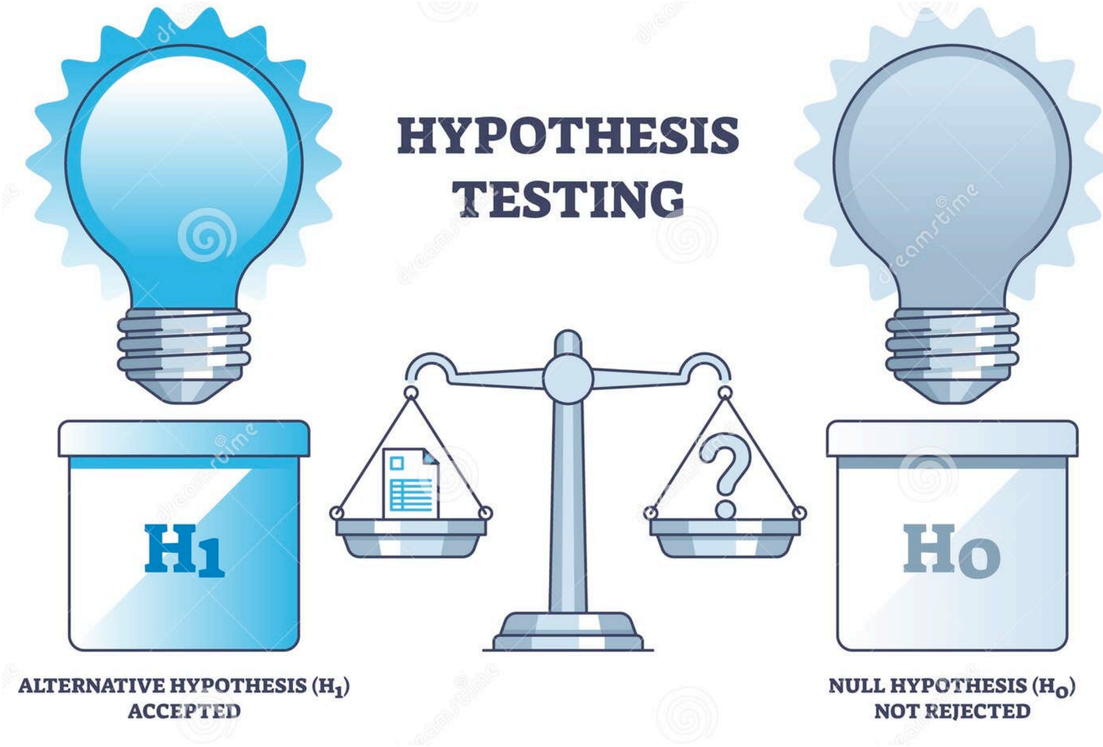
Foundations of Data Science

Hypothesis Test

尹一通、刘明谋 Nanjing University, 2024 Fall

假设检验

Hypothesis Test



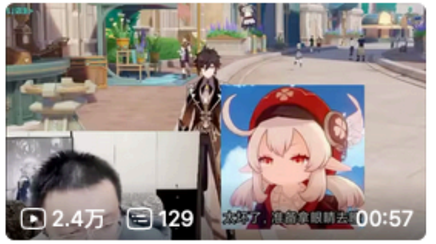
暗改概率？

怎么确定是否暗改？

- 问题：是否暗改
- 收集数据：抽卡记录
- 分析：？
  - 声明1.5%出率，10000抽中 160 / 120 / 80
  - 可以判断改低了吗？会不会只是小概率事件？
- 我们永远无法肯定地判断

严重怀疑原神圣遗物概率被米哈游暗改

UP 雅泽晶蝶 · 6-12



老四讲述圣遗物出货概率被暗改了？我也有同样的感觉 刷了快半年...

UP 二游up主 · 3-27

沉船不是因为你脸黑，将游戏概率暗改到零的Nexon，被罚款116亿

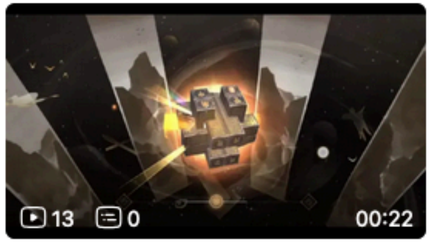
UP 3DM游戏 · 1-5

暗改概率？18发不出

UP 辉-brilliance · 4-19

策划发文增幅概率没改！宝哥找来牛津大学数学系大神计算后：石...

UP 不二青风 · 2022-3-18

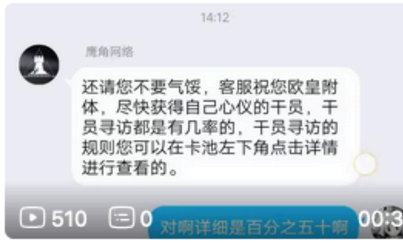


概率被暗改了？？？！！！

UP 初のの晴 · 7-12

宝哥花钱请人专门统计一万件红12上13的概率！用数据来说明是否...

UP 旭旭姥姥6868 · 2022-3-16

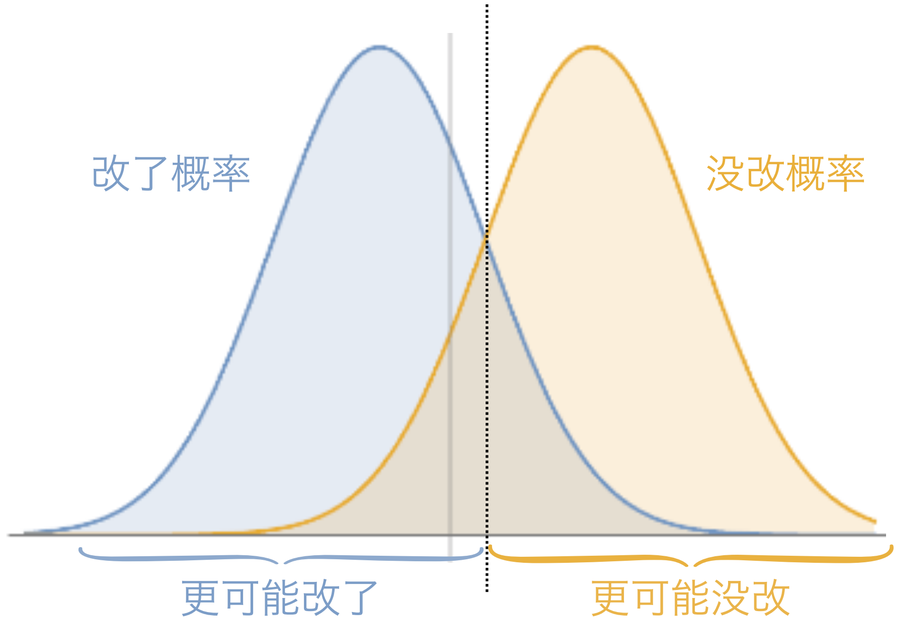


暗改概率敷衍回答还装死真有你的啊

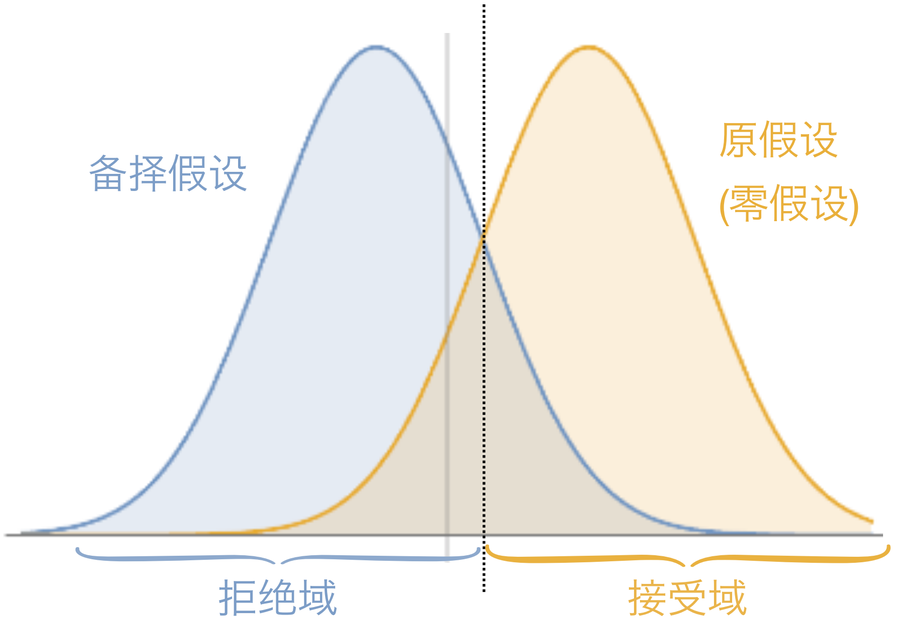
UP 风扇克星 · 2021-7-4

暗改概率？

- 我们永远无法准确地判断
- 但是我们可以量化我们的“把握”
- 假设改了：有些样本更容易出现
- 假设没改：另一些样本更容易出现
- 如果我们是二极管，那么如此猜测

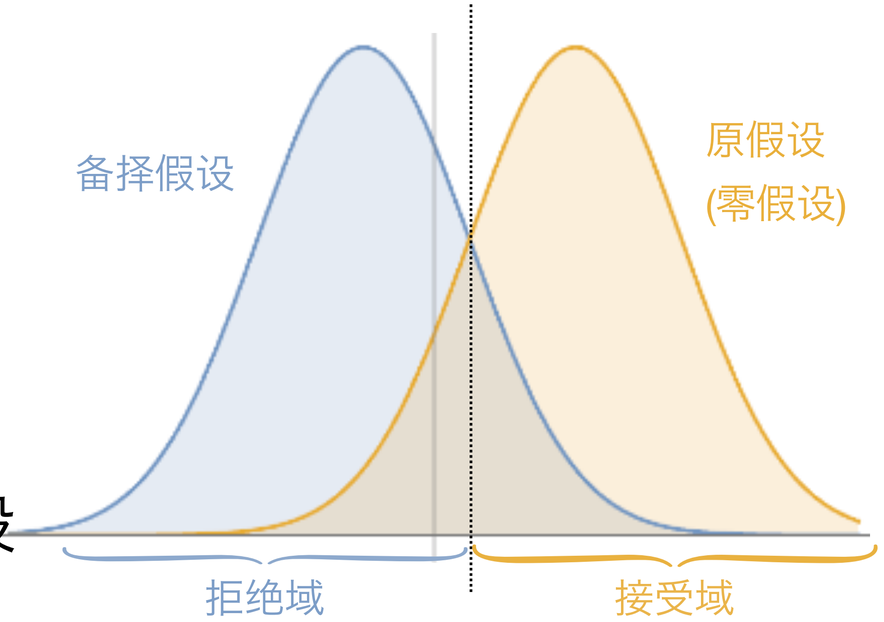


假设检验（Hypothesis test）



- 没改：原假设、零假设 (null hypothesis)
- 改了：备择假设 (alternative hypothesis)
- 检验法则(decision rule):
  - 接受域（ acceptance region）：原假设更有可能对的，应该接受原假设
  - 拒绝域（rejection region）、临界域（critical region）：应该拒绝原假设
- 我们永远无法准确地判断

假设检验（Hypothesis test）

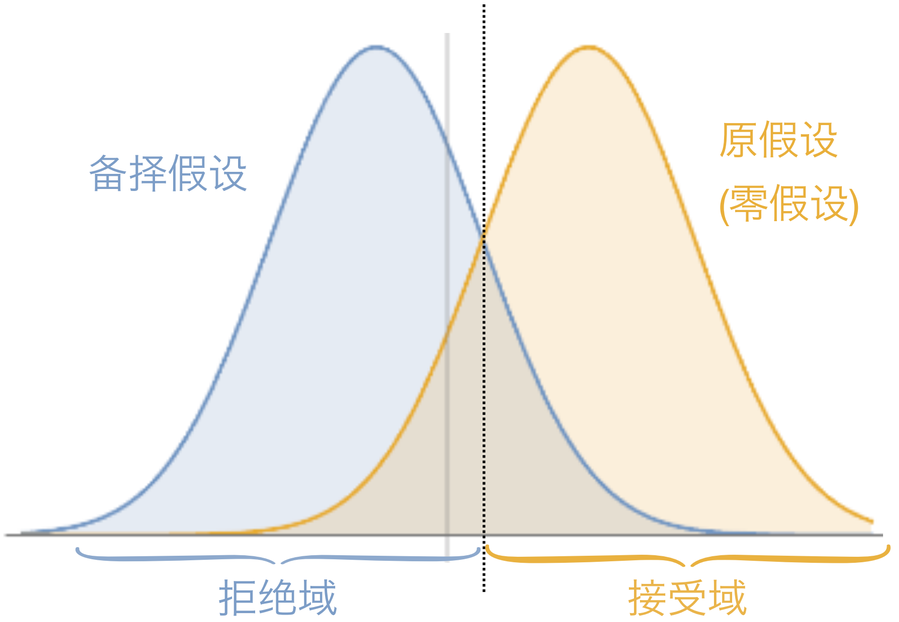


- 检验法则(decision rule):
  - 接受域（ acceptance region）：应该接受原假设
  - 拒绝域（rejection region）、临界域（critical region）：应该拒绝原假设
- 我们永远无法准确地判断

	接受原假设	拒绝原假设
原假设为真	正确	一类错误 (弃真/假阳性)
原假设为假	二类错误 (取伪/假阴性)	正确

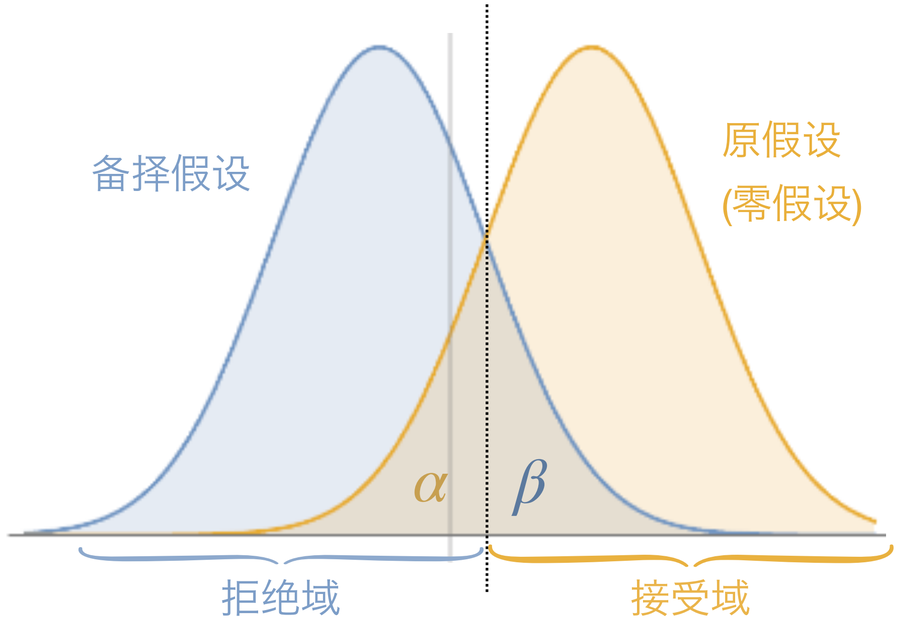
假设检验（Hypothesis test）

基本步骤



1. 提出统计假设：原假设 H_0、备择假设 H_1
2. 针对两种假设确定能区分它们的统计量
3. 根据统计量确定拒绝域和接受域
4. 采样、从样本中计算出统计值
5. 判断统计值在拒绝域还是接受域、做出决策

显著性 α 和检验功效 $1-\beta$



- 非此即彼，过于粗糙
- 边界情况？判断的可信度？出错概率？
- 犯错的概率
  - 原假设为真，犯一类错误概率 α
  - 原假设为假，犯二类错误概率 β
  - α: 显著性 (significance)
  - $1-\beta$: 检验功效 (power)
  - 置信水平 $\gamma = 1-\alpha$

	接受原假设	拒绝原假设
原假设为真	正确 1-α	一类错误 α
原假设为假	二类错误 β	正确 1-β

生活中的统计问题

- 原假设： 性别出生比 1:1
- 备择假设：性别出生比不是 1:1

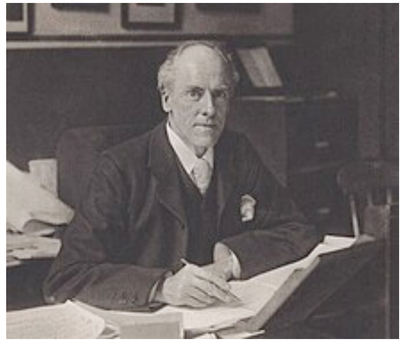
John Arbuthnot
(1667-1735)

Ronald Fisher
(1890-1962)

Jerzy Neyman
(1894-1981)

Egon Pearson
(1895-1980)

- 若原假设是对的， 该样本出现的概率是 $1/2^{82}$ (Arbuthnot 1710; Laplace 1770s)
- Fisher (1925): tests of significance， 阈值 5%.
- Neyman & Pearson (1933): significance level, α.



Karl Pearson
(1857-1936)

- 生男生女的概率？你觉得是 50% 吗?
  - 2023年全国男性人口72032万人，女性人口68935万人
  - 2021年全国出生人口性别比108.3，2019年全国出生人口性别比110.14
  - John Arbuthnot (1710): 1629 to 1710, in every year, the number of males born in London exceeded the number of females.

$$1/2^{82} = 1/4835703278458516698824704$$

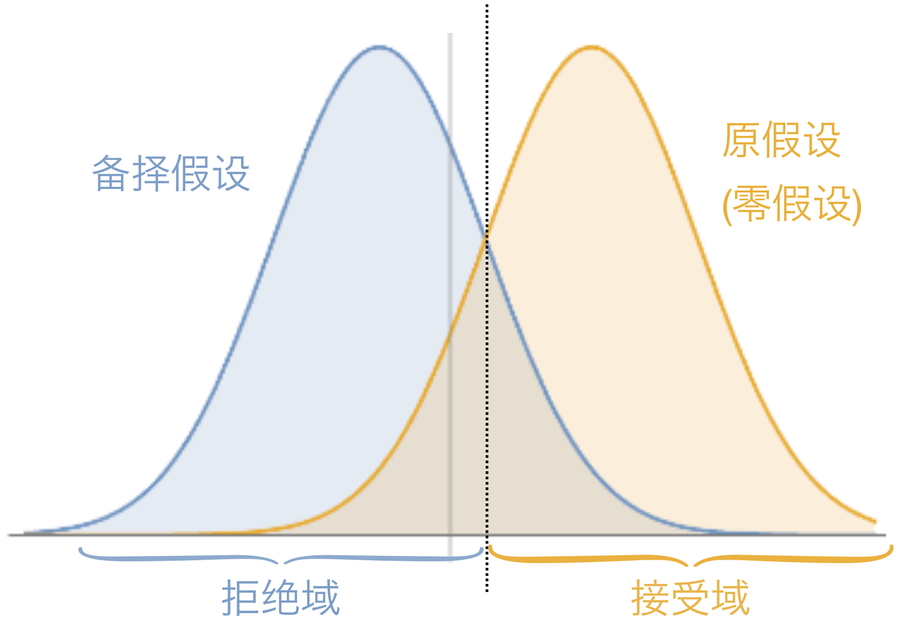
女士品茶（lady tasting tea）

Ronald Fisher, *The Design of Experiments*. (1935)

- 英式奶茶：先放奶还是先放茶，有区别吗？
  - 可乐：百事和可口的味道是一样的吗？
- 原假设：一样；备择假设：不一样
- 试验：四杯奶茶 (可口) + 四杯茶奶 (百事)，受试者随机品尝并选出四杯奶茶 (可口)
- 数据：受试者对了 k 杯
  - 若原假设为真，该样本出现的概率 $\binom{4}{k} / \binom{8}{4}$
- 实际：受试者 (藻类学家Muriel Bristol) 全对
  - 若原假设为真，该样本出现的概率是 $1/\binom{8}{4} = 1/70 \approx 1.429\,\%$

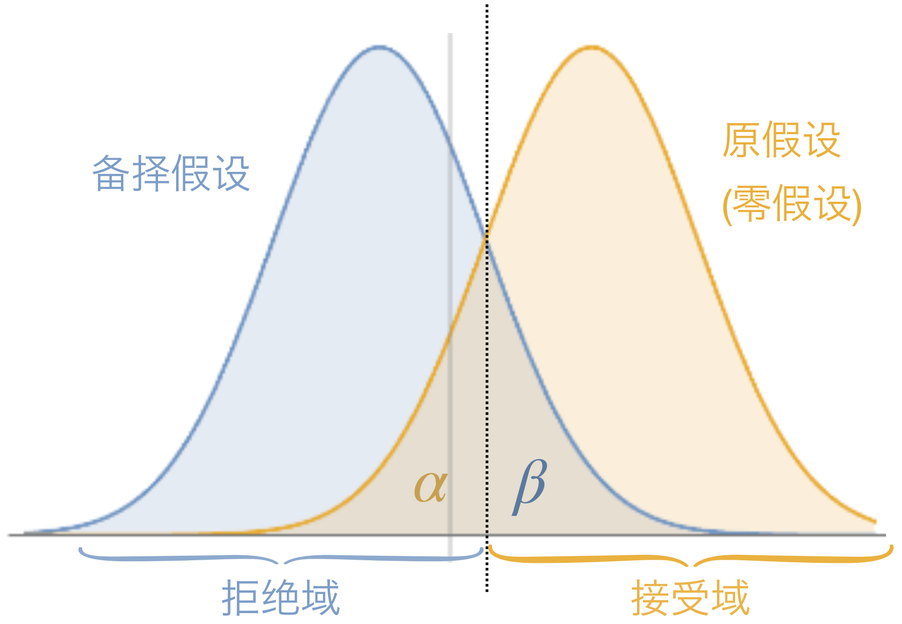
假设检验（Hypothesis test）

基本步骤



1. 提出统计假设：原假设 H_0、备择假设 H_1
2. 针对两种假设确定能区分它们的统计量
3. 规定显著性水平 α
4. 根据显著性确定拒绝域和接受域
5. 采样、从样本中计算出统计值
6. 判断统计值在拒绝域还是接受域、做出决策：“在显著性水平 α 下接受/拒绝”

显著性 α 和检验功效 $1-\beta$



- 犯错的概率
  - 犯一类错误概率 α: 显著性 (significance)
  - 无二类错误概率$1-\beta$: 检验功效 (power)
  - 置信水平 $\gamma = 1-\alpha$
- α 与 β 互相矛盾
- 两全其美：更大样本量以区分
  - 固定 α，提高样本量，使 $\beta \leq \alpha$
  - 功效 (power)：样本量由 β 确定

	接受原假设	拒绝原假设
原假设为真	正确 1-α	一类错误 α
原假设为假	二类错误 β	正确 1-β

正态总体参数检验

已知方差 σ^2，检验期望 μ

- 一洗衣粉包装机，额定标准500g/包。装袋重量服从正态分布 $N(\mu, \sigma^2)$。称得样本 $X_1, \ldots, X_n$。取显著性水平 α，问该包装机是否工作正常？

1. 原假设 $H_0 : \mu = 500$，备择假设 $H_1 : \mu \neq 500$ (双侧检验)

2. 检验统计量 $Z = \dfrac{\bar{X} - 500}{\sigma/\sqrt{n}}$。若原假设成立则 $Z \sim N(0,1)$（**Z检验**）

3. 拒绝域：$\{z : \Pr[X \leq z] \leq \alpha/2 \vee \Pr[X \geq z] \leq \alpha/2\}$, 其中 $X \sim N(0,1)$

正态总体参数检验

已知方差 σ^2，检验期望 μ

- 一洗衣粉包装机，额定标准500g/包。装袋重量服从正态分布 $N(\mu, \sigma^2)$。称得样本 $X_1, \ldots, X_n$。取显著性水平 α，问该包装机是否装少了？

1. 原假设 $H_0 : \mu = 500$，备择假设 $H_1 : \mu < 500$ (左侧检验)

2. 检验统计量 $Z = \dfrac{\bar{X} - 500}{\sigma/\sqrt{n}}$。若原假设成立则 $Z \sim N(0,1)$（**Z检验**）

3. 拒绝域：$\{z : \Pr[X \le z] \le \alpha\}$, 其中 $X \sim N(0,1)$

正态总体参数检验

已知方差 σ^2，检验期望 μ

- 一洗衣粉包装机，额定标准500g/包。装袋重量服从正态分布 $N(\mu, \sigma^2)$。称得样本 $X_1, \ldots, X_n$。取显著性水平 α，问该包装机是否装多了？

1. 原假设 $H_0 : \mu = 500$，备择假设 $H_1 : \mu > 500$ (右侧检验)

2. 检验统计量 $Z = \dfrac{\bar{X} - 500}{\sigma/\sqrt{n}}$。若原假设成立则 $Z \sim N(0,1)$（**Z检验**）

3. 拒绝域：$\{z : \Pr[X \geq z] \leq \alpha\}$, 其中 $X \sim N(0,1)$

正态总体参数检验

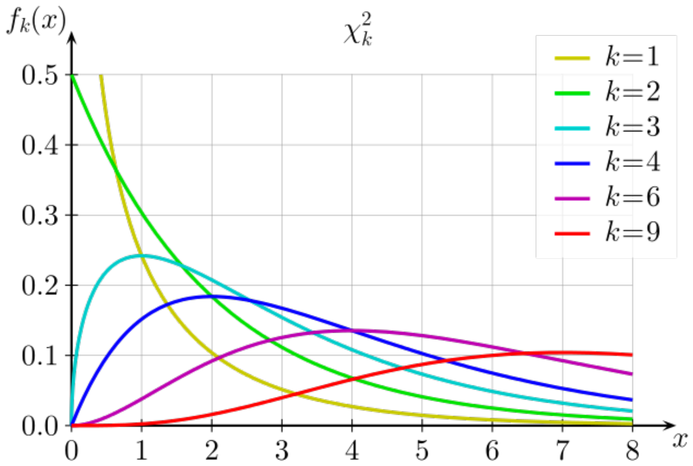
已知期望μ，检验方差 σ^2

- 一洗衣粉包装机，额定标准500g/包。装袋重量服从正态分布 $N(\mu, \sigma^2)$。称得样本 $X_1, \ldots, X_n$。已知$\mu = 500$，取显著性水平 α，检验 $\sigma^2 = \sigma_0^2$。

1. 原假设 $H_0 : \sigma^2 = \sigma_0^2$，备择假设 $H_1 : \sigma^2 \neq \sigma_0^2$ (双侧检验)

2. 什么统计量能与方差产生关联？
   - 样本方差：$S^2 = \Sigma_i (X_i - \bar{X})^2/(n-1)$，$\mathbb{E}[S^2] = \sigma^2$
3. 已知期望，检验统计量 $Z = \Sigma_i (X_i - \mu)^2/\sigma_0^2$。若原假设成立则 $Z \sim$？

Chi-squared (χ^2) distribution



- If $Z_1, \ldots, Z_k$ are **independent standard normal** random variables, then

$$Q = \sum_{i=1}^{k} Z_i^2$$

  follows the **chi-squared** (卡方) distribution with k **degrees of freedom**, denoted as $Q \sim \chi^2(k)$

- $\mathbb{E}[Z_i^2] = \mathbf{Var}[Z_i] = 1$ since $Z_i \sim N(0,1) \Longrightarrow \mathbb{E}[Q] = k$

- sum of independent $\chi^2(k)$ and $\chi^2(l)$ random variables follows $\chi^2(k+l)$

正态总体参数检验

已知期望μ，检验方差 σ^2

- 一洗衣粉包装机，额定标准500g/包。装袋重量服从正态分布 $N(\mu, \sigma^2)$。称得样本 $X_1, \ldots, X_n$。已知$\mu = 500$，取显著性水平 α，检验 $\sigma^2 = \sigma_0^2$。

1. 原假设 $H_0 : \sigma^2 = \sigma_0^2$，备择假设 $H_1 : \sigma^2 \neq \sigma_0^2$ (双侧检验)

2. 什么统计量能与方差产生关联？
   - 样本方差：$S^2 = \Sigma_i (X_i - \bar{X})^2/(n-1)$，$\mathbb{E}[S^2] = \sigma^2$

（**卡方检验**）

3. 已知期望，检验统计量 $Z = \Sigma_i (X_i - \mu)^2/\sigma_0^2$。若原假设成立则 $Z \sim \chi^2(n)$
4. 拒绝域 $\{z : \Pr[X \leq z] \leq \alpha/2 \vee \Pr[X \geq z] \leq \alpha/2\}$, 其中 $X \sim \chi^2(n)$

正态总体参数检验

未知期望μ，检验方差 σ^2

- 一洗衣粉包装机，额定标准500g/包。装袋重量服从正态分布 $N(\mu, \sigma^2)$。称得样本 $X_1, \ldots, X_n$。已知$\mu = 500$，取显著性水平 α，检验 $\sigma^2 = \sigma_0^2$。

1. 原假设 $H_0 : \sigma^2 = \sigma_0^2$，备择假设 $H_1 : \sigma^2 \neq \sigma_0^2$ (双侧检验)

2. 什么统计量能与方差产生关联？
   - 样本方差：$S^2 = \Sigma_i (X_i - \bar{X})^2/(n-1)$，$\mathbb{E}[S^2] = \sigma^2$

3. 不知期望，检验 $Z = \Sigma_i (X_i - \bar{X})^2/\sigma_0^2$。若原假设成立则 $Z \sim \chi^2(n-1)$ **（卡方检验）**

4. 拒绝域 $\{z : \Pr[X \leq z] \leq \alpha/2 \vee \Pr[X \geq z] \leq \alpha/2\}$, 其中 $X \sim \chi^2(n-1)$

正态总体的样本方差的分布

- 若 $X_1, \ldots, X_n \sim N(0,1)$, 则 $\sum_{i=1}^{n} (X_i - \bar{X})^2 \sim \chi^2(n-1)$.
- **Proof**: 样本方差 $S^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}$. 注意到 $\sum_{i=1}^{n} (X_i - \bar{X})^2 = \sum_{i=1}^{n} X_i^2 - n\bar{X}^2$

  又因为 $\bar{X} \sim N(0,1/n)$, 所以 $\sqrt{n}\bar{X} \sim N(0,1)$, 且 $n\bar{X}^2 \sim \chi^2(1)$

  改写 $\sum_{i=1}^{n} X_i^2 = (n-1)S^2 + n\bar{X}^2$, 记作 $\chi^2(n) = \underline{(n-1)S^2 + \chi^2(1)}$

  相互独立

矩生成函数 (MGF) : $M_{\chi_n^2}(t) = M_{(n-1)S^2}(t) \cdot M_{\chi_1^2}(t)$

卡方分布的矩生成函数是

$$M_{\chi_n^2}(t) = (1-2t)^{-n/2}$$

正态总体的样本方差的分布

卡方分布的矩生成函数是

$$M_{\chi_n^2}(t) = (1-2t)^{-n/2}$$

- 若 $X_1, \ldots, X_n \sim N(0,1)$, 则 $\sum_{i=1}^{n} (X_i - \bar{X})^2 \sim \chi^2(n-1)$.
- **Proof**: 记作 $\chi^2(n) = (n-1)S^2 + \chi^2(1)$.

矩生成函数 (MGF) : $M_{\chi_n^2}(t) = M_{(n-1)S^2}(t) \cdot M_{\chi_1^2}(t)$

$$(1-2t)^{-n/2} = M_{(n-1)S^2}(t) \cdot (1-2t)^{-1/2}$$

$$M_{(n-1)S^2}(t) = (1-2t)^{-n/2} \cdot (1-2t)^{1/2}$$

$$= (1-2t)^{-(n-1)/2}$$

因此 $\sum_{i=1}^{n} (X_i - \bar{X})^2 = (n-1)S^2 \sim \chi^2(n-1)$

卡方(chi-squared)分布

卡方分布的矩生成函数是

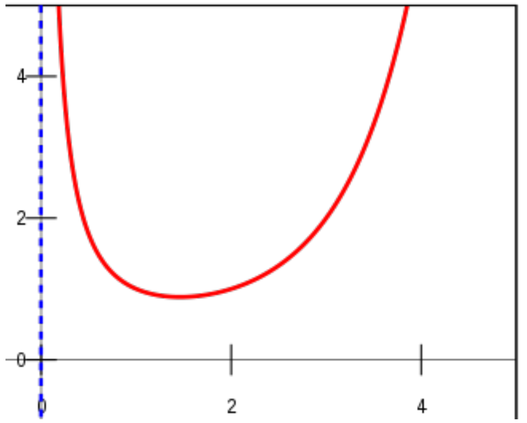
$$M_{\chi_n^2}(t) = (1-2t)^{-n/2}$$

- 令 $X \sim \chi_n^2$, 则其概率密度函数 $f(x) = \dfrac{x^{n/2-1}e^{-x/2}}{2^{n/2}\cdot\Gamma(n/2)}$, 其中 $\Gamma(n/2)$是伽马函数。

Gamma Function*

"Each generation has found something of interest to say about the gamma function. Perhaps the next generation will also."

—Philip J. Davis



- **Gamma function** $\Gamma(z)$: analytic extension of factorial $\Gamma(n) = (n-1)!$

$$\Gamma(z) = \int_0^\infty t^{z-1} \mathrm{e}^{-t}\,\mathrm{d}t, \quad \text{for } \Re(z) > 0$$

- $\Gamma(1) = \int_0^\infty \mathrm{e}^{-t}\,\mathrm{d}t = 1$ gives exponential distribution with $\lambda = 1$
- $\Gamma(k) = \int_0^\infty (\lambda t)^{k-1} \lambda \mathrm{e}^{-\lambda t}\,\mathrm{d}t = \mathbb{E}[(\lambda X)^{k-1}]$ for exponential X with $\lambda > 0$

卡方(chi-squared)分布

卡方分布的矩生成函数是

$$M_{\chi_n^2}(t) = (1-2t)^{-n/2}$$

- 令 $X \sim \chi_n^2$，则其概率密度函数 $f(x) = \dfrac{x^{n/2-1}e^{-x/2}}{2^{n/2}\cdot\Gamma(n/2)}$，其中 $\Gamma(n/2)$是伽马函数。

$$M_X(t) = \mathbb{E}[\exp(tX)] = \int_0^\infty \exp(tx)\cdot f(x)\mathrm{d}x = \int_0^\infty \exp(tx)\cdot\frac{x^{n/2-1}e^{-x/2}}{2^{n/2}\cdot\Gamma(n/2)}\mathrm{d}x$$

整理，$M_X(t) = \dfrac{1}{2^{n/2}\cdot\Gamma(n/2)}\displaystyle\int_0^\infty x^{n/2-1}e^{-(1/2-t)x}\mathrm{d}x$

$$\Gamma(z) = \int_0^\infty t^{z-1}\mathrm{e}^{-t}\,\mathrm{d}t$$

换元 $u = (1/2-t)x$，$\displaystyle\int_0^\infty x^{n/2-1}e^{-(1/2-t)x}\mathrm{d}x = \int_0^\infty \left(\frac{u}{1/2-t}\right)^{n/2-1}\frac{e^{-u}}{1/2-t}\mathrm{d}u$

卡方(chi-squared)分布

卡方分布的矩生成函数是

$$M_{\chi_n^2}(t) = (1-2t)^{-n/2}$$

- 令 $X \sim \chi_n^2$, 则其概率密度函数 $f(x) = \dfrac{x^{n/2-1}e^{-x/2}}{2^{n/2}\cdot\Gamma(n/2)}$, 其中 $\Gamma(n/2)$是伽马函数。

$$M_X(t) = \frac{1}{2^{n/2}\cdot\Gamma(n/2)}\int_0^\infty x^{n/2-1}e^{-(1/2-t)x}\mathrm{d}x$$

换元 $u = (1/2-t)x$, $\displaystyle\int_0^\infty x^{n/2-1}e^{-(1/2-t)x}\mathrm{d}x = \int_0^\infty \left(\frac{u}{1/2-t}\right)^{n/2-1}\frac{e^{-u}}{1/2-t}\mathrm{d}u$

$$M_X(t) = \frac{1}{2^{n/2}\cdot\Gamma(n/2)\cdot(1/2-t)^{n/2}}\int_0^\infty u^{n/2-1}e^{-u}\mathrm{d}u \quad = \frac{\Gamma(n/2)}{\Gamma(n/2)}(1-2t)^{-n/2}$$

$$\Gamma(z) = \int_0^\infty t^{z-1}\mathrm{e}^{-t}\,\mathrm{d}t$$

暗改概率？

- 抽卡出货量服从二项分布 $\mathrm{Bin}(n, p)$
- 中心极限定理：$X \sim \mathrm{Bin}(n, p)$，则 $X \xrightarrow{D} N\left(p, p(1-p)/n\right)$
- 数据量足够大，分组，每一组都近似服从未知正态分布
  - 我们有很多近似服从未知正态分布的样本 $X_1, \ldots, X_n$

正态总体参数检验

未知方差 σ^2，检验期望 μ

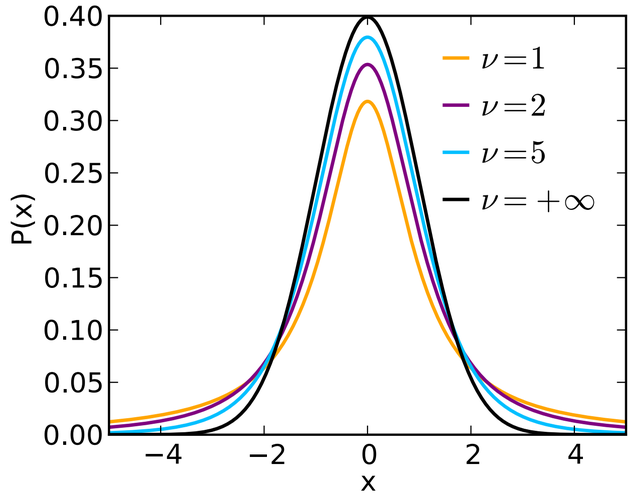
- 样本 $X_1, \ldots, X_n$ 服从某正态分布 $N(\mu, \sigma^2)$，方差 σ^2 未知。取显著性水平 α，检验 $\mu = \mu_0$

1. 原假设 $H_0 : \mu = \mu_0$，备择假设 $H_1 : \mu \neq \mu_0$ (双侧检验)

2. 检验统计量 $Z = \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}}$。若原假设成立则 $Z \sim t(n-1)$ ？ （**t检验**）

$$\bar{X} \sim N(\mu_0, \sigma^2/n), \frac{(n-1)S^2}{n\sigma^2} \sim \chi^2(n-1), \frac{(\bar{X} - \mu_0)/\sigma\sqrt{n}}{\sqrt{((n-1)S^2/n\sigma^2)/(n-1)}} = Z$$

Student's t distribution

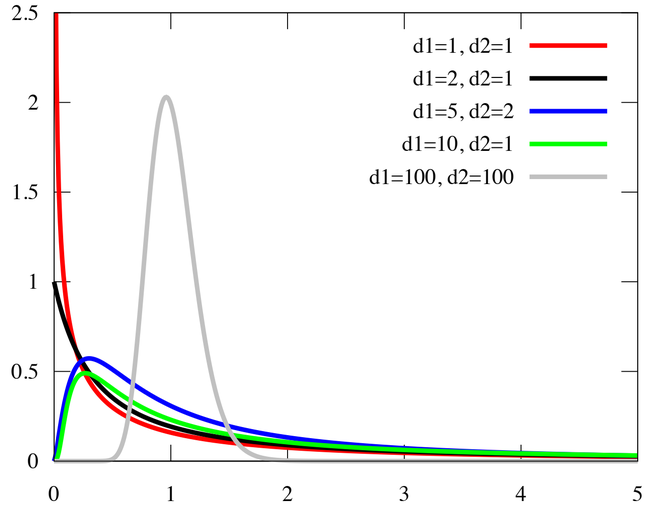


- If $X \sim N(0,1), Y \sim \chi^2(n)$ are **independent**, then

$$T = \frac{X}{\sqrt{Y/n}}$$

follows the **student's t** (学生t) distribution with n **degrees of freedom**, denoted as $T \sim t(n)$

F distribution



- If $X \sim \chi^2(n), Y \sim \chi^2(m)$ are **independent**, then

$$F = \frac{X/n}{Y/m}$$

follows the F distribution with n **and** m **degrees of freedom**, denoted as $X \sim F(n,m)$

正态总体参数检验

未知方差 σ^2，检验期望 μ

- 样本 $X_1, \ldots, X_n$ 服从某正态分布 $N(\mu, \sigma^2)$，方差 σ^2 未知。取显著性水平 α，检验 $\mu = \mu_0$

1. 原假设 $H_0 : \mu = \mu_0$，备择假设 $H_1 : \mu \neq \mu_0$ (双侧检验)

2. 检验统计量 $Z = \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}}$。若原假设成立则 $Z \sim t(n-1)$ ？ （**t检验**）

$$\bar{X} \sim N(\mu_0, \sigma^2/n), \frac{(n-1)S^2}{n\sigma^2} \sim \chi^2(n-1), \frac{(\bar{X} - \mu_0)/\sigma\sqrt{n}}{\sqrt{((n-1)S^2/n\sigma^2)/(n-1)}} = Z$$

正态总体的样本均值与样本方差

- 若 $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, 则 $\bar{X}$ 与 S^2 相互独立。
- **Proof**: 不严格证明, $f(x_1, x_2, \ldots, x_n) = (2\pi\sigma^2)^{-n/2}\exp\left(-\frac{\sum_i (x_i-\mu)^2}{2}\right)$

$$= (2\pi\sigma^2)^{-n/2}\exp\left(-\frac{\sum_i (x_i-\bar{x})^2}{2\sigma^2}\right)\exp\left(-\frac{n(\bar{x}-\mu)^2}{2\sigma^2}\right)$$

严格一点: $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$, $\bar{X} - X_j = \frac{1}{n}\sum_{i\neq j} X_i - \frac{n-1}{n}X_j$

$$\bar{X} - X_j \sim N\left(\frac{(n-1)\mu}{n} - \frac{(n-1)\mu}{n}, (n-1)\frac{\sigma^2}{n^2} + (n-1)^2\frac{\sigma^2}{n^2}\right) = N\left(0, \frac{n-1}{n}\sigma^2\right)$$

正态总体的样本均值与样本方差

- 若 $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$，则 $\bar{X}$ 与 S^2 相互独立。
- **Proof**: 严格一点：$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$，$\bar{X} - X_j = \frac{1}{n}\sum_{i \neq j} X_i - \frac{n-1}{n} X_j$

$$\bar{X} - X_j \sim N\left(\frac{(n-1)\mu}{n} - \frac{(n-1)\mu}{n}, (n-1)\frac{\sigma^2}{n^2} + (n-1)^2\frac{\sigma^2}{n^2}\right) = N\left(0, \frac{n-1}{n}\sigma^2\right)$$

两两之间的协方差：$\mathrm{Cov}(\bar{X} - X_j, \bar{X}) = \mathrm{Cov}(X_j, \bar{X}) - \mathrm{Cov}(\bar{X}, \bar{X}) = \frac{\sigma^2}{n} - \frac{\sigma^2}{n} = 0$

$$\| $$

$$\mathrm{Cov}(X_j, X_j/n) = \sigma^2/n$$

两两独立不等于相互独立，但在联合正态分布中意味着相互独立

Multivariate Normal Distribution*

- A random vector $\mathbf{Y} = (Y_1, \ldots, Y_n)$ has a multivariate normal distribution, iff there is an $\boldsymbol{A} \in \mathbb{R}^{n \times k}$, a vector $\mathbf{X} = (X_1, \ldots, X_k)$ of k independent standard normal random variables, and a vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n) \in \mathbb{R}^n$, such that

$$\mathbf{Y}^T = \mathbf{A}\mathbf{X}^T + \boldsymbol{\mu}^T$$

- If further $\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{A}^T = \mathbb{E}[(\mathbf{Y} - \boldsymbol{\mu})(\mathbf{Y} - \boldsymbol{\mu})^T]$ has full rank, then the density of $\mathbf{Y}$ is

$$f(\boldsymbol{y}) = f(y_1, \ldots, y_n) = \frac{1}{\sqrt{(2\pi)^n \det(\boldsymbol{\Sigma})}} \mathrm{e}^{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})^T}$$

- Denote $\mathbf{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Marginally, each $Y_i \sim N(\mu_i, \Sigma_{ii})$, and $\mathbf{Cov}(Y_i, Y_j) = \Sigma_{ij}$

- **Equivalent characterization**: $\forall \boldsymbol{a} \in \mathbb{R}^n$, $\langle \boldsymbol{a}, \mathbf{Y} \rangle = a_1 Y_1 + \cdots + a_n Y_n$ is normal

正态总体参数检验

未知方差 σ^2，检验期望 μ

- 样本 $X_1, \ldots, X_n$ 服从某正态分布 $N(\mu, \sigma^2)$，方差 σ^2 未知。取显著性水平 α，检验 $\mu = \mu_0$

1. 原假设 $H_0 : \mu = \mu_0$，备择假设 $H_1 : \mu \neq \mu_0$ (双侧检验)
2. 检验统计量 $Z = \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}}$。若原假设成立则 $Z \sim t(n-1)$
3. 拒绝域 $\{z : \Pr[X \leq z] \leq \alpha/2 \vee \Pr[X \geq z] \leq \alpha/2\}$, 其中 $X \sim t(n-1)$

检验比较两个正态总体

已知方差 σ_1^2, σ_2^2，检验期望差 $\mu_1 - \mu_2$

- 样本 $X_1, \ldots, X_{n_1} \sim N(\mu_1, \sigma_1^2)$，样本$Y_1, \ldots, Y_{n_2} \sim N(\mu_2, \sigma_2^2)$。取显著性水平 α，检验 $\mu_1 = \mu_2$

1. 原假设 $H_0 : \mu_1 = \mu_2$，备择假设 $H_1 : \mu_1 \neq \mu_2$ (双侧检验)

2. 检验统计量 $Z = \dfrac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$。若原假设成立则 $Z \sim N(0,1)$

3. 拒绝域： $\{z : \Pr[X \leq z] \leq \alpha/2 \vee \Pr[X \geq z] \leq \alpha/2\}$, 其中 $X \sim N(0,1)$

检验比较两个正态总体

未知方差 $\sigma_1^2 = \sigma_2^2$，检验期望差 $\mu_1 - \mu_2$

- 样本 $X_1, \ldots, X_{n_1} \sim N(\mu_1, \sigma_1^2)$，样本$Y_1, \ldots, Y_{n_2} \sim N(\mu_2, \sigma_2^2)$。取显著性水平 α，检验 $\mu_1 = \mu_2$

1. 原假设 $H_0 : \mu_1 = \mu_2$，备择假设 $H_1 : \mu_1 \neq \mu_2$ (双侧检验)

2. 检验统计量 $Z = \dfrac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right) \dfrac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}}$。若原假设成立，则 $Z \sim t(n_1 + n_2 - 2)$

3. 拒绝域：$\{z : \Pr[X \leq z] \leq \alpha/2 \vee \Pr[X \geq z] \leq \alpha/2\}$, 其中 $X \sim t(n_1 + n_2 - 2)$

Behrens–Fisher problem
(Unsolved problem in statistics)

检验比较两个正态总体

未知方差 $\sigma_1^2 \neq \sigma_2^2$，检验期望差 $\mu_1 - \mu_2$*

- 样本 $X_1, \ldots, X_{n_1} \sim N(\mu_1, \sigma_1^2)$，样本$Y_1, \ldots, Y_{n_2} \sim N(\mu_2, \sigma_2^2)$。取显著性水平 α，检验 $\mu_1 = \mu_2$

1. 原假设 $H_0 : \mu_1 = \mu_2$，备择假设 $H_1 : \mu_1 \neq \mu_2$ (双侧检验)

2. Welch's approximate *t* solution：令 $S^2 \triangleq S_1^2/n_1 + S_2^2/n_2$，$S^2$近似服从卡方分布。检验统计量 $Z = \dfrac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{S^2/?}}$

3. 近似自由度 $\ell \approx \dfrac{(g_1 + g_2)^2}{g_1^2/(n_1 - 1) + g_2^2/(n_2 - 1)}$，其中 $g_1 \triangleq S_1^2/n_1, g_2 \triangleq S_2^2/n_2$。

检验比较两个正态总体

比较方差 σ_1^2, σ_2^2

- 样本 $X_1, \ldots, X_{n_1} \sim N(\mu_1, \sigma_1^2)$，样本$Y_1, \ldots, Y_{n_2} \sim N(\mu_2, \sigma_2^2)$。取显著性水平 α，检验 $\sigma_1 = \sigma_2$

1. 原假设 $H_0 : \sigma_1 = \sigma_2$，备择假设 $H_1 : \sigma_1 \neq \sigma_2$ (双侧检验)

2. 检验统计量 $Z = \dfrac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$。若原假设成立则 $Z \sim F(n_1 - 1, n_2 - 2)$ （**F检验**）

3. 拒绝域： $\{z : \Pr[X \leq z] \leq \alpha/2 \vee \Pr[X \geq z] \leq \alpha/2\}$, 其中 $X \sim F(n_1 - 1, n_2 - 2)$

配对差异检验（paired difference test）
对比单样本检验（one-sample test）

- 现实场景往往较为复杂，没有大量理想样本
  - 比较两种肥料的效果：不同农田本身条件不同。
    - 同一块田分两半，配对比较
  - 检验药效：不同志愿者本身体质不同、病情不同。
    - 同一患者比较服药前后变化
  - 比较机器学习算法性能：不同数据集性质不同。
    - 比较每组数据在不同算法下的准确度

配对差异检验（paired difference test）

对比单样本检验（one-sample test）

- 假设机器学习算法A, B在数据集 i 上的准确度分别服从正态分布 $N(\mu_i^A, \sigma_1^2)$、$N(\mu_i^B, \sigma_2^2)$。σ_1, σ_2未知。取显著性水平 α，比较两种算法的性能 $\Sigma_i \mu_i^A / n = \Sigma_i \mu_i^B / n$?

1. 原假设 $H_0 : \Sigma_i \mu_i^A / n = \Sigma_i \mu_i^B / n$?，备择假设 $H_1 : \Sigma_i \mu_i^A / n \neq \Sigma_i \mu_i^B / n$? (双侧检验)

2. 检验统计量 $Z = \dfrac{\bar{X} - \bar{Y}}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}}$。若原假设成立，则 $Z \sim t(n_1 + n_2 - 2)$

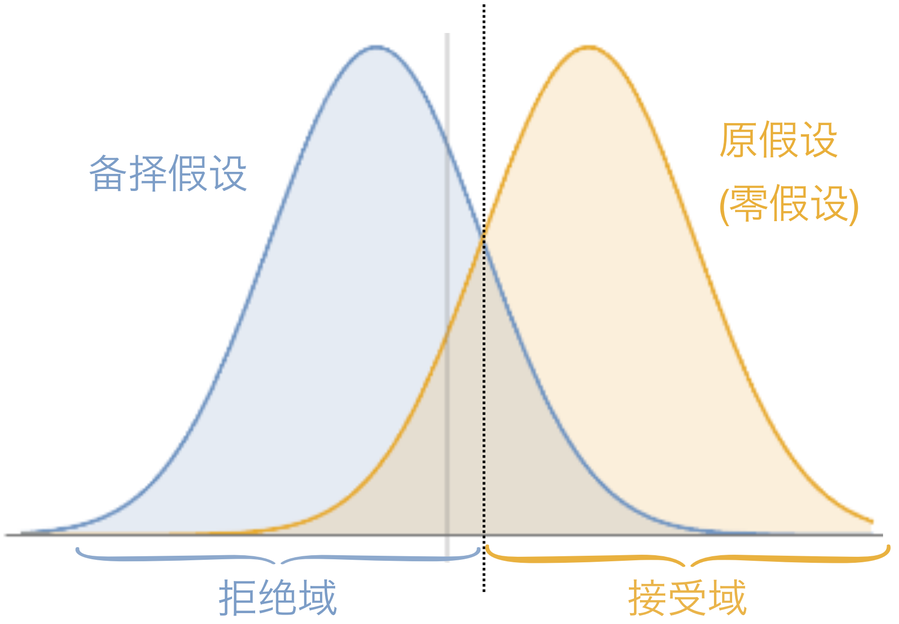
3. 拒绝域： $\{z : \Pr[X \leq z] \leq \alpha/2 \vee \Pr[X \geq z] \leq \alpha/2\}$, 其中 $X \sim t(n_1 + n_2 - 2)$

配对差异检验（paired difference test）
对比单样本检验（one-sample test）

- 现实场景往往较为复杂，没有大量理想样本
  - 比较两种不同的教学方法：
    - 不同的学生现有成绩不同，优生可能方法A更有效，差生可能应该用方法B
    - 不同知识点不一样，不能对同一个学生先后应用两种不同的方法以比较
    - 同一个学生对不同学科的天赋不同，不能对同一个学生的不同科目应用不同的教学方法
  - 复杂的场景需要更为细致的分类和精妙的实验设计

假设检验（Hypothesis test）

Neyman-Pearson’s approach



1. 提出统计假设：原假设 H_0、备择假设 H_1
2. 针对两种假设确定能区分它们的统计量
3. 规定显著性水平α
4. 根据显著性确定拒绝域和接受域
5. 从样本中计算出统计值
6. 判断统计值在拒绝域还是接受域、做出决策：“在显著性水平 α 下接受/拒绝”

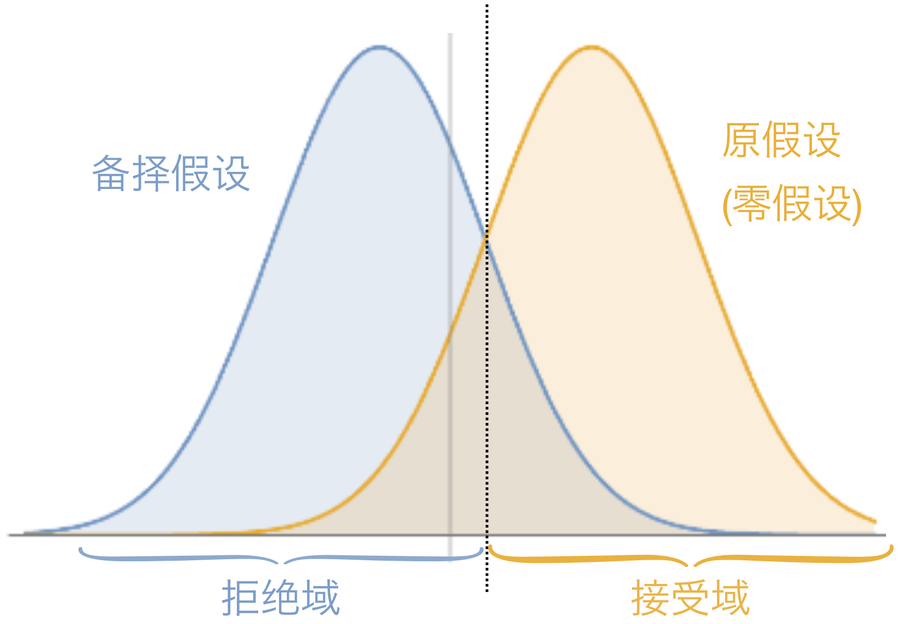
p值 (p-value / prob-value)

Fisherism

- 假设原假设为真，样本有多罕见？
- 统计假设：样本相互独立。中心极限定理：样本总体近似正态分布
  - $p \triangleq \Pr[T \geq t \,|\, H_0]$
  - $p \triangleq \Pr[T \leq t \,|\, H_0]$
  - $p \triangleq 2 \cdot \min\left\{\Pr[T \leq t \,|\, H_0], \Pr[T \geq t \,|\, H_0]\right\}$

假设检验（Hypothesis test）

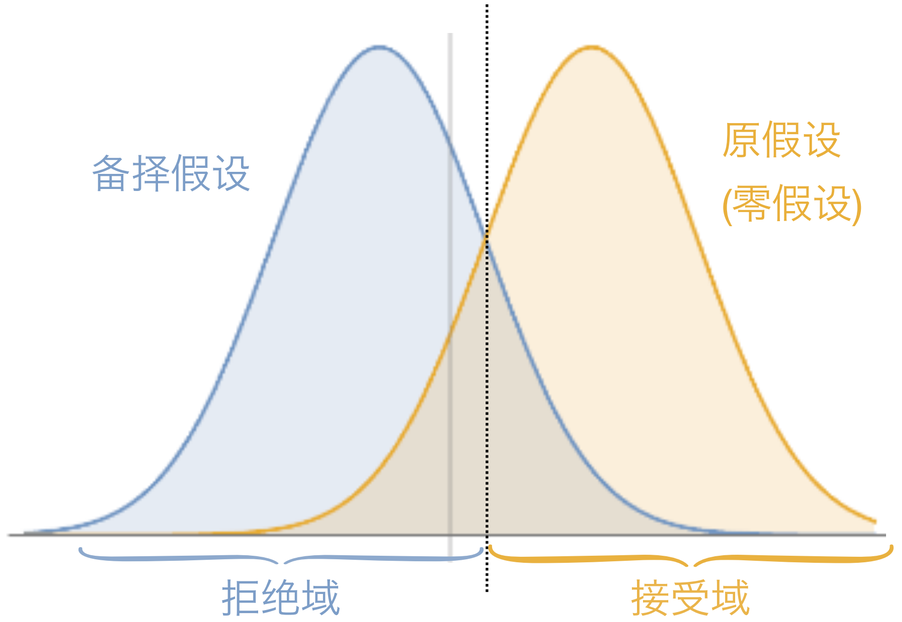
利用 p值进行检验



1. 提出统计假设：原假设 H_0
2. 规定显著性水平α
3. 采样并从样本中计算出 p值
4. 判断 p值是否超过显著性水平 α 、做出决策：“在显著性水平 α 下接受/拒绝”

假设检验（Hypothesis test）

Fisher's approach



1. 提出统计假设：原假设 H_0

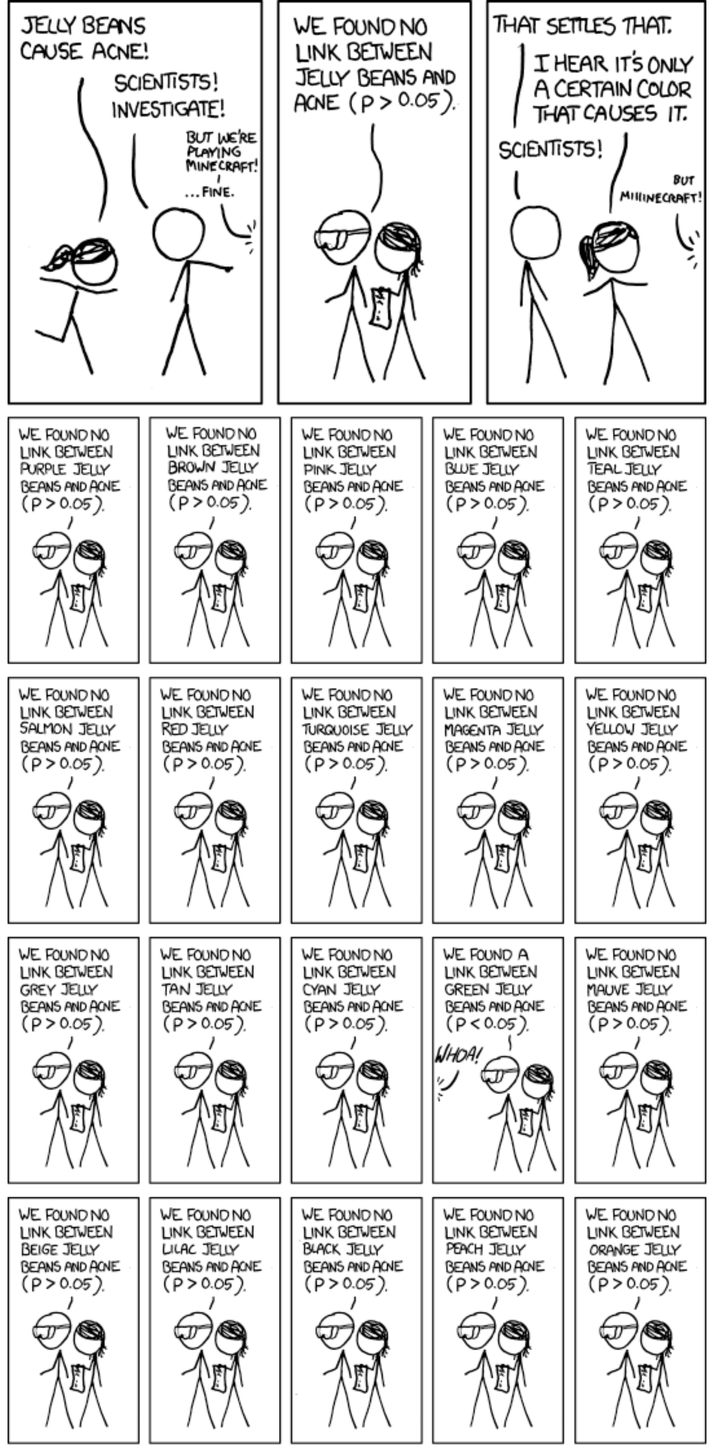
2. 采样并从样本中计算出 p值

3. 报告样本确切的 p值，而不是简单地“接受”或“拒绝”

p值的误解与误用

假设原假设为真，样本有多罕见？

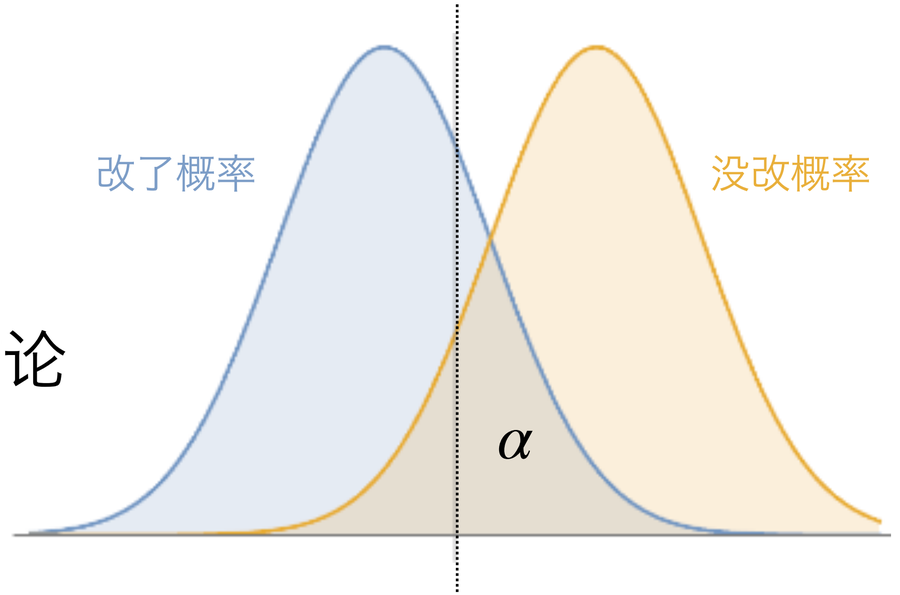
- p值小于0.01意味着原假设肯定是对的
- p值大于0.01意味着备择假设肯定是对的
- p值小于0.05意味着原假设有5%概率是对的
- p值小于0.05意味着备择假设有95%概率是对的
- p值代表犯一类错误 （弃真） 的概率
- p值越小，结论越重要。
- 在多次重复实验中可以选择性地保留那些较有意义的 p值。
- 多重比较（multiple comparisons）
  - family-wise error rate (FWER)：$1-(1-\alpha)^m$

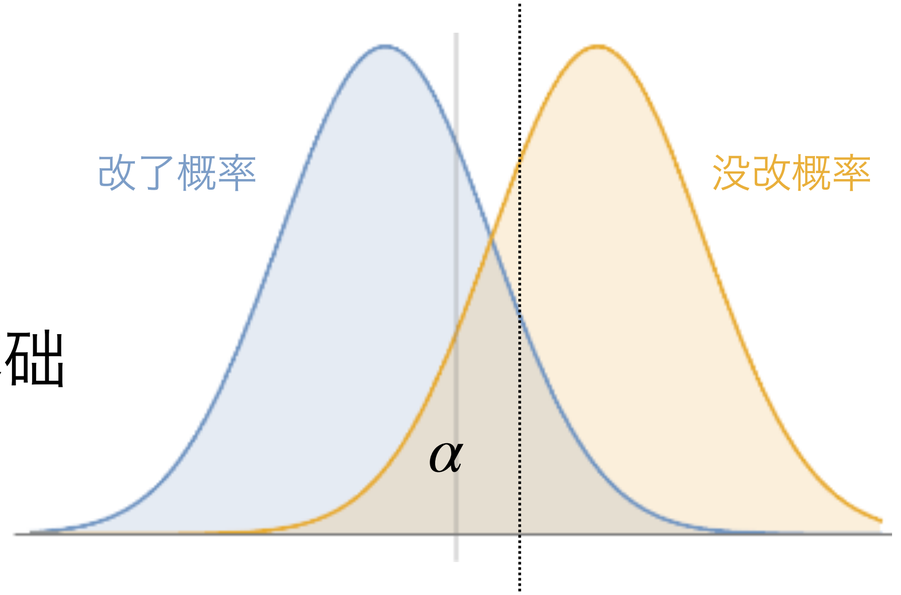


原假设 / 零假设

Null hypothesis

- 将不同的假设作为原假设，可以得出完全相反的结论
- 零假设反应了实验者的倾向
  - 同样的数据在不同倾向下有不同的解释
  - 选择较为中立或者保守的零假设
- 零假设应该是清晰的，必须为分析概率分布提供基础
- 尽可能考虑多种不同的零假设





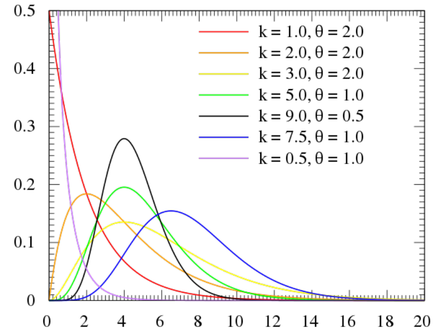
非正态总体的参数检验*

- 样本 $X_1, \ldots, X_n \sim \mathrm{Exp}(\lambda)$。取显著性水平 α, 检验 $\lambda = \lambda_0$。

1. 原假设 $H_0 : \lambda = \lambda_0$

2. 检验统计量 $\bar{X}$: $\mathbb{E}[\bar{X}] = 1/\lambda$.

Gamma Distribution*



- The random variable X has the **gamma distribution** with parameters $k, \lambda > 0$, denoted $\Gamma(k, \lambda)$ or $\mathrm{Gamma}(k, \lambda)$, if it has the density

$$f_X(x) = \frac{1}{\Gamma(k)}(\lambda x)^{k-1} \cdot \lambda \mathrm{e}^{-\lambda x}, \quad \text{for } x \geq 0, \quad \text{where } \Gamma(k) = \int_0^\infty t^{k-1}\mathrm{e}^{-t}\,\mathrm{d}t$$

- $\Gamma(1,\lambda)$ is the exponential distribution with parameter λ
- $\Gamma(k/2,1/2)$, for integer $k \geq 1$, is the $\chi^2(k)$ distribution
- If $X \sim \Gamma(\alpha, \lambda)$ and $Y \sim \Gamma(\beta, \lambda)$ are independent, then $X + Y \sim \Gamma(\alpha + \beta, \lambda)$
  - $\sum_{i=1}^{k} X_i \sim \Gamma(k, \lambda)$ if $X_1, \ldots, X_k$ are i.i.d. exponential random variables with parameter λ

非正态总体的参数检验*

- 样本 $X_1, \ldots, X_n \sim \mathrm{Exp}(\lambda)$。取显著性水平 α，检验 $\lambda = \lambda_0$。

1. 原假设 $H_0 : \lambda = \lambda_0$

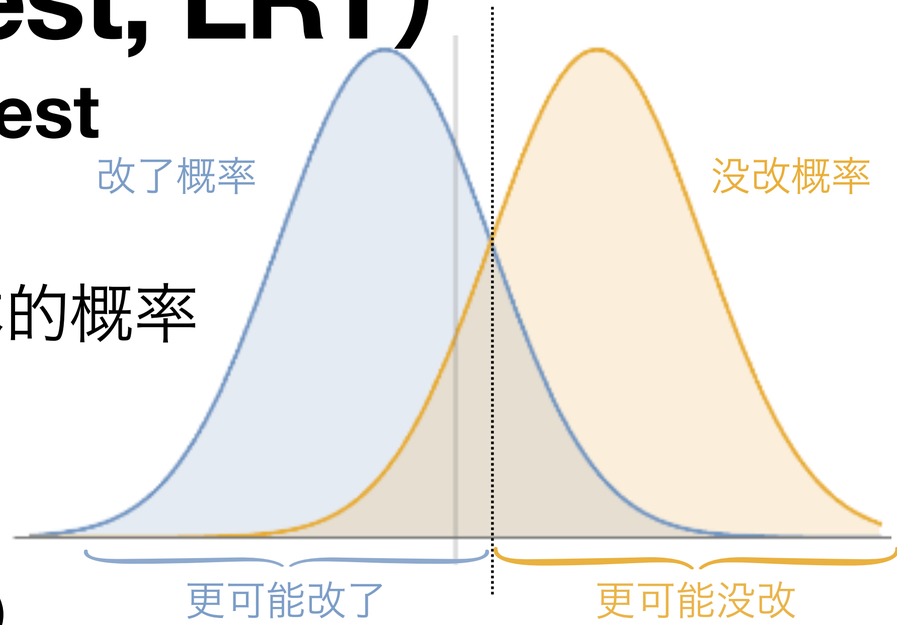
2. 检验统计量 $\bar{X}$: $\mathbb{E}[\bar{X}] = 1/\lambda$.

   - $\mathrm{Exp}(\lambda) = \Gamma(1,\lambda)$, $\chi^2(n) = \Gamma(n/2,1/2)$。因此 $n\bar{X} \sim \Gamma(n,\lambda)$, $2\lambda n\bar{X} \sim \Gamma(n,1/2) = \chi^2(2n)$

3. p值: $2 \cdot \min\left\{\Pr[\chi^2 \le 2\lambda n\bar{X}], \Pr[\chi^2 \ge 2\lambda n\bar{X}]\right\}$, 其中 $\chi^2 \sim \chi^2(2n)$

似然比检验（likelihood ratio test, LRT）

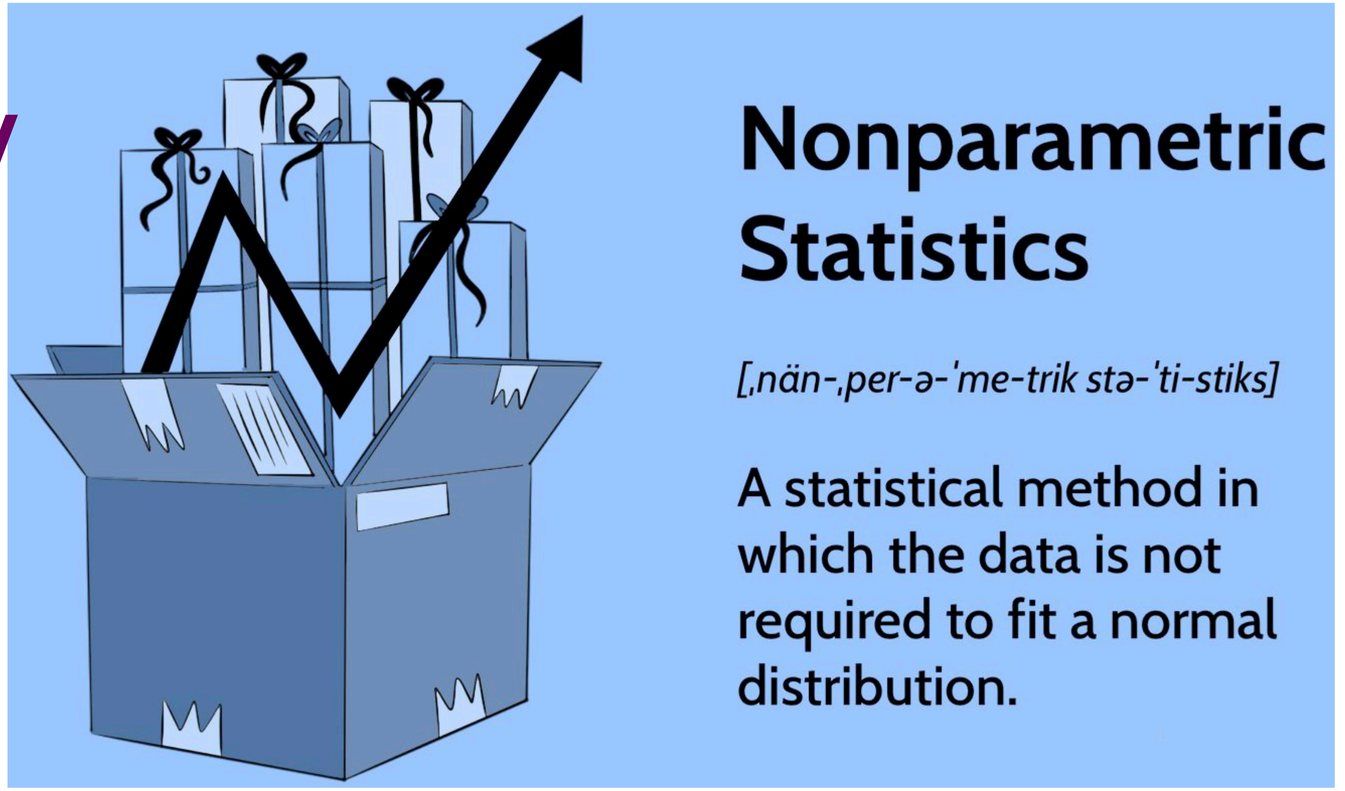
Wilks test / Lagrange multiplier test / Wald test



- 似然函数 $L(x;\theta_0), L(x;\theta_1)$ 分别表示两种假设下样本的概率
- $L(x;\theta_0) \geq (\leq) L(x;\theta_1)$ 则更应该支持 $H_0(H_1)$
  - $L(x;\theta_0)/L(x;\theta_1) \geq (\leq) 1$ 则更应该支持 $H_0(H_1)$
- 显著性 α 和功效 $1-\beta$ 也可以写成这样比值的形式
  - 可以根据似然比来接受/拒绝原假设
- **Neyman-Pearson 引理**: 给定显著性 α，LRT是功效 $1-\beta$ 最高的检验方法

非参数假设检验

Non-Parametric Test / Distribution-Free Test



拟合优度检验（goodness of fit）

- 德军轰炸分布是否服从泊松分布？期末成绩得分是否符合正态分布？

1. 将所有可能的结果分成不相交的事件 $\mathscr{E}_1, \ldots, \mathscr{E}_m$
2. (可选)通过点估计确定猜测分布的部分参数。假设服从分布 $p(\mathscr{E}_i)$
3. 检验统计量：$\chi^2 = \sum_{i=1}^{m} \frac{\left(p(\mathscr{E}_i) - E_i\right)^2}{p(\mathscr{E}_i)}$，$E_i$ 是样本中事件 $\mathscr{E}_i$ 发生的频率

Pearson's chi-squared test

- 考虑二项分布 $X \sim \mathrm{Bin}(n,p)$

  根据中心极限定理, $Z = \dfrac{X-np}{\sqrt{np(1-p)}} \xrightarrow{D} N(0,1)$ 因此 $Z^2 \xrightarrow{D} \chi^2(1)$。

  令 $p_1 = p, \quad p_2 = 1-p, Y = n - X\,.$

  注意到 $Z^2 = \dfrac{(X-np_1)^2}{np_1} + \dfrac{(Y-np_2)^2}{np_2} \xrightarrow{D} \chi^2(1)$

拟合优度检验（goodness of fit）

Pearson’s chi-squared test

- 德军轰炸分布是否服从泊松分布？期末成绩得分是否符合正态分布？

1. 将所有可能的结果分成不相交的事件 $\mathscr{E}_1, \ldots, \mathscr{E}_m$
2. (可选)通过点估计确定猜测分布的部分参数。假设服从分布 $p(\mathscr{E}_i)$
3. 检验统计量：$\chi^2 = \sum_{i=1}^{m} \frac{\left(p(\mathscr{E}_i) - E_i\right)^2}{p(\mathscr{E}_i)} \sim \chi^2(m - k - 1)$，目标分布未定参数数目 k
4. p值/显著性/接受/拒绝

独立性检验（statistical independence）
列联表（contingency table / cross tabulation / crosstab）

- 两种因素对样本有影响吗？数据样本的特征和它的分类是否相互独立？

1. 根据 r 种分类和 c 种特征建表，计数 $r \times c$ 种样本的数目

2. 从样本中估计分类与特征的分布函数 $p_{i\cdot}$ 、 $p_{\cdot j}$

3. 若分类与特征相互独立，则 $n_{ij} \approx np_{i\cdot}p_{\cdot j}$ 。检验统计量

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij} - np_{i\cdot}p_{\cdot j})^2}{np_{i\cdot}p_{\cdot j}}$$

4. 若相互独立，则统计量 $\chi^2 \approx \chi^2((r-1)(c-1))$。 p值/显著性/接受/拒绝。

同质性检验（statistical homogeneity）

列联表（contingency table / cross tabulation / crosstab）

- 两种因素对样本有影响吗？两种机器学习算法在不同数据集上有性能差异吗？

1. 根据 2 种算法和 c 个数据集建 $2 \times c$ 的表
2. 从样本中估计每组数据的正确率 p_j
3. 若没有性能差异，则 $n_{1j} \approx n_{2j} \approx p_j$。检验统计量 $\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{c} \frac{(n_{ij} - p_j)^2}{p_j}$
4. 若相互独立，则统计量 $\chi^2 \approx \chi^2(c-1)$。 p值/显著性/接受/拒绝。

符号检验（sign test）

- 检验总体的中位数 $m = m_0$？ 检验上了《数据科学基础》的学生GPA变高?

  1. 对每一个样本 $X_1, \ldots, X_n$ 检验是否 $X_1 < m_0$，并记录 - / +
  2. 若原假设成立，则 #- $\sim \mathrm{Bin}(n, 1/2)$
  3. p值/显著性/接受/拒绝

---

  1. 对每一个受试者检验修课前后GPA变化，并记录 - / +
  2. 若原假设成立，则 #+ $\sim \mathrm{Bin}(n, p)$，其中 $p > 1/2$
  3. p值/显著性/接受/拒绝

秩和检验（rank-sum test）

Mann–Whitney–Wilcoxon U test

- 两个总体 $\mathscr{D}_1, \mathscr{D}_2$ 是否差不多: $\Pr[X > Y] = \Pr[Y > X]$?其中 $X \sim \mathscr{D}_1, Y \sim \mathscr{D}_2$
  1. 采样样本 $X_1, \ldots, X_{n_1} \sim \mathscr{D}_1, Y_1, \ldots, Y_{n_2} \sim \mathscr{D}_2$
  2. 将 $X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2}$ 按大小顺序排序，并记录每个样本的排名
  3. 检验统计量 $U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1, U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$。$R_1, R_2$ 分别表示两组样本的排名之和，且 $U_1 + U_2 = n_1 n_2$。
  4. 若原假设成立，检验 $\min\left\{U_1, U_2\right\}$。小样本时查表，大样本时中心极限定理
  5. p值/显著性/接受/拒绝

符号秩检验（Wilcoxon signed-rank test）

Wilcoxon T-test

- 样本$X_1, \ldots, X_n$来自总体 $\mathscr{D}$，检验 $\mathscr{D}$ 是否关于 0 轴对称。

1. 将样本按照 $|X_1|, \ldots, |X_n|$ 排序
2. 检验 $T^+ \triangleq \sum_{i:X_i>\theta} R_i$，其中 R_i 表示 X_i 在上述排序中的排名
3. 若原假设成立：当 n 较小时，查表；当 n 较大时，中心极限定理: $T^+ \approx N$

   $$\mathbb{E}[T^+] = n(n+1)/4 \; ; \; \mathrm{Var}(T^+) = n(n+1)(2n+1)/24$$

4. p值/显著性/接受/拒绝