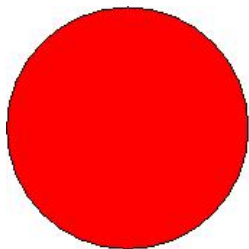


WHY THE EMPIRICAL SCIENCES NEED STATISTICS SO DESPERATELY

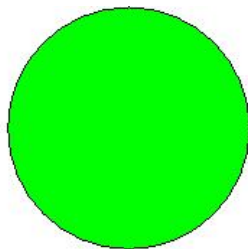
Olle Häggström

6ECM

Krakow, June 2012



80%



20%

What is the geodesic distance, in kilometers,
between Krakow and Gothenburg?

Science is what we have learned about how to keep from fooling ourselves.

Richard Feynman

Coin tossing

A coin is said to be **fair** if

$$\mathbb{P}(\text{heads}) = \mathbb{P}(\text{tails}) = \frac{1}{2}.$$

Suppose we toss a coin 10 times, obtaining 8 heads and 2 tails.

Is this a statistically significant deviation from fairness?

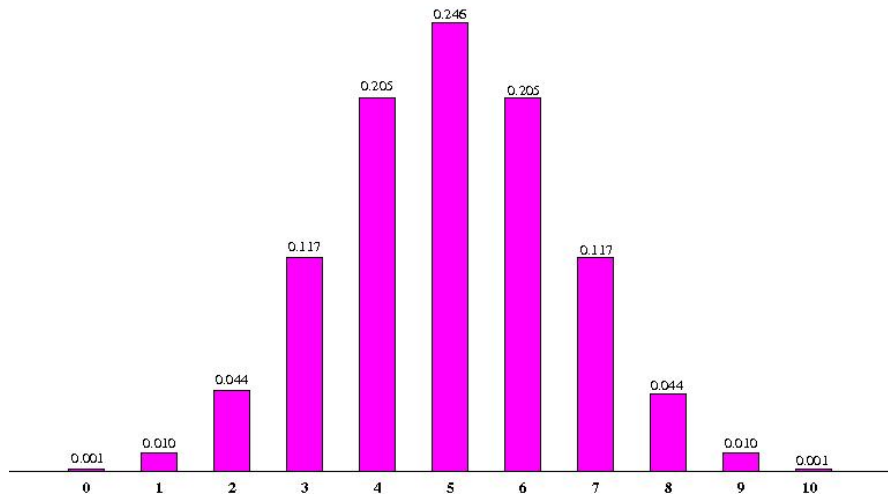
Statistical significance is about detecting deviations from what we call **the null hypothesis** H_0 .

In this case, $H_0 = \{\text{the coin is fair}\}$.

The **p-value** is defined (usually, in simple situations) as the probability of obtaining as extreme data as we actually did get, *given that the null hypothesis is true*.

If the p-value falls below a prescribed threshold (usually 0.05), the result is said to be **statistically significant**.

Under the null hypothesis $H_0 = \{\text{the coin is fair}\}$,
the number of heads in 10 tosses is $\text{Bin}(10, \frac{1}{2})$.



The fallacy of the transposed conditional...

...is the tempting mistake of confusing $\mathbb{P}(\text{data}|H_0)$ with $\mathbb{P}(H_0|\text{data})$.

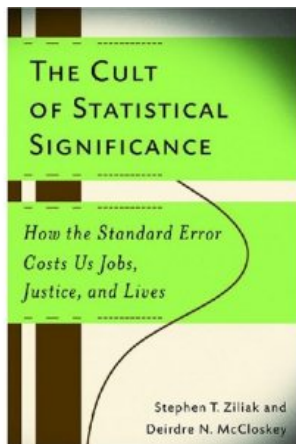
In general, $\mathbb{P}(A|B) \neq \mathbb{P}(B|A)$. Consider for instance the experiment of picking a male Swedish citizen NN at random. Let $A = \{\text{NN is prime minister}\}$ and $B = \{\text{NN is bald}\}$. Then

$$\mathbb{P}(A|B) \approx \frac{1}{500\,000},$$

whereas

$$\mathbb{P}(B|A) = 1.$$





S.T. Ziliak and D. McCloskey (2008)

*The Cult of Statistical Significance:
How the Standard Error Costs Us
Jobs, Justice and Lives*

Suppose a new drug for reducing blood pressure is being tested and that the fact of the matter is that the drug does have a positive effect (as compared to placebo) but that the effect is so small that it is of no practical relevance to the patients' health or well-being.

If the clinical study involves sufficiently many patients, then the effect will nevertheless with high probability be detected, and statistical significance will be obtained.

To commit **sizeless science** is to be so obsessed with *statistical* significance that you forget to ask yourself whether the detected effect is large enough to be of *subject matter* significance.

Ziliak and McCloskey found that among the 369 papers involving regression analysis in the *American Economic Review* between 1980 and 1999, 276 committed sizeless science.

Daniel Carat, debating clinical evidence for the effects of antidepressants, in the *New York Review of Books*, August 18, 2011:

“Even including the negative data, all twelve antidepressants were statistically superior to placebo. Furthermore, [...] Dr. Kirsch’s judgement about the lack of ‘clinical’ significance was based on an arbitrary cut-off point suggested by the UK’s National Institute for Health and Clinical Excellence, a cut-off point with little if any scientific validity.”

Swedish national public radio in the morning of February 16, 2010:

“The critical examination of the United Nations climate panel IPCC and their work continues, and now it is one of the leading characters who says in an interview with British BBC that he is no longer so sure about global warming. It is none other than the director of CRU at the University of East Anglia, professor Phil Jones, who says that there are no statistically significant proofs of global warming during the last 15 years.”

From the BBC interview with Phil Jones, February 13, 2010:

Q: Do you agree that from 1995 to the present there has been no statistically-significant global warming?

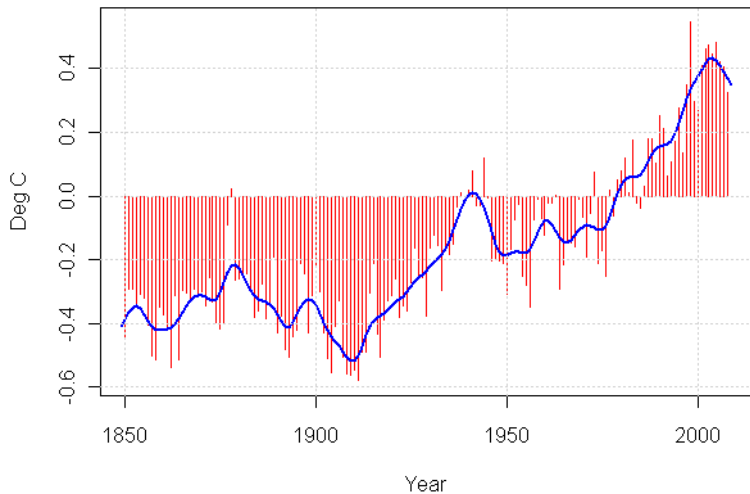
A: Yes, but only just. I also calculated the trend for the period 1995 to 2009. This trend (0.12C per decade) is positive, but not significant at the 95% significance level. [...] Achieving statistical significance in scientific terms is much more likely for longer periods, and much less likely for shorter periods.

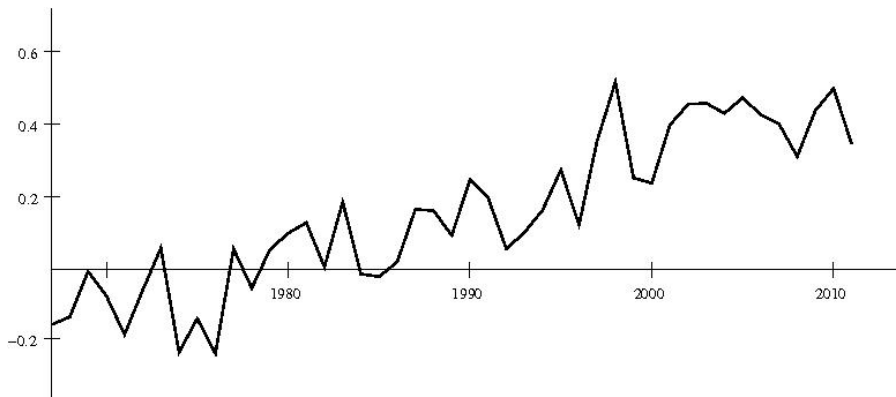
[...]

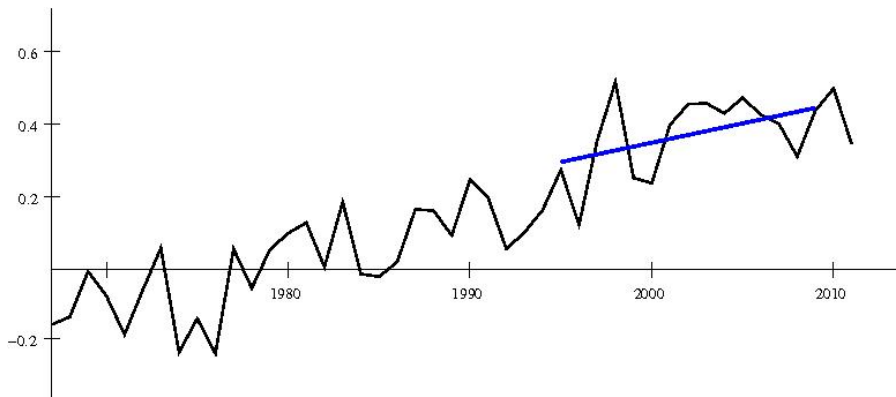
Q: How confident are you that warming has taken place and that humans are mainly responsible?

A: I'm 100% confident that the climate has warmed. As to the second question, I would go along with IPCC Chapter 9 – there's evidence that most of the warming since the 1950s is due to human activity.

HadCRUT Annual Temperature Anomaly







The Duhem–Quine thesis: hypotheses cannot be tested in isolation.

$$H_0 = H_0^0 \cap H_0^1 \cap H_0^2 \cap \dots$$

HadCRUT3 global mean temperature, 1995-2009

	annual	biannual	monthly
slope ($^{\circ}\text{C}/\text{yr}$)	0.0108	0.0108	0.0107
RMS residual	0.0927	0.1011	0.1181
p-value	0.072	0.017	0.00000040

What to do?

Richard Gill, *Statistica Neerlandica* 2011:

“Real world problems are often brought to a statistician because the person with the question, for some reason or other, thinks the statistician must be able to help them. The client has often already left out some complicating factors, or made some simplifications, which he thinks that the statistician doesn't need to know. [...] The first job of the statistical consultant is to undo the pre-processing of the question which his client has done for him.”