Stat C131A: Statistical Methods for Data Science

Lecture11: Hypothesis Testing IV

Oct 2 025

Today

- Revise degrees of freedom in t-test
- Comparing across GROUPS
- Introduce **Confidence Intervals**

t-stat

- The simpler case: **pooled** populations
  - $$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{s^2(\frac{1}{n_1} + \frac{1}{n_2})}}$$
- More nuanced: **Welch's formula**
  - $$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1{}^2}{n_1} + \frac{s_2{}^2}{n_2}}}$$

$\overline{X}_1$ = sample mean of group 1

$\overline{X}_2$ = sample mean of group 2

s_1^2 = sample variance of group 1

s_2^2 = sample variance of group 2

n_1 = sample size of group 1

n_2 = sample size of group 2

s^2 = pooled variance estimate (pooled test only)

df = degrees of freedom

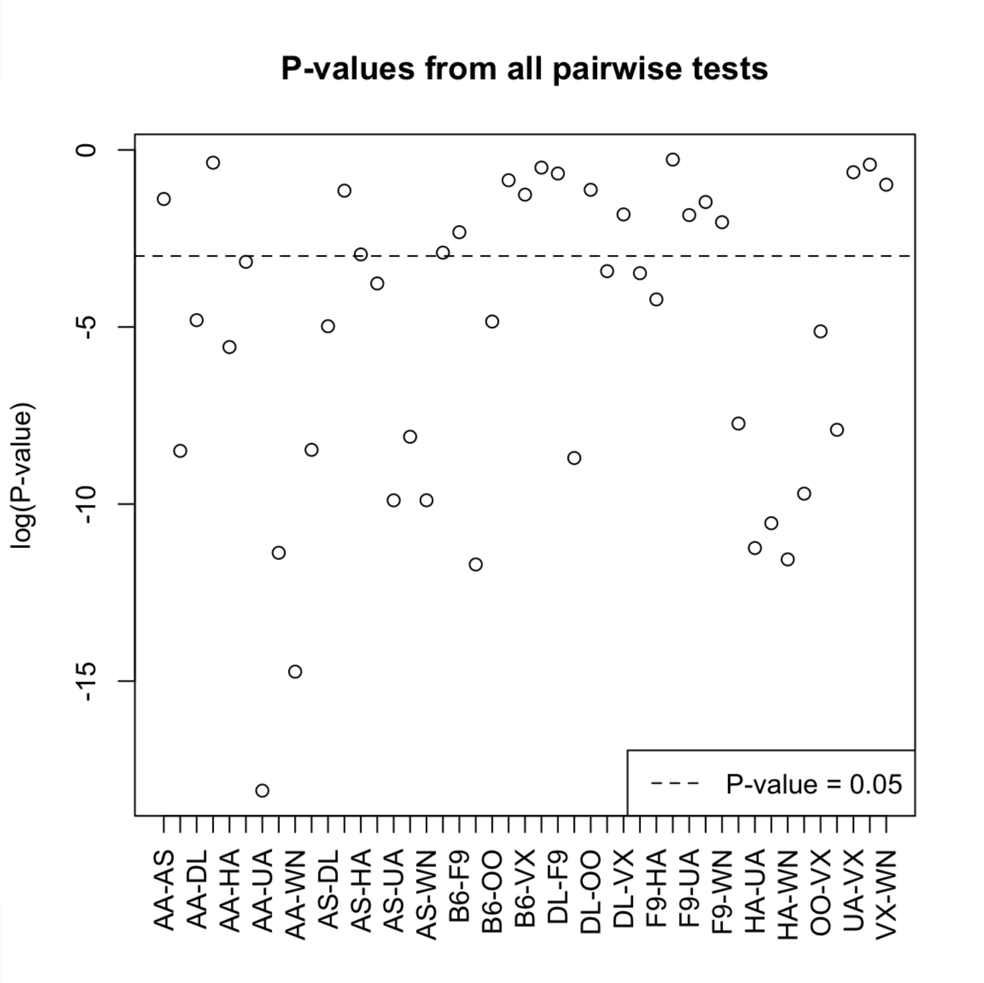
Degrees of freedom

- **Pooled two sample t-test**
  - Assumes both groups have equal population variance.
  - Uses a pooled variance estimate that combines both samples.
  - Degrees of freedom = $n_1 + n_2 - 2$
  - More powerful if assumption is true
    - But can be misleading if variances differ significantly
- **Welch's two sample t-test**
  - Makes no assumption of equal variance.
  - Standard error computed separately from each sample's variance.
  - Degrees of freedom adjusted using **Welch–Satterthwaite formula**
    - Each group's variance estimate contributes uncertainty.
    - If one group is very small or very noisy, its contribution reduces the effective df.
    - The formula is a weighted average of the sample variances' variances.
    - The df becomes fractional
  - More robust in practice, especially with unequal n's or variances.
    - Slightly less power if variances truly equal.

Comparing Groups

Flight delays across airlines

- Suppose we are analyzing **average flight delays across multiple airlines**.
- For each airline pair, we test:
  - H_0: Airline A and Airline B have the same mean delay.
  - H_1: Airline A and Airline B have different mean delays.
- With 6 airlines, that's C(6,2) = **15 pairwise comparisons**.



Type I error Recap

- A **Type I Error** occurs when we:
  - Reject the null hypothesis (claim a difference),
  - But in reality, the null is true (there is no difference).
- Probability of a Type I Error in a single test = **α** (e.g., 0.05).
- Example:
  - We say *Airline A is slower than Airline B,*
  - But in fact, their mean delays are the same.

The Multiple Testing Problem

- Running many tests **inflates the risk of at least one false positive**.
  - Analogy: flipping a coin once → 50% chance of heads. Flip it 15 times → chance of ≥1 head is much higher.
- **Familywise Error Rate (FWER)** = probability of at least one Type I Error across the whole "family" of tests.
  - $FWER = 1 - (1 - \alpha)^m$
- Example:
  - m = 15 tests, a = 0.05
  - $FWER = 1 - \left(1 - 0.05\right)^{15} \approx 54\%$
  - Meaning: even if all airlines are identical, we will often see at least one "significant" difference just by chance

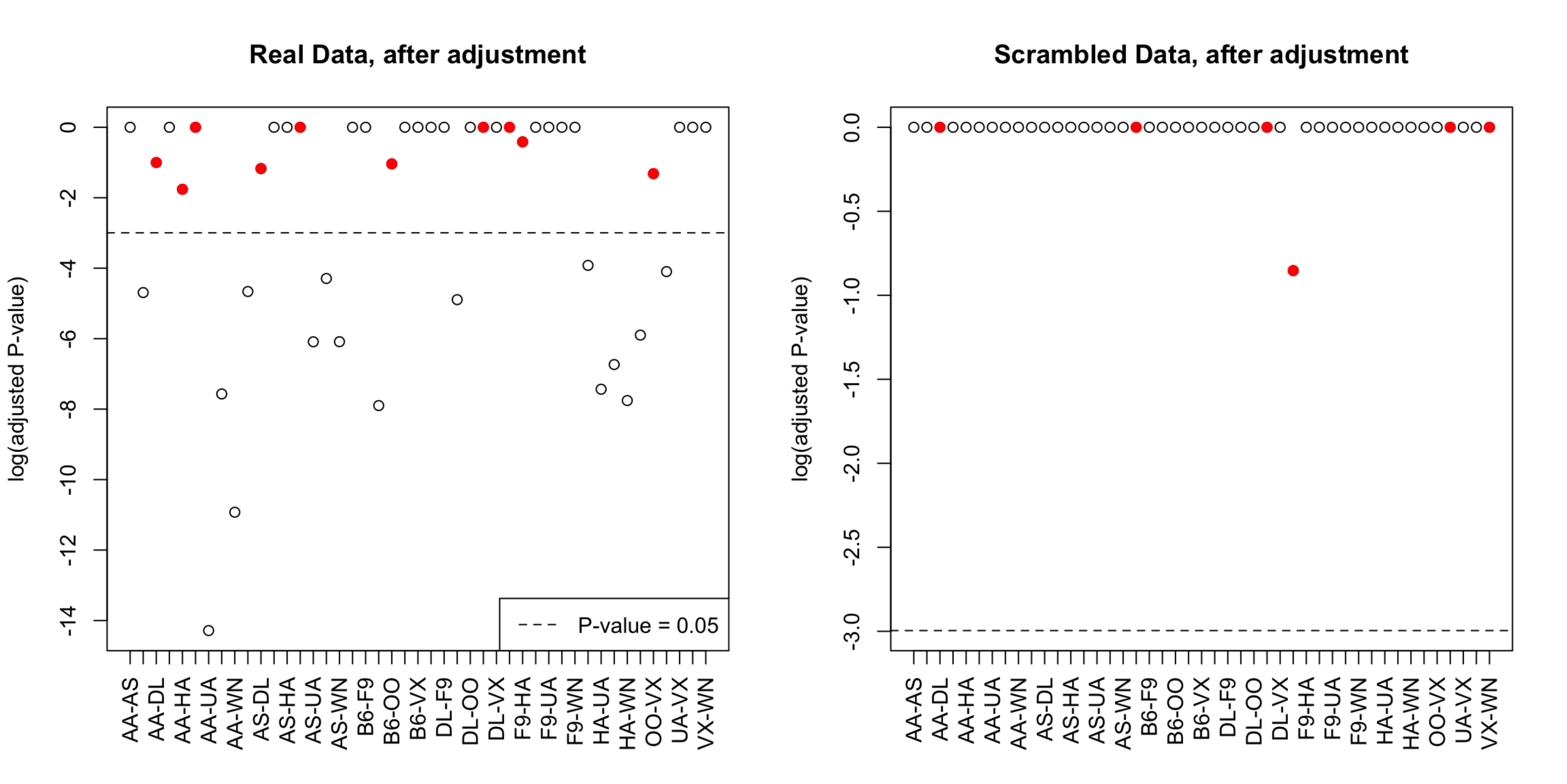
Error Inflation

- ★ If we only had **1 test**, probability of a false alarm = 5%.
- ★ If we have **15 tests**, probability of ≥1 false alarm ≈ 54%.

Control the inflation of error

- **Goal:** keep the chance of at least one false positive ≤ a (e.g., 0.05).
- One simple method: **Bonferroni Correction**
  - If you are running m tests, make each test stricter by dividing a.
  - *a' = a / m*
- Example:
  - a = 0.05, m = 15 → a' ≈ 0.0033
- Now, each test needs $p < 0.0033$ to be considered significant

Bonferroni corrected



Effect

- **Controls the overall error rate**: ensures the probability of any false positive is ≤ α.
- **Very conservative**:
  - Reduces Type I Errors,
- But increases Type II Errors (missed true differences)
  - WHY ?
- Works best when:
  - Number of tests is not huge,
  - You really care about avoiding false positives.

CONFIDENCE INTERVALS

Confidence Intervals

- A **confidence interval (CI)** gives us a *range of plausible values* for a parameter.
- Instead of just reporting a point estimate, we say:
  "We're 95% confident the true value lies between X and Y."
- Built from:

  Point estimate (like a mean or proportion).

  Standard error (SE), which measures variability.

  A multiplier (Z* or t*) that adjusts for desired confidence level.

- $P(V_1 \leq \theta \leq V_2) = 0.95.$

Confidence, Not Certainty

- A 95% CI does **not** mean “95% of the population lies here.”
- Instead:

  If we repeated the sampling process many times,

  Then 95% of the intervals we construct will contain the true parameter.
- Confidence is about the **procedure**, not the single dataset.

Quantiles

- Confidence intervals are based on **quantiles** of the sampling distribution.
- Example: a 95% CI for a mean uses the **2.5% and 97.5% quantiles** of the t-distribution
- Quantiles tell us "cut-off values" where a certain % of the distribution lies below them.
  - $(q(\alpha/2)\ ,\ q(1-\alpha/2))$
  - For standard normal distribution
    - $q(0.025) \approx -1.96$
    - $q(0.975) \approx +1.96$
- Where did 1.96 come from ?

For mean of one group

- We want to estimate the **population mean** μ of a single group.
- From data, we have the **sample mean** $\bar{X}$.
- But $\bar{X}$ is just one sample; it will vary.
- So we use a **confidence interval** to give a plausible range for μ.

CI of One Mean

- CI:
  - $\bar{X} \pm t^* \times SE$
    - Where:
      - $\bar{X}$ = sample mean
      - $SE = s / \sqrt{n}$ (sample standard deviation over square root of sample size)
      - t^* = cutoff from t-distribution (depends on $df = n-1$, confidence level)
- $\overline{X} \pm t_{\frac{\alpha}{2},\mathrm{df}} \times \frac{s}{\sqrt{n}}$
-

Interpretation

- If CI includes the hypothesized value (like 0 or a benchmark), the result is not significant
  - WHY ?
- If CI is entirely above or below, we conclude that the mean is likely different.
  - WHY ?
- Example:
  - $\bar{X}$ = 21 minutes, n = 36, s = 3.9.
  - 95% CI = (19.9, 22.1).
  - Interpretation: We're 95% confident the *true mean delay* is between 19.9 and 22.1 minutes

t versus Z

- If population σ is known → use Z.
- If σ is unknown (the usual case) → estimate with sample s → use t
  - t accounts for added uncertainty from estimating σ.
- For large n, $t \approx Z$. For small n, t is wider.

Confidence Interval

- We have a random sample of size *n* from a population
- We want to estimate the **population mean (μ)**
- From the data, we compute:
  - Sample mean: $\bar{X}$
  - Sample standard deviation: s

Confidence Interval

- By the Central Limit Theorem (CLT), the sample mean $\bar{X}$ is approximately normally distributed if *n* is large enough.
- If the population is normal, then $\bar{X}$ is exactly normal, even for small *n*.
- Distribution of $\bar{X}$:
  - Mean = μ
  - Standard deviation = $\sigma / \sqrt{n}$
- But usually, we do **not know σ**

Confidence Interval

- In practice, we do **not** know the population standard deviation (σ).
- We estimate it using the **sample standard deviation (s)**.
- Substituting s for σ changes the distribution:
- It's no longer exactly normal,
- It follows a **t-distribution with df = n − 1**.
- $t = (\bar{X} - \mu) / (s/\sqrt{n})$

Confidence Interval

- We know that most of the time, the standardized statistic lies between $-t^*$ and $+t^*$
- $P(-t^* \leq (\bar{X} - \mu)/(s/\sqrt{n}) \leq t^*) = 0.95$
- Here:
  - t^* = critical value from t distribution,
  - Depends on confidence level (e.g., 95%) and df = $n - 1$.

Confidence Interval

$$-t \le \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}} \le t$$

$$-t \cdot \frac{s}{\sqrt{n}} \le \overline{X} - \mu \le t \cdot \frac{s}{\sqrt{n}}$$

$$\overline{X} - t \cdot \frac{s}{\sqrt{n}} \le \mu \le \overline{X} + t \cdot \frac{s}{\sqrt{n}}$$

R example