# Hypothesis tests and p-values: z and t tests, power and multiple comparisons

Ten problems on hypothesis tests: a one-sample z test on a latency benchmark with its two-sided p-value, one-sided against two-sided tests and why the side is chosen first, the confidence interval as the set of nulls a test keeps, a t test on nine measurements with the t critical value, an A/B test of two conversion rates with the pooled standard error, the power of a test against a given effect, the sample size for 80% power, McNemar's test for two classifiers scored on the same examples, the false-positive rate of twenty tests and the Bonferroni fix, and an exact permutation test on three runs per model, with worked solutions and the mistakes of reading the p-value as the probability that the null is true, choosing the side after seeing the data, using z instead of t on a small sample, taking one group's standard error for a difference, and testing twenty metrics and reporting the one that passed.

**BEFORE YOU START**

A hypothesis test asks whether the data are surprising if nothing interesting is going on. The "nothing interesting" is the null hypothesis $H_0$: the new model is no faster, the new button converts no better, the two classifiers are equally accurate. The test turns the data into one statistic whose distribution under $H_0$ is known, usually a standard normal or a $t$ distribution because of the central limit theorem, and the p-value is the probability, computed as if $H_0$ were true, of a statistic at least as extreme as the one observed. A small p-value says the data would be unusual under $H_0$; it does not say how likely $H_0$ is, and a large one does not say the effect is zero. These ten problems run the z test, the t test, the two-proportion test behind every A/B test and the paired test behind every model comparison by hand, then work out power, sample size, the cost of running many tests at once, and an exact test that needs no normal approximation at all.

- **The recipe.** State $H_0$ and the alternative $H_1$, and pick the level $\alpha$ (usually $0.05$) before looking at the data. Compute a test statistic, find its p-value under $H_0$, and reject $H_0$ when $p \le \alpha$.
- **z test for a mean** with known $\sigma$: $z = (\bar{x} - \mu_0)/(\sigma/\sqrt{n})$. Two-sided, $p = 2\big(1 - \Phi(|z|)\big)$; one-sided for $H_1 : \mu > \mu_0$, $p = 1 - \Phi(z)$. The two-sided $5\%$ cutoff is $|z| > 1.96$, the one-sided one $z > 1.645$.
- **t test** when $\sigma$ is estimated by the sample standard deviation $s$: $t = (\bar{x} - \mu_0)/(s/\sqrt{n})$ has a $t$ distribution with $n - 1$ degrees of freedom, whose tails are heavier than the normal's. For $8$ degrees of freedom the two-sided $5\%$ cutoff is $2.306$.
- **Two proportions.** For $\hat{p}_1 = x_1/n_1$ and $\hat{p}_2 = x_2/n_2$ under $H_0 : p_1 = p_2$, pool them, $\hat{p} = (x_1 + x_2)/(n_1 + n_2)$, and use $z = (\hat{p}_2 - \hat{p}_1)\big/\sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n_2)}$.

- **Confidence interval duality.** A two-sided test at level $\alpha$ rejects $\mu_0$ exactly when $\mu_0$ lies outside the $1-\alpha$ confidence interval $\bar{x} \pm z_{\alpha/2}\,\sigma/\sqrt{n}$.
- **Errors and power.** A type I error rejects a true $H_0$, and happens with probability $\alpha$. A type II error keeps a false one. The power is the probability of rejecting when a specific alternative is true. For a two-sided z test with true shift $\delta$ standard errors, power $= \Phi(\delta - 1.96) + \Phi(-\delta - 1.96)$.
- **Sample size** for power $1-\beta$ against a difference $\Delta$: $n = \left((z_{\alpha/2} + z_{\beta})\,\sigma/\Delta\right)^2$, rounded up. For 5% two-sided and 80% power, $z_{\alpha/2} + z_{\beta} = 1.96 + 0.8416$.
- **Many tests.** With $m$ independent true nulls each tested at $\alpha$, $\Pr(\text{at least one rejection}) = 1 - (1-\alpha)^m$. Bonferroni tests each at $\alpha/m$, which keeps that probability at most $\alpha$.

Builds on: Sampling distributions and the central limit theorem: standard errors, confidence intervals and the √N rule, Continuous random variables: densities, CDFs, expectation and the uniform, exponential and normal, Discrete random variables: expectation, variance and the binomial, geometric and Poisson

## Problems

**1.** A serving stack is documented at a mean latency of 200 ms with a standard deviation of 30 ms. After an update, 36 requests average 210 ms. Test $H_0 : \mu = 200$ against $H_1 : \mu \neq 200$ at the 5% level, assuming $\sigma = 30$ is unchanged. difficulty ·

**2.** For the data of Problem 1, find the p-value of the one-sided test of $H_1 : \mu > 200$. Then suppose the sample mean had been 209 ms instead, so $z = 1.8$: find the one-sided and two-sided p-values, and say which one may be reported. difficulty ·

**3.** For the data of Problem 1, find the 95% confidence interval for $\mu$. Use it to decide, without computing a new statistic, whether a two-sided 5% test rejects $\mu_0 = 200$ and $\mu_0 = 205$, and check the second with its p-value. difficulty ··

**4.** Nine runs of a training job take $\bar{x} = 10.6$ hours on average, with sample standard deviation $s = 1.5$ hours. The previous version averaged 9.5 hours. Test $H_0 : \mu = 9.5$ against $H_1 : \mu \neq 9.5$ at the 5% level. difficulty ··

**5.** In an A/B test, 200 of 2000 users convert with the old checkout and 240 of 2000 with the new one. Test whether the conversion rates differ, two-sided at 5%. difficulty ··

**6.** For the test of Problem 1 ($n = 36$, $\sigma = 30$, two-sided 5%), find the probability of a type I error, and the power if the true mean latency is 212 ms. difficulty ··

**7.** How many requests are needed for the test of Problem 1 to have 80% power against a 12 ms change in mean latency, still two-sided at 5% with $\sigma = 30$? Check the power at that $n$. difficulty ··

**8.** Two classifiers are scored on the same 500 test examples. Both are right on 400, both wrong on 40, only A is right on 40 and only B is right on 20. Test whether their accuracies differ with McNemar's test, using the normal approximation and the exact binomial version, and compare with an unpaired two-proportion test. difficulty ···

**9.** A team tracks 20 metrics in an experiment where, in fact, nothing changed, and tests each at the 5% level, independently. Find the probability that at least one metric comes out significant and the expected number that do. Repeat with the Bonferroni level, and say whether a metric with $p = 0.004$ survives it. difficulty ··

**10.** Three seeds of model A score 12, 14 and 15; three seeds of model B score 9, 10 and 13. Run an exact permutation test of whether A scores higher, using the difference in mean scores, one-sided and two-sided. What is the smallest p-value this design can produce? difficulty ···

## Answers

1. $z = 2$, $p \approx 0.0455 < 0.05$: reject $H_0$; the update changed the mean latency
2. $p = 0.0228$ one-sided for $z = 2$; for $z = 1.8$, $0.0359$ one-sided and $0.0719$ two-sided. The one-sided value is valid only if $H_1 : \mu > 200$ was fixed before the data were seen
3. The interval is $[200.2, 219.8]$: it rejects 200 and keeps 205. For 205, $z = (210 - 205)/5 = 1$ and $p = 2(1 - \Phi(1)) \approx 0.317$
4. $t = 2.2 < 2.306$ and $p \approx 0.059 > 0.05$: do not reject at the $5\%$ level
5. $z = 0.02/0.00989 \approx 2.02$ and $p \approx 0.043$: the new checkout converts better, just significantly
6. Type I error $0.05$; power $\approx 0.670$
7. $n = 50$; then $\delta = 12/(30/\sqrt{50}) \approx 2.83$ and the power is $\Phi(2.83 - 1.96) \approx 0.807$
8. McNemar: $z \approx 2.58$, $p \approx 0.010$ (exact $0.013$), so A is more accurate. The unpaired test gives $z \approx 1.82$, $p \approx 0.068$
9. $\Pr(\text{at least one}) \approx 0.64$ and one false positive expected; with Bonferroni, $0.0488 \leq 0.05$. The metric with $p = 0.004$ does not survive, since $0.004 > 0.0025$
10. One-sided $p = 2/20 = 0.10$, two-sided $p = 4/20 = 0.20$; the smallest possible one-sided p-value is $1/20 = 0.05$

## Worked solutions

### Problem 1

> A serving stack is documented at a mean latency of 200 ms with a standard deviation of 30 ms. After an update, 36 requests average 210 ms. Test $H_0 : \mu = 200$ against $H_1 : \mu \neq 200$ at the 5% level, assuming $\sigma = 30$ is unchanged.

| | |
|---|---|
| $\sigma/\sqrt{n} = 30/6 = 5$ | The standard error of the mean of 36 requests. |
| $z = (210 - 200)/5 = 2$ | The sample mean is two standard errors above the null value. |
| $1 - \Phi(2) \approx 0.02275$ | The upper tail of the standard normal beyond 2. |
| $p = 2(1 - \Phi(2)) \approx 0.0455$ | Two-sided: a mean of 190 ms would have been just as surprising, so both tails count. |
| $z = 2$, $p \approx 0.0455 < 0.05$: **reject $H_0$; the update changed the mean latency** | Equivalently, $\lvert z \rvert = 2 > 1.96$. The p-value is the chance of data this extreme if the mean were still 200 ms, not the chance that it is 200 ms (Mistake 1). |

### Problem 2

> For the data of Problem 1, find the p-value of the one-sided test of $H_1 : \mu > 200$. Then suppose the sample mean had been 209 ms instead, so $z = 1.8$: find the one-sided and two-sided p-values, and say which one may be reported.

| | |
|---|---|
| $p_{\text{one}} = 1 - \Phi(2) \approx 0.0228$ | Only the upper tail counts when $H_1$ says the mean went up. |
| $z = (209 - 200)/5 = 1.8$ | The new sample mean. |
| $p_{\text{two}} = 2(1 - \Phi(1.8)) \approx 0.0719$ | Not significant at 5%. |
| $p_{\text{one}} = 1 - \Phi(1.8) \approx 0.0359$ | Significant at 5%, if the one-sided test had been chosen. |

| | |
|---|---|
| $p = 0.0228$ **one-sided for** $z = 2$**; for** $z = 1.8$, $0.0359$ **one-sided and** $0.0719$ **two-sided. The one-sided value is valid only if** $H_1 : \mu > 200$ **was fixed before the data were seen** | A one-sided test is right when only one direction matters, such as "is it slower?". Picking the side that the data happen to favor doubles the false-positive rate (Mistake 2). |

### Problem 3

> For the data of Problem 1, find the 95% confidence interval for $\mu$. Use it to decide, without computing a new statistic, whether a two-sided 5% test rejects $\mu_0 = 200$ and $\mu_0 = 205$, and check the second with its p-value.

| | |
|---|---|
| $1.96 \cdot 5 = 9.8$ | The margin of error: 1.96 standard errors. |
| $210 \pm 9.8 = [200.2,\ 219.8]$ | The 95% confidence interval. |
| $200 \notin [200.2,\ 219.8]$ | So the test rejects $\mu_0 = 200$, as in Problem 1. |
| $205 \in [200.2,\ 219.8]$ | So the test keeps $\mu_0 = 205$. |
| **The interval is** $[200.2, 219.8]$**: it rejects** $200$ **and keeps** $205$**. For** $205$**,** $z = (210 - 205)/5 = 1$ **and** $p = 2(1 - \Phi(1)) \approx 0.317$ | The interval is exactly the set of null values a two-sided 5% test would keep, and it also says how big the change plausibly is, which the p-value does not. |

### Problem 4

> Nine runs of a training job take $\bar{x} = 10.6$ hours on average, with sample standard deviation $s = 1.5$ hours. The previous version averaged 9.5 hours. Test $H_0 : \mu = 9.5$ against $H_1 : \mu \neq 9.5$ at the 5% level.

| | |
|---|---|
| $s/\sqrt{n} = 1.5/3 = 0.5$ | The estimated standard error. |
| $t = (10.6 - 9.5)/0.5 = 2.2$ | The statistic, with $\sigma$ replaced by $s$. |

| | |
|---|---|
| $\mathrm{df} = n - 1 = 8$, cutoff $t_{0.025,8} \approx 2.306$ | Estimating $\sigma$ from nine runs adds uncertainty, so the $t$ cutoff is wider than 1.96. |
| $p = 2\Pr(T_8 > 2.2) \approx 0.059$ | The two tails of the $t$ distribution with 8 degrees of freedom beyond $\pm 2.2$. |
| $t = 2.2 < 2.306$ **and** $p \approx 0.059 > 0.05$: **do not reject at the** $5\%$ **level** | The evidence that the job got slower is suggestive, not conclusive. Treating 2.2 as a z statistic gives $p \approx 0.028$ and the opposite conclusion (Mistake 3). |

### Problem 5

> In an A/B test, 200 of 2000 users convert with the old checkout and 240 of 2000 with the new one. Test whether the conversion rates differ, two-sided at $5\%$.

| | |
|---|---|
| $\hat{p}_1 = 0.10$, $\hat{p}_2 = 0.12$, difference $0.02$ | Two percentage points. |
| $\hat{p} = 440/4000 = 0.11$ | Under $H_0$ the rates are equal, so pool all 4000 users to estimate the common rate. |
| $\hat{p}(1-\hat{p}) = 0.11 \cdot 0.89 = 0.0979$ | The Bernoulli variance at the pooled rate. |
| $\mathrm{SE} = \sqrt{0.0979\,(1/2000 + 1/2000)} \approx 0.00989$ | The variance of a difference of independent estimates is the sum of their variances. |
| $z = 0.02/0.00989 \approx 2.02$ **and** $p \approx 0.043$: **the new checkout converts better, just significantly** | The unpooled standard error $\sqrt{0.1 \cdot 0.9/2000 + 0.12 \cdot 0.88/2000}$ is also about 0.00989 here, so the choice barely matters at these sizes. Using one group's standard error alone gives $z \approx 2.86$ (Mistake 4). |

## Problem 6

For the test of Problem 1 ($n = 36$, $\sigma = 30$, two-sided 5%), find the probability of a type I error, and the power if the true mean latency is 212 ms.

| | |
|---|---|
| $\Pr(\text{type I}) = \alpha = 0.05$ | By construction: under $H_0$, $\lvert Z\rvert > 1.96$ has probability 0.05. |
| Reject when $\bar{x} > 209.8$ or $\bar{x} < 190.2$ | $200 \pm 1.96 \cdot 5$, the cutoffs on the scale of the sample mean. |
| $\Pr(\bar{x} > 209.8) = \Phi\big((212 - 209.8)/5\big) = \Phi(0.44)$ | Under the alternative, $\bar{x} \sim N(212, 5^2)$. |
| $\Phi(0.44) \approx 0.670$ and $\Pr(\bar{x} < 190.2) = \Phi(-4.36) \approx 7 \times 10^{-6}$ | The lower tail is negligible. |
| **Type I error** 0.05; **power** $\approx 0.670$ | The same as $\Phi(\delta - 1.96) + \Phi(-\delta - 1.96)$ with $\delta = 12/5 = 2.4$. A real 12 ms slowdown is missed one time in three, so a p-value above 0.05 from this design is weak evidence of no change. |

## Problem 7

How many requests are needed for the test of Problem 1 to have 80% power against a 12 ms change in mean latency, still two-sided at 5% with $\sigma = 30$? Check the power at that $n$.

| | |
|---|---|
| $z_{\alpha/2} = 1.96$, $z_\beta = 0.8416$ | $\Phi(0.8416) \approx 0.80$. |
| $(1.96 + 0.8416) \cdot 30/12 \approx 7.004$ | The number of standard deviations of the data the effect must be, scaled up. |
| $n = 7.004^2 \approx 49.06$ | The formula $\big((z_{\alpha/2} + z_\beta)\sigma/\Delta\big)^2$. |
| $n = 50$ | Round up: 49 requests fall just short of 80%. |

| | |
|---|---|
| $n = 50$; **then** $\delta = 12/(30/\sqrt{50}) \approx 2.83$ **and the power is** $\Phi(2.83 - 1.96) \approx 0.807$ | Halving the effect to detect quadruples the sample size, since $n$ grows like $1/\Delta^2$. |

### Problem 8

Two classifiers are scored on the same 500 test examples. Both are right on 400, both wrong on 40, only A is right on 40 and only B is right on 20. Test whether their accuracies differ with McNemar's test, using the normal approximation and the exact binomial version, and compare with an unpaired two-proportion test.

| | |
|---|---|
| Accuracies $440/500 = 0.88$ and $420/500 = 0.84$ | A is right on $400 + 40$, B on $400 + 20$. |
| Only the 60 disagreements carry information | On the 440 examples where the models agree, the difference in accuracy is 0. Under $H_0$, each disagreement favors A with probability $\frac{1}{2}$. |
| $z = (40 - 20)/\sqrt{60} \approx 2.58$, so $p \approx 0.0098$ | Under $H_0$, A's wins are $\mathrm{Bin}(60, \frac{1}{2})$, with mean 30 and standard deviation $\sqrt{15}$; $(40 - 30)/\sqrt{15}$ is the same $z$. |
| Exact: $p = 2\Pr(\mathrm{Bin}(60, \frac{1}{2}) \geq 40) \approx 0.0135$ | The normal approximation is a little optimistic here. |
| **McNemar:** $z \approx 2.58$, $p \approx 0.010$ **(exact** $0.013$**), so A is more accurate. The unpaired test gives** $z \approx 1.82$, $p \approx 0.068$ | The unpaired test treats the two accuracies as independent samples, but they come from the same examples. Pairing removes the shared difficulty of the examples, which is why it is the right test, and the more powerful one. |

### Problem 9

A team tracks 20 metrics in an experiment where, in fact, nothing changed, and tests each at the 5% level, independently. Find the probability that at least one metric comes out significant

and the expected number that do. Repeat with the Bonferroni level, and say whether a metric with $p = 0.004$ survives it.

| | |
|---|---|
| $\Pr(\text{no false positives}) = 0.95^{20} \approx 0.358$ | Each true null is kept with probability 0.95, independently. |
| $1 - 0.95^{20} \approx 0.642$ | At least one false positive is more likely than not. |
| $20 \cdot 0.05 = 1$ | The expected number of false positives, by linearity, independent or not. |
| $0.05/20 = 0.0025$, and $1 - 0.9975^{20} \approx 0.0488$ | Bonferroni: test each metric at $\alpha/m$. |
| $\Pr(\text{at least one}) \approx 0.64$ **and one false positive expected; with Bonferroni,** $0.0488 \le 0.05$. **The metric with** $p = 0.004$ **does not survive, since** $0.004 > 0.0025$ | Bonferroni holds the chance of any false alarm at 5% even when the tests are dependent, at the price of power (Mistake 5). |

### Problem 10

Three seeds of model A score 12, 14 and 15; three seeds of model B score 9, 10 and 13. Run an exact permutation test of whether A scores higher, using the difference in mean scores, one-sided and two-sided. What is the smallest p-value this design can produce?

| | |
|---|---|
| Observed difference: $41/3 - 32/3 = 3$ | The two groups' means. |
| There are $\binom{6}{3} = 20$ ways to label three of the six scores "A" | Under $H_0$ the labels are exchangeable, so each split is equally likely. |
| Splits with mean difference at least 3: $\{13, 14, 15\}$ and $\{12, 14, 15\}$ | The difference is at least 3 exactly when the "A" scores sum to at least 41, since the total is 73. |

| | |
|---|---|
| Splits at least 3 apart in either direction: those two and their mirror images $\{9, 10, 12\}$, $\{9, 10, 13\}$ | Sums at most 32. |
| **One-sided $p = 2/20 = 0.10$, two-sided $p = 4/20 = 0.20$; the smallest possible one-sided p-value is $1/20 = 0.05$** | Three seeds per model can never reach significance below 0.05, however large the gap. The test is exact: no normal approximation or equal-variance assumption is needed. |

## Where this goes wrong

### 1. The p-value read as the probability that the null is true

A p-value of 0.0455 invites the sentence "there is only a 4.6% chance the update did nothing".

| | |
|---|---|
| $p = \Pr(\text{data this extreme} \mid H_0) \approx 0.0455$ | Right so far: Problem 1. |
| "So $\Pr(H_0 \mid \text{data}) \approx 0.0455$." | That flips the conditional. Getting $\Pr(H_0 \mid \text{significant})$ needs Bayes' rule, a prior and the power. |
| Among significant results, only 5% are false alarms ✗ | Suppose 90% of tested changes do nothing, and the test has 80% power for the rest. False alarms then make up $0.9 \cdot 0.05 = 0.045$ of all tests and true discoveries $0.1 \cdot 0.8 = 0.08$, so $\Pr(H_0 \mid \text{significant}) = 0.045/0.125 = 0.36$: over a third of the significant results are false alarms. |

### 2. Choosing the side after seeing the data

The data in Problem 2 point upward, so a one-sided test toward "up" looks natural.

| | |
|---|---|
| $z = 1.8$, two-sided $p \approx 0.072$ | Right so far: Problem 2. |
| "The mean went up, so test $H_1 : \mu > 200$." | The direction was chosen by the data. Had the mean gone down, the same analyst would have tested $\mu < 200$. |

| | |
|---|---|
| $p = 0.036$, significant at 5% ✗ | The rule "reject if $z > 1.645$ or $z < -1.645$, whichever way the data went" is a two-sided test at level $\Pr(\lvert Z\rvert > 1.645) \approx 0.10$, not 0.05. The honest p-value is 0.072. |

### 3. z instead of t with nine observations

The t statistic in Problem 4 has the same form as a z statistic, so the normal table seems to apply.

| | |
|---|---|
| $t = (10.6 - 9.5)/(1.5/3) = 2.2$ | Right so far: Problem 4. |
| "Compare with the normal: $2.2 > 1.96$." | $s$ was estimated from the same nine runs, so the statistic varies more than a standard normal. The reference is $t$ with 8 degrees of freedom. |
| $p = 2(1 - \Phi(2.2)) \approx 0.028$, reject ✗ | The right p-value is $2\Pr(T_8 > 2.2) \approx 0.059$, and 2.2 falls short of the cutoff 2.306. For large $n$ the two agree; at $n = 9$ they give opposite conclusions. |

### 4. One group's standard error used for the difference

Each conversion rate has a standard error, and the formula $\sqrt{\hat{p}(1-\hat{p})/n}$ is the familiar one.

| | |
|---|---|
| $\hat{p} = 0.11$, $n = 2000$ per group | Right so far: Problem 5. |
| "SE $= \sqrt{0.11 \cdot 0.89/2000} \approx 0.0070$." | That is the standard error of one rate. The difference of two independent rates has variance equal to the sum of theirs, which adds the second $1/2000$. |
| $z = 0.02/0.0070 \approx 2.86$, $p \approx 0.004$ ✗ | The right standard error is $0.0070\sqrt{2} \approx 0.00989$, giving $z \approx 2.02$ and $p \approx 0.043$. The mistake makes the result look ten times more significant than it is. |

## 5. Testing twenty metrics and reporting the one that passed

Each metric was tested correctly at 5%, so any one that comes out significant seems to be a real finding.

| | |
|---|---|
| Each test alone has a 5% false-positive rate | Right so far. |
| "Report the metric with $p = 0.03$ as a significant improvement." | The chance that some metric out of twenty passes by luck is not 5%. |
| The improvement is significant at the 5% level ✗ | With 20 true nulls, $1 - 0.95^{20} \approx 0.64$, and one false positive is expected (Problem 9). At the Bonferroni level 0.0025, $p = 0.03$ is nowhere near significant; or pick one primary metric before the experiment starts. |

---

**More at eulerpad.app.** This is one of 82 free practice sets with worked solutions at eulerpad.app/practice, alongside guides to the topics behind them.

**EulerPad**, the app, reads your handwritten math on iPad, iPhone and Android and checks it line by line. It's free to start.

This set is also Chapter 18 of *Probability & Statistics for Machine Learning*, a 250-problem workbook with worked solutions: eulerpad.app/workbooks

eulerpad.app/practice/hypothesis-tests-and-p-values