应用统计

清华大学数学科学系　　梁 恒

办公室：数学系荷二办公楼215#

电　话：010-62798741

Email:liangh@mail.tsinghua.edu.cn

概率与统计

- 分赌本问题
- 甲、乙两个赌徒进行一场9局5胜制的赌博每人有本金100元，胜者得到全部200元。当赌博在甲3:1领先时，被迫停止时，200元本金如何分配？
- Pascal的求解

统计：收集与分析数据的科学与艺术

- 统计和随机（概率）的概念已经深入到生活中的方方面面

  降水概率，病人的存活率，彩票 ……

- 对数据的理解

  Polya关于医生的玩笑：一名医生安慰他的病人说：你患了一种非常严重的病，患这种病的人只有十分之一能活下来。但是你不用担心，你到我这里来是十分幸运的，因为……

统 计 观 念

统计学与概率论的宗旨都是把**不确定现象量化**

差别在于：
概率论是数学，其基本特征是从法则到结果(from rules to results)，而统计学是一门科学，其基本特征是从结果到法则(from results to rules)。

统计学研究的主要内容是搜集和分析数据。
通过对数据的分析，从中提炼有用的信息，达到对未知事务的推断、对未来可能发生事件的预测等等。

统计学能够发挥作用的领域不胜枚举

- 科学：实证的科学研究离不开搜集和分析数据；
- 技术：技术的创新和改进离不开作试验和对试验数据的分析；
- 工农业生产：改进质量或提高产量离不开作试验和对试验数据的分析；
- 经济金融：对经济金融形势的分析与展望需要建立模型，离不开对大量数据的分析；投资、保险、股票等；
- 政府或公司的管理和决策——进行量化的管理和决策；
- 天气、水文、地震等的预报：建立模型，对大量数据进行分析；
- 医药疗效评估：FDA（食品药物管理局）对新的药物或治疗方法的效果评估有非常严格的统计标准；
- 国家或行业标准的制定：其中有大量的统计方法；
- 人口或其他社会现象（大选、热点问题）的调查：抽样调查；
- 社会卫生医疗和收入保障体系的制定；
- 法律法规的制定；

统计应用问题 例 1

- 1947年印度刚成立，首都发生了暴乱，一个称为红色堡垒的地方聚集了大量难民。政府有责任给难民提供食品，并将这个任务交给了承包商。由于没有任何关于难民人数的信息，政府被迫接受承包商所提出的食品和日用品的账单，这笔开支数量庞大。

政府无从检验承包商是否故意夸大商品的需求量，从中牟取暴利。政府为了避免受到承包商的蒙蔽，必须知道红色堡垒中避难者比较准确的人数。

困难在于因为宗教等原因，外人无法进入红色堡垒；同时也没有任何避难人数的先验信息

统计学家帮助政府完成了这项任务。

统计应用问题 例 1

- 统计学家利用了承包商交给政府的账单，这些账单记录了各种生活用品的供应量，如米、豆类和盐等。

- 假设全体避难者一天所需的米、豆和盐的总量为R, P, S。由消费调查，每人每天所需这些食物的量分别为r, p, s。

  因而R/r，P/p，S/s 提供了人数的估计值。

  统计学家计算发现三个比值中S/s最小， R/r最大。

  与盐相比，商品中大米的量可能被夸大了。考虑到当时印度盐的价格非常低，利润少，米的价格高，利润大。因此，统计学家提出用S/s作为避难人数的估计值，得到了很好的近似。

应用问题 例 2

莎士比亚的新诗

- 1985年11月14日，研究莎士比亚的学者Taylor从某图书馆发现一首写在纸片上的九节没有署名的新诗，共429个单词。他猜测是莎士比亚的作品。如何验证？

- Ronald Thisted, Bradley Efron,
- Did Shakespeare Write a Newly-Discovered Poem?
- Biometrika, 74(3), 1987, 445-455

莎士比亚著作中的用词总数为884647，其中31534是不同的，这些词的出现频数如下：

使用1次 14376； 使用2次 4343； 使用3次 2292

使用4次 1463； 使用5次 1043； 使用6次 837

……

使用大于100次 846 新诗所含429个单词中有258个是不同的

应用问题 例 2

莎士比亚的新诗

Table 1. *Number, m_x, of distinct words in the Taylor poem that appeared exactly x times in the Shakespearean canon*

x	0	1	2	3	4	5	6	7	8	9	Row total
0+	9	7	5	4	4	2	4	0	2	3	40
10+	1	0	3	0	1	1	1	2	1	0	10
20+	2	2	1	5	3	1	0	2	2	3	21
30+	4	1	1	1	2	1	0	0	3	3	16
40+	1	2	0	0	2	1	1	2	1	1	11
50+	0	1	1	1	1	0	0	1	0	2	7
60+	0	1	0	0	1	1	0	0	1	0	4
70+	0	0	1	0	0	1	0	0	1	1	4
80+	0	0	1	1	0	0	0	0	0	0	2
90+	0	0	0	1	0	1	1	0	0	0	3

For example, 9 distinct words in the poem appeared zero times in the canon, 7 appeared once each, etc.

应用问题 例 2

莎士比亚的新诗

Table 2. *Estimated expectation, $\hat{v}_x$, for the corresponding count m_x in Table 1, assuming Shakespearean authorship for the Taylor poem*

x	0	1	2	3	4	5	6	7	8	9
0–9	6·97	4·21	3·33	2·84	2·53	2·43	2·16	2·01	1·87	1·76
10–19	1·62	1·50	1·52	1·51	1·36	1·38	1·33	1·28	1·25	1·22
20–29	1·18	1·16	1·13	1·11	1·09	1·06	1·04	1·02	1·00	0·98
30–39	0·96	0·94	0·93	0·91	0·90	0·88	0·86	0·85	0·83	0·82
40–49	0·80	0·79	0·77	0·76	0·75	0·74	0·73	0·72	0·70	0·69
50–59	0·68	0·67	0·66	0·65	0·64	0·63	0·62	0·61	0·60	0·59
60–69	0·58	0·57	0·56	0·55	0·54	0·53	0·52	0·51	0·50	0·50
70–79	0·49	0·48	0·48	0·47	0·47	0·46	0·45	0·45	0·44	0·44
80–89	0·43	0·42	0·42	0·41	0·41	0·40	0·39	0·39	0·38	0·38
90–99	0·37	0·36	0·36	0·35	0·35	0·34	0·34	0·33	0·32	0·32

应用问题 例 2

莎士比亚的新诗

Table 3. *The eight poems analysed in this paper*

Abbreviation	Description	Total length	Number distinct words
1. JON	Ben Jonson; 'An Elegy'	411	243
2. MAR	C. Marlowe: four poems	495	272
3. DON	J. Donne; 'The Ecstacy'	487	252
4. CYM	Shakespeare: from 'Cymbeline'	323	215
5. PUC	from 'A Midsummer Night's Dream'	234	156
6. PHO	'The Phoenix and Turtle'	352	216
7. SON	Sonnets, Nos. 12–15	448	264
8. TAY	Taylor poem	429	258

应用问题 例 2

莎士比亚的新诗

	不同单词使用的频数				
莎士比亚作品中单词使用的次数	本约翰逊（哀歌）	马洛（四首诗）	多恩（狂喜）	新发现的诗	基于莎士比亚作品的期望值
0	8	10	17	9	6.97
1	2	8	5	7	4.21
2	1	8	6	5	3.33
3-4	6	16	5	8	5.36
5-9	9	22	12	11	10.24
10-19	9	20	17	10	13.96
20-29	12	13	14	21	10.77
30-39	12	9	6	16	8.87
40-59	13	14	12	18	13.77
60-79	10	9	3	8	9.99
80-99	13	13	10	5	7.48
不同单词数	243	272	252	258	258
单词总数	411	495	487	429	…

Table 4. *Words in the eight poems categorized according to their Shakespearean frequencies*

Poem	Category of x										
	0	1	2	3–4	5–9	10–19	20–29	30–39	40–59	60–79	80–99
1. JON	8	2	1	6	9	9	12	12	13	10	13
2. MAR	10	8	8	16	22	20	13	9	14	9	5
3. DON	17	5	6	5	12	17	14	6	12	3	10
4. CYM	7	4	3	5	13	17	9	12	17	4	4
5. PUC	1	4	0	3	9	6	9	4	5	9	3
6. PHO	14	5	5	9	8	18	13	7	13	8	5
7. SON	7	8	1	5	16	14	12	13	12	13	8
8. TAY	9	7	5	8	11	10	21	16	18	8	5

As in Table 1, except that the counts have been summed over 11 categories.

Table 5. *Estimated expected values for the counts in Table* 4, *assuming Shakespearean authorship*

Poem	Category of x										
	0	1	2	3–4	5–9	10–19	20–29	30–39	40–59	60–79	80–99
1. JON	6·68	4·03	3·19	5·14	9·81	13·16	9·94	8·18	12·68	9·17	6·83
2. MAR	8·04	4·86	3·85	6·19	11·81	15·91	12·03	9·92	14·92	10·72	8·26
3. DON	7·91	4·78	3·78	6·09	11·62	15·59	11·77	9·68	14·99	10·83	8·06
4. CYM	5·25	3·17	2·51	4·04	7·71	10·35	7·82	6·44	9·99	7·23	5·39
5. PUC	3·79	2·29	1·81	2·91	5·57	7·47	5·65	4·66	7·22	5·23	3·91
6. PHO	5·72	3·46	2·73	4·40	8·40	11·28	8·52	7·02	10·87	7·87	5·87
7. SON	7·28	4·40	3·48	5·60	10·69	14·52	11·10	9·06	13·71	10·02	7·96
8. TAY	6·97	4·21	3·33	5·36	10·24	13·96	10·77	8·87	13·77	9·99	7·48

应用问题 例 2

莎士比亚的新诗

4. MODELLING AND TESTING

We wish to test whether the observed counts m_x for each of our eight poems fit the predicted values $\hat{\nu}_x$ based on the assumption of Shakespearean authorship. Our tests will rely upon the following regression model, that for $x = 0, 1, \ldots, 99$, the m_x have independent Poisson distributions with means μ_x, where

$$\log \mu_x = \log \hat{\nu}_x + \beta_0 + \beta_1 \log (x+1). \tag{4·1}$$

The quantities $\hat{\nu}_x$ are considered to be constants in what follows, for example having the values shown in Table 2 for the Taylor poem. The null hypothesis

$$H_0: \beta_0 = \beta_1 = 0 \tag{4·2}$$

corresponds to $\mu_x = \hat{\nu}_x$, that is, to perfect agreement with Shakespearean usage.

Model (4·1) is motivated in § 6. It is a generalized linear model of the simplest sort, as described by McCullagh & Nelder (1983, Ch. 2). Expression (4·1) can be rewritten as

$$\mu_x / \hat{\nu}_x = e^{\beta_0}(x+1)^{\beta_1}.$$

统计应用问题 例 3

小儿麻痹症疫苗的有效性

- 问题： 小儿麻痹疫苗问世后，1954年进行了一项研究以评价它在预防幼儿麻痹及死亡方面的有效性。两组幼儿参加了这项研究。一组按规定接受三次疫苗，另一组则不接受疫苗。后一组作为证实疫苗有效性的对照是必须的。比较的最重要的判据是两组中发生麻痹以及死亡的幼儿数。由于小儿麻痹症发病率极低，两组都需要大量的幼儿以保证有足够的病例发生，从而为比较提供可靠的基础。Meier的文章称该项研究是“有史以来最大规模的公共卫生试验”。两组人数都略多于200000名小孩。

- 决定每个小孩是否接受疫苗使用了随机化的方法。这里使用的是分层随机化。全美国的许多学校参加了这项计划，在每个参加学校分别进行随机化抽样，使得每个学校中接受疫苗（试验组）和没有接受疫苗（对照组）的小孩数目大致相等。从而相对高发病率地区和相对低发病率地区的学校都有大致相等数目的随机选择的试验组和对照组小孩。

统计应用问题 例 3

小儿麻痹症疫苗的有效性

表 试验组和对照组小儿麻痹发病率

组别	幼儿人数	发病人数	发病率（每十万人）
试验组	200745	33	16
对照组	201229	1.15	57

- 每一位不接种的小孩接受三次生理盐水（医学上称为安慰剂）的注射。
- 该项试验中安慰剂的目的是为了使幼儿、家长、注射者，以及当某一幼儿患病时为其治疗的大夫都不知道这个小孩接受的是疫苗还是生理盐水。
- 两组幼儿的发病率是否有本质的差异？ 差异大小的点估计和区间估计是什么？回答这些问题是统计推断的重要内容。

随机化方法（思想）的应用

- 对敏感问题的随机化处理

- 美国某大学想了解学生中吸食毒品的情况。如果直接问“你吸食毒品吗？”很难得到真是的回答。

- 利用随机化的策略，列出如下两个问题

  S：你吸食毒品吗？

  T：你的电话号码的尾数是偶数吗？

要求被提问者抛掷一个硬币，出现正面回答S，出现反面回答T。假设回收到500份问卷，其中198回答“是”。

可得吸食毒品的学生比例的估计值 73/250。

统计应用问题 例 4

吸烟与健康（吸烟者的死亡率）

- **问题：** 1951年到1959年期间，曾经有过7次大规模的对吸烟男性死亡率的比较研究。1次在英国，1次在加拿大，5次在美国。除了一些微调，研究计划基本是一致的。首先，给选定组别的人送一份调查表，询问最近及过去的吸烟习惯以及其他一些情况，如年龄等。启动一套程序以保证一旦回答问卷的人死了，这一消息会马上被报告、记录，并得到死亡原因诊断（通过死亡证明书或尸体解剖报告）。研究涉及的人数最少有34000，最多达到448000。

- 这些研究包含众多的死亡率和死亡原因可以进行比较的样本
  （1）不同类型的吸烟者——不吸烟者、吸香烟者、吸雪茄者、吸烟斗者、混吸着；
  （2）给定类型的不同吸烟量；
  （3）给定吸烟类型和吸烟量，不同的开始吸烟的年龄；
  （4）通过戒烟的时间和戒烟前的吸烟量，对戒烟者分类。

统计应用问题 例 4

吸烟与健康（吸烟者的死亡率）

- **基于观察的研究** 当人们希望通过这些分组之间的比较得出结论时，他们发现，吸烟研究与前面口感舒适度与小儿麻痹等研究之间有一个主要的逻辑上的差别。后两种研究，研究人员能够决定哪一组对象接受怎样的试验。可以通过随机化处理保证各组间除了试验方法以外没有系统差异。

- 可是在吸烟研究中，研究人员无法指定对象分组。分组依靠的是对象（即吸烟者）的吸烟习惯。这样，除了吸烟以外，不同类型的吸烟者之间可能存在多方面的系统差异对死亡率产生影响。例如，吸雪茄和烟斗的人的年龄通常会比不吸烟者大得多。年轻人更倾向于吸香烟。众所周知，中年以后的死亡率随年龄逐步增高。所以，简单的死亡率比较会有利于吸香烟者，而严重的不利于吸雪茄和烟斗的人。进而，不同类型的吸烟者其饮食习惯、体育锻炼以及其他众多可因素都可能影响死亡率。

- 为了避免这些偏差，研究人员尝试将外部条件相似的人群按吸烟习惯分组，并调整死亡率。但是，这样做大大提高了统计分析的复杂度，同时也缺乏充分的说服力。因为很难证明考虑到了所有重要的外界因素使得样本均匀，以及是否做了正确的度量和调整。类似吸烟的这种研究通常称为是基于观察的。它的意思是提醒人们，研究人员缺乏为了进行比较而创造分组的能力，而是不得不受数据的限制。

统计学有自己独特的思维方式与方法

统计学本质上是一门应用性、方法性的学科

统计的目的是回答实际领域中提出的各种问题，对科学结论提供定量分析（而不是单纯定性分析）的依据；为发现新的理论模型提供线索；预测未来，为决策提供支持等。因此**统计学以问题为导向，而不以理论为导向。**

统计的对象—数据是部分的、具有不确定性

因此所有的统计结论都可能是错的！你不可能得到绝对正确的结论，只能设法尽量降低因犯错误所造成的损失。

没有一种统计方法是“最好的”

对同一组数据有可能用不同的方法去分析，甚至得到相互矛盾的结论！任何一组数据都有一定的背景。用什么方法分析可以得到“好的”或“更好的”结果？这需要大量统计分析的实践经验。在某种程度上可以说：统计既是科学，又是艺术。

对数据的考察（第一手数据）

- 一个统计学者被邀请分析某落后地区一些人类测量学方面的数据。测定的10个特征中有一个是体重。原始的测量记录为：7.6, 6.5, 8.1, 7.4, …等。这里的重量单位是英石，1英石等于14磅。7.6*14=106.4磅…

- 统计学者开始拿到的是经过换算得到的以磅为单位的体重记录。但是他认为应该查看原始记录。在查看整个记录时，他发现了一个奇异点，所有的重量测量值里小数点后面从来没有出现过7，8，9三个数字！进一步调查发现，这一地区在测量重量时将一英石分为7个单位，并非使用的10进制。

- 正确的体重转换应该是7*14+(6/7)*14=110磅…

统计应用问题 例 5

解读数据--研究生入学的性别歧视

1973年,共有8442男生，4321女生申请加州大学Berkeley分校的研究生院。
男生录取比例大约 44%，女生录取比例大约35%

Science

7 February 1975:

Vol. 187 no. 4175 pp. 398-404

Sex Bias in Graduate Admissions: Data from Berkeley

P. J. Bickel, E. A. Hammel, J. W. O'Connell

统计应用问题 例 5

解读数据--研究生入学的性别歧视

加州大学Berkeley分校6个最大专业研究生入学录取比例，男生为44.5%，女生为30.3%。是否存在对女性考生的歧视？

加州大学 Berkeley 分校 6 个最大专业的研究生入学资料

	男		女	
专业	申请人数	录取百分比	申请人数	录取百分比
A	825	62	108	82
B	560	63	25	68
C	325	37	593	34
D	417	33	375	35
E	191	28	393	24
F	373	6	341	7

研究生入学的性别歧视

观察数据发现男生偏向报考容易的专业，而女生则相反

1. A、B 两个专业容易考取。51.5%的男生申请，女生申请率只有7.25%；
2. 其他四个专业较难考取，90%以上的女生申请这四个专业。

简单的看入学率是不合理的，简单的看各系的录取率同样不全面。
更合理的考察应该是加权入学率，即综合考虑到各系的规模和录取率。

	A	B	C	D	E	F
申请人数	933	585	918	792	584	714
申请比例	0.206	0.129	0.203	0.175	0.129	0.158
男生录取率	0.62	0.63	0.37	0.33	0.28	0.06
女生录取率	0.82	0.68	0.34	0.35	0.24	0.07

男生的加权平均入学率=0.39；　　女生的加权平均入学率=0.43

哪一组显得更随机一些

A

0 0 0 0 1 0 1 0 1 1 1 0 0 1 0 1 0 0 0 0 1 0 1 1 1
0 0 0 0 1 0 0 1 0 0 1 1 1 1 0 0 0 0 0 0 1 0 1 1 0
1 0 0 1 1 1 1 0 1 1 1 1 1 0 1 1 0 0 0 0 1 0 0 1 0
0 1 1 0 0 0 0 1 0 0 0 0 0 1 0 0 1 1 0 1 0 0 1 1 0
1 1 1 0 0 0 0 0 0 0 1 0 1 0 1 1 0 1 1 0 0 1 0 1 0
0 1 1 0 1 1 0 0 0 0 1 0 1 1 1 1 1 0 1 0 1 1 1 1 1
1 0 0 0 1 0 1 0 1 1 1 0 0 1 0 0 1 1 1 1 1 0 0 0 0
0 0 1 1 1 1 0 1 0 0 0 0 0 0 1 1 0 0 1 0 1 1 0 1 0

B

1 1 1 1 0 0 1 0 0 1 0 0 1 0 1 0 0 0 1 0 0 1 1 1 1
0 0 1 0 1 0 0 0 1 0 0 1 1 0 1 0 1 1 1 0 0 1 1 1 0
0 1 1 1 1 0 1 1 0 1 0 1 1 1 1 0 1 1 0 1 0 0 1 0 0
0 1 1 1 0 0 1 1 0 1 0 1 1 0 1 0 1 0 1 1 0 1 0 0 1
1 0 1 0 1 0 0 1 1 1 0 1 1 0 1 0 0 1 0 0 0 0 0 1 1
0 1 0 1 1 0 1 1 0 0 0 0 1 1 1 0 1 0 1 1 0 0 1 1 0
0 1 1 1 0 0 1 1 1 1 0 1 1 0 0 0 1 1 1 0 0 1 1 0 1
0 1 1 1 0 0 1 1 0 0 0 1 1 1 1 1 0 1 0 1 0 0 1 1 1

掷硬币的随机性鉴别

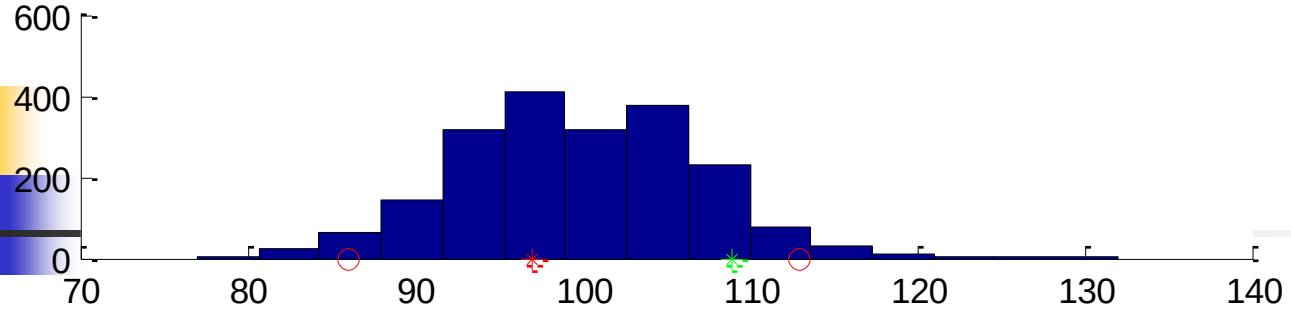
1. 定义关于 200 位 0-1 序列的统计量；找到公平抛掷(即 b(1, 0.5))假设下统计量的分布，统计量的分布可能可以解析地算出，也可用直方图等近似；

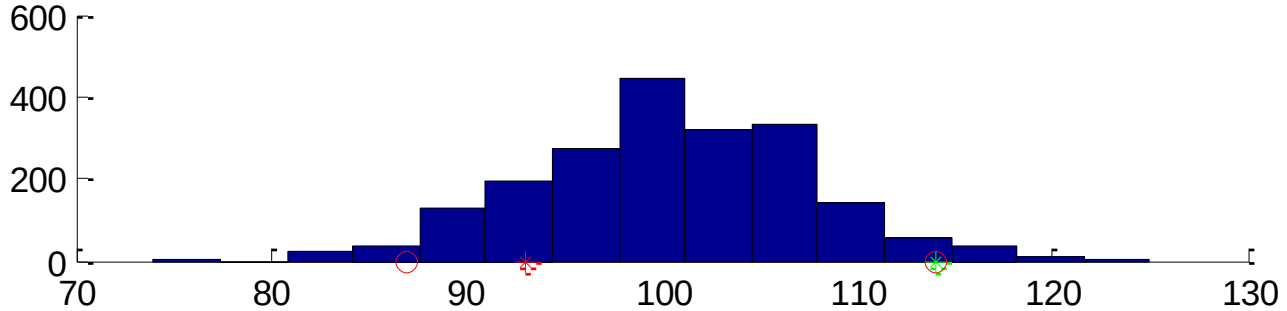
2. 提供几个对这个问题不一定是很好的统计量供大家参考

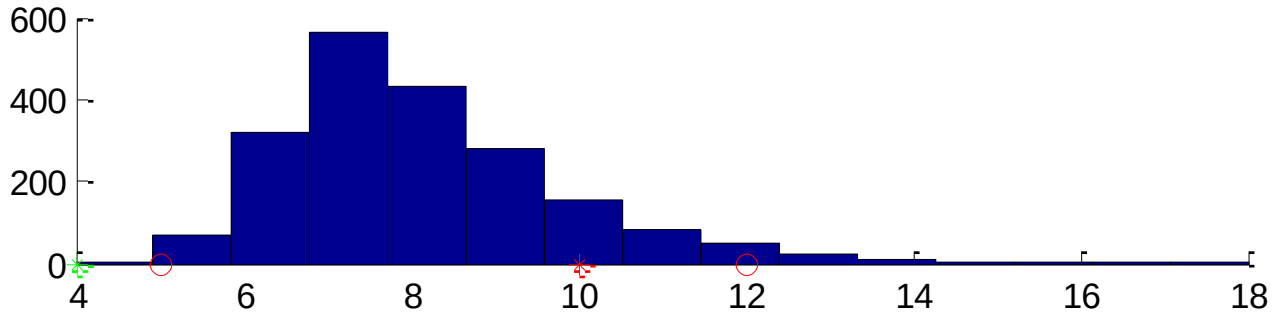
   (1) 正面的个数；
   (2) 最长0或1串的长度；
   (3) 0-1变化次数，比如01001的切换次数为4， 0-1-0-1 ；

3. 争取提出更多适合的统计量进行判断。

统计量的分布与经验分布函数







nn=2000; % sample size

作业与考试

- 总评分
  - 考试 60%
  - 作业、报告与考勤等 40%