统计学是数据分析的基石。学了统计学，你会发现很多时候的分析并不靠谱。比如很多人都喜欢用平均数去分析一个事物的结果，但是这往往是粗糙的，不准确的。如果学了统计学，那么我们就能以更多更科学的角度看待数据。

大部分的数据分析，都会用到统计方面的以下知识，可以重点学习：

基本的统计量：均值、中位数、众数、方差、标准差、百分位数等

概率分布：几何分布、二项分布、泊松分布、正态分布等

总体和样本：了解基本概念，抽样的概念

置信区间与假设检验：如何进行验证分析

相关性与回归分析：一般数据分析的基本模型

通过基本的统计量，你可以进行更多元化的可视化，以实现更加精细化的数据分析。这个时候也需要你去了解更多的 Excel 函数来实现基本的计算，或者 python、R 里面一些对应的可视化方法。

有了总体和样本的概念，你就知道在面对大规模数据的时候，怎样去进行抽样分析。

你也可以应用假设检验的方法，对一些感性的假设做出更加精确地检验。

利用回归分析的方法，你可以对未来的一些数据、缺失的数据做基本的预测。

了解统计学的原理之后，你不一定能够通过工具实现，那么你需要去对应的找网上找相关的实现方法，也可以看书。先推荐一本非常简单的：吴喜之-《统计学·从数据到结论》。也可以看《商务与经济统计》，结合业务能更容易理解。

另外，如何精力允许，请掌握一些主流算法的原理，比如线性回归、逻辑回归、决策树、神经网络、关联分析、聚类、协同过滤、随机森林。再深入一点，还可以掌握文本分析、深度学习、图像识别等相关的算法。关于这些算法，不仅需要了解其原理，你最好可以流畅地阐述出来，还需要你知晓其在各行业的一些应用场景。如果现阶段不是工作刚需，可不作为重点。

本文算是一个知识点汇总，不做细致展开，让大家了解统计学有哪几大块，每一类分别用于什么样的分析场景。后面几篇会以实际案例的方式，细致讲讲描述性统计、概率分布等。

知识点汇总：

1.集中趋势

2.变异性

3.归一化

4.正态分布

5.抽样分布

6.估计

7.假设检验

8.T 检验

一、集中趋势

1.众数

出现频率最高的数；

2.中位数

把样本值排序，分布在最中间的值；

样本总数为奇数时，中位数为第(n+1)/2 个值；

样本总数为偶数时，中位数是第 n/2 个，第(n/2)+1 个值的平均数；

3.平均数

所有数的总和除以样本数量；

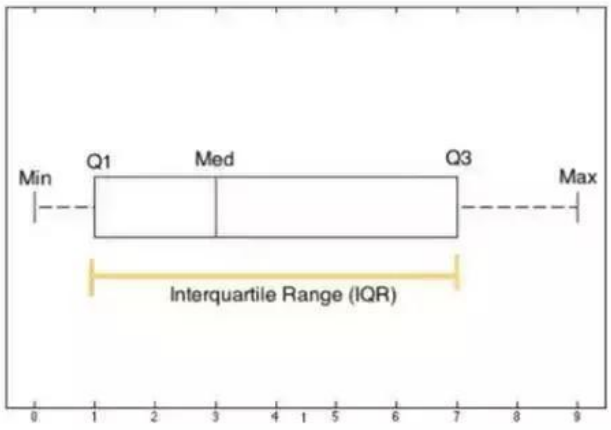
现在大家接触最多的概念应该是平均数，但有时候，平均数会因为某些极值的出现收到很大影响。举个小例子，你们班有 20 人，大家收入差不多，19 人都是 5000 左右，但是有 1 个同学创业成功了，年入 1 个亿，这时候统计你们班同学收入的“平均数”就是 500 万了，这也很好的解释了，每年各地的平均收入数据出炉，小伙伴们直呼给祖国拖后腿了，那是因为大家收入被平均了，此时，“中位数”更能合理的反映真实的情况；

二、变异性

1.四分位数

上面说到了“中位数”，把样本分成了 2 部分，再找个这 2 部分各自的“中位数”，也就把样本分为了 4 个部分，其中 1/4 处的值记为 Q1，2/4 处的值记为 Q2，3/4 处的值记为 Q3

2.四分位距 IQR=Q3-Q1



3.异常值

小于 Q1-1.5(IQR)或者大于 Q3+1.5(IQR);

对于异常值，我们在数据处理的环节就要剔除；

4.方差

$$\sigma^2 = \frac{\sum_{i=0}^{n}(x_i - \bar{x})^2}{n}$$

5.平方偏差

方差的算术平方根

6.贝塞尔矫正：修正样本方差

实际在计算方差时，分母要用 n-1，而不是样本数量 n。原因在于，比如在高斯分布中，我们抽取一部分的样本，用样本的方差表示满足高斯分布的大样本数据集的方差。由于样本主要是落在 x=u 中心值附近，那么样本如果用如下公式算方差，那么预测方差一定小于大数据集的方差（因为高斯分布的边沿抽取的数据也很少）。为了能弥补这方面的缺陷，那么我们把公式的 n 改为 n-1，以此来提高方差的数值，这种方法叫贝塞尔矫正系数。

三、归一化

1.标准分数

一个给定分数 距离 平均数 多少个标准差？

标准分数是一种可以看出某分数在分布中相对位置的方法。

标准分数能够真实的反映一个分数距离平均数的相对标准距离。

$$Z=\frac{x-\mu}{\sigma}$$

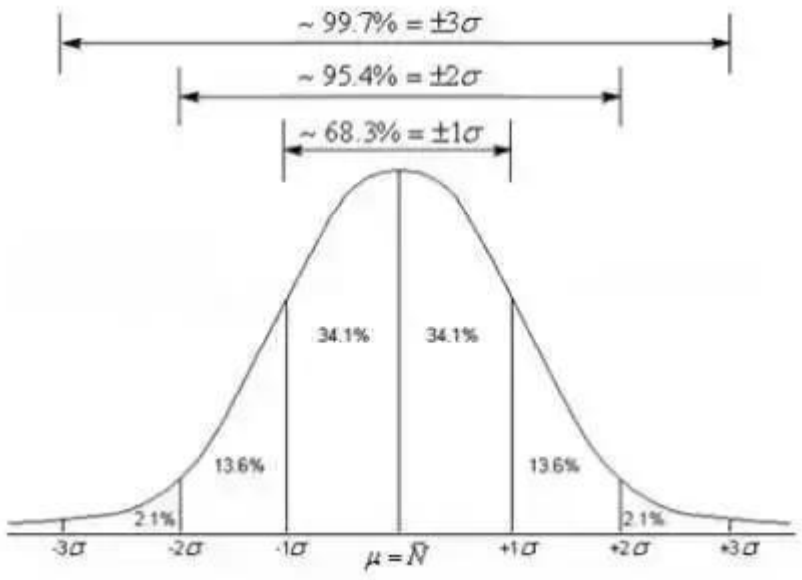
四、正态分布

1.定义：随机变量 X 服从一个数学期望为μ，方差为σ⊃2;的正态分布，记为 N(μ,σ⊃2;)

随机取一个样本，有 68.3%的概率位于距离均值μ有 1 个标准差σ内；

有 95.4%的概率位于距离均值μ有 2 个标准差σ内；

有 99.7%的概率位于距离均值μ有 3 个标准差σ内；



五、抽样分布

1.中心极限定理

设从均值为μ，方差为σ⊃2;的任意一个总体中抽取样本量为 n 的样本，当 n 充分大时，样本均值的抽样分布近似服从均值为μ、方差为σ⊃2;/n 的正态分布

2.抽样分布

设总体共有 N 个元素，从中随机抽取一个容量为 n 的样本，在重置抽样时，共有 N·n 种抽法，即可以组成 N·n 不同的样本，在不重复抽样时，共有 N·n 个可能的样本。每一个样本都可以计算出一个均值，这些所有可能的抽样均值形成的分布就是样本均值的分布。但现实中不可能将所有的样本都抽取出来，因此，样本

均值的概率分布实际上是一种理论分布。数理统计学的相关定理已经证明：在重置抽样时，样本均值的方差为总体方差的 1/n。

举个例子：

48 盆 MM 豆，计算出每盆有几个蓝色的 MM 豆，48 个数据构成了总体样本。然后随机选择五盆，计算五盆中含有蓝色 MM 豆的平均数，然后反复进行了 50 次。这就是 n 为 5 的样本均值抽样。

$$\sigma_M = \frac{\sigma}{\sqrt{N}}$$

六、估计

1. 误差界限

$$Z^* \cdot \frac{\sigma}{\sqrt{n}}$$

2. 置信度

We are some % sure the true population parameter falls within a specific range

我们有百分之多少确信总体中的值落在一个特定范围内；

一般情况下，取 95%的置信度就可以；

3. 置信区间

$$\bar{x} \pm Z^* \cdot \frac{\sigma}{\sqrt{n}}$$

七、假设检验

"大多数鸡有两只脚吗？" 这个问题的难点在于，我们很难说清楚"大多数鸡有两只脚" 为什么是对的。

- 显著水平

首先，何谓"大多数"呢？每个人的看法可能都不一样。

因此，需要挑选一个显著性水平（Alpha level），这里我们假定

$$\alpha = 0.5$$

问题转化为，"超过50%的鸡有两只脚吗？"

1.问题：什么是显著性水平？

显著性水平是估计总体参数落在某一区间内，可能犯错误的概率，也就是 Type I Error

A Type II Error is when you fail to reject the null when it is actually false.

- 零假设和对立假设

由于我们很难证明某种说法是对的（大多数鸡有2只脚）。

因而我们设法寻找该说法的对立面（大多数鸡少于2只脚）错误的证据。

如果我们能够设法证明该说法的对立面是错的，那么就相当于证明了该说法本身是对的。

所以，建立两个互相对立的假设。

零假设： 超过50%的鸡少于两只脚

对立假设： 超过50%的鸡有两只脚

4个小时后，得到的样本如下

64.3% 的样本（鸡）有2只脚，35.7%的样本（鸡）少于2只脚。

- 统计学结论

拒绝零假设（大多数鸡少于两只脚）

相当于接受对立假设(大多数鸡有2只脚）

最后，祝大家鸡年大吉，积（鸡）极（头）向上。😄

2. 如何选择备选检验和零假设？

一个研究者想证明自己的研究结论是正确的，备择假设的方向就要与想要证明其正确性的方向一致；

同时将研究者想收集证据证明其不正确的假设作为原假设 H0

八、T 检验

1. 主要用于样本含量较小（例如 n<30），总体标准差σ未知的正态分布。

流程如下：

是用 t 分布理论来推论差异发生的概率，从而比较两个平均数的差异是否显著；

一般检验水准α取 0.05 即可；

计算检验统计量的方法根据样本形式不同;

2. 独立样本 T 检验：

现在要分析男生和女生的身高是否相同两者的主要区别在于数据的来源和要分析的问题。

$$t = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

问题：为什么 T 检验查表时候要 n-1?

样本均值替代总体均值损失了一个自由度

3. 配对样本 t 检验

分析人的早晨和晚上的身高是否不同，于是找来一拨人测他们早上和晚上的身高，这里每个人就有两个值，这里出现了配对

$$\frac{(x_2 - x_1) - (\mu_2 - \mu_1)}{\frac{\sqrt{(s_1^2 + s_2^2}}{n}}$$

样本误差（Standard Error）

$$SE = \frac{\sigma}{\sqrt{n}}$$

$$SE = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

4. Pooled variance 合并方差

当样本平均数不一样，但实际上认为他们的方差是一样的时候，需要合并方差

不要被公式吓到，他的本质是两个样本方差加权平均

Independent Samples *t* Test by Hand (Definitional Formula)			
Variance (computed for each group)	Pooled Variance	Standard Error of the Difference	Definitional *t* Test
$s_x^2 = \frac{\Sigma X^2 - \frac{(\Sigma X)^2}{n}}{n-1}$	$S_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$	$SE_{M_1-M_2} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$	$t = \frac{M_1 - M_2}{SE_{M_1-M_2}}$

$$s = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$$

5. Cohen's d

效应量(effect size):提示组间真正的差异占统计学差异的比例，值越大，组间差异越可靠。

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s}$$

$$\frac{\text{mean difference}}{\text{standard deviation}} \text{ or } \frac{\text{M2 - M1}}{\text{pooled standard deviation}}$$