CHARTA LEARNING · LESSON NOTES

Hypothesis testing

CFA Level I · Quantitative Methods

You will learn: Write a null and an alternative hypothesis, read a p-value and the two kinds of error, test a mean and a correlation, and choose a non-parametric test when it is the better tool.

Running example: Examples use made-up figures, such as a fund with 36 monthly returns averaging 1.5% with a sample standard deviation of 3.6%.

1. The null and the alternative

- A hypothesis test works like a courtroom: innocent until proven guilty. The null hypothesis H_0 is the starting position, held until the evidence says otherwise. The alternative H_a is what you suspect; you conclude it only if the null is rejected.
- The null always contains the equals sign (equal to, at most, or at least), because the test needs one exact value to start from. Whatever you hope to show goes in the alternative, never the null.
- Recipe: suspicion becomes the alternative; its negation, with the equals sign, becomes the null; the tail follows from the alternative.
- One-tailed: you suspect greater than or smaller than. Two-tailed: you suspect simply different. Choose the direction before you see the data.

WRITING THE PAIR

Suspicion: strategy earns positive alpha	H_a: $\alpha > 0$	**One-tailed**
Null	H_0: $\alpha \leq 0$	
Wrong version	your claim in the null: H_0: $\alpha > 0$	**arranges a win by default**
Doubt a brochure's 8% average return	H_a: $\mu \neq 8\%$; H_0: $\mu = 8\%$	**Two-tailed**

Watch out: Putting your own claim in the null.

2. Making the decision

- The significance level α is chosen in advance (often 5%): the chance you accept of rejecting a null that is actually true. The test statistic measures how far the sample is from the null, allowing for noise. If it lands beyond the critical value, reject; otherwise fail to reject.
- The p-value answers: if the null were true, how often would we see a result at least this extreme? If the p-value is below α, reject. It is not the probability that the null is true, and it says nothing about the size or importance of the effect.
- Failing to reject is an acquittal, not proof that the null is true. With a small sample you can nearly always fail to reject: say not shown, rather than shown not to exist.

A TEST RETURNS A P-VALUE OF 0.03

Level α = 5%	0.03 < 0.05	**Reject**
Level α = 1%	0.03 > 0.01	**Fail to reject**
Meaning	data this extreme about 3% of the time if H_0 were true	**not a 3% chance H_0 is true**

Watch out: Reading the p-value as the probability that the null is true, and failure to reject as proof.

3. Type I and Type II errors

- Type I error: rejecting a true null (convicting an innocent person). Its probability is α, which you chose. Type II error: failing to reject a false null (letting a guilty person go). Its probability is β, which you do not choose.
- The power of the test, 1 − β, is the probability of rejecting a false null.
- Tightening α lowers Type I errors but raises Type II errors, and loosening does the opposite. Pick by which mistake costs more. The only real way to cut both is a larger sample.

DECISION AGAINST THE TRUTH YOU CANNOT SEE	
Null true, you reject	**Type I error (α)**
Null false, you fail to reject	**Type II error (β)**
Null false, you reject	**Correct: power**
Null true, you fail to reject	**Correct**

4. Testing a mean

- If the population variance is known and the population is normal, use a z-test. Usually the variance is unknown and estimated by the sample standard deviation s, which calls for a t-test. When in doubt, use t.
- The statistic is a distance measured in standard errors, where the standard error is s ÷ √n. Degrees of freedom are n − 1. Small samples face a higher t critical value; as the sample grows, t converges on z.
- Two independent groups: two-sample t-test. Matched pairs (before and after): subtract within each pair, then run an ordinary t-test on the differences with the null that the average difference is zero.

$t = (\bar{x} - \mu_0) \div (s \div \sqrt{n})$, with $n - 1$ degrees of freedom

36 MONTHLY RETURNS, $\bar{X}$ = 1.5%, S = 3.6%; H_0: M = 0, H_a: M ≠ 0, 5%, 35 DF		
Standard error	3.6 ÷ √36 = 3.6 ÷ 6	**0.60%**
Test statistic	(1.5 − 0) ÷ 0.60	**2.50**
Critical value, 35 df, two-tailed 5%	2.50 > 2.030	**Reject H_0**
If the mean had been 1.1%	t = 1.1 ÷ 0.6	**1.83, fail to reject**
Critical values at 5%, two-tailed	n = 6 / 31 / 121 / z	**2.571 / 2.042 / 1.980 / 1.960**

5. Testing a correlation

- A sample correlation r can be non-zero by pure chance, so test it. The null is that the true correlation is zero; the alternative is that it is not (two-tailed). The statistic follows a t distribution with n − 2 degrees of freedom, as two things are estimated.
- A bigger sample or a stronger correlation gives a bigger t. A weak correlation can be significant in a large sample, while a strong one is not significant in a small sample.
- Significance answers whether a relationship exists, not whether it is large enough to act on. Squaring r gives the share of variation explained.

$t = r \times \sqrt{(n - 2)} \div \sqrt{(1 - r^2)}$, with $n - 2$ degrees of freedom

TWO-TAILED TEST AT 5%		
r = 0.55, 12 pairs: t	0.55 × 3.162 ÷ √(1 − 0.3025) = 1.739 ÷ 0.835	**2.08**
Critical t, 10 df	2.08 < 2.228	**Not significant**
r = 0.10, 500 pairs: t	0.10 × 22.32 ÷ 0.995 = 2.232 ÷ 0.995	**2.24**
Critical t, 498 df	2.24 > 1.965	**Significant**
Variation explained	r = 0.55; r = 0.10	**About 30%; 1%**

Watch out: Using the wrong degrees of freedom (n − 1 for a mean, n − 2 for a correlation), and confusing significant with important.

6. Parametric and non-parametric tests

- A parameter describes a population, like a mean or variance. A parametric test is a question about a parameter and assumes the data have a certain shape, usually near normal. The z-test, t-test, chi-square test (one variance) and F-test (two variances) are parametric.
- Use a non-parametric test when the data are ranks, when there are outliers in a small sample, or when the question is not about a parameter. It makes few assumptions and works on ranks or counts. The price is power when the assumptions do hold.

MATCH THE QUESTION TO THE TEST

Mean of 60 returns above zero?	a parameter, big sample	**t-test**
Two funds, equal variances?	both near normal	**F-test**
Do two rankings of funds match?	the data are ranks	**Spearman rank correlation**
Median of 12 returns, two extreme	outliers dominate	**Wilcoxon signed-rank test**
Beat benchmark in most months?	only direction counts	**Sign test**
Two classifications independent?	counts in a table	**Chi-square test of independence**

Key takeaways

- The null keeps the equals sign; your suspicion goes in the alternative; pick the tail before the data.
- Reject when the p-value is below α or the statistic passes the critical value.
- The p-value is not the probability that the null is true; failing to reject is not proof.
- Type I rejects a true null (probability α); Type II fails to reject a false null (β); power is $1 - \beta$.
- Mean: $t = (\bar{x} - \mu_0) \div (s \div \sqrt{n})$, $n - 1$ df. Correlation: $t = r\sqrt{(n - 2)} \div \sqrt{(1 - r^2)}$, $n - 2$ df.
- Significant is not the same as important; use a non-parametric test for ranks, outliers or counts.

Quick check

1. A manager claims to beat the benchmark. What is the alternative hypothesis?
   Answer: H_a: mean excess return > 0 (null: ≤ 0).
2. A test gives p = 0.04. What is the decision at a 1% level?
   Answer: 0.04 > 0.01, so fail to reject.
3. A correlation is tested with 25 pairs. How many degrees of freedom?
   Answer: $n - 2 = 23$.