Inference about Two Populations

Applied Statistics

Fall 2025

目录

大纲 Outline

1. 两个均值的差异：独立样本 The difference between two means: independent samples
2. 配对比较检验 Paired Comparison test
3. 方差的比较：两个方差的比率 Comparing variances: the ratio of two variances

1 双样本推断

1.1 为何需要两总体推断 Why Two-Population Inference

- 我们经常需要比较两个总体均值的差异。
  We often need to compare differences in means of two populations.
- 示例：
  Examples:
  - 性别工资差距：男性与女性的平均工资
    Gender wage gap: average wage of males vs females
  - 新药有效性：治疗组与对照组
    Effectiveness of a new drug: treatment group vs control group
  - 政策实施前后
    Before vs after a policy
- 在所有这些问题中，关键问题不是“均值是多少?”，而是“两个均值有多大差异?”
  In all these examples, the key question is not “What is the mean?” but “How different are the two means?”

1.2 独立样本与配对样本 Independent Samples vs Paired Samples

1.2.1 两种研究设计类型 Two Types of Study Design

- **独立样本 Independent samples:** 一组中的观测值与另一组中的观测值相互独立。
  the observations in one group are independent of those in the other group.
- **配对样本 Paired samples:** 一个样本中的每个观测值与第二个样本中的一个观测值相匹配。不独立。
  each observation in one sample is matched with an observation in a second sample. Not independent.

1.2.2 对比示例 Comparison Examples

问题	独立样本	配对样本
训练能提高分数吗	比较两个班级，训练组与未训练组	比较相同的学生训练前后
药物能降低血压吗?	比较治疗组和对照组	比较病人治疗前后的血压

表 1: 独立样本与配对样本的对比 Comparison of Independent and Paired Samples

注意：区分取决于数据是如何收集的。
Note: The distinction depends on how the data are collected.

1.3 比较两个独立样本的均值 Comparing two means from two independent samples

1.3.1 参数与统计量 Parameters and Statistics

- 感兴趣的参数是 $\mu_1 - \mu_2$
  Parameters of interest $\mu_1 - \mu_2$.
- $\mu_1 - \mu_2$ 回答了这个问题：平均而言，总体 1 比总体 2 大（或小）多少？
  $\mu_1 - \mu_2$ answers the question: On average, how much larger (or smaller) is population 1 compared to population 2?
- 每个样本都被视为来自不同总体的样本。
  Each is considered to be a sample from a distinct population.

总体或处理	参数	统计量	样本量
1	μ_1	$\bar{x}_1$	n_1
2	μ_2	$\bar{x}_2$	n_2

表 2: 两总体推断的参数与统计量

1.3.2 抽样分布 Sampling Distribution

- 我们依赖 $\bar{x}_1 - \bar{x}_2$ 的抽样分布来对 $\mu_1 - \mu_2$ 进行推断。
  We rely on the sampling distribution of $\bar{x}_1 - \bar{x}_2$ to make inference on $\mu_1 - \mu_2$.
- 幸运的是，中心极限定理适用于 $\bar{x}_1 - \bar{x}_2$。
  Luckily the central limit theorem applies to $\bar{x}_1 - \bar{x}_2$.
- 即使个体总体不服从正态分布，随着样本量的增加，$\bar{x}_1 - \bar{x}_2$ 的抽样分布也趋近于正态分布。
  The sampling distribution of $\bar{x}_1 - \bar{x}_2$ approaches a normal distribution as the sample size increases even if the individual populations are not Normal.
- 极限分布为 $N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$，其中 n_1 和 n_2 是各自的样本量。
  The limiting distribution is $N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$ where n_1 and n_2 are the respective sample sizes.

1.3.3 两样本推断 Two-sample Inference

- 两样本推断是关于总体均值差异 $\mu_1 - \mu_2$ 的。
  Two-sample Inference is about the difference in population means $\mu_1 - \mu_2$.

- 以下两样本统计量服从标准正态分布：
  Two-Sample Statistic below follows a standard normal distribution:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

- 这种情况主要是理论上的。在实践中，总体方差几乎从不知道。
  This case is mainly theoretical. In practice, population variances are almost never known.

1.4 两样本 t 统计量 Two-sample t Statistic

1.4.1 统计量公式

- 当总体方差未知时，我们使用样本计算的方差 s_1 和 s_2 来替代 σ_1 和 σ_2。
  When the population variances are unknown, we use variance calculated from samples s_1 and s_2 to replace σ_1 and σ_2.

- 统计量 $\bar{x}_1 - \bar{x}_2$ 的标准误为：
  The standard error of the statistic $\bar{x}_1 - \bar{x}_2$ is:

$$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

- 检验统计量为：
  The test statistic is:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- 分母衡量来自两个样本的总抽样不确定性。
  The denominator measures total sampling uncertainty coming from both samples.

1.4.2 自由度 Degrees of Freedom

- 这个统计量没有精确的 t 分布。它只有近似的 t 分布，且自由度也是近似的。
  This statistic does not have an exact t distribution. It only has approximately a t distribution with an approximation for the degrees of freedom.

- 我们只能近似自由度 k。
  We can only approximate the degree of freedom k.

- 在实践中，软件会自动计算自由度。你不需要手动计算。
  In practice, software computes the degrees of freedom automatically. You do not need to calculate this by hand.

- 或者我们可以采用保守方法：使用 $n_1 - 1$ 和 $n_2 - 1$ 中较小的一个作为自由度。
  Or we can use a conservative approach: using the smaller of $n_1 - 1$ and $n_2 - 1$ for the degrees of freedom.

1.4.3 软件对自由度的近似 Software Approximation for the Degrees of Freedom

- 有效自由度：大多数统计软件使用 Satterthwaite 近似来计算有效自由度。它根据数据计算得出。
  The effective degree of freedom: Most statistical software uses Satterthwaite's approximation for an effective degree of freedom. It is calculated from the data.

- 公式：
  Formula:

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

- 当两个样本量 n_1 和 n_2 都大于或等于 5 时，这个近似相当准确。
  It is quite accurate when both sample sizes n_1 and n_2 are 5 or larger.

- **注意**：你不需要记忆这个公式。重要的是理解为什么当变异性高时自由度较小。
  Note: You are not expected to memorize this formula. What matters is understanding why the degrees of freedom are smaller when variability is high.

1.5 两样本置信区间 Two-sample Confidence Interval

1.5.1 公式

- 两个均值之间差异的两样本 t 区间：
  Two-Sample t Interval for a Difference Between Means:

$$(\bar{x}_1 - \bar{x}_2) \pm t^*_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

1.5.2 示例：长叶松树胸径差异 Example: Longleaf Pine DBH Difference

两样本置信区间计算 Two-sample Confidence Interval Calculation

背景：为 Wade Tract Preserve 北部和南部长叶松树的平均胸径差异构建并解释一个 90% 的置信区间。

Background: Construct and interpret a 90% confidence interval for the difference in the mean DBH for longleaf pines in the northern and southern halves of the Wade Tract Preserve.

描述性统计：

Descriptive Statistics:

变量 Variable	N	均值 Mean	标准差 StDev
North	30	23.70	17.50
South	30	34.53	14.26

计算：

- 由于条件满足，我们可以为差异 $\mu_1 - \mu_2$ 构建一个两样本 t 区间。我们将使用保守的自由度 $df = 30 - 1 = 29$。
  Since the conditions are satisfied, we can construct a two-sample t interval for the difference $\mu_1 - \mu_2$. We'll use the conservative $df = 30 - 1 = 29$.
- 查表得 $t^* = 1.699$（对于 $df = 29$，双侧 90% 置信区间）。
  From t-table, $t^* = 1.699$ (for $df = 29$, two-sided 90% confidence interval).
- 置信区间：

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = (34.53 - 23.70) \pm 1.699 \sqrt{\frac{14.26^2}{30} + \frac{17.50^2}{30}}$$

$$= 10.83 \pm 7.00 = (3.83, 17.83)$$

解释：我们有 90% 的把握认为从 3.83 到 17.83 厘米的区间包含了两个均值之间的差异。这个区间表明，南部树木的平均直径比北部树木的平均直径大 3.83 到 17.83 厘米。

Interpretation: We are 90% confident that the interval from 3.83 to 17.83 centimeters captures the difference in two means. This interval suggests that the mean diameter of the southern trees is between 3.83 and 17.83 cm larger than the mean diameter of the northern trees.

1.5.3 检查所需条件 Checking the Required Conditions

- **随机性 Random:** 数据来自 30 棵树的随机样本，分别来自森林的北部和南部。
  The data come from random samples of 30 trees, one from the northern half and one from the southern half of the forest.

- **正态性 Normal:** 箱线图中观察到的偏度使我们有理由相信 DBH 测量的总体分布可能不是正态的。然而，由于两个样本量都至少为 30，我们可以安全地使用 t 检验。
  Skewness seen in the boxplots gives us reason to believe that the population distributions of DBH measurements may not be Normal. However, since both sample sizes are at least 30, we are safe using t procedures.

- **独立性 Independent:** 研究者从森林的北部和南部独立抽取样本。
  Researchers took independent samples from the northern and southern halves of the forest.

1.6 两样本 t 显著性检验 Two-Sample t Significance Test

1.6.1 检验步骤

- 假设随机性、正态性和独立性条件得到满足。
  Suppose the Random, Normal, and Independent conditions are met.

- 为检验假设 $H_0 : \mu_1 - \mu_2 = 0$，计算 t 统计量：
  To test the hypothesis $H_0 : \mu_1 - \mu_2 = 0$, compute the t statistic:

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

- 通过计算在备择假设 H_a 指定的方向上获得这么大或更大的 t 统计量的概率来找到 P 值。使用由技术近似或 $n_1 - 1$ 和 $n_2 - 1$ 中较小者作为自由度的 t 分布。
  Find the P-value by calculating the probability of getting a t statistic this large or larger in the direction specified by the alternative hypothesis H_a. Use the t distribution with degrees of freedom approximated by technology or the smaller of $n_1 - 1$ and $n_2 - 1$.

1.6.2 示例：CEO 绩效 Example: CEO Performance

两样本 t 检验（方差不等）Two-sample t test (unequal variances)

背景：一项研究比较了家族企业两种类型 CEO 的绩效：家族成员或来自外部的职业经理人。收集家族企业样本的利润信息和 CEO 类型。

Background: One study compares the performance of two types of CEOs for family-owned businesses: family members or professional managers from outside. Collect profit information and CEO type from a sample of family-owned businesses.

问题：我们能否在 5% 的显著性水平上得出结论，认为两种类型 CEO 的绩效存在显著差异？

Question: Can we conclude at the 5% significance level that there are significant differences in the performance of the two types of CEOs?

描述性统计：

Descriptive Statistics:

变量 Variable	观测数 Obs	均值 Mean	标准差 Std. Dev.
Offspring	42	-0.1	1.946138
Outsider	98	1.235918	2.834568

两样本 t 检验结果：

Two-sample t test results:

- $t = -3.2196$
- Satterthwaite 近似的自由度 $= 110.749$
- 双侧检验的 P 值 $= 0.0017$

```
Two-sample t test with unequal variances

Variable     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
Offspr~g      42        -.1    .3002956    1.946138   -.7064593    .5064593
Outsider      98    1.235918    .2863346    2.834568    .6676234    1.804213
---------+--------------------------------------------------------------------
combined     140    .8351429    .2252244    2.664891    .3898342    1.280452
---------+--------------------------------------------------------------------
    diff           -1.335918    .4149277               -2.158146   -.5136909
------------------------------------------------------------------------------
    diff = mean(Offspring) - mean(Outsider)                       t =  -3.2196
Ho: diff = 0                     Satterthwaite's degrees of freedom =  110.749

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0008         Pr(|T| > |t|) = 0.0017          Pr(T > t) = 0.9992
```

结论：由于 P 值 $= 0.0017 < 0.05$，我们拒绝原假设。有足够的证据表明，家族成员 CEO 和外部职业经理人 CEO 的绩效存在显著差异。

Conclusion: Since P-value $= 0.0017 < 0.05$, we reject the null hypothesis. There is sufficient evidence to conclude that there is a significant difference in performance between family member CEOs and outside professional manager CEOs.

1.7 两样本 t 检验的稳健性 Robustness of the two-sample t test

- 两样本 t 检验比单样本 t 方法更稳健。
  The two-sample t procedures are more robust than the one-sample t methods.
- 当两个样本量相等且两个样本分布相似时，它们是最稳健的。
  They are the most robust when both sample sizes are equal and both sample distributions are similar.
- 相等的样本量增加了对非正态性的稳健性。
  Equal sample sizes increases robustness against non-normality.
- 当计划一个两样本研究时，如果可能，选择相等的样本量。
  When planning a two-sample study, choose equal sample sizes if you can.

1.8 t 检验的使用指南 Using the t Procedures

- 除了小样本情况外，数据来自感兴趣总体的简单随机样本这一条件比正态总体条件更重要。
  Except in the case of small samples, the condition that the data are SRSs from the populations of interest is more important than the norm population.
- 样本量之和小于 15：如果数据看起来接近正态，则使用 t 检验。如果数据明显偏斜或存在异常值，不要使用 t 检验。
  Sum of the sample sizes less than 15: Use t procedures if the data appear close to Normal. If the data are clearly skewed or if outliers are present, do not use t.
- 样本量之和至少为 15 且小于 40：t 检验可以使用，除非存在异常值或强偏度。
  Sum of the sample sizes at least 15 and less than 40: The t procedures can be used except in the presence of outliers or strong skewness.
- 大样本：即使分布明显偏斜，当样本量之和很大（大约 $n_1 + n_2 \geq 40$）时，也可以使用 t 检验。
  Large samples: The t procedures can be used even for clearly skewed distributions when the sum of the sample sizes is large, roughly $n_1 + n_2 \geq 40$.

1.9 小样本推断 Inference for Small Samples

- 当我们没有足够的观测值来检查分布形状时，只有极端的异常值才会显现。这些小样本需要特别注意。
  When we do not have enough observations to examine distribution shapes, only extreme outliers will stand out. These small samples require special care.
- 显著性检验的功效往往较低。（功效衡量其检测备择假设的能力。）
  The power of significance tests tends to be low. (The power measures its ability to detect an alternative hypothesis.)
- 置信区间的误差边际往往较大。
  Margins of error of confidence intervals tend to be large.
- 如果效应很大，即使样本量很小，它仍然应该是明显的。
  If an effect is large, it should still be evident, even with small sample sizes.
- 对于小样本，“不显著”通常意味着“低功效”，而不是“没有效应”。
  With small samples, “no significance” often means “low power,” not “no effect.”

1.10 合并两样本 t 检验 Pooled Two-Sample t Procedures

1.10.1 概述

- 存在另一种版本的两样本 t 检验：假设**方差相等**（“合并两样本检验”）。
  There are another version of the two-sample t-test: one assuming **equal variance** (“pooled two-sample test”).
- 我们到目前为止讨论的是“不等”方差检验，不假设两个总体方差相等。
  The one we discussed so far is the “unequal” variance test, not assuming equal variance for the two populations.
- 它们有略微不同的公式和自由度。
  They have slightly different formulas and degrees of freedom.
- 合并 t 检验假设总体方差相等，这通常是不现实的。
  The pooled t-test assumes equal population variances, which is often unrealistic.
- 如今，通常更倾向于使用不等方差 t 检验（Welch 检验）。
  Today, the unequal-variance t-test (Welch's test) is generally preferred.

1.10.2 合并方差估计量 Pooled Variance Estimator

- 如果假设方差相等，我们可以使用合并方差估计量：
  If assume equal variance, we can use the pooled variance estimator:

$$s_p^2 = \frac{\sum_{i=1}^{n_1}(X_{1i} - \bar{x}_1)^2 + \sum_{i=1}^{n_2}(X_{2i} - \bar{x}_2)^2}{n_1 + n_2 - 2} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

1.10.3 检验统计量与置信区间

- 检验统计量服从自由度为 $n_1 + n_2 - 2$ 的 t 分布：
  The test statistic follows a t-distribution with degrees of freedom $n_1 + n_2 - 2$:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

- 置信区间变为：
  The confidence interval becomes:

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2}(n_1 + n_2 - 2)\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

1.10.4 示例：钙补充剂与血压 Example: Calcium Supplement and Blood Pressure

合并两样本 t 检验 Pooled Two-Sample t Test

背景：增加饮食中的钙摄入量能降低血压吗？一项随机对照实验让一组 10 名黑人男性服用钙补充剂 12 周，另一组 11 名黑人男性服用外观相同的安慰剂。以下是血压降低的汇总统计：

Background: Does increasing the amount of calcium in our diet reduce blood pressure? A randomized comparative experiment gave one group of 10 black men a calcium supplement for 12 weeks and a control group of 11 black men a placebo that appeared identical. The summary statistics are given below for the decreases in blood pressure.

组别 Group	n	均值 Mean	s
Calcium	10	5.000	8.743
Placebo	11	-0.273	5.901

检验： $H_0 : \mu_1 = \mu_2$ 对 $H_a : \mu_1 > \mu_2$（单侧）

Test: $H_0 : \mu_1 = \mu_2$ against $H_a : \mu_1 > \mu_2$ (one-sided).

计算：

- 合并方差：
  Pooled variance:
  $$s_p^2 = \frac{(10-1)8.743^2 + (11-1)5.901^2}{(10+11-2)} = 54.536$$
- 合并标准差：$s_p = \sqrt{54.536} = 7.385$
- t 统计量：
  t statistic:
  $$t = \frac{5.000 - (-0.273)}{7.385\sqrt{\frac{1}{10} + \frac{1}{11}}} = 1.634$$
- 自由度：$df = 10 + 11 - 2 = 19$
- P 值（来自统计软件）：$p = 0.059$

结论：有一些证据，但不是令人信服的证据表明钙补充剂能降低血压。
Conclusion: There is some evidence but not convincing evidence that calcium supplement reduces blood pressure.

1.11 使用哪种检验？Which test to use?

1.11.1 合并方差 t 检验 Pooled Variance t Test

- 合并方差 t 检验长期以来一直是教科书中两样本 t 检验的标准版本。
  The pooled variance t test has long been the standard version of the two-sample t test in textbooks.
- 当样本量相近时，它们对非正态性和不等标准差都具有相当的稳健性。
  They are reasonably robust against both non-normality and unequal standard deviations when the sample sizes are close.
- 它也更容易计算。
  It is also practically easier to calculate.

1.11.2 不等方差 t 检验 Unequal Variance t Test

- 当样本量差异很大时，除非样本量非常大，否则等方差 t 检验对不等标准差变得敏感。
  When the samples are quite different in size, the equal variances t-test becomes sensitive to unequal standard deviations unless the samples are very large.

- 不等标准差相当常见：当中心增大时，数据的散布往往会增加。
  Unequal standard deviations are quite common: the spread of data tend to increase when the center gets larger.
- 过去，人们通常使用两样本方差检验来检查是否可以假设总体方差相等。
  It was once common to use a two-sample variance test to check if we can assume equal population variances.
- 然而，缺乏不等方差的证据并不等同于有相等方差的证据。此外，方差检验假设正态总体，并且对该假设的违反不稳健。
  However, lack of evidence for non-equal variances is not the same as evidence for equal variances. Also, the variance tests assume a normal population and are not robust to violation of such an assumption.
- **因此，除非另有要求，否则使用不等方差 t 检验。**
  Thus, use the unequal variance t-test unless required otherwise.

2 均值差异的推断:配对比较检验 Inference of Mean Differences: Paired Comparison test

2.1 概述

- 第二种均值差异的类型适用于配对样本。
  The second type of difference in means applies to paired samples.
- 示例：同一受试者在两个时期观测的数据，或具有相似特征的匹配受试者。
  Examples: data from the same subjects observed in two periods, or subjects matched to have similar characteristics.
- 配对设计通过控制个体差异来减少变异性。
  Paired designs reduce variability by controlling for individual differences.

2.2 推断方法

- 直接查看差异：$\mu_D = \mu_1 - \mu_2$
  Directly look at the difference: $\mu_D = \mu_1 - \mu_2$.
- 假设直接关于 μ_D，例如 $H_1 : \mu_D > 0$
  The hypothesis is directly about μ_D, e.g. $H_1 : \mu_D > 0$.

- 我们首先计算每对数据的差异，然后计算差异的均值 $\bar{x}_D$ 和标准差 s_D。
  We start with calculate the differences for each pair of data and calculate the mean $\bar{x}_D$ and standard deviation s_D of the differences.
- 其余的推断与单样本 t 检验相同。
  The rest of the inference is the same as a one-sample t test.

2.3 检验统计量与置信区间

- 差异总体均值的检验统计量为：
  The test statistic for the mean of the population of differences is:

$$t = \frac{\bar{x}_D - \mu_D}{s_D/\sqrt{n}}$$

  服从自由度为 $n-1$ 的学生 t 分布。
  a Student t distributed with $n-1$ degrees of freedom.

- 置信区间估计量：
  The confidence interval estimator:

$$\bar{x}_D \pm t_{\alpha/2}\frac{s_D}{\sqrt{n}}$$

2.4 示例：专业与薪资 Example: Major and Salary

配对 t 检验 Paired t Test

背景：我们想知道金融专业毕业生的薪资是否高于营销专业毕业生。随机抽取 50 名最近毕业的 MBA，其中一半主修金融，一半主修营销，记录最高薪资（包括福利）报价。我们能否推断金融专业 MBA 的薪资高于营销专业 MBA？
Background: Suppose that we want to know if finance majors were being offered higher salaries than marketing majors. Randomly sampled 50 recently graduated MBAs half of whom majored in finance and half in marketing and record the highest salary (including benefits) offer. Can we infer that finance majors obtain higher salary offers than do marketing majors among MBAs?
配对设计：我们随机抽取一对金融和营销专业学生，他们的 GPA 在 3.92 到 4 之间（基于最高 4 分），再抽取一对 GPA 在 3.84 到 3.92 之间的金融和营销专业学生。继续这个过程，直到第 25 对金融和营销专业学生，他们的 GPA 在 2.0 到 2.08 之间。匹配是通过选择具有相似 GPA 的金融和营销专业学生来进行的。所得数据是匹配对。
Paired Design: We randomly sample a finance and a marketing major whose grade point average (GPA) falls between 3.92 and 4 (based on a maximum of 4), a finance

and a marketing major whose GPA is between 3.84 and 3.92. We continue this process until the 25th pair of finance and marketing majors are selected whose GPA fell between 2.0 and 2.08. The matching is conducted by selecting finance and marketing majors with similar GPAs. The resulting data are matched pairs.

问题：我们能否从这些数据中得出结论，金融专业毕业生的薪资高于营销专业毕业生？

Question: Can we conclude from these data that finance majors draw larger salary offers than do marketing majors?

配对 t 检验结果：

Paired t test results:

- 差异均值：$\bar{x}_D = 5064.52$
- 差异标准差：$s_D = 6646.895$
- 标准误：1329.379
- t 统计量：$t = 3.8097$
- 自由度：$df = 24$
- 双侧 P 值：0.0009

Paired t test

Variable	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
Finance	25	65438.2	4218.918	21094.59	56730.78	74145.62
Market~g	25	60373.68	4333.321	21666.61	51430.14	69317.22
diff	25	5064.52	1329.379	6646.895	2320.816	7808.224

mean(diff) = mean(**Finance** - **Marketing**) t = **3.8097**
Ho: mean(diff) = 0 degrees of freedom = **24**

Ha: mean(diff) < 0 Ha: mean(diff) != 0 Ha: mean(diff) > 0
Pr(T < t) = **0.9996** Pr(|T| > |t|) = **0.0009** Pr(T > t) = **0.0004**

结论：在显著性水平 $\alpha = 0.05$ 下，P 值 $= 0.0009 < 0.05$，我们拒绝原假设。有足够的证据表明金融专业 MBA 的薪资高于营销专业 MBA。

Conclusion: At significance level $\alpha = 0.05$, P-value $= 0.0009 < 0.05$, we reject the null hypothesis. There is sufficient evidence to conclude that finance majors obtain higher salary offers than marketing majors among MBAs.

2.5 独立样本与配对样本的对比 Independent Samples vs Matched Pairs

对比：配对检验与独立样本检验 Comparison: Paired vs Independent Samples Test

如果忽略配对性并将数据视为两个独立样本会怎样？

What if we ignore the pairedness and treat the data as two independent samples?

不等方差两样本 t 检验结果：

Two-sample t test with unequal variances results:

- t 统计量：$t = 0.8374$
- 自由度：47.9657
- 双侧 P 值：0.4065

Two-sample t test with unequal variances

Variable	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
Finance	25	65438.2	4218.918	21094.59	56730.78	74145.62
Market~g	25	60373.68	4333.321	21666.61	51430.14	69317.22
combined	50	62905.94	3014.711	21317.23	56847.65	68964.23
diff		5064.52	6047.888		-7095.798	17224.84

diff = mean(Finance) - mean(Marketing) t = 0.8374
Ho: diff = 0 Satterthwaite's degrees of freedom = 47.9657

Ha: diff < 0 Ha: diff != 0 Ha: diff > 0
Pr(T < t) = 0.7967 Pr(|T| > |t|) = 0.4065 Pr(T > t) = 0.2033

对比：在我们的薪资示例中，两个检验统计量的分子相同。检验统计量的变化是由于标准误不同：

- 独立样本（合并方差）：$\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} = 6048$
- 配对样本：$\frac{s_D}{\sqrt{n_D}} = 1329$

解释：匹配对减少了变异（已经通过我们匹配的信息部分解释），因此标准误更小。如果匹配的变量解释了结果变量的变异，它将改善比较。如果公司在招聘时不考虑 GPA，那么在这里匹配还有效吗？

Interpretation: The matched pairs reduced the variation (already explained partially by the information we are matching on) and thus smaller standard errors. If the variable being matched on explains the variation in outcome variables, it would improve the comparison. If companies did not consider GPA in hiring, would matching still work here?

3 两个总体方差相等的检验 Testing for the Equality of two Population Variances

3.1 概述

- 比较两个总体方差：
  Comparing two population variances:
  - 检验不同生产方法的质量一致性
    to test the quality consistency of different production methods
  - 比较两个投资组合的风险
    to compare the risks of two investment portfolio
- 我们将描述用于比较两个正态总体散布的 F 检验。
  We will describe the F test for comparing the spread of two Normal populations.

3.2 F 分布 F Distribution

- F 分布由两个自由度定义：v_1 是分子 χ^2 的自由度，v_2 是分母 χ^2 的自由度。
  An F distribution is defined by two degrees of freedom: v_1 is the numerator χ^2 degree of freedom and v_2 is the denominator χ^2 degree of freedom.

$$F = \frac{\chi^2_{v_1}/v_1}{\chi^2_{v_2}/v_2}$$

- F 分布从零开始（非负）且不对称。
  The F distribution starts at zero (is non-negative) and is not symmetrical.

Table 6(a)

Critical Values of F, $A = .05$

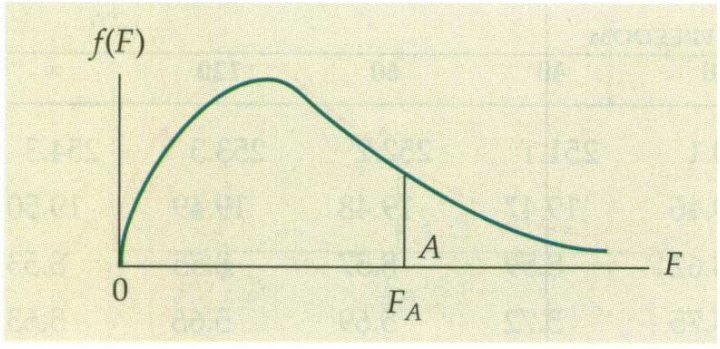


图 1: F Distribution

3.3 两个方差比率的推断 Inference about the ratio of two variances

3.3.1 理论基础

- 推断基于方差的比率。我们知道对于正态总体：
  The inference is based on the ratio of the variances. We know that for normal populations:
  $$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$
- 方差的比率因此服从自由度为 (n_1-1, n_2-1) 的 F 分布：
  The ratio of the variances thus follows an F-distribution with (n_1-1, n_2-1) degrees of freedom:
  $$F = \frac{\frac{(n_1-1)s_1^2}{\sigma_1^2}/(n_1-1)}{\frac{(n_2-1)s_2^2}{\sigma_2^2}/(n_2-1)} = \frac{s_1^2/s_2^2}{\sigma_1^2/\sigma_2^2}$$

3.3.2 假设检验

- 我们的原假设通常是检验两个方差是否相等：
  Our null hypothesis is usually to test for the equality of two variances:
  $$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$$
- 在原假设下，检验统计量简化为：
  Under the null hypothesis, the test statistic simplifies to:
  $$F = \frac{s_1^2}{s_2^2}$$
  它服从自由度为 $df_1 = n_1 - 1$ 和 $df_2 = n_2 - 1$ 的 F 分布。
  which is F-distributed with degrees of freedom $df_1 = n_1 - 1$ and $df_2 = n_2 - 1$.

3.4 示例:比较两台容器填充机器 Example: Comparing two container-filling machines

方差比检验 Variance Ratio Test

背景：比较两台容器填充机器的总体方差。我们有 $s_2^2 = 0.4528$。回忆一下 $s_1^2 = 0.6333$，且 $n_1 = n_2 = 25$。
Background: To compare the population variances of two container-filling machines. We have $s_2^2 = 0.4528$. Recall that $s_1^2 = 0.6333$ and $n_1 = n_2 = 25$.
问题：我们能否在 5% 的显著性水平上推断第二台机器在一致性方面更优（即方

差更小)?

Question: Can we infer at the 5% significance level that the second machine is superior in its consistency (i.e., has smaller variance)?

方差比检验结果：

Variance ratio test results:

- $F = \frac{s_1^2}{s_2^2} = \frac{0.6333}{0.4528} = 1.3988$
- 自由度：$df_1 = 24, df_2 = 24$
- 备择假设 H_a : ratio < 1（即 $\sigma_1^2 < \sigma_2^2$）的 P 值：0.7915
- 备择假设 H_a : ratio > 1（即 $\sigma_1^2 > \sigma_2^2$）的 P 值：0.2085
- 备择假设 H_a : ratio $\neq 1$（即 $\sigma_1^2 \neq \sigma_2^2$）的 P 值：0.4170

结论：如果我们想检验第二台机器是否更优（方差更小），我们使用左尾检验。P 值 $= 0.7915$ 远大于 0.05，因此没有足够证据支持第二台机器方差更小的声称。

Conclusion: If we want to test whether the second machine is superior (smaller variance), we use the left-tailed test. P-value $= 0.7915$ is much larger than 0.05, so there is not enough evidence to support the claim that the second machine has smaller variance.

3.5 方差比较的注意事项 Cautions for variance comparisons

- F 检验对非正态性极其敏感，即使在大样本中也是如此。
  The F test is extremely sensitive to non-normality, even in large samples.
- 在实践中，很难判断一个显著的 F 值是不等总体方差的证据，还是仅仅是总体不服从正态分布的证据。
  It is difficult in practice to tell whether a significant F-value is evidence of unequal population variances or simply evidence that the populations are not Normal.
- 当你拥有大样本时，检查数据是否服从正态分布是有帮助的，所以大样本在这方面有帮助。
  When you have large samples, it helps to check if the data is normally distributed so large samples help in this sense.
- 在应用工作中，方差比较通常是描述性的，而不是推断性的。
  In applied work, variance comparisons are often descriptive rather than inferential.

4 非正态分布的推断 Inference for Non-Normal Distributions

- 如果数据偏斜，你可以尝试转换变量使其更接近正态性（例如，对数变换）。
  If the data are skewed, you can attempt to transform the variable to bring it closer to Normality (e.g., logarithm transformation).

- 除了正态分布之外的其他分布可能很好地描述你的数据。许多非正态模型已经被开发出来以提供推断方法。
  A distribution other than a Normal distribution might describe your data well. Many non-Normal models have been developed to provide inference procedures too.

- 你总是可以使用无分布（“非参数”）推断方法，它不假设总体的任何特定分布。
  You can always use a distribution-free (“nonparametric”) inference procedure that does not assume any specific distribution for the population.

- 然而，这样的方法通常比分布驱动的检验（例如 t 检验）功效更低。
  However, such a procedure is usually less powerful than distribution-driven tests (e.g., t-test).

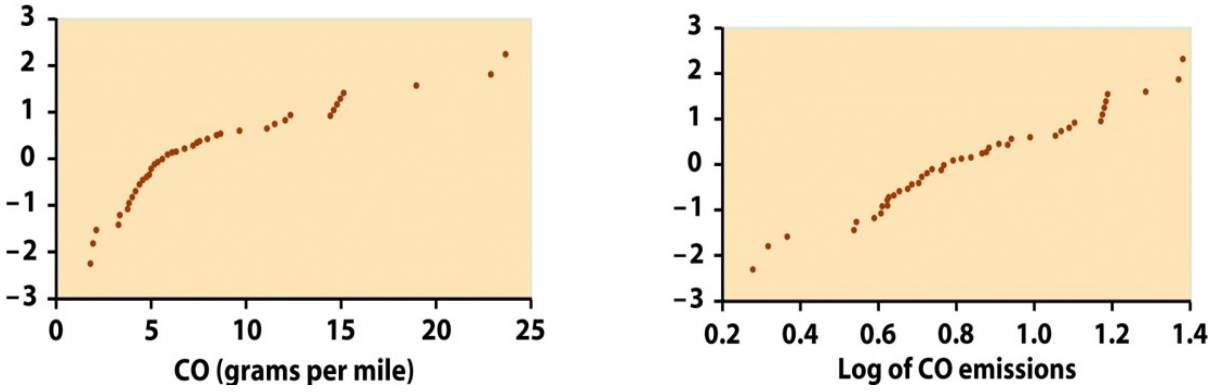
4.1 数据转换 Transforming Data

- 最常见的转换是对数（log）
  The most common transformation is the logarithm (log).

- 适用于右偏分布：拉回分布的右尾。
  Good for right-skewed distributions: pull in the right tail of a distribution.

- 数据值必须全部为正。
  The data values must all be positive.

- **注意：**转换数据会改变参数的**解释：**置信区间不在原始尺度上。
  Caution: Transforming data changes the interpretation of the parameter: confidence interval not in the original scale.



(a) (b)

图 2: 对数数据转换可视化对比

总结 Summary

- **两总体推断概述:**
  - 比较两个总体均值差异有两种主要设计：独立样本和配对样本。
    There are two main designs for comparing differences in two population means: independent samples and paired samples.
  - 正确识别数据类型对于选择适当的检验至关重要。
    Correctly identifying the data type is crucial for choosing the appropriate test.
- **独立样本均值的比较:**
  - 参数：$\mu_1 - \mu_2$。
    Parameter: $\mu_1 - \mu_2$.
  - 当总体方差未知时，使用两样本 t 检验。
    When population variances are unknown, use two-sample t-test.
  - 检验统计量：$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$。
  - 自由度近似（Satterthwaite 近似）或保守估计（较小样本量减 1）。
    Degrees of freedom approximated (Satterthwaite) or conservatively (smaller sample size minus 1).
  - 置信区间：$(\bar{x}_1 - \bar{x}_2) \pm t^*_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$。
- **配对样本均值的比较:**
  - 参数：$\mu_D = \mu_1 - \mu_2$。
    Parameter: $\mu_D = \mu_1 - \mu_2$.
  - 计算每对数据的差异，然后进行单样本 t 检验。
    Compute differences for each pair, then perform one-sample t-test on the differences.
  - 检验统计量：$t = \frac{\bar{x}_D - \mu_D}{s_D/\sqrt{n}}$，自由度 $df = n - 1$。
  - 置信区间：$\bar{x}_D \pm t_{\alpha/2}\frac{s_D}{\sqrt{n}}$。
  - 配对设计通过控制混杂变量减少变异性，提高检验功效。
    Paired designs reduce variability by controlling for confounding variables, increasing test power.
- **合并与不等方差 t 检验:**
  - 合并 t 检验假设方差相等，使用合并方差 s_p^2。
    Pooled t-test assumes equal variances, uses pooled variance s_p^2.

  - 不等方差 t 检验（Welch 检验）不假设方差相等，更通用，通常更受推荐。
    Unequal variance t-test (Welch's test) does not assume equal variances, more general, usually preferred.
  - 除非有强有力的先验理由相信方差相等，否则使用不等方差检验。
    Use unequal variance test unless there is strong prior reason to believe variances are equal.

- **两个方差的比较:**
  - 使用 F 检验比较两个正态总体的方差。
    Use F-test to compare variances of two normal populations.
  - 检验统计量：$F = \frac{s_1^2}{s_2^2}$，服从 $F(df_1, df_2)$ 分布。
  - F 检验对非正态性非常敏感，需谨慎使用。
    F-test is very sensitive to non-normality, use with caution.
  - 在应用工作中，方差比较通常是描述性的。
    In applied work, variance comparisons are often descriptive.

- **检验选择指南:**

数据类型	感兴趣的参数	推荐检验
独立样本，方差未知，不假设等方差	$\mu_1 - \mu_2$	不等方差两样本 t 检验（Welch）
独立样本，方差未知，假设等方差	$\mu_1 - \mu_2$	合并两样本 t 检验 t-test
配对样本	μ_D	配对 t 检验
比较两个正态总体的方差	σ_1^2/σ_2^2	F 检验

表 3: 两总体检验选择指南

- **稳健性与条件:**
  - 两样本 t 检验比单样本 t 检验更稳健。
    Two-sample t-tests are more robust than one-sample t-tests.
  - 当样本量相等且分布相似时最稳健。
    Most robust when sample sizes are equal and distributions similar.
  - 随机抽样条件比正态性条件更重要（大样本时）。
    Random sampling condition is more important than normality condition (for large samples).
  - 小样本时需谨慎：功效低，误差边际大。
    For small samples: low power, large margins of error.

- **非正态数据的处理:**
  - 考虑数据转换（如对数变换）使数据更接近正态。
    Consider data transformation (e.g., log) to make data more normal.
  - 考虑非参数方法（如 Mann-Whitney U 检验）。
    Consider nonparametric methods (e.g., Mann-Whitney U test).
  - 注意转换会改变参数的解释。
    Note: transformation changes interpretation of parameters.

- **关键公式 Key Formulas:**
  - 两样本 t 统计量（不等方差）：$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$
  - 两样本置信区间：$(\bar{x}_1 - \bar{x}_2) \pm t^*_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
  - 合并方差：$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$
  - 合并 t 统计量：$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$
  - 配对 t 统计量：$t = \frac{\bar{x}_D - \mu_D}{s_D/\sqrt{n}}$
  - F 检验统计量：$F = \frac{s_1^2}{s_2^2}$

本章核心要点 Core Takeaways

- **正确识别数据类型：**区分独立样本与配对样本至关重要，因为检验方法不同。
  Correctly identify data type: Distinguishing between independent and paired samples is crucial as test methods differ.
- **选择适当的检验：**
  - 独立样本 → 两样本 t 检验
    Independent samples → Two-sample t-test
  - 配对样本 → 配对 t 检验
    Paired samples → Paired t-test
  - 比较方差 → F 检验（谨慎使用）
    Compare variances → F-test (use with caution)
- **理解检验假设：**
  - 随机抽样是最重要的条件。
    Random sampling is the most important condition.

  - 正态性假设在大样本时不太关键（中心极限定理）。
    Normality assumption is less critical for large samples (CLT).
  - 方差齐性假设影响检验选择。
    Homogeneity of variance assumption affects test choice.

- **解释结果：**
  - 置信区间提供差异的大小和不确定性。
    Confidence intervals provide magnitude and uncertainty of difference.
  - P 值提供反对原假设的证据强度。
    P-values provide strength of evidence against null hypothesis.
  - 统计显著不等于实际重要。
    Statistical significance does not equal practical importance.

- **考虑稳健性和功效：**
  - 两样本 t 检验相对稳健。
    Two-sample t-tests are relatively robust.
  - 配对设计通常提高功效。
    Paired designs usually increase power.
  - 小样本时功效低，需谨慎解释” 不显著” 结果。
    Low power for small samples, interpret ”non-significant” results cautiously.

- **应用建议：**
  - 如果可能，使用配对设计控制混杂因素。
    If possible, use paired design to control confounders.
  - 如果可能，使用相等样本量提高稳健性。
    If possible, use equal sample sizes to increase robustness.
  - 除非有充分理由，否则使用不等方差 t 检验。
    Use unequal variance t-test unless there is good reason not to.
  - 谨慎使用方差比较检验（F 检验对非正态性敏感）。
    Use variance comparison tests cautiously (F-test sensitive to non-normality).

- **实际问题解决步骤：**

1. 确定研究问题和数据类型。
   Identify research question and data type.
2. 检查假设条件（随机性、正态性、独立性、方差齐性）。
   Check assumptions (randomness, normality, independence, homogeneity of variance).
3. 选择适当的检验。
   Select appropriate test.
4. 进行计算或使用统计软件。
   Perform calculations or use statistical software.
5. 解释结果（置信区间、P 值、效应大小）。
   Interpret results (confidence intervals, P-values, effect sizes).
6. 得出结论并考虑实际意义。
   Draw conclusions and consider practical significance.