Statistics for Data Science

Chris Qi

Ph.D. University of Maryland

Why statistics (and probabilities)

- help us make better and wiser decisions under uncertainty and with limited information (rainfall, odds of winning, toss a coin...)

- Overview of today's lesson
  - Basic statistics
  - Sampling, random variable
  - Probability distribution
  - Confidence interval

Two types of statistics

- 描述性统计 （descriptive statistics ）
  - 平均数（mean），众数(mode)，中位数(median)（central tendency）
  - range, 分位数(quantile), 标准差(standard deviation) （dispersion）
- 推论统计（inference statistics）
  - Use sample to infer population
  - very costly, time-consuming, not feasible to survey each individual
  - So, we use a sample to estimate the parameters of a population
- Backbone of data science

standard deviation

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{n}}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

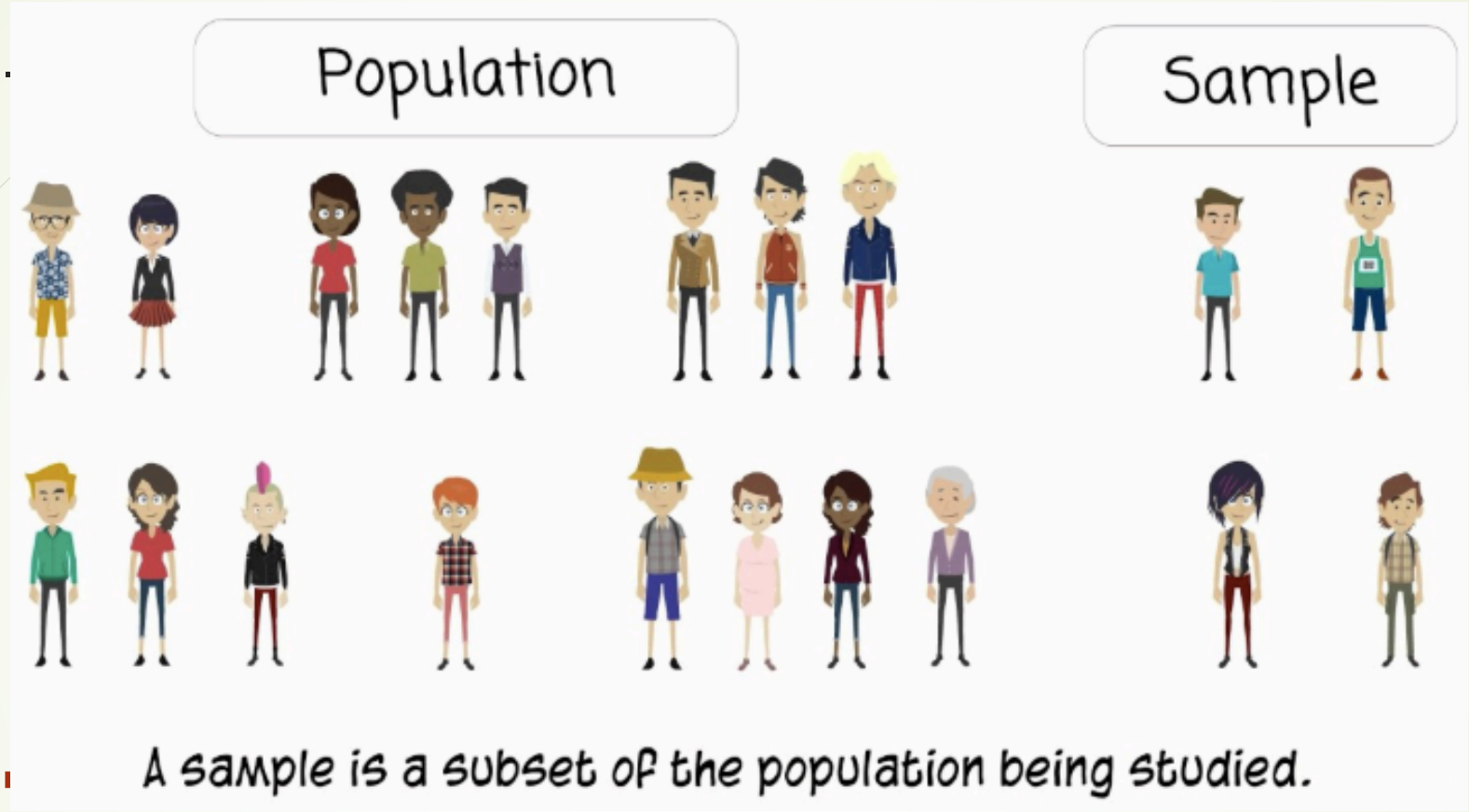
descriptive statistics

Name	Size	x_i -mu	(x_i -mu)^2					
Alice	40	-0.636	0.405		Median	40		
Ben	39	-1.636	2.678		Mean	40.64		
Cindy	42	1.364	1.860		Mode	40		
Dan	43	2.364	5.587		SD	21.69		
Casey	40	-0.636	0.405					
Peter	42	1.364	1.860					
Jenny	40	-0.636	0.405					
Maggie	45	4.364	19.041					
Jackie	34	-6.636	44.041					
Kevin	32	-8.636	74.587					
Simon	50	9.364	87.678					
			238.545					

$$
\begin{aligned}
\sigma^2 &= \frac{\sum (X-\mu)^2}{N} && (1)\\
&= \frac{\sum (X^2 - 2\mu X + \mu^2)}{N} && (2)\\
&= \frac{\sum X^2}{N} - \frac{2\mu \sum X}{N} + \frac{N\mu^2}{N} && (3)\\
&= \frac{\sum X^2}{N} - 2\mu^2 + \mu^2 && (4)\\
&= \frac{\sum X^2}{N} - \mu^2 && (5)
\end{aligned}
$$

inference statistics

- 推论统计（inference statistics）
  - Use sample to infer population
  - very costly, time-consuming, not feasible to survey each individual
  - So, we use a sample to estimate the parameters of a population



- 幸存者偏差，二战盟军统计学家沃尔德
- 老物件、双盲实验、成功者的故事（从大学退学的特征）
- 不能只看贼吃肉，不看贼挨揍

Random variable (随机变量)

- map outcome of a random process to numbers （随机过程-->数字）
- e.g: toss a coin, or a dice

- why it is useful?
  - with numbers, we can do more things easily
- often uppercase
- P(X<k)
- F(x)=P(X<x), x的分布函数（distribution function）

Discrete and Continuous

- random variable
  - Discrete variable： you can list the values, and count the number of values 离散型随机变量，有限，可列举。抛硬币、做重复实验直到成功的次数
  - Continuous variable 连续型随机变量，取值不可列 。某个时间段的降雨量

Discrete variable

- distribution law 分布律
- list values that can be assigned to X. 1,2 ,3, 4, 5, 6
- P(X=x)=p
- list
- graph

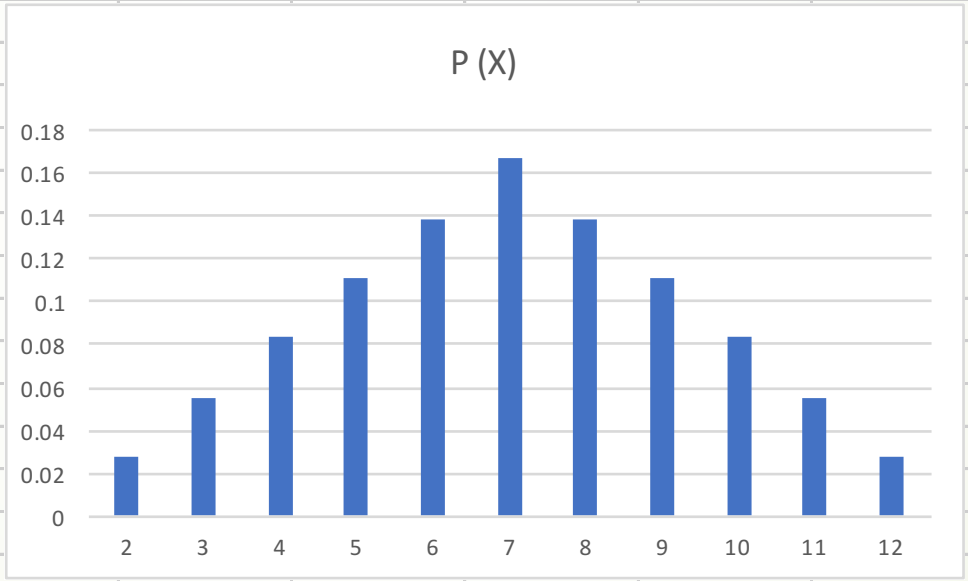
probability distribution -binomial distribution(二项分布)

- 是*n*个独立的 是/非试验 中成功的次数的离散概率分布，其中每次试验的成功概率为*p*。这样的单次成功/失败试验又称为伯努利试验。
- consider: X, the number of “head” after 3 flips of a fair coin.
- HHH HHT HTH THH THT HTT TTH TTT
- P(X=0)=1/8
- P(X=1)=3/8
- P(X=2)=3/8
- P(X=3)=1/8

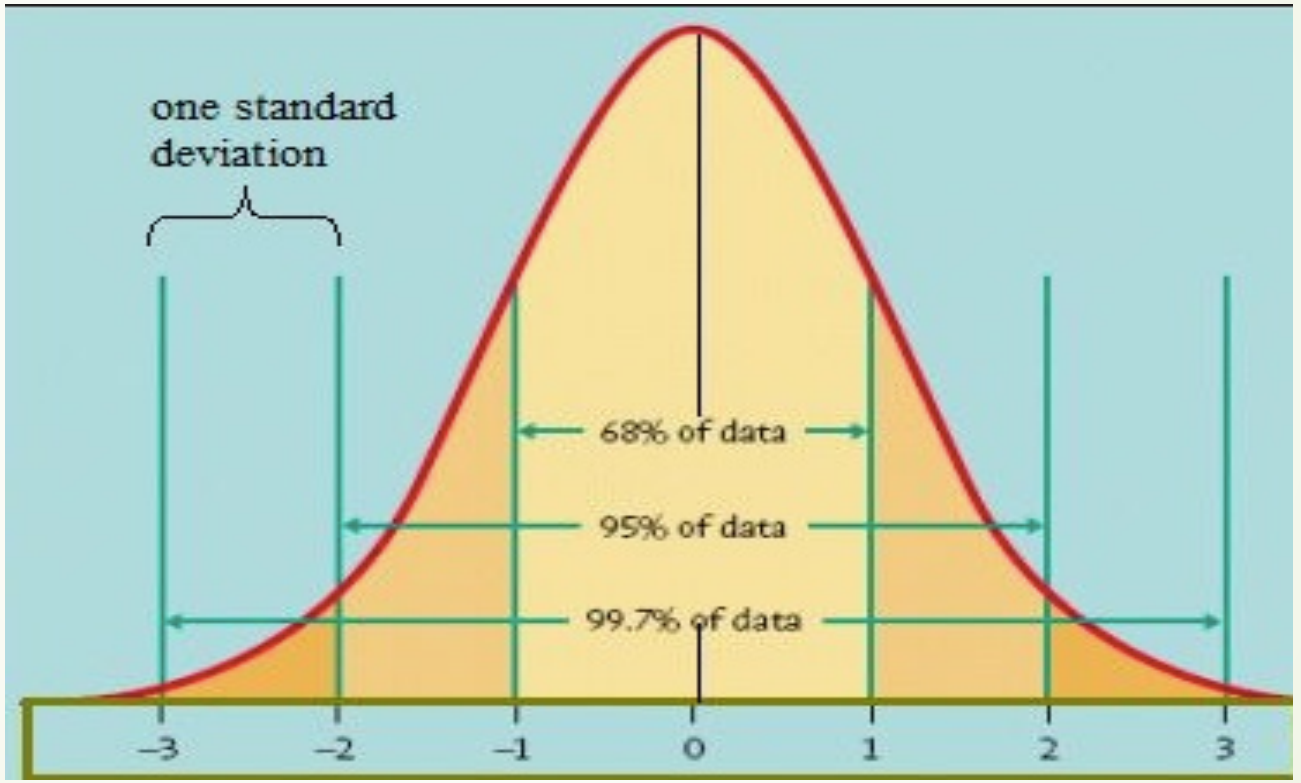
Discrete variable example

- sum of two dices

X	P (X)
2	1/36
3	1/18
4	1/12
5	1/9
6	5/36
7	1/6
8	5/36
9	1/9
10	1/12
11	1/18
12	1/36



Normal distribution

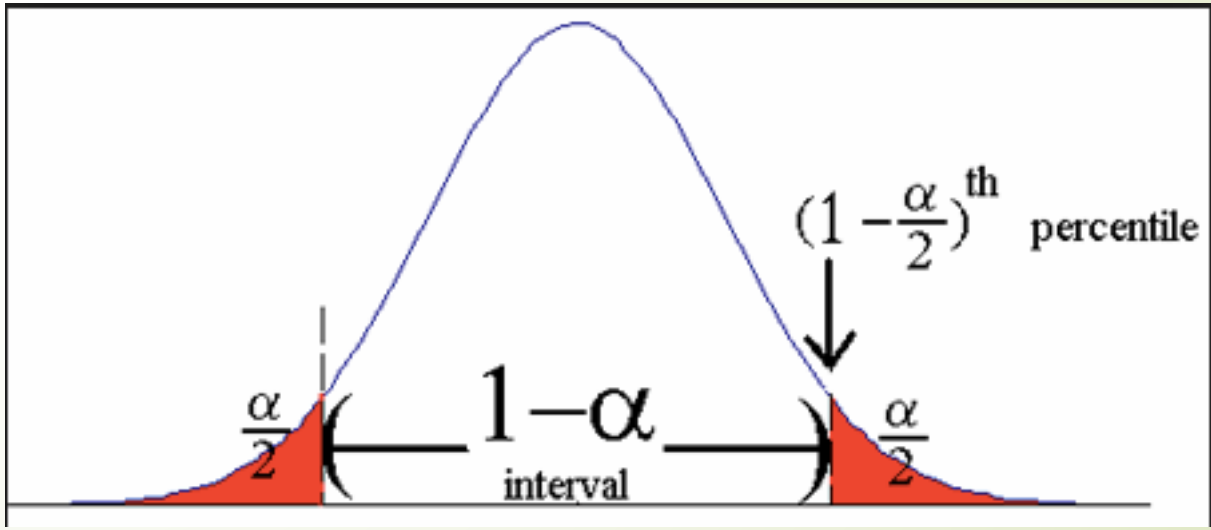


Confidence Interval （置信区间）

- 用样本统计量估计总体参数，难免存在误差
- 我们需要确定一个合理的区间，使参数真值落在其中的概率(1-α)达到指定要求。
- θ
- $P(\underline{\theta} < \theta < \bar{\theta}) = 1 - \alpha$
  - θ，参数真值，未知，需要估计
  - $\bar{\theta}$ upper bound, $\underline{\theta}$ lower bound
  - $1 - \alpha$ confidence level 置信水平
- $\bar{\theta} - \underline{\theta}$ 区间长度，反映精度
- $1 - \alpha$，反映信度
- α越小，$1 - \alpha$越大，信度就越大，但是精度就会越小，置信区间扩大

Confidence Interval example

- Q: x_1 x_2 x_3 $x_{4...}$ x_n 来自总体X ~ N (μ,1)的样本，求μ的置信度为 1- α的置信区间
  - 直观理解？
- $\bar{X} \sim N(\mu, \frac{1}{n})$ ☞ $\frac{\bar{X}-\mu}{1/\sqrt{n}} \sim N(0, 1)$
- So~
- $P\left\{-\mu_{1-\frac{\alpha}{2}} < \frac{\bar{X}-\mu}{1/\sqrt{n}} < \mu_{1-\alpha/2}\right\}$= 1- α



所以 μ 的置信度为 $1-\alpha$ 的置信区间为

$$(\bar{X}-\frac{1}{\sqrt{n}}u_{1-\alpha/2},\ \bar{X}+\frac{1}{\sqrt{n}}u_{1-\alpha/2})$$

若取 $n=16$，$\alpha=0.05$，查表得到 $u_{1-\alpha/2}=1.96$

则 μ 的置信度为 95% 的置信区间为

$$(\bar{X}-0.49,\ \bar{X}+0.49)$$

- https://en.wikipedia.org/wiki/Standard_normal_table

置信区间的含义

反复抽取容量为 16 的样本，每次都可以根据样本观测值$x_1, x_2, \cdots, x_n$算得样本均值 $\bar{x}$，得到一个区间

$$(\bar{x} - 0.49,\ \bar{x} + 0.49)$$

此区间可能包含未知参数μ的真值，也可能没包含.

而包含未知参数μ的区间个数约占95%，不包含未知参数μ的区间个数约占5%.

Summary

- Descriptive statistics
  - basic statistics: mean, mode, median, standard deviation
- Inference statistics
- Confidence Interval
- Random variable
  - sampling issue
- Probability distribution