**21. PROC MEANS均值以及均值的T检验**

**（一）PROC MEANS过程步**

由PROC UNIVARIATE过程步生成的大多数统计描述，用PROC MEANS过程步也可以实现。区别是，UNIVARIATE是做更深入的统计分析；如果只是需要计算少数的统计量，PROC MEANS更适合（不能做图形输出）。

基本语法：

PROC MEANS data = 数据集 *statistic-keywords*;

CLASS *variable;*

VAR *variable-list*;

说明：

（1）CLASS指定分组变量，VAR指定要做统计分析的变量；

（2）默认置信水平是0.05（即95%的置信限），若要设定在统计量关键词位置加上，例如，ALPHA =0.1;

（3）若不加统计量关键词，默认输出：均值、非缺失值个数、标准差、最小值、最大值。可选的统计量关键词包括：

|  |  |  |  |
| --- | --- | --- | --- |
| 关键词 | 说明 | 关键词 | 说明 |
| CLM | 双边置信限 | RANGE | 极差 |
| CSS | 校正平方和 | SKEWNESS | 偏度 |
| CV | 变异系数 | STDDEV | 标准差 |
| KURTOSIS | 峰度 | STDERR | 均值的标准误 |
| LCLM | 置信下限 | SUM | 求和 |
| MAX | 最大值 | SUMWGT | 求权重和 |
| MIN | 最小值 | UCLM | 置信下限 |
| MEAN | 均值 | USS | 未修正的平方和 |
| MODE | 众数 | N | 非缺失值个数 |
| NMISS | 缺失值个数 | PROBT | T检验的概率 |
| T | T值 | Q1或P25 | 第1四分位数 |
| MEDIAN  或P50 | 中位数 | Q3或P75 | 第3四分位数 |
| P1 | 1%分位数 | P5 | 5%分位数 |
| P10 | 10%分位数 | P90 | 90%分位数 |
| P95 | 95%分位数 | P99 | 99%分位数 |

**例1** 儿童书作家考察市面上儿童书的页数作为出书的参考，搜集数据（C:\MyRawData\Picbooks.dat）如下：



读入数据，计算数据个数、均值、中位数，以及90%的置信限。

**代码**：

**data** booklengths;

infile 'c:\MyRawData\Picbooks.dat';

input NumberOfPages @@;

**run**;

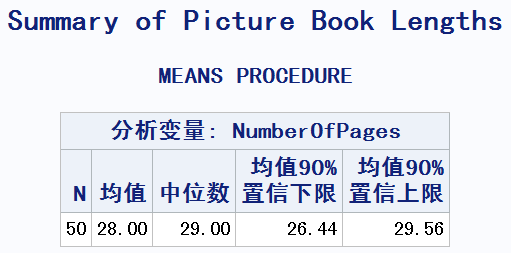
\*Produce summary statistics;

**proc** **means** data = booklengths N MEAN MEDIAN CLM ALPHA = **0.10** MAXDEC = **2**;

title 'Summary of Picture Book Lengths';

**run**;

运行结果：



说明：有90%的把握说“儿童书的页数范围是：[26.44, 29.56]”.

**（二）假设检验的P值法**

**一、 什么是假设检验？**

实际中，我们只能得到抽取的样本（部分）的统计结果，要进一步推断总体（全部）的特征，但是这种推断必然有可能犯错，犯错的概率为多少时应该接受这种推断呢？

为此，统计学家就开发了一些统计方法进行统计检定，通过把所得到的统计检定值，与统计学家树立了一些随机变量的概率分布进行对比，我们可以知道在百分之多少的机遇下会得到目前的结果。

倘若经比较后发现，涌现这结果的机率很少，即是说，是在时机很少、很罕有的情况下才出现；那我们便可以有信念地说，这不是巧合，该推断结果是具有统计学上的意义的。否则，就是推断结果不具有统计学意义。

**二、假设检验的基本思想——小概率反证法思想**

小概率思想是指小概率事件（P<α, α=0.05或0.01）在一次试验中基本上不会发生。反证法思想是先提出原假设（H0），再用适当的统计方法确定假设成立的可能性（P值）大小，如可能性小（P≤α），则认为原假设不成立，若可能性大，则还不能认为备择假设（H1）成立。

**三、原假设与备择假设**

原假设与备择假设是是完备且相互独立的事件组，一般，

原假设（H0）——研究者想收集证据予以反对的假设；

备择假设（H1）——研究者想收集证据予以支持的假设；

假设检验的P值是由检验统计量的样本观察值得出的原假设可被拒绝的最小显著水平。

假设检验判断方法有：临界值法、P值检验法。

**四、假设检验分类及步骤（以T检验为例）**

1. 双侧检验

I. 原假设H0: μ=μ0, 备择假设H1: μ≠μ0;

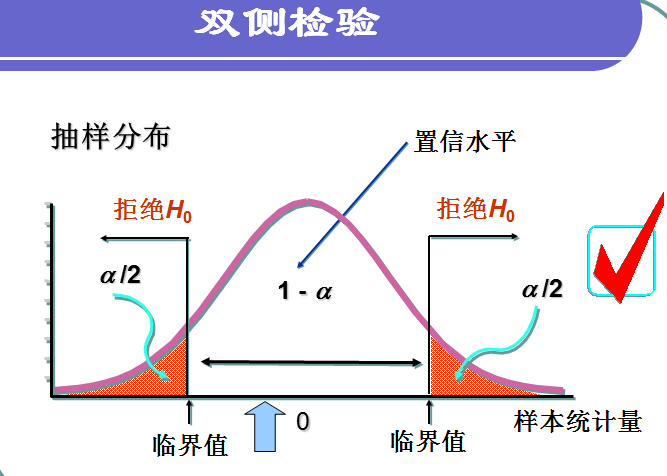
Ⅱ. 根据样本数据计算出统计量t的观察值t0;

Ⅲ. P值 = P{|t| ≥ |t0|} = t0的双侧尾部的面积；

Ⅳ. 若P值≤α（在右尾部分），则在显著水平α下拒绝H0;

若P值>α，则在显著水平α下接受H0;

注意：α为临界值，看P值在不在阴影部分（拒绝域），空白部分为接受域。



2. 左侧检验

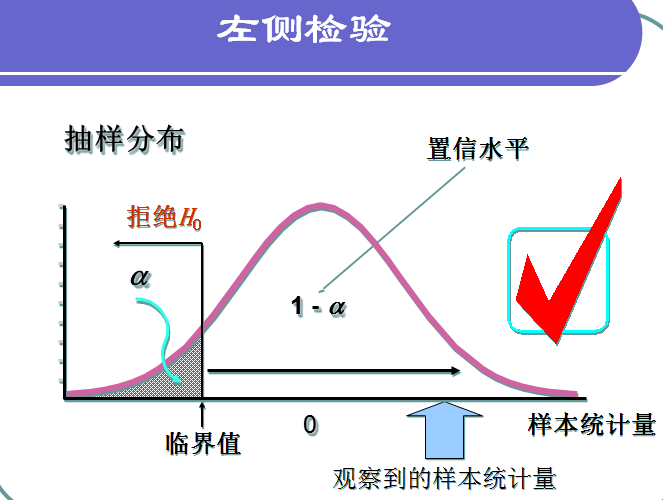
I. 原假设H0: μ≥μ0, 备择假设H1: μ<μ0;

Ⅱ. 根据样本数据计算出统计量t的观察值t0（< 0）;

Ⅲ. P值 = P{t ≤ t0} = t0的左侧尾部的面积；

Ⅳ. 若P值≤α（在左尾部分），则在显著水平α下拒绝H0;

若P值>α，则在显著水平α下接受H0;



3. 右侧检验

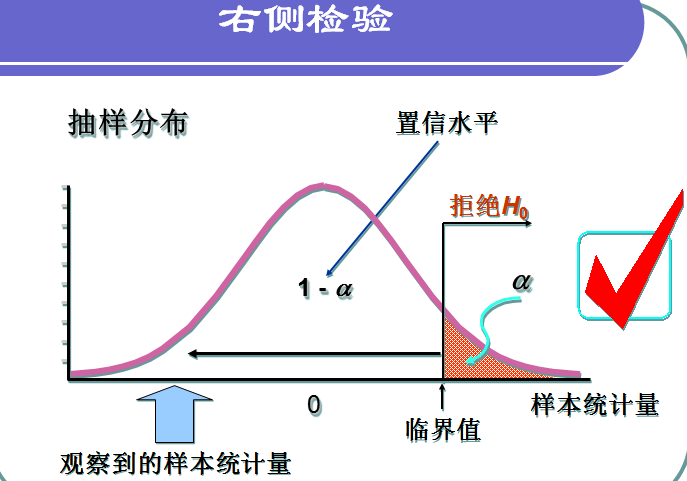
I. 原假设H0: μ≤μ0, 备择假设H1: μ>μ0;

Ⅱ. 根据样本数据计算出统计量t的观察值t0（> 0）;

Ⅲ. P值 = P{t ≥ t0} = t0的右侧尾部的面积；

Ⅳ. 若P值≤α（在右尾部分），则在显著水平α下拒绝H0;

若P值>α，则在显著水平α下接受H0;



**（三）T分布与T检验**

**一、T分布**

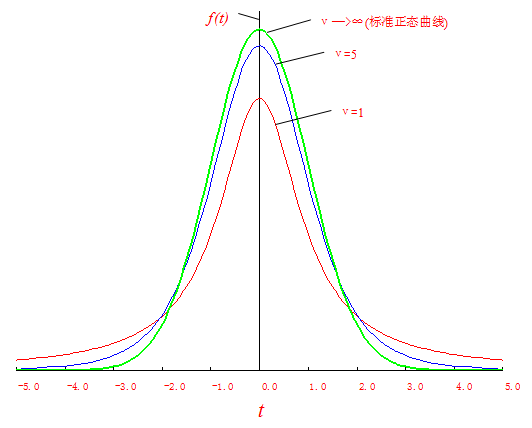
若样本均数服从正态分布， 经过U变换, 可以变成标准正态分布N(0, 12), 也成为U分布.

实际工作中，由于总体标准差未知，用样本标准差代替，则不再服从标准正态分布，而是服从T分布：

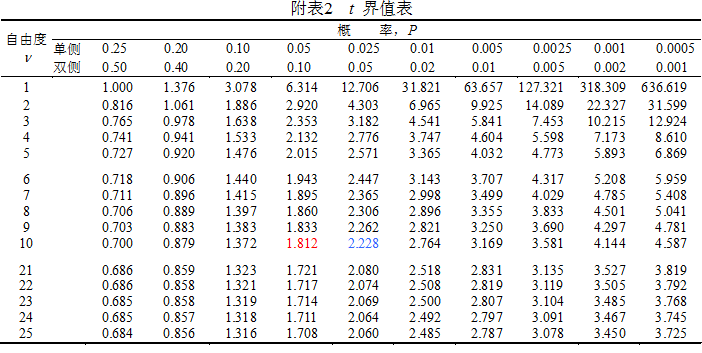


其中，S为样本方差，n为样本含量，v为自由度。

T分布只有一个参数——自由度v. v→∞时，T分布无限接近标准正态分布。



T分布的图形



说明：

单侧概率（单侧尾部面积）用表示；

双侧概率（双侧尾部面积）用表示；

例如，t0.05,10=1.812, 则P(t≤-1.812)=P(t≥1.812)=0.05

t0.05/2,10=2.228, 则P(t≤-2.228)+P(t≥2.228)=0.05

**二、T检验**

T检验，是一种参数假设检验，用来检验“**单样本均数与已知均数**、**两独立样本均数**、**配对设计资料的均数**”是否存在差异，这种差异是否能推论至总体。

T检验适用于样本含量较小（比如n<60，大样本数据可以用U检验），适用条件：

① 数据服从正态分布（做正态性检验）；

② 满足方差齐性（即两样本的总体方差相等，做F检验）；

若数据不满足①，②，可以尝试对数据做变量变换：对数变换、平方根变换、倒数变换、平方根反正弦变换等。

正态性检验：

（1）W检验，Shapiro-Wilk检验是基于次序统计量对它们期望值的回归而构成的。所用检验统计量为W，又称为W检验。在样本量3≤n≤50时使用。

（2）D检验，Kolmogorov-Smirnov检验的统计量为D，所以也称D检验，在样本量50≤n≤1000时使用。

假设检验H0：数据总体服从正态分布，H1：数据总体不服从正态分布。当P≤α时，拒绝H0，认为样本所来自的总体不服从正态分布；当P＞α时，不拒绝H0，认为样本所来自的总体服从正态分布。

F检验

要求两样本数据的总体均服从正态分布，统计量F为为较大的方差与较小的方差的比值：



假设检验H0：两总体方差相等；H1：两总体方差不相等。取α=0.10（α较大以减少II类错误），当P≤α时，拒绝H0，认为两总体方差不相等；当P＞α时，不拒绝H0，认为两总体方差相等。

**（四）均值的T检验**

T检验“**单样本均数与已知均数**、**两独立样本均数**、**配对设计资料的均数**”是否存在差异。

例如，检查学生成绩平均分是否在某个分值之上；比较同一老师教的两个班的学生平均分是否存在差异；正常饲料组和维E缺乏组大白鼠肝中维生素A含量的均值是否存在差异。

T检验在SAS中用PROC TTEST过程步实现。

**一、单样本与指定均值**

基本语法：

PROC TTEST data = 数据集 H0 = mu0 *options*;

VAR *variable*;

说明：“H0=mu0”，对变量的均值与指定均值mu0（默认是H0=0），做T检验。原假设是μ=μ0.

**二、两独立样本的均值**

基本语法：

PROC TTEST data = 数据集 *options*;

CLASS *variable*;

VAR *variable*;

说明：CLASS语句指定分组变量，进行组间均值的比较。

**三、配对设计资料的均值**

基本语法：

PROC TTEST *options*;

PAIRED *variable1 \* variable2*;

说明：配对均值T检验，是检验两个变量各数据的差的均值是否等于0.

**四、Options可选项**

（1）ALPHA = n

指定显著水平α;

（2）CI = type

指定标准差的置信区间的类型，EQUAL（默认）为

equal-tailed的置信区间，UMPU为基于一致最优无偏检验的

置信区间，NONE不输出置信区间；

（3）SIDES = type

指定单侧、双侧检验，默认type=2为双侧检验，L为左侧

检验，U右侧检验；

**五、绘制T检验的图形**

在TTEST语句中使用参数PLOTS = (绘图类型)即可。

基本语法：

PROC TTEST data = 数据集 PLOTS = (*绘图类型*);

可选的绘图类型：

ALL或NONE——绘制全部图形或不绘制任何图形；

BOXPLOT——盒形图；

HISTOGRAM——直方图（包括正态分布、核密度线）；

INTERVALPLOT——均值的置信区间的图形；

QQPLOT——QQ图；

SUMMARYPLOT——在一张图中绘制直方图和盒形图；

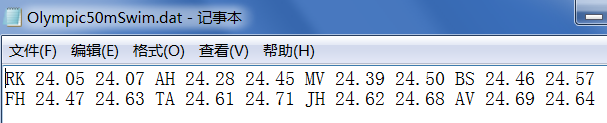
AGREEMENTPLOT——AGREEMENT图；

PROFILESPLOT——PROFILESPLOT图；

注意：TTEST过程步都默认绘制QQ图和SUMMARYPLOT图，配对T检验还默认绘制AGREEMENTPLOT图和PROFILESPLOT图。另外，指定绘制图形类型之后，那些默认图仍然会绘制，除非加上(ONLY):

PROC TTEST data = 数据集 PLOTS(ONLY) = (*绘图类型*);

**例2** 50米女子自由泳的数据（C:\MyRawData\Olympic50mSwim.dat），变量包括姓名、决赛用时、半决赛用时：



读入数据，用配对T检验考察决赛和半决赛用时有无明显差异。

**代码：**

**data** Swim;

infile 'c:\MyRawData\Olympic50mSwim.dat';

input Swimmer $ FinalTime SemiFinalTime @@;

**run**;

**proc** **ttest** data = Swim PLOTS(ONLY) = (SUMMARYPLOT QQPLOT);

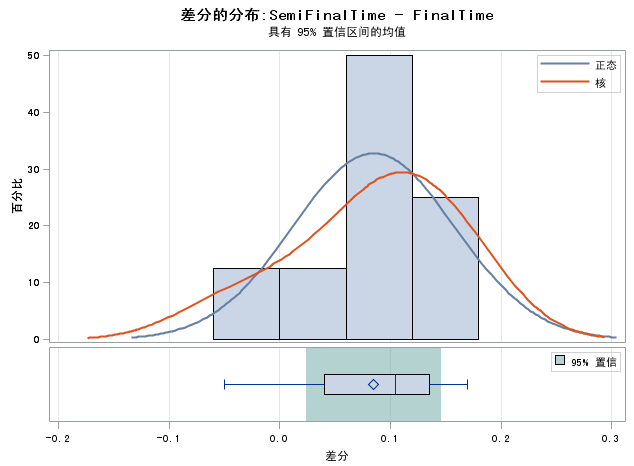
paired SemiFinalTime \* FinalTime;

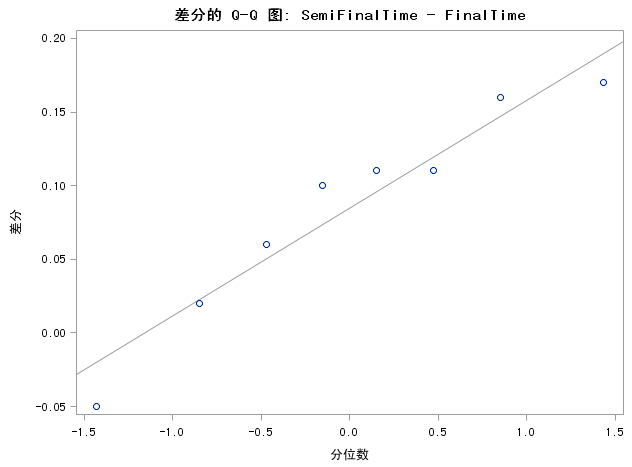
title '50m Freestyle Semifinal vs. Final Results';

**run**;

运行结果：







程序说明：

（1）总决赛用时与半决赛用时之差的均值为0.0850, 其95%置信区间为[0.239, 0.1461]; 标准差为0.0731, 其95%置信区间为[0.0483, 0.1488];

（2）自由度为7的t值=3.29，P值=0.0133 小于显著性水平α=0.05（P值在双尾部分“拒绝域”），故拒绝原假设H0. 结论：两个均值不相同，其结果有统计学意义。