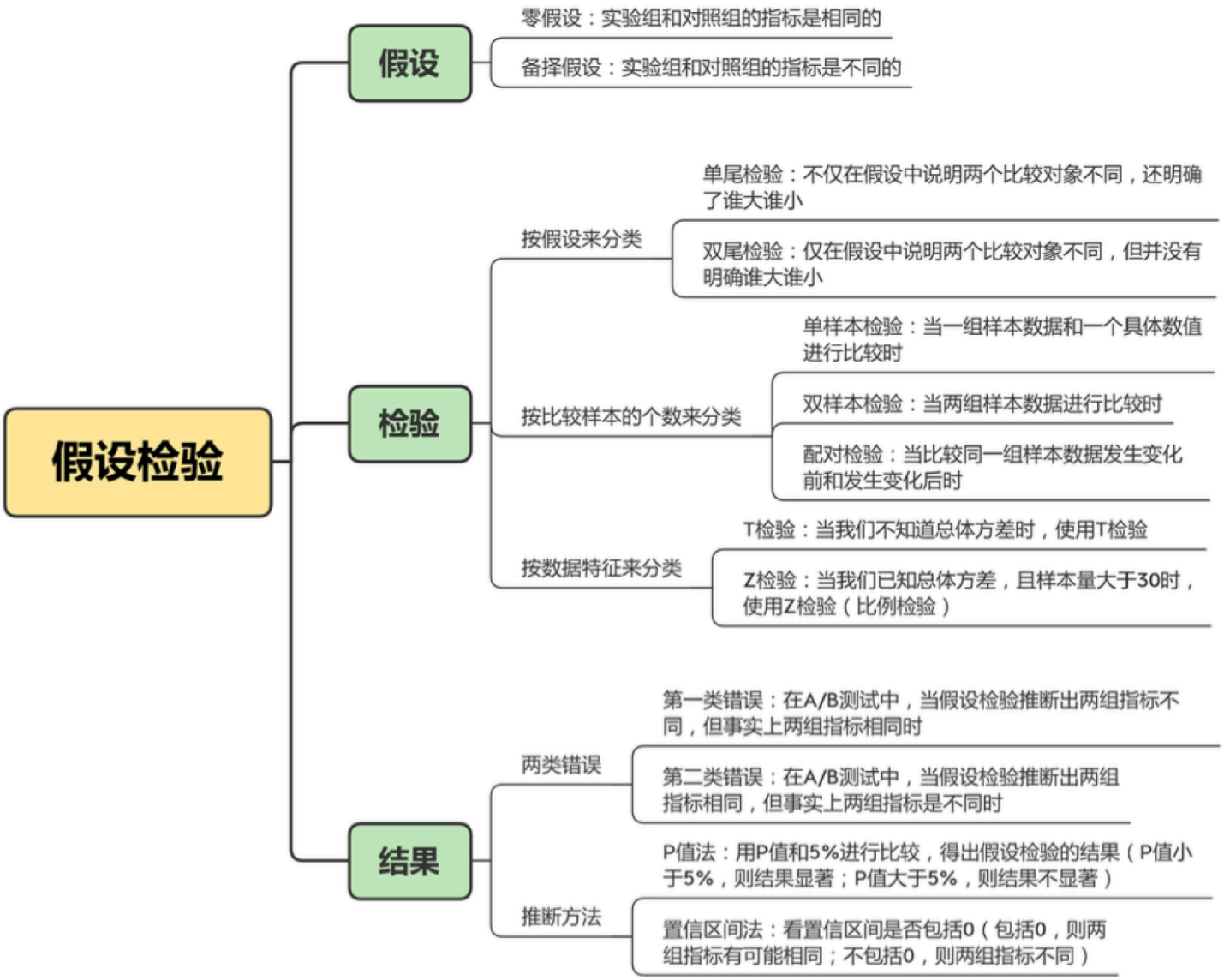
补充：统计学原理

1. 假设检验



1.1 建立假设

在假设检验中，常把一个被检验的假设称为原假设，用H_0表示，通常将不应轻易加以否定的假设作为原假设. 当H_0被拒绝时而接收的假设称为备择假设，用H_1表示，它们常常成对出现。

1.2 选择检验统计量，给出拒绝域形式

由样本对原假设进行判断总是通过一个统计量完成的，该统计量称为检验统计量. 比如，在例 7.1.1中，样本均值 $\bar{x}$ 就是一个很好的检验统计量，因为要检验的假设是正态总体的均值，在方差已知场合，样本均值 x 是总体均值的充分统计量、使原假设被拒绝的样本观测值所在区域称为拒绝域，一般它是样本空间的一个子集，并用 W 表示，在例 7.1.1中，样本均值 $\bar{x}$ 愈大，意味着总体均值 θ 也大，样本均值 $\bar{x}$ 愈小，意味着总体均值 θ 也小，因此，在样本均值的取值中有一个临界值 c（待定），所以拒绝域为

$$\{W = |(x_1, \cdots, x_n); \bar{x} \leqslant c\} = \{\bar{x} \leqslant c\}$$

是合理的.

当拒绝域确定了，检验的判断准则跟着也确定了：

- 如果 $(x_1, \cdots, x_n) \in W$ ，则认为 H_0 不成立；
- 如果 $(x_1, \cdots, x_n) \in \overline{W}$ ，认为 H_0 成立；

1.3 选择显著性水平

假设检验会推断出两种结果:

1. 接受零假设，拒绝备择假设，也就是说实验组和对照组的指标是相同的
2. 接受备择假设，拒绝零假设，也就是说实验组和对照组的指标是不同的

A/B测试的可能结果

	两组指标事实上不同	两组指标事实上相同
假设检验推断出两组指标不同	推断正确	第一类错误（Type I Error）或假阳性（False Positive）
假设检验推断出两组指标相同	第二类错误（Type II Error）或假阴性（False Negative）	推断正确

1.3.1 第一类错误(Type I Error)

定义：统计上的定义是拒绝了事实上是正确的零假设

> 在 A/B 测试中，零假设是两组的指标是相同的，当假设检验推断出两组指标不同，但事实上两组指标相同时，就是第一类错误。我们把两组指标不同称作阳性(Positive)。所以，第一类错误口叫假阳性(False Positive)。

> 发生第一类错误的概率用α表示，也被称为显著水平(Significance Level)。“显著”是指错误发生的概率大，统计上把发生率小于 5% 的事件称为小概率事件，代表这类事件不容易发生。因此显著水平一般也为 5%。

1.3.2 第二类错误(Type II Error)

定义：统计上的定义是接受了事实上是错误的零假设。

> 在 A/B 测试中，当假设检验推断出两组指标相同，但事实上两组指标是不同时，就是第二类错误。
>
> 我们把两组指标相同称作阴性(Negative)，所以第二类错误口叫假阴性(FalseNegative)。发生第二类错误的概率用β表示，统计上一般定义为 20%。

1.3.3 势函数

利用这个势函数容易写出其犯两类错误的概率分别为

$$\alpha(\theta) = \Phi\left(\frac{c-\theta}{4/5}\right), \quad \theta \in \Theta_0 \tag{7.1.4}$$

$$\beta(\theta) = 1 - \Phi\left(\frac{c-\theta}{4/5}\right), \quad \theta \in \Theta_1 \tag{7.1.5}$$

1.4 给出拒绝域

1.4.1 p值

定义：在统计上，P 值就是当零假设成立时，我们所观测到的样本数据出现的概率

> 在 A/B 测试 的语境下，P 值就是当对照组和实验组指标事实上是相同时，在 A/B 测试中用样本数据所观测到的“实验组和对照组指标相同”出现的概率。

> p值是在零假设成立的情况下观测到样本数据或更极端情况发生的概率：

> 当你进行假设检验时，你首先假设零假设是正确的。然后，你使用收集到的样本数据计算一个统计量，例如t值或Z值，根据这个统计量计算出一个p值。p值表示，如果零假设是正确的，那么观测到的样本数据或更极端情况发生的概率有多大。换句话说，p值衡量了你的观测数据与零假设一致的程度
>
> 与此相反的是，当我们在 A/B 测试中观测到“实验组和对照组指标不同”的概率(P 值) 很大，比如 70%，那么在零假设成立时，我们观测到这个事件还是很有可能的。所以这个时候我们接受零假设，拒绝备择假设，即两组指标是相同的。

在统计中，我们会用 P 值和显著水平α进行比较，口因为α一般取 5%，所以就用 P 值和5% 进行比较，就可以得出假设检验的结果了:

- 当P值小于5%时，我们拒绝零假设，接受备择假设，得出两组指标是不同的结论，口叫做结果显著
- 当P值大于5%时，我们接受零假设，拒绝备择假设，得出两组指标是相同的结论，口叫做结果不显著

1.4.2 置信区间

置信区间是一个范围，一般前面会跟着一个百分数，最常见的是 95% 的置信区间。这是什么意思呢?在统计上，对于一个随机变量来说，有 95% 的概率包含总体平均值(Population mean)的范围，就叫做 95% 的置信区间。

置信区间的统计定义其实不是特别好懂，其实你可以直接把它理解为随机变量的波动范围，95% 的置信区间就是包含了整个波动范围的 95% 的区间。

> 置信水平表示置信区间包含真正的实验效应的频率（100次有多少次）

A/B 测试本质上就是要判断对照组和实验组的指标是否相等，那怎么判断呢?

答案就是计算实验组和对照组指标的差值δ。因为指标是随机变量，所以它们的差值δ也会是随机变量，具有一定的波动性。

这就意味着，我们就要计算出δ的置信区间，然后看看这个置信区间是否包括 0。

- 如果包括 0 的话，则说明 δ 有可能为 0，意味着两组指标有可能相同
- 如果不包括 0，则说明两组指标不同

> 例如，计算得出两组指标差值δ的 95% 置信区间为$[0.005, 0.011]$，不包含 0，也可以推断出两组指标显著不同。

若实验组和对照组分别的置信区间有95%区域不重叠，则实验效应应该是统计显著的，此时p值 < 0.05.

!理解95%：95%表示经过许多研究计算得到的95%置信区间，例如进行100次研究计算，会得到100个对应的95%置信区间，而在这100个95%置信区间中，有多少频率、有几个置信区间会包含真正的实验效应。

2. 区间估计类型

2.1 * 区间估计概念及构造置信区间的方法

2.1.1 区间估计概念

设 θ 是总体的一个参数,$x_1,\cdots,x_n$ 是样本，所谓区间估计就是要找两个统计量元 $\hat{\theta}_L = \hat{\theta}_L(x_1,\cdots,x_n)$ 和 $\hat{\theta}_U = \hat{\theta}_U(x_1,\cdots,x_n)$, 使得 $\hat{\theta}_L < \hat{\theta}_U$, 在得到样本观测值之后，就把 θ 估计在区间 $\left[\hat{\theta}_L,\hat{\theta}_U\right]$ 内. 显然, 作为区间估计通常要求区间 $\left[\hat{\theta}_L,\hat{\theta}_U\right]$ 盖住 θ 的概率 $P(\hat{\theta}_L \leqslant \theta \leqslant \hat{\theta}_U)$ 尽可能大, 但这必然导致区间长度增大, 为平衡此矛盾, 把区间 $\left[\hat{\theta}_L,\hat{\theta}_U\right]$ 盖住 θ 的概率 (以后称为置信水平) 事先给定, 这就引入如下置信区间的概念.

定义 6.5.1. 设 θ 是总体的一个参数, 其参数空间为 Θ, $x_1,\cdots,x_n$ 是来自该总体的样本, 对给定的一个 $\alpha(0<\alpha<1)$, 若有两个统计量 $\hat{\theta}_2 = \hat{\theta}_2(x_1,\cdots,x_n)$ 和 $\hat{\theta}_U = \hat{\theta}_U(x_1,\cdots,x_n)$, 若对任意的 $\theta \in \Theta$, 有

$$P_\theta\left(\hat{\theta}_L \leqslant \theta \leqslant \hat{\theta}_U\right) \geqslant 1-\alpha \tag{6.5.1}$$

则称随机区间 $\left[\hat{\theta}_{\mathrm{L}},\hat{\theta}_{\mathrm{U}}\right]$ 为 θ 的**置信水平为 $1-\alpha$ 的置信区间**, 或简称 $\left[\hat{\theta}_{\mathrm{L}},\hat{\theta}_{\mathrm{U}}\right]$ 是 θ 的 $1-\alpha$ **置信区间**, $\hat{\theta}_{\mathrm{L}}$ 和 $\hat{\theta}_{\mathrm{U}}$ 分别称为 θ 的 (双侧) **量信下限**和**量信上限**.

置信水平 $1-\alpha$ 有一个频率解释: 在大量重复使用 θ 的置信区间 $\left[\hat{\theta}_{\mathrm{L}},\hat{\theta}_{\mathrm{U}}\right]$ 时, 每次得到的样本观测值是不同的, 从而每次得到的区间估计值也是不一样的. 对一次具体的观测值而言, θ 可能在 $\left[\hat{\theta}_{\mathrm{L}},\hat{\theta}_{\mathrm{U}}\right]$ 内, 也可能不在. 平均而言, 在这大量的区间估计观测值中, 至少有 $100(1-\alpha)\%$ 包含 θ. 下例中的图6.5.1和图6.5.2直观地显示了该种频率意义.

同等置信区间、置信上限、置信下限

定义 6.5.2. 沿用定义6.5.1的记号, 如对给定的 $\alpha(0<\alpha<1)$ 对任意的 $\theta\in\Theta$, 有

$$P_\theta\left(\hat{\theta}_L\leqslant\theta\leqslant\hat{\theta}_U\right)=1-\alpha \tag{6.5.2}$$

则称 $\hat{\theta}_L$ 为 θ 的 $1-\alpha$ **同等置信区间**.

在一些实际问题中, 人们感兴趣的有时仅仅是未知参数的一个下限或一个上限. 譬如, 对某种产品的平均寿命来说, 我们希望它越大越好, 因此人们关心的是它的 0.90 置信下限是多少, 此下限标志了该产品的质量, 它的一般定义如下.

定义 6.5.3. 设 $\hat{\theta}_L=\hat{\theta}_L(x_1,\cdots,x_n)$ 是统计量, 对给定的 $\alpha\in(0,1)$ 和任意的 $\theta\in\Theta$ 有

$$P_\theta\left(\tilde{\theta}_L\leqslant\theta\right)\geqslant 1-\alpha,\qquad\forall\theta\in\Theta \tag{6.5.3}$$

则称 $\hat{\theta}_L$ 为 θ 的置信水平为 $1-\alpha$ 的 (**单侧**) **置信下限**. 假如等号对一切 $\theta\in\Theta$ 成立, 则称 $\hat{\theta}_L$ 为 θ 的 $1-\alpha$ **同等置信下限**.

类似地, 对某些指标人们希望它越小越好. 比如, 某种药品的毒性, 这引出了置信上限的概念.

定义 6.5.4. 设 $\hat{\theta}_U=\hat{\theta}_U(x_1,\cdots,x_n)$ 是统计量, 对给定的 $\alpha\in(0,1)$ 和任意的 $\theta\in\Theta$, 有

$$P_\theta\left(\hat{\theta}_U\geqslant\theta\right)\geqslant 1-\alpha \tag{6.5.4}$$

则称 $\hat{\theta}_U$ 为 θ 的置信水平为 $1-\alpha$ 的 (**单侧**) **置信上限**. 若等号对一切 $\theta\in\Theta$ 成立, 则称 $\hat{\theta}_U$ 为 $1-\alpha$ **同等置信上限**.

不难看出, 单侧置信下限和单侧置信上限都是置信区间的特殊情形. 因此, 寻求置信区间的方法可以用来寻找置信限. 接下来我们主要介绍寻找置信区间的方法.

2.1.2 *构造置信区间 - 轴度量法

构造未知参数 θ 的置信区间的最常用的方法是**枢轴置法**, 其步骤可以概括为如下三步:

1. 设法构造一个样本和 θ 的函数 $G=G(x_1,\cdots,x_n,\theta)$ 使得 G 的分布不依赖于未知参数. 一般称具有这种性质的 G 为枢轴量.
2. 适当地选择两个常数 c、d, 使对给定的 $\alpha(0<\alpha<1)$, 有

$$P(c\leqslant G\leqslant d)=1-\alpha \tag{6.5.5}$$

3. 假如能将 $c\leqslant G\leqslant d$ 进行不等式等价变形化为 $\hat{\theta}_L\leqslant\theta\leqslant\hat{\theta}_U$, 则有

$$P_\theta\left(\hat{\theta}_L\leqslant\theta\leqslant\hat{\theta}_U\right)=1-\alpha \tag{6.5.6}$$

这表明 $\left[\hat{\theta}_L,\hat{\theta}_v\right]$ 是 θ 的 $1-\alpha$ 同等置信区间.

上述构造置信区间的关键在于构造枢轴量 G, 故把这种方法称为**枢轴置法**. 枢轴量的寻找一般从 θ 的点估计出发. 而满足6.5.5的 c、d 可以有很多, 选择的目的是希望6.5.6中的平均长度 $E_\theta\left(\tilde{\theta}_U-\hat{\theta}_L\right)$ 尽可能短. 假如可以找到这样的 c、d 使 $E_\theta\left(\tilde{\theta}_U-\hat{\theta}_L\right)$ 达到最短当然是最好的, 不过在不少场合很难做到这一点. 故常这样选择 c 和 d, 使得

$$P_\theta(G<c)=P_\theta(G>d)=\alpha/2 \tag{6.5.7}$$

这样得到的置信区间称为**等尾置倍区间**. 实用的置信区间大都是等尾置信区间.

2.2 单个正态总体参数的置信区间

2.2.1 方差σ已知

在这种情况下, 由于 μ 的点估计为 $\bar{x}$, 其分布为 $N\left(\mu, \sigma^2/n\right)$, 因此枢轴量可选为 $G = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$, c 和 d 应满足 $P(c \leqslant G \leqslant d) = \Phi(d) - \Phi(c) = 1-\alpha$, 经过不等式变形可得

$$P_\mu(\bar{x} - d\sigma/\sqrt{n} \leqslant \mu \leqslant \bar{x} - c\sigma\sqrt{n}) = 1 - a$$

该区间长度为 $(d-c)\sigma\sqrt{n}$. 由于标准正态分布为单峰对称的, 从图 6.5.3 上不难看出在 $\Phi(d) - \Phi(c) = 1-\alpha$ 的条件下, 当 $d = -c = u_{1-\alpha/2}$ 时, $d-c$ 达到最小, 由此给出了 μ 的 $1-\alpha$ 同等置信区间为

$$\left[\bar{x} - u_{1-\alpha/2}\sigma/\sqrt{n}, \quad \bar{x} + u_{1-\alpha/2}\sigma/\sqrt{n}\right] \tag{6.5.8}$$

这是一个以 $\bar{x}$ 为中心, 半径为 $u_{1-\alpha/2}\sigma/\sqrt{n}$ 的对称区间, 常将之表示为 $\bar{x} \pm u_{1-\alpha/2}\sigma/\sqrt{n}$.

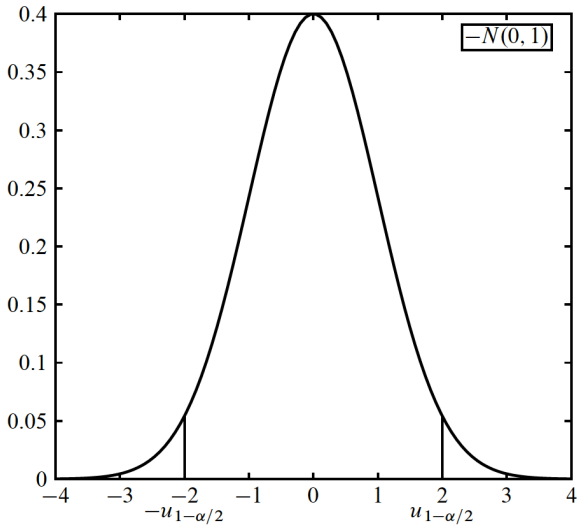


图 6.5.3: 标准正态分布示意图

2.2.2 方差σ未知

这时可用 t 统计量, 因为 $t = \frac{\sqrt{n}(\bar{x}-\mu)}{s} \sim t(n-1)$, 因此 t 可以用来作为枢轴量. 完全类似于上一小节, 可得到 μ 的 $1-\alpha$ 置信区间为

$$\left[\bar{x} - t_{1-\alpha/2}(n-1)s/\sqrt{n}, \bar{x} + t_{1-\alpha/2}(n-1)s/\sqrt{n}\right] \tag{6.5.9}$$

此处 $s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$ 是 σ^2 的无偏估计.

2.2.3 大样本置信区间

在样本容量充分大时, 可以用渐近分布来构造近似的置信区间. 一个典型的例子是关于比例 p 的置信区间.

设 $x_1, \cdots, x_n$ 是来自二点分布 $b(1,p)$ 的样本, 现要求 p 的 $1-\alpha$ 置信区间. 由中心极限定理知, 样本均值 $\bar{x}$ 的渐近分布为 $N\left(p, \frac{p(1-p)}{n}\right)$, 因此有

$$u = \frac{\bar{x}-p}{\sqrt{p(1-p)/n}} \dot{\sim} N(0,1)$$

这个 w 可作为枢轴量, 对给定 α, 利用标准正态分布的 $1-\alpha/2$ 分位数 $u_{1-\alpha/2}$ 可得

$$P\left(\left|\frac{\bar{x}-p}{\sqrt{p(1-p)/n}}\right| \leqslant u_{1-\alpha/2}\right) \approx 1-\alpha$$

括号里的事件等价于

$$(\bar{x}-p)^2 \leqslant u_{1-\alpha/2}^2 p(1-p)/n$$

记 $\lambda = u_{1-\alpha/2}^2$, 上述不等式可化为

$$\left(1+\frac{\lambda}{n}\right)p^2 - \left(2\bar{x}+\frac{\lambda}{n}\right)p + \bar{x}^2 \leqslant 0$$

左侧的二次多项式的判别式

$$\left(2\bar{x}+\frac{\lambda}{n}\right)^2 - 4\left(1+\frac{\lambda}{n}\right)\bar{x}^2 = \frac{4\bar{x}(1-\bar{x})}{n} + \frac{\lambda^2}{n^2} > 0,$$

故此二次多项式是开口向上并与 x 轴有两个交点的曲线 (见图6.5.5). 记此两个交点为 p_L 和 p_U, 则有

$$P(p_L \leqslant p \leqslant p_U) = 1-\alpha$$

这里 p_L 和 p_U 是该二次多项式的两个根, 它们可表示为

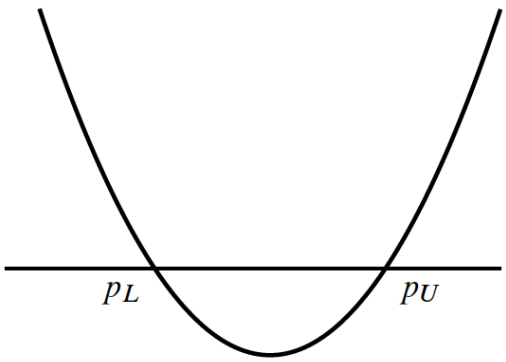


图 6.5.5: 二次多项式及其根示意图

$$p = \frac{1}{1+\frac{\lambda}{n}}\left(\bar{x}+\frac{\lambda}{2n} \pm \sqrt{\frac{\bar{x}(1-\bar{x})}{n}+\frac{\lambda^2}{4n^2}}\right)$$

由于 n 比较大, 在实用中通常略去 λ/n 项, 于是可将置信区间近似为

$$\left[\bar{x}-u_{1-\alpha/2}\sqrt{\frac{\bar{x}(1-\bar{x})}{n}}, \bar{x}+u_{1-\alpha/2}\sqrt{\frac{\bar{x}(1-\bar{x})}{n}}\right] \tag{6.5.11}$$

2.3 两个正态总体的置信区间 - $\mu_1 - \mu_2$

设 $x_1, \cdots, x_m$ 是来自 $N(\mu_1, \sigma_1^2)$ 的样本, $y_1, \cdots, y_n$ 是来自 $N(\mu_2, \sigma_2^2)$ 的样本, 且两个样本相互独立. $\bar{x}$ 与 $\bar{y}$ 分别是它们的样本均值, $s_x^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \bar{x})^2$ 和 $s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$. 分别是它们的样本方差. 下面讨论两个均值差和两个方差比的置信区间.

2.3.1 方差已知

此时有 $\bar{x} - \bar{y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}\right)$, 取枢轴量为

$$u = \frac{\bar{x} - \bar{y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} \sim N(0, 1),$$

沿用前面多次用过的方法可以得到 $\mu_1 - \mu_2$ 的 $1 - \alpha$ 置信区间为

$$\left[\bar{x} - \bar{y} - u_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}, \bar{x} - \bar{y} + u_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}\right],$$

该区间称为二样本 u 区间.

2.3.2 方差未知 $\sigma_1^2 = \sigma_2^2 = \sigma^2$

此时有

$$\bar{x} - \bar{y} \sim N\left(\mu_1 - \mu_2, \left(\frac{1}{m} + \frac{1}{n}\right)\sigma^2\right),$$

$$\frac{(m-1)s_x^2 + (n-1)s_y^2}{\sigma^2} \sim \chi^2(m + n - 2)$$

由于 $\bar{x}, \bar{y}, s_x^2, s_y^2$ 相互独立, 故可构造如下服从 t 分布 $t(m + n - 2)$ 的枢轴量

$$t = \sqrt{\frac{mn(m+n-2)}{m+n}} \frac{\bar{x} - \bar{y} - (\mu_1 - \mu_2)}{\sqrt{(m-1)s_x^2 + (n-1)s_y^2}} \sim t(m + n - 2)$$

记 $s_w^2 = \frac{(m-1)s_x^2 + (n-1)s_y^2}{m+n-2}$, 则 $\mu_1 - \mu_2$ 的置信区间为

$$\left[\bar{x} - \bar{y} - \sqrt{\frac{m+n}{mn}} s_w t_{1-\alpha/2}(m + n - 2), \bar{x} - \bar{y} + \sqrt{\frac{m+n}{mn}} s_w t_{1-\alpha/2}(m + n - 2)\right].$$

2.3.3 当m和n 都很大时的近似置信区间

可以证明:

$$\frac{\overline{x}-\overline{y}-(\mu_1-\mu_2)}{\sqrt{\frac{s_x^2}{m}+\frac{s_y^2}{n}}} \sim N(0,1)$$

由此可给出$\mu_1-\mu_2$的$1-\alpha$近似置信区间为:

$$[\overline{x}-\overline{y}-\mu_{1-\alpha/2}\sqrt{\frac{s_x^2}{m}+\frac{s_y^2}{n}}, \overline{x}-\overline{y}+\mu_{1-\alpha/2}\sqrt{\frac{s_x^2}{m}+\frac{s_y^2}{n}}]$$

2.4 比例p检验

比例 p 可看作某事件发生的概率，即可看作二点分布 $b(1,p)$ 中的参数. 作 n 次独立试验，以 x 记该事件发生的次数，则 $x \sim b(n,p)$. 我们可以根据 x 检验关于 p 的一些假设. 先考虑如下单边假设检验问题.

$$H_0 : p \leqslant p_0 \quad \text{vs} \quad H_1 : p > p_0 \tag{7.3.8}$$

直观上看，一个显然的检验方法是取如下的拒绝域 $W = \{x \geqslant c\}$，由于 x 只取整数值，故 c 可限制在非负整数中. 然而，一般情况下对给定的 α，不一定能正好取到一个 c 使

$$P(x \geqslant c; p_0) = \sum_{i=c}^{n} \binom{n}{i} p_0^i (1-p_0)^{n-i} = \alpha \tag{7.3.9}$$

能恰巧使得7.3.9成立的 c 值是罕见的. 这是在对离散总体作假设检验中普遍会遇到的问题，在这种情况下，较常见的是找一个 c_0，使得

$$\sum_{i=c_0}^{n} \binom{n}{i} p_0^i (1-p_0)^{n-i} > a > \sum_{i=c_0+1}^{n} \binom{n}{i} p_0^i (1-p_0)^{i-1}$$

于是，若取 $c = c_0$，这相当于把检验的显著性水平提高了一些，由 α 提高到到 $\sum_{i=c_0}^{n} \binom{n}{i} p_0^i (1-p_0)^{n-i}$ 若取 $c = c_0 + 1$，此时相当于把显著性水平由 α 降低到 $\sum_{i=c_0+1}^{n} \binom{n}{i} p_0^i (1-p_0)^{i-1}$，因为后者可保证 (7.3.9) 的左侧不大于 α，故取 $c = c_0 + 1$ 可得水平为 α 的检验，

对检验问题

$$H_0 : p \geqslant p_0 \quad \text{vs} \quad H_1 : p < p_0 \tag{7.3.10}$$

处理方法是类似的，检验的拒绝域为 $W = \{x \geqslant c\}$，c 为满足

$$\sum_{i=0}^{c} \binom{n}{i} p_0^i (1-p_0)^{n-i} \leqslant a$$

的最大正整数. 对检验问题

$$H_0 : p = p_0 \quad \text{vs} \quad H_1; p \neq p_0, \tag{7.3.11}$$

检验的拒绝域 $W = \{x \leqslant c_1\}$ 或 $\{x \geqslant c_2\}$，其中 c_1 为满足

$$\sum_{i=0}^{c_1} \binom{n}{i} p_0^i (1-p_0)^{n-i} \leqslant \frac{\alpha}{2}$$

的最大整数，c_2 为满足

$$\sum_{i=c_2}^{n_1} \binom{n}{i} p_0^i (1-p_0)^{n-i} \leqslant \frac{\alpha}{2}$$

的最小整数.

2.4.1 大样本情况

前一小节我们介绍了对二点分布参数 p 的检验问题，我们看到临界值的确定比较繁琐，使用不太方便. 如果样本量较大，我们可用近似的检验方法——大样本检验. 其一般思路如下：设 $x_1,\cdots,x_n$，是来自某总体的样本，又设该总体均值为 θ，方差为 θ 的函数，记为 $\sigma^2(\theta)$，譬如，对二点分布 $b(1,\theta)$，其方差 $\theta(1-\theta)$ 是均值 θ 的函数，则对下列三类假设检验问题：

(1) $H_0:\theta\leqslant\theta_0$ vs $H_0:\theta>\theta_0$;

(2) $H_0:\theta\geqslant\theta_0$ vs $H_0:\theta<\theta_0$;

(3) $H_0:\theta=\theta_0$ vs $H_0:\theta\neq\theta_0$

在样本容量 n 充分大时，利用中心极限定理，$\bar{x}\dot{\sim}N\left(\theta,\sigma^2(\theta)/n\right)$，故在 $\theta=\theta_0$ 时，可采用如下检验统计量

$$u=\frac{\sqrt{n}(\bar{x}-\theta_0)}{\sqrt{\sigma^2(\theta_0)}}\dot{\sim}N(0,1) \tag{7.3.12}$$

近似地确定拒绝域. 对应上述三类检验问题的拒绝域依次分别为

$$W=\{u\geqslant u_{1-\alpha}\}$$
$$W=\{u\leqslant u_a\}$$
$$W=\left\{|u|\geqslant u_{1-a/2}\right\}$$

2.5 总结

2.5.1 单个正态总体

σ	检验方式	统计量	H_0	拒绝域
已知	z检验	$z=\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$	$\mu\leq\mu_0$	$W=\{z\geq z_{1-\alpha}\}$
			$\mu\geq\mu_0$	$W=\{z\leq z_{-\alpha}\}$
			$\mu=\mu_0$	$W=\{\|z\|\geq z_{1-\alpha/2}\}$
未知	t检验	$t=\frac{\sqrt{n}(\bar{x}-\mu_0)}{s}$	$\mu\leq\mu_0$	$W=\{t\geq t_{1-\alpha}(n-1)\}$
			$\mu\geq\mu_0$	$W=\{t\leq t_{-\alpha}(n-1)\}$
			$\mu=\mu_0$	$W=\{\|t\|\geq t_{1-\alpha/2}(n-1)\}$
大样本	z检验			

2.5.2 两个正态总体

> 此处**u检验**也叫**z检验**

表 7.2.2: 两个正态总体均值的假设检验

检验法	条件	原假设 H_0	备择假设 H_1	检验统计量	拒绝域
u 检验	σ_1,σ_2 已知	$\mu_1 \leqslant \mu_2$	$\mu_1 > \mu_2$	$u=\frac{(\bar{x}-\bar{y})}{\sqrt{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}}}$	$\{u \geqslant u_{1-\alpha}\}$
		$\mu_1 \geqslant \mu_2$	$\mu_1 < \mu_2$		$\{u \leqslant u_{\alpha}\}$
		$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$		$\{\lvert u\rvert \geqslant u_{1-\alpha/2}\}$
t 检验	σ_1,σ_2 未知	$\mu_1 \leqslant \mu_2$	$\mu_1 > \mu_2$	$t=\frac{(\bar{x}-\bar{y})}{s_w\sqrt{\frac{1}{m}+\frac{1}{n}}}$	$\{t \geqslant t_{1-\alpha}(m+n-2)\}$
	$\sigma_2 \geqslant \mu_2$	$\mu_1 < \mu_2$		$\{t \leqslant t_{\alpha}(m+n-2)\}$	
	$\sigma_1=\sigma_2$	$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$		$\{\lvert t\rvert \geqslant t_{1-\alpha/2}(m+n-2)\}$
大样本检验	σ_1,σ_2 未知 m,n 充分大	$\mu_1 \leqslant \mu_2$	$\mu_1 > \mu_2$	$u=\frac{(\bar{x}-\bar{y})}{\sqrt{\frac{s_x^2}{n}+\frac{s_y^2}{m}}}$	$\{u \geqslant u_{1-\alpha}\}$
		$\mu_1 \geqslant \mu_2$	$\mu_1 < \mu_2$		$\{u \leqslant u_{\alpha}\}$
		$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$		$\{\lvert u\rvert \geqslant u_{1-\alpha/2}\}$
近似 t 检验	σ_1,σ_2 未知 m,n 不很大	$\mu_1 \leqslant \mu_2$	$\mu_1 > \mu_2$	$t=\frac{(\bar{x}-\bar{y})}{\sqrt{\frac{s_x^2}{n}+\frac{s_y^2}{m}}}$	$\{t \geqslant t_{1-\alpha}(l-1)\}$
		$\mu_1 \geqslant \mu_2$	$\mu_1 < \mu_2$		$\{t \leqslant t_{\alpha}(l-1)\}$
		$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$		$\{\lvert t\rvert \geqslant t_{1-\alpha/2}(l-1)\}$

2.5.3 比例p检验

一般考虑为大样本情况，用z检验方法

检验方式	统计量	H_0	拒绝域
z检验	$u=\frac{\sqrt{n}(\bar{x}-\theta_0)}{\sqrt{\sigma^2(\theta_0)}} \sim N(0,1)$	$\mu \leq \mu_0$	$W=\{z \geq z_{1-\alpha}\}$
		$\mu \geq \mu_0$	$W=\{z \leq z_{-\alpha}\}$
		$\mu = \mu_0$	$W=\{\lvert z\rvert \geq z_{1-\alpha/2}\}$

3. AB实验的建设检验

A/B 测试的语境中，假设一般是指关于实验组和对照组指标的大小的推断。

在假设检验中的“假设”是一对:零假设(Null Hypothesis)和备择假设(Alternative Hypothesis)，它们是完全相反的。

在 A/B 测试的语境下，零假设指的是实验组和对照组的指标是相同的，备择假设指的是实验组和对照组的指标是不同的。

3.1 单尾、双尾检验

单尾检验(One-tailed Test)和双尾检验(Two-tailed Test)这两个概念。

- 单尾检验口叫单边检验(One-sided Test)，它不仅在假设中说明了两个比较对象不同，并且还明确了谁大谁小，比如实验组的指标比对照组的指标大。
- 双尾检验口叫双边检验(Two-sided Test)，指的是仅仅在假设中说明了两个比较对象不同，但是并没

有明确谁大谁小。

> 单边检验可能需要对数据有主观意识的代入，例如指标应该更大，或者更小，而双边检验不带有这种主观想法，只是单纯利用数据来判断相同还是不同。

3.1.1 在 A/B 测试的实践中，更推荐使用双尾检验。

原因如下：

- 第一个原因是，双尾检验可以让数据自身在决策中发挥更大的作用。

  > 我们在实践中使用 A/B 测试，就是希望能够通过数据来驱动决策。**我们要尽量减少在使用数据前产生的任何主观想法来干扰数据发挥作用。**
  >
  > 所以，双尾检验这种不需要我们明确谁大谁小的检验，更能发挥数据的作用。

- 第二个原因：双尾检验可以帮助全面考虑变化带来的正、负面结果。

  > 双尾检验可以同时照顾到正面和负面的结果，更接近多变的现实情况。但是单尾检验只会适用于其中一种，而且通常是我们期望的正面效果。

3.2 其他检验

检验有很多种，单尾检验和双尾检验，是从“假设”的角度来分类的。除此之外， 常见的“检验”还可以根据比较样本的个数进行分类，包括**单样本检验(One-Sample Test)、 双样本检验(Two-Sample Test)和配对检验(Paired Test)**

3.2.1 各个检验的使用范围

- 当两组样本数据进行比较时，就用双样本检验

  > 比如 A/B 测试中实验组和对照组的比较。

- 当一组样本数据和一个具体数值进行比较时，就用单样本检验

  > 比如，我想比较极客时间用户的日均使用时间有没有达到 15 分钟，这个时候，我就可以把一组样本数据(抽样所得的极客时间用户的每日使用时间)和一个具体数值15来进行比较。

- 当比较同一组样本数据发生变化前和发生变化后时，就用配对检验

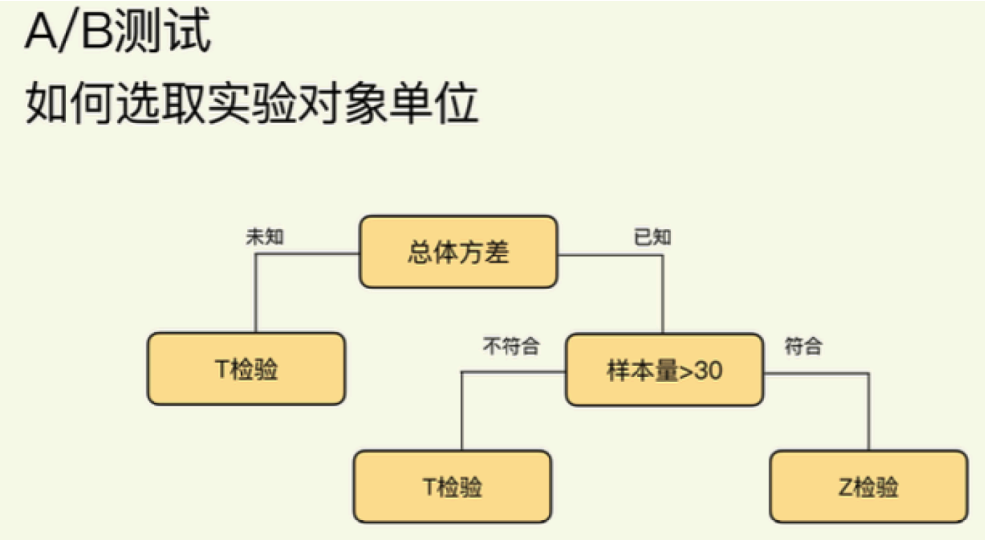
> 比如，我现在随机抽取 1000 个极客时间的用户，给他们“全场专栏一律 1 折”这个优惠，然后在这1000 个人中，我们会比较他们在收到优惠前一个月的日均使用时间，和收到优惠后一个月的日均使用时间。

> **在 A/B 测试中，使用双样本检验。**

3.3 T检验、Z检验

主要看样本量的大小和是否知道总体方差(Population Variance)：

- 当我们不知道总体方差时，使用 T 检验
- 当我们已知总体方差，且样本量大于 30 时，使用 Z 检验



> 在统计中我们习惯说样本量大于30 就是很大的样本，就可以用样本方差来近似总体方差，这样我们就知道总体方差，就可以用Z检验了，但其实30只是经验值，大于30的总体方差也是样本方差近似的，所以如果准确的说的话样本量大于30，在总体方差未知的情况下，也要用T检验。

这些理论具体到 A/B 测试实践中，一个经验就是：**均值类指标一般用 T 检验，概率类指标一般用 Z 检验(比例检验)。**

> 在样本量足够大的情况下 T 分布近似于Z分布，所以如果你不知道该用t检验还是z检验，而**样本量够大时，直接用T分布即可**；

> 比例检验(Proportion Test) 是，专指用于**检验概率类指标的 Z 检验**。

3.4 FAQ

> **AB测试是否也能转换成单样本检验?比如AB两组样本，用A样本的均值标准差，和B样本做单样本检验? 通常用excel的Z.TEST时会这么干，会有什么问题吗**
>
> 双样本检验是两个有波动性的随机变量在比较，单样本检验时一个随机变量和个常数比较，你把其中一个变量简化成一个常数肯定会丢失掉原数据的一些特征嘛结果肯定没有双样本检测准确的，所以A/B测试是不推荐单样本检测的。

> **如果不只两个实验可以用t或z检验吗? 一个对照组两个实验组，用实验组分别和对照组做假设检验吗?**
>
> 对的!你说的是A/B/n测试，这里面有不止一个实验组，这是后就要用实验组分别和对照组做假设检验

> **如何检验两个样本比率是否发生变化**
>
> 用独立性检验也就是卡方检验来验证两个样本比率是否发生变化。

3.5 统计功效

3.5.1 定义与理解

统计功效Power，又被称作 Statistical Power：**如果变体之间存在真实差异，检出出这个有意义的差值的概率（统计上指当真实有差异的时候拒绝零假设的概率）**

Power 的本质是概率，在 A/B 测试中，如果实验组和对照组的指标事实上是不同的，Power 指的就是通过 A/B 测试探测到两者不同的概率。

> 在实验组和对照组中事实上确实存在差异时，AB测试准确检测出差异的概率。
>
> Power越大，就越能探测到两组的不同。**把 Power 看成 A/B 测试的灵敏度就可以了**

$\alpha = P(\text{reject null} \mid \text{null true})$

$\beta = P(\text{fail to reject} \mid \text{null false})$

$1-\beta = \text{sensitivity}$： 通常为80%

> 小样本：
>
> - α 小
> - β 高
>
> 大样本：
>
> - α 不变
> - β 更小

统计功效是在测试中检测出选件之间转化率真实差异的概率。由于转化事件存在随机性，因此即使两个选件之间的转化率在长期测试中存在实际差异，该测试可能也不会显示具有统计意义的显著差异。可以认为这就是运气不好或纯属偶然。我们将这种未能检测到转化率真实差异的情况称为漏报或 II 类错误。

3.5.2 例子

我们先把用户分为对照组和实验组，其中:

- 对照组是正常的用户注册流程，输入个人基本信息一短信/邮箱验证注册成功实验组是，在正常的用户注册流程中，还加入了微信、微博等第三方账号登录的功能用户可以通过第三方账号一键注册登录
- 相信不用我说，你也能猜到，实验组用户的注册率肯定比对照组的要高，因为实验组帮用户省去了繁琐的注册操作。这就说明，在事实上这两组用户的注册率是不同的

那么，现在如果 A/B 测试有 80% 的 Power，就意味着这个 A/B 测试有 80% 的概率可以准确地检测到这两组用户注册率的不同，得出统计显著的结果。换句话说，这个 A/B 测试有 20% 的概率会错误地认为这两组用户的注册率是相同的
可见，Power 越大，说明 A/B 测试越够准确地检测出实验组与对照组的不同(如果两组事实上是不同的)

> 当对照实验的置信区间包含0，并不意味着置信区间中的零比其他值更有可能出现，实验很可能没有足够的统计功效。