Bayes' Theorem

Frequentist vs. Bayesian Statistics

- Common situation in science: We have some data and we want to know the true physical law describing it. We want to come up with a model that fits the data.
  - Example: We look at n=10 random galaxies and find that m=4 are spirals. So what's the true ratio of spirals in the universe, r?

Frequentist:

•There are 'true', fixed parameters in a model (though they may be unknown at times).

•Data contain random errors which have a certain probability distribution (Gaussian for example)

•Mathematical routines analyze the probability of getting certain data, given a particular model (If I flip a fair coin, what's the probability of me getting exactly 50% heads and 50% tails?)

Bayesian:

•There are no 'true' model parameters. Instead all parameters are treated as random variables with probability distributions.

•Random errors in data have no probability distribution, but rather the model parameters are random with their own distributions

•Mathematical routines analyze probability of a model, given some data (If I flip a coin and get X heads and Y tails, what is the probability that the coin is fair?). The statistician makes a guess (prior distribution) and then updates that guess with the data

- Both approaches are addressing the same fundamental problem, but attack it in reverse orders (probability of getting data, given a model, versus probability of a model, given some data). Its quite common to get the same basic result out of both methods, but many will argue that the Bayesian approach more closely relates to the fundamental problem in science (we have some data, and we want to infer the most likely truth)

Bayes' Theorem

•The primary tool of Bayesian statistics. Allows one to estimate the probability of measuring/observing something given that you have already measured/observed some other relevant piece of information

$$p(B \mid A) = p(A \mid B)\frac{p(B)}{p(A)}$$

- P(B|A)=probability of measuring B given A
- P(A|B)=probability of measuring A given B
- P(B)=*prior* probability of measuring B, before any data is taken
- P(A)=*prior* probability of measuring A, before any data is taken

A simple example

- Drug Testing:
  - Let say 0.5% of people are drug users
  - Our test is 99% accurate (it correctly identifies 99% of drug users and 99% of non-drug users)
  - What's the probability of being a drug user if you've tested positive?
  - Our Bayes' theorem reads:

`

$$p(user \mid pos) = p(pos \mid user)\frac{p(user)}{p(pos)} = 0.99 \times \frac{0.005}{0.01 \times 0.995 + 0.99 \times 0.005} = 0.33$$

•p(pos|user)=0.99 (99% effective at detecting users)
•p(user)=0.005 (only .5% of people actually are users)
•p(pos)=0.01*0.995+0.99*0.005 (1% chance of non-users, 99.5% of the population, to be tested positive, plus 99% chance of the users, 0.5% of the population, to be tested positive

•Only a 33% chance that a positive test is correct
•This example assumes we know something about the general population (users vs nonusers), but we usually don't!

Example: Galaxy Populations

- Looked at n=10 random galaxies.
- Found m=4 spirals.
- What's the ratio of spirals in the universe, r? We are introducing an unknown model parameter.
- Bayes' Theorem reads:

$$p(r \mid data) = p(data \mid r)\frac{p(r)}{p(data)}$$

  - p(r|data)=probability of getting r, given our current data (what we want to know)
  - p(data|r)=probability of measuring the current data for a given r
  - p(r)=probability of r before any data is taken (known as a *prior*)
  - p(data)=prior probability of measuring the data. This acts as a normalizing constant, and is defined as

$$p(data) = \int_0^1 p(data \mid r)\,p(r)\,dr$$

  - In other words, it's the probability of getting finding the data considering all possible values of r

$$p(r \mid data) = p(data \mid r)\frac{p(r)}{\int_0^1 p(data \mid r)p(r)dr}$$

- Since there are only two possible measurements (spiral or no spiral), *p(r|data)* is adequately described by a binomial distribution

$$p(data \mid r) = \frac{n!}{m!(n-m)!} r^m (1-r)^{n-m} = \frac{10!}{4!6!} r^4 (1-r)^6$$

- We'll assume that before any data was taken, we figured all possible values of r were equally likely, so we'll set *p(r)=1* (our *prior*)
- *p(data)* is just an integral, and we find

$$\int_0^1 p(data \mid r)p(r)dr = \frac{10!}{4!6!}\int_0^1 r^4 (1-r)^6 dr = \frac{10!}{4!6!}\frac{1}{2310}$$

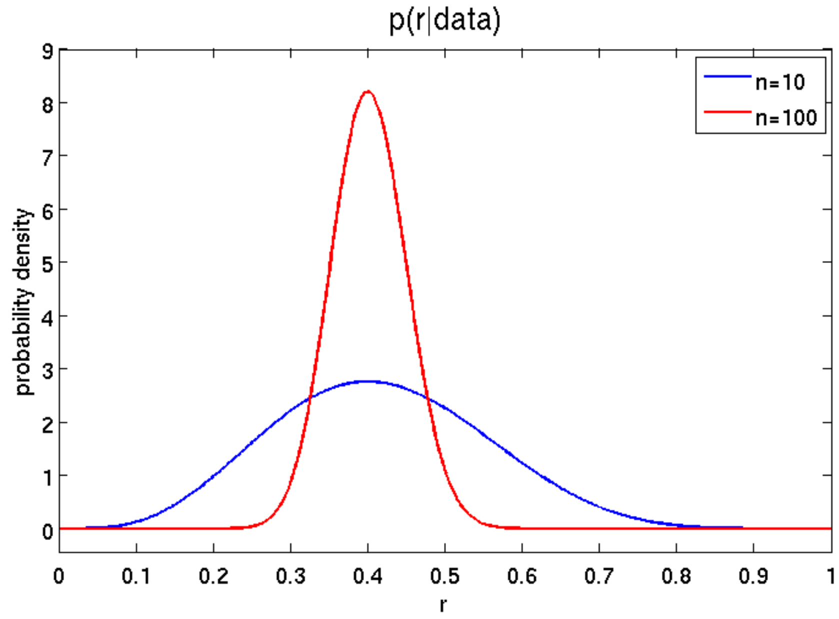
$$p(r \mid data) = p(data \mid r)\frac{p(r)}{\int_0^1 p(data \mid r)p(r)dr}$$

•Putting all this together and simplifying, we get:

$$p(r \mid data) = 2310r^4(1-r)^6$$

•This is just a probability distribution for r, centered around 0.4 as we would expect.

•Also, as expected, more data makes the result more robust (red curve).



The role of priors

- In previous example, we assumed that all values of r were equally likely before we took any data. Often, we'll know something else (apart from the data) which we'll want to incorporate into our prior (physics, models, a hunch, etc.)
- As an example, lets say we run a cosmological simulation which suggests r~0.7$\pm$0.05. We'll use this as our prior, p(r), and estimate it as a Gaussian distribution centered around 0.7, with σ = 0.05.

$$p(data \mid r) = \frac{n!}{m!(n-m)!} r^m (1-r)^{n-m}$$

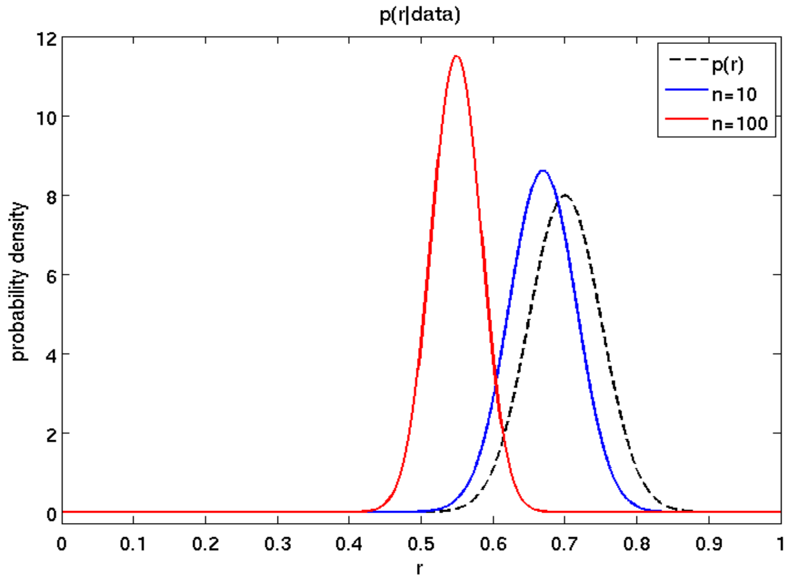
Same as before

$$p(r) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{(r-0.7)^2}{2\sigma^2} \right\}$$

New prior

The role of priors

- Our new distribution



- Notice the profound effect the prior can have on the result. The more data one has, the more the prior is overwhelmed, but it clearly plays a powerful (and potentially dangerous) role in low sample sizes
- Priors can be very controversial, especially when you have no extra information on which to base your prior. Uniform priors, like we originally chose, are considered too agnostic, even though they may seem like the safest approach.