CHARTA LEARNING · LESSON NOTES

Hypothesis Testing

FRM Part I

1 A story: skill or luck?

- Average beat the market by 0.8% a month for 36 months
- Skill, or a lucky run? Zero-skill people sometimes do this
- Ask: how surprising is this IF there were no skill?
- Hypothesis testing makes that question exact

IN THE LESSON

- Imagine a fund manager who has beaten the market by eight tenths of a percent a month, on average, over three years.
- Is she skilled, or did she just have a lucky run?
- Even a person with no skill at all, flipping coins to pick stocks, would sometimes show a record like that.
- So the question is never whether the average is above zero.
- The question is how surprising an average this large would be if there were no skill at all.
- If it would be very surprising, we start to believe in skill.
- If it would happen fairly often by chance, we stay doubtful.
- Hypothesis testing turns that everyday reasoning into a clear procedure with numbers.

2 The logic: assume the default, test the surprise

- Presume the boring explanation: no effect, no skill
- Measure how far the sample sits from that world
- Distance in standard errors is the test statistic
- Far enough, by a level fixed in advance: reject

IN THE LESSON

- A hypothesis test works like a trial.
- The accused is presumed innocent, and the jury convicts only if the evidence would be very unlikely for an innocent person.
- In the same way, we assume the boring explanation, that nothing special is going on.
- Then we compute how far our sample result lies from what that boring world predicts, measured in standard errors.
- If the result is far out in the tail, the boring explanation is a poor fit, and we reject it.
- If it is close, we keep it, but we have not proved it.

3 Null and alternative hypotheses

- Claims about the population parameter, not the sample
- H0 (null): the default, always holds the =, ≤ or ≥
- H1 (alternative): the strict inequality or ≠
- The claim you want evidence for goes in H1
- Fix both before looking at the data

IN THE LESSON

- Every test starts with two claims about a population parameter, such as the true mean return, and never about the sample figure itself.
- The null hypothesis is the default position, the one the test tries to reject.
- It always carries the equality: the mean equals zero, or is at most zero, or is at least zero.
- The alternative hypothesis is its complement and holds the strict inequality or the not equal sign.
- Put the claim you want evidence for into the alternative, because only a rejection counts as evidence for it.
- A fund manager with skill means a mean above zero, so that claim goes in the alternative, and the null says the mean is zero or lower.

4 Choosing the hypotheses

The claim to be proved sits in H1; the equality sits in H0

Situation	Hypotheses
Fund beats its benchmark	H0: μ ≤ 0 vs H1: μ > 0
Desk profit differs from zero	H0: μ = 0 vs H1: μ ≠ 0
Loss rate below the limit	H0: μ ≥ L vs H1: μ < L

IN THE LESSON

- Here are three situations.
- First, we suspect a fund beats its benchmark.
- The claim to prove goes in the alternative: the mean excess return is above zero, and the null is that it is zero or below.
- Second, a risk manager asks whether a trading desk's average daily profit differs from zero in either direction.
- The alternative uses the not equal sign, and the null says the mean is exactly zero.
- Third, a bank wants to show that its average loss rate is below a regulatory limit.
- The claim, below the limit, is the alternative, and the null says the loss rate is at or above the limit.
- In every case, the null holds the equality and the claim to be proved sits in the alternative.

5 What a test result really says

- Reject H0: evidence for the alternative
- Fail to reject: the data were not strong enough
- That never proves the null true
- Fix hypotheses and level before seeing the data

IN THE LESSON

- Notice the careful language.
- A test ends in one of two ways.
- We reject the null, which counts as evidence for the alternative.
- Or we fail to reject the null, which only means the data were not strong enough at the chosen significance level.
- It never proves the null true.
- A jury that says not guilty has not shown the accused is innocent.
- This matters for the fund manager.
- If her record does not clear the bar, we have not shown she has no skill.
- We may just have too little data.
- And because the data can tempt us, the hypotheses and the significance level must be fixed before we look at the numbers, never adjusted afterwards to get a result we like.

6 The test statistic

z = (estimate − null value) ÷ SE
SE = s ÷ √n

Estimate	What the sample gives us
Value under the null	What the null hypothesis claims
Standard error	Typical sampling noise
Test statistic	Distance in standard errors

IN THE LESSON

- The workhorse is the test statistic.
- Take your sample estimate, such as the sample mean return.
- Subtract the value the null claims, such as zero.
- Then divide by the standard error, which measures how much the estimate would bounce around from sample to sample.
- The standard error of a mean is the sample standard deviation divided by the square root of the sample size.
- The result counts how many standard errors the estimate lies from the null value.
- With a reasonably large sample, and under the null, this statistic behaves like a standard normal variable, which is why we compare it with normal critical values.
- A statistic near zero is unremarkable.
- A statistic of two or more in size is unusual.

7 Example 1: is the fund manager skilled?

36 months: mean 0.8%, std dev 2.4%. H0: μ = 0 (illustration)

Standard error	2.4% ÷ √36	**0.4%**
Test statistic z	(0.8% − 0) ÷ 0.4%	**2.0**
Critical value (2-sided 5%)	from the normal table	**±1.96**
Decision	2.0 > 1.96	**Reject**

IN THE LESSON

- Our illustration, with made up round numbers.
- Thirty six months of excess returns have a mean of eight tenths of a percent and a standard deviation of two point four percent.
- The null says the true mean is zero.
- First, the standard error is two point four divided by the square root of thirty six, which is six, giving point four percent.
- Second, the test statistic is eight tenths minus zero, divided by point four, which equals two point zero.
- Third, the two sided critical value at five percent is one point nine six.
- Finally, two point zero is larger than one point nine six, so we reject the null.
- At the five percent level the record is unlikely to be pure luck.

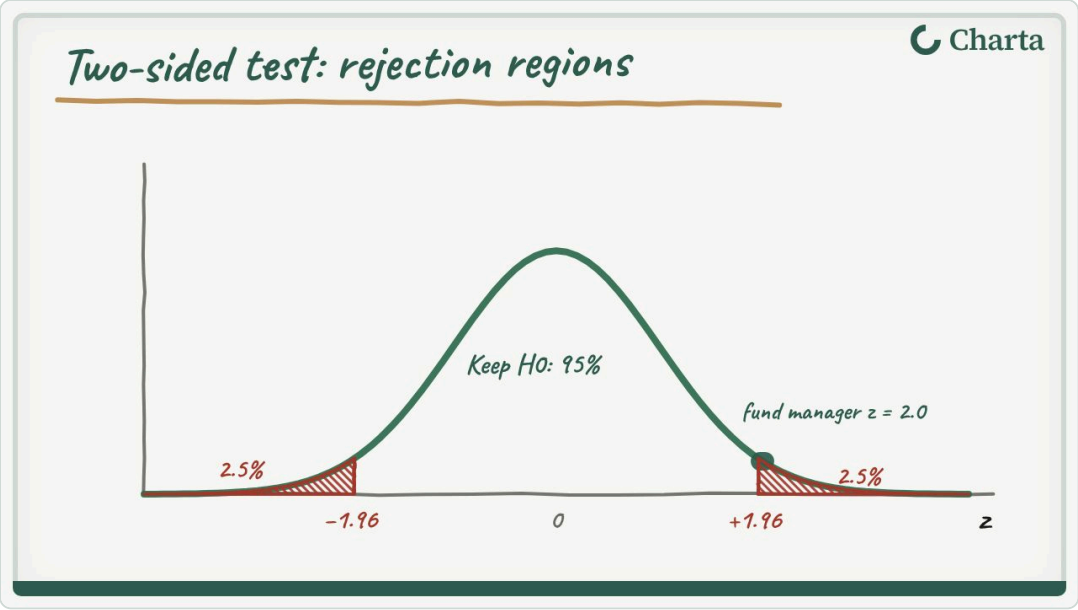
8 One-sided and two-sided tests

- Two-sided: H1 is μ ≠ μ0; α ÷ 2 in each tail
- One-sided: H1 is μ > μ0 or μ < μ0; all of α in one tail
- At 5% with a normal statistic: two-sided ±1.96
- At 5% with a normal statistic: one-sided 1.645

IN THE LESSON

- The alternative hypothesis decides the shape of the test.
- A two sided test has the not equal sign.
- It rejects when the statistic is far out in either tail, so the significance level alpha is split in half, with alpha over two in each tail.
- A one sided test has a greater than or a less than sign.
- It puts all of alpha in a single tail.
- At five percent with a normal statistic, the critical values are one point nine six for a two sided test and one point six four five for a one sided test.
- The one sided cutoff is closer to zero because the whole five percent sits on one side, so a one sided test rejects more easily in the direction it looks.

9 Two-sided test: rejection regions



IN THE LESSON

- Here is the standard normal curve, which describes the test statistic when the null is true.
- The middle area is where ordinary sample noise lives.
- Most of the time the statistic lands there, and we keep the null.
- The hatched areas are the rejection regions.
- They begin at minus one point nine six and at plus one point nine six, and each holds two and a half percent of the probability, so together they hold five percent.
- That total is the significance level.
- If the statistic lands in either hatched region, a result this extreme would happen only five times in a hundred under the null, so we reject.
- A statistic of two point zero, from our fund manager, falls just inside the right hand region.

10 When is a one-sided test appropriate?

- Only if theory or the decision rules out the other side
- Example: a risk limit that matters only when exceeded
- Tail chosen after the sign: 5% test becomes ~10%
- In doubt, use the two-sided test

IN THE LESSON

- Choose a one sided test only when theory or the decision itself rules out the other direction before the data are seen.
- A risk limit is a good example.
- A desk is breaching its limit only if losses are bigger than the limit, so a test of whether average loss exceeds the limit looks in one direction.
- A more profitable result is simply not a problem.
- Using a one sided test because it is easier to pass is not allowed.
- The classic error is to look at the sign of the estimate first, and then pick the tail that agrees with it.
- That secretly turns a five percent test into roughly a ten percent test.
- When there is any doubt, or the direction is not decided in advance, use the two sided test.

11 Example 2: same data, different tail

Mean 0.72%, SE 0.4%, 5% level (illustration)

Test statistic z	0.72% ÷ 0.4%	**1.8**
Two-sided: 1.8 vs 1.96	inside the keep zone	**Keep H0**
One-sided: 1.8 vs 1.645	beyond the cutoff	**Reject H0**
Lesson	choose the tail first	

IN THE LESSON

- Now a lesson in why the tail must be chosen first.
- Suppose the mean is point seven two percent instead, with the same standard error of point four.
- The test statistic is point seven two divided by point four, which is one point eight.
- Under a two sided test at five percent, the cutoff is one point nine six, and one point eight is smaller, so we keep the null.
- Under a one sided test the cutoff is one point six four five, and one point eight is larger, so we reject.
- Same data, opposite answers.
- That is why the choice is allowed only when made in advance for a real reason.
- Choosing after the fact turns the five percent test into roughly a ten percent test.

12 Two kinds of error

- Type I: reject a true null (false alarm); probability = α
- Type II: keep a false null (missed finding); probability = β
- Size of the test = α, the level you choose
- Power = 1 − β: chance of rejecting a false null

IN THE LESSON

- Because the test uses a sample, it can go wrong in two ways.
- A type one error rejects a null that is actually true.
- It is a false alarm, such as declaring a skilless manager skilled.
- The probability of a type one error is called the size of the test, and it equals the significance level alpha that you chose.
- A type two error fails to reject a null that is actually false.
- It is a missed discovery, such as overlooking a truly skilled manager.
- Its probability is called beta.
- The power of the test is one minus beta, the chance of correctly rejecting a false null.
- You never know which world you are in, so you design the test to control both mistakes.

13 The four outcomes of a test

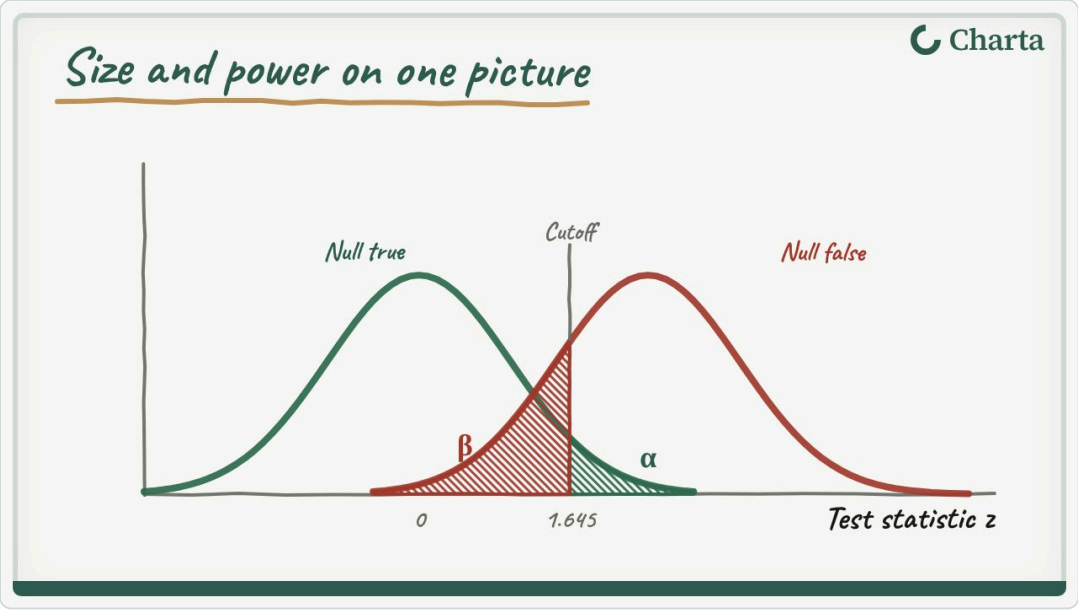
Truth is unknown; the test controls the error rates

Truth, decision	Result
Null true, we reject	Type I error, chance α
Null true, we keep	Correct, chance 1 − α
Null false, we reject	Correct: power = 1 − β
Null false, we keep	Type II error, chance β

IN THE LESSON

- Here is every possibility.
- If the null is true and we reject it, that is a type one error, with probability alpha.
- If the null is true and we keep it, we are right, and that happens with probability one minus alpha.
- If the null is false and we reject it, we are right again, and the probability of that is the power, one minus beta.
- If the null is false and we keep it, that is a type two error, with probability beta.
- Power concerns only worlds where the null is false.

14 Size and power on one picture



IN THE LESSON

- Two curves tell the story.
- The green curve is the distribution of the test statistic when the null is true.
- The red curve is the distribution when the null is false and the true mean sits above it.
- The vertical line is the cutoff for a one sided test.
- To its right, the hatched green sliver under the null curve is alpha, the size, the chance of a false alarm.
- To its left, the hatched region under the red curve is beta, the chance of keeping a false null, a missed finding.
- Everything under the red curve to the right of the line is power.
- Slide the line right to cut alpha, and beta grows.
- Lower the false alarms and you raise the misses.

15 What raises power?

- Lower α: β rises, power falls (for a given sample)
- More power: bigger sample, bigger true effect, less noise
- A low-power test that fails to reject tells you little
- Fail to reject is not evidence of no effect

IN THE LESSON

- For a given sample size, the two errors trade off.
- Lowering alpha makes rejection harder, so beta rises and power falls.
- But you can improve power without paying in false alarms.
- Power rises with a larger sample, which shrinks the standard error.
- It rises with a larger true departure from the null, because a big effect is easy to spot.
- And it rises when the underlying data are less variable, since there is less noise to hide in.
- When a low power test fails to reject, that tells you little, because a false null would often survive it.

16 Example 3: power and sample size

One-sided 5% test, H0: μ = 0, true mean 0.6%, std dev 2.4%

n = 36: cutoff for the mean	1.645 × 0.4%	**0.658%**
n = 36: power	1 − Φ((0.658 − 0.6) ÷ 0.4)	**44%**
n = 144: cutoff for the mean	1.645 × 0.2%	**0.329%**
n = 144: power	1 − Φ((0.329 − 0.6) ÷ 0.2)	**91%**

IN THE LESSON

- Let's compute power.
- The null says the mean is zero, with a one sided test at five percent.
- The true mean is point six, the standard deviation is two point four.
- With thirty six observations the standard error is point four, so we reject when the sample mean exceeds one point six four five times point four, which is point six five eight.
- The true mean is only point six, so power is about forty four percent.
- Now quadruple the sample to one hundred forty four.
- The standard error halves to point two, and the cutoff drops to point three two nine.
- The cutoff now lies one point three five five standard errors below the true mean, so power jumps to about ninety one percent.

17 Confidence intervals: same ingredients

Interval = estimate ± critical value × SE
95%: ±1.96 × SE 99%: ±2.576 × SE

Estimate	The sample estimate
Critical value	Taken at α ÷ 2 in each tail
Standard error	Same standard error as the test
Confidence interval	Plausible values for the truth

IN THE LESSON

- A confidence interval is built from the very same ingredients as a test: the estimate, its standard error and a critical value.
- A two sided interval at confidence level one minus alpha equals the estimate, plus or minus the critical value times the standard error.
- The critical value is taken at alpha over two in each tail, so one point nine six for ninety five percent confidence and two point five seven six for ninety nine percent.
- The interval is a range of plausible values for the true parameter.
- Because the pieces are shared, tests and intervals always agree, and we use that link on the next slide.

18 Example 4: interval and test agree

Mean 0.8%, SE 0.4%. Null value is 0 (illustration)

95% interval	0.8 ± 1.96 × 0.4	**0.016 to 1.584**
Is 0 inside?	No: reject at 5%	**Reject**
99% interval	0.8 ± 2.576 × 0.4	**−0.23 to 1.83**
Is 0 inside?	Yes: keep at 1%	**Keep**

IN THE LESSON

- Back to our fund manager.
- The mean is point eight percent and the standard error is point four.
- For a ninety five percent interval the margin is one point nine six times point four, which is point seven eight four.
- The interval runs from point zero one six to one point five eight four.
- Zero, the value in the null, lies outside it, so a two sided five percent test rejects, matching the earlier statistic of two.
- Now widen to ninety nine percent.
- The margin is two point five seven six times point four, about one point zero three.
- The interval runs from minus point two three to one point eight three.
- Zero is now inside, so the one percent test fails to reject.
- A ninety nine percent interval matches a one percent test.

19 Reading an interval correctly

- Reject at α exactly when μ0 is outside the (1 − α) interval
- 95% interval ↔ 5% test; 99% interval ↔ 1% test
- The interval shows every value that would be rejected
- Not: 95% chance the truth is inside THIS interval

IN THE LESSON

- Here is the link in one sentence.
- A two sided test of the null that the mean equals some value, at significance level alpha, rejects exactly when that value lies outside the one minus alpha interval, and fails to reject when it lies inside.
- So a ninety five percent interval matches a five percent test, and a ninety nine percent interval matches a one percent test.
- There is a common misreading to avoid.
- A ninety five percent interval does not mean there is a ninety five percent probability that the true value is inside this one interval.
- The true value is fixed.
- The ninety five percent describes the method: across many repeated samples, about ninety five percent of intervals built this way would cover the truth.

20 The p value

p (two-sided) = 2 × (1 − Φ(|z|))
p (one-sided) = 1 − Φ(z)

Test statistic	The test statistic
Normal cumulative	Area to the left of a value
p value	Chance of a result this extreme
Test level, alpha	Reject when p is below it

IN THE LESSON

- Reporting only reject or keep throws information away.
- The p value tells you how convincing the evidence is.
- It is the probability, calculated assuming the null is true, of getting a test statistic at least as extreme as the one observed.
- It is also the smallest significance level at which you could reject.
- The rule is simple: reject the null when the p value is below your chosen alpha.
- For a normal statistic, a one sided p value is the tail area beyond the statistic, in the direction of the alternative.
- A two sided p value is twice the tail area beyond the absolute value of the statistic.
- Here the cumulative function of the standard normal gives the area to the left.

21 Example 5: from statistic to p value

Test statistic z = 2.0, from the normal table (illustration)

Tail area beyond 2.0	1 − Φ(2.0)	**0.0228**
Two-sided p value	2 × 0.0228	**0.0455**
Compare with α = 5%	0.0455 < 0.05	**Reject**
Compare with α = 1%	0.0455 > 0.01	**Keep**

IN THE LESSON

- Take the test statistic of two point zero from the fund manager.
- The normal table gives the area to the left of two as point nine seven seven two five, so the tail area beyond two is point zero two two seven five, rounded to point zero two two eight.
- For a one sided test, that is the p value.
- For a two sided test we double it, since we count extremes in both tails, and get point zero four five five.
- Compare with five percent.
- Point zero four five five is below point zero five, so we reject.
- Compare with one percent.
- Point zero four five five is above point zero one, so we keep the null.
- The same evidence is significant at five percent but not at one percent.

22 What a p value is not

- Not the probability that the null is true
- Not the probability that chance alone made the result
- Statistically significant ≠ economically important
- Traps: forgetting to double for two-sided; wrong tail

IN THE LESSON

- A p value is easy to misread.
- It is not the probability that the null is true.
- The calculation assumes the null is true, so it cannot also tell you how likely that assumption is.
- Nor is it the probability that the result arose by chance alone.
- It is the chance of data this extreme given the null.
- A small p value shows statistical significance, but that is not the same as economic importance.
- With a huge sample, a trivially small average gain can have a tiny p value and still be worth nothing after costs.
- Two traps: forgetting to double the tail area in a two sided test, and using the wrong tail.

23 Confidence intervals for one-sided tests

- Two-sided at α ↔ (1 − α); one-sided at α ↔ 1 − 2α
- 5% one-sided uses the 90% interval; 1% uses the 98%
- H1: μ > μ0: reject if the lower limit is above μ0
- H1: μ < μ0: reject if the upper limit is below μ0
- Reject at 95% confidence = 5% significance level

IN THE LESSON

- Intervals can also carry out one sided tests.
- A two sided test at significance alpha matches a one minus alpha interval.
- A one sided test at alpha puts the whole alpha in one tail, so it matches a two sided interval whose confidence level is one minus two times alpha.
- A five percent one sided test therefore uses the ninety percent two sided interval, and a one percent one sided test uses the ninety eight percent interval.
- For an alternative of greater than, reject if the lower limit exceeds the null value.
- For an alternative of less than, reject if the upper limit is below it.
- Rejecting at the ninety five percent confidence level means significance of five percent.
- Using the ninety five percent interval for a five percent one sided test is the common slip.

24 Example 6: a one-sided test by interval

Mean 0.72%, SE 0.4%. H1: μ > 0 at 5% (illustration)

90% interval (right one)	0.72 ± 1.645 × 0.4	**0.062 to 1.378**
Lower limit above 0?	0.062 > 0	**Reject**
95% interval (the slip)	0.72 ± 1.96 × 0.4	**−0.064 to 1.504**
Zero inside, so we'd keep	wrong tool for 1-sided	**Missed**

IN THE LESSON

- Return to the case with a mean of point seven two and a standard error of point four.
- We test the null that the mean is at most zero against the alternative that it is above zero, at five percent.
- A five percent one sided test needs the ninety percent interval.
- The margin is one point six four five times point four, which is point six five eight.
- The interval runs from point zero six two to one point three seven eight.
- The lower limit, point zero six two, is above zero, so we reject.
- This agrees with the earlier one sided statistic of one point eight.
- Now see the slip.
- The ninety five percent interval runs from minus point zero six four to one point five zero four.
- It includes zero, so we would wrongly fail to reject.

25 Testing whether two means differ

- 1. H0: μ1 − μ2 = 0 against H1 (two-sided unless justified)
- 2. Choose the level and the critical value
- 3. Difference in sample means and its standard error
- 4. Statistic = (difference − 0) ÷ SE
- 5. Compare with critical value or p value; conclude

IN THE LESSON

- Often the question compares two groups: do two fund managers, or two market regimes, have different average returns?
- Follow a fixed sequence of five steps.
- First, state the null that the difference of the means is zero, against the alternative that it is not zero, or a one sided version if justified in advance.
- Second, choose the significance level and critical value.
- Third, compute the difference in sample means and its standard error.
- Fourth, form the test statistic: the difference minus zero, divided by the standard error.
- Fifth, compare it with the critical value, or compute the p value, and state the conclusion in words.

26 The standard error of a difference

SE = √(s1² ÷ n1 + s2² ÷ n2)
z = (mean1 − mean2 − 0) ÷ SE

Difference in means	Gap between the sample means
Variance 1	Noise in group one
Variance 2	Noise in group two
Sample sizes	Bigger samples, smaller SE
Rule	Add variances, never subtract SEs

IN THE LESSON

- For two independent samples, the two estimates are separate pieces of noise, and variances of independent pieces add.
- So the variance of the difference equals the first sample variance divided by its sample size, plus the second sample variance divided by its sample size.
- The standard error is the square root of that sum.
- Notice that it is larger than either standard error alone.
- Then the test statistic is the difference of the sample means, minus the claimed difference of zero, divided by that standard error.
- Do not subtract the standard errors, and do not pool the variances unless you have good reason to believe they are equal.

27 Example 7: two managers compared

A: 0.9%, sd 2.4, n 36. B: 0.5%, sd 3.0, n 36 (illustration)

Difference in means	0.9 − 0.5	**0.4**
Standard error	√(2.4²÷36 + 3.0²÷36)	**0.64**
Test statistic	0.4 ÷ 0.64	**0.62**
Decision at 5%	0.62 < 1.96	**Keep H0**

IN THE LESSON

- Manager A has a mean monthly return of point nine percent, a standard deviation of two point four, and thirty six months.
- Manager B has a mean of point five, a standard deviation of three point zero and thirty six months.
- The null says the means are equal.
- The difference is point four.
- The variance term for A is five point seven six over thirty six, which is point one six.
- For B it is nine over thirty six, point two five.
- They add to point four one, and the square root is about point six four.
- The statistic is point four divided by point six four, which is about point six two.
- That is well inside plus or minus one point nine six, so we keep the null.
- The gap looks big but is lost in the noise.

28 Paired samples and pooling

- Independent samples: variances of the two means add
- Paired samples: test the mean of the pairwise differences
- Linked data have covariance; ignoring it misstates the SE
- Pool variances only if you believe they are equal

IN THE LESSON

- Two further points complete the picture.
- First, the formula you just saw assumes the samples are independent.
- If the samples are paired, for example the same stocks measured before and after a rule change, or two managers measured over the same months, the observations are linked.
- Then the covariance matters, and the correct approach is to compute the difference for each pair and test whether the mean of those differences is zero.
- That is a one sample test on the differences.
- Second, some textbooks pool the two variances into one combined estimate.
- Do that only when you believe the true variances are equal.
- Otherwise use the formula with separate variances, which is safe in more situations.

29 Many tests at once

- Screen 20 rules; none truly works
- Each test has a 5% false-alarm chance
- Expect 20 × 5% = 1 false rejection on average
- Reporting only the winner is data snooping

IN THE LESSON

- Now a warning that matters a great deal in finance.
- Suppose you screen twenty trading rules, and none of them truly works.
- Each test at five percent has a five percent chance of a false alarm.
- Run enough of them and the chance that at least one passes becomes high.
- If you then report only the winner, a pure accident looks like a discovery.
- This habit is called data snooping.
- It happens whenever people search across many strategies, factors or models and show only the best.
- The expected number of false rejections is the number of tests times alpha, so twenty tests at five percent give one false alarm on average, even when nothing is real.

30 The chance of at least one false alarm

Chance of ≥ 1 false alarm = 1 − (1 − α)^{m}
Bonferroni: each at α ÷ m, or multiply each p by m
(first line assumes independent true nulls)

Test level, alpha	Level of each single test
Number of tests	How many tests were run
Chance of a false alarm	Grows fast as tests pile up

IN THE LESSON

- Here is the arithmetic.
- For a number of independent true nulls, each tested at level alpha, the chance that none is rejected is one minus alpha, raised to the power of the number of tests.
- So the chance that at least one is falsely rejected is one minus that.
- For a handful of tests the figure is modest, but it grows quickly.
- The simplest fix is the Bonferroni correction.
- Test each hypothesis at alpha divided by the number of tests, or, equivalently, multiply each p value by the number of tests.
- This protects against false positives.
- The price is lower power, because each individual test now demands stronger evidence.

31 Example 8: twenty tests and Bonferroni

20 independent true nulls, each tested at 5% (illustration)

P(at least 1 false alarm)	$1 - 0.95^{20}$	**64%**
Expected false alarms	20×0.05	**1**
Bonferroni cutoff per test	$0.05 \div 20$	**0.0025**
Best rule's adjusted p value	0.01×20	**0.20**

IN THE LESSON

- Back to the twenty trading rules, each tested at five percent.
- If none works, the chance that all twenty are correctly kept is point nine five to the twentieth power, about point three six, so the chance of at least one false alarm is about sixty four percent.
- The expected number of false alarms is twenty times point zero five, which is one.
- The Bonferroni cutoff for each test is point zero five divided by twenty, which is point zero zero two five.
- Now suppose the best rule shows an unadjusted p value of point zero one.
- Multiplied by twenty, its adjusted p value is point two, far above five percent.
- The best of many tests overstates the evidence.

32 Four mistakes to avoid

- Tail picked after the data: decide the tail in advance
- No rejection ≠ proof: say 'not enough evidence'
- p value ≠ chance the null is true: it assumes the null
- Best of many tests, raw p: adjust, e.g. $\alpha \div m$

IN THE LESSON

- Let's collect the four classic mistakes.
- First, choosing the tail after you see the sign of the estimate.
- The fix is to decide in advance, and use two sided when unsure.
- Second, treating a failure to reject as proof that the null is true, especially with a low power test.
- The fix is to say the evidence was not strong enough.
- Third, reading the p value as the probability that the null is true.
- It is the chance of data this extreme if the null holds.
- Fourth, reporting the best of many tests with its raw p value, which overstates the evidence.
- The fix is to adjust, for instance by dividing alpha by the number of tests.

33 Recap

- H0 holds the =, H1 the claim; two-sided splits α
- Type I = α, Type II = β, power = 1 − β, up with n
- Intervals: 1 − α or 1 − 2α; p value: extreme if H0 holds
- Two means: $SE = \sqrt{s1^2 \div n1 + s2^2 \div n2}$
- Many tests: adjust, e.g. $\alpha \div m$

IN THE LESSON

- Let's recap.
- The null hypothesis holds the equality and the alternative holds the claim you want evidence for, both fixed before the data.
- Two sided tests split alpha across two tails, one sided tests use one tail and must be justified in advance.
- Type one error is a false alarm with probability alpha, type two error is a miss with probability beta, and power is one minus beta, helped by larger samples.
- A two sided test at alpha matches a one minus alpha interval, and a one sided test matches one minus two alpha.
- The p value is the chance of an extreme result if the null holds.
- For two means, use the standard error of a difference.
- And with many tests, adjust.

34 Quiz: test yourself

1. Hypotheses for 'excess return is positive'?
   Answer: H0: μ ≤ 0; H1: μ > 0 (claim in H1)
2. Interval for a one-sided 5% test?
   Answer: 90% two-sided: 1 − 2 × α. Lower limit above μ0 rejects
3. Two-sided p value of 0.03?
   Answer: Reject at 5%, keep at 1%. Not P(null true)

IN THE LESSON

- Time for three quick questions.
- Pause the video after each one and answer before you listen to the answer.
- Question one.
- You want evidence that a strategy's average excess return is positive.
- How do you set the hypotheses?
- Question two.
- You run a one sided test at five percent using a confidence interval.
- Which interval do you use?
- Question three.
- A two sided test gives a p value of point zero three.
- What do you conclude at five percent and at one percent, and what does the number not mean?
- Take your time, and then check your answers against the board before we close the lesson.