1 Estimation

对于一个总体, 当然这个总体满足一定的分布 $X \sim \mathcal{D}(\theta)$, 其中 X 是随机变量, $\mathcal{D}(\theta)$ 是任意一种参数分布, θ 是这个分布的参数。如果我们知道总体的所有单位的观测值, 那么这个分布的参数就已经是确定的了。例如正态分布 $X \sim \mathcal{N}(\mu, \sigma^2)$，那么自然有

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i, \quad \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2,$$

问题在于, 我们拿到的只是 n 个样本的抽样总体, 因此我们的对参数的估计是不准确的（因为我们只能拿到部分的信息)。

1.1 点估计

估计不准确? 去他妈的, 我管你那么多, 我能给你估计一个合理的参数就不错了。所谓估计, 那本质上就是我现在有一堆抽样观测值 x_i, 然后把他们搞在一起, 弄成一个值, 然后把这个值作为我估计的值。所以可以说点估计本质上就是用一个统计量来估计总体参数。用观测值构造的统计量最简单当然是我们最爱的**矩**啦！然后用样本的矩来估计总体的矩，美哉美哉!

好，没错，你估计出总体的矩了，这有什么用? 你最终估计出总体分布的参数了吗? 你说，估计出来了，正态分布的两个参数分别就是一阶原点矩（总体均值）和二阶中心矩（总体方差)，这不就估计出来了? 我说，那对于那些参数不是啥啥矩的的分布呢，阁下当如何应对?

你说，客官莫急，我自有锦囊妙计！你不等我反应便当即提问：试问我们最终要做什么事情?——“估计出总体分布的参数”。那我随便把观测数据搞在一起弄一个值出来，算不算一个估计?

我想了想，确实算。不对不对，不能算，你这也太随便了吧，至少得弄个准一点的估计吧!

哈哈哈，你脸上暗自窃笑，追问道，那怎么才称得上“准”呢?

我搅拌脑汁，突然冒出灯泡——你看，所谓分布，我可以想象成随机变量的一种分布形状。假设现在我已经确定我的参数了，那么这个形状就确定了，而抽样分布也形成一种分布的形状。我只需要比较一下这两者的匹配程度即可! 我给你举个抛硬币的例子，假设这枚硬币正面朝上的概率为 p，反面朝上的概率为 $1-p$，那么这个分布的形状就是 $(p, 1-p)$。假设现在我投掷了 10 次，有 7 次正面，3 次反面，此时我观测的形状就是 $(0.7, 0.3)$。而我只需要调整我的待估计参数 p，使得其形状能够匹配上 $(0.7, 0.3)$，这样就可以说我估计出来的 p 是“准”的估计!

哈哈哈哈，你大笑，不错不错，问题在于，什么是你衡量分布“匹配度”的标准呢?

1.1.1 MDE

我想到了一个最简单的：就是对每个随机变量的取值都计算其距离：

$$L(\theta) = \sum_i |P(X = x_i|\theta) - p_i|^2.$$

如果把分布写成向量的形式，总体分布向量 $\boldsymbol{x}$（是一个与参数 θ 有关的向量）和抽样分布向量 $\hat{\boldsymbol{x}}$ 的距离

$$L(\theta) = \|\boldsymbol{x} - \hat{\boldsymbol{x}}\|^2,$$

所以估计的参数 θ 就是最小化这个距离

$$\theta = \arg\min_\theta \|\boldsymbol{x} - \hat{\boldsymbol{x}}\|.$$

此外，既然可以最小距离，那也可以直接最大化相似度（点积）

$$\theta = \arg\max_\theta \boldsymbol{x} \cdot \hat{\boldsymbol{x}}.$$

1.1.2 MLE

如果不用向量这种图形化的语言来观察两个分布，而是这样想：观测值是总体分布的抽样结果，那么就可以算出这个结果出现的概率！假设只抽一个样本，那么这个样本的分布就是服从整体的分布，例如 $x \sim \mathcal{B}(1, p)$, 那么结果要么是 1, 要么是 0. 假设结果为 1, 则可以计算出现这个结果的概率

$$P(X = 1|p) = p^1 \cdot (1-p)^0 = p,$$

如果一次抽样有多个独立同分布的样本，那么抽样的结果应该是一个**抽样分布**。在这个投掷硬币的例子中，抽样分布是一个伯努利分布。设随机变量 Y 表示抽样结果中正面朝上的个数，则 7 次正面、3 次反面的概率

$$P(Y = 7|p) = \binom{10}{7} \cdot p^7 \cdot (1-p)^3 = 120p^7 \cdot (1-p)^3 \equiv L(p),$$

记为**似然函数** $L(p)$, 估计参数就是最大化使这个结果出现的可能性

$$p = \arg\max_p L(p) = \arg\max_p p^7 \cdot (1-p)^3.$$

当然为了求其最大值，使用取对数求导的方法:

$$\begin{aligned}(\ln L(p))' &= (7\ln p + 3\ln(1-p))' \\ &= \frac{7}{p} + \frac{3}{1-p},\end{aligned}$$

另之等于 0, 可以解得 $p = 0.7$, 美哉快哉！

1.1.3 KL 散度

我感觉我泉思喷涌，因为一想到我要做的实际实际上是调整参数 θ 使得总体分布 $F(\theta)$ 和样本分布 $\hat{F}$ 的“匹配度”最大，我就想到了更多的方法！这里我要说的就是 KL 散度。

KL 散度是从信息论的角度出发来定义分布 B 相对于分布 A 的信息损失量（信息熵）。回顾一下**信息量**的定义，对于一个系统，如果一件事件发生的概率为 p，则其信息量为 $-\log p$。这样的定义满足了这样几个性质：当事件发生的概率为 1 时，我们知道它一定会发生，因此它发生与否都没有带来更多的信息，因此其信息量为 0; 当事件发生的概率趋于 0 时，一个几乎不可能发生的事情竟然发生了，那么他的信息量应该趋于无穷大。**信息熵**衡量了一个随机变量的平均信息量，即 $H(X) = -\sum_i p(x_i) \ln p(x_i)$. 这样你应该能理解 KL 散度的定义了：

$$D_{\mathrm{KL}}(A \| B) = \sum_i A(x_i) \cdot \ln\left(\frac{A(x_i)}{B(x_i)}\right).$$

1.1.4 卡方距离

与 MDE 直接算平均距离不同，卡方距离是“相对”距离，或者更准确地说应该是相对频次。样本量为 n理论 x_i 的频次应该是 np_i，实际统计出的频次是 f_i，则其卡方距离

$$\chi^2 = \sum_i \frac{(f_i - np_i)^2}{np_i}.$$

记住我们给出这些统计量的定义是为了进行点估计，效果可嘉。但是上述方法你不觉得有点随便吗? 哪里随便了，我问。你看，你是利用样本的观测值来估计总体的参数，本质上是用样本观测值的一个统计量来估计总体参数，那么这个统计量肯定服从一个抽样分布吧，在 MLE 的时候已经提到了。我说，对啊，那咋了? 呵呵，要是这个分布很狗屎呢，比如很偏、方差很大之类? 嗷嗷，有道理啊！那怎么办呢? 你自信微笑，没事儿，一个值直接估计固然有些勉强，对于这种拿不准的事情，我们只需要油滑一些——给出一个估计的区间就行！

1.2 区间估计

区间估计的想法其实很简单。假设现在老板说：啊，小伙子，给我估计一下这个分布的参数，我允许你有 5% 的误差，要有九五成的把握，咳咳。你想想，你要是只丢给老板一个值的话，这包不准的啊，你估计完全正确的概率是 0, ZERO！那咋办? 你灵机一动，那我不给一个值了，而是给出一个估计的区间，这样真实值总有概率落在我的区间里面嘛！而我呢，只需要精密地调整我的区间位置，使让它尽量小一点，并且真实值落在里面的概率要高于九五成即可！显然区间越小容错越小，那么我要做的就是调整区间使得真实值落在里面的概率恰好是 95%。

问题来了，啥是“真实值落在我区间的概率”? 我区间不是给定的吗，而且真实值不也是确定的吗? 这不都是确定的，何来概率一说? 嘿嘿，你微微一笑，这就是随机变量的确定的观测值和其随机性的辨证统一关系的体现了。啥意思，我问? 就是说，你以为你的区间是给定的，但其实它也是某种随机变量，只需要注意到你通过样本搞出来的统计量也是一个随机变量！我们以这个统计量 $Y = y$ 作为中点，以一定的误差值 E（边际误差）左右延展构成最终的估计区间 $[Y \pm E]$, 那么这个区间也具有随机性。而我们之前

说的“真实值落在区间的概率”是哪个随机变量的概率呢? 是**“真实值落在区间中”**，也就是 $|\theta - Y| \leq E$，也就是

$$P\{|\theta - Y| \leq E\} = 1 - \alpha.$$

注意看⚠，由于统计量 Y 服从一个抽样分布，所以其累计密度函数 F 是可以得到的，我们要做的就是调整 E 的值，使得上述等式成立即可，哦耶✌！当然，为了更好得使用标准的分布，我们通常会构造一个**枢轴量**。例如对于估计正态分布的参数 μ，在已知 σ^2 的情况下，我们直接使用统计量 $\overline{X}$ 来估计，也就是

$$P\{|\mu - \overline{X}| \leq E\} = 1 - \alpha$$

然后把它搞成正态分布参数的形式（由于 $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$）

$$P\left\{\left|\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}\right| \leq \frac{E}{\sigma/\sqrt{n}}\right\} = 1 - \alpha$$

即

$$P\left\{\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq -\frac{E}{\sigma/\sqrt{n}}\right\} \equiv F\left(-\frac{E}{\sigma/\sqrt{n}}\right) = \frac{\alpha}{2},$$

这里

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

就是我们构造的枢轴量，是一个很标准的分布，主要就是为了方便。

然后就是为了快速确定 E 的值, 统计学家搞出来了个**“分位数”**的概念, 就不用写累计分布函数的庞杂的式子了。分位数的定义很简洁，对于累计分布函数 $F(x) = P\{X \leq x\}$，给定一个累计概率 α，满足 $F(x_\alpha) < \alpha$，就称 x_α 是随机变量 X 的 α 分位数。由上面那个式子，可以显然得到

$$-\frac{E}{\sigma/\sqrt{n}} = z_{\alpha/2} \quad \Longrightarrow \quad E = -\frac{\sigma}{\sqrt{n}} z_{\alpha/2}.$$

所以最后估计的区间（**置信区间**）为

$$\left[p \pm \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\right]$$

1.3 评价估计量的标准

上述点方案都是通过“准”的方式选择估计量，但是评价一个估计量的好坏还不至于此，下面是一些其他的指标:

Definition 1.1 无偏性

估计量的数学期望等于被估计的总体参数

Definition 1.2 有效性

对同一总体参数的两个无偏点估计量，有更小标准差的估计量更有效。

Definition 1.3 一致性

随着样本容量的增大，估计量的值越来越接近被估计的总体参数。

	待估参数	其他参数	枢轴变量及其分布

1.4 样本容量的确定

2 Hypothesis Test

假设检验的思想同样是一种“油滑”的思想😎，还是拿牛马的例子来说。现在老板有一批货，这批货有一个指标 X 满足正态分布 $X \sim \mathcal{N}(\mu, \sigma^2)$。现在你统计了这批货的平均指标 $\bar{x}$，发现 $\bar{x} \neq \mu$，因此老板怀疑这批货的指标有问题。不会有问题的老板！你说。咋可能，明明你统计出来的值和我想要的不一样。你娓娓道来，老板，你看，这个货物的指标并不是一个固定的值，而是在你想要的值 μ 左右波动。你看，现在我这批货有 n 个样本，假设这批货真的有问题没达到您想要的指标，那我们就假设它的指标为 μ_0，并且我们假定这种波动的效果是一样的（就是同方差）。假设每个样本独立同分布，那么我统计出的平均指标 $\overline{X}$ 是一个随机变量，并且也服从正态分布 $\overline{X} \sim \mathcal{N}\left(\mu_0, \frac{\sigma^2}{n}\right)$。这就是说平均指标也是一个上下波动的值，老板你看，你说我这批货指标有问题是不是太过武断了😁？

老板作深思状🤔，确实，那你这样一说你的指标随便取什么值都有可能一样，这显然不合适吧？如果这个偏差太离谱了话，怎么看你这批货都有问题啊！

你顺势说道，没错老板，就是这个意思，只要我的偏差不要太离谱，那我都可以接受这批货的指标是没问题的。因此，我先假设我的货就是没问题的（这样就把我可接受的的分布确定了），即假设

$$H_0 : \mu = \mu_0,$$

那么我这批货应该服从均值为 μ 的正态分布 $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$，即 $\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$，这是一个已知的分布。我们可以设置一个离谱值 Δ，只要最终我的结果在 $[\mu - \Delta, \mu + \Delta]$ 内，那我就可以接受这批货的指标是没问题的。假如我们设定的离谱值使得我这批货统计出的平均指标太离谱的概率小于一个水平 α，也就是犯错误的概率为 α，即

$$P\{|\overline{X}-\mu|>\Delta\}<\alpha,$$

那么我们就有 $1-\alpha$ 的把握说我们这批货的指标确实是没问题的。

2.1 第 I 类错误和第 II 类错误

好，你假设了 $\mu=\mu_0$，但如果我假设

$$H_1:\mu\neq\mu_0$$

怎么办? 也就是说现在有两种可能的情况——要么我这批指标货没问题（H_0），要么有问题（H_1）。当我们判断我们这批货到底有没有问题时，我们可能判断错❌，也可能判断对✓。于是我们可以组合出下面四种情况

	没问题	有问题	
没问题	✅正确判断 $1-\alpha$	❌第 I 类错误 α	判断结果
有问题	❌第 II 类错误 β	✅正确判断 $1-\beta$	
	真实情况		

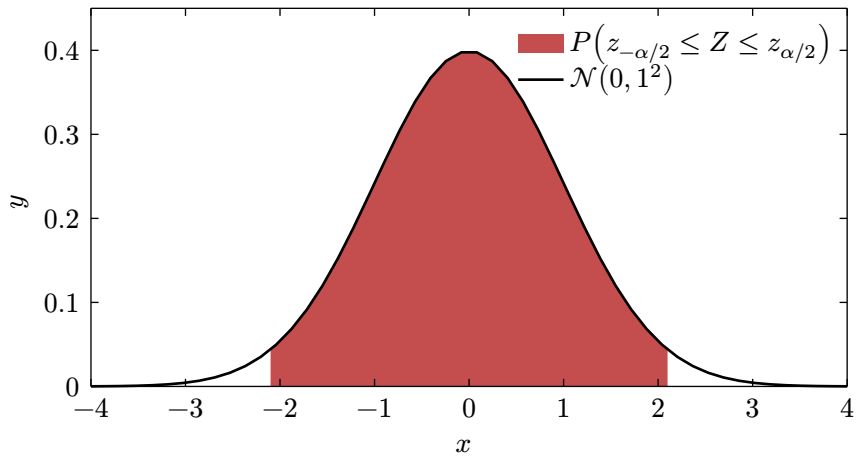
其中 α 和 β 分别为**弃真错误**和**取伪错误**允许的概率。假设检验的方法是承担第 I 类错误的风险（通过设定 α 值），但没办法承担第 II 类错误的风险，因为通常我们没法知道备择假设成立时统计量的分布。

2.2 检验的一般步骤

我们把上面的式子先标准化一下

$$P\left\{Z\equiv\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}<-\frac{\Delta}{\sigma/\sqrt{n}}\equiv z_{\alpha/2}\right\}<\frac{\alpha}{2}. \tag{2.1}$$

这个式 (2.1) 展示了下图中的红色区域。



假设平均值的观测值 $\bar{x}$ 落在这个红色区域内，即

$$\frac{\overline{x}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2},$$

则说明我们这批货的平均指标还算正常（95% 的把握认为）；如果超出了这个值，说明假设 H_0 就不成立了（黑话说的是“拒绝了原假设”），说明我们这批货质量确实不行。

⚠注意上述的推导是基于双侧置信区间（双尾检验）。

当然，你也可以反过来做，也就是拿着现有的统计值 $\overline{x}$，计算出 $z = \frac{\overline{x}-\mu}{\sigma/\sqrt{n}}$ 值，然后看他是否在红色区域内，即先计算

$$p = P\{|Z| > z\}$$

这个值代表了比我们现在的值还离谱的概率。比较 p 值与 α 的大小，如果 $p < \alpha$ 说明比我们还离谱的值太少了，比老板允许的离谱值 α 还少，说明我们这批货质量确实不行。

2.3 卡方检验

还记得小节 1.1.4 中我们定义了一个卡方距离吗？也就是实际频次相对于理论频次的“吻合程度”。在 Z 检验中，我们通过给出样本统计量的抽样分布，通过假定原假设成立来假设该统计量的分布参数，通过给出允许的弃真容错概率 α 来计算统量的观测值是否在拒绝域中。走完上述流程，我们检验了统计量的观测值是否有 $1-\alpha$ 的把握是没问题的。在我们现在要说的卡方检验中，这个统计量服从卡方分布，也就是我们的主角——卡方距离。

Definition 2.1 **卡方分布**

n 个独立同分布于标准正态分布的随机变量 $X_1, X_2..., X_n$ 的平方和 $\sum_{i=1}^{n} X_i^2$ 构成的一组新的随机变量，其分布规律称为卡方分布 $\chi^2(n)$。

卡方统计考察的是这样一个问题：设总体中类别 i 的概率为 p_i，满足 $\sum_{i=1}^{k} p_i = 1$. 从该总体中抽取 n 个独立观测，记观测频数为 f_i，满足 $\sum_{i=1}^{k} f_i = n$. 检验的原假设

$$H_0 : p_i = \pi_i, \quad i = 1, 2, ..., k$$

其中 π_i 是事先给定的理论概率。下面给出卡方统计量的定义。

Definition 2.2 **卡方统计量**

卡方统计量定义为

$$X^2 = \sum_{i=1}^{k} \frac{(f_i - n\pi_i)^2}{n\pi_i},$$

其中 f_i 是观测频次，$n\pi_i$ 是理论频次。若原假设成立，则统计量 X^2 服从自由度为 $k-1$ 的卡方分布 $\chi^2(k-1)$。

容易证明卡方统计量服从自由度为 $k-1$ 的卡方分布，只需要注意到进行 n 次实验，k 个互斥类别出现的频数 $(f_1, f_2, ..., f_k)$ 服从自由度为 $k-1$ 的多项分布。当实验次数 n 较大时，每个观测频数 f_i 近似服从正态分布，并且

$$\mathbb{E}(f_i) = n\pi_i, \quad \mathrm{Var}(f_i) = n\pi_i.$$

可以标准化

$$Z_i = \frac{f_i - n\pi_i}{\sqrt{n\pi_i}} \sim \mathcal{N}(0,1),$$

因此

$$X^2 = \sum_{i=1}^{k} Z_i^2 \sim \chi^2(k-1).$$

若理论分布 π_i 中有 r 个参数由参数需由样本估计，则自由度减少为 $k-1-r$，证明思路类似，但需考虑参数估计带来的约束。

除此之外，任何可以被构造为卡方分布的统计量都可以使用卡方检验，例如：
1. 检验某个连续变量的分布是否与理论分布一致；
2. 检验某个分类变量各类的出现概率是否等于指定概率；
3. 检验某两个分类变量是否相互独立。如吸烟是否与呼吸道疾病有关；
4. 检验控制某种或某几种分类因素的作用以后，另两个分类变量是否相互独立。

2.3.1 卡方拟合优度检验

也就是假定理论的分布（或者分类变量每个类别的概率），做 n 次实验，统计每个值（或每个类别）出现的频率。

$$H_0 : \text{观察分布等于期望分布}$$

值（类别）	期望频数	实际频数
1	20	18
2	20	19
3	20	23
⋮	⋮	⋮

自由度 $\mathrm{df} = k - 1$。

2.3.2 卡方独立性检验

独立性检验只是多了一个步骤，也就是需要计算理论分布，再按照上面同样的步骤检验。参考方差分析中因子的概念 Definition 3.1，卡方拟合优度检验只有一个因子（或者说一种分类方式），可以直接构造卡方统计量

$$X^2 = \sum \frac{(A-E)^2}{E}.$$

但是当有多个因子时，假设每个类别的理论分布（期望频数）都是固定的，那么我们可以直接写成笛卡尔积的形式：

值（类别）	期望频数	实际频数
(1,1,1)	20	18
(1,1,2)	20	19
(1,2,1)	20	23
⋮	⋮	⋮

然后我们又可以美汁汁儿计算卡方统计量来检验实际频数是否拟合期望频数了，嘻嘻😁。

问题是在独立检验中，我们压根不知道理论频数，我们只知道分类变量的实际频数！

A\B	1	2	A 的实际频数
1	43	96	139
2	28	84	112
B 的实际频数	71	180	总计 251

好在我们并不需要知道真正的理论分布，别忘了我们的目的是判别这两个分类自变量是否独立！因此我只可以直接假设因子的期望频数就是实际频数，即

$$P\{A=1\} = \frac{139}{251}, \quad P\{A=2\} = \frac{112}{251};$$
$$P\{B=1\} = \frac{71}{251}, \quad P\{B=2\} = \frac{180}{251}.$$

给出我们的原假设

$$H_0 : \text{A 与 B 独立}$$

那就应当满足

$$P\{A=i \wedge B=j\} = P\{A=i\} \cdot P\{B=j\}$$

这样我们就可以计算出每个因变量的理论频数

A\B	1	2
1	$251 \times \frac{139}{251} \times \frac{71}{251} \approx 39.32$	$96 \times \frac{139}{251} \times \frac{180}{251} \approx 99.68$
2	$28 \times \frac{112}{251} \times \frac{71}{251} \approx 31.68$	$84 \times \frac{112}{251} \times \frac{180}{251} \approx 80.32$

这样我们就得到了

(A,B)	期望频数	实际频数
$(1,1)$	43	39.32
$(1,2)$	96	99.68
$(2,1)$	28	31.68
$(2,2)$	84	80.32

自由度为行数-1✖列数-1：

$$\mathrm{df} = (R-1)(C-1) = 1.$$

2.4 Z 检验

2.4.1 单总体均值的差异检验

2.4.2 双总体均值的差异检验

2.5 T 检验

通过对卡方检验我们理解了，这个某某检验无非是把要检验的量构造成某某分布。T 检验当然是对于能搞成 t 分布的检验，其实跟 Z 检验差不多，无非是此时方差是通过样本算出来的而不是已知的。

Definition 2.3 **T 分布**

设随机变量 X 服从标准正态分布 $\mathcal{N}(0,1)$，随机变量 Y 服从自由度为 n 的卡方分布 $\chi^2(n)$，且 X 与 Y 相互独立，则随机变量

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

的分布称为自由度为 n 的 T 分布，记为 $T \sim t(n)$。

2.5.1 单样本 t 检验

2.5.2 配对样本 t 检验

2.5.3 独立样本 t 检验

2.6 F 检验

3 ANOVA

3.1 因素

在小节 2 我们知道了

Definition 3.1 因素/因子

称影响因变量（通常是频数）的分类自变量为因素或因子。

3.2 单因素方差分析

3.3 双因素方差分析

4 Linear Regression

5 Cluster Analysis

6 Discriminant Analysis

7 Priciple Component Analysis

8 Factor Analysis