**一、引言**

**1、背景与目标**

为保证数据报表准确性、及时发现并修正报表异常数据，现对报表数据进行监控、并根据异常发生可能性大小进行分级报警。

**2、参考资料**

（1）3σ原则

<https://baike.baidu.com/item/3%CF%83%E5%87%86%E5%88%99/9361985>

（2）正态分布检验

正态分布检验：<https://www.sohu.com/a/140979052_489312>

Q-Q图检验：<https://www.jianshu.com/p/09748f57160d?utm_campaign>

**二、基本步骤**

1、数据准备

① 工作日vs周末：两者数据差异较大，所以将数据标记出星期，后续可按每周X观测数据。例：计算日期为周一，则获取周一历史数据进行计算；

② 法定假日vs正常排班：法定假日由于调休的影响不可按照正常工作日/周末处理，对其进行特殊处理：调休的日期（非周六日时）记为周六，调班日期记为周五。

2、数据报警

① 输出可能异常的数据；② 输出对应的错误发生概率；③ 根据不同重要级别发送不同的报警方式。

3σ：数据在（μ - 3σ,μ + 3σ)范围外---高优先级：错误可能性为99.73%以上，电话、邮件报警；

2σ：数据在（μ - 2σ,μ + 2σ)范围外---中优先级：错误可能性为95.45%以上，电话、邮件报警；

1σ：数据在（μ - σ,μ + σ)范围外---低优先级：错误可能性为68.27%以上，邮件报警。

3、备选方案

算法预测数据未计算出来时，延用上周区间结果。如今天周一，延用上周一区间。

**三、监控设计方案**

**1、整体规划**

① 获取数据；

② 计算指标概率分布情况（或pdf图），其中小概率数据认为是异常数据；

③ 设定小概率数据标准，如5%；

④ 计算得出小概率数据范围。

若为正态分布，②③④步可直接替换为3σ原则，即：

数值分布在（μ-σ,μ+σ)中的概率为68.27%；

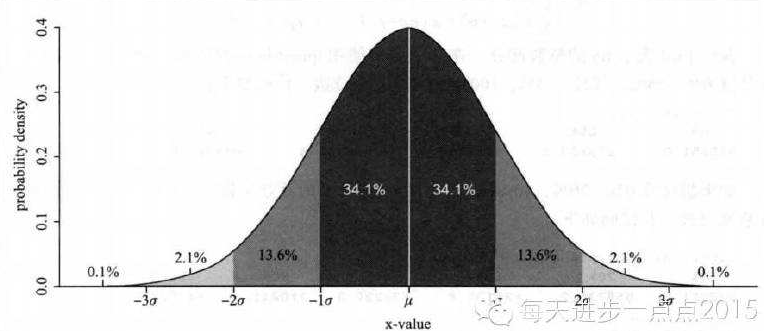
数值分布在（μ-2σ,μ+2σ)中的概率为95.45%；

数值分布在（μ-3σ,μ+3σ)中的概率为99.73%；

判断报表中当天指标数值是否在上述规定的区间内，如果不在则落入小概率数据范围，数据异常。

μ：标准值，使用当天算法预测值；

σ：过去一段时间内报表中指标数值标准差。



本期待监测的3个指标根据前期数据调研可知（请见附：正态分布检验）：

非标准正态分布，但与正态分布近似，本期暂按照正态分布进行处理；

有些指标呈现双峰正态分布或截断正态分布形式，后续再优化。

所以根据以上说明，认为本期待监测指标为正态分布，可使用3σ原则。

有一点问题还需确定：计算标准差使用历史近多少天数据合理？如果天数选取太长，数据量级差异过大；天数选取太短，数据波动过大。

**2、历史数据天数确定**

确定天数的标准：设t天指标数值为Xt，t-1天为Xt-1，... t-n天为Xt-n，Xt ~ Xt-n数据同分布。

同分布即意味着：

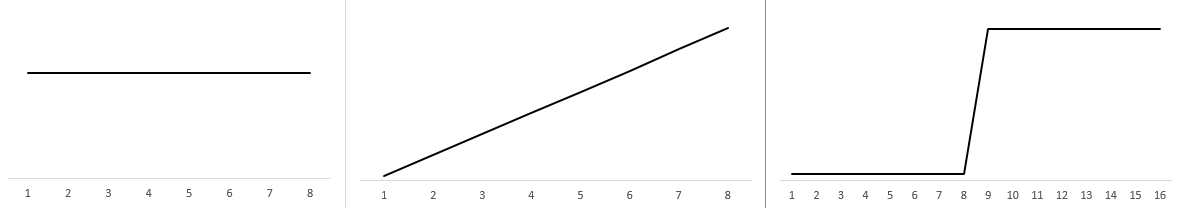
均值相等：E[Xt] = E[Xt-n] ；

方差（标准差）相等：σ[Xt] = σ[Xt-n] ；

当前历史数据可通过一些变换使其变为均值、方差近似的数据，再判断近多少天数据是同分布。

（1）均值相等：

数据简化后可能呈现出以下3种情况：



将Xt ~ Xt-n的数据每一项与前一项做差：Yt = Xt - Xt-1，在3种情况下分别为：

Yt ~ （0，σ），此时可取历史所有数据，均值均为0；

Yt ~ （μ，σ），此时均值不为0但相等；

Yt ~ （μ，σ），数据计算到阶跃点时，均值会有较大波动。

说明：数据不会如图那么平滑，所以减完后仍会有方差。

（2）方差相等：

暂不考虑， 后期再优化。

均值处理中第3种情况做差后均值不等，历史数据不全部属于同一分布，处理方式如下：

**方案1：**假定每两个阶跃点间的数据属于同一分布，获取距离今天最近的一个阶跃点与今天之间的数据计算方差。

获取距离今天最近的点至今的**天数计算方式**为：

① 统计数据范围：18年11月以来每周X数据（太早的数据体量差异过大），不含当天。

例：当天是周一，则选取18年11月至今每周一数据；

② 计算每两项数据间的差值（取绝对值）：Kt = | Xt - Xt-1 |；

③ 获取差值top5天数：选取差值top5的距离计算日期天数，判断差值top1天数是否 > 4天。

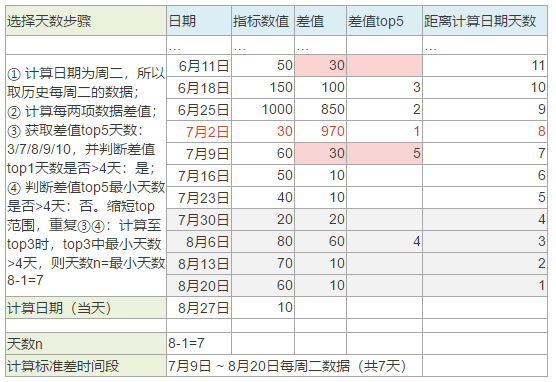
* 其中4天是指4个周X（如4个周二），避免天数太少数据波动大；
* 如有差值相等的情况，优先取距离计算日期天数小的（见如下标粉色单元格）。

④ 获取天数n（第一部分）：差值top1天数 > 4天时，判断差值top5最小天数是否 ＞ 4天，若 > 4天，则天数n = 最小天数 - 1，若 ≤ 4天，则缩小差值top范围，缩小步长 = 1，top4、... top2，重复③④，直到满足条件时截止。

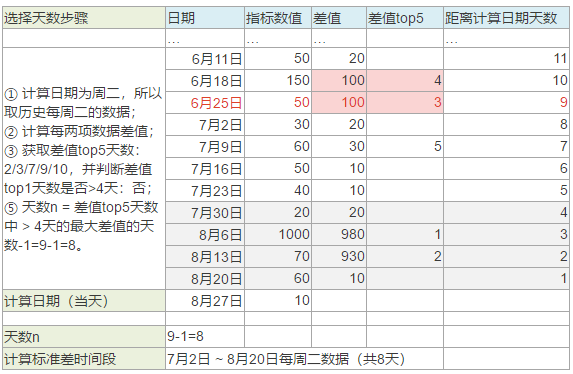
⑤ 获取天数n（第二部分）：差值top1天数 ≤ 4天时，则天数n = （差值top5天数中 > 4天的最大差值的天数） - 1。

举例如下表：

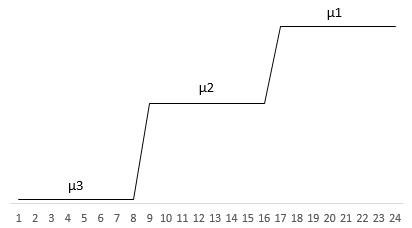
例1：差值top1天数 > 4天



例2：差值top1天数 ≤ 4天



**方案2**：按阶跃点将数据分段，假定每两个阶跃点之间的数据属于同一分布，以最近一段数据均值μ1为标准，将历史每一段数据均调整为均值为μ1的数列，使其成为同分布，此时可使用历史所有数据计算标准差。



**调整为同一均值的计算方式**：

① 统计数据范围：每周X数据，不含当天。

例：当天是周一，则选取历史每周一数据；

② 计算每两项数据间的差值（取绝对值）：Kt = | Xt - Xt-1 |；

③ 获取分段数据：选取差值top5日期，就近每两个点间的数据为一组，可分为6组；

④ 计算每段数据均值：设离今天最近的一段开始，第 i 段均值为μi，i=1、2、...、6；

⑤ 数据调整为同一分布：以离今天最近的一段数据均值μ1为标准，对其他每段数据的值进行处理：设 Xj 为第 i 段的第 j 个数，则对Xj处理后的Yj = Xj \* μ1 / μi ，对每个数据进行处理后即得到同一分布的新数列；

例：μ1=100，μ2=80，第二段共有3个数据80/90/70，以μ1为标准对3个数据进行处理：Y1=80\*100/80、Y2=90\*100/80、Y3=70\*100/80；

**四、可优化项**

方案1

① 最小天数：目前是取4天；

② 差值最大的top数：目前是取top5。

③ 确定历史天数的方式可优化：目前认为取差值最大的几个点中距离今天最近的点，是否还有更合理的方式。

方案2

① 分段数量：目前分为6段，后面可根据情况优化。

② 数据同分布处理：目前仅对数据做了差值处理，使得均值尽量一致，但方差未做处理。

其他

① 数据分布：目前是按标准正态分布处理，但数据实际上不标准，如双峰正态分布、截断正态分布；