【机器学习】一文详尽系列之模型评估指标

深度学习自然语言处理 2019-11-23

以下文章来源于Datawhale ，作者阿泽

Datawhale

一个专注于AI领域的开源组织，汇集了众多领域院校和知名企业的优秀学习者，聚合了...

> 阅读大概需要12分钟🤔
> 跟随小博主，每天进步一丢丢 😄

来自：Datawhale
作者：阿泽

在机器学习领域通常会根据实际的业务场景拟定相应的不同的业务指标，针对不同机器学习问题如回归、分类、排序，其评估指标也会不同。

准确率、精确率、召回率、F1值

定义

- 准确率（Accuracy）：正确分类的样本个数占总样本个数，$A = \frac{TP+TN}{N}$
- 精确率（Precision）：预测正确的正例数据占预测为正例数据的比例，$P = \frac{TP}{TP+FP}$
- 召回率（Recall）：预测为正确的正例数据占实际为正例数据的比例，$R = \frac{TP}{TP+FN}$
- F1 值（F1 score）：$F_1 = \frac{2}{\frac{1}{P}+\frac{1}{R}} = \frac{2\times P\times R}{P+R}$

计算

背景：假如有 100 个广告，某用户对 80 个不感兴趣，对其中 20 个感兴趣，目标是找出所有用户感兴趣的广告，现在挑出 40 个，其中 10 个感兴趣的，请问如何评估一下他的工作。

	实际正类	实际负类
预测正类	TP=10	FP=30

	实际正类	实际负类
预测负类	FN=10	TN=50

通过混淆矩阵，我们可以算出来 $A=0.6$ ， $P=0.25$ ， $R=0.5$ ， $F_1=0.33$

优缺点

准确率、精确率、召回率、F1 值主要用于分类场景。

准确率可以理解为预测正确的概率，其缺陷在于：当正负样本比例非常不均衡时，占比大的类别会影响准确率。如异常点检测时：99.9% 的都是非异常点，那我们把所有样本都视为非异常点准确率就会非常高了。

精确率可以理解为预测出的东西有多少是用户感兴趣的，召回率可以理解为用户感兴趣的东西有多少被预测出来了。一般来说精确率和召回率是一对矛盾的度量。为了更好的表征学习器在精确率和召回率的性能度量，我们引入 F1 值。

在个别领域可能我们对精确率和召回率的偏重不同，故我们引入 F_β，来表达出对精确率和召回率的不同偏好。

$$F_\beta=\frac{(1+\beta^2)*P*R}{(\beta^2*P)+R}$$

$\beta>1$ 时精确率影响力更大，$\beta<1$ 是召回率影响更大。

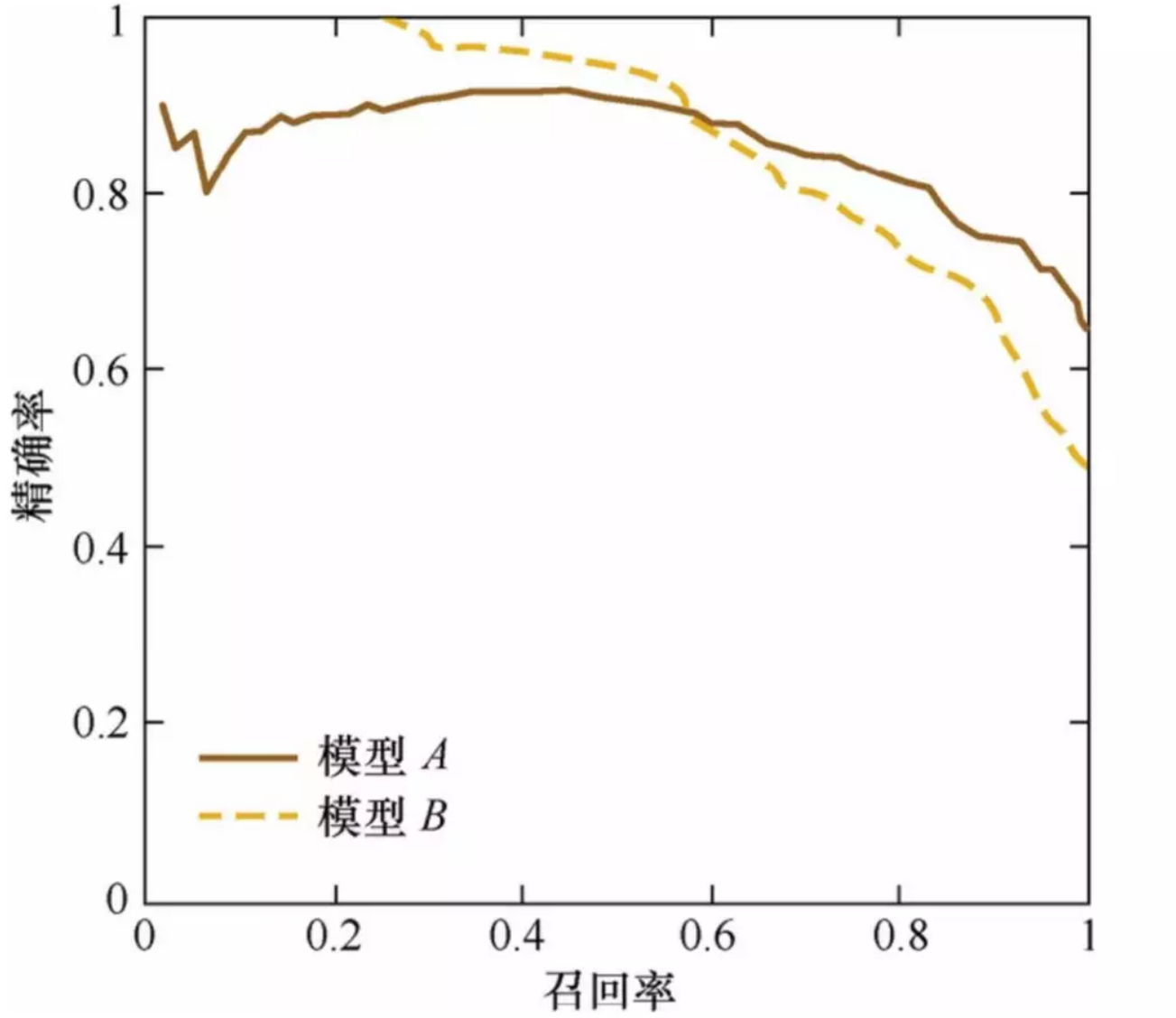
P-R、ROC、AUC

定义

- P-R 曲线：横轴召回率，纵轴精确率。
- ROC（receiver operating characteristic curve接收者操作特征曲线）：采用不分类阈值时的TPR（真正例率）与FPR（假正例率）围成的曲线，以FPR为横坐标，TPR为纵坐标。如果ROC 是光滑的，那么基本可以判断没有太大的overfitting。
- AUC（area under curve）：计算从（0, 0）到（1, 1）之间整个ROC曲线一下的整个二维面积，用于衡量二分类问题其机器学习算法性能的泛化能力。其另一种解读方式可以是模型将某个随机正类别样本排列在某个随机负类别样本之上的概率。

计算

P-R



P-R 曲线上的点代表不同阈值下模型将大于阈值的结果视为正样本，小于阈值的为负样本。

我们可以看到不同召回率下模型 A 和模型 B 的精确率表现不同，所以如果只对某点来衡量模型的性能是非常片面的，而只有通过 P-R 曲线的整体表现才能够进行更为全面的评估。

ROC、AUC

除了 F1 和 P-R 曲线外，ROC 和 AUC 也可以综合反应一个模型的性能。二分类真实值：$y_{true} = [0, 0, 1, 1]$ 分为正样本的概率：$scroe = [0.1, 0.4, 0.35, 0.8]$

针对 $score$ 对数据进行排序，将阈值一次取为 $score$ 值，故阈值依次取值为 0.1，0.35，0.4，0.8

然后我们依次计算不同阈值下的 TPR 和 FPR。我们以 $score = 0.35$ 为例

	实际正类	实际负类
预测正类	TP=2	FP=0
预测负类	FN=1	TN=1

真阳性率：

$$TPR = \frac{TP}{TP + TN} = 1$$

假阳性率：

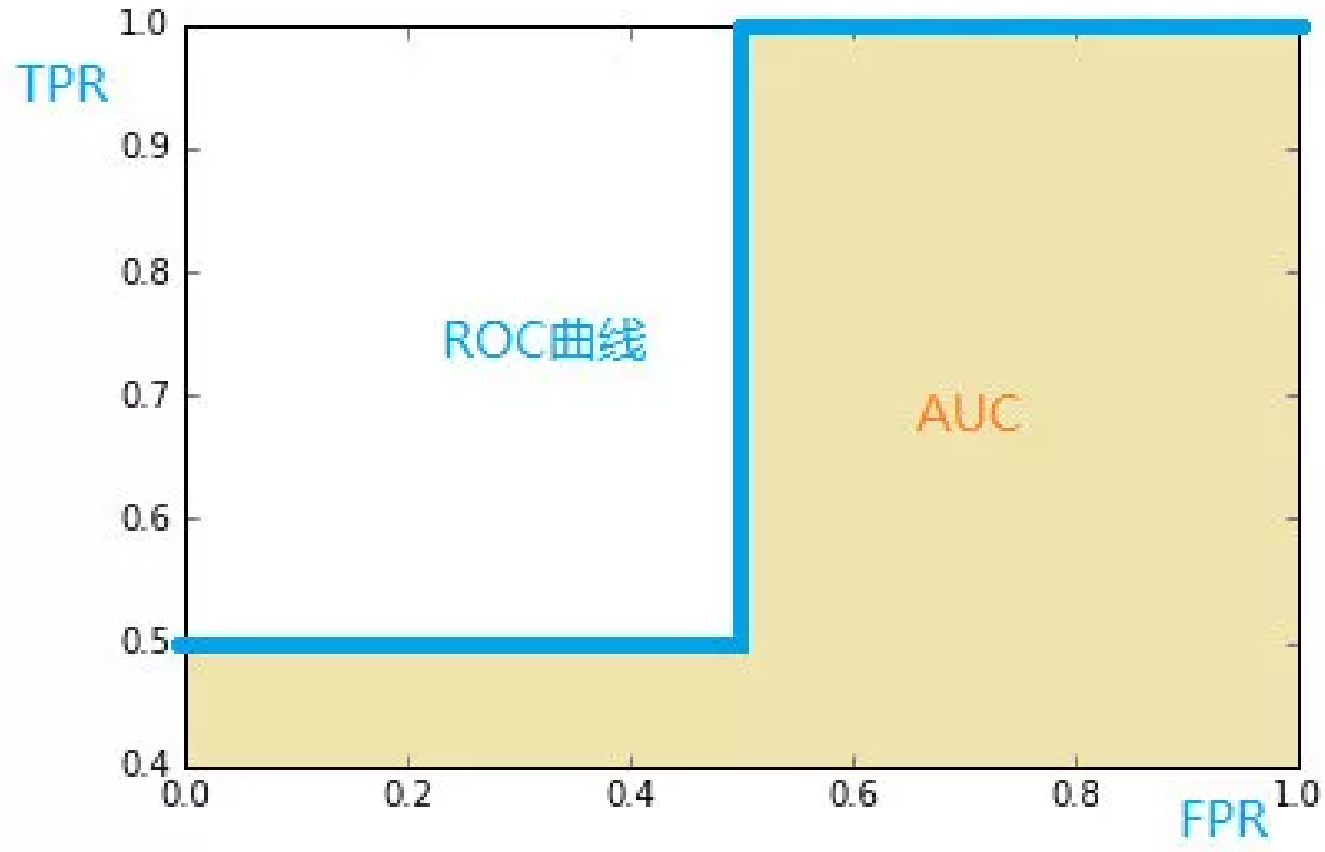
$$FPR = \frac{FP}{FP + TN} = 0.5$$

即可得到一个点的坐标。

计算完四种阈值后得到：

阈值	**0.8**	**0.4**	**0.35**	**0.1**
FPR	0	0.5	0.5	1
TPF	0.5	0.5	1	1

画图如下：



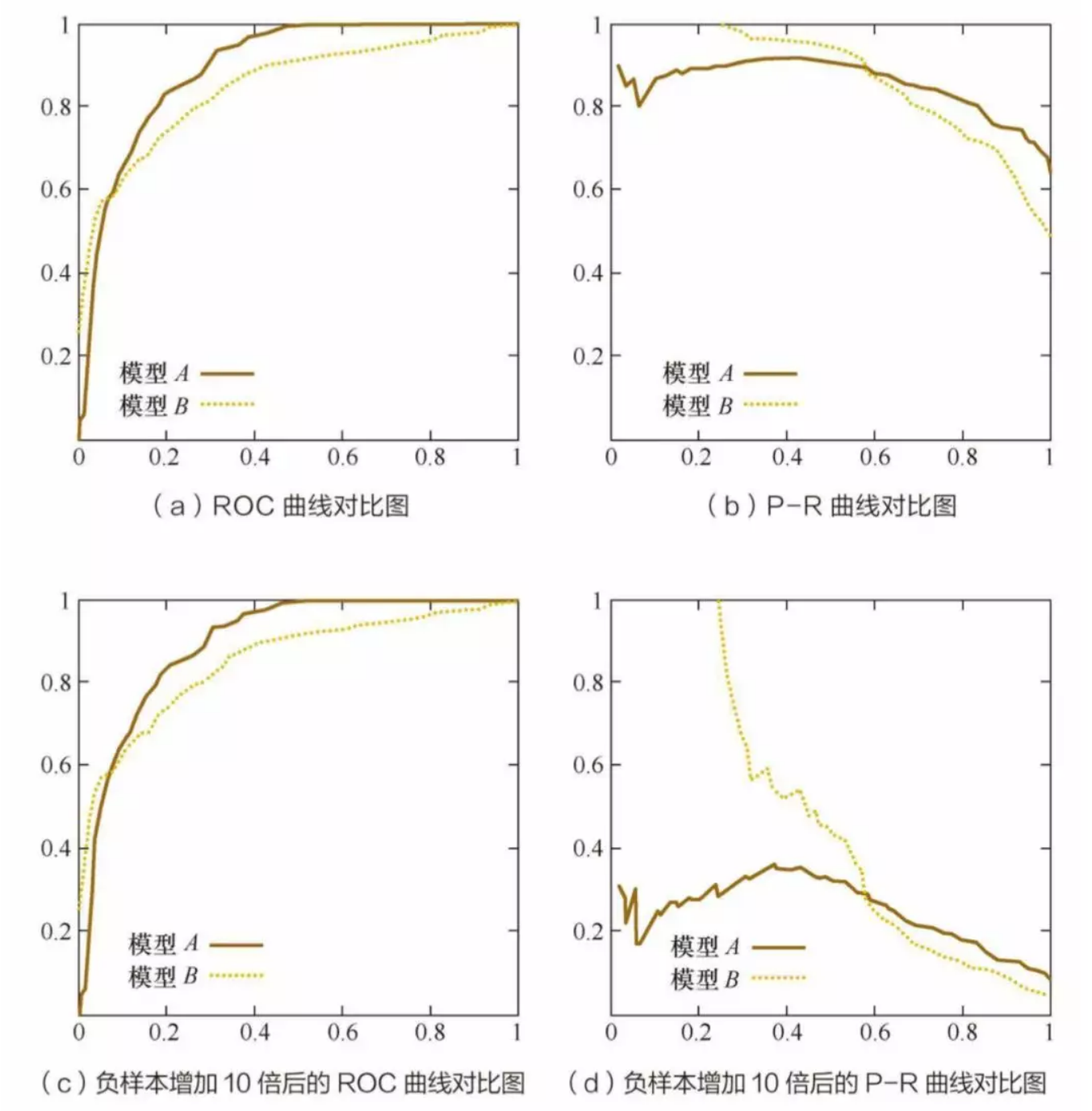

```
fpr: [ 0.   0.5  0.5  1. ]
tpr: [ 0.5  0.5  1.   1. ]
thresholds: [ 0.8   0.4   0.35  0.1 ]
```

我们可以看到 ROC 曲线是通过移动分类器的阈值来生成曲线上的关键点。ROC 曲线一般都处于 $y = x$ 直线的上方，所以AUC的取值一般是 0.5 ~ 1，AUC 越大，说明分类性能更好。

优缺点

P-R、ROC、AUC 主要用于分类场景。

相比 P-R 曲线来说，ROC 曲线有一个很大的特点：ROC 曲线的形状不会随着正负样本分布的变化而产生很大的变化，而 P-R 曲线会发生很大的变化。



如上图测试集负样本数量增加 10 倍以后 P-R 曲线发生了明显的变化，而 ROC 曲线形状基本不变。在实际环境中，正负样本的数量往往是不平衡的，所以这也解释了为什么 ROC 曲线使用更为

广泛。

MSE、RMSE、MAE、R2

定义

- MSE(Mean Squared Error) 均方误差，$P = \frac{1}{m}\sum(y_i - y)^2$
- RMSE(Root Mean Squared Error) 均方根误差，$P = \sqrt{\frac{1}{m}\sum(y_i - y)^2}$
- MAE(Mean Absolute Error) 平均绝对误差，$P = \frac{1}{m}\sum|y_i - y|^2$
- R^2，决定系数，$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum(y_i - \overline{y})^2}{\sum(y_i - \hat{y})^2} = 1 - \frac{MSE(y, \hat{y})}{Var(y)}$

优缺点

MSE、$RMSE$、MAE、R^2 主要用于回归模型。

MSE 和 RMSE 可以很好的反应回归模型预测值和真实值的偏离成都，但如果存在个别离群点的偏离程度非常大时，即使其数量非常少也会使得RMSE指标变差（因为用了平方）。解决这种问题主要有三个方案：

1. 如果认为是异常点时，在数据预处理的时候就把它过滤掉；
2. 如果不是异常点的话，就提高模型的预测能力，将离群点产生的原因建模进去；
3. 此外也可以找鲁棒性更好的评价指标，如：MAE ，。

余弦距离的应用

样本间的距离有不同的定义方式，常见的有欧式距离、曼哈顿距离、汉明距离、余弦距离等等。这里我们主要介绍下余弦距离及其应用。

定义

余弦相似度的定义如下：

$$||A - B||_2 = \frac{A \cdot B}{||A||_2||B||_2}$$

取值范围为：$[-1, 1]$。

如果我们想得到类似距离的表示，只需要将 1 减去余弦相似度即可：

$$dist(A, B) = 1 - ||A - B||_2 = 1 - \frac{A \cdot B}{||A||_2||B||_2}$$

其取值范围为：$[0,2]$。

我们要注意，虽然我们称其为余弦距离，但其并不是严格定义的距离。我们知道距离的严格定义需要满足：非负性，对称性，三角不等式。

- 非负性：

$$dist(A,B) \geq 0$$

特别的：

$$dist(A,B) = 0 \Leftrightarrow ||A||_2||B||_2 = AB \Leftrightarrow A = B$$

- 对称性：

$$dist(A,B) = 1 - \frac{A \cdot B}{||A||_2||B||_2} = 1 - \frac{B \cdot A}{||B||_2||A||_2} = dist(B,A)$$

- 三角不等式：

  给出反例：

$$A = (1,0), B = (1,1), C = (0,1)$$
$$dist(A,B) = 1 - \frac{\sqrt{2}}{2}$$
$$dist(B,C) = 1 - \frac{\sqrt{2}}{2}$$
$$dist(A,C) = 1$$

  因此有：

$$dist(A,B) + dist(B,C) = 2 - \sqrt{2} < 1 = dist(A,C)$$

通过以上证明我们可以看出来，余弦距离是不满足距离的定义的。

优缺点

我们知道余弦相似度关注的是两个向量之间的角度关系，并不关心其绝对大小。在推荐系统的最直接的优点在于：不同用户对电影的打分力度不同，有的严一点平均打分低，有的松一点平均打分都很高，用余弦相似性可以排除打分程度的干扰，关注相对差异。

总的来说欧式距离体现的数值上的绝对差异，而余弦距离体现方向上的相对差异。

A/B测试

A/B 测试是验证模型最终效果的主要手段。当进行 A/B 测试时，通常会采用两个（或多个）组：A 组和 B 组。第一个组是对照组，第二个组会改变其中一些因素。

为什么需要 A/B 测试

1. 离线评估无法消除模型过拟合的影响，因此得出的离线评估结果无法完全替代线上评估结果；
2. 离线评估无法完全还原线上的工程环境，如：数据丢失、标签缺失等情况；
3. 某些评估指标离线状态下无法评估，比如：用户点击率、留存时长、PV 访问量等。

理论基础

中心极限定理：给定一个任意分布的总体，每次从这些总体中随机抽取 n 个抽样，一共抽 m 次。然后把这 m 组抽样分别求出平均值。这些平均值的分布接近正态分布。

中心极限定理是 A/B 测试分析数据的基础，我们可以通过随机抽取样本来估计出总体样本的均值和方差。

设计原则

对用户进行分桶，将用户分成实验组和对照组，对实验组的用户用新模型，对照组用就模型。分桶过程中注意样本的**独立性**和采样方式的**无偏性**，从而确保同一用户只能被分到一个桶中。

假设检验

假设检验的基本原理是先对总体的特征作出某种假设，然后通过抽样研究的统计推理，对此假设应该被拒绝还是接受作出推断。假设检验意味着我们需要给出一个决定：到底是相信原假设，还是相信备择假设。

其大概步骤为：

1. 提出问题（给出零假设和备选假设，两个假设互补）；
2. 收集证据（零假设成立时，得到样本平均值的概率：p 值）；
3. 判断标准（显著水平 α，0.1% 1% 5%）；
4. 做出结论（p<=α，拒绝零假设，否则接受）。

假设检验的精髓在于，根据已有数据信息构造出合理的检验统计量，当我看到这个统计量大于某一个数值的时候的就舍弃原假设，不然我就相信它。

常见假设检验的种类包括：t 检验，z 检验，卡方检验。

t 检验

也称学生检验，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。目的在于比较样本均数，所代表的未知总体均数 μ 和已知总体均数 μ_0 的比较。

适用条件：

1. 已知一个总体均数；
2. 可得到一个样本均数及该样本标准差；
3. 样本来自正态或近似正态总体。

步骤：

1. 建立假设 H_0：$\mu_1 = \mu_2$，即先假定两个总体平均数之间没有显著差异；
2. 计算统计量 T 值，对于不同类型的问题选用不同的统计量计算方法；
3. 根据自由度 $df = n - 1$，查 T 值表，找出规定的 T 理论值并进行比较。理论值差异的显著水平为 0.01 级或 0.05 级；
4. 比较计算得到的t值和理论T值，推断发生的概率，依据给出的T值与差异显著性关系表作出判断。

z 检验

z 检验是一般用于大样本(即样本容量大于 30)平均值差异性检验的方法。它是用标准正态分布的理论来推断差异发生的概率，从而比较两个平均数的差异是否显著。

步骤：

1. 建立虚无假设 H_0：$\mu_1 = \mu_2$，即先假定两个平均数之间没有显著差异；
2. 计算统计量 Z 值，对于不同类型的问题选用不同的统计量计算方法；
3. 比较计算所得 Z 值与理论 Z 值，推断发生的概率，依据 Z 值与差异显著性关系表作出判断。

卡方检验

前两个都是正态分布检验，卡方检验属于非参数检验。主要是比较两个及两个以上样本率(构成比)以及两个分类变量的关联性分析。其根本思想就是在于比较理论频数和实际频数的吻合程度问题。

卡方检验是以 χ^2 分布为基础的一种常用假设检验方法，它的无效假设 H_0 是：观察频数与期望频数没有差别。

卡方检验的基本思想是：首先假设 H_0 成立，基于此前提计算出 χ^2 值，它表示观察值与理论值之间的偏离程度。根据 χ^2 分布及自由度可以确定在 H0 假设成立的情况下获得当前统计量及更极端情

况的概率 P。如果 P 值很小，说明观察值与理论值偏离程度太大，应当拒绝零假设，表示其具有显著性差异；否则就接受零假设。

χ^2 值表示观察值与理论值之间的偏离程度，其大致步骤如下：

1. 设 A 代表某个类别的观察频数，E 代表基于 $H+0$ 计算出的期望频数，A 与 E 之差称为残差；
2. 残差可以表示某一个类别观察值和理论值的偏离程度，但如果将残差简单相加以表示各类别观察频数与期望频数的差别，则有一定的不足之处。因为残差有正有负，相加后会彼此抵消，总和仍然为 0，为此可以将残差平方后求和；
3. 另一方面，残差大小是一个相对的概念，相对于期望频数为 10 时，期望频数为 20 的残差非常大，但相对于期望频数为 1000 时 20 的残差就很小了。考虑到这一点，人们又将残差平方除以期望频数再求和，以估计观察频数与期望频数的差别。

进行上述操作之后，就得到了常用的 χ^2 统计量，其公式如下：

$$\chi^2=\sum\frac{(A-E)^2}{E}=\sum_{i=1}^{k}\frac{(A_i-E_i)^2}{E_i}=\sum_{i=1}^{k}\frac{(A_i-np_i)^2}{np_i}$$

A_i 为 i 水平的观察频数，E_i 为 i 水平的期望频数，n 为总频数，p_i 为 i 水平的期望频率。i 水平的期望频数 E_i 等于总频数 n 乘 i 水平的期望概率 p_i，k 为单元格数。当 n 比较大时，χ^2 统计量近似服从 k-1 (计算 E_i 时用到的参数个数)个自由度的卡方分布。

例子——独立性检验：

某机构欲了解现在性别与收入是否有关，他们随机抽样 500 人，询问对此的看法，结果分为“有关、无关、不好说“三种答案，图中为调查得到的数据：

	A	B	C	D	E
1			抽样数据		
2	性别	有关	无关	不知道	合计
3	男	120	60	50	230
4	女	100	110	60	270
5	合计	220	170	110	500

1. 零假设 H0：性别与收入无关。

2. 确定自由度为 (3-1)×(2-1)=2，选择显著水平 α=0.05。

3. 求解男女对收入与性别相关不同看法的期望次数，这里采用所在行列的合计值的乘机除以总计值来计算每一个期望值，在单元格 B9 中键入“=B5*E3/E5”，同理求出其他值。

	A	B	C	D	E
1			抽样数据		
2	性别	有关	无关	不知道	合计
3	男	120	60	50	230
4	女	100	110	60	270
5	合计	220	170	110	500
6					
7			期望值		
8	性别	有关	无关	不知道	合计
9	男	101.2	78.2	50.6	230
10	女	118.8	91.8	59.4	270
11	合计	220	170	110	500
12					

4. 利用卡方统计量计算公式计算统计量，在单元格 B15 中键入 “=(B3-B9)^2/B9”，其余单元格依次类推，结果如下所示：

	A	B	C	D	E
1			抽样数据		
2	性别	有关	无关	不知道	合计
3	男	120	60	50	230
4	女	100	110	60	270
5	合计	220	170	110	500
6					
7			期望值		
8	性别	有关	无关	不知道	合计
9	男	101.2	78.2	50.6	230
10	女	118.8	91.8	59.4	270
11	合计	220	170	110	500
12					
13			统计量		
14	性别	有关	无关	不知道	合计
15	男	3.49249	4.235806	0.007115	7.73541
16	女	2.975084	3.608279	0.006061	6.589424
17	合计	6.467574	7.844084	0.013175	14.32483

5. 最后得出统计量为 14.32483，而显著水平为 0.05 自由度为 2 卡方分布的临界值为 5.9915。

6. 比较统计量度和临界值，统计量 14.32483 大于临界值 5.9915，故拒绝零假设。

参考

https://wiki.mbalib.com/wiki/%E5%8D%A1%E6%96%B9%E6%A3%80%E9%AA%8C
