JOINT INSTITUTE
交大密西根学院

# The Fisher Test

# Hypotheses and Testing

In this section we will discuss the second major statistical method for gaining information on a probability distribution: ***hypothesis testing***. The goal is to reject or fail to reject statements (hypotheses) based on statistical data.

We will present three approaches:

(i) Fisher's null hypothesis testing,

(ii) Neyman–Pearson decision theory,

(iii) The amalgam of (i) and (ii) that is still used and sometimes taught today, called Null Hypothesis Significance Testing.

In our initial discussion, a hypothesis will be a statement about a population parameter, denoted $\boldsymbol{\theta}$.

The hypothesis will compare $\theta$ to a ***null value***, denoted $\boldsymbol{\theta_0}$.

# Fisher's Null Hypothesis Test

We consider a single hypothesis that compares a population parameter $\theta$ to a given null value $\theta_0$.

This hypothesis will be denoted by $H_0$ and is called the ***null hypothesis***.

Our goal is to find statistical evidence that allows us to reject the null hypothesis.

The process of using statistical data to decide whether or not a hypothesis should be rejected is called "performing a hypothesis test".

Null hypotheses take one of three forms:

- $H_0 : \theta = \theta_0$
- $H_0 : \theta \leq \theta_0$
- $H_0 : \theta \geq \theta_0$

# Fisher's Null Hypothesis Test

14.1. Example. We want to find evidence that a new car design has a mean mileage greater than 26 mpg. Therefore, we set up the null hypothesis

$$H_0 : \mu \leq 26. \qquad (14.1)$$

反证法?

Our goal is to gather data that allows us to ***reject $H_0$***.

14.2. Remark. A hypothesis test is based on rejecting a hypothesis because it is possible to gather statistical evidence that a certain claim is likely to be false, while it is impossible for statistical evidence to directly prove that a claim is true. This will become more clear soon.

Suppose that the hypothesis (14.1) is given. We then take a random sample and calculate $\overline{X}$. If the value of $\overline{X}$ is much greater than 26, there is reason to believe that $H_0$ is false.

# The *P*-Value for a One-Tailed Test

The test of a hypothesis of the form

$$H_0 : \theta \leq \theta_0 \qquad \text{or} \qquad H_0 : \theta \geq \theta_0$$

is said to be a ***one-tailed test***.

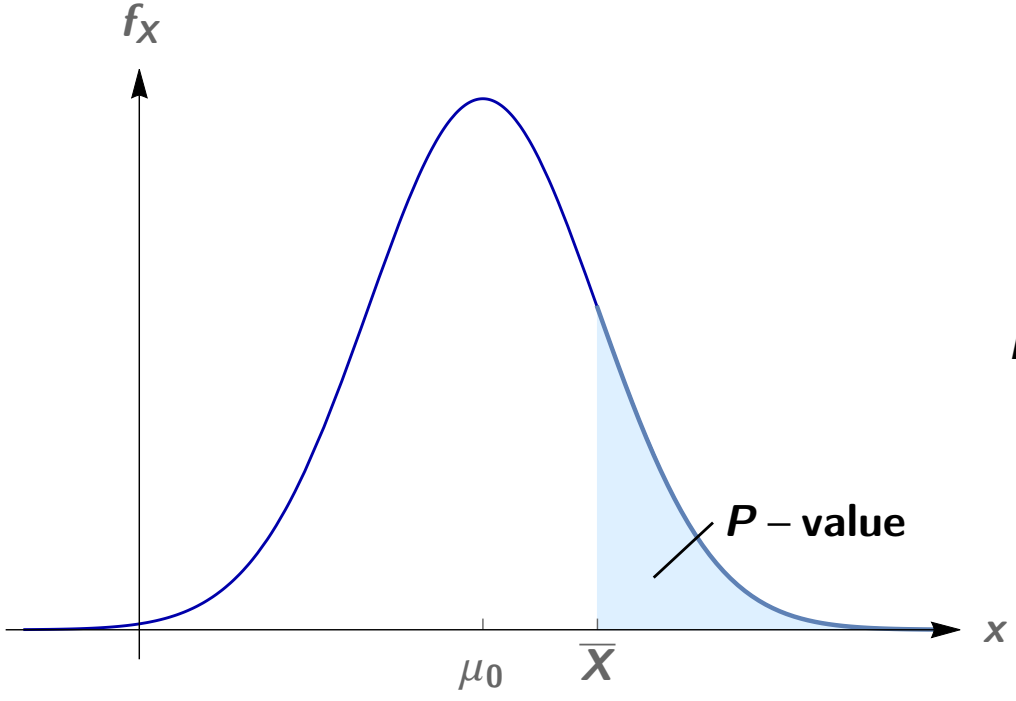
In our current example (14.1) we take a random sample of size $n$ and find the value $\overline{x}$ for the sample mean. We then find the probability of obtaining the measured value of $\overline{x}$ or a larger result if $\theta = \theta_0$. This is said to be the ***significance*** or ***P-value*** of the test.

Note that finding the probability that we obtain $\overline{x}$ or a greater result if $\mu = 26$ is an upper bound on the probability given $\mu \leq 26$:

$$P[\overline{X} \geq \overline{x} \mid \mu \leq 26] \leq P[\overline{X} \geq \overline{x} \mid \mu = 26]$$

# The *P*-Value for a One-Tailed Test

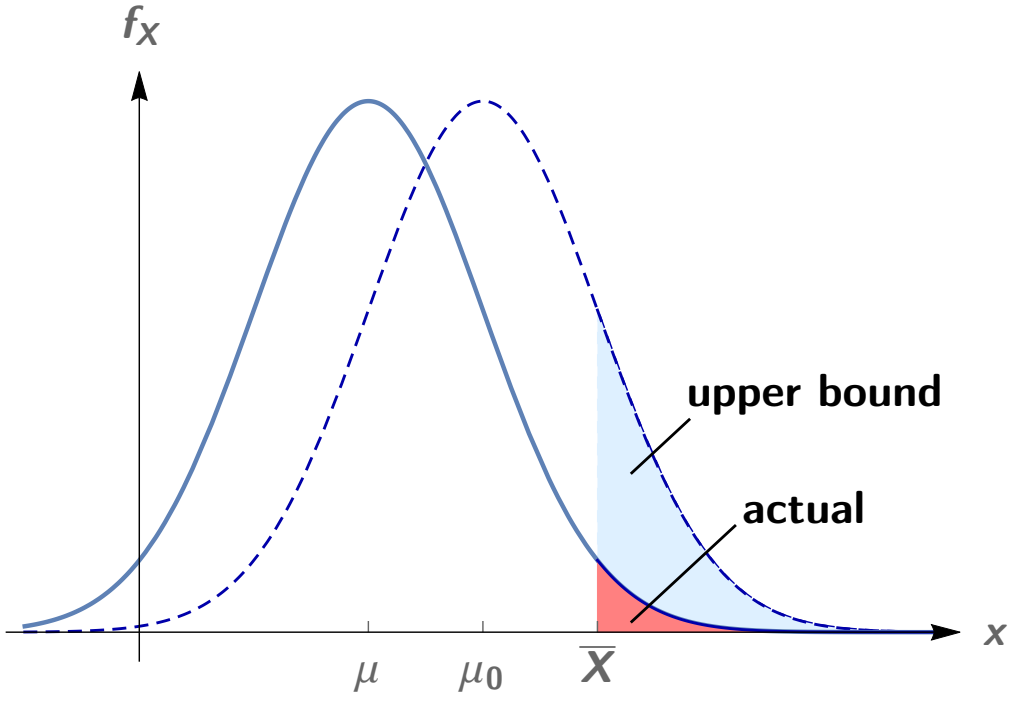
This is illustrated in the sketch below. The sample mean $\overline{X}$ follows a normal distribution with mean $\mu$. If $\mu = \mu_0$, the probability of obtaining a value of $\overline{X}$ at least equal to the measured $\overline{x}$ is indicated by the shaded region.




$H_0 : \mu \leq \mu_0$

# The *P*-Value for a One-Tailed Test

If $\mu < \mu_0$, the probability will be smaller, since the density curve will be shifted to the left.




$H_0 : \mu \leq \mu_0$

# The $P$-Value and Rejecting the Null Hypothesis

The $P$-value is therefore an upper bound of the probability of obtaining the data if $H_0$ is true. If $D$ represents the statistical data,

$$P[D \mid H_0] \leq P\text{-value}$$

and we will reject $H_0$ if this value is small.

We then say that we either

- ***fail to reject $H_0$*** or
- ***reject $H_0$ at the [P-value] level of significance***.

The $P$-value is also called the level of significance of the test.

The statistic on which the $P$-value is based is called the ***test statistic***. In our discussion so far, the test statistic has been the sample mean.

# A One-Tailed Test Based on the Normal Distribution

14.3. Example. Continuing from Example 14.1, we may assume that the mileage of cars currently has a standard deviation of 5 miles per gallon and that this will also be true for the new design. Furthermore, we suppose that the gas mileage follows a normal distribution.

We take a sample of 36 cars and find their gas mileages. We decide to base our rejection of $H_0$ on the sample mean.

If $\mu = 26$ and $\sigma = 5$, the sample mean is normally distributed with $\mu = 26$ and standard deviation $\sigma/\sqrt{n} = 5/6$.

Suppose that we find a sample mean $\overline{x} = 28.04$ mpg.

# A One-Tailed Test Based on the Normal Distribution

We now calculate the $P$-value of the test, i.e., the probability of obtaining this or a larger value of the sample mean if $H_0$ were true.

$$\begin{aligned} P[\overline{X} \geq 28.04 \mid \mu \leq 26, \sigma = 5] &\leq P[\overline{X} \geq 28.04 \mid \mu = 26, \sigma = 5] \\ &= P\left[\frac{\overline{X} - 26}{5/6} \geq \frac{28.04 - 26}{5/6}\right] \\ &= P[Z \geq 2.45] = 1 - P[Z \leq 2.45] \\ &= 1 - 0.9929 = 0.0071. \end{aligned}$$

This is the $P$-value of the test. Since it is very small, we decide to reject the null hypothesis at the 0.7% level of significance.

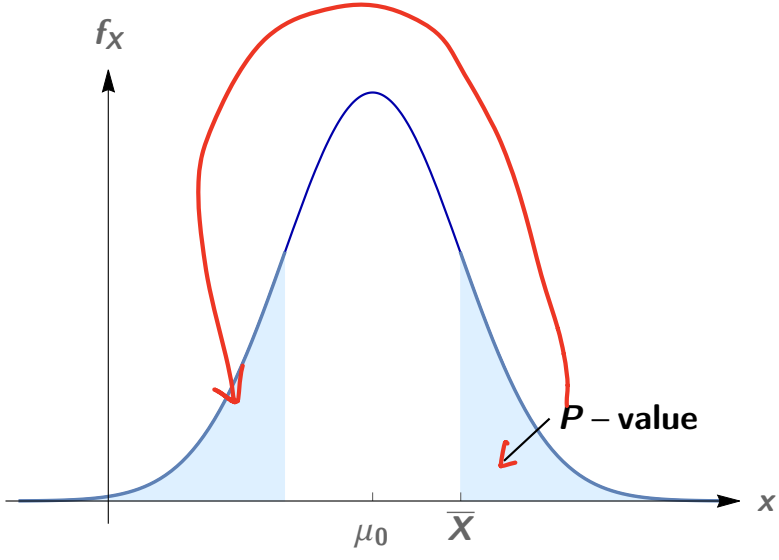
We may say that there is evidence that the gas mileage of the cars of new design is greater than 26 mpg.

# Two-Tailed Tests

If we are testing a hypothesis of the form

$$H_0 : \theta = \theta_0$$

we say we are performing a ***two-tailed test***. In this case, the $P$-value is twice the value of a one-tailed test, since there is evidence that the null hypothesis is false if the statistic differs from $\theta_0$ significantly, regardless of whether the statistic is greater or smaller.




$$H_0 : \mu = \mu_0$$

# A Twp-Tailed Test Based on the Normal Distribution

14.4. Example. The burning rate of a rocket propellant is being studied. Specifications require that the mean burning rate must be 40 cm/s. Furthermore, suppose that we know that the standard deviation of the burning rate is approximately $\sigma = 2\,\text{cm/s}$. The experimenter decides to base the test on a random sample of size $n = 25$. The null hypothesis is

$$H_0 : \mu = 40\,\text{cm/s}$$

If $H_0$ is true, the sample mean is normally distributed with mean $\mu_0 = 40\,\text{cm/s}$ and variance $\sigma^2/n$; thus she will use the test statistic

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$$

which is standard normal if $H_0$ is true.

# The $Z$-Test

Twenty-five specimen are tested, and the sample mean burning rate obtained is $\overline{x} = 41.25\,\text{cm/s}$. The value of the test statistic is

$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{41.25 - 40}{2/\sqrt{25}} = 3.125.$$

Then

$$P[Z \geq 3.125 \mid H_0] = 1 - P[Z \leq 3.125 \mid H_0] = 1 - 0.9991 = 0.0009$$

Since this is a two-tailed test, the $P$-value is twice this number and she decides to reject $H_0$ at the 0.18% level of significance.

There is evidence that the burning rate is not $40\,\text{cm/s}$.

# Fisher's Null Hypothesis Test

## 14.5. Remarks.

- Fisher originally recommended rejecting $H_0$ if the $P$-value is less than 5%, i.e., $P < 0.05$. However, he later changed his mind and advocated quoting the actual $P$-value and deciding whether or not to reject $H_0$ on a case-by-case basis.

- According to Fisher, this type of test should only be used if very little is known about the parameter $\theta$. The hypothesis test is just a first step in investigating $\theta$. Confidence intervals and other techniques give far more information in practice.

- Fisher also observed that a single significant test should not be enough to comprehensively reject $H_0$. Only multiple, independent significant test should be enough to allow the conclusion that $H_0$ is actually false.

# Does a small $P$-value provide evidence that $H_0$ is false?

But there is also a more fundamental issue: what a researcher wants is, given data $D$, the probability that $H_0$ is true, i.e.,

$$P[H_0 \mid D].$$

But the $P$-value is just the converse probability, i.e.,

$$P[D \mid H_0].$$

It is easy to write down Bayes's theorem and see that

$$P[H_0 \mid D] = \frac{P[D \mid H_0] \cdot P[H_0]}{P[D \mid H_0] \cdot P[H_0] + P[D \mid \neg H_0] \cdot P[\neg H_0]}.$$

Since $P[\neg H_0] = 1 - P[H_0]$ we find that

$$P[H_0 \mid D] = \frac{P[D \mid H_0] \cdot P[H_0]}{P[D \mid H_0] \cdot P[H_0] + P[D \mid \neg H_0](1 - P[H_0])}.$$

# Is Hypothesis Testing logical?

Then if $P[H_0] \neq 0$ then

$$P[H_0 \mid D] = \frac{P[D \mid H_0] \cdot P[H_0]}{P[D \mid H_0] \cdot P[H_0] + P[D \mid \neg H_0](1 - P[H_0])}$$
$$= \frac{1}{1 + \frac{P[D|\neg H_0]}{P[D|H_0]} \frac{1-P[H_0]}{P[H_0]}}$$

This shows the following:

- If $P[H_0]$ is small, even a large $P$-value does not mean that $H_0$ is likely to be true.
- If $P[H_0]$ is large, even a small $P$-value does not mean that $H_0$ is likely to be false.

Hence, it is possible that we have data which is very unlikely given $H_0$, but that in fact $H_0$ given the data is very likely (and vice-versa).

In short, classical hpyothesis testing does not take the probability of $H_0$ being true in the first place into account.

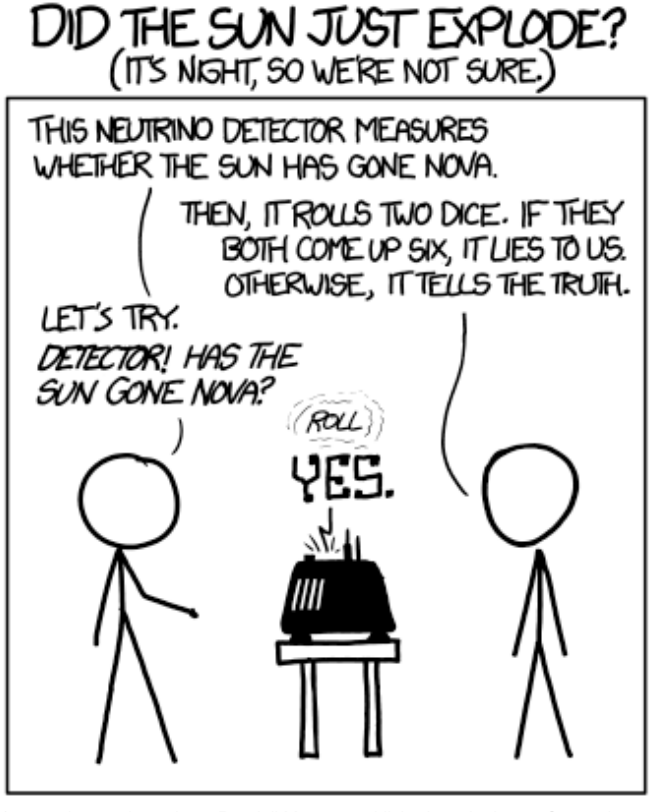
# Bayesian vs. Frequentist Statistics

This problem has given rise to ***Bayesian statistics*** which attempts to assign a ***prior probability*** to $H_0$ and deduce a ***posterior probability*** based on experimental results. However, finding this prior probability is often tricky. There are, broadly speaking, two groups of statisticians:

- ***Frequentists***, who mainly ignore the problems mentioned here or claim that they are not relevant in their specific research (for example, because they consider $P[H_0]$ to not be small in their experiments).
- ***Bayesians*** who claim to understand the logical inconsistencies and intend to compensate for them via prior and posterior probability distributions. While theoretically pure, this may be difficult to implement in practice.

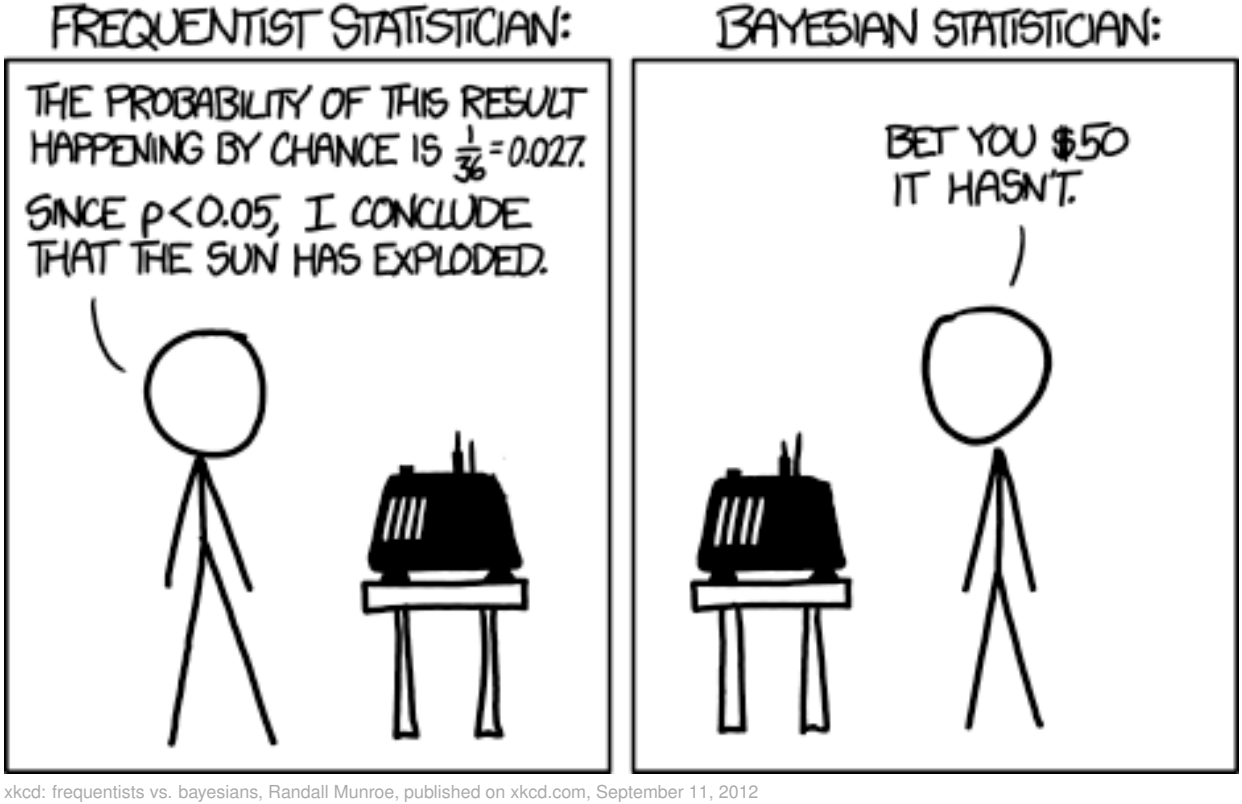
Of course, these are extreme characterizations. In practice, every statistician knows Bayes's theorem and will apply it as much as possible and no statistician entirely rejects frequentist methods.

# Bayesian vs. Frequentist Statistics




xkcd: frequentists vs. bayesians, Randall Munroe, published on xkcd.com, September 11, 2012

# Bayesian vs. Frequentist Statistics




xkcd: frequentists vs. bayesians, Randall Munroe, published on xkcd.com, September 11, 2012

# Does a small $P$-value provide evidence that $H_0$ is false?

A more serious example:

> *For young women of age 30 the incidence of live-born infants with Down's syndrome is 1/885, and the majority of pregnancies are normal. Even if the two conditional probabilities of a correct test result, given either an affected or a normal fetus, were 99.5 percent, the probability of an affected child, given a positive test result, would be only 18 percent. [...]*
>
> *Thus, if we substitute "The fetus is normal" for $H_0$, and "The test result is positive (i.e. indicating Down's syndrome)" for $D$, we have $P[D \mid H_0] = 0.005$, which means $D$ is a significant result, while $P[H_0 \mid D] = 0.82$ (i.e., $1 - 0.18$).*

Pauker, S. P., & Pauker, S. G. (1979). ***The amniocentesis decision: An explicit guide for parents.*** In C. J. Epstein, C. J. R. Curry, S. Packman, S. Sherman, & B .D . Hall (Eds.), ***Birth defects: Original article series: Volume 15. Risk, communication, and decision making in genetic counseling*** (pp. 289-324). New York: The National Foundation.

# Is Rejecting $H_0$ Trivial?

Tukey and others have argued that a null hypothesis of the form $H_0 : \mu = \mu_0$ is never true in practice; at some point in the decimal expansion of the null value and the (unknown) true value, a difference will occur with probability 1.

Therefore, testing to reject $H_0$ is pointless: a significant result can always be obtained if the sample size $n$ is chosen large enough. Conversely, a failure to reject $H_0$ simply means that the sample size wasn't large enough.

Hence, if $H_0$ is rejected, that does not show that $H_0$ was false (by the above argument this was obvious anyway) but only that the researcher was clever enough to put together a test with enough power to detect this.

One solution for this problem is to avoid two-tailed tests entirely.

# Interpretation of the $P$-value

Suppose that you perform a Fisher test comparing a mean $\mu$ to a null value $\mu_0$

$$H_0 : \mu \leq \mu_0.$$

After obtaining your data and from it the sample mean $\overline{X}$, you find that $\overline{X} > \mu_0$ and calculate a $P$-value of 0.3%.

Which of the following statements are correct?

(1) There is at least a 99.7% chance that $H_0$ is true.

(2) There is at least a 0.3% chance that $H_0$ is true.

(3) If $H_0$ were true, there would be at most a 0.3% chance of obtaining a value of $\overline{X}$ equal or greater to the one measured.

(4) If $H_0$ were false, there would be at least a 99.7% chance of obtaining a value of $\overline{X}$ equal to the one measured or greater.