假设检验入门
与实验数据处理

大纲

一、假设检验从何而来

二、假设检验与其P值

三、几种经典假设检验

四、所用数据从何而来

假设检验从何而来

背景补充

从抛硬币开始...

- 假设有一个硬币，记抛出正面为1，抛出反面为0，那么抛十次，我们就能得到一个类似于这样的序列：
- $\{1, 0, 1, 1, 0, 1, 0, 0, 0, 1\}$
- 我们可以把十次当中第i次的结果记为x_i，则对于任意x_i，其取值是**不唯一**、**随机**的，我们就可以将这样的一种有不同可能的取值描述为一个“**随机变量**”，记作
- $X = \begin{cases} 0 \\ 1 \end{cases}$ 或 $X = \{0, 1\}$

从离散随机变量到连续随机变量

离散

- 抛硬币、年级等等
- 可能的取值“**可数**”：从一个结果出发，可以不重复地遍历所有可能
- 任意**两个值之间没有定义**：
  - 处于正面和反面之间？
  - 大二点五？

连续

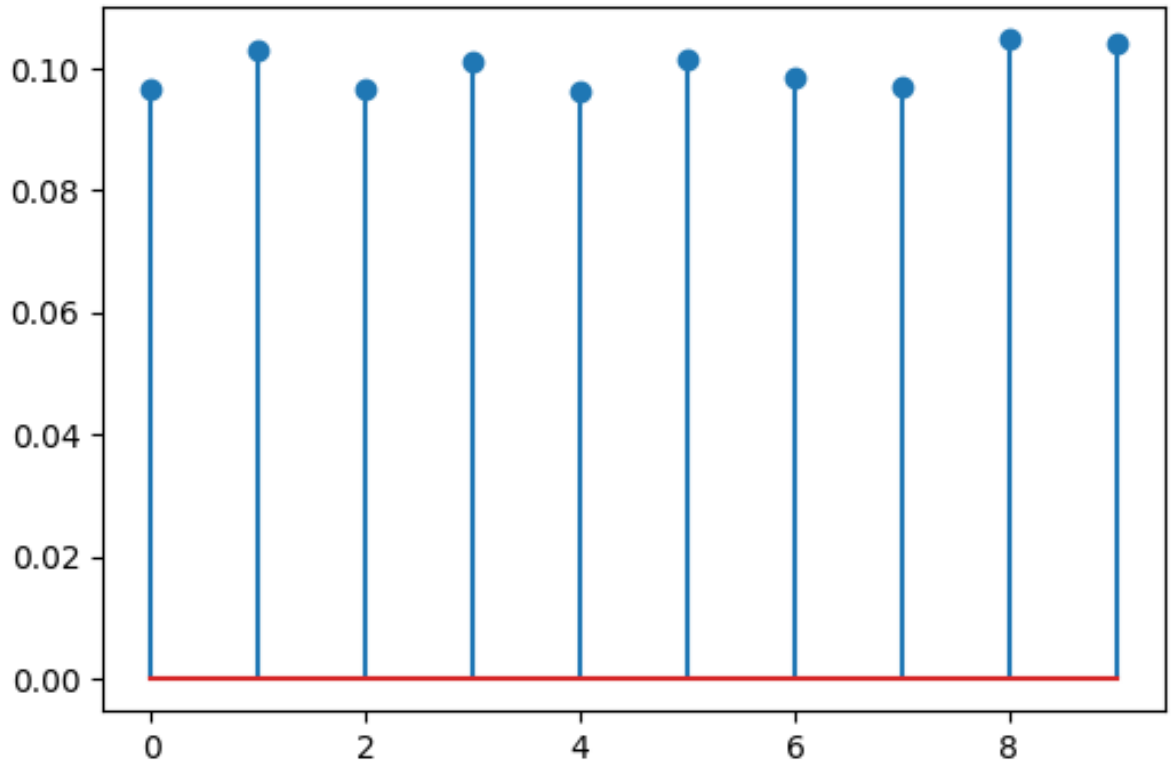
- 身高、绩点、气温等等
- 可能的取值“**不可数**”：没有一种方法能够不重不漏地遍历所有可能
- 任意**两个值之间可以无限细分**：
  - 0到1之间的实数与$-\infty$至$+\infty$一样多
  - 不存在身高完全一样的人

如何描述随机变量

取值 + 概率

- 概率的**总和为1**
- **概率分布**：描述随机变量各个取值的概率

离散随机变量的概率分布看起来如下

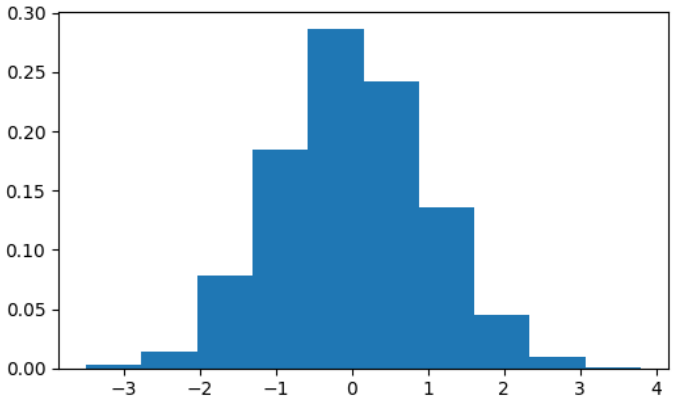


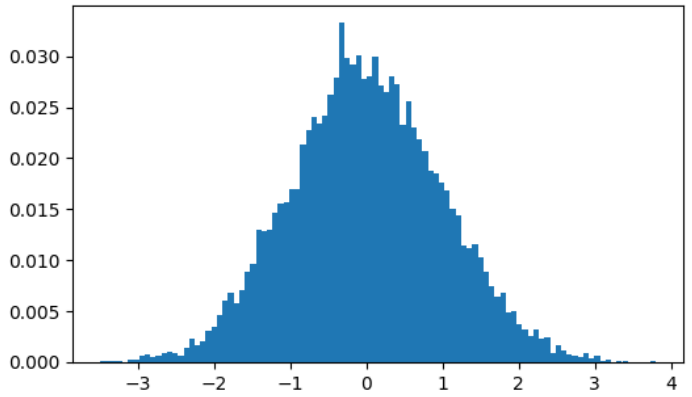
如何描述随机变量

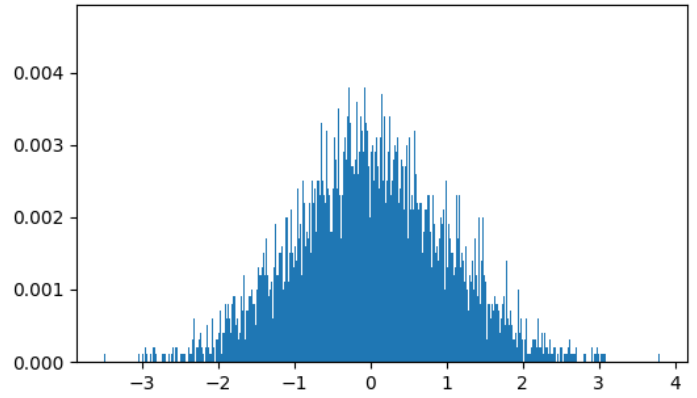
取值 + 概率

- **概率分布**：描述随机变量各个取值的概率
  - 但是对于连续型随机变量，由于其取值可以无限细分，因此相当于有无穷多的值可取，每个值的概率就→ 0

连续随机变量的概率分布随着取值的细分趋向于0





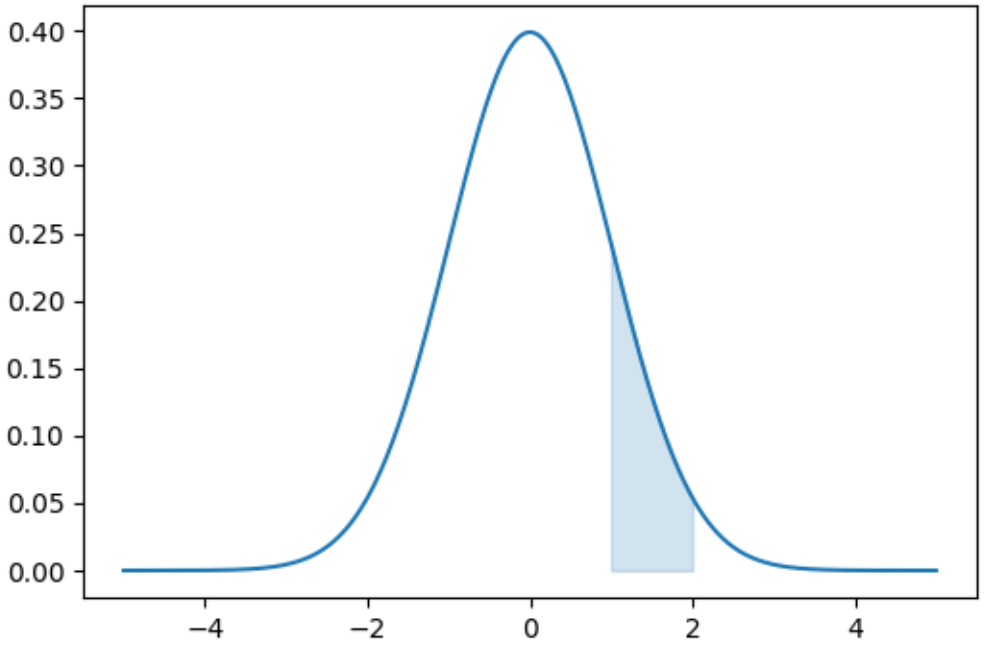


如何描述随机变量

取值 + 概率

- **概率分布**：描述随机变量各个取值的概率
  - 但是对于连续型随机变量，由于其取值可以无限细分，因此相当于有无穷多的值可取，每个值的概率就→ 0
  - 因此我们采用**概率密度分布**来描述其概率分布

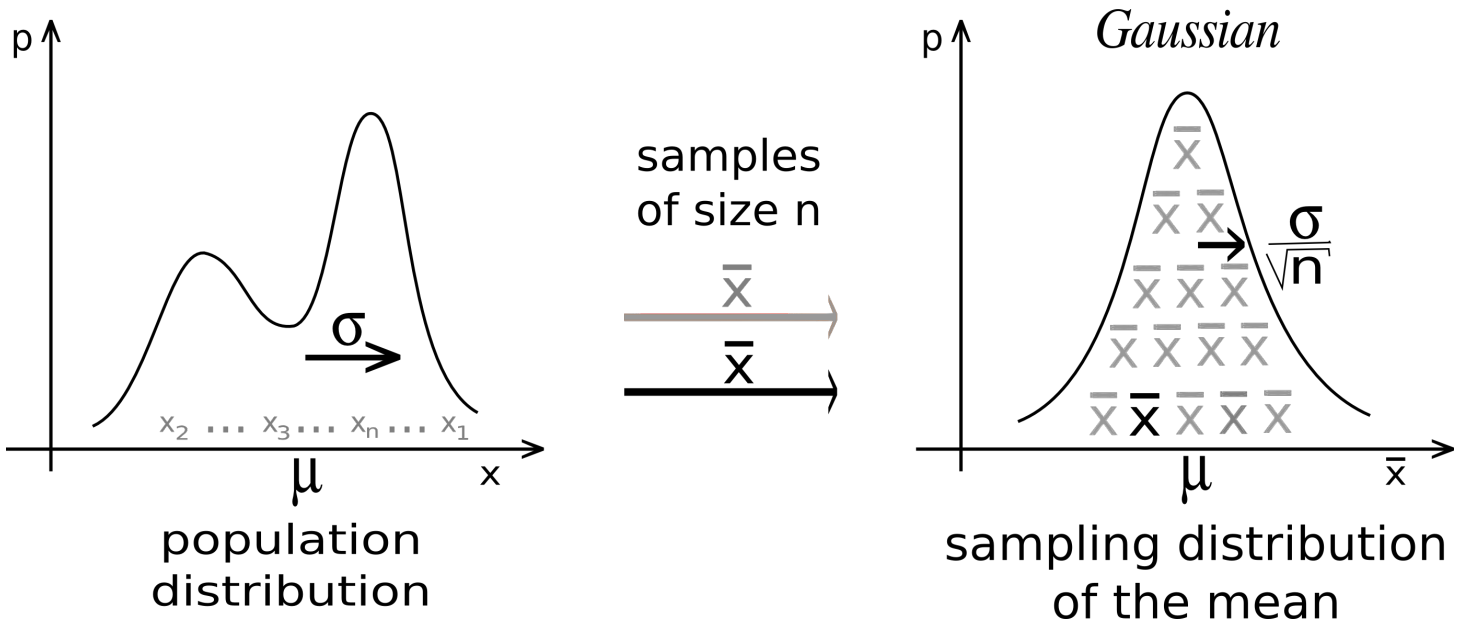
概率密度分布用**面积**来表示一段范围内取值的概率大小



正态分布——最常见的分布类型

又称为高斯分布、正规分布

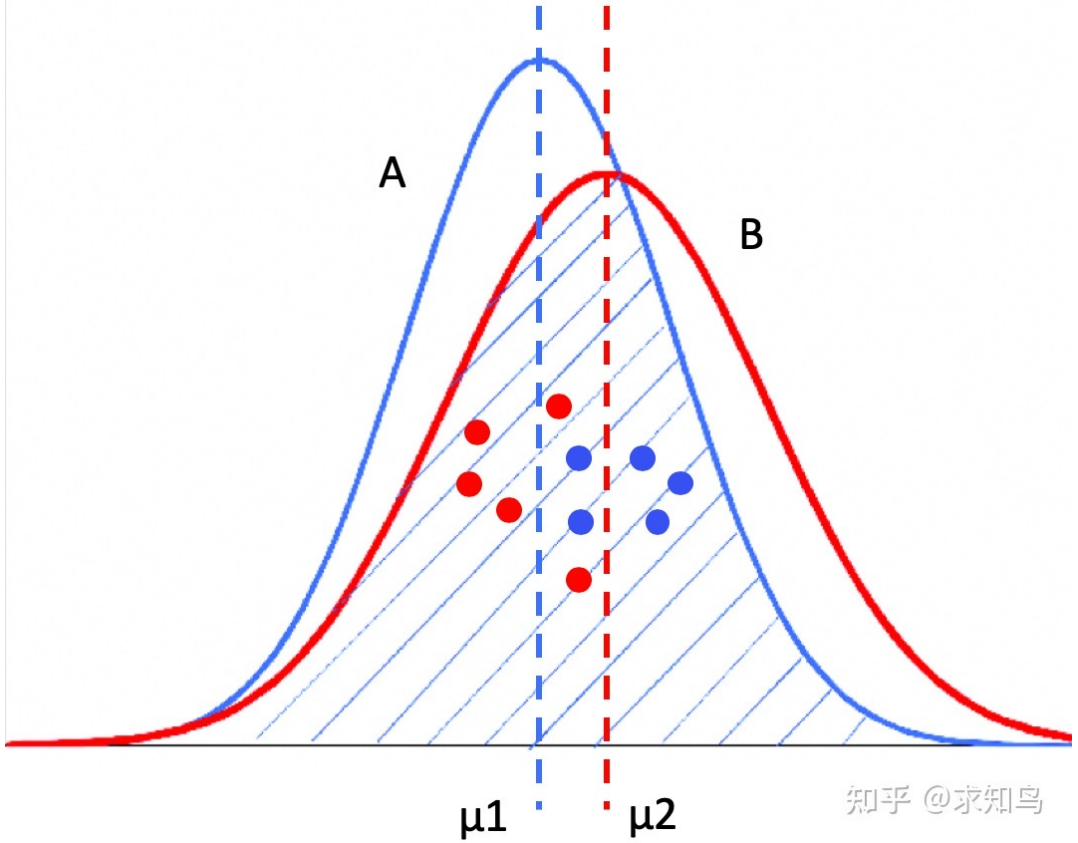
- 不管总体遵从什么类型的分布，依照中心极限定律，从中采样得到的**样本均值**依然**服从正态分布**



为什么需要假设检验

描述一个正态分布只需要均值和标准误两个参数

- 但是我们没法直接知道总体分布的这两个参数，只能从抽样所得结果中去“猜”
  - 一个样本点什么都猜不了
  - 两个样本点只能猜均值
  - 所以至少需要三个样本点
- 依概率抽样意味着存在得出**完全相反**结论的可能
  - 总体分布形如曲线，但是如果单从抽样结果出发我们就会得到“蓝色均值大于红色”这样的错误结论



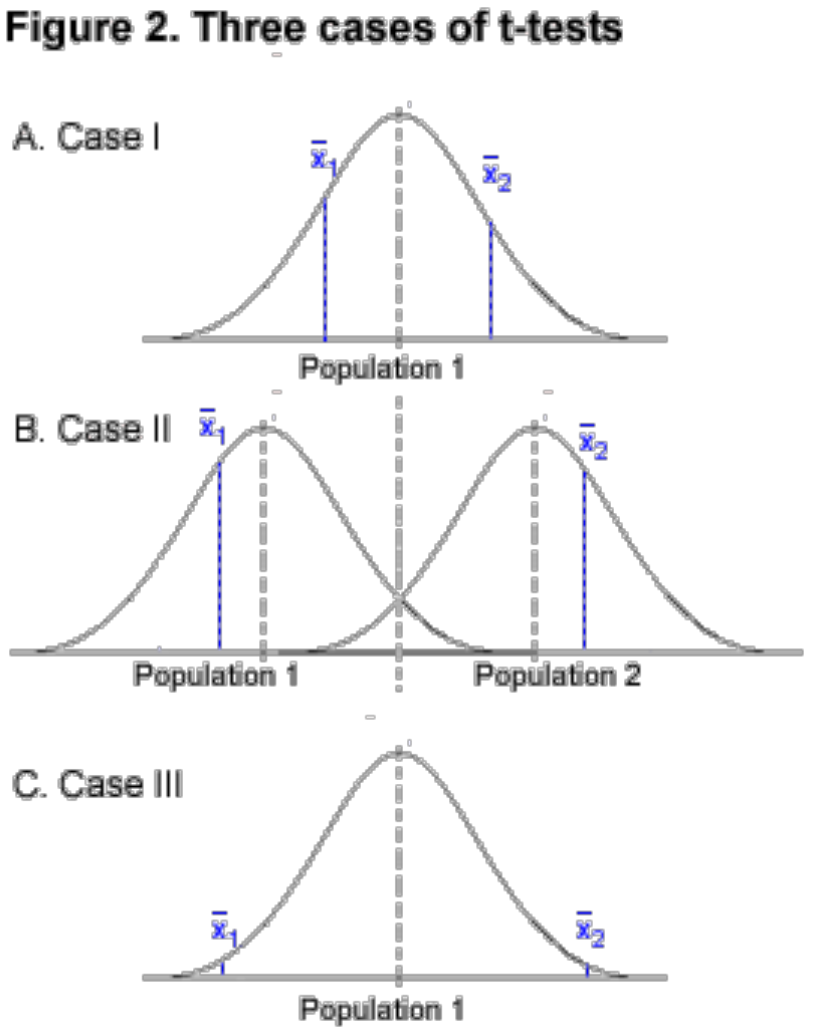
假设检验中的P值

重要指标

假设检验怎么做——从最简单的开始

t-test （t-检验）

- 根本问题：抽样结果是否**有差异**，即是否**服从同一分布**
- 我们的**零假设**：
  - 两组样本服从同一分布，没有差异（独立样本）
  - 样本服从均值为零的分布（单样本/成对样本）
  - 或者任何希望反驳的分布情况
- 我们的**备选假设**：
  - 两组样本服从不同分布，有差异（独立样本）
  - 样本服从均值非零的分布（单样本/成对样本）
  - 或者与零假设不同的、希望证明的分布情况



假设检验怎么做——怎么来“猜”

P值：在假定零假设成立的情况下，得到抽样结果的可能有多大

- 一个离散随机变量的例子：
- 一颗正六面体骰子，连续掷出了10个1，问其是均匀的可能性有多大？
  - 零假设：骰子是均匀的，即掷出1-6的可能性都是$1/6$
  - 备选假设：骰子不均匀，即掷出1-6的概率不相等
  - 计算：连续十次掷出1的概率为

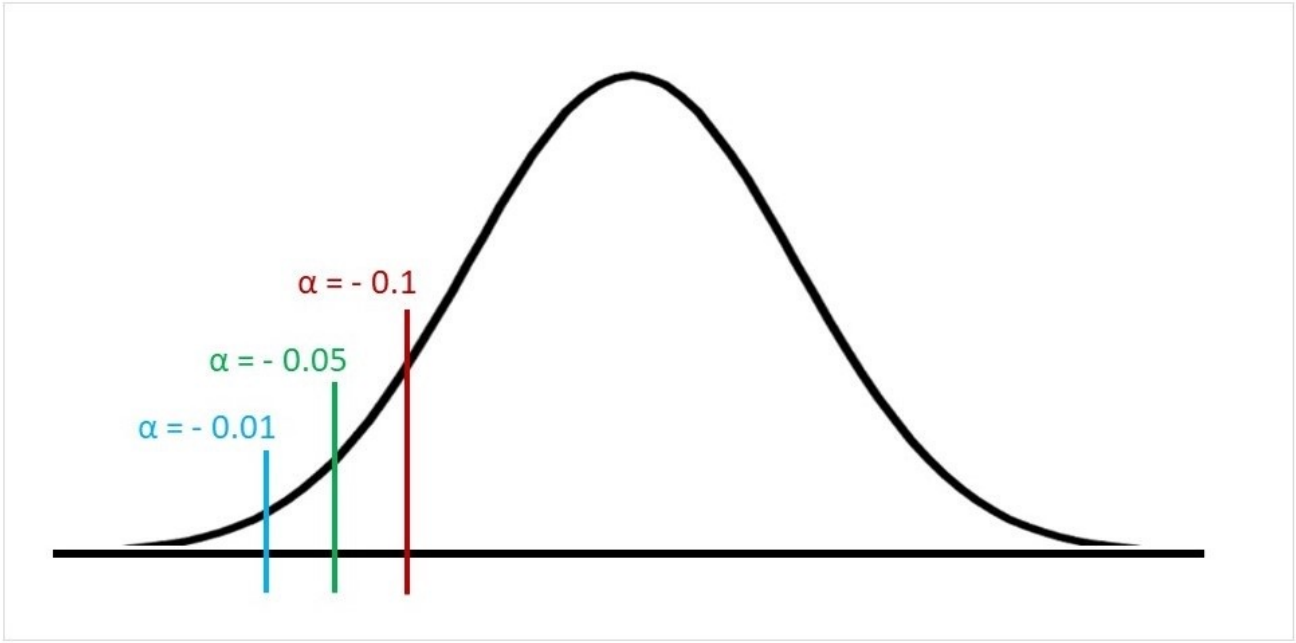
$$P(x_{1..10}=1)=\prod_{i=1}^{10}P(x_i=1)=\left(\frac{1}{6}\right)^{10}=\frac{1}{6^{10}}$$

  - 因此在这种情况下P值就是$\frac{1}{6^{10}}$，也就是说当我们拒绝零假设时，只有$\frac{1}{6^{10}}$的概率犯错

假设检验怎么做——何时拒绝零假设

为了控制“误杀”零假设的犯错概率，我们需要决定显著性水平

- 显著性水平为**预先决定**的拒绝零假设时的P值的**阈值**
- 低于显著性水平=拒绝零假设
- 显然所取**样本越多，P值越小**，因此要选择合适的显著性水平
- 常见的几个显著性水平有：
  - 0.1
  - 0.05
  - 0.01
  - …



Carsten Grube. Statistical Data Analysis. dataZ4s.com

几种经典假设检验

方法选择

选择合适的假设检验——单样本、双样本

t-test (t-检验)

- 单样本
  - 生物化学课程成绩，etc.（标准误由样本方差估计）
  - 零假设：服从的分布均值为0，与其没有差异
  - 备选假设：服从的分布均值不为0
- 成对样本
  - 参加学业辅导前后的成绩变化，etc.
  - 将成对的观测值通过相减/相除化为单样本，假设也同单样本
- 独立样本
  - 参加学业辅导与不参加学业辅导的人群成绩对比，etc.（参数由样本估计）
  - 零假设：两个变量服从同一个分布，没有差异
  - 备选假设：两个变量服从不同分布

选择合适的假设检验——多样本

ANOVA（Analysis of Variance，方差分析）

- 单因素方差分析 One-way ANOVA
  - 两个以上实验分组（pH，研磨方法，etc.）
  - 零假设：所有分布的均值没有差异，即实验分组对观测结果无影响
  - 备选假设：至少有一个分组服从不同的分布
- 双因素方差分析 Two-way ANOVA
  - 有两组互相独立的实验分组（pH+研磨方法，etc.）
  - 零假设：任一分组对应分布均值无差异，且两个实验分组互不影响（2+1）
  - 备选假设：有至少一个分组对结果造成影响，或两个分组间有关联
- 具体过程与t-检验不太相同，但结果显著性都以P值衡量

选择合适的假设检验——其他方法

方法很多，熟悉原理，仔细选择

- 卡方检验
  - 类似于ANOVA，但是观测值不连续，无法求方差，只能依照其分类比例计算
- 回归分析
  - 探究两个连续随机变量之间是否有关联性
- 所有假设检验都有共同点，需要视具体情况分析！

有差异 = 显著 ≠ 差异大

不能反驳零假设 = 不显著 ≠ 接受零假设

所用数据从何而来

数据处理

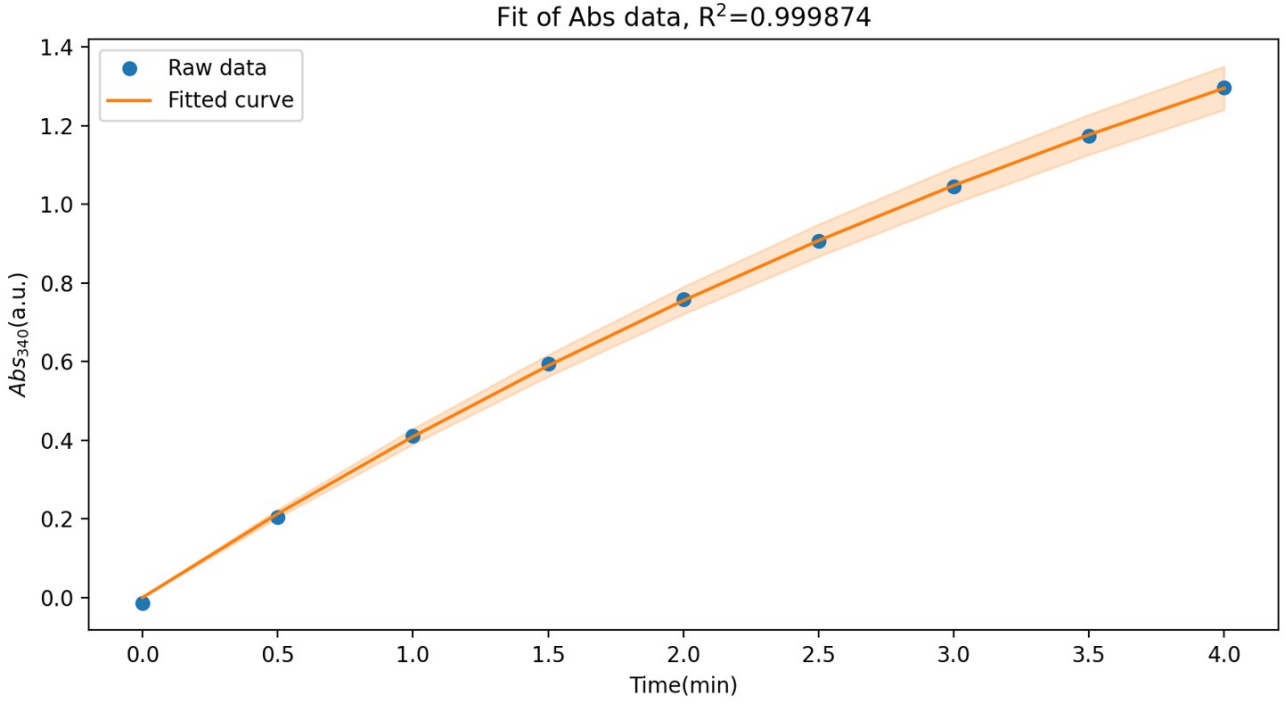
要做分析先要找好对象

探究研磨方式/pH对酶活的影响

- 实验变量：研磨方式或pH
- 研究对象：单位酶活
- 原始数据：吸光度值-时间曲线
  - 首先从吸光度值-时间曲线计算出单位酶活（假设所有人用量均为$3\mu L$）
  - 所用模型为

$$Abs_{340} = (1 - e^{-K_{tot} \cdot t}) S_0 \cdot \varepsilon \cdot L$$

  - 推导过程可以参考提交数据的网页上侧边栏，待定项为总酶活K_{tot}与起始(有效)底物浓度S_0（可以自行比较如果固定S_0其拟合结果如何）

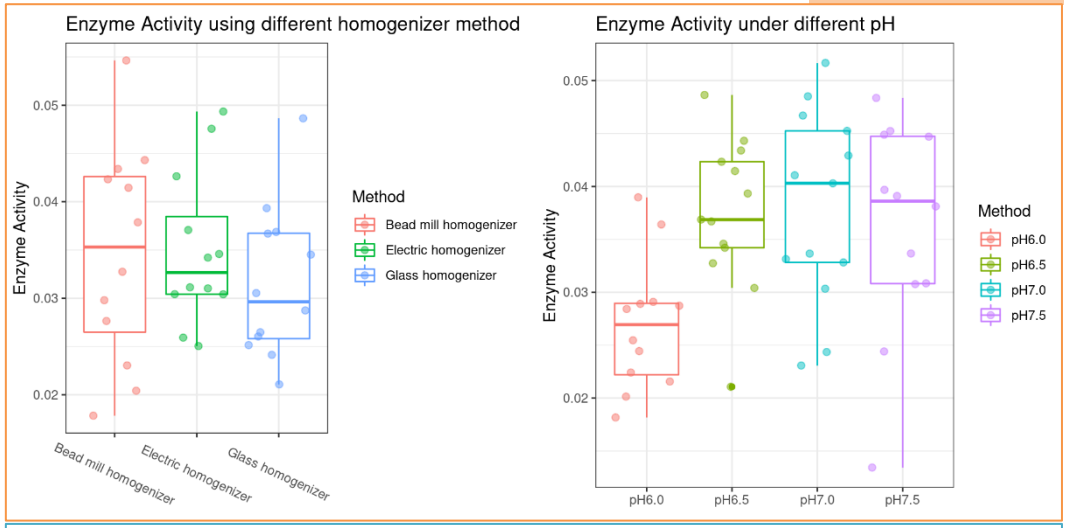


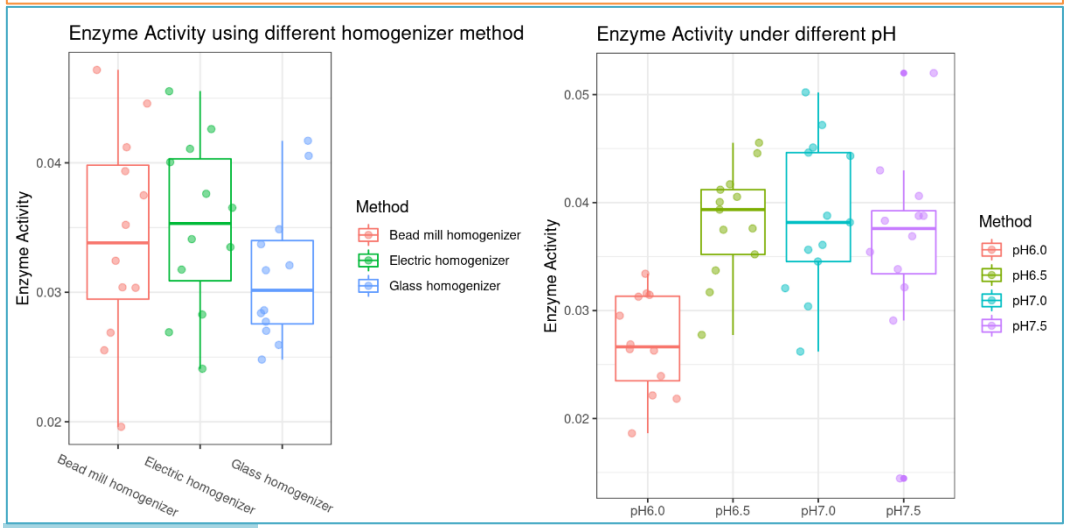
进一步处理——归一化

不同组之间可能体系、样品量有差异

- 我们希望减少这部分差异带来的实验性因素的影响，同时也要去除重复数值的偏置影响（去重）
- 假设酶活只随处理的不同而变化
- 将不重复的每组结果取均值，计算此均值与全班均值的偏差比例，再将对应数据点除去此比例，使得组间酶活的均值位于同一水平

归一化前





归一化后

给同学们的数据形式

纵表

- 每行对应一个数值
- 数值之后有相关分组信息
- 允许嵌套分组/任意增加记录字段

横表

- 每行有多个数值
- 一列为一组数值
- 较难嵌套分组

	Method	Enzyme_Activity_tot	Enzyme_Activity_avg	Enzyme_Activity_avg_norm	S_estimate	R_squared	Method_Group	Day	Group	Name	Id
1	Electric homogenizer	0.111193709922668	0.0370645699742228	0.0317685672646368	0.262011043060187	0.999669189063822	0	Fri	1	8d51340a	2000012136
2	Glass homogenizer	0.145928928868958	0.0486429762896526	0.0416925831133258	0.23600097745046	0.998707113503391	0	Fri	1	8d51340a	2000012136
3	Bead mill homogenizer	0.0893881697813553	0.0297960565937851	0.0255386216211035	0.241228434356425	0.999878228596774	0	Fri	1	8d51340a	2000012136
4	pH6.0	0.076401538195408	0.0254671793984693	0.0218282797378607	0.205830670446861	0.999573448302901	1	Fri	1	8d51340a	2000012136
5	pH6.5	0.145928928868958	0.0486429762896526	0.0416925831133258	0.23600097745046	0.998707113503391	1	Fri	1	8d51340a	2000012136
6	pH7.0	0.120897518364911	0.0402991727883037	0.0345409910953997	0.276036913334945	0.99959597289406	1	Fri	1	8d51340a	2000012136
7	pH7.5	0.134105088978416	0.0447016963261388	0.0383144563006653	0.261370596085891	0.999888511431921	1	Fri	1	8d51340a	2000012136
8	Electric homogenizer	0.127905471598561	0.0426351571995202	0.0365431963806219	0.276173632063776	0.999874056650278	0	Fri	1	91093c91	2000012278
9	Glass homogenizer	0.117996361484802	0.0393321204949339	0.0337121168941961	0.234946210192544	0.999852897194734	0	Fri	1	91093c91	2000012278
10	Bead mill homogenizer	0.113581888950824	0.0378606296502746	0.0324508812745634	0.210053776627696	0.999859874365636	0	Fri	1	91093c91	2000012278

Showing 1 to 10 of 86 entries

Previous 1 2 3 4 5 … 9 Next

欢迎自行探索其他数据处理方法

方法	总酶活	归一化前单位酶活	归一化后单位酶活	拟合底物浓度	R^2	方法分组	其他信息			
	Method	Enzyme_Activity_tot	Enzyme_Activity_avg	Enzyme_Activity_avg_norm	S_estimate	R_squared	Method_Group	Day	Group	Name	Id
1	Electric homogenizer	0.111193709922668	0.0370645699742228	0.0317685672646368	0.262011043060187	0.999669189063822	0	Fri	1	8d51340a	2000012136
2	Glass homogenizer	0.145928928868958	0.0486429762896526	0.0416925831133258	0.23600097745046	0.998707113503391	0	Fri	1	8d51340a	2000012136
3	Bead mill homogenizer	0.0893881697813553	0.0297960565937851	0.0255386216211035	0.241228434356425	0.999878228596774	0	Fri	1	8d51340a	2000012136
4	pH6.0	0.076401538195408	0.0254671793984693	0.0218282797378607	0.205830670446861	0.999573448302901	1	Fri	1	8d51340a	2000012136
5	pH6.5	0.145928928868958	0.0486429762896526	0.0416925831133258	0.23600097745046	0.998707113503391	1	Fri	1	8d51340a	2000012136
6	pH7.0	0.120897518364911	0.0402991727883037	0.0345409910953997	0.276036913334945	0.99959597289406	1	Fri	1	8d51340a	2000012136
7	pH7.5	0.134105088978416	0.0447016963261388	0.0383144563006653	0.261370596085891	0.999888511431921	1	Fri	1	8d51340a	2000012136
8	Electric homogenizer	0.127905471598561	0.0426351571995202	0.0365431963806219	0.276173632063776	0.999874056650278	0	Fri	1	91093c91	2000012278
9	Glass homogenizer	0.117996361484802	0.0393321204949339	0.0337121168941961	0.234946210192544	0.999852897194734	0	Fri	1	91093c91	2000012278
10	Bead mill homogenizer	0.113581888950824	0.0378606296502746	0.0324508812745634	0.210053776627696	0.999859874365636	0	Fri	1	91093c91	2000012278

Showing 1 to 10 of 86 entries

Previous 1 2 3 4 5 … 9 Next

谢谢