心理学探新　2023,Vol. 43,No. 1,77 - 83
PSYCHOLOGICAL EXPLORATION

SNARC 效应量大小及置信区间的可信问题的实例分析*

何　华
（苏州大学教育学院,苏州 215123）

摘　要:通过实例分析 SNARC 效应量和置信区间可信度问题。(1)效应量是心理学实验结果报告中一个非常重要的部分,依据 SNARC 效应量的特殊性提出一种新的统计方法(在线性回归模型中引入混合虚拟变量)处理 Aleottia 等(2020)的开放实验数据,尽管该方法比一般方法更复杂。(2)参数的区间估计是一种基本统计推断形式 。根据枢轴量分布,置信区间在一定置信度下可估计总体参数所在的可能范围。文章通过构建一个特殊实例分析了置信区间的估计过程,并和假设检验、贝叶斯统计进行对比分析,结果显示,虽然置信区间可以用来估计参数,但是存在依据某置信区间无法作出正确估计的情况,通过贝叶斯统计能分析出其中原因。

关键词:SNARC 效应;虚拟变量;假设检验;置信区间;先验分布;后验分布;贝叶斯统计
中图分类号:B841.2　　**文献标识码**:A　　**文章编号**:1003 - 5184(2023)01 - 0077 - 07

1　SNARC 效应量大小分析

SNARC 效应(spatial - numerical association of response codes effect,SNARC effect)本质上反映的是数量具有空间表征,即,在人脑中存在一条心理数字线,小数(1,2,3,4)表征在左侧空间,而大数(6,7,8,9)表征在右侧空间(Dehaene,Bossini,& Giraux,1993)。其获得的一个最为简单的实验流程为:“ + ”在屏幕中央呈现 300*ms*,消失后随机出现一个小数(或大数)300*ms*,被试既快又准确地以双手食指按键判断该数是大于 5 还是小于 5(刺激 5 不出现在屏幕上)。这个实验用到的任务是数字大小判断(当然还存在其他种类的任务),实验一般采用重复测量实验设计,数据分析采用重复测量方差分析和线性回归分析(Lorch & Myers,1990)。

线性回归分析最终会得到一个一元一次线性回归方程,该方程被用来描述 SANRC 效应。SNARC 效应的线性回归方程一般是指,反应时差异(右侧减左侧)在数字(1,2,3,4,6,7,8,9)上的线性回归,并可以一般性的一元一次线性回归模型描述:

$$\mathrm{DRT} = \beta_0 + \beta_1 \mathrm{N} + \mu \tag{1}$$

其中,N 为数字 1、2、3、4、6、7、8、9,是预测变量;DRT 是右侧反应时减去左侧反应时的差值。该线性模型中的回归系数 β_1 可以为正或负值,正负表明 SNARC 效应的方向不同。SNARC 效应具有普遍性和跨文化性,西方研究者在以西方人为被试和以数字(1,2,3,4,6,7,8,9)为实验材料时,均可稳定获得数字 SNARC 效应,在描述 SNARC 效应的一元一次线性回归方程中,回归系数是负值。而在以希伯来语或波斯语为母语的被试身上,也可稳定获得数字 SNARC 效应,但方向则和西方人的相反,因此回归系数是正值。SNARC 效应的效应量大小就是该线性方程的回归系数大小,或者为该线性方程的斜率的绝对值,与正负号无关。另外,对该方程要进行回归方程的显著性检验和回归系数的显著性检验,检验是欲以证实因变量确实是和自变量呈现线性关系。因为只有一个预测变量,所以这两种检验结果是一致的。

查阅国内外文献发现,需要比较 SNARC 效应量大小的研究为数众多,如,跨通道研究需要比较在单通道和跨通道条件下 SNARC 效应量大小的改变,以此来判断一个通道的数字加工是否影响到了另一个通道的数字加工;还有,不同空间维度(水平、垂直和矢量维度)的 SNARC 效应比较,以此来判断哪个维度下的 SNARC 效应量可能是最大的;另外还有发展性的研究(如,小学生、初中生和大学生),等等。因为具体的研究内容不同,这些研究中

* **基金项目**:江苏省教育学会“十三五”教育科研规划重点课题(19B1N2SZ18)。
通讯作者:何华,E - mail:tougaolunwen2006@163.com。

的 SNARC 效应量的比较具有各自的具体意义。但是其中的共同之处在于,需要通过考察 SNARC 效应量的改变来分析出某个自变量因素是否在其中起了作用。另外还有一个统计上的共同需求,就是如何采取合适有效的统计方法来进行比较分析。前述提到的研究中,需要比较的 SANRC 效应量大小至少是两组之间,这是最少的和最简单的,当然,也可能有四组甚至更多组之间,这种情形就比较复杂了。考虑到简单和具有代表性,选取三组之间的 SNARC 效应量进行比较分析,而且以 Aleottia, Girolamob, Massaccesic 和 Priftis 发表在 2020 年《Cognition》上的论文"Numbers around Descartes: A preregistered study on the three - dimensional SNARC effect"里的开放式数据为分析对象。分析过程中曾向论文作者 Aleottia 和 Priftis(通讯作者)咨询该研究相关事宜以准确把握其数据含义。

Aleottia 等(2020)的研究是有关数字的三维空间表征问题,为当前首个对三维(水平、垂直和矢量)SNARC 效应的研究。Aleottia 等(2020)指出了前人研究存在的一些问题后,创造性设计出一种特殊反应装置(如下图 1 所示),真正解决了三维空间上的按键反应问题。研究得到,三维 SNARC 效应的效应量值是基本相等的(且都是负值),存在三维的数字的心理空间表征。Aleottia 等(2020)采用了频率主义(经典统计分析)和计算贝叶斯因子两种方法对结果进行了分析并得到最终结论,这两种分析方法下的结果大多完全一致,但从 Aleottia 等(2020)的结果分析可以看到,在面对经典统计分析下的边缘显著情况,他们都会进一步进行贝叶斯分析以期得到更准确的判断。具体而言,为探索三个空间维度上 SNARC 效应的情况(即,三个维度上的 SNARC 效应的效应量或回归系数大小是否有显著差异),Aleottia 等(2020)进行了如下分析:对每个维度都进行了线性回归分析,并通过经典的单样本 t - 检验和贝叶斯单样本 t 检验均得到回归方程是显著的。进一步,为了比较三个维度上 SNARC 效应量大小,Aleottia 等(2020)通过经典配对样本 t 检验和贝叶斯配对样本 t 检验均得到,三个维度上的 SNARC 效应量大小无显著差异。经典配对样本 t 检验方法因简便实用而基本为前人研究所采用。但是,多重比较分析容易增加犯 I 型错误的概率,Aleottia 等(2020)所做经典配对样本 t 检验存在不可靠的可能性。比较 SNARC 效应量还可以有其他统计思路,如,在线性回归模型中引入虚拟变量方法,下面详细介绍。

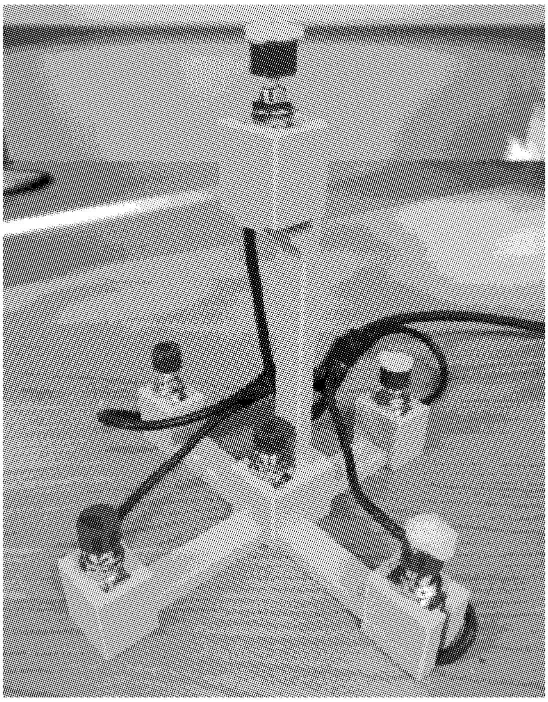

图 1 Aleottia 等(2020)研究中专门设计的反应盒

如果两个或多个样本组中的线性模型设定是相同的(即为同向),则这几组之间的回归系数大小是可以比较的,而且这种比较在多数实证分析中都是非常必要的。在数字加工的 SNARC 效应研究中也存在同样情况。以 Aleottia 等(2020)的研究为例,其实验得到三个维度(水平轴、垂直轴和矢量轴)的 SNARC 效应量大小显著且方向一致,则可确定数字认知存在心理的三维空间表征,因此比较三个维度(水平轴、垂直轴和矢量轴)的 SNARC 效应量大小是可行且必要的。由(1)式,三个维度的 SNARC 效应量的比较其实就是比较如下三个模型的回归系数大小是否存在差异:

水平轴 SNARC 效应模型: $DRT_H = \beta_{0H} + \beta_{1H}N + \mu_H$ (1a)

垂直轴 SNARC 效应模型: $DRT_V = \beta_{0V} + \beta_{1V}N + \mu_V$ (1b)

矢量轴 SNARC 效应模型: $DRTs = \beta_{0S} + \beta_{1S}N + \mu_S$ (1c)

考虑到三个模型中的回归系数和常数项不等,采取加法乘法式混合引入虚拟变量得到如下线性模型(Wooldridge, 2009/2010)。

$$DRT = \beta_0 + \delta_1 d_1 + \delta_2 d_1 * N + \delta_3 d_2 + \delta_4 d_2 * N + \beta_1 N + \mu$$

（2）

该模型以水平轴为对照，且 d1（矢量轴）和 d2（垂直轴）均为虚拟变量。当两个虚拟变量都取 0 时，则表示为水平轴 SNARC 效应模型，且为对照组；当垂直轴取 1，矢量轴取 0，则表示为垂直轴 SNARC 效应模型；当垂直轴取 0 时，矢量轴取 1，则表示矢量轴 SNARC 效应模型。N 为数字 1、2、3、4、6、7、8、9。因此，结合两个虚拟变量的不同取值，可得到：

水平轴 SNARC 效应模型（$d_1=d_2=0$）：

$$DRT=\beta_0+\beta_1 N+\mu \tag{2a}$$

矢量轴 SNARC 效应模型（$d_1=1, d_2=0$）：

$$DRT=(\beta_0+\delta_1)+(\delta_2+\beta_1)N+\mu \tag{2b}$$

垂直轴 SNARC 效应模型（$d_1=0, d_2=1$）：

$$DRT=(\beta_0+\delta_3)+(\delta_4+\beta_1)N+\mu \tag{2c}$$

据此，检验垂直轴、矢量轴和水平轴之间的回归系数大小是否存在显著差异，就是检验 $H_0: \delta_2$ 或 $\delta_4=0$。依此类推，如果（2）式中的虚拟变量 d_1 为水平轴、d_2 为矢量轴，垂直轴为参照记为 0，当两个虚拟变量都取 0 则表示垂直轴，这样垂直轴就是对照组；当水平轴取 1，矢量轴取 0，则表示水平轴（$d_1=1, d_2=0$）；当水平轴取 0，矢量轴取 1，则表示矢量轴（$d_1=0, d_2=1$）。据此，检验水平轴、矢量轴和垂直轴之间的回归系数大小是否存在显著差异，最终都可归结为检验 $H_0: \delta_2$ 或 $\delta_4=0$。

下面即是利用 Aleotti 等（2020）的公开数据通过 SPSS（SPSS Statistics 21，IBM，NY，USA）对模型（2）进行回归分析和 F 检验，最终得到方差分析表（表 1）和两个回归系数表（表 2 和表 3）。可以看到，不论以哪个轴为参照轴，方差分析表都是相同的（即都是表 1 形式），这表明存在 SNARC 效应。而且，当以水平轴为参照时，在回归系数大小比较上，其他两轴均与其无显著差异（表 2）；当以垂直轴为参照时，在回归系数大小比较上，其他两轴均与其无显著差异（表 3）。结果表明，存在数字的空间三维表征。因此，可以看到，通过引入虚拟变量对回归系数大小的分析和 Aleotti 等（2020）的分析结果是一致的。

表 1　方差分析表

	平方和	*df*	均方	*F*	Sig.
回归	23546.837	5	4709.367	11.167	0.000
残差	7590.847	18	421.714		
总计	31137.684	23			

表 2　回归系数表（水平轴为参照）

	非标准化系数		标准系数	*t*	Sig.
	B	标准误差			
（常量）	66.406	15.114		4.394	0.000
N	-11.804	2.651	-0.897	-4.452	0.000
d_1	-21.506	21.374	-0.281	-1.006	0.328
d_2	-38.798	21.374	-0.508	-1.815	0.086
d_1*N	1.006	3.749	0.079	0.268	0.791
d_2*N	2.417	3.749	0.190	0.645	0.527

表 3　回归系数表（垂直轴为参照）

	非标准化系数		标准系数	*t*	Sig.
	B	标准误差			
（常量）	27.609	15.114		1.827	0.084
N	-9.386	2.651	-0.714	-3.541	0.002
d_1	38.798	21.374	0.508	1.815	0.086
d_2	17.292	21.374	0.226	0.809	0.429
d_1*N	-2.417	3.749	-0.19	-0.645	0.527
d_2*N	-1.411	3.749	-0.111	-0.376	0.711

由前面分析可知，引入交乘项检验某个或某几个变量的系数是否存在组间差异，只需在一般线性模型中加入交乘项即可，但这一方法背后存在隐含的假设条件。对于数字认知的 SNARC 效应的一般线性模型（1）式，这个隐含的假设条件是，三组的误差项应具有相同的分布（因为估计时是将三组样本混合在一起进行估计的），即，$\sigma_H{}^2=\sigma_V{}^2=\sigma_S{}^2$。国内夏帆（2008）给出了证明，一元线性回归模型中引入一个虚拟变量时，其中的虚拟变量回归参数 b 的显著性检验和两正态独立总体均值比较 t 检验是等效的。通过前面实例分析得到，混合引入两个虚拟变量后的结果也是等效的。引入虚拟变量的好处之一是可能会减弱假阳性。

2　置信区间估计的可信问题分析

第一部分的分析中提到贝叶斯检验在心理学研究中已开始得到应用。贝叶斯学派认为概率是主观、先验和可变化的，是从有到变化；而频数派认为概率是从无到有到相对稳定，是客观的。若只从客观性角度看，似乎频数派的统计方法的可信度相对更高，但实际情况是否真是如此呢？而且现在心理学实验的检验结果报告除了列出 *t* 值（或 *z* 值）大小，还需要需要提供“*Cohen's d* 值”，也有进一步提供置信区间的。因此，和第一部分一样，下面仍从实例出发对置信区间的可信问题进行分析。需要指出的是，虽然依据实例分析统计问题研究是初步的，但从数学角度来看，其不失为一种发现某些统计问题甚至质疑某些统计观点的重要手段。

区间估计与假设检验是传统的推断统计中两种重要的统计方法。参数估计包括参数的点估计和区间估计。设 $x_1, x_2, \cdots x_n$ 是来自总体 X 的样本,总体 X 分布函数为 $F(x, \theta)$(θ 为未知参数),现建立两个统计量 $T_1(x_1, x_2, \cdots x_n)$ 及 $T_2(x_1, x_2, \cdots x_n)$ 并满足不等式 $T_1(x_1, x_2, \cdots x_n) < T_2(x_1, x_2, \cdots x_n)$,则称 $[T_1, T_2]$ 为随机区间 RI(Random Interval, RI)。设 $\alpha(0 < \alpha < 1)$ 为一给定常数,若满足关系式 $P\{T_1 \leqslant \theta \leqslant T_2\} = l - \alpha$, RI 作为参数 θ 的估计,则称 RI 是参数 θ 置信水平为 $1 - \alpha$ 的区间估计,此时 RI 被称为置信区间 CI(Confidence Interval, CI), α 为显著性水平, $T_1(x_1, x_2, \cdots x_n)$ 和 $T_2(x_1, x_2, \cdots x_n)$ 分别为上、下置信限。根据枢轴量分布,置信区间在一定置信度下可估计总体参数的可能范围。

2.1 例题及解答

例 1:假设某私立学校规定学生升学标准之一为近 6 次英语考试平均分不低于 101 分(总分为 120),对 A 学生进行了 6 次考试后得到如下分数:85,90,95,96,101,103。问 A 学生在 $\alpha = 0.05$ 下是否达到英语入学标准?

解法一:

样本均值和标准差分别为 $\bar{X} = 95, s = 6.723$,置信度为 0.95 的均值置信区间为 $(95 - 2.015 \times 6.723/\sqrt{5}, +\infty)$,计算得到 $(88.942, +\infty)$,101 落入其中,依据该置信区间 A 学生达到入学标准(置信区间法)。

假设检验解法如下:

①构建假设 $H_0: \mu_0 \geqslant 101, H_1: \mu_0 < 101$

②选取样本统计量 $t = \dfrac{\bar{x} - \mu_0}{\dfrac{S}{\sqrt{n-1}}}$

③对显著性水平 $\alpha = 0.05$,查 t 值表确定临 界值 $t_{0.05}(5) = -2.015$

④计算样本统计量 $t = \dfrac{95 - 101}{6.723/\sqrt{5}} = -1.996$

由于 $-1.996 > -2.015$,接受原假设 H_0, A 学生达到入学标准。

解法二:

置信度为 0.95 的均值置信区间为 $(-\infty, 95 + 2.015 \times 6.723/\sqrt{5})$,计算得到 $(-\infty, 101.058)$,101 落入其中,依据该置信区间 A 学生达到入学标准(置信区间法)。

假设检验解法如下:

①构建假设 $H_0: \mu_0 \leqslant 101, H_1: \mu_0 > 101$

②选取样本统计量 $t = \dfrac{\bar{x} - \mu_0}{\dfrac{S}{\sqrt{n-1}}}$

③对显著性水平 $\alpha = 0.05$,查 t 值表确定临 界值 $t_{0.05}(5) = 2.015$

④计算样本统计量 $t = \dfrac{95 - 101}{6.723/\sqrt{5}} = -1.996$

由于 $-1.996 < 2.015$,接受原假设 H_0, A 学生未达到入学标准。

2.2 分析与讨论

可以看到,假设检验下两种解法产生冲突。解法一拒绝域为 $(-\infty, -t_\alpha)$(记为 R_1),解法二拒绝域为 $(t_\alpha, +\infty)$(记为 R_2),解法一和二接受域存在一个交集区间 $(-t_\alpha, t_\alpha)$(记为 $R_{交}$)。当样本观测值落入 $R_{交}$ 时,两种解法都是接受原假设,从而得出截然相反的结果,这表明 $R_{交}$ 是无法作出接受 H(原假设)或 $\bar{H}$(备择假设)判断的区间。但两种解法的置信区间都得到 A 学生达到了入学标准,未产生冲突,置信区间的估计过程似乎可信。

解法二下的置信区间和假设检验产生冲突,因为 101 和置信区间的位置关系不可忽视。为更一般性讨论这个问题,若取某个置信区间(置信度为 $1 - \alpha$)并构建原假设 $H_0: p = 0$,在拒绝和接受 H_0 下,依据样本观测值和置信区间的位置关系进行估计时,以下两种结果均可能出现:

①接受 H_0,置信区间为 (a_1, b_1)

假设 0 落在 (a_1, b_1) 中,当 a_1, b_1 中的一个或都距离 0 很近,若取置信区间 $(a_1 = -0.02, b_1 = 0.04)$,假设检验和区间估计结论则可能产生冲突。当 a_1, b_1 都距离 0 较远,若取置信区间 $(a_1 = -15, b_1 = 20)$,假设检验和区间估计结论则趋向一致,但由于区间估计的上下限距离 0 较远,没有哪个区间值能有更大把握 在 0 附近。

②拒绝 H_0,置信区间为 (a_2, b_2)

假设 0 落在 (a_2, b_2) 外,但 a_2, b_2 中的一个或都距离 0 很近,若取置信区间 $(a_2 = 0.02, b_2 = 0.04)$,其区间内任意值均在 0 附近,这时,区间估计与假设检验结论可能存在冲突。而若取置信区间 $(a_2 = 15, b_2 = 20)$,区间上下限都和 0 相距较远,区间估计和假设检验结论则趋向一致。

因此,①和②都存在产生冲突的可能,解法二属于前述 1 中可能产生冲突的一种情形,从而对置信区间可信存有怀疑。

2.3　例 1 再分析——贝叶斯统计(解法三)

从例 1 所给 6 次考试的样本观测值看,A 学生仅有一次考试略高于 101,一次刚好合格,其余 4 次均不合格,这六次考试平均分数也只有 95,远低于入学标准线。分析例 1 的解法一中的置信区间(88.942,101.058)也可发现,101 虽然落入其中,但在这个区间中处于非常偏右的位置(距离区间上限很近),样本观测值大多数都小于 101。这种样本信息提示,$H_0:\mu_0 \leqslant 101$,$H_1:\mu_0 > 101$ 很可能会得到支持,即,A 学生未达到入学标准,这个结果和解法二的假设检验法相一致。由于贝叶斯统计方法的基本思想就是充分利用样本信息和先验分布,下面采用贝叶斯统计方法(解法三)对例 1 进行再分析。

第一步　假设存在先验分布

心理统计与测量中,学生考试成绩一般服从正态分布。假设 A 学生考试分数符合正态分布,$X \sim N(100, 7^2)$,该正态分布作为先验分布。

第二步　依据后验分布 进行计算

将例 1 中全部 6 个样本观测值作为样本信息并结合先验分布计算得到后验分布。后验分布依据下面三个公式:

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{ML} \qquad (1)$$

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2} \qquad (2)$$

$$\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n \qquad (3)$$

其中,μ_N为后验分布平均值,$\sigma_N{}^2$为后验分布方差,μ_0为先验分布平均值,$\sigma_0{}^2$为先验分布方差,σ^2为样本方差,μ_{ML}为样本平均值,x_n为第 n 个样本观测值,N 为样本容量。

例 1 中 6 个样本观测值平均分为 95,方差为 45.199,据式(1)和(2)(即结合先验分布计算公式)可得后验分布为 $X_N \sim N(95.666, 42.602)$。由于例 1 所给样本观测值只有 6 个,后验分布均值 95.666 与 6 个样本观测值均值(95)很接近,但小于先验分布均值 100 和入学标准 101 分,而 100 已很接近 101 分,所以假设的先验分布似乎未影响后验分布。95.666 小于 101(但变异减小),所以 A 学生未达到入学标准。本次分析支持前述解法二的假设检验法。进一步,在当前样本信息下 A 学生考试成绩的先验分布需要满足什么条件才可能使后验分布均值达到入学标准呢?根据式(1)列出如下二元一次不等式方程式(4)计算先验分布均值和方差范围。变形后得到 $45.2\mu_0 - 45.2 \times 101 \geqslant 600 \times \sigma_0{}^2$,最终得到,对于先验分布,当 $120 \geqslant \mu_0 \geqslant 101$ 且 $1.43 \geqslant \sigma_0{}^2 \geqslant 0$ 时,A 学生后验分布均值才可能达到入学标准 101 分。考虑极端情形($\sigma_0{}^2 = 1.43$),从 A 学生实际考试成绩来说,标准差为 1.2 的考试成绩意味,相差一分就会相差近一个标准差,实际出现这种情况的可能性不大。所以综合前面分析,就当前样本信息来说 A 学生实际未达到入学标准 101 分。

$$\frac{45.2}{6 \times \sigma_0^2 + 45.2} \times \mu_0 + \frac{6 \times \sigma_0^2}{6 \times \sigma_0^2 + 45.2} \times 95 \geqslant 101 \quad (4)$$

解法三在利用样本信息上又更进一步,还考虑到先验分布,因此处于相对更可信地位。但解法三支持解法二的假设检验法表明,解法一和二的置信区间估计的可信性值得怀疑。

2.4　讨论

例 1 解答共采用了三种方法,最后通过贝叶斯统计分析才给出当前样本信息下相对确定性结果(通过计算贝叶斯因子结果也一样)。概括而言这些解法主要回答如何依据样本信息进行方向判断(大于还是小于英语入学标准 101 分)。解答过程中虽利用置信区间进行估计,但其可信受到怀疑,所以对置信区间涵义及其可信问题还需在概念上进一步反思和澄清:

(1)区间估计是构造出包含参数的样本范围进行估计(置信区间),例如,对 $\alpha = 0.05$ 时的置信区间(0.1,0.2),认为参数有 95% 的可信度落在 0.1 ~0.2 之间是不够精确的,因为这种表述将参数看作随机变量,置信区间覆盖某参数的可能性才是对置信区间的精确理解;假设检验是对于给定参数值(原假设),寻找与该参 数不一致的样本(拒绝域)。置信区间是参数空间的子集,是一个随机区间,具体区间范围由样本观测值决定;假设检验的拒绝域是样本空间的子集,相对来说是确定的集合,没有随机性。区间估计与假设检验二者存在联系:置信区间与接受域对应,置信水平 $1-\alpha$ 与显著性水平 α 对应。区间估计属经典统计理论中的估计方法,是根据样本信息对总体分布或总体特征进行推断,只用到了总体信息和样本信息。而贝叶斯理论在使用前述两种信息以外,还使用了先验信息。先验信息是人们在试验之前对某问题的了解和已掌握的相关统计信息。与经典统计理论观点不同的是,贝叶斯理论认为 总体中的未知参数 θ 是随机变量,将先验信息(先验概率分布)加入统计推断中可提高统计推

断的质量。

(2)如果不考虑例 1 中样本观测值和置信区间的实际具体含义,推广到一般情形时,置信区间上下限值与参数之间的远近关系是置信区间不可信的重要原因。例 1 解法一和二的置信区间下的结果都是一致的,但解法二中的样本观测值的置信区间结果和假设检验结果是相冲突的,这令假设检验和置信区间估计的可信性都受到怀疑,但不能最终确定是哪个出了问题。解法一和二的置信区间虽然是不一致的,但结论相同,让人觉得置信区间是可信的。进一步通过贝叶斯分析后得到两种置信区间可信的存在基础是:考试成绩变异小,以及除这 6 次考试之外其他足够多次数的考试成绩都能≥101;但是这种情况实际发生的可能性未知,所以解法一和二的置信区间似乎又是不可信,因此这意味,通过贝叶斯统计方法找到置信区间估计可信或不可信的原因所在。但需要通过已有知识经验获得样本的先验分布(本质上即是需要符合一定条件的足够考试次数或实验重复次数)。另外,如果基于样本观测值和置信区间在例 1 中实际具体含义,则 101 分在例 1 中具体含义是英语入学分数,其落在置信区间(88.942,101.058)中。虽从连续变量角度看区间上限 101.058 是合理的,但作为实际考试分数,101.058 是不存在的。所以从样本实际情况来看 101 分作为边界更加合理,而恰巧这个分值又是区间上限,所以 A 学生很有可能就是未达到入学标准,例 1 中的先验分布变异很小而后验分布变异还会更小可进一步支持该观点。

(3)置信区间按某估计程序被建构和实际理解使用某个置信区间之间存在较大差异。按频数说,实验重复很多次后,利用很多样本信息可以计算得到很多个置信区间,真值存在于某一置信度(如 95%)下的置信区间里。但就某一个置信区间而言,真实值或在里面或不在里面,这种情形不好用概率去描述,这个置信区间里的样本观测值也不存在分布信息。因为实验重复很多次后,概率意义才成立,一定置信度下的置信区间才能建构出来并具有价值(总体参数在其中)。从这个角度出发,例 1 前两种解法可以理解为是通过置信区间的建构实现对真值估计:由解法一和解法二的置信区间法可知,两种解法构建的置信区间都包含了真值。这样,两者对真值进行估计的过程中,在目前样本信息下出现了两个置信区间的交集区间($R_{交}$)。理论上这两个置信区间不能相交,相交则意味 $R_{交}$ 置信度会低于 95%(样本信息完全相同),而作为 95% 置信区间子集 $R_{交}$ 置信度实际上就是 95%,于是矛盾产生了。但是,实验重复很多次后就不存在这样的问题,而这种理论上的构建和实际某样本信息下构建某个置信区间过程之间的差异一直存在,是抽样研究所固有的特点。

3 结束语

(1)包含虚拟变量的线性回归模型是一种特殊的回归模型,引入虚拟变量增加了线性回归模型的复杂性,使一个线性回归方程能融合至少两个线性回归方程在其中,对回归问题的描述(尤其是包含定性问题时)更凝聚清晰和接近现实。一元线性回归方程中的斜率反映 SNARC 效应量,其差异性检验通常使用 t 检验,这种方法简便易行,但并未充分利用线性回归方程的自身特性,存在增加假阳性的可能。在 SNARC 效应的一元线性回归模型中引入虚拟变量后,通过使用回归模型的总体显著性检验 F 检验可以替代 t 检验,即与比较 两个独立的正态分布的总体的均值的 t 检验结果是一致的,可以降低结果的假阳性。但对于 SNARC 效应量的分析来说,引入虚拟变量至线性回归模型后增加了分析的难度。

(2)由前述分析,置信区间估计的不可信主要看参数和置信区间的上下限之间的位置关系以及实验重复次数,在具体问题中还离不开样本观测值的具体含义。但最关键一点,置信区间可以反映样本信息并在实验重复多次后能反映总体参数(一定条件下样本可以替代总体),这是置信区间合理可信的基础。所以,总体上置信区间合理可信,且含有重要量化信息(如,可利用置信区间研究效应重复性问题),可以一定程度上补充经典假设检验。可信的终极价值应是指向对真值的更准确估计,并随样本信息增加,该分配多大概率(或把握)给这些样本信息以表明依据它们所估计的真值能准确到什么程度。对于频数说的置信区间,实验重复很多次后才能赋予置信区间某个置信度(如,95%)下包含真值。所以在建构或估计真值过程中基于样本信息的可信都是相对的,但在样本信息不是很足够时就能充分利用已有知识经验实现估计的统计学理论或方法可以胜出,因而贝叶斯统计理论有其相对优势,将贝叶斯统计引入心理学实验结果的检验并报告出来是非常有必要的。

参考文献

巢思娇.(2019).*跨视听通道下小学生、初中生与大学生的负数SNARC效应*(硕士学位论文).苏州大学.

李贵玉,顾昕.(2021).贝叶斯统计方法的应用与现状.*心理学探新*,*41*(5),466－473.

夏帆.(2008).虚拟变量回归与t检验等效性的探讨.*统计研究*,*25*(10),109－110.

赵琪,李龙凤,邓之君,陈英和.(2017).数量空间表征的特点:跨视听通道下的SNARC效应.*心理发展与教育*,*33*(2),129－138.

Aleotti, S. , Girolamo, F. D. , Massaccesi, S. , & Priftis, K. (2020). Numbers around descartes: A preregistered study on the three－dimensional SNARC effect. *Cognition*, *195*, 0010－0277.

Dehaene, S. , Bossini, S. , & Giraux, P. (1993). The mental representation of parity and number magnitude. *Journal of Experimental Psychology General*, *122*(3), 371－396.

Kruschke, J. K. , & Liddell, T. M. (2018). The bayesian new statistics: Hypothesis testing, estimation, meta－analysis, and power analysis from a bayesian perspective. *Psychonomic Bulletin & Review*, *25*, 178－206.

Morey, R. D. , Hoekstra, R. , Rouder, J. N. , et al. (2015). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, *23*(1), 103－123.

Shaki, S. , & Gevers, W. (2011). Cultural characteristics dissociate magnitude and ordinal information processing. *Journal of Cross Cultural Psychology*, *42*(4), 639－650.

Case Study on SNARC Effect Size and the Credibility About Estimation of Confidence Interval

He Hua

(School of Education, Soochow University, Suzhou 215123)

Abstract: Two problems (SNARC effect size and confidence interval) are analyzed by some cases. (1) Effect size is a very important part in the report of the psychological experiment results, the research inquires into the statistical methods of comparing the size on SNARC effect in the field of number cognition, which is different from the common definitions and methods on effect size. The research raises a new statistical method (Mixed dummy variable being introduced into linear regression model) to process the open data in Aleottia et. al. (2020) study, although more complicated than general methods. (2) Interval estimation of parameters is a basic form of statistical inference. According to the distribution of the pivot quantity, confidence intervals can estimate the possible range of the population parameter undera certain confidence. The research analyzes the estimation process of a certain confidence interval by constructing a special example, and makes a comparative analysis with Null Hypothesis Significance Testing and Bayesian statistics. The results show that although the confidence interval is reliable in population, there is a situation that the accurate judgment cannot be made according to a certain confidence interval, and the reason can be analyzed through Bayesian statistics.

Key words: SNARC effect; dummy variable; Null Hypothesis Significance Testing; Confidence Interval; prior distribution; posterior distribution; Bayesian Statistics