

The logic behind testing hypotheses

We toss a coin 10 times and get 7 tails. Is this sufficient evidence to conclude that the coin is biased?

The **null hypothesis**, H_0 , states that "nothing extraordinary is going on". So in this case

$$H_0: P(T) = \frac{1}{2}$$

The **alternative hypothesis**, H_A , states that there is a different chance process that generates the data. Here we can take

$$H_A: P(T) \neq \frac{1}{2}$$

Hypothesis testing proceeds by collecting data and evaluating whether the data are compatible with H_0 or not (in which case one **rejects** H_0).

The logic behind testing hypotheses

A different example: A company develops a new drug to lower blood pressure. It tests it with an experiment involving 1,000 patients.

In this case "nothing extraordinary going on" means that the drug has no effect. So

H_0 : no change in blood pressure H_A : blood pressure drops

Note that in this case the company would like to reject H_0 !

So the logic of testing is typically indirect: One assumes that nothing extraordinary is happening and then hopes to reject this assumption H_0 .

Setting up a test statistic

A **test statistic** measures how far away the data are from what we would expect if H_0 were true.

The most common test statistic is the **z-statistic**:

$$z = \frac{\text{observed} - \text{expected}}{\text{SE}}$$

‘Observed’ is a statistic that is appropriate for assessing H_0 . In the example of the 10 coin tosses, appropriate statistics would be the number of tails or the percent of tails.

‘Expected’ and SE are the expected value and the SE of this statistic, *computed under the assumption that H_0 is true*.

In the example: Using the formulas for the sum of 0/1 labels we get

‘expected’ = $10 \times \frac{1}{2} = 5$ and $\text{SE} = \sqrt{10} \sqrt{\frac{1}{2} \times \frac{1}{2}} = 1.58$. So

$$z = \frac{7 - 5}{1.58} = 1.27$$

p-values measure the evidence against H_0

Large values of $|z|$ are evidence against H_0 : The larger $|z|$ is, the stronger the evidence.

The strength of the evidence is measured by the

p-value (or: observed significance level):

The p-value is the probability of getting a value of z as extreme or more extreme than the observed z , assuming H_0 is true.

But if H_0 is true, then z follows that standard normal curve, according to the central limit theorem, so the p-value can be computed with normal approximation:

The smaller the p-value, the stronger the evidence against H_0 . Often the criterion for rejecting H_0 is a p-value smaller than 5%. Then the result is called **statistically significant**.

p-values measure the evidence against H_0

In the example:

Note that the p-value **does not** give the probability that H_0 is true, as H_0 is either true or not - there are no chances involved. Rather, it gives the probability of seeing a statistic as extreme, or more extreme, than the observed one, assuming H_0 is true.

Distinguishing Coke and Pepsi by taste

It has been said that it is difficult to distinguish Coke and Pepsi by taste alone, without the visual cue of the bottle or can.

In an experiment that I did in a class at Stanford, 10 cups were filled at random with either Coke or Pepsi. A student volunteer tasted each of the 10 cups and correctly named the contents of seven. Is this sufficient evidence to conclude that the student can tell apart Coke and Pepsi?

"Nothing extraordinary is going on" means that the student does not have any special ability to tell them apart and is just guessing.

To write this down formally we introduce 0/1 labels since we are counting correct answers: 1 = correct answer, 0 = wrong answer

$$H_0: P(0) = P(1) = \frac{1}{2} \qquad H_A: P(1) > \frac{1}{2}$$

This is a **one-sided test**: the alternative hypothesis for $P(1)$ we are interested in is on one side of $\frac{1}{2}$.

Distinguishing Coke and Pepsi by taste

Since we are looking at the sum of ten 0/1 labels, the z-statistic is the same that we had for coin-tossing:

$$z = \frac{\text{observed sum} - \text{expected sum}}{\text{SE of sum}} = \frac{7 - 5}{1.58} = 1.27$$

But since we do a one-sided test instead of a two-sided test, the p-value is only half as large:

Since 10.2% is not smaller than 5%, we don't reject H_0 : We are not convinced that the student can distinguish Coke and Pepsi.

Distinguishing Coke and Pepsi

A two-sided alternative might also be appropriate:

$$H_A: P(1) \neq \frac{1}{2}$$

H_A corresponds to a student who is more likely than not to distinguish Coke and Pepsi, but who may confuse them. Such a student might get one correct answer (say).

One has to carefully consider whether the alternative should be one-sided or two-sided, as the p-value gets doubled in the latter case.

It is not ok to change the alternative afterwards in order to get the p-value below 5%.

The t-test

The health guideline for lead in drinking water is a concentration of not more than 15 parts per billion (ppb).

Five independent samples from a reservoir average 15.6 ppb. Is this sufficient evidence to conclude that the concentration μ in the reservoir is above the standard of 15 ppb?

Recall our model for measurements:

$$\text{measurement} = \mu + \text{measurement error}$$

So it may be that the concentration μ is below 15 ppb, but measurement error results in an average of 15.6 ppb.

$$H_0: \mu = 15 \text{ ppb} \quad H_A: \mu > 15 \text{ ppb}$$

We can try a z-test for the average of the measurements:

$$z = \frac{\text{observed average} - \text{expected average}}{\text{SE of average}} = \frac{15.6 \text{ ppb} - 15 \text{ ppb}}{\text{SE of average}}$$

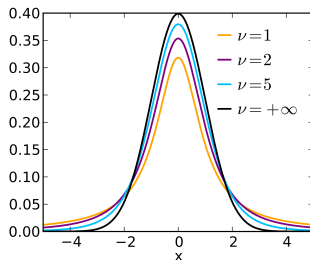
since the measurement error has expected value zero.

The t-test

SE of average = $\frac{\sigma}{\sqrt{n}}$, but the standard deviation σ of the measurement error is unknown.

We can estimate σ by s , the sample standard deviation of the measurements. However:

If we estimate σ and n is small ($n \leq 20$), then the normal curve is not a good enough approximation to the distribution of the z-statistic. Rather, an appropriate approximation is **Student's t-distribution with $n - 1$ degrees of freedom**:



The t-test

The fatter tails account for the additional uncertainty introduced by estimating σ by

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}.$$

Using the **t-test** in place of the z-test is only necessary for small samples: $n \leq 20$ (say).

In that case it is also better to replace the confidence interval $\bar{x} \pm z$ SE by

$$\bar{x} \pm t_{n-1} \text{SE}$$

More on testing

- ▶ *Statistically significant does not mean that the effect size is important:*

Suppose the sample average shows a lead concentration that is only slightly above the health standard of 15 ppb: say the sample average is 15.05 ppb.

That may not be of practical concern, even though the test may be highly significant: Statistical significance convinces us that there is an effect, but it doesn't say how big the effect is.

Reason: A large sample size n makes $SE = \frac{\sigma}{\sqrt{n}}$ small, so even a small exceedance over the limit by (say) 0.05 ppb may give a statistically significant result.

Therefore it is helpful to complement a test with a confidence interval: In the above case a 95% confidence interval for μ might be [15.02 ppb, 15.08 ppb].

More on testing

- ▶ There is a general connection between confidence intervals and tests:
A 95% confidence interval contains all values for the null hypothesis that will not be rejected by a two-sided test at a 5% significance level.
(A 5% **significance level** means that the threshold for the p-value is 5%).
- ▶ There are two ways that a test can result in a wrong decision:
 H_0 is true, but was erroneously rejected \rightarrow Type I error ('false positive')
 H_0 is false, but we fail to reject it \rightarrow Type II error
Rejecting H_0 if the p-value is smaller than 5% means $P(\text{type I error}) \leq 5\%$

The two-sample z-test

Last month, the President's approval rating in a sample of 1,000 likely voters was 55%. This month, a poll of 1,500 likely voters resulted in a rating of 58%. Is this sufficient evidence to conclude that the rating has changed?

We want to assess whether

p_1 = proportion of all likely voters approving last month
is equal to

p_2 = proportion of all likely voters approving this month

"nothing unusual is going on" means $p_1 = p_2$. It's common to look at the difference $p_2 - p_1$ instead:

$$H_0 : p_2 - p_1 = 0$$

$$H_1 : p_2 - p_1 \neq 0$$

p_1 is estimated by $\hat{p}_1 = 55\%$, p_2 by $\hat{p}_2 = 58\%$. The central limit theorem applies to the difference $\hat{p}_2 - \hat{p}_1$ just as it does to \hat{p}_1 and \hat{p}_2 . So we can use a z-test:

The two-sample z-test

We can use a z-test for the difference $\hat{p}_2 - \hat{p}_1$:

$$z = \frac{\text{observed difference} - \text{expected difference}}{\text{SE of difference}} = \frac{(\hat{p}_2 - \hat{p}_1) - (p_2 - p_1)}{\text{SE of difference}}$$

An important fact is that if \hat{p}_1 and \hat{p}_2 are independent, then

$$\text{SE}(\hat{p}_2 - \hat{p}_1) = \sqrt{(\text{SE}(\hat{p}_1))^2 + (\text{SE}(\hat{p}_2))^2}. \quad \text{So}$$

$$z = \frac{(\hat{p}_2 - \hat{p}_1) - 0}{\sqrt{\sqrt{\frac{p_1(1-p_1)}{1000}}^2 + \sqrt{\frac{p_2(1-p_2)}{1500}}^2}} = \frac{0.03}{0.0202} = 1.48$$

The two-sample z-test

The confidence interval for $p_2 - p_1$ is

$$(\hat{p}_2 - \hat{p}_1) \pm z \text{SE}(\hat{p}_2 - \hat{p}_1) = [-1\%, 7\%] \quad \text{when } z = 2$$

We can improve the estimate of $\text{SE}(\hat{p}_2 - \hat{p}_1)$ somewhat by using the fact that $p_1 = p_2$ on H_0 . Since there is a common proportion we can estimate it by **pooling** the samples:

$0.55 \times 1000 = 550$ voters approve in the first sample, 870 in the second, so in total there are 1420 approvals out of 2500. So the **pooled estimate** of $p_1 = p_2$ is $\frac{1420}{2500} = 56.8\%$.

So we estimate $\text{SE}(\hat{p}_2 - \hat{p}_1)$ by $\sqrt{\frac{0.568(1-0.568)}{1000} + \frac{0.568(1-0.568)}{1500}} = 0.02022$, which essentially gives the same answer in this case.

The two-sample z-test

The two-sample z-test is applicable in the same way to the difference of two sample means in order to test for equality of two population means.

If the two samples are independent, then again

$$SE(\bar{x}_2 - \bar{x}_1) = \sqrt{(SE(\bar{x}_1))^2 + (SE(\bar{x}_2))^2}$$

and $SE(\bar{x}_1) = \frac{\sigma_1}{\sqrt{n_1}}$ is estimated by $\frac{s_1}{\sqrt{n_1}}$.

If the sample sizes n_1, n_2 are not large, then the p-value needs to be computed from the t-distribution.

The pooled standard deviation

If one has reason to assume that $\sigma_1 = \sigma_2$ (or if this has been checked), then one may use the **pooled estimate** for $\sigma_1 = \sigma_2$ given by

$$s_{\text{pooled}}^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

However, the advantages of using s_{pooled}^2 are small and the analysis rests on the assumption that $\sigma_1 = \sigma_2$. For these reasons the pooled t-test is usually avoided.

All of the above two-sample tests require that the two samples are independent. They are also applicable in special situations where the samples are dependent, e.g. to compare the treatment effect when subjects are randomized into treatment and control groups.

The paired-difference test

Do husbands tend to be older than their wives?

The ages of five couples:

Husband's age	Wife's age	age difference
43	41	2
71	70	1
32	31	1
68	66	2
27	26	1

The two-sample t-test is not applicable since the two samples are not independent. Even if they were independent, the small differences in ages would not be significant since the standard deviations are large for husbands and also for the wives.

The paired-difference test

Since we have paired data, we can simply analyze the differences obtained from each pair with a regular t-test, which in this context of **matched pairs** is called **paired t-test**:

H_0 : population difference is zero

$t = \frac{\bar{d}-0}{SE(\bar{d})}$, where d_i is the age difference of the i th couple.

$SE(\bar{d}) = \frac{\sigma_d}{\sqrt{n}}$. Estimate σ_d by $s_d = 0.55$. Then $t = \frac{1.4-0}{0.55/\sqrt{5}} = 5.69$

The independence assumption is in the sampling of the couples.

The sign test

What if didn't know the age difference d_i but only if the husband was older or not?

We can test

H_0 : half the husbands in the population are older than their wives
using 0/1 labels and a z-test, just as we tested whether a coin is fair:

$$z = \frac{\text{sum of } 1s - \frac{n}{2}}{\text{SE of sum}} = \frac{5 - \frac{5}{2}}{\sqrt{5 \cdot \frac{1}{2}}} = 2.24 \quad \text{since } \sigma = \frac{1}{2} \text{ on } H_0.$$

The p-value of this **sign-test** is less significant than that of the paired t-test. This is because the latter uses more information, namely the size of the differences. On the other hand, the sign test has the virtue of easy interpretation due to the analogy to coin tossing.