社会统计学及SPSS软件应用
STATISTICS WITH SPSS

Instructor:王荣欣

Email: rxwang@qq.com

周二3-4节、单周周四3-4节，3A106-2

2020年9月24日

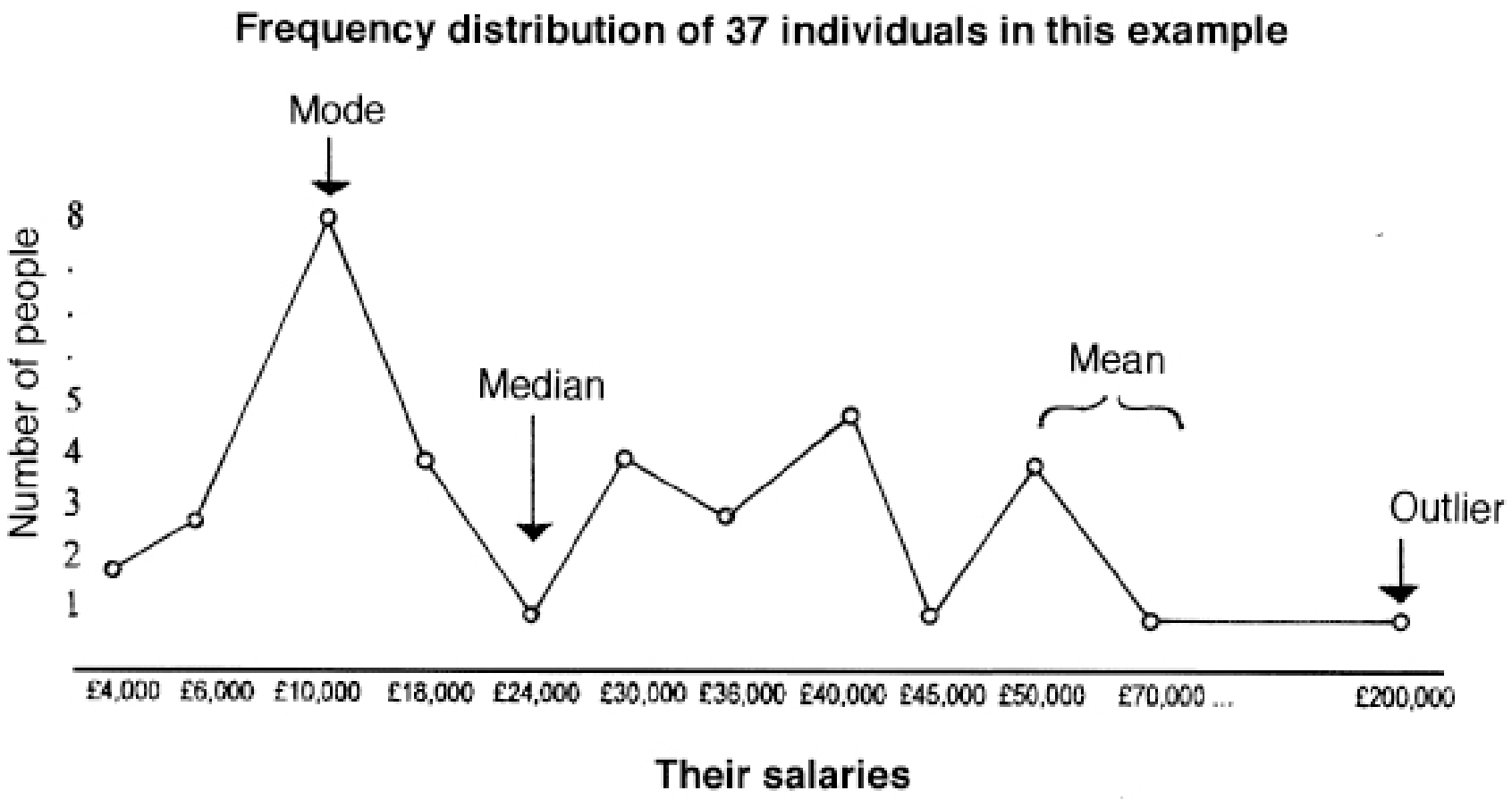
CONTENTS

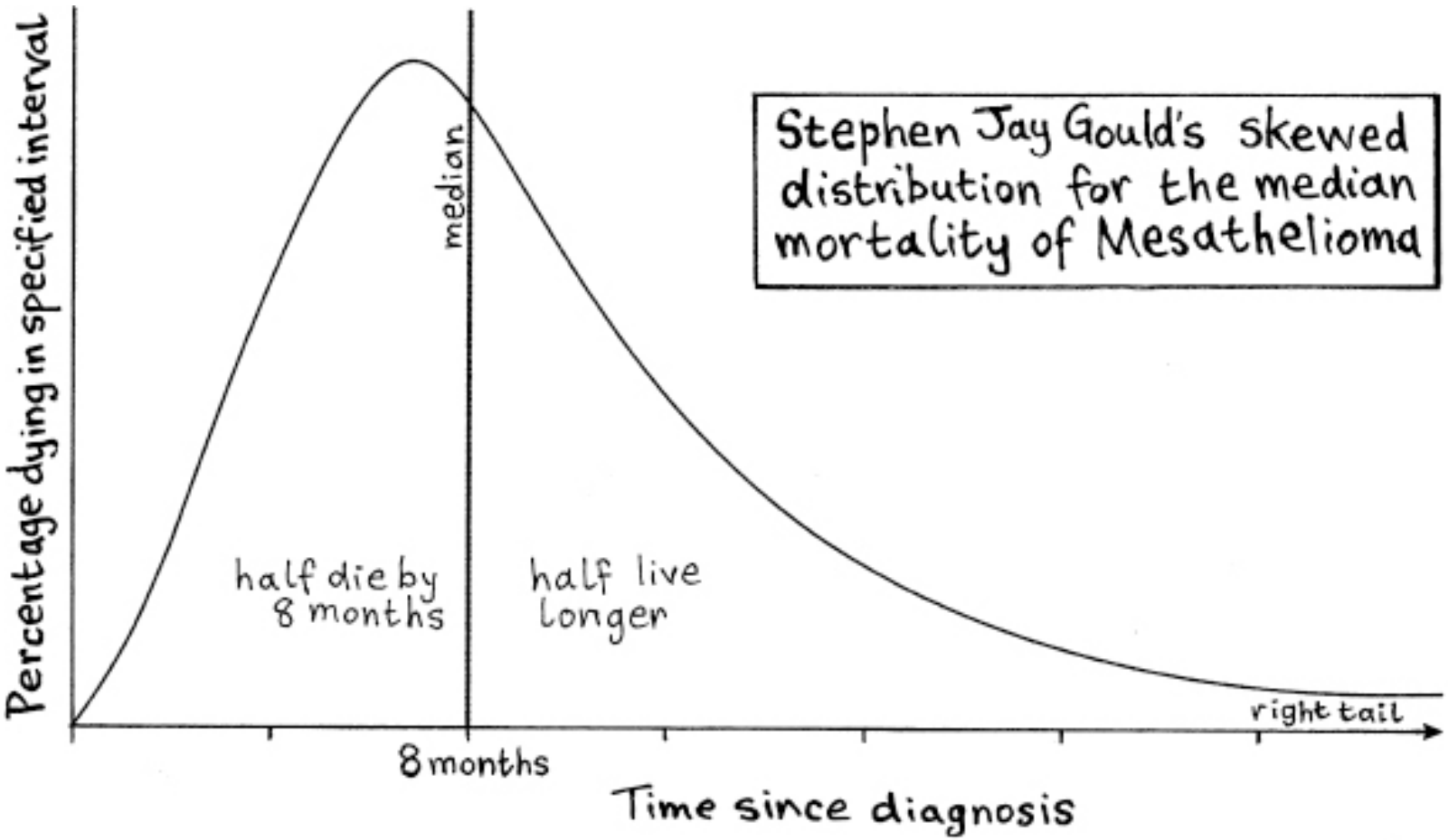
- Binomial Distribution
- Central Limit Theorem
- Sampling distribution

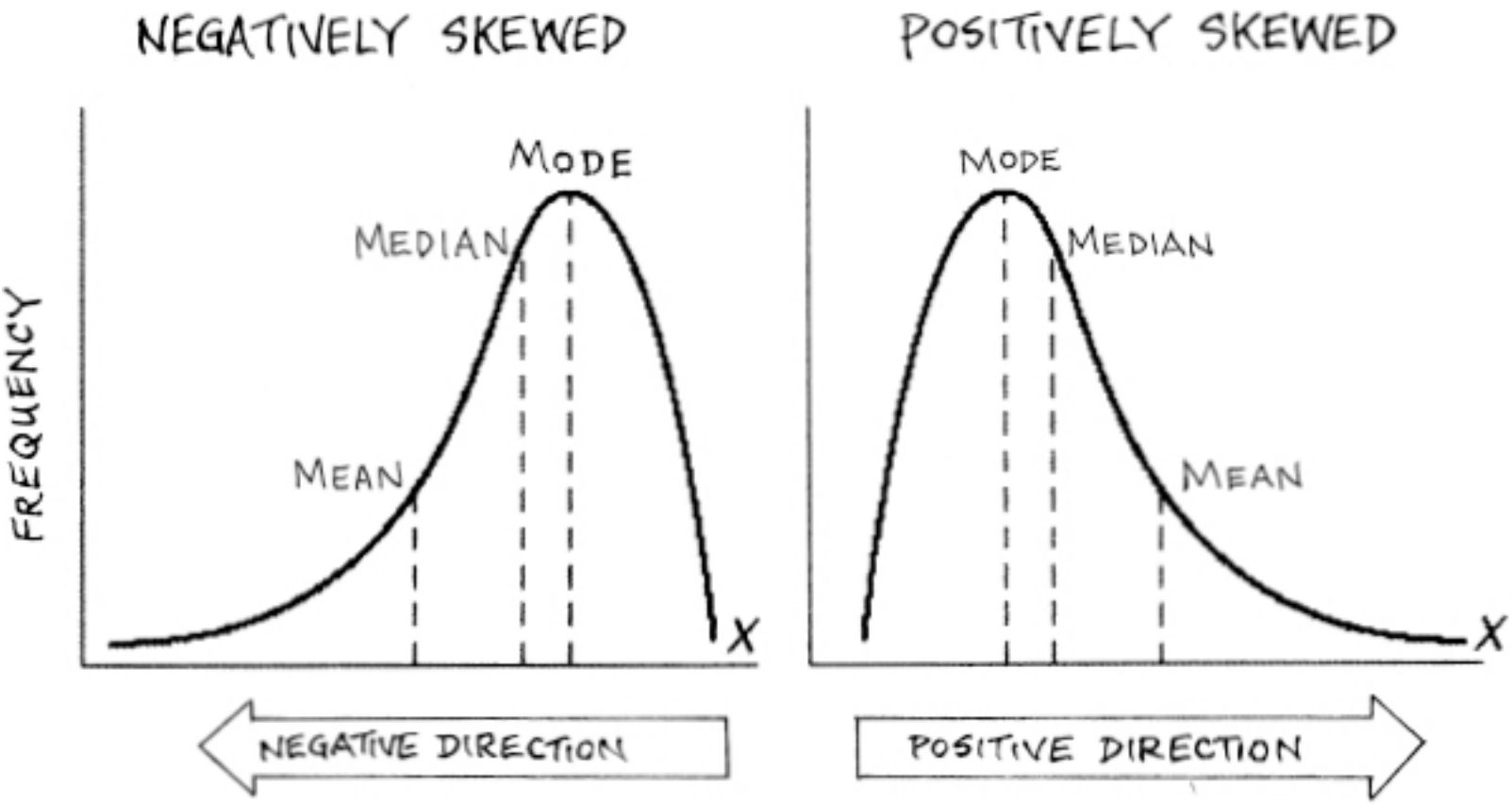
X = One person

Number of people	Salary	
XX	£4,000	
XXXXXX	£6,000	
XXXXXXXX	£10,000	—— mode: the one occurring most frequently
XXXX	£18,000	
X	£24,000	—— median: the one in the middle with twenty people above and twenty people below
XXXX	£30,000	
XXX	£36,000	
XXXXX	£40,000	
XX	£45,000	
XXXX	£50,000	
X	£70,000	
X	£200,000	

The mean value = **£60,400**
The modal value (with eight people) = **£10,000**
The median value = **£24,000**


Frequency distribution of 37 individuals in this example
Mode
Median
Mean
Outlier
Number of people
8
5
4
3
2
1
£4,000
£6,000
£10,000
£18,000
£24,000
£30,000
£36,000
£40,000
£45,000
£50,000
£70,000 ...
£200,000
Their salaries

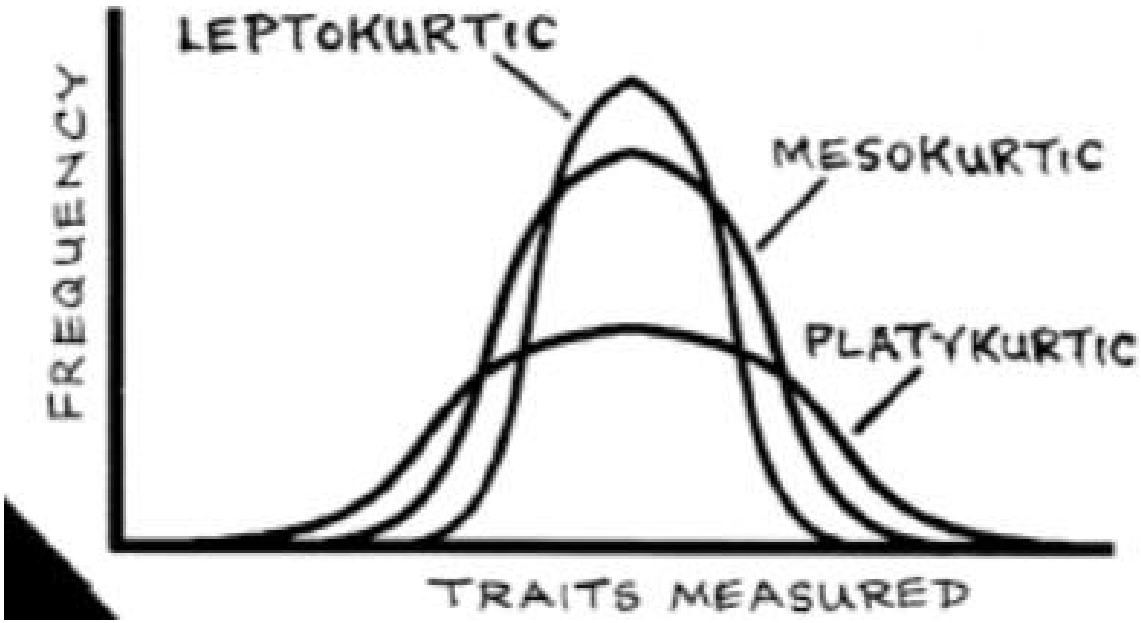

Stephen Jay Gould's skewed distribution for the median mortality of Mesathelioma
Percentage dying in specified interval
median
half die by 8 months
half live longer
right tail
8 months
Time since diagnosis


NEGATIVELY SKEWED
MODE
MEDIAN
MEAN
FREQUENCY
X
NEGATIVE DIRECTION
POSITIVELY SKEWED
MODE
MEDIAN
MEAN
X
POSITIVE DIRECTION

Leptokurtic Distributions

- These have high kurtosis, and thus long tails. Gosset drew these lepping (leaping) kangaroos to help him remember that leptokurtic distributions have long tails.


PLATYKURTIC
LEPTOKURTIC
"STUDENT'S"
ILLUSTRATIONS
FROM HIS PAPER
"ERRORS OF
ROUTINE ANALYSIS"
1927.


LEPTOKURTIC
MESOKURTIC
PLATYKURTIC
FREQUENCY
TRAITS MEASURED

1. Nominal variable
   - Each category is listed with its corresponding frequency - the number of observations falling into that category.
   - Frequency distributions can be portrayed in bar chart.（柱状图或条形图）
2. Distributions for interval and ratio variables can be portrayed in histogram.（直方图）
   - 直方图的纵轴有两种处理方式：一是代表频数，二是代表密度。

箱线图（boxplot）：展示连续型变量。箱线图的基本三要素：

- 箱子中间的一条线（数据的中位数）
- 箱子的上下限（数据的上四分位数和下四分位数）。箱子的高度反映数据的波动程度。
- 箱子上下方的两条线

STATA命令（CGSS2003）

1 以下展示如何进行列联表分析，给出chi2 gamma taub系数。

tab educ sex, row col chi2 gamma taub

2 因变量为连续变量的列联表分析

tab edu, sum(incmonth)

tab sex, sum(incmonth)

统计学的基本概念

1. 描述性统计（descriptive statistics）
2. 统计推断（inferential statistics or statistical inference）是通过样本统计量来推断未知的总体参数。

- 统计的核心：推论性统计
- 抽样数据所代表的总体是什么情况？
  - 不应存在针对样本下结论的研究。
- 样本在多大程度上可以反映样本所来自的总体？
  - 统计的显著性检验（通过统计显著性，推翻原假设）（用样本识别总体）

SAMPLE AND POPULATION

- Statistics ($\bar{y}$, s or r) are to samples what parameters (μ, σ, ρ) are to populations.
- Every sample drawn from the population has its own statistics ($\bar{y}$, s or r), which is used to estimate the parameter (μ, σ, ρ) of its population.
- 总体的概括性的、相对稳定的特征，称为总体参数（**population parameter**）
- 通过样本计算得到的样本特征，称为样本统计量（**sample statistic**）。

Binomial Distribution

Binomial distribution is a discrete probability distribution and represents the probability of two outcomes, which may or may not occur. It describes the possible number of times that a particular event will occur in a sequence of observations. The distribution was introduced by the Swiss mathematician Jacques Bernoulli.

- **The binomial distribution** (The term binomial means ”two number”)

$$(p+q)^n \tag{1.1}$$

is determined by the number of observation n and the probability of occurrence, denoted by p+q (the two possible outcomes).

A BINOMIAL EXPERIMENT

1. There can be only two outcomes per trial - call them success and failure
2. There must be n repeated, independent trials
3. The probability of success in each trial must be constant

Binomial Distribution

- If n is 2

$$(p+q)^2 = p^2 + 2pq + q^2 \tag{1.2}$$

- If n is 3

$$(p+q)^3 = p^3 + 3p^2q + 3pq^2 + q^3 \tag{1.3}$$

- If n is 5

$$(p+q)^5 = p^5 + 5p^4q + 10p^3q^2 + 10p^2q^3 + 5pq^4 + q^5 \tag{1.4}$$

Binomial Distribution

- If n is 2

$$(p+q)^2 = p^2 + 2pq + q^2 \tag{1.2}$$

- If n is 3

$$(p+q)^3 = p^3 + 3p^2q + 3pq^2 + q^3 \tag{1.3}$$

- If n is 5

$$(p+q)^5 = p^5 + 5p^4q + 10p^3q^2 + 10p^2q^3 + 5pq^4 + q^5 \tag{1.4}$$

Binomial Distribution

- If n is 2

$$(p+q)^2 = p^2 + 2pq + q^2 \tag{1.2}$$

- If n is 3

$$(p+q)^3 = p^3 + 3p^2q + 3pq^2 + q^3 \tag{1.3}$$

- If n is 5

$$(p+q)^5 = p^5 + 5p^4q + 10p^3q^2 + 10p^2q^3 + 5pq^4 + q^5 \tag{1.4}$$

Binomial Distribution

- If p=q=0.5

$$\begin{aligned}(p+q)^2 &= p^2+2pq+q^2 \\ &= \frac{1}{4}+\frac{2}{4}+\frac{1}{4}\end{aligned} \tag{1.5}$$

- If p=q=0.5

$$\begin{aligned}(p+q)^5 &= p^5+5p^4q+10p^3q^2+10p^2q^3+5pq^4+q^5 \\ &= \frac{1}{32}+\frac{5}{32}+\frac{10}{32}+\frac{10}{32}+\frac{5}{32}+\frac{1}{32}\end{aligned} \tag{1.6}$$

Binomial Distribution

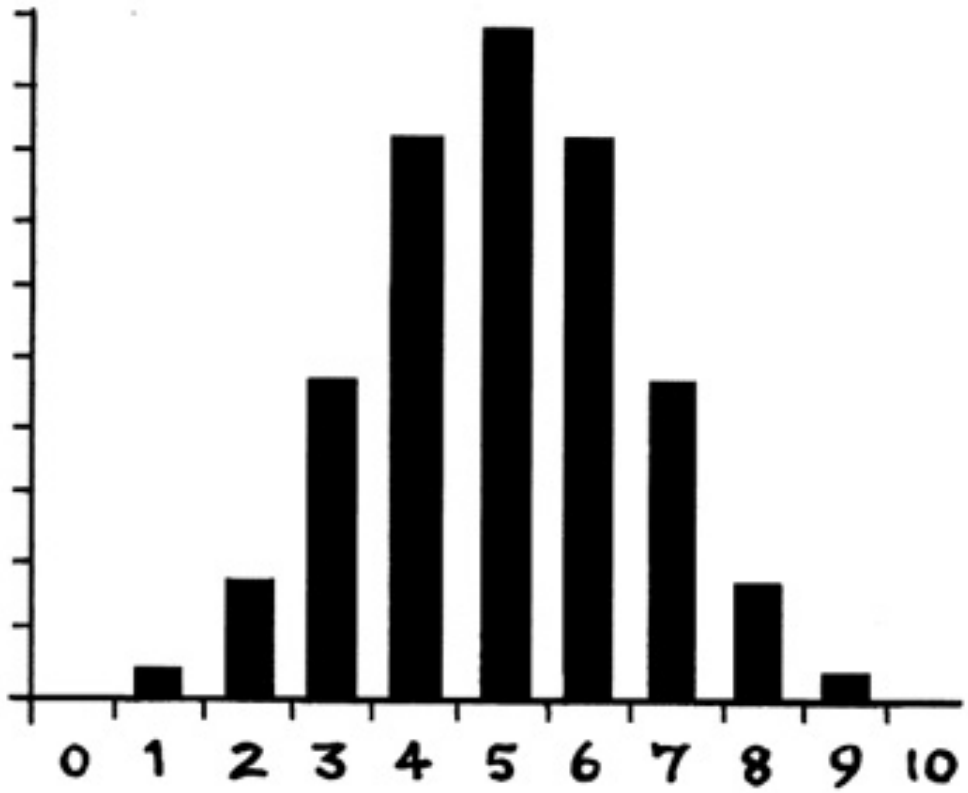
- If p=q=0.5

$$\begin{aligned}(p+q)^2 &= p^2+2pq+q^2 \\ &= \frac{1}{4}+\frac{2}{4}+\frac{1}{4}\end{aligned} \tag{1.5}$$

- If p=q=0.5

$$\begin{aligned}(p+q)^5 &= p^5+5p^4q+10p^3q^2+10p^2q^3+5pq^4+q^5 \\ &= \frac{1}{32}+\frac{5}{32}+\frac{10}{32}+\frac{10}{32}+\frac{5}{32}+\frac{1}{32}\end{aligned} \tag{1.6}$$

EXPANDED BINOMIAL DISTRIBUTION OF n=10



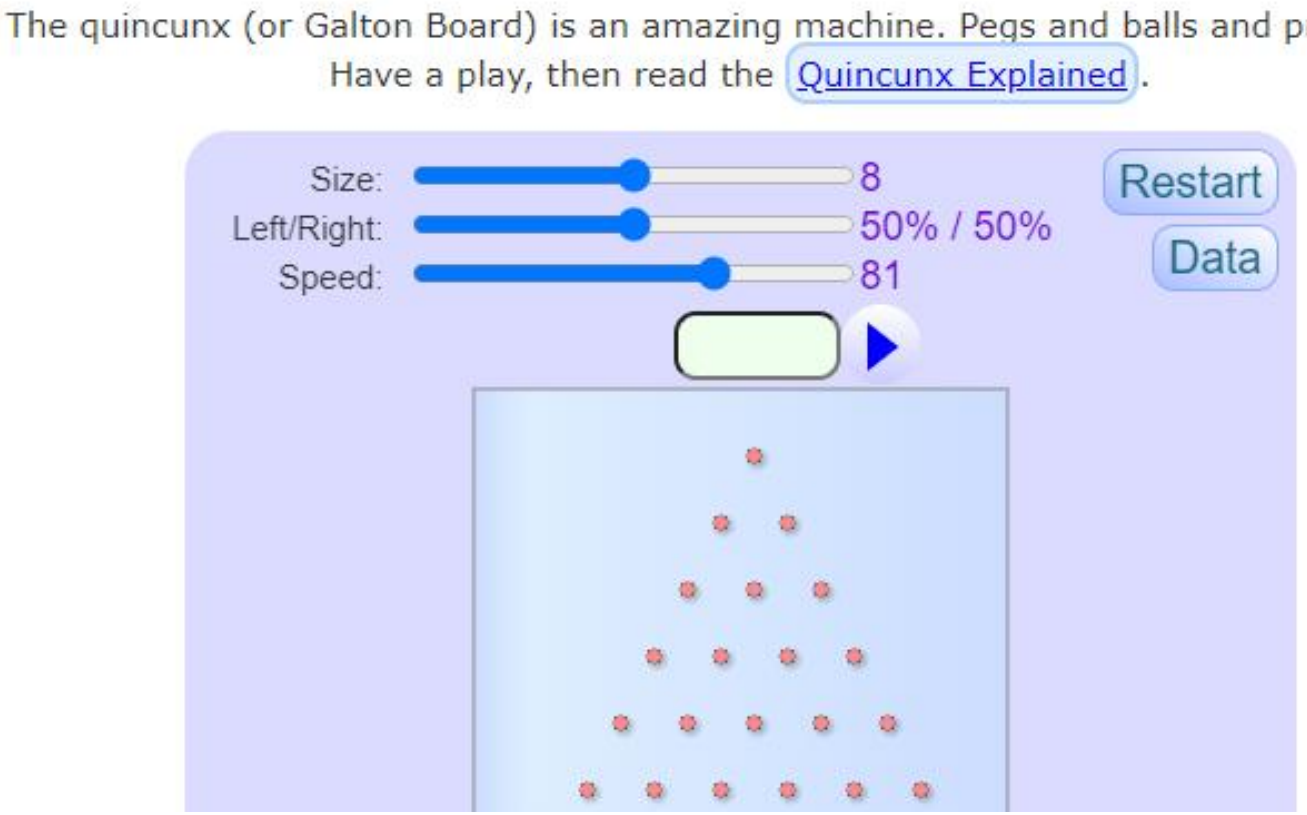


Figure 1.1: quincunx (or Galton Board)

https://www.mathsisfun.com/data/quincunx.html

统计学的核心问题和核心思路

- **核心问题：从样本向总体的过渡**

  用样本结果来推断总体结果。

Example 1.1

估计福州市2018年的人均月收入。总体=600万人。假定总体是得不到的。对总体的估计，需要通过能够代表总体的样本来得到。

比如，现在在福州随机抽取1000人，从中得到1000人的人均月收入为5000元。由此回答，福州（600万人）的人均月收入是多少？

统计学的核心问题和核心思路

- **核心问题：从样本向总体的过渡**
  用样本结果来推断总体结果。

Example 1.1

估计福州市2018年的人均月收入。总体=600万人。假定总体是得不到的。对总体的估计，需要通过能够代表总体的样本来得到。

比如，现在在福州随机抽取1000 人，从中得到1000人的人均月收入为5000元。由此回答，福州（600万人）的人均月收入是多少？

核心问题：从样本向总体的过渡

Example 1.2

估计一个培训班学生的平均年龄。总体=250人。假定总体是得不到的。对总体的估计，需要通过能够代表总体的样本来得到。

比如，现在在班级随机抽取25人，从中得到他们的平均年龄为27岁。由此回答，班级学生的平均年龄是多少？

在250名学生中抽取25人，共有？种可能的样本。

例子：估计湖中的鱼苗数目

Example 1.3

鱼塘中有若干条鱼苗，打捞起*50*条鱼苗，做上记号后再将其放回。

过一段时间后，重新打捞起*100*条鱼苗，发现其中有记号的鱼苗有*10*条。

请问，鱼塘中大概有多少条鱼苗？

例子：估计湖中的鱼苗数目

- 样本：100条；样本中有标记的10条。
- 总体：N=？；总体中有标记的50条。

大数定理（Law of Large Numbers）

- 当样本足够大时，样本均值将落在总体均值的附近。
- 大数定理的应用

CENTRAL LIMIT THEOREM

- It showed that the larger the sample size, the more closely the data will conform to the normal distribution.
- The **sampling distribution** of means gets closer and closer to the normal curve as the sample size increases, despite any departure from normality in the population distribution.
- The mathematical underpinnings of this theorem state that data which are influenced by a very large number of many small and unrelated random effects will be approximately normally distributed.

抽样分布

- 假设我们对总体进行重复抽样，每次用同样的公式计算样本统计量，那么从所有这些样本中得到的统计量就构成了一个分布，该分布被称为抽样分布。
- 对于一个总体而言，**所有可能样本**的统计量数值的概率分布被称作抽样分布。
  （A sampling distribution is a mathematical description of **all possible sampling event** outcomes and the probability of each one.）
- 根据中心极限定理，无论总体分布如何，只要样本规模足够大，统计量的抽样分布就接近正态分布。

SAMPLING DISTRIBUTION

- 对于一个总体而言，所有可能样本的统计量数值的概率分布被称作抽样分布。

Example 1.4

样本平均数的抽样分布是：

$$\bar{y} \sim N(\mu, \frac{\sigma_Y^2}{n})$$

[等式左边是样本统计量，右边是总体参数，这就搭建了样本与总体之间的桥梁]

样本统计量（平均值）服从以总体平均数μ为中心、方差为$\frac{\sigma_Y^2}{n}$的正态分布。σ_Y^2为总体的方差。

标准误 STANDARD ERROR

所有可能样本的（某个变量的）平均数的标准差为 $\sigma(\bar{y}) = \frac{\sigma_Y}{\sqrt{n}}$，也称为标准误（standard error, 简称SE）。

Example 1.5

举例：计算*SE*

sum cfps2012_age

mean cfps2012_age

display 16.87916/sqrt(35722)

中心极限定理

1. 所有可能样本平均数$\bar{y}$的分布是正态分布。
2. 所有可能样本平均数的平均数等于总体平均数。

$$E(\bar{y}) = \bar{\bar{y}} = \mu$$

3. 所有可能样本平均数的方差等于总体的方差除以样本规模。

INFERENTIAL STATISTICS

1. Hypothesis testing is a scientific procedure for making rational decisions about two different claims. （先假设总体的情况，然后进行抽样，检验原有的假设是否成立）
2. Estimation theory is a branch of statistics that deals with estimating the values of parameters. （先看样本情况，再问总体的情况）

假设检验 HYPOTHESIS TESTING

- **假设检验**

$$z = \frac{\bar{y}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

Example 1.6

零假设（*null hypothesis*）：总体的平均年龄μ=*24*岁
备择假设：μ不等于*24*岁
已知总体方差σ^2=*100*, 样本规模=*25*
如果我们的样本来自这个总体，那么它的平均数就应该在*24*岁附近，即落在一个标准差内的概率为*68%*，落在*1.96*个标准差内的概率为*95%*。
如果样本平均数小于*26*岁，*z=1*, 发生的概率为*68%*。如果样本平均数大于*28*岁，*z=2*, 即可能性不到*5%*。

NORMAL DISTRIBUTION

The area to the right of the mean is 50 percent.

1. The area within 1 standard deviation of the mean (that is, $\mu - \sigma < x < \mu + \sigma$) is approximately 68 percent.
2. The area within 2 standard deviation of the mean (that is, $\mu - 2\sigma < x < \mu + 2\sigma$) is approximately 95 percent.
3. The area within 3 standard deviation of the mean (that is, $\mu - 3\sigma < x < \mu + 3\sigma$) is approximately 99 percent.

假设检验 HYPOTHESIS TESTING

1. 在假定H_0成立的条件下，计算某个统计量的值，并确定它的概率分布。
2. 计算由样本得到的统计量的值所发生的概率，又称之为显著性水平（significance level），一般用α表示。
3. 若统计量的值所发生的概率低于我们事先设定的概率标准（如0.10、0.05和0.01），就说明统计显著，于是倾向于拒绝或否定原假设。

假设检验 HYPOTHESIS TESTING

1. z test
2. t test: 若不知道总体标准差，用样本标准差s代替总体方差，此时所有可能样本平均数的分布满足t分布，即t检验。

假设检验 HYPOTHESIS TESTING

在假定总体平均数和方差已知（即所有可能样本平均数的分布已知）的情况下，（在抽取之前）尽管我们不知道抽中的样本平均数会落在哪里，但我们知道它落在任何位置上的可能性有多大。

参考文献

- Magnello, Eileen and Borin Van Loon. 2009. Introducing Statistics: A Graphic Guide. London: Icon Books Ltd.