习系主任曼纽拉 · 维罗索(Manuela Veloso)教授的Keynote报告将它作为人工智能领域的重要进展介绍，称其“使间隔理论复兴(*renascence*)”“为学习算法设计带来了新洞察(*new insight*)”。

然而，笔者仍有隐忧。虽然2013理论相应的泛化误差界在当时是最紧致的，但今后会不会有人基于其他的间隔物理量获得更紧的界，导致我们关于“AdaBoost为何未发生过拟合”的答案和“最大化平均间隔同时最小化间隔方差”的算法指导思想被推翻呢?

六年后，在2019年底的NeurIPS会议上，丹麦奥胡斯大学的阿兰 · 格洛隆德(Allan Grønlund)、卡斯柏 · 拉森(Kasper G. Larsen)、莱尔 · 卡玛(Lior Kamma)、亚历山大 · 马塞厄森(Alexander Mathiasen)与加州大学伯克利分校的杰拉尼 · 纳尔逊(Jelani Nelson)合作发表了一篇论文（见图7）。纳尔逊是美国总统奖和斯隆研究奖得主，拉森在STOC和FOCS曾两获最佳学生论文奖，是理论计算机科学界的新星，卡玛则毕业于以色列魏兹曼研究所这个计算机科学重镇。理论计算机科学家出手机器学习理论问题，是近年来的一个重要趋势。这篇论文最终证明了2013年我们给出的已经几乎是最紧的泛化误差上界，至多再改进一个log因子，并且这个上界已经与下界匹配，理论上不可能有更好的结果！