Applied Hypothesis Testing

Advanced Module A0

Digital Finance

Course Overview

Topics Covered

- Hypothesis testing framework
- Type I and Type II errors
- p-values and significance
- One-sample t-test
- Two-sample t-test
- Paired t-test

Advanced Topics

- One-way ANOVA
- Post-hoc tests
- Power analysis
- Effect sizes
- Multiple comparisons
- Practical examples in R

Focus on practical application and interpretation

Hypothesis Testing Framework

Key Components

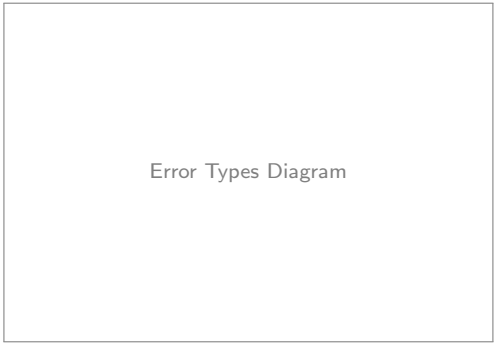
- Null hypothesis (H_0)
- Alternative hypothesis (H_1 or H_a)
- Test statistic
- Significance level (α)
- p-value
- Decision rule

The foundation of statistical inference

Type I and Type II Errors

Two ways to be wrong:

- **Type I Error** (α) = **False Alarm**
  - Reject true H_0
  - Like convicting an innocent person
  - Controlled by significance level
- **Type II Error** (β) = **Missed Detection**
  - Fail to reject false H_0
  - Like letting a guilty person go free
  - Related to statistical power



Balance between error types is crucial for study design

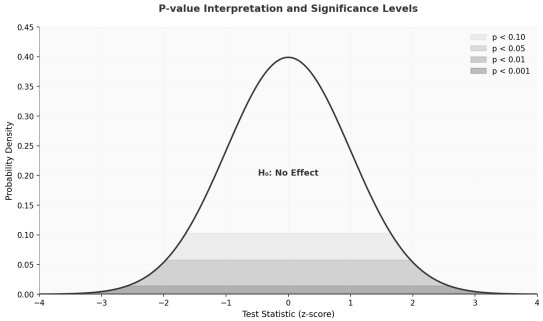
Understanding p-values

p-value Definition

- Probability of observing data at least as extreme as the actual observed results
- Calculated under the assumption that H_0 is true
- **NOT** the probability that H_0 is true

Interpretation

- $p < 0.05$: "Significant"
- $p < 0.01$: "Highly significant"
- $p < 0.001$: "Very highly significant"



p-values measure evidence against H_0, not effect size

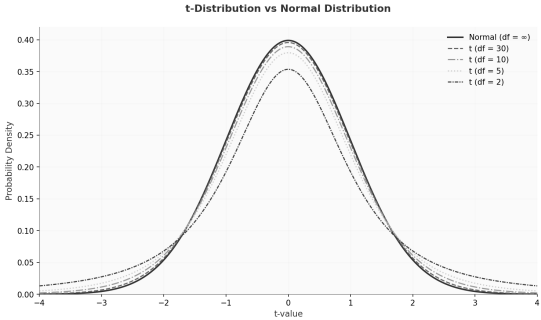
Introduction to t-tests

When to Use t-tests

- Comparing means
- Small sample sizes
- Unknown population variance
- Approximately normal data

The t-distribution

- Similar to normal distribution
- Heavier tails
- Approaches normal as df increases



William Sealy Gosset ("Student"), 1908

One-sample t-test: Theory

Hypotheses

- H_0: $\mu = \mu_0$
- H_1: $\mu \neq \mu_0$ (two-tailed)
- H_1: $\mu > \mu_0$ (one-tailed)
- H_1: $\mu < \mu_0$ (one-tailed)

Test Statistic ("How surprising is our data?")

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{\text{observed} - \text{expected}}{\text{typical variation}}$$

where:

- $\bar{x}$ = sample mean
- μ_0 = hypothesized mean
- s = sample standard deviation
- n = sample size

Foundation for comparing a sample to a known value

Assumptions

- Random sampling
- Independent observations
- Approximately normal distribution
- Or large sample size (n $>$ 30)

Degrees of Freedom

$$df = n - 1$$

Decision Rule

- Reject H_0 if $|t| > t_{critical}$
- Or if p-value $< \alpha$

One-sample t-test: R Example

Example: Testing Mean Height

```r
# Sample data
heights <- c(172, 168, 175,
             170, 165, 178,
             169, 171, 173,
             167)

# Population mean = 170 cm
# One-sample t-test
t.test(heights,
       mu = 170,
       alternative = "two.sided")

# Output:
# t = 0.949, df = 9
# p-value = 0.367
# 95% CI: [168.2, 173.4]
# Sample mean: 170.8
```

Interpretation

- Sample mean: 170.8 cm
- Test statistic: t = 0.949
- p-value = 0.367 > 0.05
- Cannot reject H_0
- No evidence that mean differs from 170 cm

Confidence Interval

- 95% CI: [168.2, 173.4]
- Contains $\mu_0 = 170$
- Consistent with hypothesis test

Always check assumptions before interpreting results

Two-sample t-test: Theory

Purpose

- Compare means of two independent groups
- Test for differences between populations

Hypotheses

- H_0: $\mu_1 = \mu_2$
- H_1: $\mu_1 \neq \mu_2$

Test Statistic (equal variances)

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

Pooled Standard Deviation

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

Assumptions

- Independent samples
- Normal distributions
- Equal variances (or use Welch's test)

Degrees of Freedom

- Equal variance: $df = n_1 + n_2 - 2$
- Welch's test: complex formula

Most common test for comparing two groups

Two-sample t-test: R Example

```r
# Treatment groups
control <- c(23, 25, 21,
             22, 26, 20)
treatment <- c(28, 30, 32,
               29, 31, 27)

# Test for equality of variances
var.test(treatment, control)
# F = 1.08, p = 0.94

# Two-sample t-test
t.test(treatment, control,
       var.equal = TRUE)

# Output:
# t = 5.99, df = 10
# p-value = 0.00013
# Mean difference: 7.17
```

Results Interpretation

- Control mean: 22.8
- Treatment mean: 29.5
- Difference: 6.7 units
- p-value < 0.001
- Strong evidence against H_0
- Treatment effect is significant

Effect Size

```r
# Cohen's d
d <- (mean(treatment) -
      mean(control)) / s_p
# d = 2.45 (large effect)
```

Always report effect size alongside p-values

Paired t-test: Theory

When to Use

- Before-after measurements
- Matched pairs
- Repeated measures
- Cross-over designs

Approach

- Calculate differences: $d_i = x_{1i} - x_{2i}$
- Test if mean difference $= 0$
- Reduces to one-sample t-test

Test Statistic

$$t = \frac{\bar{d} - 0}{s_d / \sqrt{n}}$$

where:

- $\bar{d}$ = mean of differences
- s_d = standard deviation of differences
- n = number of pairs

Advantages

- Controls for individual differences
- More powerful than independent samples
- Smaller sample size needed

Paired design removes between-subject variability

Paired t-test: R Example

```r
# Weight loss study
before <- c(82, 78, 85,
            90, 76, 88)
after <- c(78, 76, 80,
           84, 73, 82)

# Paired t-test
t.test(before, after,
       paired = TRUE)

# Output:
# t = 5.77, df = 5
# p-value = 0.0022
# Mean difference: 5.17 kg

# Effect size
diff <- before - after
cohen_d <- mean(diff) /
           sd(diff)
# d = 2.36 (large)
```

Data Structure

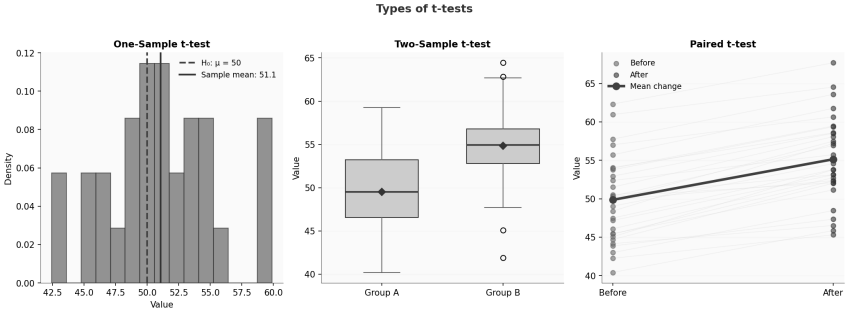
Subject	Before	After
1	82	78
2	78	76
3	85	80
4	90	84
5	76	73
6	88	82

Interpretation

- Average weight loss: 5.17 kg
- p-value = 0.002
- Significant weight reduction
- Large effect size (d = 2.36)

Paired design is ideal for before-after comparisons

Types of t-tests: Summary



Choose the appropriate test based on

study design

Analysis of Variance (ANOVA): Concept

Purpose

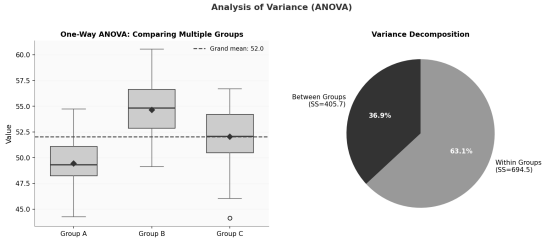
- Compare means of 3+ groups
- Extension of two-sample t-test
- Test overall difference

Hypotheses

- H_0: $\mu_1 = \mu_2 = ... = \mu_k$
- H_1: At least one mean differs

Key Idea ("Signal vs Noise")

- **Signal:** Between-group variance (are groups different?)
- **Noise:** Within-group variance (random variation)
- **F-ratio:** Signal / Noise
- Large F = groups really differ



ANOVA tests for any difference among groups

ANOVA: Variance Decomposition

Sum of Squares

- **SST** (Total): Total variability
- **SSB** (Between): Group differences
- **SSW** (Within): Random error
- SST = SSB + SSW

Mean Squares

- MSB = SSB / (k-1)
- MSW = SSW / (N-k)

F-statistic

$$F = \frac{MSB}{MSW} = \frac{\text{Between-group variance}}{\text{Within-group variance}}$$

Degrees of Freedom

- $df_{between}$ = k - 1
- df_{within} = N - k
- df_{total} = N - 1

Decision

- Large F $\rightarrow$ reject H_0
- Compare to F-distribution

F-ratio measures signal-to-noise ratio

One-Way ANOVA: R Example

```r
# Three teaching methods
method_A <- c(78, 82, 75,
              80, 77)
method_B <- c(85, 88, 83,
              87, 84)
method_C <- c(72, 74, 70,
              73, 71)

# Combine data
scores <- c(method_A,
            method_B,
            method_C)
method <- factor(rep(
  c("A", "B", "C"),
  each = 5))

# ANOVA
model <- aov(scores ~ method)
summary(model)
```

ANOVA Table

Source	df	SS	MS	F
Method	2	338.5	169.3	31.7
Error	12	64.0	5.3	

p-value < 0.001

Interpretation

- $F(2,12) = 31.7$
- $p < 0.001$
- Strong evidence of differences
- Need post-hoc tests

Group Means

- Method A: 78.4
- Method B: 85.4
- Method C: 72.0

ANOVA tells us differences exist, not which groups differ

ANOVA Assumptions

Key Assumptions

1. Independence of observations
2. Normality within groups
3. Homogeneity of variances

Checking Assumptions

- **Normality**: Q-Q plots, Shapiro-Wilk test
- **Equal variances**: Levene's test, Bartlett's test
- **Independence**: Study design

Robustness

- ANOVA fairly robust to normality violations
- Sensitive to unequal variances
- Very sensitive to dependence

Alternatives

- Welch's ANOVA (unequal variances)
- Kruskal-Wallis test (non-parametric)
- Transformation of data

Always check assumptions before interpreting ANOVA

Multiple Comparisons Problem

The Problem

- ANOVA: overall difference
- Which groups differ?
- Multiple pairwise tests
- Inflated Type I error

Example: 4 groups

- $\binom{4}{2} = 6$ comparisons
- If $\alpha = 0.05$ per test
- Family-wise error rate:

$$1 - (1 - 0.05)^6 = 0.265$$

- 26.5% chance of false positive!

Solutions

- **Bonferroni**: $\alpha_{adj} = \alpha/m$
- **Tukey HSD**: All pairwise
- **Dunnett**: Compare to control
- **Scheffé**: All contrasts

Trade-offs

- Control Type I error
- Reduced power
- Conservative corrections

Balance between Type I error control and statistical power

Post-hoc Tests: R Example

```r
# Tukey HSD test
TukeyHSD(model)

# Output (abbreviated):
#         diff    p adj
# B-A     7.0    0.0003
# C-A    -6.4    0.0008
# C-B   -13.4    0.0001

# Bonferroni correction
pairwise.t.test(scores,
                method,
                p.adj = "bonf")

# Dunnett test (A as control)
library(multcomp)
dunnett <- glht(model,
  linfct = mcp(
    method = "Dunnett"))
summary(dunnett)
```

Choose post-hoc test based on research questions

Tukey HSD Results

Comparison	Difference	95% CI	p-adj
B - A	7.0	[3.8, 10.2]	0.0003
C - A	-6.4	[-9.6, -3.2]	0.0008
C - B	-13.4	[-16.6, -10.2]	0.0001

Interpretation

- All pairwise differences significant
- Method B $>$ Method A $>$ Method C
- Adjusted for multiple comparisons
- Family-wise error controlled at 0.05

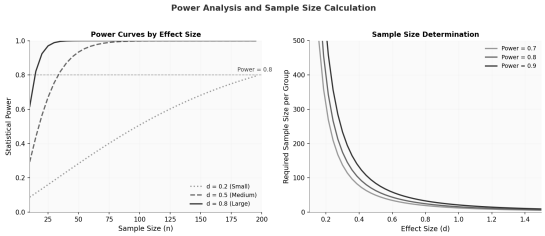
Statistical Power

Definition ("Your ability to find what's really there")

- Power = 1 - β = 1 - P(missed detection)
- Probability of detecting a real effect
- Want power $\geq$ 0.80 (80% chance to detect)

Factors Affecting Power

- Effect size (larger $\rightarrow$ more power)
- Sample size (larger $\rightarrow$ more power)
- Significance level (α)
- Variability (less $\rightarrow$ more power)



Power analysis essential for study design

Power Analysis in R

```r
library(pwr)

# Two-sample t-test power
# Effect size d = 0.8
# n per group = 30
pwr.t.test(n = 30,
           d = 0.8,
           sig.level = 0.05,
           type = "two.sample")
# Power = 0.86

# Sample size for power=0.8
pwr.t.test(d = 0.5,
           power = 0.8,
           sig.level = 0.05,
           type = "two.sample")
# n = 64 per group
```

ANOVA Power

```r
# 3 groups, effect size f=0.4
pwr.anova.test(k = 3,
               f = 0.4,
               sig.level = 0.05,
               power = 0.8)
# n = 21 per group

# Detectable effect
pwr.anova.test(k = 3,
               n = 15,
               sig.level = 0.05,
               power = 0.8)
# f = 0.47
```

Planning Studies

- Estimate effect size from pilot/literature
- Calculate required sample size
- Consider practical constraints

Always conduct power analysis before data collection

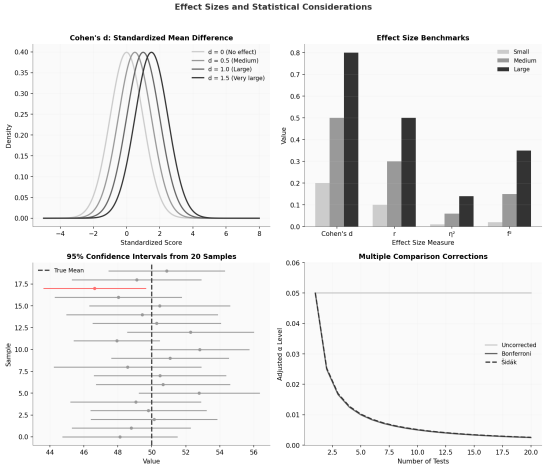
Effect Sizes

Why Effect Sizes? ("How big is the effect?")

- p-value = "Is it real?" (yes/no)
- Effect size = "How big?" (magnitude)
- With large n, tiny effects become significant
- Effect sizes are comparable across studies

Common Measures

- **Cohen's d**: Mean difference in SD units (small=0.2, medium=0.5, large=0.8)
- **r**: Correlation coefficient
- η^2: Variance explained (ANOVA)



Always report effect sizes with hypothesis tests

Complete Example: Clinical Trial

Study Design

- Compare 3 treatments
- Primary outcome: Pain score
- n = 20 per group
- $\alpha = 0.05$

```r
# Load data
data <- read.csv("trial.csv")

# Check assumptions
bartlett.test(pain ~ treatment,
              data = data)
# p = 0.34 (equal var OK)

# ANOVA
model <- aov(pain ~ treatment,
             data = data)
summary(model)
# F(2,57) = 8.42, p = 0.0006
```

```r
# Effect size
library(effectsize)
eta_squared(model)
# eta2 = 0.228 (large)

# Post-hoc tests
TukeyHSD(model)
# Drug A vs Placebo: p=0.001
# Drug B vs Placebo: p=0.023
# Drug A vs Drug B: p=0.142

# Power analysis (post-hoc)
pwr.anova.test(k = 3,
               n = 20,
               f = 0.54,
               sig.level = 0.05)
# Power = 0.84
```

Conclusions

- Both drugs superior to placebo
- No difference between drugs
- Adequate power achieved

Complete analysis includes assumptions, effect size, and power

Best Practices

Before Analysis

- Define hypotheses clearly
- Choose appropriate test
- Set significance level
- Conduct power analysis
- Plan for multiple comparisons

During Analysis

- Check all assumptions
- Use appropriate corrections
- Calculate effect sizes
- Document all decisions

Reporting Results

- State hypotheses tested
- Report test statistics
- Include exact p-values
- Provide effect sizes
- Show confidence intervals
- Interpret in context

Common Pitfalls

- p-hacking
- Ignoring assumptions
- Multiple testing without correction
- Confusing significance with importance

Good statistical practice ensures reproducible research

Summary

Key Concepts

- Hypothesis testing framework
- Type I and Type II errors
- p-values and their interpretation
- Effect sizes complement p-values

Tests Covered

- One-sample t-test
- Two-sample t-test
- Paired t-test
- One-way ANOVA
- Post-hoc comparisons

Important Reminders

- Always check assumptions
- Consider study design
- Plan sample size with power analysis
- Report complete results
- Statistical vs practical significance

Next Topics

- Two-way ANOVA
- Non-parametric tests
- Multiple regression
- Mixed models

Statistical inference is the foundation of empirical research