平均数差异显著性检验统计检验力和效果大小的估计原理与方法

胡竹菁

（江西师范大学 心理学院，南昌 330022）

摘　要：该文以平均数差异显著性检验为例，对实验数据进行假设检验后，继续对其统计检验力和效果大小进行估计的基本原理和方法作一介绍。

关键词：平均数差异显著性检验；假设检验；统计检验力；效果大小

中图分类号：B841.2　　**文献标识码**：A　　**文章编号**：1003－5184(2010)01－0068－06

1　引言

在心理学研究中，应用 Z 检验、t 检验、F 检验或卡方检验等推断统计方法会得到对某种类型的样本统计量（如样本平均数、方差、相关系数、比率或实计数等）的几个样本统计量之间的差异是否显著的假设检验结果，但是，假设检验结果并不能有助于了解它们之间的差异到底有多大，其差异显著性有多重要。因此，现代心理统计学的发展要求通过计算统计检验力（或效果大小）来达到这些目的。当今国内外心理与教育统计学最重要的发展趋势之一是有关“统计检验力（power of test）和效果大小（effect size）”的计算方法问题越来越重要了。随着心理学科的发展，目前美国心理学会的学术期刊已经明确要求研究者在投稿时需要在文章中提供有关“统计力”和“效果大小”等方面的数据。例如，美国心理协会最新版（第 5 版）的《写作手册》一书中，明确要求：“作者对于自己的研究假设进行检验时，必须考虑采取严格的统计力（statistics power）。我们可以通过特定的 α 水平、效果大小和样本大小来决定统计力，而这关系到正确地拒绝作者想要检验的假设的可能性……为了让读者能够充分地了解到你的研究发现的重要性，在你的结果段落中呈现效果大小（effect size）的索引或关系强度（strength of a relationship）是必要的。你可以使用一些一般效果大小的估计值来估计你研究结果的效果大小或关系强度，包括（但不是受限于）：γ^2，η^2，ω^2……Cohen 的 d 值和 k 值……”[1]。在我国现有的心理或教育统计教材中，比较重视如何控制 α 型错误的问题，对于如何计算和控制 β 型错误并由此提高统计检验力问题则较少介绍[2－5]。鉴于目前国外较为著名的心理统计学教材[6－9]都重视统计检验力和效果大小的估计原理与估计方法，文章拟以平均数差异显著性检验为例，对实验数据进行假设检验后，继续对其统计检验力和效果大小进行估计的基本原理和方法进行介绍。

2　统计检验力的含义与估计原理

对平均数进行差异显著性检验时，通常将检验的虚无假设设为 $H_0:\mu_1=\mu_0$，或 $\mu_1-\mu_0=0$，这是假设 μ_1 与 μ_0 在统计学意义上，两个平均数之间实质上是没有显著差异的；而将虚无假设的反面设为 $H_1:\mu_1\neq\mu_0$，这是假设 μ_1 与 μ_0 在统计学意义上，两个平均数之间实质上是有显著差异的。根据假设检验的结果，无论是拒绝或者不拒绝虚无假设，都有可能或者犯 α 错误或者犯 β 错误。通常情况下 α 和 β 不可能同时增大或减小，α 和 β 的相互关系通常有如表 1 和图 1 所示。

表 1　假设检验的各种可能结果

	接受 H_0	拒绝 H_0
H_0 为真	正确决策，概率＝$1-\alpha$＝置信度	第一类错误，概率＝α＝检验水平
H_0 为假	第二类错误，概率＝β	正确决策，概率＝$1-\beta$＝统计检验力

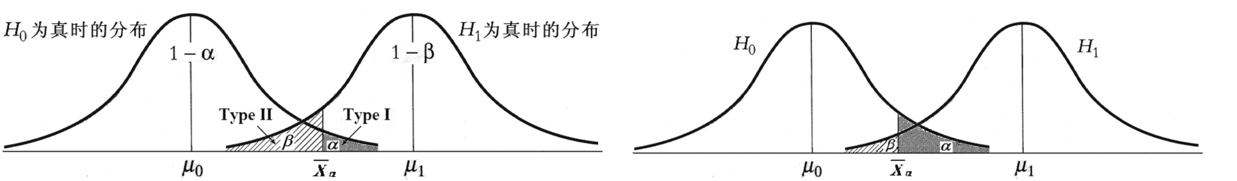


图 1　α 和 β 关系示意图

当虚无假设是 $H_0: \mu_1 = \mu_0$ 时，虚无假设分布(null hypothesis distribution, NHD)是以零为中心的正态分布，以对平均数的检验为例，所谓虚无假设分布，就是指当虚无假设 H_0 为真时，$\overline{X} - \mu_0$ 或 $\overline{X}_1 - \overline{X}_2$ 的分布。由于可以通过预先设定 α 水平的方式来控制当虚无假设为真时拒绝它可能会犯错误的概率。因此，在此基础上得到的虚无假设差异显著性检验的 Z 统计量分布(或 t 统计量分布)在置信度范围内也是以零为中心的分布(a central Z (or t) distribution)。

由表 1 可知，当虚无假设 H_0 为假(备择假设 H_1 为真)时，接受 H_0 就会犯 β 型错误；拒绝 H_0，则是做出了正确的决策，其概率等于 $1-\beta$。换言之，当 H_1 为真，即 μ_1 与 μ_0 确实有差异时，μ_1 与 μ_0 的距离即表示 μ_1 与 μ_0 的真实差异，以 $1-\beta$ 的概率接受 H_1。$1-\beta$ 反映着正确辨认真实差异的能力，统计学中称之为统计检验力(power of test)或效力。也可以把统计检验力 $1-\beta$ 定义为："在虚无假设 H_0 为假(备择假设 H_1 为真)时，正确拒绝 H_0 的概率"。

如果想要知道犯 β 型错误的概率是多少，就要知道备择假设的分布情况(alternative hypothesis distribution, AHD)，以对平均数差异显著性的检验为例，所谓备择假设分布，就是指当虚无假设 H_0 为假，而备择假设 H_1 为真时，$\overline{X} - \mu_0$ 或 $\overline{X}_1 - \overline{X}_2$ 的分布。遗憾的是，由于与虚无假设相对立的备择假设 $H_1: \mu_1 \neq \mu_0$ 是虚无假设 $H_0: \mu_1 = \mu_0$ 的补集，两个平均数之间不相等的值几乎是无限多的，因此，它们之间的差值到底是多少是不确定的。在理论上，备择假设 $H_1: \mu_1 \neq \mu_0$ 的分布值是一个不是以零为中心的分布(a noncentral Z(or t) distribution)，它是以什么值为中心也有着无数多个选择，因此，一般情况下，难以对 β 值或 $(1-\beta)$ 值作出准确的估计。

另一方面，从总体上说，对统计检验力的分析是一个近似值的分析，因此可以假设备择假设分布(AHD)是一种正态分布，在此基础上得到的备择假设差异显著性检验的 Z 统计量分布(或 t 统计量分布)也是正态分布。如果能找到某种备择假设分布的集中值，就能对 β 或 $(1-\beta)$ 值作出某种近似的估计。

虽然在一般情况下由于不知道备择假设分布是怎样的而无法精确计算 β 型错误的值，但在以进行平均数差异检验为目的的抽样实验中总会得到一个 Z(或 t)统计量，如果备择假设分布服从正态分布，那么就可以利用这个在一次性抽样中所得到的 Z 值来作为估计备择假设分布的中心点，以此作为该次实验中备择假设分布的期望值(The expected Z value，或称为备择假设分布的平均值)，通常用希腊字母"δ"表示备择假设分布的期望 Z 值。并假定在备择假设的分布中，大于这个 Z 统计量的备择假设的数量与小于这个 Z 统计量的备择假设的数量各占 50%，由此来对可能犯 β 型错误的概率进行近似的估计。

由于 δ 是表示某种特殊的备择假设分布的平均 Z 值，因此它的计算方法也与 Z 值的计算方法密切关联。由于在估计 δ 值时，一方面假定备择假设的分布是正态分布，另一方面也用一次抽样中所获得的假设检验的 Z 统计量来作为 δ 值，因此，两个独立样本平均数差异显著性检验的 δ 值的计算方法将以下列公式 1 所示的计算 Z 统计量的公式为基础。

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \qquad \text{(公式 1)}$$

如前所述，在两个独立样本平均数差异显著性检验中，δ 是表示某种特殊的备择假设分布的期望 Z 值(在 t 检验中就是期望 t 值)，而样本平均数 $\overline{X}_1$ 的期望值是 μ_1，样本平均数 $\overline{X}_2$ 的期望值是 μ_2，与此相应，计算两个独立样本平均数差异显著性检验的 δ 值的公式有如下列公式 2 所示。

$$\delta = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \qquad \text{(公式 2)}$$

例如，当对两个实验分组的实验数据进行平均数差异显著性检验时，如果用公式 1 计算的平均数差异显著性检验统计量为 $Z=3.0$，那么就可以把"$Z=3.0$"视为是公式 2 所得到的"δ 值"的计算结果，并且把"所有可能的备择假设的分布"假设为是以 3.0 为中心的正态分布，这时，这次抽样检验结果 $Z=3.0$ 就被视为该次实验中备择假设分布的期望值，这有如图 2 所示。

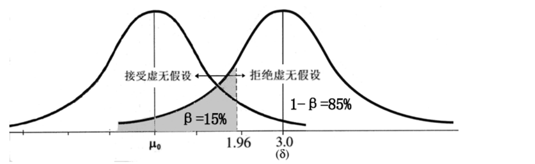


图 2　$\delta=3.0$ 时的 δ 值分布及统计检验力示意图

在 $\alpha=0.05$ 水平上进行双侧检验时，作出接受或拒绝虚无假设的临界值是 $|Z_{\frac{\alpha}{2}}|=1.96$。以此为分界点，通常的统计决策是：当实际得到的 Z 值小于 1.96 时，就认为没有充分理由拒绝虚无假设，这时在虚无假设为假，备择假设为真时有可能犯"拒真的 β 型错误"，其可能犯 β 型错误的概率值有如图 2 中阴影部分所示；而当实际得到的 Z 值大于 1.96，就会拒绝虚无假设，这时在虚无假设为假时就作出了正确的决策，由于在正态分布条件下与 $\alpha=0.05$ 相应的 Z 值是 1.96，因此，其正确拒绝虚无假设的概率值有如图 2 中以 1.96 为分界线所示的右边部分的面积，它等于 $(1-\beta)$。由于做出决策的临界值定为 1.96，它离备择假设分布的期望 Z 值 $\delta=3.0$ 约一个标准差，即 $3.0-1.96=1.04$，查正态分布表可知，阴影部分的面积(也就是犯 β 型错误的概率)约

为 0.15，统计检验力($1-\beta$)则为 $1-0.15=0.85$。

3　两独立样本平均数差异显著性检验统计检验力的估计方法

根据上述统计检验力的估计原理，可以通过以下几个步骤来估计两个独立样本平均数差异显著性检验统计检验力 $1-\beta$ 的值：

步骤 1：根据已知条件建立需要检验的假设；

步骤 2：用相应的公式计算 Z 统计量；

步骤 3：确定做出统计决策的 α 水平及相应的临界值；

步骤 4：计算实际得到的 Z 值与 α 水平临界值的差；

步骤 5：根据 Z 值与 α 水平临界值的差查正态分布表，确定可能犯的 β 型错误或统计检验力 $1-\beta$ 的概率。

例如，有研究者在甲乙两校中分别抽取 100 名 16 岁的男生进行智商测查，测得甲乙两校该年龄组男生总智商的平均分分别为 115 分和 110 分。根据常模，该年龄组男生总智商的标准差是 15 分。那么，求取甲乙两校 16 岁男生平均智商差异显著性检验统计检验力的过程有如下述。

解：已知数据是：

$$n_1=100,\overline{X}_1=115,\sigma_1=15$$

$$n_2=100,\overline{X}_2=111,\sigma_2=15$$

步骤 1：建立假设：$\begin{matrix} H_0:\mu_1=\mu_2 \\ H_1:\mu_1\neq\mu_2 \end{matrix}$

步骤 2：用公式 1 计算检验统计量：

$$Z=\frac{(\overline{X}_1-\overline{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}=\frac{115-111}{\sqrt{\frac{15^2}{100}+\frac{15^2}{100}}}=\frac{4}{2.12}=1.89$$

步骤 3：确定做出统计决策的 α 水平并做出决策：在显著性水平 $\alpha=0.05$ 时，其临界值为 $|Z_{\frac{\alpha}{2}}|=1.96$，由于 $Z=1.89<Z_{\frac{\alpha}{2}}=1.96$，因此，正常情况下做出的决策是：没有充分的把握推论差异不是由偶然因素造成的，换言之，没有充分理由去拒绝虚无假设而接受虚无假设。由此推断两校男生的总智商没有显著差异，其差异是由于偶然因素造成的。

在一般情况下，假设检验的过程至此已经完成。但是，如果要知道假设检验结果的统计检验力是多少，则还需要做以下的工作。

步骤 4：计算实际得到的 Z 值与 α 水平临界值的差，得到 $1.89-1.96=-0.07$。

步骤 5：根据 Z 值与 α 水平临界值的差查正态分布表，确定可能犯的 β 型错误或统计检验力 $1-\beta$ 的概率：查正态分布表可知，从中心点为零到右边 0.07 个标准差所占的面积是 0.0279，加上中心点左边的 0.50 的面积，共有曲线下面积 $0.50+0.0279=0.5279$，约等于 53%（如图 3 所示）。这就是本例情况下估计出来的犯 β 型错误的概率值，其拒绝错误的虚无假设，接受正确备择假设的概率是($1-\beta$)，这就是统计检验力，在本例中，它的值是 $1-0.53=0.47$。

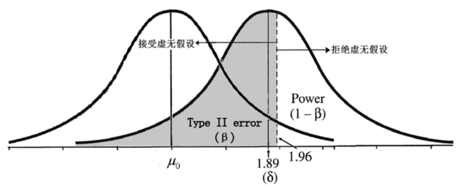


图 3　本例中的备择假设分布图($\alpha=0.05$)

如果用 Z_β 来表示图 3 中“拒绝或接受虚无假设”分界点到该正态分布曲线中点的距离，那么前述求取 β 或 $1-\beta$ 过程中的前 4 步也可以用下列公式 3 计算。

$$Z_\beta=\frac{(\overline{X}_1-\overline{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}-Z_\alpha \qquad \text{（公式 3）}$$

在本例中，假设 $u_1=u_2$，当显著性水平为 $\alpha=0.05$ 时，$Z_\alpha=1.96$，代入公式 3 可得：

$$Z_\beta=\frac{(\overline{X}_1-\overline{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}-Z_\alpha=\frac{115-111}{\sqrt{\frac{15^2}{100}+\frac{15^2}{100}}}-1.96=1.89-1.96=-0.07$$

由此得到的 $|Z_\beta|=0.07$ 与前述步骤 1 至步骤 4 得到的结果是一样的。

上述计算表明，根据本例中的数据对平均数差异显著性进行检验后，得到的 Z 统计量是 $Z=1.89$，当显著性水平为 $\alpha=0.05$ 时，$Z=1.89<Z_{\frac{\alpha}{2}}=1.96$，因此，接受虚无假设，认为两个样本平均数之间没有显著差异，作出这种决策的统计检验力约为 47%。这种计算平均数差异显著性检验统计检验力的过程只是评估方法之一，另外一种评估平均数差异显著性检验统计检验力的方法是表格换算法，换算的根据之一是效果大小(effect size)d 值的评估值。正因为统计检验力 $1-\beta$ 值与效果大小有着密切关联，因此，著名统计学家 J. Cohen(1992)把统计检验力 $1-\beta$ 定义为：“显著性检验的统计检验力是在给定总体效果大小(effect size)，显著性水平 α 和样本容量 N 的条件下拒绝 H_0 的概率”[10]。

4　两独立样本平均数差异显著性检验效果大小的估计方法

效果大小(effect size,有时也被译为效应大小或效应值)是反映统计检验效果大小或处理效应大小的重要指标,它表示不同处理下的总体平均数之间差异的大小,可以在不同研究之间进行比较。考虑到“效应”一词在统计学中很多地方使用,如方差分析中的主效应,各因子效应等,为了不使读者产生不必要的混淆,用效果大小一词来表示“effect size”的内涵。效果大小反映了两个总体受某种事物的影响后的差异程度。平均数差异显著性检验的效果大小一般用符号 d 表示。

在两个独立样本的方差和样本容量都相等的条件下,公式 2 可作如下推导:

$$\delta=\frac{(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2+\sigma_2^2}{n}}}=\frac{(\mu_1-\mu_2)}{\sqrt{\frac{2\sigma^2}{n}}}=\frac{(\mu_1-\mu_2)}{\sqrt{\frac{2}{n}}\sqrt{\sigma^2}}=\frac{(\mu_1-\mu_2)}{\sigma\sqrt{\frac{2}{n}}}=\frac{(\mu_1-\mu_2)\sqrt{\frac{n}{2}}}{\sigma}$$

,由此可得:

$$\delta=\frac{(\mu_1-\mu_2)}{\sigma}\sqrt{\frac{n}{2}} \qquad (公式 4)$$

令
$$d=\frac{(\mu_1-\mu_2)}{\sigma} \qquad (公式 5)$$

则有
$$\delta=d\sqrt{\frac{n}{2}} \qquad (公式 6)$$

进而有
$$d=\frac{\delta}{\sqrt{\frac{n}{2}}}=\delta\sqrt{\frac{2}{n}} \qquad (公式 7)$$

由公式 6 可知,在实际计算过程中,δ 值是 d 值与样本量除以 2 之后的算术平方根两部分的乘积。这表明,在其他条件不变的情况下,d 值越大,δ 值也会越大,而 δ 值越大,则统计检验力 $1-\beta$ 也越大,因此,d 值与统计检验力 $1-\beta$ 之间存在正相关的关系。

d 值与统计检验力 $1-\beta$ 之间有着密切关系,但也要注意这两者之间的区别:统计检验力 $1-\beta$ 受样本容量影响较大,而 d 值则不受样本容量影响。可以通过上述公式 5 在不考虑样本容量的情况下来计算某个心理研究中的 d 值。我们也可以通过把 d 值视为两个总体分布的重叠量的途径来理解效果大小的内涵。图 4 列出了四种 d 值在两个总体中的重叠情形。

图 4 表明,表示效果大小的 d 值越大,重叠程度就越小,平均数差异显著性检验的效果就会越明显;d 值越小则相反。也可以这样来理解,即不管你取哪种样本,d 值总是作为一种标准的平均数差异的估计,它与当前样本无关。显然,传统的推断统计量 Z、t、F 或 χ^2 值及相应的概率值 p 值只是说明平均数的差异如何,但这种差异脱离样本推广到不同的抽样群体时,差异究竟有多大则需要用反映效果大小的 d 值来描述。比如说,一般情况下,男性平均身高要比女性要高,反映这种平均数的 d 值就可以这样来解释,d 值越大则更容易在实际生活中看到男性比女性要高,d 值越小则在实际生活中看到男性比女性高的可能性就会小些。换言之,d 值更能说明在实际生活中所关心的差异,而不是数据上的差异显著问题。d 值能知道观测到的差异是不是事实上的差异。

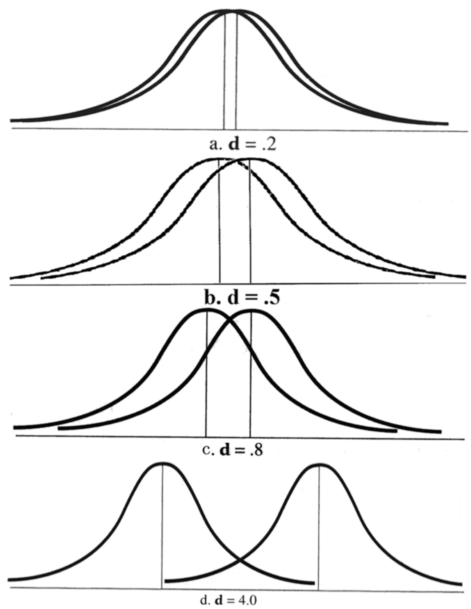


图 4　作为效果大小函数的两总体重叠图

J. Cohen(1992)认为,$d=0.2$ 视为是低效果,$d=0.5$ 视为是中等程度的效果大小,$d=0.8$ 视为是高效果。

上述公式 5 就是求效果大小的具体方法,利用这一公式,可以对前面求统计检验力的相应数据求该研究的效果大小 d 值。已知的条件是:

$$n_1=100,\overline{X}_1=115,\sigma_1=15$$

$$n_2=100,\overline{X}_2=111,\sigma_2=15$$

由于 $\sigma_1=\sigma_2=15$,因此可以利用公式 5 来计算该研究的 d 值,将数据代入公式后得:

$$d=\frac{(\mu_1-\mu_2)}{\sigma}=\frac{115-111}{15}=0.27$$

即该研究的效果大小是 $d=0.27$,根据 J. Cohen 的观点,这属于较小程度的效果大小。

5　影响平均数差异显著性检验统计检验力的其他因素

前面论述了效果大小与统计检验力的相互关系,即在其他条件不变的情况下,效果大小 d 值越

大，统计检验力 $1-\beta$ 值也越大。此外，统计检验力 $1-\beta$ 还受“α 水平”和“样本量的大小”这两个因素的影响。

公式 6 即 $\delta=d\sqrt{\frac{n}{2}}$ 表明，δ 值是 d 值和样本容量除 2 后的算术平方根的乘积。这表明，样本容量也是影响 δ 值的因素之一，样本容量越大，δ 值也越大。

对于另外一个即“α 水平”因素，需要作进一步的说明。正常情况下，当选择 $\alpha=0.05$ 的临界水平时，在正态分布的情况下，其相应的 Z 值为 1.96。以此为分界点，通常的统计决策是：当实际得到的 Z 值大于 1.96 时就会拒绝虚无假设；实际得到的 Z 值小于 1.96 时就会接受虚无假设。例如，前面所举的例子中的检验结果是 $Z=1.89$，由于小于 $Z_{(\alpha=0.05)}=1.96$ 的临界值，因此一般是接受虚无假设，认为甲乙两校男生在总智商方面没有显著差异。做出这种决策在虚无假设为假时就有可能犯 β 型错误。如图 3 所示，犯 β 型错误的概率约等于 53%，与此相应的统计检验力的值就是 $1-0.53=0.47$。

如果做出决策时把 α 水平从 $\alpha=0.05$ 改为 $\alpha=0.10$ 的水平进行双尾检验(其接受或拒绝虚无假设的临界点等同于在 $\alpha=0.05$ 水平下进行单尾检验时的临界点)情况会是怎样呢?

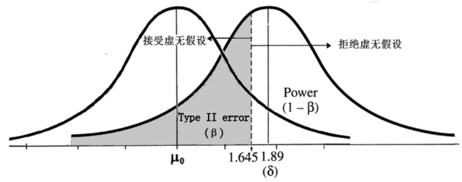
我们仍然用前面所举的例子为例。通过前两个求解步骤得到的检验统计量仍为 $Z=1.89$。

步骤 3：确定做出统计决策的 α 水平并做出决策：查正态分布表可知，在显著性水平 $\alpha=0.10$(双侧)时，其临界值为 $Z_{\frac{\alpha}{2}}=1.645$，由于 $Z=1.89>Z_{\frac{\alpha}{2}}=1.645$，因此，正常情况下做出的决策是拒绝虚无假设而接受备择假设，认为甲乙两校男生在总智商方面，甲校男生的总智商比乙校男生的总智商要更好。

步骤 4：计算实际得到的 Z 值与 α 水平临界值的差，得到 $1.89-1.645=0.245$。

步骤 5：根据 Z 值与 α 水平临界值的差查正态分布表，确定可能犯的 β 型错误或统计检验力 $1-\beta$ 的概率：查正态分布表可知，查正态分布表可知，从中心点为零到左边 0.245 个标准差所占的面积是 0.0968，约等于 0.10，如图 5 所示，其阴影部分的概率值为 $0.50-0.10=0.40$。

与前述一样，$Z=1.89$ 的检验统计量作为估计备择假设分布的中心点(或期望值 δ)，建构一个正态分布(1.89 左右两边各占 50% 的面积)，以 $Z=1.645$ 作为接受或拒绝假设的分界点，如果实际获得的 Z 值小于 $Z=1.645$(图 5 左边阴影部分)就接受虚无假设，这时如果虚无假设为假则犯 β 型错误的概率值约为 40%；如果实际获得的 Z 值大于 $Z=1.645$(图 5 右边部分)就拒绝虚无假设，这时如果虚无假设为假则是作出了正确的决策，其正确决策的概率值是 $(1-\beta)=1-0.40=60\%$。



zz 图 5　备择假设分布图($\alpha=0.10$ 双侧检验)

需要特别强调的是，当用上述五个步骤的方法或公式 3 来计算 β 型错误率或统计检验力 $1-\beta$ 时，一定要根据是接受还是拒绝虚无假设的假设检验结果，来确定通过 Z 统计量与显著性水平临界值之差查 Z 分布表后所得到的概率值的两种不同意义。

意义 1：当根据平均数差异显著性检验结果接受虚无假设(如本文例子在 $\alpha=0.05$ 双侧检验时)，在虚无假设是真时，这是正确决策；在虚无假设是假时，就做了错误的决策，犯错误的概率是 β；在本文例子中，检验结果得到的 Z 统计量即 $Z=1.89$ 与显著性水平 $\alpha=0.05$ 双侧检验的临界值 $Z_{\frac{\alpha}{2}}=1.96$ 之差为 0.07，查 Z 分布表可得对应于平均数到 0.07 个标准差的面积是约 0.03，如图 3 所示，由于是接受虚无假设，那么，临界值 $Z_{\frac{\alpha}{2}}=1.96$ 左边阴影部分的面积都属于接受域，图中 1.89 是这次实验中备择假设分布的期望值，该点两边各占 50% 的面积，而临界点 1.98 这个值在中心点 1.89 的右边，因此，阴影部分的面积就应该是 $0.50+0.03=0.53$，如果虚无假设是假，接受它做出错误决策的概率为 $\beta=0.53$。

意义 2：当根据平均数差异显著性检验结果拒绝虚无假设(如本文例子在 $\alpha=0.10$ 双侧检验时)，在虚无假设是真时，错误决策的概率是显著性水平 α；在虚无假设是假时，就做了正确的决策，正确决策的概率是 $1-\beta$。在本文例子中，检验结果得到的 Z 统计量即 $Z=1.89$ 与显著性水平 $\alpha=0.10$ 双侧检验的临界值 $Z_{\frac{\alpha}{2}}=1.645$ 之差为 0.245，查 Z 分布表可得对应于平均数到 0.245 个标准差的面积是约 0.10。如图 5 所示，与图 3 相似，临界值 $Z_{\frac{\alpha}{2}}=1.645$ 这次实验中是备择假设分布的期望值期望值，该点两边各占 50% 的面积，其左边阴影部分的面积属于接受域，由于 1.645 这个值在中心点 1.89 的左边，因此，阴影部分接受域的面积就应该是 $0.50-0.10=0.40$，如果虚无假设是假，这时拒绝它所做出正确决

策的概率为 $1-\beta=1-0.40=0.60$。

统计检验力与 α 水平的关系还可以通过下列图 6 来进一步理解。该图表明，对于任何的 δ 值，以及由横轴表示的相应于该 δ 的统计检验力的值，其变化将是 α 的函数。例如，在 $\delta=2.6$ 的情况下，当 $\alpha=0.10$ 时，其统计检验力为 0.83；当 α 水平降低到 $\alpha=0.05$ 时，统计检验力也降低到 0.74；当 α 水平进一步降低到 $\alpha=0.01$ 时，统计检验力就降低到只有 0.50 了。

上述分析表明，统计检验力 $1-\beta$ 与统计检验过程中进行决策时的 α 水平是有密切关联的，统计检验力 $1-\beta$ 的值随着 α 值的减小而降低。随着 δ 值的增大而增大。

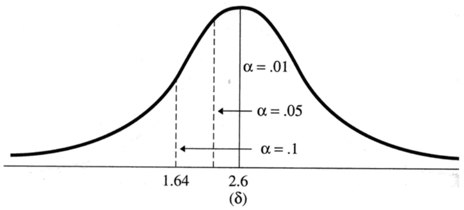


图 6　作为固定 δ 值的 α 的函数的统计检验力示意图

6　总　结

以平均数差异显著性检验为例介绍了假设检验后估计统计检验力和效果大小的基本原理和估计方法。

估计统计检验力的基本原理是以对两种假设分布的分析为基础。

虚无假设分布是以零为中心的正态分布，由于我们可以通过预先设定 α 水平的方式来控制当虚无假设为真时拒绝它可能会犯错误的概率。因此，在此基础上得到的虚无假设差异显著性检验的 Z 分布（或 t 分布值）在置信度范围内也是以零为中心的分布（a central Z (or t) distribution）。

由于两个平均数之间不相等的值几乎是无限多的，因此，备择假设 $H_1:\mu_1\neq\mu_0$ 的分布值是一个不是以零为中心的分布，它以什么值为中心也有着无数多个选择，因此，一般情况下是难以对 β 或 $(1-\beta)$ 值作出准确的估计的。当如果假设备择假设分布服从正态分布，用在一次性抽样中所得到的 Z 值来作为估计备择假设分布的中心点，以此作为该次实验中备择假设分布的期望值 δ，并假定在备择假设的分布中，大于 δ 值的备择假设的数量与小于 δ 值的备择假设的数量各占 50% 时，就可以由此来对可能犯 β 型错误的概率进行近似的估计。

统计检验力与效果大小具有密切的关系。效果大小 d 值是反映统计检验效果大小或处理效应大小的重要指标，是指两个总体受某种事物的影响后的差异程度。它与显著性水平 α 和样本容量 n 都是统计检验力评估中的重要影响因素。

总之，当检验结果是接受虚无假设时，人们更关注犯 II 型错误的概率 β 值的大小，或其反面即统计检验力 $1-\beta$ 值的大小，β 值越小（也即统计检验力越大），那么犯纳伪的可能性越小；当检验结果是拒绝虚无假设时，人们更关注两个平均数之间的实际差异即效果大小 d 值的大小，d 值越大，反映实际差异越大。

参考文献

1　美国心理协会．美国心理协会写作手册（第五版）．陈玉玲等译．重庆大学出版社，2008：14－15
2　张厚粲，孟庆茂，冯伯麟．心理与教育统计学．北京：北京师范大学出版社，1988.
3　张敏强．教育与心理统计学（修订版）．北京：人民教育出版社，2002.
4　甘怡群，等．心理与行为科学统计．北京：北京大学出版社，2005.
5　舒华，张亚旭著．心理学研究方法．北京：人民教育出版社，2008.
6　Barry H. Cohen. Explaining psychological statistics. 3rd Edition. New York University，2008.
7　Arthur Aron，Elaine N. Aron. Statistics for psychology. 影印版．北京：世界图书出版公司，2006.
8　David C. Howell. Fundamental statistics for the behavioral sdience. Brooks/Cole Publishing Company，1999.
9　Richard P. Runyon. Fundamentals of behavioral statistics. 北京：人民邮电出版社，2004.
10　Cohen J. A power primer. Psychological Bulletin，1992，112：155－159.

The Principle and Method of Estimating the Statistical Power and Effect Size When Make Z Test

Hu Zhujing

(School of Psychology,Jiangxi Normal University,Nanchang 30022)

Abstract: This paper taked an example of the statistical analysis of significant difference between mean values, after hypothesis testing the experimental data, this paper briefly introducted the basic principle and methods of estimating the power of test and effect size.

Key words: statistical analysis of significant difference between mean values; hypothesis testing; the power of test; effect size