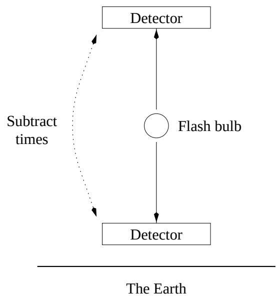
## *A Hypothesis Testing Problem*

Does gravity slow down light? We might try to find out with an apparatus like this:




A flash of light goes both up and down.

We record the arrival times at the two detectors, and take the difference of the upward time minus the downward time to see if the speeds differ.

After eliminating all sources of systematic error (eg, unequal distances), we make $n$ measurements of the arrival time difference.

If the average of these measurements is not zero, can we conclude that gravity affects light speed? How far from zero does the average have to be before we conclude that there is a real difference?

## *The Logic of Hypothesis Testing*

All measurements have errors, so in our experiment we don't expect to find that the arrival times are *exactly* equal. Looking at the average difference reduces variability, but doesn't eliminate it.

*The average of the measured differences will be non-zero even if the true difference is zero.*

To avoid announcing amazing discoveries every time we run an experiment, we want to declare that we have found a real difference in arrival times only if a difference as big as we see *is unlikely to occur by chance if the real difference is zero*.

The probability of a difference as big as we saw occurring when the *null hypothesis* of zero real difference is true is called the *P-value*.

## *P-Values From Test Statistics*

To find a $P$-value, we select a *test statistic* that measures how extreme our results are. We must know the sampling distribution of the test statistic *under the null hypothesis*.

Here, based on measurements $x_1, \ldots, x_n$ of the time difference, we will use the test statistic

$$z = \frac{\bar{x}}{\sigma/\sqrt{n}}$$

where $\sigma$ is the standard deviation of the error in one measurement (assume we know this).

If $\bar{x}$ has an approximately normal distribution, the distribution of $z$ will be $N(0,1)$, if the null hypothesis that $\mu_X = 0$ is true.

From our data, we calculate $z$, and then find out the probability that a $z$ this big or bigger would occur if the null hypothesis is true. This is our $P$-value.

## *What Does the P-Value Mean?*

A small $P$-value (close to zero) is evidence that the null hypothesis is false. People often speak of *rejecting* the null hypothesis if the $P$-value is smaller than some critical value. Use of a critical value of 0.05 is traditional, but **not** for any good reason.

Why might a small $P$-value lead you to doubt the null hypothesis? Because it can only have come about in two ways:

1) An unlikely event occurred (getting a big value for the test statistic by chance).

2) The null hypothesis is false.

Note that if the null hypothesis is false, a large value for the test statistic may not be unlikely. We may therefore prefer explanation (2) to explanation (1).

## *What Do We Mean by "Big"*

The $P$-value is the probability of getting a test statistic as big or bigger as we observed, under the null hypothesis. But what counts as big?

In our experiment, we might find that $z$ is a large *positive* number or a large *negative* number — corresponding to the average upward travel time being *greater* or *less* than the average downward travel time. Should both count as big?

If we count only large positive (or only large negative) values as big, we are using a *one-sided* test (also called a *one-tailed* test).

If we count large values in either direction as big, we are using a *two-sided* (or *two-tailed*) test.

## *Calculating the P-Value*

For our experiment, suppose we know that the standard deviation of the measurement error is $\sigma = 0.1$. We do $n = 4$ measurements, getting values of

$$0.21,\ -0.15,\ 0.19,\ 0.02$$

The mean of these is $\bar{x} = 0.0675$. The test statistic is

$$z \;=\; \frac{\bar{x}}{\sigma/\sqrt{n}} \;=\; \frac{0.0675}{0.1/\sqrt{4}} \;=\; 1.35$$

Under the null hypothesis that the mean is zero, the *one-sided* $P$-value (in the positive direction) is

$$P(z \geq 1.35) \;=\; 0.0885$$

The *two-sided* $P$-value is

$$P(z \geq 1.35 \text{ or } z \leq -1.35) \;=\; 2\times 0.0885 \;=\; 0.1770$$

What would you conclude?

## *What Does the P-Value NOT Mean?*

The $P$-value is **not** the probability that the null hypothesis is true. Such a probability is not meaningful if we are viewing probability in terms of frequencies in repeatable situations.

A small $P$-value does **not** mean we are compelled to reject the null hypothesis by mathematics. Sometimes, common sense will compel us to reject the null hypothesis if the $P$-value is small, but not always.

The $P$-value is **not** the probability of the data that we observed, nor the probability of the test statistic that we computed from this data. (These probabilities are zero for data and test statistics with continuous distributions.)

The $P$-value is the probability under the null hypothesis of the observed data, *or data more extreme* (as determined by the test statistic).

## *Null and Alternative Hypotheses*

One way of thinking about what whether to do a one-sided or two-sided test is by considering both *null* and *alternative* hypotheses.

For our experiment, if the true average difference is travel times is $\mu$, we might use the following hypotheses:

$$H_0: \quad \mu = 0$$
$$H_a: \quad \mu \neq 0$$

This is a two-sided alternative, leading to a test where we look at both tails of the test statistic's distribution.

We could instead use a one-sided alternative:

$$H_0: \quad \mu = 0$$
$$H_a: \quad \mu > 0$$

With the one-sided test, we are supposed to ignore results going the other way! Reasonable?