

t-test

1、适用条件：

t检验 的适用条件和前提要求如下：

1. **正态性假设**：t检验假设数据来自正态分布。这是因为t检验的理论基础是基于正态分布的性质。

- **独立样本t检验**：假设每个样本数据来自正态分布。

假设样本 X_1, X_2, \dots, X_n 和 Y_1, Y_2, \dots, Y_m 分别来自两个正态分布：

$$X_i \sim N(\mu_X, \sigma_X^2) \quad Y_i \sim N(\mu_Y, \sigma_Y^2)$$

- **配对样本t检验（通常相关系数不为0）**：假设成对差值数据来自正态分布。

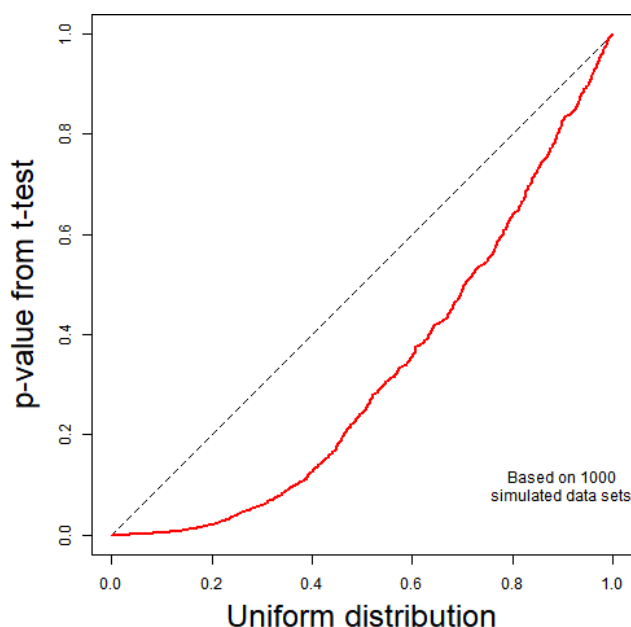
假设差值 $D_i = X_i - Y_i$ 来自正态分布：

$$D_i \sim N(\mu_D, \sigma_D^2)$$

2. **样本独立性**：样本数据应是相互独立的。也就是说，一个样本中的数据不会影响另一个样本的数据。

$$P(X_i, X_j) = P(X_i) \cdot P(X_j) \quad \text{for } i \neq j$$

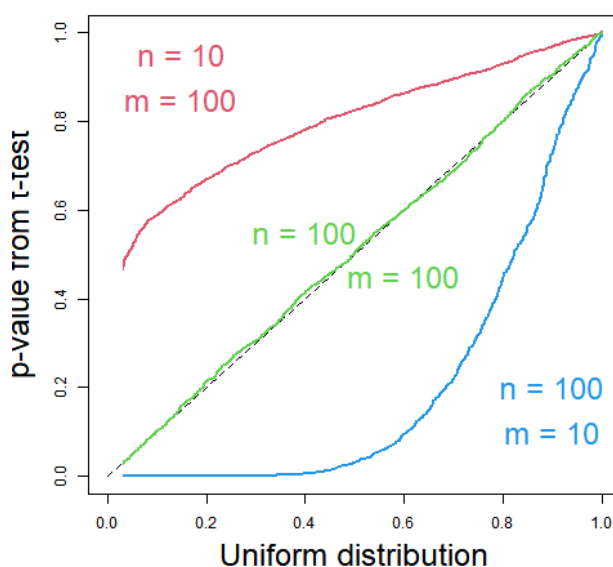
$$P(Y_i, Y_j) = P(Y_i) \cdot P(Y_j) \quad \text{for } i \neq j$$



$X \sim N(0, 1), Y \sim N(0, 1), \rho(X, Y) = 0.5$, 这个条件下两随机变量不独立, **p值** 就距离理想情况较远。

3. **方差齐性 (仅适用于独立样本t检验)** : 对于独立样本t检验, 假设两个样本的方差相等。如果方差不等, 可以使用Welch's t检验。

$$\sigma_1^2 = \sigma_2^2$$

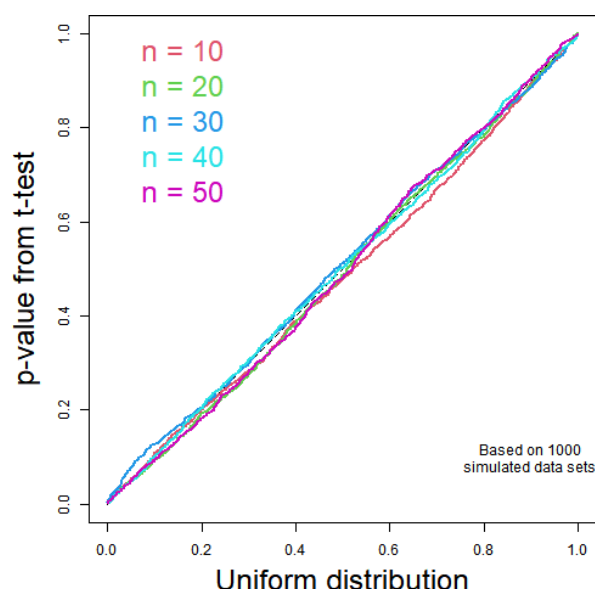


上图所示为 $X \sim N(0, 1), Y \sim N(0, 10)$ 下分别取样 n, m 数量后进行**t检验** 后的 **p值** 可以看出

- **对角线 (45度线, 虚线)** : 代表理想情况下, p值完全服从均匀分布的情况。
- **红色线 ($n = 10, m = 100$)** : 这条线明显偏离对角线, 尤其是在低p值区间 (<0.5) , 表明在样本量不对称且总体标准差差异较大的情况下, t检验的p值分布明显偏离均匀分布。这意味着在这种情况下, t检验可能会较频繁地产生较大的p值, 从而不容易拒绝零假设。
- **绿色线 ($n = 100, m = 100$)** : 这条线大致接近对角线, 表明在样本量相同且较大时, 即使总体标准差有很大差异, t检验的p值分布仍接近均匀分布。这显示了t检验在样本量较大且均衡时的鲁棒性。
- **蓝色线 ($n = 100, m = 10$)** : 这条线在低p值区间明显偏离对角线, 表明在样本量不对称且总体标准差差异较大的情况下 (尤其是y的样本量较少) , t检验的p值分布明显偏离均匀分布。这意味着在这种情况下, t检验较容易产生较小的p值, 从而更容易拒绝零假设。

4. **数据的尺度**: t检验通常用于连续数据或近似连续的数据。

- 对于小样本（通常 $n < 30$ ），正态性假设尤其重要。
- 对于大样本（ $n \geq 30$ ），根据中心极限定理，即使数据不完全符合正态分布，t检验也能获得较为可靠的结果。



上图为 $X, Y \sim \Gamma(1, \frac{1}{10})$ 分布下的，正态性假设被违反时**t检验 p值** 的分布情况

横轴 (Uniform distribution)：表示理想情况下均匀分布的p值，从0到1。这实际上是理论上的均匀分布的分位数。

纵轴 (p-value from t-test)：表示从t检验中得到的p值。

尽管 **t检验** 假设数据来自正态分布，但在样本量足够大的情况下（通常 $n > 30$ ），t检验对正态性假设的违反并不敏感。根据 **中心极限定理**，样本均值的分布将趋近于正态分布，即使原始数据不完全符合正态分布。因此，在大样本情况下，t检验仍然可以有效。

补充说明：

- **小样本情况下**：如果样本量较小且数据明显偏离正态分布，可以考虑使用非参数检验，如 Mann-Whitney U检验（对于独立样本）或 Wilcoxon符号秩检验（对于配对样本）。
- **正态性检验**：可以使用 Shapiro-Wilk检验或 Kolmogorov-Smirnov检验等方法检验数据是否来自正态分布。

总结：t检验假设数据来自正态分布，特别是在样本量较小的时候。然而，在样本量较大时，即使数据不完全符合正态分布，t检验仍然具有较强的鲁棒性，可以有效地使用。

理解p值在零假设成立时接近均匀分布的数学基础，可以通过假设检验的定义和概率论来解释。这种理解主要基于p值的性质和其累积分布函数（CDF）的性质。

2、性质：p值在零假设 H_0 成立时接近均匀分布

p值的定义

假设我们有一个统计量 T ，其分布在零假设 H_0 成立时已知。p值定义为：

$$p = P(T \geq T_{\text{obs}} | H_0)$$

其中， T_{obs} 是观测到的统计量值。

p值的分布

在零假设 H_0 成立时，统计量 T 的分布是已知的。如果我们将 T 转换为p值，这个p值的分布函数（CDF）在零假设成立时是均匀分布的。具体地说：

$$U = F_T(T_{\text{obs}})$$

其中， F_T 是统计量 T 的分布函数。

对于任何连续分布的统计量 T ，其累积分布函数 F_T 在0到1之间均匀分布。也就是说：

$$U \sim \text{Uniform}(0, 1)$$

数学证明

设 T 是一个随机变量，其累积分布函数为 $F_T(t)$ 。我们可以证明 $U = F_T(T)$ 是均匀分布的。

1. 设 $U = F_T(T)$ ，我们想求 U 的分布函数。

2. 对于任意 u 在区间 $[0, 1]$ 上，有：

$$P(U \leq u) = P(F_T(T) \leq u)$$

3. 由于 F_T 是单调递增函数，反函数 F_T^{-1} 存在，因此：

$$P(F_T(T) \leq u) = P(T \leq F_T^{-1}(u))$$

4. 因为 F_T 是 T 的分布函数，所以：

$$P(T \leq F_T^{-1}(u)) = F_T(F_T^{-1}(u)) = u$$

由此可见， U 在 $[0, 1]$ 上均匀分布，即：

$$P(U \leq u) = u, \quad \text{for } 0 \leq u \leq 1$$

总结

- 当零假设 H_0 成立时，p值的分布应该是均匀分布。
- 这种均匀性确保了在零假设下，任何给定的显著性水平 α 都能够正确地控制第一类错误率（Type I error rate）。
- 数学证明基于统计量的分布函数转换性质，说明了p值在0到1之间均匀分布的合理性。

因此，p值接近均匀分布是零假设成立时假设检验的一项重要性质，这确保了检验过程的正确性和公平性。

3、t检验的步骤和计算方法

单样本t检验（One-sample t-test）

用于检验样本均值与已知值（或假设的总体均值）之间的差异。

步骤

1. 设定假设：

- 零假设 H_0 ：样本均值等于总体均值， $\mu = \mu_0$
- 备择假设 H_1 ：样本均值不等于总体均值， $\mu \neq \mu_0$

2. 计算检验统计量：

$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

- \bar{x} 是样本均值
- μ_0 是假设的总体均值
- s 是样本标准差
- n 是样本大小

3. 计算自由度：

$$df = n - 1$$

4. 查找临界值或计算p值：

- 根据计算出的t值和自由度，在t分布表中查找临界值，或者使用统计软件计算p值。

5. 作出结论：

- 如果p值小于设定的显著性水平（如0.05），则拒绝零假设；否则，不拒绝零假设。

独立两样本t检验 (Independent two-sample t-test)

用于比较两个独立样本的均值是否存在显著差异。

步骤

1. 设定假设：

- 零假设 (H_0)：两个样本均值相等, $\mu_1 = \mu_2$
- 备择假设 (H_1)：两个样本均值不等, $\mu_1 \neq \mu_2$

2. 计算检验统计量：

- 若假设两个样本的方差相等：

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

其中, s_p 是合并标准差 (pooled standard deviation)：

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

- \bar{x}_1 和 \bar{x}_2 是两个样本的均值
 - s_1 和 s_2 是两个样本的标准差
 - n_1 和 n_2 是两个样本的样本大小
- 若不假设两个样本的方差相等 (Welch's t-test)：

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

3. 计算自由度：

- 假设方差相等时：

$$df = n_1 + n_2 - 2$$

- 假设方差不等时 (Welch's t-test)：

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{\left(\frac{s_1^2}{n_1} \right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2} \right)^2}{n_2 - 1}}$$

4. 查找临界值或计算p值：

- 根据计算出的t值和自由度，在t分布表中查找临界值，或者使用统计软件计算p值。

5. 作出结论：

- 如果p值小于设定的显著性水平（如0.05），则拒绝零假设；否则，不拒绝零假设。

配对样本t检验 (Paired t-test)

用于比较同一组对象在不同条件下的均值差异（例如，前后测试的结果）。

步骤

1. 设定假设：

- 零假设 (H_0)：配对差异的均值为零， $\mu_d = 0$
- 备择假设 (H_1)：配对差异的均值不等于零， $\mu_d \neq 0$

2. 计算检验统计量：

$$t = \frac{\bar{d}}{s_d / \sqrt{n}}$$

- \bar{d} 是配对差异的均值（即每对数据的差异的均值）
- s_d 是配对差异的标准差
- n 是配对数

3. 计算自由度：

$$df = n - 1$$

4. 查找临界值或计算p值：

- 根据计算出的t值和自由度，在t分布表中查找临界值，或者使用统计软件计算p值。

5. 作出结论：

- 如果p值小于设定的显著性水平（如0.05），则拒绝零假设；否则，不拒绝零假设。

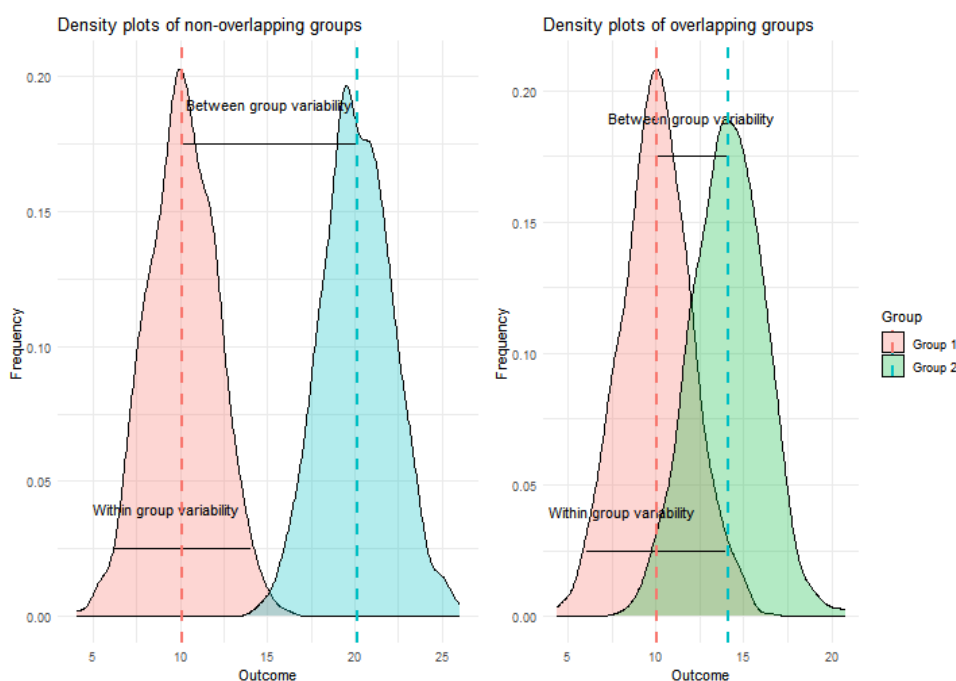
p值的计算

p值的计算基于t统计量和其对应的自由度，可以通过查找t分布表或使用统计软件来完成。具体来说：

- **单侧检验：**
 - 若t值为正，p值为t值对应的上尾概率。
 - 若t值为负，p值为t值对应的下尾概率。
- **双侧检验：**
 - p值为单侧检验p值的两倍。

通过这些步骤和公式，t检验可以帮助我们评估样本均值之间是否存在显著差异，从而作出统计推断。

4、图像说明



这个图像标题为“t-test groups visualized”，展示了t检验中比较两组数据时的情况。图像分为两部分，分别展示了两种不同的组间关系：非重叠组（non-overlapping groups）和重叠组（overlapping groups）。

左图：非重叠组（Non-overlapping groups）

- **组间差异 (Between group variability)**：两组数据的均值差异显著，图中两组数据的分布几乎没有重叠。
- **组内差异 (Within group variability)**：每组数据自身的变异情况（即每组数据的标准差）。
- 红色和蓝色分布代表两组数据（Group 1 和 Group 2），组间差异较大，这通常会导致t检验中容易拒绝零假设（即认为两组均值存在显著差异）。

右图：重叠组 (Overlapping groups)

- **组间差异 (Between group variability)**：两组数据的均值差异较小，图中两组数据的分布有较多重叠。
- **组内差异 (Within group variability)**：每组数据自身的变异情况（即每组数据的标准差）。
- 红色和绿色分布代表两组数据（Group 1 和 Group 2），组间差异较小，这通常会导致t检验中不容易拒绝零假设（即认为两组均值可能没有显著差异）。

总结

这个图像通过展示两种不同情况下的密度图，直观地解释了t检验的概念：

1. **组间差异 (Between group variability)**：反映两组均值的差异。当组间差异大时，t值较大，容易拒绝零假设。
2. **组内差异 (Within group variability)**：反映每组数据内部的变异情况。当组内差异小而组间差异大时，容易发现显著差异。

这两张图通过箱线图的方式直观地展示了t检验在不同组间差异情况下的表现，帮助理解t检验的应用和结果解释。

图1：非重叠组 (Non-overlapping groups)

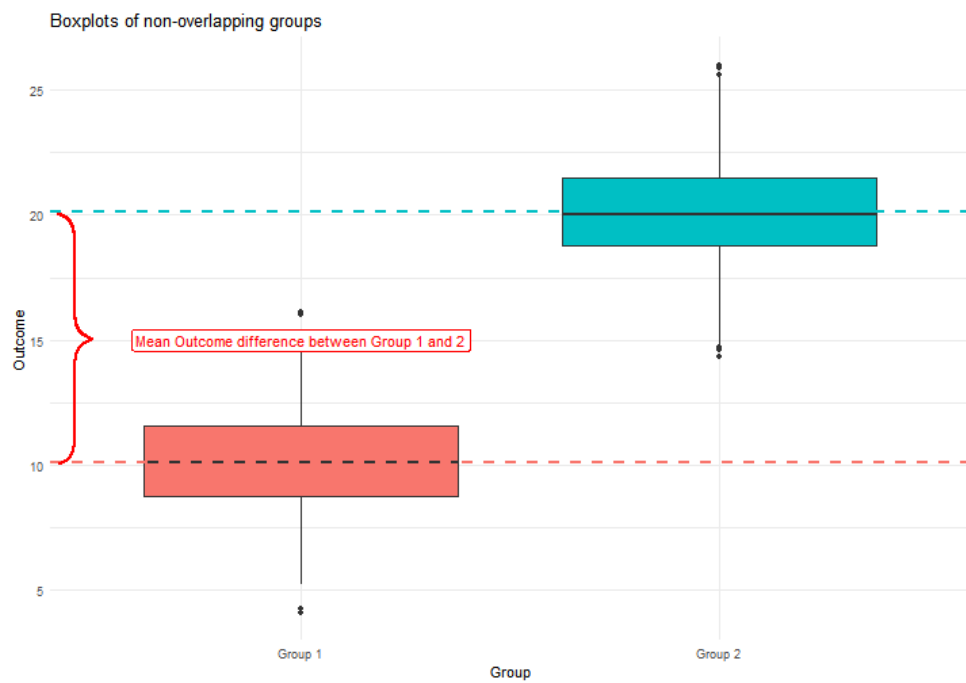
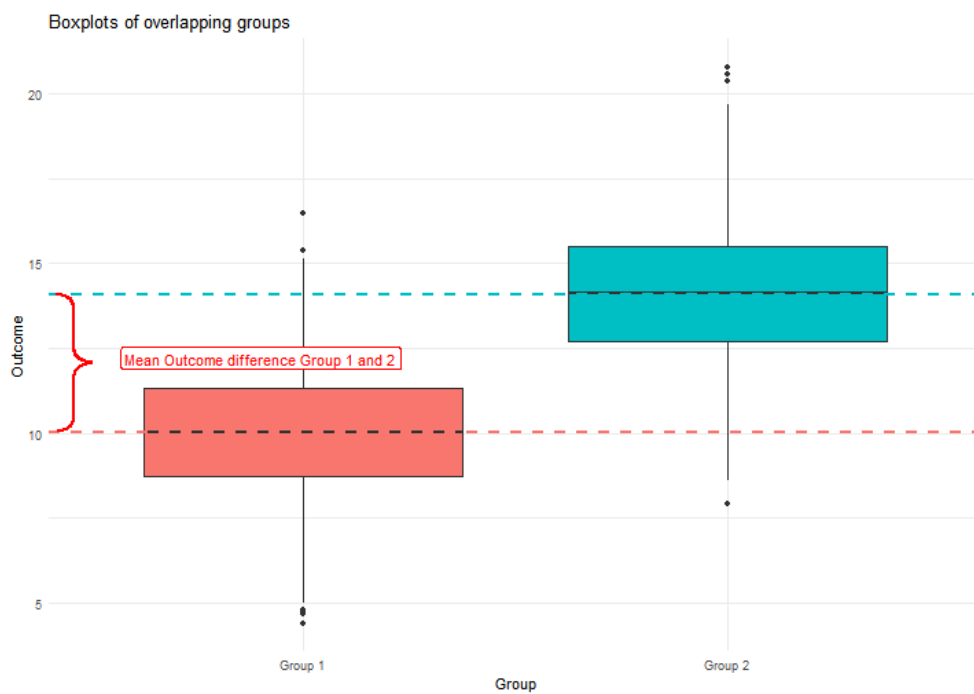


图2：重叠组 (Overlapping groups)



解释

非重叠组 (Non-overlapping groups)

- **组1 (红色)**：数据分布在较低范围。
- **组2 (蓝色)**：数据分布在较高范围。
- **组间差异**：两组均值的差异显著，几乎没有重叠。
- **结论**：在这种情况下，t检验能够明显检测到两组均值的显著差异，容易拒绝零假设。

重叠组 (Overlapping groups)

- **组1 (红色)** : 数据分布在中等范围。
- **组2 (蓝色)** : 数据分布在稍高范围。
- **组间差异**: 两组均值差异较小, 存在较多重叠。
- **结论**: 在这种情况下, t检验可能难以检测到显著差异, 可能不拒绝零假设。

结合说明

1. 组内差异 (Within group variability) :

- 这两个图中, 每组数据的箱线图展示了组内的变异情况。
- 组内差异反映了数据的离散程度, 决定了t检验统计量的分母部分 (标准误)。

2. 组间差异 (Between group variability) :

- 组间差异反映了两组均值的差异, 决定了t检验统计量的分子部分。
- 在非重叠组中, 组间差异大, 导致t值较大, 容易拒绝零假设。
- 在重叠组中, 组间差异小, 导致t值较小, 可能不拒绝零假设。

3. t检验的原理:

- t检验的统计量计算公式为:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

- 其中, \bar{x}_1 和 \bar{x}_2 是两组的均值, s_p 是合并标准差。
- 当组间差异 (均值差异) 显著大于组内差异时, t值较大, p值较小, 容易拒绝零假设。
- 当组间差异不明显时, t值较小, p值较大, 难以拒绝零假设。

总结

通过结合这两张图, 可以更全面地理解t检验在不同组间差异情况下的表现。非重叠组展示了显著组间差异导致t检验容易拒绝零假设, 而重叠组展示了较小组间差异导致t检验难以发现显著差异。这种可视化方式直观地展示了t检验的基本原理和应用效果, 有助于更好地理解统计检验的逻辑和结果解释。

5、t检验的用途和实际例子

t检验的用途

t检验是一种用于比较两个样本均值是否存在显著差异的统计方法，常用于以下几种情境：

1. **单样本t检验 (One-sample t-test)**：比较样本均值与已知总体均值。
2. **独立两样本t检验 (Independent two-sample t-test)**：比较两个独立样本的均值。
3. **配对样本t检验 (Paired t-test)**：比较同一组对象在不同条件下的均值。

实际例子

我们以一个具体的例子说明t检验的应用：

例子：新药物对血压的影响

假设我们想研究一种新药物对降低高血压患者血压的效果。我们可以设计一个实验，随机选取高血压患者，分成两组：一组接受新药物（实验组），另一组接受安慰剂（对照组）。经过一段时间的治疗后，我们测量两组患者的血压下降幅度。

步骤

1. 收集数据：

- 实验组（接受新药物）的血压下降数据：

$$X = [15, 20, 25, 22, 18, 16, 19, 21, 24, 23]$$

- 对照组（接受安慰剂）的血压下降数据：

$$Y = [5, 8, 12, 10, 7, 6, 9, 11, 13, 14]$$

2. 设定假设：

- 零假设 H_0 ：新药物与安慰剂对血压的影响没有差异， $\mu_1 = \mu_2$ 。
- 备择假设 H_1 ：新药物对血压的影响不同于安慰剂， $\mu_1 \neq \mu_2$ 。

3. 选择检验方法：

- 由于我们有两个独立的样本组，采用独立两样本t检验。

4. 计算检验统计量：

- 计算两个样本的均值和标准差：
 - 实验组均值 $\bar{x}_1 = 20.3$ ，标准差 $s_1 = 3.23$ 。
 - 对照组均值 $\bar{x}_2 = 9.5$ ，标准差 $s_2 = 3.01$ 。

- 计算合并标准差（假设方差相等）：

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{(10 - 1)3.23^2 + (10 - 1)3.01^2}{10 + 10 - 2}} = 3.12$$

- 计算t统计量：

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{20.3 - 9.5}{3.12 \sqrt{\frac{1}{10} + \frac{1}{10}}} = \frac{10.8}{1.39} = 7.77$$

5. 计算自由度：

$$df = n_1 + n_2 - 2 = 10 + 10 - 2 = 18$$

6. 查找临界值或计算p值：

- 查找t分布表，或者使用统计软件计算p值。
- 通过统计软件，我们得到p值非常小，远小于0.05。

7. 作出结论：

- 由于p值小于0.05，我们拒绝零假设，认为新药物对血压的影响显著不同于安慰剂。

总结

这个例子展示了t检验在医学研究中的应用。通过t检验，我们能够判断新药物是否在统计学上显著降低了血压。类似地，t检验也广泛应用于心理学、教育学、市场研究等领域，用于比较两个组之间的均值差异，从而为决策提供数据支持。

6、t 分布表

以下是一个常见的t分布表，它列出了不同自由度（degrees of freedom, df）和显著性水平（ α ）下的临界值（critical value）。

t分布临界值表

df / α	0.10	0.05	0.025	0.01	0.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
15	1.341	1.753	2.131	2.602	2.947
20	1.325	1.725	2.086	2.528	2.845
25	1.316	1.708	2.060	2.485	2.787
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
50	1.299	1.676	2.009	2.403	2.678
60	1.296	1.671	2.000	2.390	2.660
80	1.292	1.664	1.990	2.374	2.639
100	1.290	1.660	1.984	2.364	2.626

如何使用t分布表

- 1. **确定自由度 df** : 根据样本量计算自由度。例如，在独立两样本t检验中，自由度 $df = (n_1 + n_2 - 2)$ 。
- 2. **选择显著性水平 α** : 通常使用0.10、0.05、0.025、0.01或0.005。
- 3. **查找临界值**: 在表中找到对应自由度和显著性水平下的t临界值。

计算出的t检验统计量的用途

- **比较计算出的t值和表中的临界值：**

- 如果计算出的t值大于表中的临界值，则拒绝零假设。
- 如果计算出的t值小于或等于表中的临界值，则不拒绝零假设。

数学意义

t检验统计量 t 的数学意义如下：

- t值表示样本均值与假设总体均值（或两个样本均值差异）相差的标准误数量。
- t分布在零假设为真时近似呈正态分布，但在小样本情况下更准确。

公式：

$$t = \frac{\text{样本均值差异}}{\text{标准误}}$$

- 样本均值差异：两个样本均值之间的差异。
- 标准误：样本均值差异的标准误。

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

其中 s 是样本标准差， n 是样本量。

示例应用

在新药物对血压影响的例子中：

- 计算的t值为7.77，自由度df为18。
- 查表（显著性水平0.05），df为18时，临界值约为2.101。
- 由于 $7.77 > 2.101$ ，拒绝零假设，认为新药物显著降低血压。

总结

t检验通过比较样本均值差异与标准误的比值（t值），结合t分布表，判断是否拒绝零假设。这种方法广泛用于各种科学研究中，以检测两组数据均值是否存在显著差异。

7、p 值的作用

在t检验中，p值的计算和使用是非常重要的，它提供了对结果的显著性进行解释的一种方式。虽然可以通过比较计算得到的t统计量和t分布表中的临界值来做决定，但p值提供了一种更灵活和直

接的方法来理解结果的显著性。

p值的作用

1. 解释显著性:

- p值表示在零假设 (H_0) 为真的情况下, 观察到或更极端的结果出现的概率。
- 小p值 (通常小于显著性水平, 如0.05) 表示观察到的数据与零假设不符, 建议拒绝零假设。

2. 灵活性:

- 使用p值可以避免查找t分布表的麻烦。
- 直接使用p值与显著性水平比较, 可以方便地在任何显著性水平下做出判断, 而不需要每次都查表。

t检验中的p值计算

t检验中的p值计算基于t统计量和自由度, 可以通过统计软件或查表得到。

步骤总结

1. 计算t统计量:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{SE}$$

其中, (SE) 是标准误。

2. 计算自由度:

- 对于两独立样本t检验, 假设方差相等时, 自由度为:

$$df = n_1 + n_2 - 2$$

3. 计算p值:

- 使用统计软件或查表, 根据计算出的t统计量和自由度, 找到相应的p值。
- 例如, 在Python中, 可以使用 `scipy.stats.ttest_ind` 函数来计算p值。

4. 比较p值与显著性水平:

- 如果p值小于显著性水平 (如0.05), 则拒绝零假设。
- 如果p值大于显著性水平, 则不拒绝零假设。

具体例子

继续之前的新药物对血压影响的例子：

数据：

- 实验组（新药物）： $X = [15, 20, 25, 22, 18, 16, 19, 21, 24, 23]$
- 对照组（安慰剂）： $Y = [5, 8, 12, 10, 7, 6, 9, 11, 13, 14]$

步骤：

1. 计算t统计量：

- 样本1均值 $\bar{x}_1 = 20.3$
- 样本2均值 $\bar{x}_2 = 9.5$
- 合并标准差 $s_p = 3.12$
- 标准误 $SE = 1.39$
- t统计量 $t = \frac{20.3-9.5}{1.39} = 7.77$

2. 计算自由度：

- $df = 10 + 10 - 2 = 18$

3. 计算p值：

- 使用统计软件（例如Python中的SciPy库）：

```
from scipy import stats
t_stat, p_value = stats.ttest_ind(X, Y, equal_var=True)
print("t统计量:", t_stat)
print("p值:", p_value)
```

假设得到的p值为 $p < 0.0001$

4. 比较p值与显著性水平：

- 由于p值非常小，小于0.05，我们拒绝零假设，认为新药物对降低血压有显著影响。

总结

- **p值的计算**：t检验中，p值是通过计算t统计量和自由度，并查找t分布得到的。
- **p值的使用**：比较p值与显著性水平，可以判断是否拒绝零假设。

- **t值与p值的关系：**计算出的t统计量用来查找p值，p值提供了一种灵活且直接的显著性解释方法。