北京航空航天大學
BEIHANG UNIVERSITY

Lattice Parton Collaboration

第七届格点量子场论暑期学校

数据分析基础 (一)

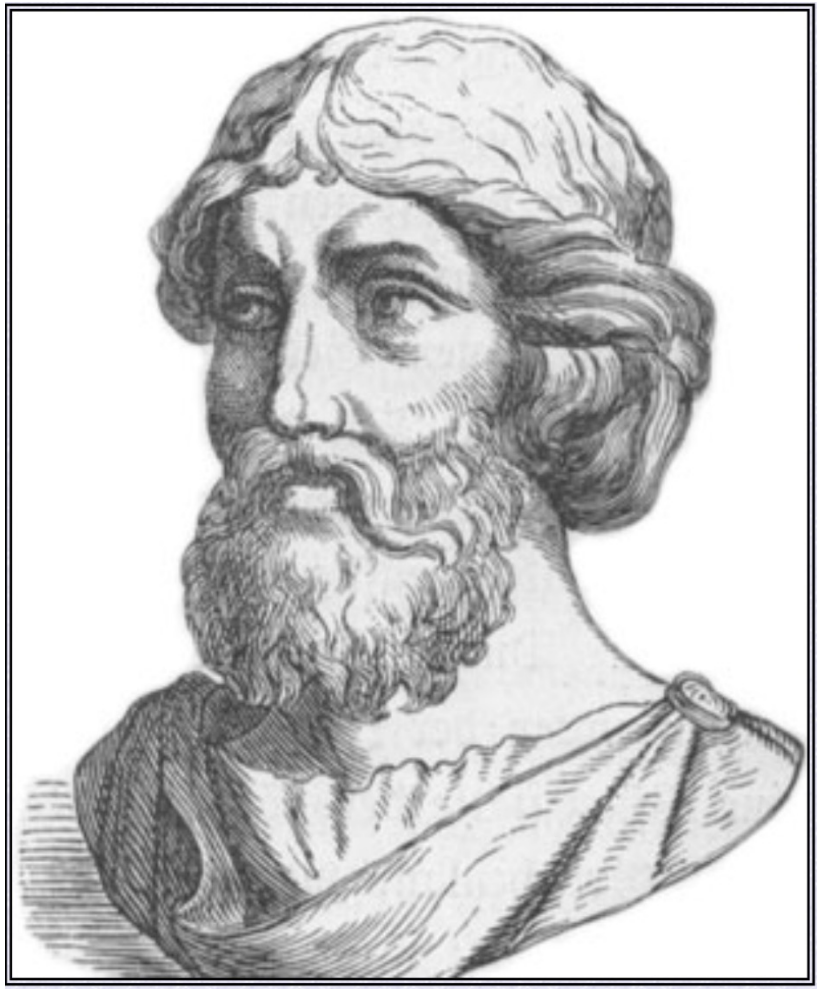
张其安

北京航空航天大学

Aug. 12, 2024

OUTLINE

- 数据分析的数学基础
  - 采样与统计
  - 重采样
  - 误差与误差传递
  - 数据拟合
- 格点谱学数据分析实例：从两点关联函数到强子谱



All things are number

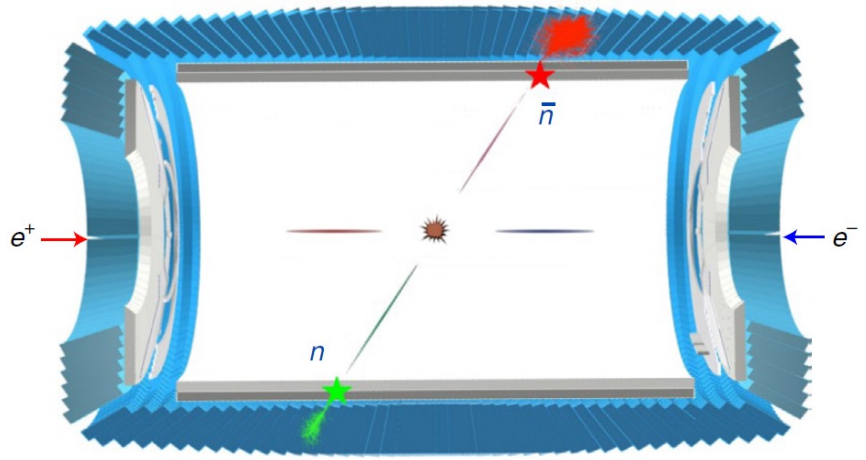
万物皆数

毕达哥拉斯（570BC - 495BC）

➢ 物理学是一门实验科学，高能粒子物理更是如此……

实验测量出每个事例的特征量(能动量，末态粒子数…)

理论预言出上述各特征量的分布，而且可能还会包含某些如相互作用耦合常数等自由参数。



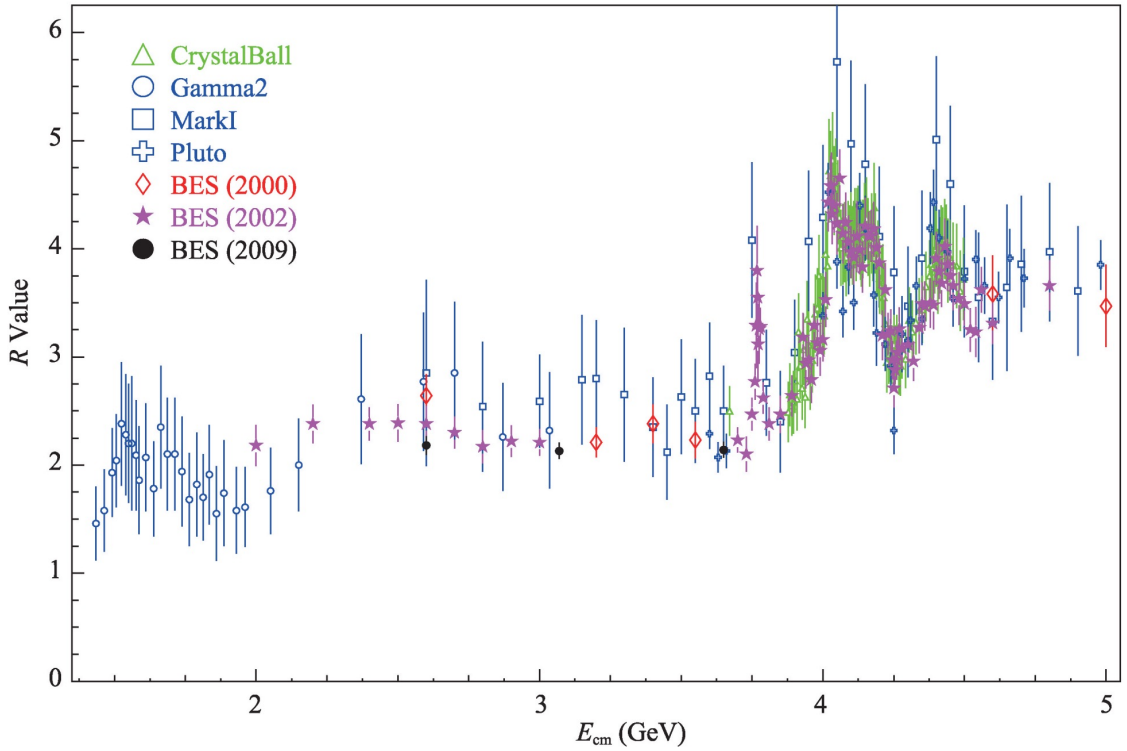
收集数据
统计分析

→

估计参数值与相应的误差范围，检验在何种程度上理论与实验数据相符。

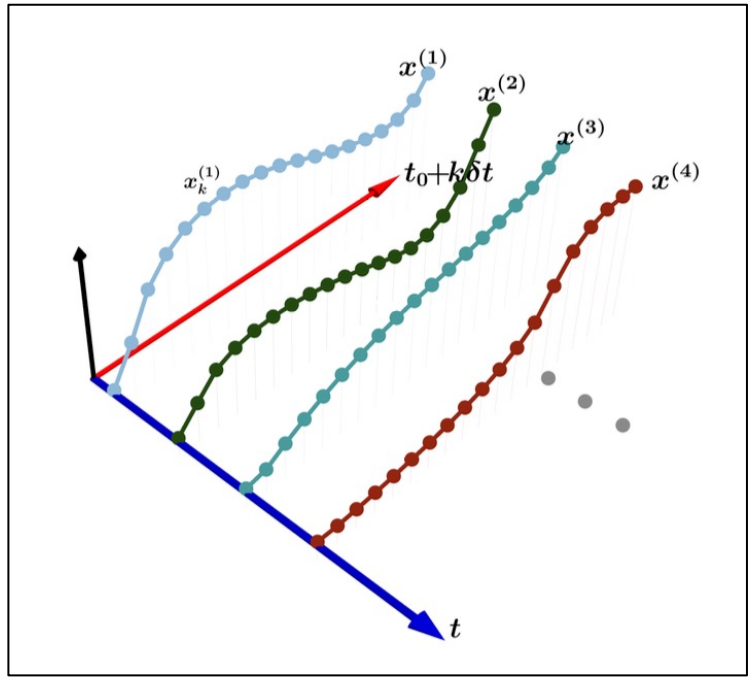
问题：如何统计实验测量的一个个数据点？如何将实验数据与理论模型联系起来？如何通过参数评价理论与实验的相符程度？

➢ 人们通常无法获取能够完整描述一个过程的实验数据



北京谱仪上的R值测量（5GeV以下）

选择一系列能量点进行取数



格点组态的生成

满足QCD作用量的一系列路径

……

采样

- 假设你是钢丝绳的工厂的质检员，需要在生产中检测产品的质量（断裂强度）。

总体 (Population)：工厂生产的所有钢丝绳（无穷）

样本 (Samples)：从中随机抽取n根钢丝绳进行检验，记做$(\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots\ldots, \boldsymbol{x_n})$，对它们做m次测试，得到结果：

$$\begin{aligned}
\boldsymbol{x}^{(1)} &= (x_1^{(1)}, x_2^{(1)}, \ldots, x_n^{(1)}), \\
\boldsymbol{x}^{(2)} &= (x_1^{(2)}, x_2^{(2)}, \ldots, x_n^{(2)}), \\
&\ldots\ldots \\
\boldsymbol{x}^{(m)} &= (x_1^{(m)}, x_2^{(m)}, \ldots, x_n^{(m)}),
\end{aligned}$$

通常我们认为同一根钢丝上的m次测量是关联的，对它求平均来得到这一根钢丝的断裂强度。这样我们可以得到描述独立的n根钢丝绳的测试结果的样本：

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_n) = \frac{1}{m}\sum_{j=1}^{m} (x_1^{(j)}, x_2^{(j)}, \ldots, x_n^{(j)}).$$

假设总体平均值(population mean)为$\hat{x}$，这是我们希望得到但永远也不可能得到的。可以通过样本平均值(sample mean)$\bar{x}$（或写作E(x)）来进行估计：

$$\langle \bar{x} \rangle = \sum_{i=1}^{n} \frac{x_i}{n} \xrightarrow{n \to \infty} \hat{x}.$$

同样，总体方差(population variance)$\sigma^2(x)$也是不可知的，它与样本方差(sample variance)$\sigma^2(\bar{x})$的关系是：

$$\sigma^2(\bar{x}) = \langle (\bar{x} - \langle \bar{x} \rangle)^2 \rangle = \frac{1}{n} \sigma^2(x),$$

其中总体方差$\sigma^2(x)$的定义为：

$$\sigma^2(x) = \frac{1}{n} \left\langle \sum_{i=1}^{n} (x_i - \hat{x})^2 \right\rangle$$

这样我们就可以通过有限次采样来估计生产的总体的质量。

如果随着时间增长，设备老化导致的产品质量也会发生变化。为了描述增加的这一个变量的影响，我们考虑协方差。

协方差用于衡量两个随机变量联合变化的情况，它描述两个随机变量间关系强度和方向的统计量，通过衡量两个变量联合变化的情况，帮助我们理解数据间的关联性。

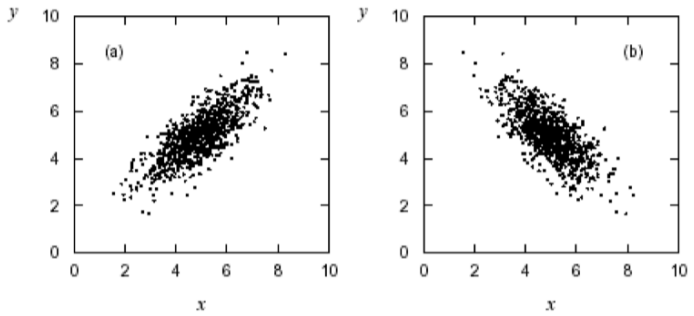
$$\begin{aligned} \operatorname{cov}(x, y) &= \overline{(x-\bar{x})(y-\bar{y})} \\ &= \overline{xy} - \bar{x}\,\bar{y} \end{aligned}$$

如果x和y相互独立，则$\mathbf{cov}(\boldsymbol{x}, \boldsymbol{y}) = \mathbf{0}$. 否则说明x和y之间存在关联。

定义相关系数：

$$\rho_{xy} = \frac{\operatorname{cov}(x, y)}{\sigma(x)\sigma(y)} \qquad -1 \leqslant \rho_{xy} \leqslant 1$$

$\rho = 0.75$



$\rho = -0.75$

$\rho = 0.95$

$\rho = 0.25$

-1代表完全负相关，+1代表完全正相关，0则表示不相关。

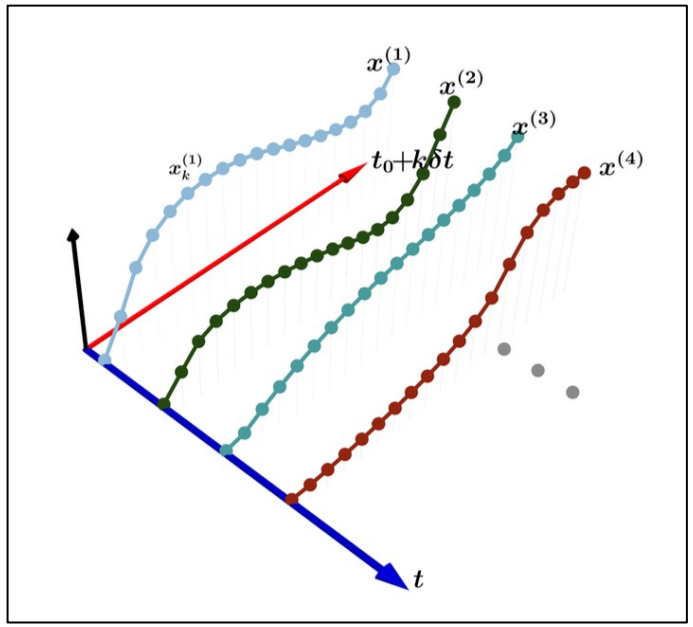
工厂生产的所有钢丝绳 ➡ 路径积分表述下满足QCD作用量的所有可能的路径

采样用来检测的钢丝绳 ➡ 在格点上产生的规范场组态，可以看作是对路径积分中所有可能路径的采样

在钢丝绳上检测断裂强度 ➡ 在组态上设计算法计算关联函数

- 通过在有限多的组态上测量得到的关联函数的信息，可以得到路径积分下关联函数的数值结果。

$$\langle\Omega|\,T\phi_H(x_1)\phi_H(x_2)\,|\Omega\rangle = \lim_{T\to\infty(1-i\epsilon)} \frac{\int \mathcal{D}\phi\,\phi(x_1)\phi(x_2)\exp\left[i\int_{-T}^{T} d^4x\,\mathcal{L}\right]}{\int \mathcal{D}\phi\,\exp\left[i\int_{-T}^{T} d^4x\,\mathcal{L}\right]}.$$

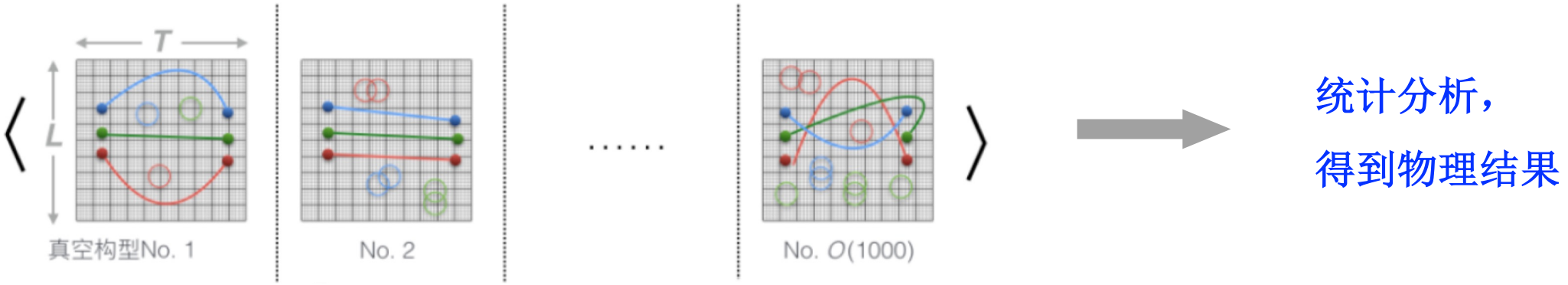


➢ 路径积分表述下的关联函数：

$$\langle \mathcal{O}(U,q,\bar{q}) \rangle = \frac{\int [\mathcal{D}U] e^{-S^{\text{latt}}_{\text{glue}}} \prod_q \det\left(D^{\text{latt}}_{\mu}\gamma_\mu + am_q\right) \tilde{\mathcal{O}}(U)}{\int [\mathcal{D}U] e^{-S^{\text{latt}}_{\text{glue}}} \prod_q \det\left(D^{\text{latt}}_{\mu}\gamma_\mu + am_q\right)}$$

规范场组态　费米子传播子　流结构

在每个组态上重复计算并统计平均，就可以计算得到完整QCD路径积分的统计平均值和统计误差。



很多时候，只知道平均值和方差是远远不够的，我们希望知道数据的更多信息，

并希望能够利用这些数据做一系列后续分析……

➢ 假如有一组样本x满足正态分布，中心值为1，标准差等于0.5. 问：$\log x$的中心值和标准差是多少？

样本x

```
1.7434 0.3460 0.5621 0.9716 1.1384
0.9272 1.5119 1.2523 1.4899 1.3410
1.8583 1.1405 1.0252 0.8401 1.4321
1.1676 1.6836 1.4797 0.8726 0.3985
1.3597 0.7728 0.0932 0.3319 1.3032
0.7917 1.2121 1.1983 1.0782 0.6962
1.0600 -0.1333 1.2507 0.6059 0.5737
1.3577 1.6586 1.6925 1.1836 0.4975
0.8620 0.7378 1.0138 0.5564 0.8839
2.0981 0.5269 0.1714 1.0816 0.9179
```

$\log x$

```
0.5558 -1.0613 -0.5761 -0.0288 0.1296
-0.0756 0.4134 0.2250 0.3987 0.2934
0.6196 0.1315 0.0249 -0.1742 0.3592
0.1549 0.5209 0.3919 -0.1362 -0.9201
0.3072 -0.2578 -2.3726 -1.1031 0.2648
-0.2335 0.1923 0.1809 0.0753 -0.3621
0.0582 nan 0.2237 -0.5010 -0.5557
0.3058 0.5060 0.5262 0.1686 -0.6982
-0.1485 -0.3041 0.0137 -0.5863 -0.1234
0.7410 -0.6408 -1.7637 0.0785 -0.0857
```

- 样本x中存在一个小于0的值，它不会影响对x的线性操作，但是当遇到一些非线性操作时会出问题.
- 在上面的样本中，每一次采样约有4.56%的概率得到负数（$2-\sigma$）.

这种情况应该如何处理？

重采样 (Resampling)

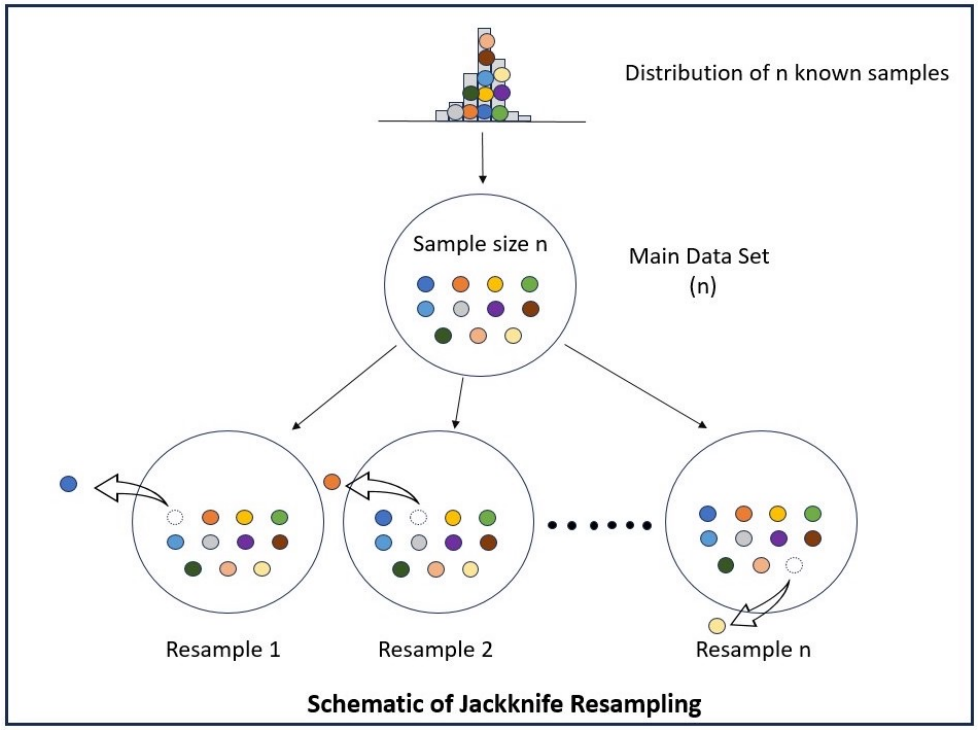
在前面的问题中，如果我们能有办法把x的分布收窄（标准差减小），这样求log就不容易遇到问题了。然后求完log以后我们再想办法恢复回原来的宽度。

在统计学中，重采样是一种通过对已有的样本进行重新组合而产生一系列新样本的方法。注意它不会创造新的样本，而是对原来样本的某种变换。

常用的重采样方法有两种：**Jackknife resampling**和**Bootstrap resampling**

➢ Jackknife resampling

在原样本$\boldsymbol{x}$中依次剔除一个观测值，然后剩余的观测值组成一个新的样本$\boldsymbol{x}_{(i)}$。此时所有的$\bar{\boldsymbol{x}}_{(i)}$组成一个新的样本$\boldsymbol{x}_{\text{jack}}$。



- 原样本$\boldsymbol{x}$和新样本$\boldsymbol{x}_{\text{jack}}$的平均值是相同的：

$$\text{原样本：}\quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{i\in[n]} x_i,$$

$$\text{新样本：}\quad \bar{x}_{(i)} = \frac{1}{n-1}\sum_{j\in[n],j\neq i} x_j, \qquad \bar{x}_{\text{jack}} = \frac{1}{n}\sum_{i=1}^{n}\bar{x}_{(i)}.$$

- 新样本$\boldsymbol{x}_{\text{jack}}$的方差被收窄了n(n−1)倍：

$$\sigma^2(\bar{x}_{\text{jack}}) = \frac{n-1}{n}\sum_{i=1}^{n}(\bar{x}_i - \bar{x}_{\text{jack}})^2 = \frac{1}{n(n-1)}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{1}{n(n-1)}\sigma^2(x)$$

样本x

```
1.7434 0.3460 0.5621 0.9716 1.1384
0.9272 1.5119 1.2523 1.4899 1.3410
1.8583 1.1405 1.0252 0.8401 1.4321
1.1676 1.6836 1.4797 0.8726 0.3985
1.3597 0.7728 0.0932 0.3319 1.3032
0.7917 1.2121 1.1983 1.0782 0.6962
1.0600 -0.1333 1.2507 0.6059 0.5737
1.3577 1.6586 1.6925 1.1836 0.4975
0.8620 0.7378 1.0138 0.5564 0.8839
2.0981 0.5269 0.1714 1.0816 0.9179
```

$\log x$

```
0.5558 -1.0613 -0.5761 -0.0288 0.1296
-0.0756 0.4134 0.2250 0.3987 0.2934
0.6196 0.1315 0.0249 -0.1742 0.3592
0.1549 0.5209 0.3919 -0.1362 -0.9201
0.3072 -0.2578 -2.3726 -1.1031 0.2648
-0.2335 0.1923 0.1809 0.0753 -0.3621
0.0582 nan 0.2237 -0.5010 -0.5557
0.3058 0.5060 0.5262 0.1686 -0.6982
-0.1485 -0.3041 0.0137 -0.5863 -0.1234
0.7410 -0.6408 -1.7637 0.0785 -0.0857
```

$\bar{x} = \text{nan}$

$\sigma(x) = \text{nan}$

Jackknife Resampling

```
0.9974 1.0259 1.0215 0.0131 0.0097
1.0141 1.0021 1.0074 0.0026 0.0056
0.9951 1.0097 1.0121 0.0157 0.0037
1.0091 0.9986 1.0028 0.0151 0.0245
1.0052 1.0172 1.0311 0.0259 0.0064
1.0168 1.0082 1.0085 0.0109 0.0186
1.0113 1.0357 1.0075 0.0204 0.0210
1.0053 0.9991 0.9984 0.0088 0.0226
1.0154 1.0179 1.0123 0.0214 0.0148
0.9902 1.0222 1.0295 0.0108 0.0141
```

$\log x_{\text{jack}}$

```
-0.0026 0.0256 0.0213 0.0131 0.1296
0.0140 0.0021 0.0074 0.0026 0.2934
-0.0050 0.0097 0.0120 0.0157 0.3592
0.0091 -0.0014 0.0028 0.0151 -0.9201
0.0052 0.0171 0.0306 0.0259 0.2648
0.0167 0.0082 0.0085 0.0109 -0.3621
0.0113 0.0351 0.0074 0.0204 -0.5557
0.0053 -0.0009 -0.0016 0.0088 -0.6982
0.0153 0.0178 0.0122 0.0214 -0.1234
-0.0099 0.0220 0.0291 0.0108 -0.0857
```

$\bar{x}_{\text{jack}} = 0.0122$

$\sigma(x_{\text{jack}}) = 0.0096$

- $n_{\mathrm{samp}} = 50$:

$\bar{x}_{\mathrm{jack}} = 0.0122$

$\sigma(x_{\mathrm{jack}}) = 0.0096$

$$\bar{x}_{\mathrm{jack}} = \bar{x}, \sigma(x_{\mathrm{jack}}) = \frac{\sigma(x)}{\sqrt{n(n-1)}}$$

$\bar{x} = 0.0122$

$\sigma(x) = 0.4752$

- $n_{\mathrm{samp}} = 100000$:

$\bar{x} = 0.0004, \sigma(x) = 0.5003$

接近理论值：$\log(1 \pm 0.5) = 0 \pm 0.5$

Jackknife resampling是将样本的分布转换成样本平均值的分布

➢ Bootstrap resampling

在原样本$\boldsymbol{x}$中随机取出m个观测值组成一个新的样本$\boldsymbol{x_{(i)}}$，重复$\mathbf{N_{boot}}$次。此时所有的$\boldsymbol{\bar{x}_{(i)}}$组成一个新的样本$\boldsymbol{x}_{\mathrm{boot}}$。

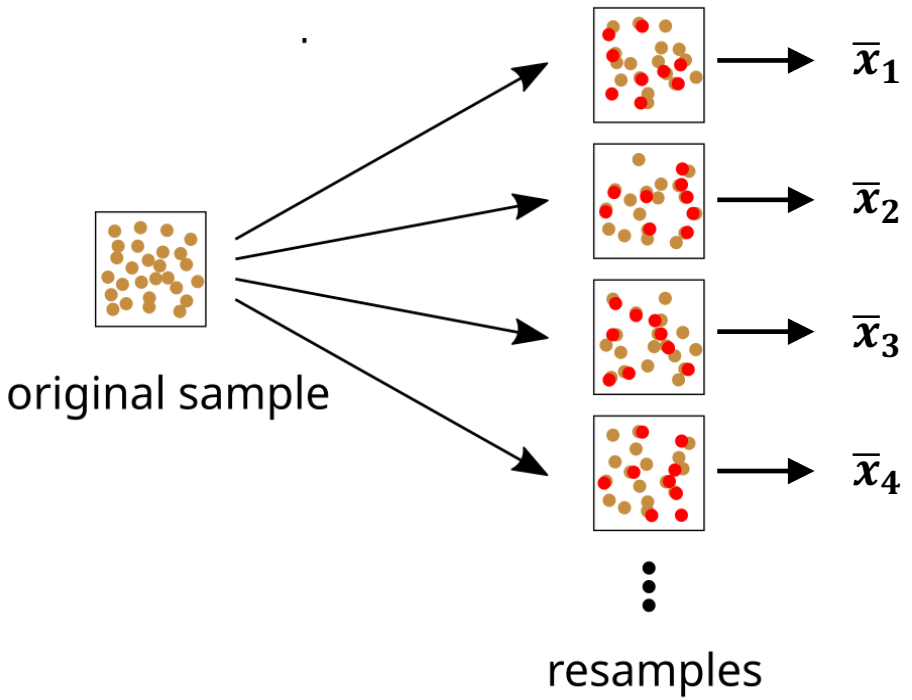
$$\bar{x}_{\mathrm{boot}} = \frac{1}{N_{\mathrm{boot}}}\sum_{i=1}^{N_{\mathrm{boot}}} \bar{x}_i = \frac{1}{N_{\mathrm{boot}} * m}\sum_{i=1}^{N_{\mathrm{boot}}*m} x_i$$

$$\simeq \frac{1}{N}\sum_{i=1}^{N} x_i = \bar{x}$$

当$\mathbf{N_{boot}} * \boldsymbol{m}$足够大时，可以认为样本中的每一个观测值都出现了$\mathbf{N_{boot}} * \boldsymbol{m/N}$次

$$\sigma^2(\bar{x}_{\mathrm{boot}}) = \frac{N-1}{N}\sigma^2(\bar{x})$$

当N足够大时，bootstrap后的方差与原样本的方差是一致的



➢ Jackknife与Bootstrap resampling的对比

Jackknife

- ✓ 定义简单，容易实现
- ✓ 定义明确，没有任意性。在做cross check时容易检验
- × 当样本量比较小的时候，扔掉一个对整个数据的影响较大，此时jackknife不太稳定
- × 对于一些非线性统计（比如对中位数的估计等）不太适用

Bootstrap

- ✓ 更灵活，有更大的自由度。 Jackknife实际Bootstrap的一种特殊情况
- ✓ 当样本量比较小的时候，比Jackknife方法更稳定
- ✓ 在一些非线性统计问题中更适用
- × 更复杂
- × 在样本量较大的时候需要更多的计算资源

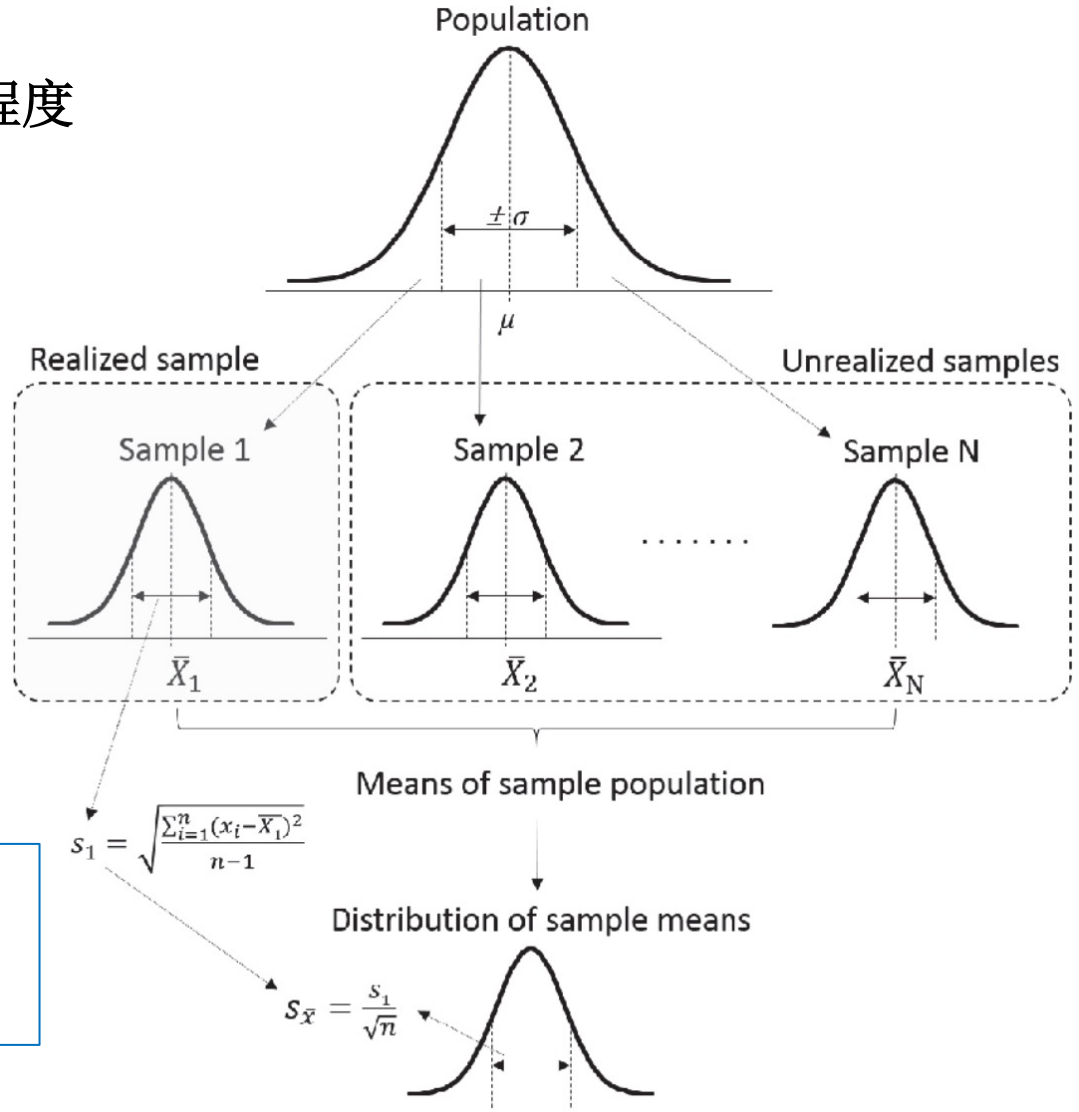
标准差 (standard deviation, SD)和标准误 (standard error of mean, SEM)

标准差 (standard deviation)：描述一个样本分布的分散程度

$$s = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{N-1}}$$

标准误 (standard error of mean, SEM)：描述样本中各个子集的差别，也就是样本平均值的分布。根据CLT可知，$\bar{x}$的标准误近似满足正态分布。

中心极限定理 (CLT)：当N足够大时样本的均值将近似满足正态分布



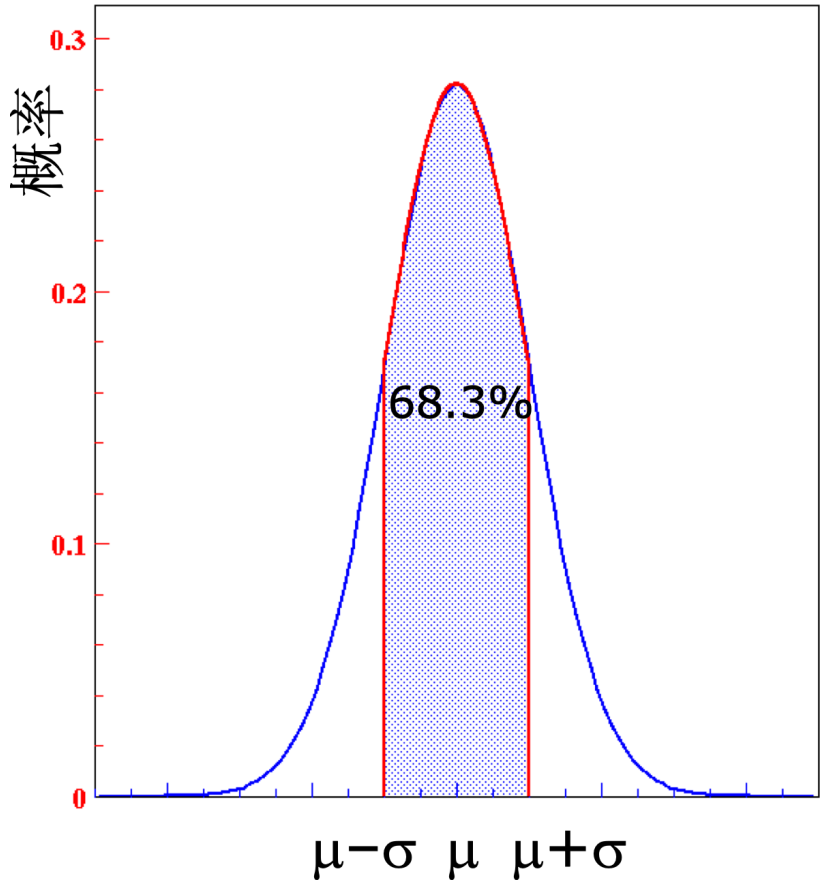
- 重采样的作用是将原本的样本分布变成了样本平均值的分布
- 在格点计算中，这是有实际的物理意义的：

$$\langle\Omega|\,T\phi_H(x_1)\phi_H(x_2)\,|\Omega\rangle = \lim_{T\to\infty(1-i\epsilon)} \frac{\int \mathcal{D}\phi\,\phi(x_1)\phi(x_2)\exp\left[i\int_{-T}^{T} d^4x\,\mathcal{L}\right]}{\int \mathcal{D}\phi\,\exp\left[i\int_{-T}^{T} d^4x\,\mathcal{L}\right]}.$$

  - $D\phi\,\exp[\text{xxx}]$项对应规范场组态，表示对路径积分的一次采样
  - $\phi(x_1)\phi(x_2)$可以看作是在规范场组态上的测量

关联函数是在路径积分平均（积分）的意义上定义的，而在一个组态上测量的结果对应的只是一次采样，只有采样的平均值平均值才可以看作是一次测量。而重采样就是把不同组态的结果多次采样得到的样本平均值的分布，这个分布才对应关联函数的测量。

➢ 我们经常看到，实验测量x的结果偏离理论n-σ.



- 假设一物理量的真值为X，实验对其进行测量，得到的结果发表为$\mu \pm \sigma$
- 该结果应该理解为：物理量的真值落在$\mu \pm \sigma$的概率范围内的概率为68.3%，或者说真值落在范围外的概率为31.7%
- **Evidence：3-σ** 99.73002%; **Discovery: 5-σ** 99.99994%
- 注意：在不知道真值的情况下，没有误差的结果是没有意义的

重采样后的误差估计

我们通常所说的“误差”，大多数时候指的是标准误 (standard error of mean, SEM)。我们在每一个数据点上要估计的是关联函数的平均值，所以我们要的误差实际上是平均值的误差。

- 重采样之前原始数据的SEM：

$$\sigma^2(\bar{x}) = \sigma^2\left(\frac{1}{N}\sum_{i=1}^{N} x_i\right) = \sigma^2\left(\sum_{i=1}^{N} x_i\right)/N^2$$
$$\simeq N\sigma^2(x_i)/N^2 = \sigma^2(x_i)/N$$

SD

即： $\mathbf{SEM} = \sqrt{\sigma^2(\bar{x})} = \mathbf{SD} / \sqrt{N}.$

用到的关系：

$$\sigma^2(x) = \mathrm{cov}(x, x) = E[(x - E(x))^2] = E(x^2) - [E(x)]^2$$

$$\sigma^2(Ax + By) = A^2\sigma^2(x) + B^2\sigma^2(y) + 2AB\,\mathrm{cov}(x, y)$$

- **Jackknife重采样后的SEM：**

  重采样后的新分布记做 $\boldsymbol{x_{-i}} = \frac{1}{N-1}\sum_{j\neq i} x_j$, 重采样不改变样本平均值，因此$\boldsymbol{E(x) = E(x_{-i})}$

  重采样后样本的方差：

$$\sigma^2(x_{-i}) = E[(x_{-i} - E(x))^2] = \frac{\sigma^2(x_i)}{(N-1)^2}$$

其中

$$\begin{aligned} x_{-i} - E(x_{-i}) &= \frac{1}{N-1}\sum_{j\neq i} x_i - E(x) \\ &= \frac{\sum_{j\neq i} x_i}{N-1} - E(x) \\ &= \frac{NE(x) - x_i}{N-1} - E(x) \\ &= \frac{E(x) - x_i}{N-1} \end{aligned}$$

即：

$$\sigma^2(\bar{x}) = \frac{\sigma^2(x_i)}{N} = \frac{(N-1)^2\sigma^2(x_{\text{jack}})}{N}$$

SEM(x) = SD(x_{jack}) * $\sqrt{N-1}$. for large N

- **Bootstrap重采样后的SEM：**

  Bootstrap后的每一个样本实际上是对原样本平均值的一次测量，即bootstrap后的样本为原样本平均值的分布。所以新样本的方差就是原样本平均值的方差。

  即： $\mathrm{SEM}(x) = \mathrm{SD}(x_{\mathrm{boot}})$.

误差的传递

假设$\boldsymbol{x}=(\boldsymbol{x_1},\boldsymbol{x_2},\ldots\ldots,\boldsymbol{x_n})$满足某一分布$\boldsymbol{f}(\boldsymbol{x})$，我们也许并不全部知道该函数形式，但假设我们知道平均值$\boldsymbol{\mu}=\boldsymbol{E}(\boldsymbol{x})$和协方差$\boldsymbol{V_{ij}}=\mathbf{cov}(\boldsymbol{x_i},\boldsymbol{x_j})$。

现考虑一函数$\boldsymbol{y}(\boldsymbol{x})$，方差$\boldsymbol{V}(\boldsymbol{y})=\boldsymbol{E}\left(\boldsymbol{y^2}\right)-[\boldsymbol{E}(\boldsymbol{y})]^2$是什么？

误差的传递

假设$\boldsymbol{x}=(\boldsymbol{x_1},\boldsymbol{x_2},\ldots\ldots,\boldsymbol{x_n})$满足某一分布$\boldsymbol{f}(\boldsymbol{x})$，我们也许并不全部知道该函数形式，但假设我们知道平均值$\boldsymbol{\mu}=\boldsymbol{E}(\boldsymbol{x})$和协方差$\boldsymbol{V_{ij}}=\mathbf{cov}(\boldsymbol{x_i},\boldsymbol{x_j})$。

现考虑一函数$\boldsymbol{y}(\boldsymbol{x})$，方差$\boldsymbol{V}(\boldsymbol{y})=\boldsymbol{E}(\boldsymbol{y^2})-[\boldsymbol{E}(\boldsymbol{y})]^{\boldsymbol{2}}$是什么？

将$\boldsymbol{y}(\boldsymbol{x})$在$\boldsymbol{\mu}$附近做泰勒展开：

$$y(x) \simeq y(\mu)+\sum_{i=1}^{n}\left[\frac{\partial y}{\partial x_i}\right]_{x=\mu}(x_i-\mu)$$

然后计算$\boldsymbol{E}(\boldsymbol{y})$和$\boldsymbol{E}(\boldsymbol{y^2})$

由于$\sum_i \boldsymbol{E(x_i - \mu) = 0}$, 所以

$$
\begin{aligned}
E[y(x)] &\simeq y(\mu) \\
E[y^2(x)] &\simeq y^2(\mu) + 2y(\mu)\sum_{i=1}^{n}\left[\frac{\partial y}{\partial x_i}\right]_{x=\mu} E(x_i - \mu) \\
&\quad + E\left[\left(\sum_{i=1}^{n}\left[\frac{\partial y}{\partial x_i}\right]_{x=\mu}(x_i - \mu)\right)\left(\sum_{j=1}^{n}\left[\frac{\partial y}{\partial x_j}\right]_{x=\mu}(x_j - \mu)\right)\right] \\
&= y^2(\mu) + \sum_{i,j=1}^{n}\left[\frac{\partial y}{\partial x_i}\frac{\partial y}{\partial x_j}\right]V_{ij}
\end{aligned}
$$

两项合起来给出$\boldsymbol{y(x)}$的方差

$$
\sigma^2(y) \simeq \sum_{i,j=1}^{n}\left[\frac{\partial y}{\partial x_i}\frac{\partial y}{\partial x_j}\right]_{x=\mu} V_{ij}
$$

如果$\boldsymbol{x_i}$之间是无关联的，则$\boldsymbol{V_{ij} = \sigma^2(x_i)\delta_{ij}}$，上式变为

$$
\sigma^2(y) \simeq \sum_{i=1}^{n}\left[\frac{\partial y}{\partial x_i}\right]_{x=\mu}^{2} \sigma^2(x_i)
$$

注意：

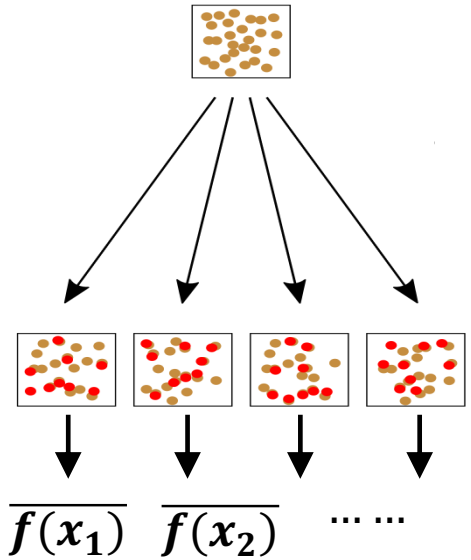
- 上述操作不需要知道x满足的分布的具体形式。
- 函数方差的计算方法是基于一阶泰勒展开，只对$y(x)$是线性变换时是精确的，近似程度在函数非线性区变化比$\sigma^2(x_i)$要大时遭到很大的破坏。
- 在经过多次非线性（像拟合）操作后，误差传递的精确性也会遭到较大的破坏。

注意：

- 上述操作不需要知道x满足的分布的具体形式。
- 函数方差的计算方法是基于一阶泰勒展开，只对$y(x)$是线性变换时是精确的，近似程度在函数非线性区变化比$\sigma^2(x_i)$要大时遭到很大的破坏。
- 在经过多次非线性（像拟合）操作后，误差传递的精确性也会遭到较大的破坏。

在实际操作时，更方便的做法是直接对重采样后的样本进行操作。此时所有的操作是作用在每一次采样的平均值上。

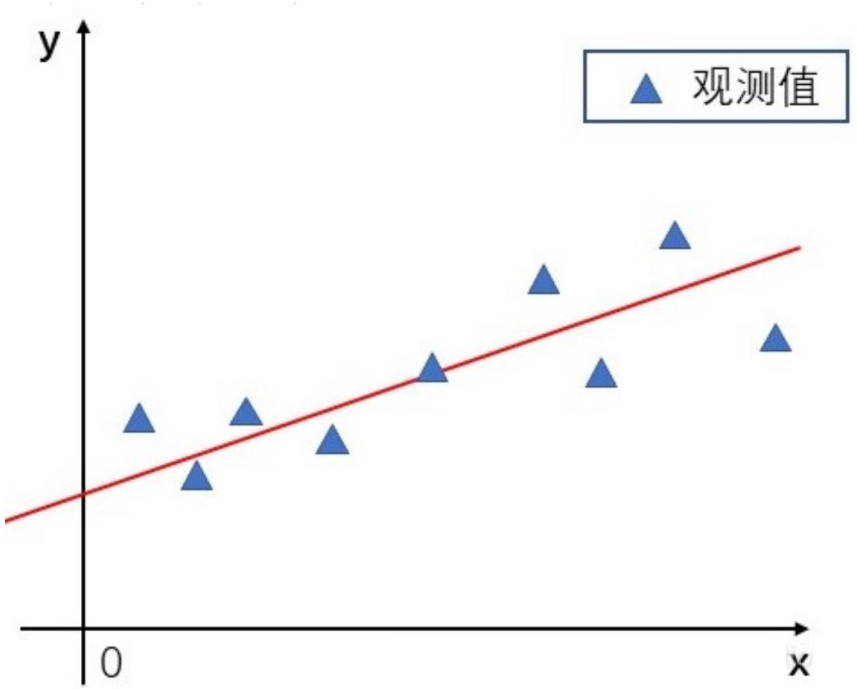
- 比如拟合，可以对重采样后的每一组样本子集做拟合，然后将拟合结果的平均值作为对这一个子集的估计。
- 多次连续拟合亦是如此。



数据拟合

假设我们观测到了一组数据(x_i, y_i)，根据数据的行为我们猜测y与x存在线性关系，这个关系如何准确的描述？对数据行为的描述又有多准确？

我们可以猜测一个模型$f(x) = kx + b$, 理论上只需要两组数据就可以定出模型中的两个参数k和b（解方程组） 。这种情况下，模型的参数有唯一确定解。



但是在实际中，我们的观测会存在误差，所以我们总会做多余观测，以期通过增加观测量来提高对参数预言的准确性。但是冗余的观测会导致我们难以找到一条直线穿过所有的点，也就是说，方程没有确定解。

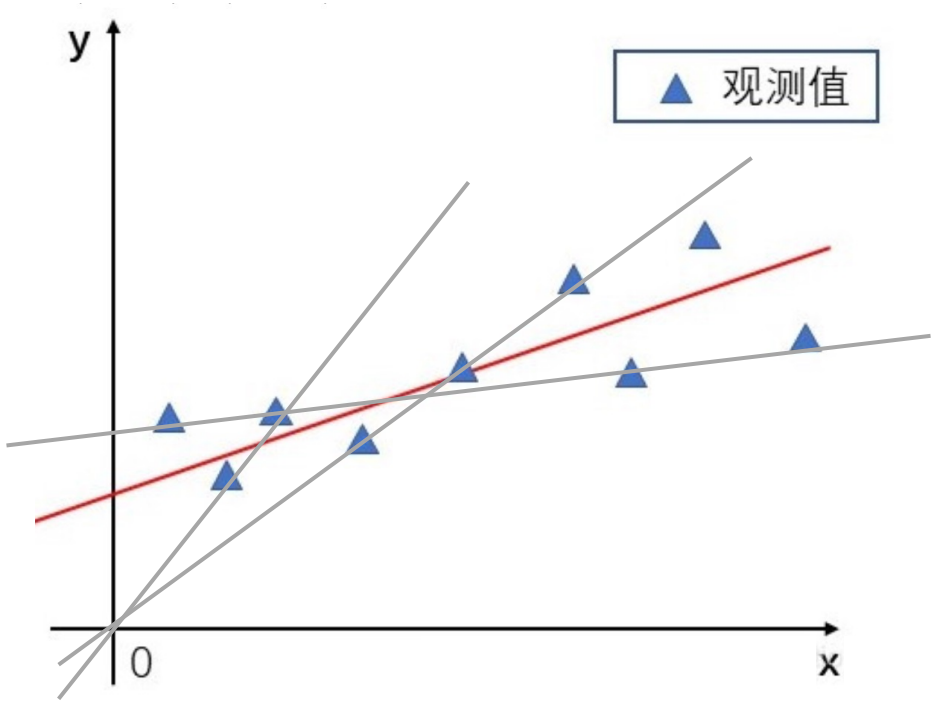
数据拟合

那我们能不能找到一种最佳方案，使得模型能尽量符合更多的观测点？

早在19世纪，勒让德就提出“让误差的平方和最小”估计出来的模型是最接近真实情形的：

$$L = \sum_{i=1}^{N} [y_i - f(x_i)]^2$$

这个准则也被称为最小二乘准则。

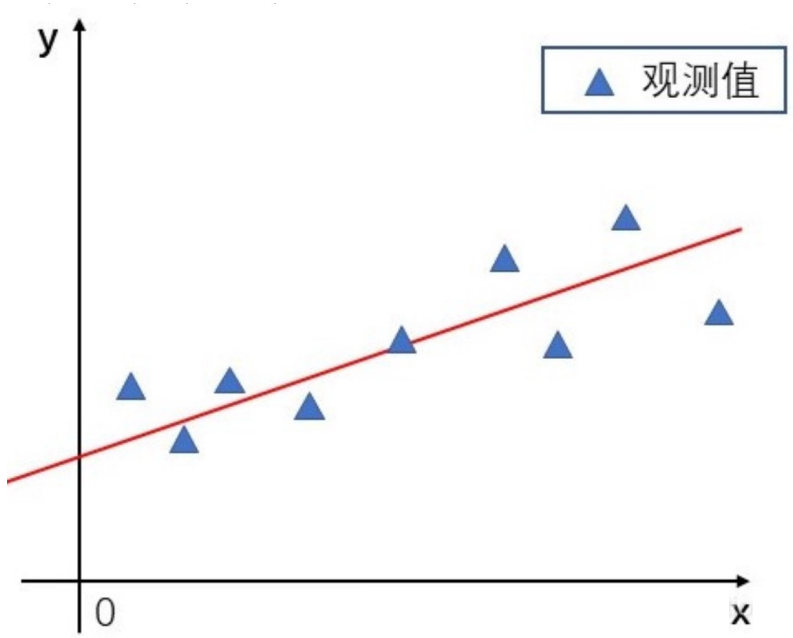


➢ 拟合χ^2的检验：当拟合出描述原始数据的曲线后，如何检验拟合的优度？

假设待拟合函数记做$f(x;\mathbf{b})$，包含P个待拟合参数b_i.

我们的目标是寻找合适的参数 **b** 来最小化以下如下目标函数：

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{y_i - f(x_i;\mathbf{b})}{\sigma_i} \right)^2$$



目标就是求解方程组：

$$\frac{\partial \chi^2}{\partial b_i} = 0, \qquad i = 1, \ldots\ldots, P$$

拟合的过程就是最小化χ^2的过程。其物理意义是，在考虑到测量误差的情况下，正确的参数值应该给出最小的残差。

➢ 带有约束的拟合 (constrained fitting) 的χ^2检验

arXiv:hep-lat/0110175

很多时候在进行拟合时，我们会根据物理的条件对拟合的参数加以限制。此时输入参数$\mathbf{b} = (\boldsymbol{b_1}, \ldots, \boldsymbol{b_i})$的预设值(priors)也会对$\chi^2$产生影响。

在基于贝叶斯统计的带有约束的曲线拟合中，χ^2的定义为

$$\chi^2 \rightarrow \chi^2_{\text{aug}} \equiv \chi^2 + \chi^2_{\text{prior}} \qquad \chi^2_{\text{prior}} \equiv \sum_i \frac{\left(b_i - \tilde{b}_i\right)^2}{\tilde{\sigma}^2_{b_i}}$$

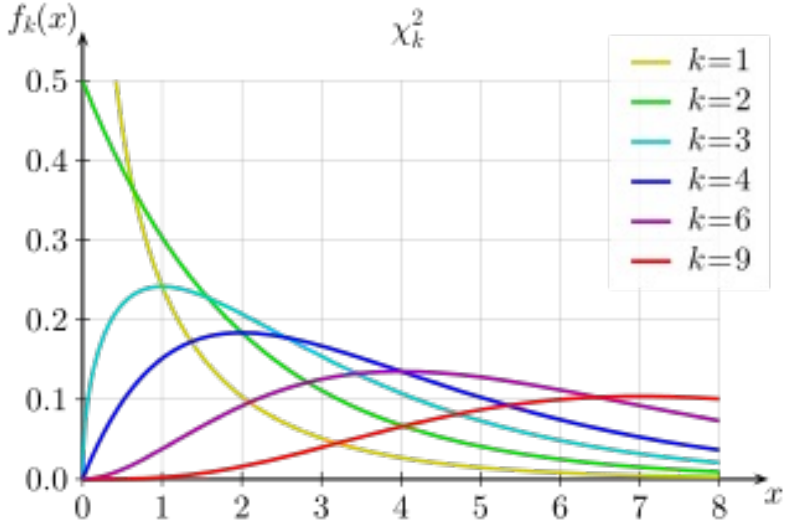
其中$\boldsymbol{b}_{\text{prior}} = \tilde{\boldsymbol{b}} \pm \tilde{\boldsymbol{\sigma}}_{\boldsymbol{b}}$是在拟合前对参数的预设值，它通常基于我们对拟合参数的物理限制的了解，给出一个合理的范围。

➢ 拟合优度的χ^2检验

从前面可以看出，如果χ^2太大，说明拟合不好。那χ^2是越小越好吗？

如果拟合很好，那么计算出的χ^2应该很好得服从相应自由度的χ^2分布，即：不会太大，也不会太小。

通常人们通过约化χ^2(reduced χ^2)来检验拟合的可靠性，即$\chi^2_{\mathrm{red}} = \chi^2/v$, 它等于$\chi^2$方除以其自由度。经验上可以通过$\chi^2_{\mathrm{red}} \sim 1$来检验拟合的好坏（并不严格，但是很方便）。



可以验证，自由度为v的χ^2分布，其均值$E(\chi^2) = v$, 方差$\sigma^2(\chi^2) = 2v$. 因此$E(\chi^2_{\mathrm{red}}) = 1$, $\sigma^2(\chi^2_{\mathrm{red}}) = 2/v$. 当自由度$v$很大时，$\chi^2$分布趋于高斯分布，有$\chi^2_{\mathrm{red}} \sim N(1, 2/v), v \gg 1$. 此时约化$\chi^2$趋于一个以1为中心的较窄的高斯分布。

- 也可以通过拟合的***P*-value**和拟合自由度来检验拟合的好坏：

Table for χ^2 and $P(\chi^2,\nu)$

	$P(\chi^2,\nu)$											
	0.99	0.95	0.94	0.9	0.70	0.50	0.30	0.10	0.06	0.05	0.01	0.001
1	0.0002	0.0039	0.0057	0.016	0.15	0.45	1.1	2.7	3.54	3.84	6.6	11
2	0.02	0.10	0.12	0.21	0.71	1.39	2.4	4.6	5.6	6.0	9.2	14
3	0.11	0.35	0.40	0.58	1.42	2.37	3.7	6.3	7.4	7.8	11	16
4	0.30	0.71	0.79	1.06	2.2	3.4	4.9	7.8	9.0	9.5	13	18
5	0.55	1.15	1.25	1.6	3.0	4.4	6.1	9.2	10.6	11.1	15	21
6	0.87	1.64	1.76	2.2	3.8	5.3	7.2	11	12.1	12.6	17	22
7	1.2	2.17	2.32	2.8	4.7	6.3	8.4	12	13.5	14.1	18	24
8	1.6	2.73	2.91	3.5	5.5	7.3	9.5	13	15.0	15.5	20	26
9	2.1	3.33	3.52	4.2	6.4	8.3	11	15	16.3	16.9	22	28
10	2.6	3.94	4.16	4.9	7.3	9.3	12	16	17.7	18.3	23	30
12	3.6	5.23	5.48	6.3	9.0	11	14	19	20.4	21.0	26	33
15	5.2	7.26	7.57	8.5	12	14	17	22	24.3	25.0	31	38
20	8	10.9	11.2	12	16	19	23	28	30.6	31.4	38	45
30	15	18.5	19.0	21	26	29	34	40	42.9	43.8	51	60
50	30	34.8	35.5	38	44	49	55	63	66.4	67.5	76	87

(ν)

If $P(\chi^2,\nu) > 0.95$ then $P(\chi^2,\nu)$ is TOO BIG	If $0.95 > P(\chi^2,\nu) > 0.05$ then the disagreement between theory and experiment is not significant	If $P(\chi^2,\nu) < 0.05$ then $P(\chi^2,\nu)$ is TOO SMALL
χ^2 TOO SMALL	χ^2 OK	χ^2 TOO BIG
"given the error bars, the theory is much too close to the experiment… error bars might be wrong or someone cheated"	"theory and experiment are consistent"	"given the error bars, there is significant disagreement between theory and experiment"
FITS TOO WELL GIVEN THE SIZE OF THE ERROR BARS	GOOD FIT	DOES NOT FIT