Introduction to Bayesian methods

Rita Almeida

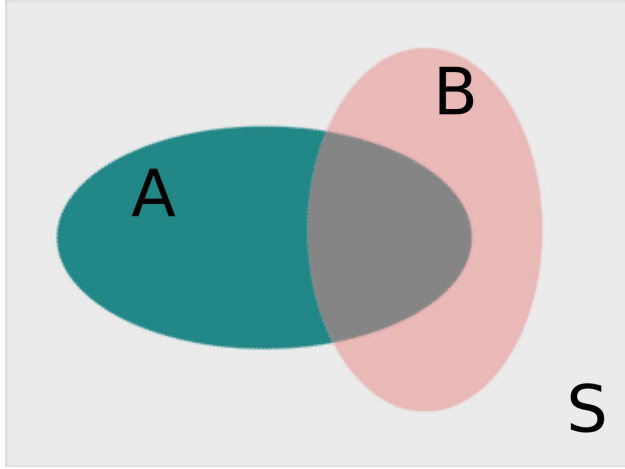
2nd of February, 2016

Bayesian methods in cognitive neuroscience and neuroimaging

- Inference
- Cognitive models
- Pattern Classification
- Image processing: segmentation, normalization
- Probabilistic tractography using DTI
- Dynamical causal modeling
- Model selection and averaging
- Source estimation

Conditional probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$



Example: Clinical depression study

Adapted from DeGroot and Schervish 2002

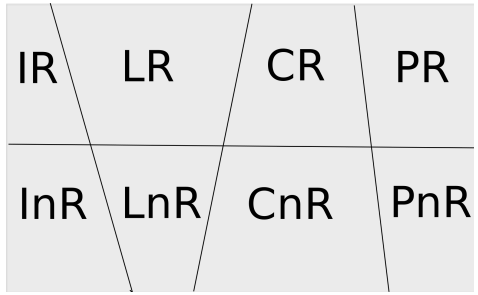
	Imipramine	Lithium	Combination	Placebo	Total
Relapse	18	13	22	24	77
No relapse	22	25	16	10	73
Total	40	38	38	34	150

Question: Select a patient at random. Find that she took the placebo. What is the probability of relapse?

Conditional probability - example

Example: Clinical depression study

Adapted from DeGroot and Schervish 2002



	Imipramine	Lithium	Combination	Placebo	Total
Relapse	18	13	22	24	77
No relapse	22	25	16	10	73
Total	40	38	38	34	150

$$\mathrm{P}(\mathrm{Placebo}) = \frac{34}{150}$$

$$\mathrm{P}(\mathrm{Relapse} \cap \mathrm{Placebo}) = \frac{24}{150}$$

$$\mathrm{P}(\mathrm{Relapse}|\mathrm{Placebo}) = \frac{\mathrm{P}(\mathrm{Relapse} \cap \mathrm{Placebo})}{\mathrm{P}(\mathrm{Placebo})} = \frac{24}{34}$$

Bayes Theorem

$$\mathrm{P}(B|A) = \frac{\mathrm{P}(A|B) \cdot \mathrm{P}(B)}{\mathrm{P}(A)}$$

Demonstration: Departing from the conditional probabilities

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)} \quad (1) \qquad \mathrm{P}(B|A) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(A)} \quad (2)$$

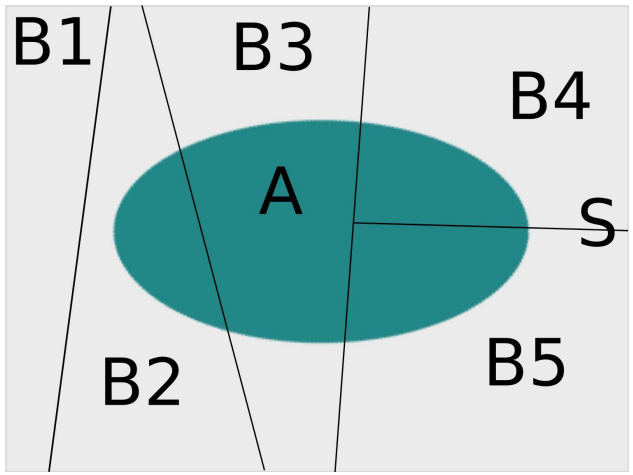
One rewrites (1) to get $\mathrm{P}(A \cap B) = \mathrm{P}(A|B)\mathrm{P}(B)$.

Replacing this in (2) one gets Bayes theorem.

Bayes Theorem

Using the Law of Total Probability:

$$P(B_i|A) = \frac{P(A|B_i) \cdot P(B_i)}{\sum_{j=1}^{k} P(A|B_j) \cdot P(B_j)}$$



Conditional version of Bayes theorem

$$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A)} \quad \rightarrow \quad \boxed{P(B|A, C) = \frac{P(A|B, C) \cdot P(B|C)}{P(A|C)}}$$

Using the conditional version of the law of total probability:

$$P(B_i|A) = \frac{P(A|B_i) \cdot P(B_i)}{\sum_{j=1}^{k} P(A|B_j) \cdot P(B_j)} \rightarrow \boxed{P(B_i|A, C) = \frac{P(A|B_i, C) \cdot P(B_i|C)}{\sum_{j=1}^{k} P(A|B_j, C) \cdot P(B_j|C)}}$$

Bayes Theorem - example

Example: Clinical trial (Adapted from DeGroot and Schervish 2002)

The probability of having a given disease is 1 in 10 000.
There is a diagnostic test available with

- Accuracy in finding that a healthy person is healthy is 0.9 (false positive rate is 0.1)
- Accuracy in finding that a sick person is sick (sensitivity, true positive rate) is 0.9

Given that someone got a positive result on the diagnostic test, what is the probability of this person NOT having the disease?

Bayes Theorem - example

- B_1 event that a person does not have the disease
- B_2 event that a person has the disease
- A event that the diagnostic test is positive

- $\mathrm{P}(B_1) = 1 - 0.0001 = 0.9999$
- $\mathrm{P}(B_2) = 0.0001$
- $\mathrm{P}(A|B_1)$=0.1
- $\mathrm{P}(A|B_2)$=0.9

Answering the question: what is $\mathrm{P}(B_1|A)$?

$$\mathrm{P}(B_1|A) = \frac{\mathrm{P}(A|B_1) \cdot \mathrm{P}(B_1)}{\sum_{j=1}^{2} \mathrm{P}(A|B_j) \cdot \mathrm{P}(B_j)} = \frac{0.1 \times 0.9999}{0.1 \times 0.0001 + 0.9 \times 0.9999} = 0.9991$$

Interpretation:

- 1 in 10 000 has the disease, but the test is positive for $\approx$ 1 in 10 persons.
- On average, for 1000 persons obtaining a positive test 1 has the disease.

Bayes theorem - continuous random variables

Events $\rightarrow$ Random variables

Example of clinic trial with random variables:
$Y = 1$ if the person has a disease
$Y = 0$ otherwise
$X = 1$ if the test is positive
$X = 0$ otherwise

For continuous variables:

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}, \qquad p(x) = \int p(x|y)p(y)dy$$

Frequentist versus Bayesian approach

- Bayesian approaches are increasingly used.
  - In particular in cognitive neuroscience and neuroimaging.
- Introductory statistics courses focus on frequentist methods.
- Comparing the approaches helps understanding.
  - Interpretation of probability is different.

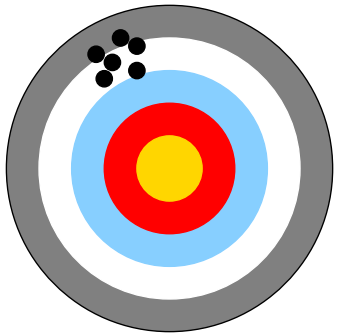
Frequentist approach

- Probabilities are relative frequencies on the limit of very many (hypothetical) replications.
  - Example: probability of obtaining tails when you toss a coin.
- Parameters characterizing the probability distribution of a variable are fixed, unknown constants. Not random variables and so they don't have associated probabilities.
- Statistical procedures are designed to have desirable long-run performance.

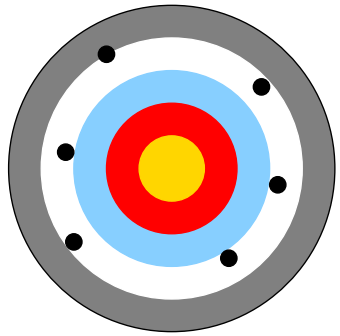
Frequentist approach - estimation example

- Estimate the mean height of a population, assumed normally distributed.
- The sample mean ($\overline{Y} = \sum_{i=1}^{n} Y_i$) is an unbiased estimator with minimum variance.

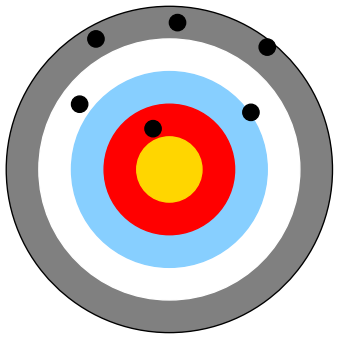
high bias, low variance



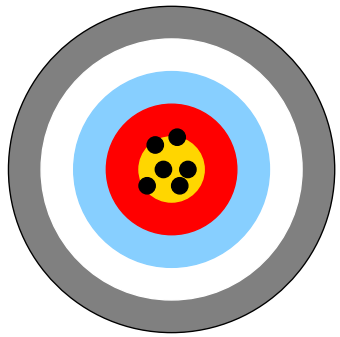
low bias, high variance



high bias, high variance

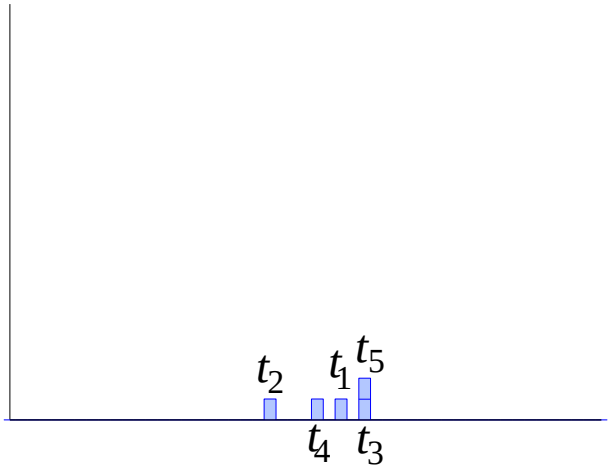


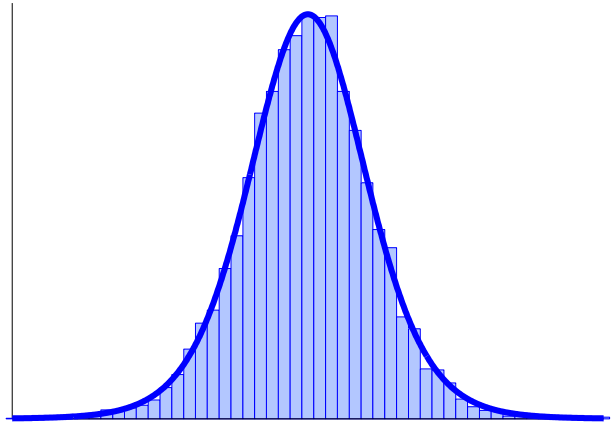
low bias, low variance



Frequentist approach - testing example

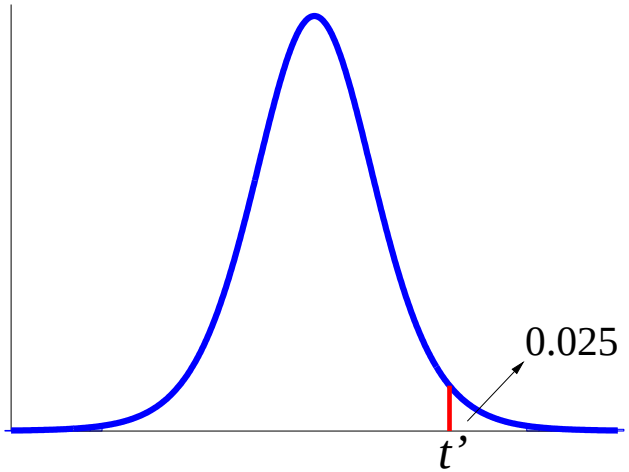
Do populations 1 and 2 have different mean heights?
Assuming $H_0 : \mu_1 = \mu_2$ is true. $t = \frac{\overline{y_1} - \overline{y_2}}{s/\sqrt{n}}$





distribution of t under H_0

We observe t':



- Under H_0 the probability of observing a value of t larger or equal to t' is 0.025: $p(t \geq t'|H_0) = 0.025$
- If we repeat the experiment under H_0, in the long-run we will find 2.5% of the times values of t equal or larger than t'.
- We do not know $p(H_0)$ or $p(H_0|t)$!

Bayesian approach

- Probabilities represent a degree of belief.
- Parameters θ can be associated with a probability distribution - $p(\theta)$
- Statistical procedures update beliefs based on the data.

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$

$$p(\theta|y) \propto p(y|\theta)p(\theta)$$

y is the data

$p(\theta)$ is the prior

$p(y|\theta)$ is the likelihood

$p(\theta|y)$ is the posterior distribution

Bayesian approach - example

- Suppose we have a random sample $y_1, y_2, \ldots, y_n$ from a normal distribution with unknown mean θ and known variance σ^2. The likelihood has the form:

$$p(y_1, y_2, \ldots, y_n|\theta) \propto \exp\left[-\frac{\sum_{i=1}^{n}(y_i - \theta)^2}{2\sigma^2}\right]$$

- Suppose we take the prior distribution to be normal with mean μ and variance ν^2:

$$p(\theta) \quad \text{is} \quad N(\mu, \nu^2)$$

- After we observe the data we update our beliefs about θ:

$$p(\theta|y_1, y_2, \ldots, y_n) \propto p(y_1, y_2, \ldots, y_n|\theta)p(\theta)$$

- The posterior represents all the current information on θ.

Bayesian approach - example

The posterior can be written:

$$p(\theta|y_1, y_2, \ldots, y_n) \propto \exp\left[-\frac{\sum_{i=1}^{n}(y_i-\theta)^2}{2\sigma^2}\right] \quad \exp\left[-\frac{(\theta-\mu)^2}{2\nu^2}\right]$$

It can be shown that $\theta|y_1, y_2, \ldots, y_n \sim N(\mu_1, \nu_1^2)$ with

$$\mu_1 = \underbrace{\frac{\frac{1}{\nu^2}}{\frac{1}{\nu^2}+\frac{n}{\sigma^2}}}_{w} \mu + \frac{\frac{n}{\sigma^2}}{\frac{1}{\nu^2}+\frac{n}{\sigma^2}} \overline{y}_n \quad \text{and} \quad \frac{1}{\nu_1^2} = \frac{1}{\nu^2} + \frac{n}{\sigma^2}$$

$$\mu_1 = w\ \mu + (1-w)\ \overline{y}_n$$

Bayesian approach - example

$$\mu_1 = w \ \mu + (1 - w) \ \overline{y}_n$$

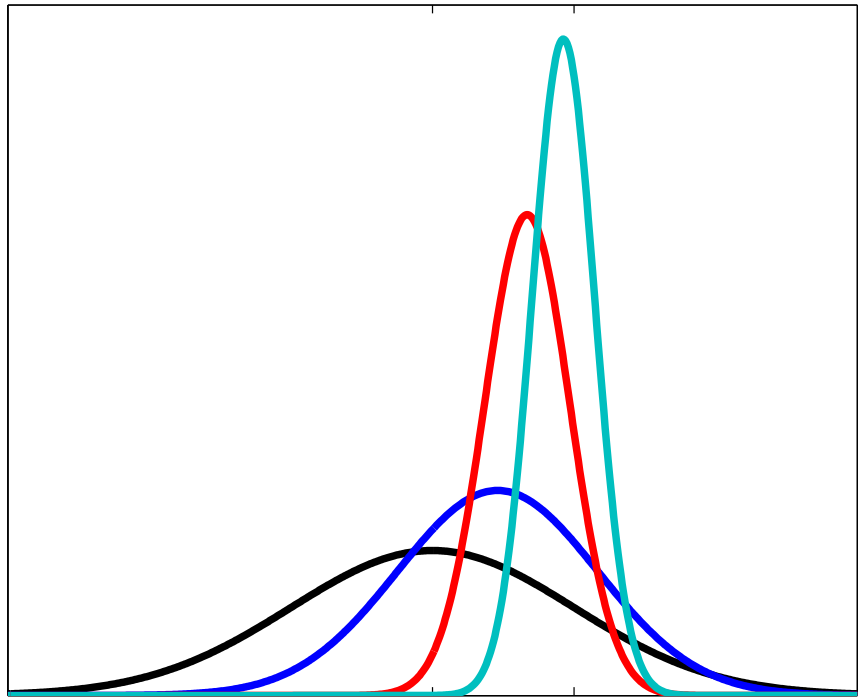
- The mean of the posterior is a weighted average of the sample and prior means
- The weight depends on the precisions.

  Precision is the inverse of the variance.

$$w = \frac{\frac{1}{\nu^2}}{\frac{1}{\nu^2} + \frac{n}{\sigma^2}}$$

- larger $n \rightarrow$ larger weight to sample average
- larger $\sigma^2 \rightarrow$ smaller weight to sample average
- larger $\nu^2 \rightarrow$ larger weight to sample average

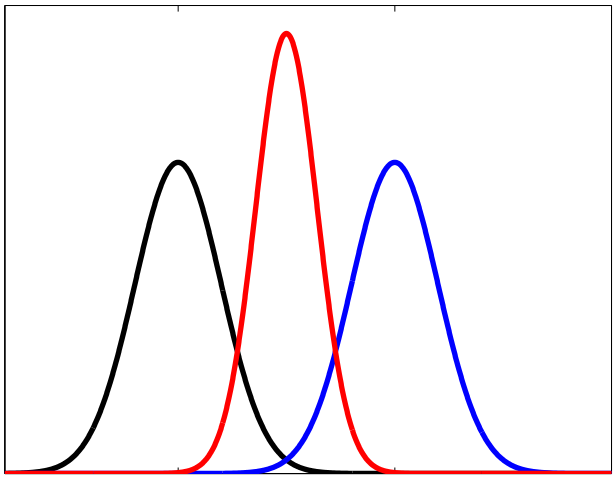
Illustration



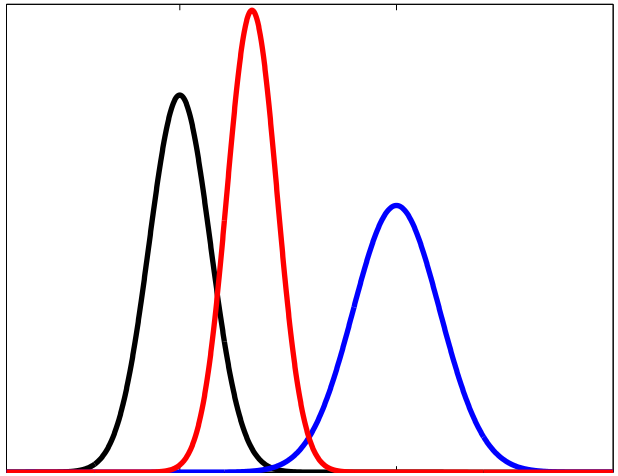
Prior and posterior n=1, n=5 and n=10

prior, likelihood, posterior

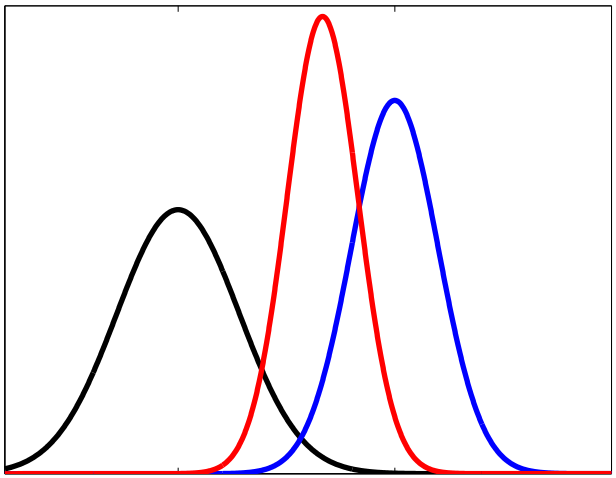
Precision of prior and data are the same



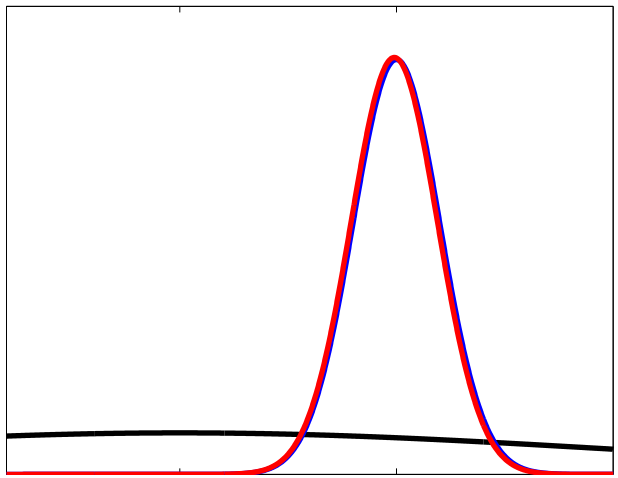
Large precision of prior



Small precision of prior

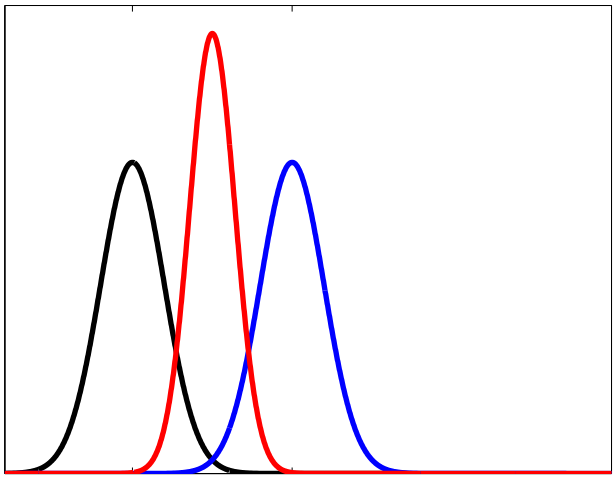


Very small precision of prior

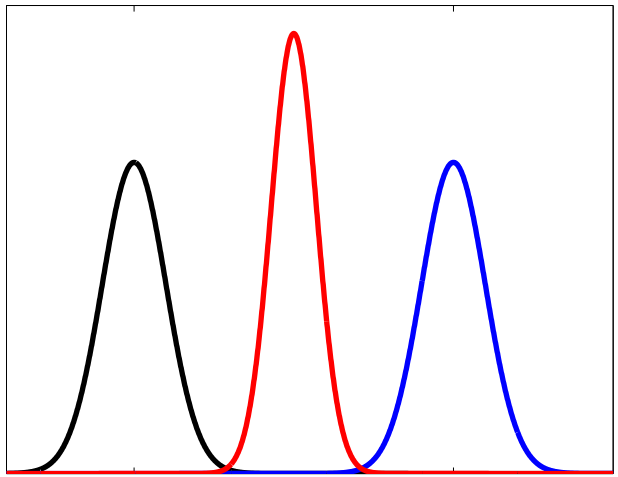


prior fixed, likelihood, posterior

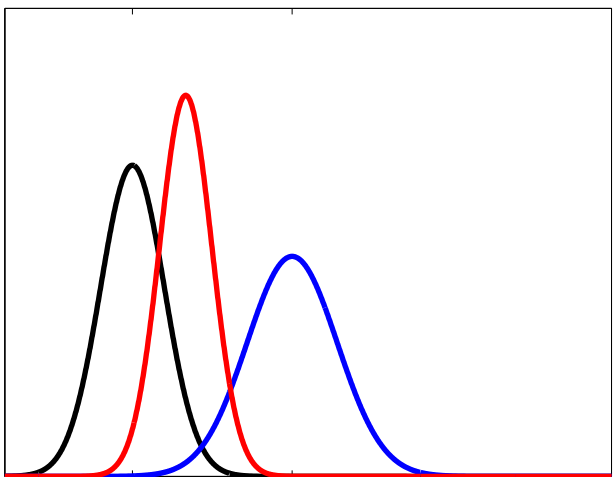
Small sample mean, small variance



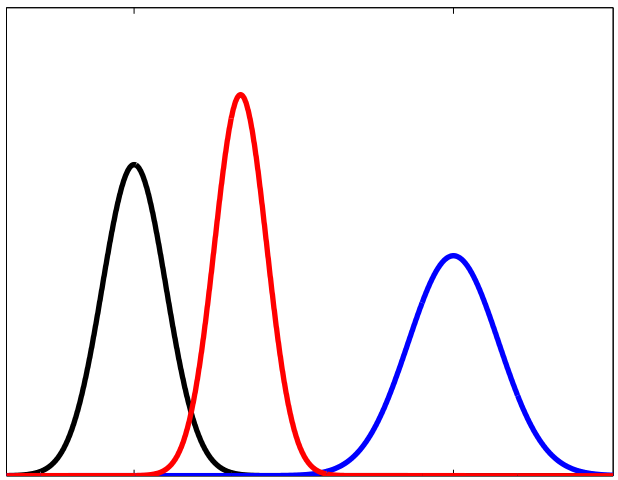
Large sample mean, small variance



Small sample mean, large variance



Large sample mean, large variance



Priors

- Priors describe knowledge before the experiment.
  - Subjective
  - Objective
- Priors can have a large influence on the posterior.
  - Small sample size
- Caution should be used so that priors do not crucially affect the posterior when there are no good reasons for that.
  - Non-informative priors can be used.
- Priors can be chosen for mathematical easiness - *conjugate priors* - the posterior belongs to the same family of distributions as the prior.
- Prior parameters can be estimated from the data - *empirical bayes*.

Posterior

- Bayesian inference is based on the posterior distribution.
- The posterior distribution and functions of the posterior distribution can be very difficult to determine.
- The functions of interest can be obtained by
  - sampling approximations
  - deterministic approximations

Summary

- Frequentists and Bayesian approaches answer different questions.

Disadvantages of Bayesian methods:

- Requires choosing a prior.
- Posterior distributions are difficult to calculate.

Summary

Advantages of Bayesian methods:

- Focus on the data collected, not on the average if one would collect many sets of data.
- Logic / more intuitive
  - Frequentist hypotheses testing does not necessarily do what one would like:
  - Does not give the probability of the null hypothesis.
  - If the sample is large enough a small effect becomes significant.
- Principled way to take into consideration prior knowledge and incorporate new evidence.
- Unified flexible approach

Reproducibility and unlikely hypothesis

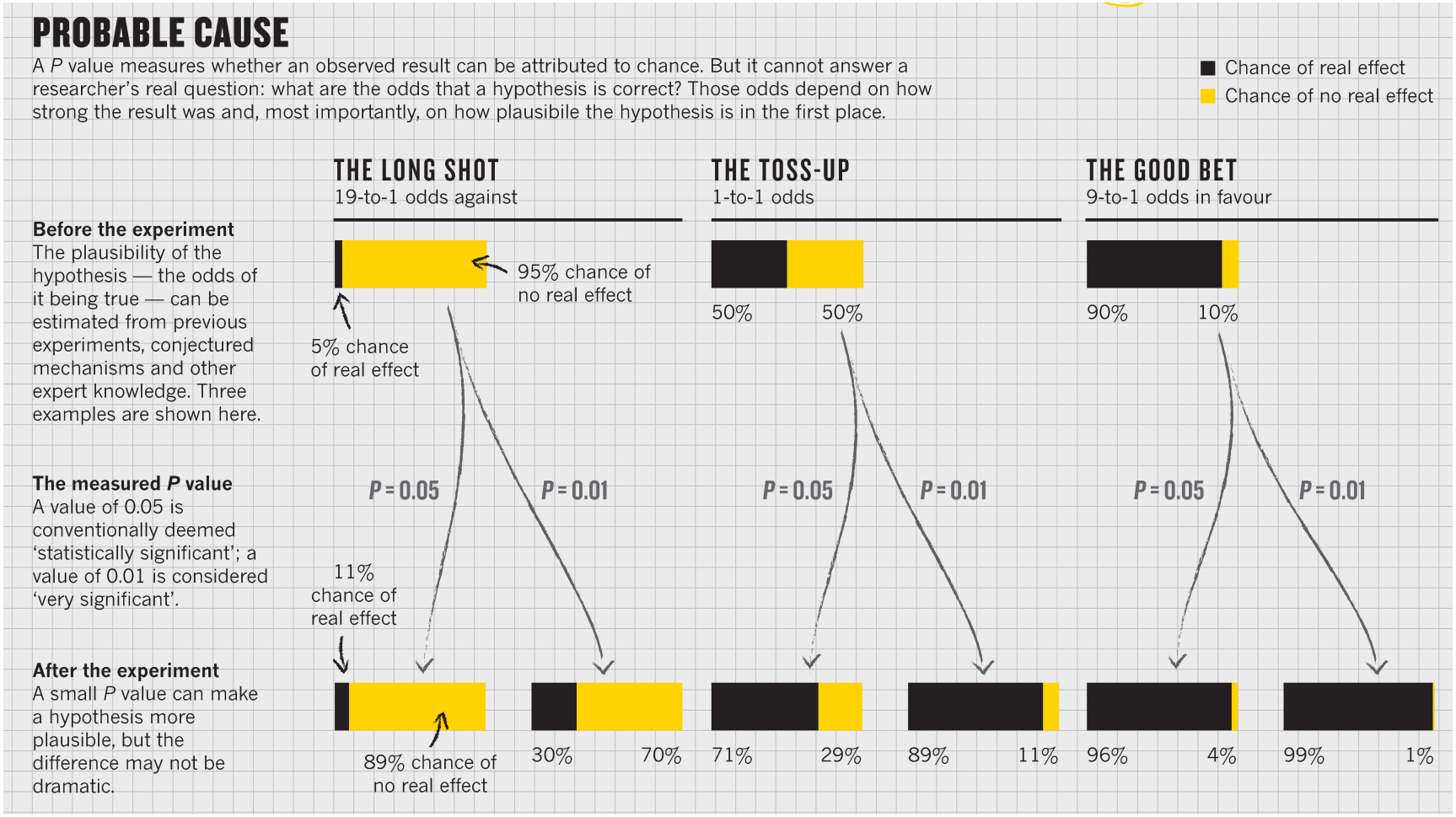


Figure adapted from Nuzzo, Nature 2014.

- The common interpretation of inference and *p* values can be misleading.