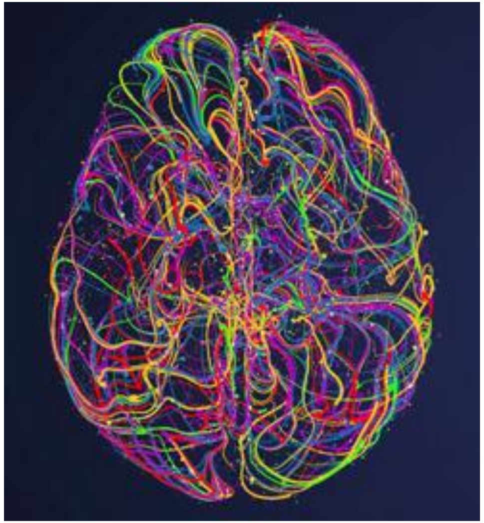


How Do We Test Our Hypotheses?

A **Bayesian** Approach

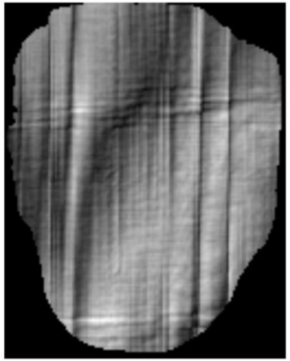
Emanuele Olivetti
olivetti@fbk.eu

NeuroInformatics Lab (NILab)
Fondazione Bruno Kessler (FBK), Trento, Italy
Centre for Mind/Brain Sciences, Univ. of Trento, Italy

FONDAZIONE BRUNO KESSLER
ICT
CENTER FOR COMMUNICATION AND INFORMATION TECHNOLOGY
UNIVERSITÀ DEGLI STUDI DI TRENTO

Motivating Example: can we *decode* Faces?

Face



Scrambled Face

Testing Hypotheses

- Is there *information* about [mental process] within brain data?
  H_1: **yes there is information** H_2: **there is no information**

- Is it possible to *decode* stimuli from brain data?
  H_1: **yes, decoding works** H_2: **no, decoding does not work**

- Can my *classifier* discriminate the category of the stimulus?
  H_1: **yes it can** H_2: **no it cannot**

Overview

Testing Hypotheses

- Classical/Frequentist
  - Significance Testing
  - Hypothesis Testing
- Bayesian
  - Bayesian Hypothesis Testing (BHT)

Classical / Frequentist

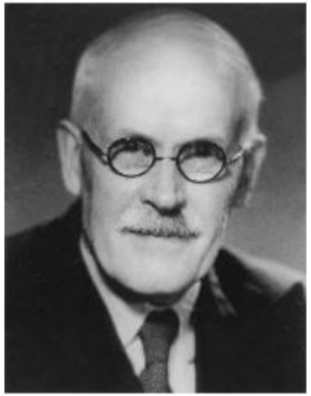
R.A.Fisher

vs.

J.Neyman

E.Pearson

VS.



Bayesian

H.Jeffreys

Ronald Aylmer Fisher (1890-1962)

R.A. Fisher: Significance Testing [Fisher, 1955]

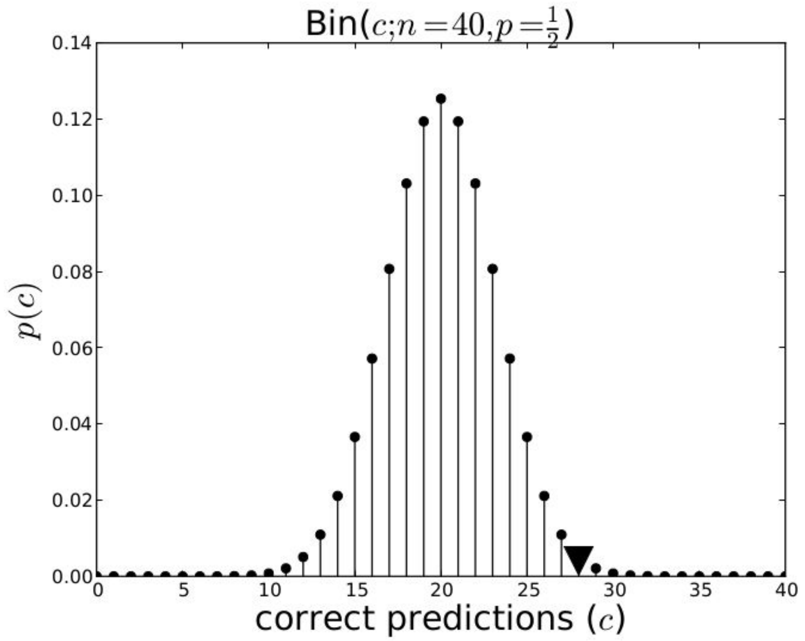
Inductive inference: from sample to population.

The Fisher's recipe

1. Set up H_0, the *null hypothesis* to be disproved with the experiment.
2. Choose a way to summarize of the data into a number, the *test statistic* T.
3. Derive the *null distribution* $p(T; H_0)$
   - Analytically.
   - By resampling.
4. Execute the experiment, collect the data and compute the actual value (T_{obs}).
5. Report the p-value $= p(T \geq T_{obs}; H_0)$ as a measure of evidence against H_0.

Fisher's Significance Testing: Example

1. **Null Hypothesis** H_0: *"the classifier predicts at chance level"*
2. **Test Statistic** T = *the number of correct predictions c* (test set size: $n = 40$).
3. **Null Distribution** $p(c; n = 40, p = \frac{1}{2}) = \mathrm{Bin}(c; n = 40, p = \frac{1}{2})$
4. **Experiment result:** $c_{obs} = \mathbf{28}$.
5. p-value $= p(T \geq 28; n = 40, p = \frac{1}{2}) = \sum_{t=28}^{40} \mathrm{Bin}(T = t; n = 40, p = \frac{1}{2}) = \mathbf{0.008}$



Fisher: Interpretation of the *p*-value

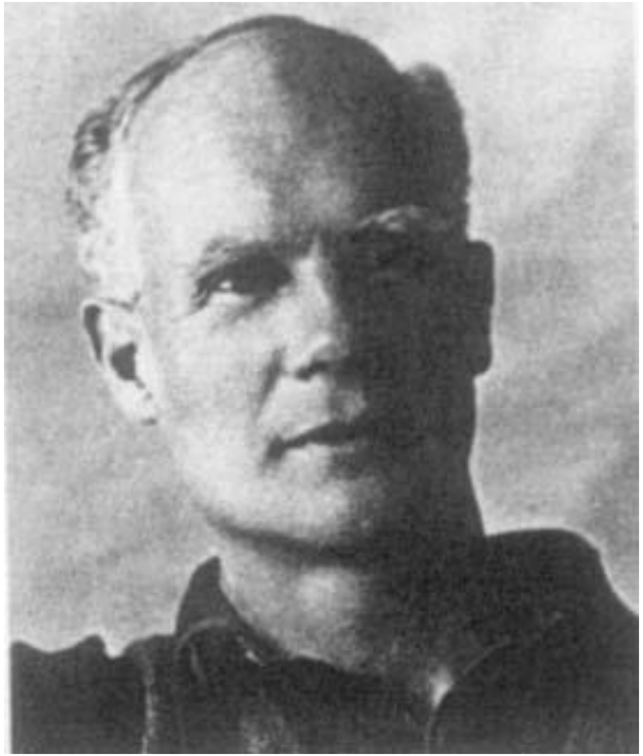
Interpretation:

- A low *p*-value means that H_0 may not be a good model.
- If H_0 is rejected, nothing is said about *what should be accepted*.

Fisher R.A., Statistical Methods for Research Workers, 1958

*"**Personally**, the writer prefers to set a low standard of significance at the 5 percent point..."*

Jerzy Neyman(1894-1981), Egon Pearson(1895-1980)



J.Neyman-E.Pearson: Hypothesis Testing

Inductive behaviour: adjusting behaviour under limited information

Neyman-Pearson recipe

1. Set up **two** *complementary* hypotheses: H_0 (*null*) and H_1 (*alternative*).
2. Choose a way to summarize of the data into a number, the *test statistic* T.
3. Derive/obtain $p(T; H_0)$ and $p(T; H_1)$
4. Decide $\alpha = p(\text{reject } H_0; H_0 \text{ true})$. Decide n (sample size). Compute $\beta = p(\text{reject } H_1; H_1 \text{ true})$.
5. Compute the for *rejection region*(s) $\mathcal{R}$ for T.
6. Run the experiment and compute the observed T_{obs}.
7. Reject H_0 and accept H_1 if $T_{obs} \in \mathcal{R}$. Or viceversa.

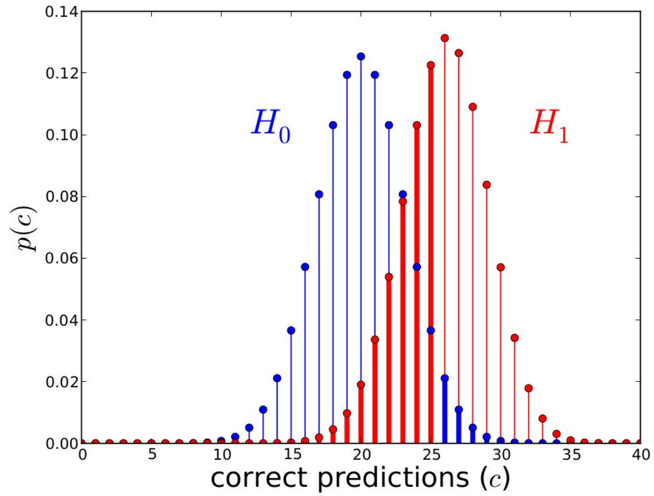
[Neyman and Pearson, 1933, Lehmann and Romano, 2005]

Neyman-Pearson: Example

1. H_0: *"the classifier predicts at **chance level**"*
   H_1: :*"the classifier predicts **better** than chance level."*
2. T = *the number of correct predictions c.*
3. H_0: $\mathrm{Bin}(c; n = 40, \mathbf{p} = \frac{1}{2})$
   H_1: $\mathrm{Bin}(c; n = 40, \mathbf{p_{MLE}} = \mathbf{0.7})$
4.

α	β	$\mathcal{R}_{c\geq}$
0.215	0.032	23
0.134	0.063	24
0.077	0.115	25
0.040	**0.193**	**26**
0.019	0.297	27

5. Rejection region $\mathcal{R} = \{c \geq 26\}$.
6. Experiment: $c_{obs} = 28$.
7. Report: H_0 **rejected** ($\alpha = 0.04$, $\beta = 0.193$, power=0.807).



FONDAZIONE BRUNO KESSLER
ICT CENTER FOR COMMUNICATION AND INFORMATION TECHNOLOGY

The Anonymous Hybrid (~1950- today 😿)

The anonymous hybrid: Fisher + Neyman-Pearson

Anonymous Hybrid's recipe

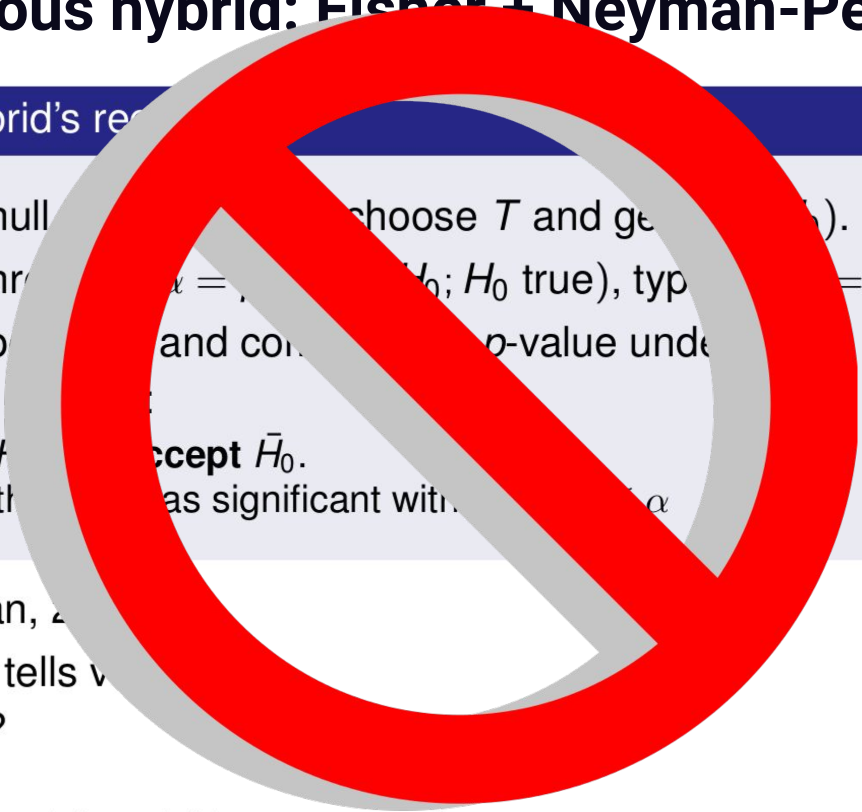
1. Set up the null hypothesis H_0, choose T and get $p(T; H_0)$.
2. Choose a threshold $\alpha = p(\text{reject } H_0; H_0 \text{ true})$, typically $\alpha = 0.05$
3. Run the experiment and compute the *p*-value under H_0.
4. If *p*-value $\leq \alpha$, then:
   - **Reject** H_0 and **accept** $\bar{H}_0$.
   - Report the result as significant with *p*-value $\leq \alpha$

Issues [Goodman, 2008]:

- α without β tells very little.
- What is $\bar{H}_0$?
  - $p \neq \frac{1}{2}$?
  - The binomial model is not correct?

FONDAZIONE BRUNO KESSLER ICT CENTER FOR COMMUNICATION AND INFORMATION TECHNOLOGY

The anonymous hybrid: Fisher + Neyman-Pearson



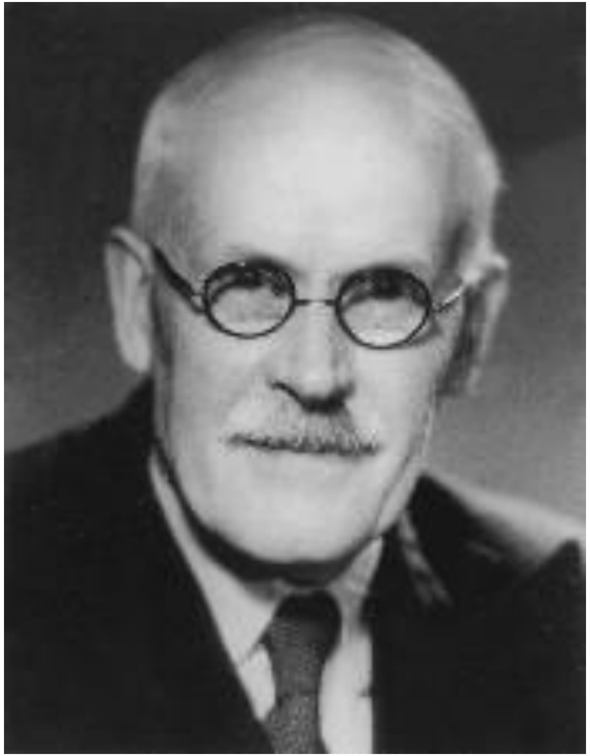
Anonymous Hybrid's re

1. Set up the null … choose T and ge…).
2. Choose a thr… $\alpha =$ … H_0; H_0 true), typ… $= 0.05$
3. Run the exp… and co… p-value unde…
4. If p-value $\leq$ … :
   - **Reject** … **ccept** $\bar{H}_0$.
   - Report th… as significant wit… α

Issues [Goodman, …

- α without β tells v…
- What is $\bar{H}_0$?
  - $p \neq \frac{1}{2}$?
  - The binomial model is not correct?

Bayesian Hypothesis Testing



Harold Jeffreys (1891 - 1989)

H.Jeffreys: Bayesian Hypothesis Testing

Bayesian recipe [Jeffreys, 1961, Kass and Raftery, 1995]

1. Set up *two (or more)* mutually exclusive hypotheses: H_1 and H_2.
2. Quantify *prior probabilities* $p(H_1)$ and $p(H_2)$ from current knowledge.
3. Model the *likelihood of the data*: $p(\text{data}|H_1)$, $p(\text{data}|H_2)$.
4. Run the experiment and collect data.
5. Compute the *posterior probability*

$$p(H_i|\text{data}) = \frac{p(\text{data}|H_i)p(H_i)}{p(\text{data}|H_1)p(H_1) + p(\text{data}|H_2)p(H_2)}$$

6. Report the posterior probabilities (or Bayes Factor).

H.Jeffreys: Bayesian Hypothesis Testing

Bayesian recipe [Jeffreys, 1961, Kass and Raftery, 1995]

1. Set up *two (or more)* mutually exclusive hypotheses: H_1 and H_2.
2. Quantify *prior probabilities* $p(H_1)$ and $p(H_2)$ from current knowledge.
3. Model the *likelihood of the data*: $p(\text{data}|H_1)$, $p(\text{data}|H_2)$.
4. Run the experiment and collect data.
5. Compute the *posterior probability*

$$p(H_i|\text{data}) = \frac{p(\text{data}|H_i)p(H_i)}{p(\text{data}|H_1)p(H_1) + p(\text{data}|H_2)p(H_2)}$$

6. Report the posterior probabilities (or Bayes Factor).

#MATHS

#MATHS

FONDAZIONE BRUNO KESSLER | ICT CENTER FOR COMMUNICATION AND INFORMATION TECHNOLOGY

H.Jeffreys: example

1. Hypotheses:
   - H_1: *"the classifier predicts at chance level"*
   - H_2: *"the classifier predicts better than chance level"*
2. Prior: $p(H_1) = 0.5$, $p(H_2) = 0.5$
3. Data Likelihoods:
   - H_1: $p(c) = \mathrm{Bin}(c|n = 40, p = 0.5)$
   - H_2: $p(c) = \mathrm{Bin}(c|n = 40, p = \pi)$
     $p(\pi) = \mathrm{Uniform}(\pi|0.5, 1)$
4. Run experiment and get the data: $c_{obs} = 28$
5. Posteriors:
   - $p(H_1|\text{data}) = 0.049$
   - $p(H_2|\text{data}) = 0.951$

How to compute the posterior probabilities?

- Prior: $p(H_1) = 0.5$, $p(H_2) = 0.5$
- Compute the data likelihood:
  - $p(\text{data}|H_1) = \text{Bin}(c = 28|n = 40, p = 0.5) = 0.005$
  - $$p(\text{data}|H_2) = \int \text{Bin}(c|n = 40, p = \pi)\,\text{Uniform}(\pi|0.5, 1)d\pi =$$
    $$= \text{...[Monte Carlo]...} = 0.097$$
- Compute the posteriors:
  - $p(H_1|\text{data}) = \frac{p(\text{data}|H_1)p(H_1)}{p(\text{data}|H_1)p(H_1)+p(\text{data}|H_2)p(H_2)} = 0.049$
  - $p(H_2|\text{data}) = \frac{p(\text{data}|H_2)p(H_2)}{p(\text{data}|H_1)p(H_1)+p(\text{data}|H_2)p(H_2)} = 0.951$

How to compute the posterior probabilities?

- Prior: $p(H_1) = 0.5,\ p(H_2) = 0.5$
- Compute the data likelihood:
  - $p(\text{data}|H_1) = \mathrm{Bin}(c = 28|n = 40, p = 0.5) = 0.005$
  - $p(\text{data}|H_2) = \int \mathrm{Bin}(c|n = 40, p = \pi)\,\mathrm{Uniform}(\pi|0.5, 1)d\pi =$

    $= \ldots[\text{Monte Carlo}]\ldots = 0.097$

```
p_data_given_H2 = 0.0
for i in range(10000):
    pi = uniform(low=0.5, high=1.0)
    p_data_given_H2 += binom.pmf(28, 40, pi) * 1.0 / (1.0 - 0.5)

p_data_given_H2 = p_data_given_H2 / 10000
```

- Compute the posteriors:
  - $p(H_1|\text{data}) = \frac{p(\text{data}|H_1)p(H_1)}{p(\text{data}|H_1)p(H_1)+p(\text{data}|H_2)p(H_2)} = 0.049$
  - $p(H_2|\text{data}) = \frac{p(\text{data}|H_2)p(H_2)}{p(\text{data}|H_1)p(H_1)+p(\text{data}|H_2)p(H_2)} = 0.951$

Take-home message

- There is more than one way to test hypotheses.
- Learning about the different frameworks is very interesting: [Christensen, 2005, Berger, 2003].
- Which hypothesis framework then?
  - Long debate...
  - My opinion: use Bayesian.

Take-home message

- There is more than one way to test hypotheses.
- Learning about the different frameworks is very interesting: [Christensen, 2005, Berger, 2003].
- Which hypothesis framework then?
  - Long debate...
  - My opinion: use Bayesian.

THANK YOU!

Berger, J. O. (2003).
Could Fisher, Jeffreys and Neyman Have Agreed on Testing?
Statistical Science, 18(1):1–32.

Berger, J. O. and Sellke, T. (1987).
Testing a Point Null Hypothesis: The Irreconcilability of P Values and Evidence.
Journal of the American Statistical Association, 82(397):112–122.

Christensen, R. (2005).
Testing Fisher, Neyman, Pearson, and Bayes.
The American Statistician, 59(2):121–126.

Fisher, R. (1955).
Statistical Methods and Scientific Induction.
Journal of the Royal Statistical Society. Series B (Methodological), 17(1):69–78.

Gigerenzer, G. (2004).
Mindless statistics.
The Journal of Socio-Economics, 33(5):587–606.

Gigerenzer, G., Krauss, S., and Vitouch, O. (2004).
The null ritual : What you always wanted to know about significance testing but were afraid to ask, pages 392–409.
Sage, Thousand Oaks (CA).

Gill, J. (1999).
The Insignificance of Null Hypothesis Significance Testing.
Political Research Quarterly, 52(3):647–674.

Goodman, S. (2008).
A Dirty Dozen: Twelve P-Value Misconceptions.
Seminars in Hematology, 45(3):135–140.

Goodman, S. N. (1999a).
Toward evidence-based medical statistics. 1: The P value fallacy.
Annals of internal medicine, 130(12):995–1004.

Goodman, S. N. (1999b).

Toward evidence-based medical statistics. 2: The Bayes factor.
Annals of internal medicine, 130(12):1005–1013.

Hubbard, R. and Bayarri, M. J. (2003).
Confusion Over Measures of Evidence (p's) Versus Errors (alpha's) in Classical Statistical Testing.
The American Statistician, 57(3):171–178.

Jeffreys, H. (1961).
Theory of Probability.
Oxford University Press, USA, 3 edition.

Kass, R. E. and Raftery, A. E. (1995).
Bayes Factors.
Journal of the American Statistical Association, 90(430):773–795.

Olivetti, E. (2020).
Multiclass decoding and bayesian hypothesis testing.
in preparation.

Olivetti, E., Greiner, S., and Avesani, P. (2012).
Testing Multiclass Pattern Discrimination.
In *Pattern Recognition in NeuroImaging (PRNI), 2012 International Workshop on*, pages 57–60. IEEE.

Sellke, T., Bayarri, M. J., and Berger, J. O. (2001).
Calibration of p Values for Testing Precise Null Hypotheses.
The American Statistician, 55(1):62–71.

Bayesian Concepts

- $p(X)$ = my degree of belief/knowledge in X.
- Everything is a random variable, including distribution's parameters and hypotheses.
- Prior probabilities must be defined.
- The Bayesian approach provides a belief calculus.

H.Jeffreys: example

1. Hypotheses:
   - H_1: *"the classifier predicts at chance level"*
   - H_2: *"the classifier predicts better than chance level"*
2. Prior: $p(H_1) = 0.5$, $p(H_2) = 0.5$
3. Data Likelihoods:
   - H_1: $c \sim \mathrm{Bin}(n = 40, p = 0.5)$
   - H_2: $c \sim \mathrm{Bin}(n = 40, p = \pi)$

     $\pi \sim \mathrm{Uniform}(0.5, 1)$
4. Run experiment and get the data: $c_{obs} = 28$
5. Posteriors:
   - $p(H_1|\text{data}) = 0.049$
   - $p(H_2|\text{data}) = 0.951$

Bayes Factor: $BF_{21} = \frac{p(\text{data}|H_2)}{p(\text{data}|H_1)} = 19.14$

How to interpret the Bayes Factor?

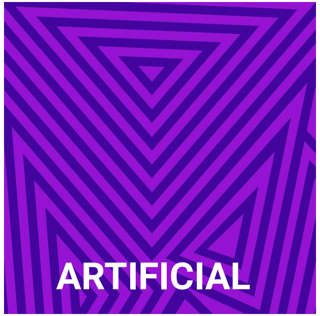
From [Jeffreys, 1961, Kass and Raftery, 1995]

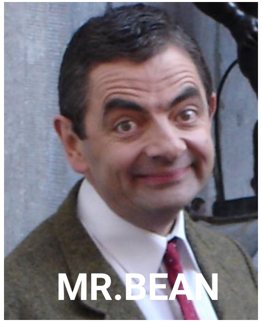
BF_{21}	**Evidence**
< 1	Negative (supports H_1)
1 to 3	Bare Mention
3 to 10	Substantial
10 to 30	Strong
30 to 100	Very Strong
> 100	Decisive

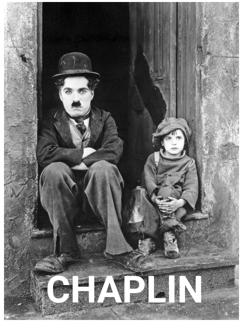
p-value=0.05 $\leftrightarrow$ BF_{21} 2.5-3.4
p-value=0.005 $\leftrightarrow$ BF_{21} 14-26 [Benjamin et al. 2017]

FONDAZIONE BRUNO KESSLER ICT CENTER FOR COMMUNICATION AND INFORMATION TECHNOLOGY

Another Example: can we *decode* short video-clips from MEG?







[Arto Klami et al. ICANN/PASCAL2 Challenge: MEG Mind Reading]

How do we interpret this decoding result?

		Predicted					
		Art.	Nat.	Foo.	Bean	Cha.	
True	Art.	56	55	36	3	0	150
	Nat.	30	96	21	4	0	151
	Foo.	33	22	46	1	0	102
	Bean	4	3	3	95	20	125
	Cha.	1	0	0	11	113	125
		124	176	106	114	123	

Q: can the classifier decode all categories of stimulus?

How do we interpret this decoding result?

		Predicted					
		Art.	**Nat.**	**Foo.**	**Bean**	**Cha.**	
True	**Art.**	56	55	36	3	0	150
	Nat.	30	96	21	4	0	151
	Foo.	33	22	46	1	0	102
	Bean	4	3	3	95	20	125
	Cha.	1	0	0	11	113	125
		124	176	106	114	123	

Bayesian Hypohesis Testing

1. $p(\{\{Art.\},\{Nat.\},\{Foo.\},\{Bean\},\{Cha.\}\}|\mathbf{N}) = 0.735$
2. $p(\{\{Art., Foo.\},\{Nat.\},\{Bean\},\{Cha.\}\}|\mathbf{N}) = 0.264$
3. $p(\{\{Art., Nat.\},\{Foo.\},\{Bean\},\{Cha.\}\}|\mathbf{N}) = 0.001$
4. $p(\{\{Art., Nat., Foo.\},\{Bean\},\{Cha.\}\}|\mathbf{N}) \approx 10^{-10}$

13. $p(\{\{Art., Nat., Foo.\},\{Bean, Cha.\}\}|\mathbf{N}) \approx 10^{-40}$

... (52 hypotheses) ...

[Olivetti et al., 2012, Olivetti, 2020]