

Introduction to Probabilistic (Bayesian) Modeling and Inference

Marek J. Drużdżel

Wydział Informatyki

Politechnika Białostocka

m.druzdzel@pb.edu.pl

<http://aragorn.wi.pb.bialystok.pl/~druzdzel/>

Session overview

- **Fundamentals:**
 - Joint probability distribution
 - Marginal probability distribution
 - Conditional probability distribution
- Bayes theorem, prior and posterior probability distribution
- Subjectivist Bayesian approach to probability

Generally, an introductory session,
bringing us all to the same page,
hopefully not too boring



What I want you to know after this session?

- Relate what you already know from probability theory to practice.
- Understand the fundamental role of the joint probability distribution.
- Understand the intuition behind Bayes theorem and be able to apply it whenever appropriate.

Motivation

Uncertainty manifested in data

	Age	Sex	Smoking_Status	Lung_Cancer
1	43	Male	Smoker	Yes
2	55	Female	NonSmoker	Yes
3	27	Female	Smoker	No
4	18	Male	NonSmoker	No
5	81	Female	Smoker	No

9873	72	Male	NonSmoker	Yes

Data like the above are not at all atypical.

Some sources of uncertainty:

- Errors in measurement (e.g., cancer misdiagnosed).
- Subjects providing wrong information (e.g., smoking status, age).
- Latent variables that we did not control for (e.g., asbestos exposure).
- Subject selection (possible bias).
- Bad luck.
- ...

Why statistics?

“... in this world nothing can be said to be certain, except death and taxes” --- Benjamin Franklin in a letter to his friend M. Le Roy

(*) The Complete Works of Benjamin Franklin, John Bigelow (ed.), New York and London: G.P. Putnam's Sons, 1887, Vol. 10, page 170

- In other words, “Uncertainty is prominent around us.”
- It is an inherent part of all information and all knowledge.
- We need to deal with uncertainty in empirical work.
- Because this class focuses on analytics, we are going to review some basic tools for looking at data and making inferences from data.

Why statistics 😊?



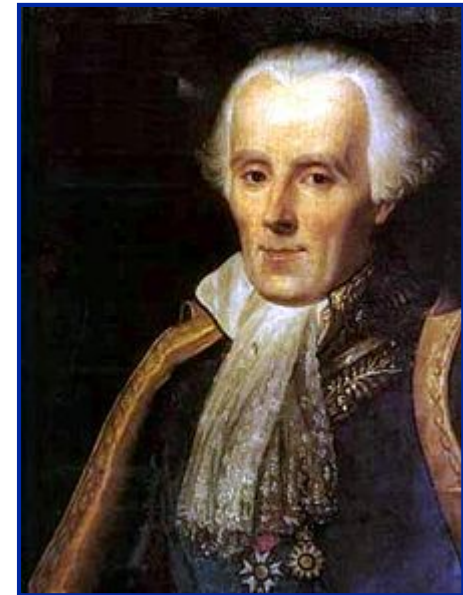
“Our statistician will drop in and explain why you have nothing to worry about.”

Why probability theory and statistics?

“The theory of probabilities is basically only common sense reduced to a calculus.”

(“... la théorie des probabilités n'est, au fond, que le bon sens réduit au calcul.”)

— Pierre-Simon Laplace, “Philosophical Essay on Probabilities” (1814)



Uncertainty manifested in data

Even though a behavior may be unpredictable in the short run, it may have a regular and predictable pattern in the long run.

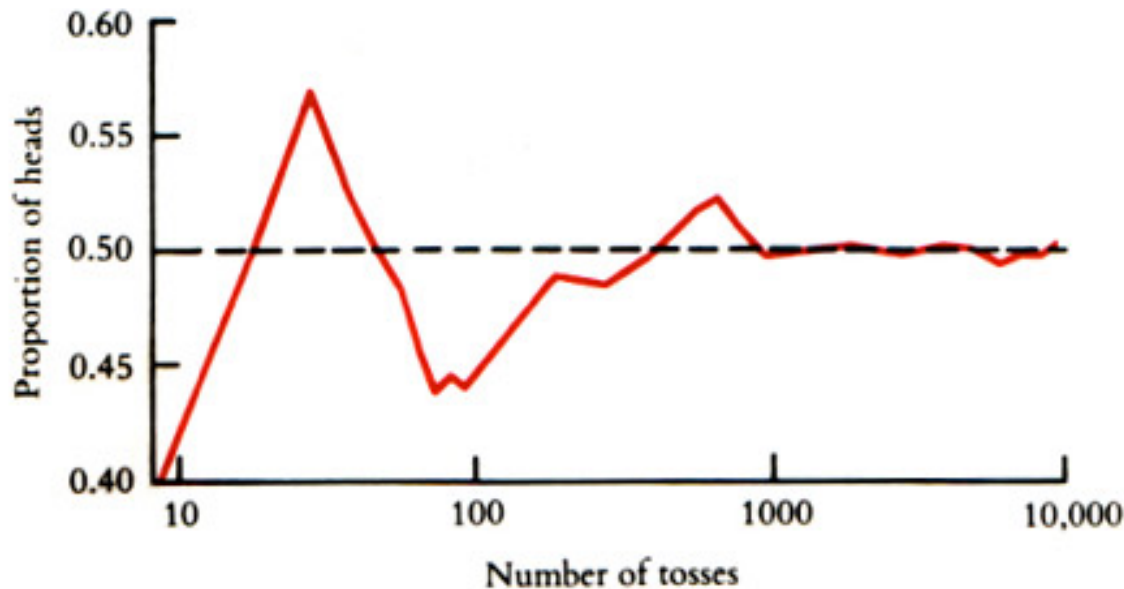


Figure 7.2 Percent of heads versus number of tosses in Kerrich's coin-tossing experiment. [David Freedman et al., *Statistics* Norton, 1978.]



A crash review of probability theory and statistics

Why probability theory and statistics?

- “Statistics is the study of the collection, organization, analysis, and interpretation of data.” Dodge, Y. (2003) *The Oxford Dictionary of Statistical Terms*
- Statistics is **the** mathematical discipline for processing and interpreting data, it is closely related probability theory.
- Departure from probability theory leads to provable anomalies (e.g., “Dutch book” argument).
- All (with some exceptions) knowledge is uncertain and, hence, best expressed by means of probabilities and probability distributions.

Probability

$$P(X|\xi)$$

X is an event (defined on a discrete or continuous domain)
 ξ denotes background information/knowledge

e.g., Probability of heads in
a coin toss



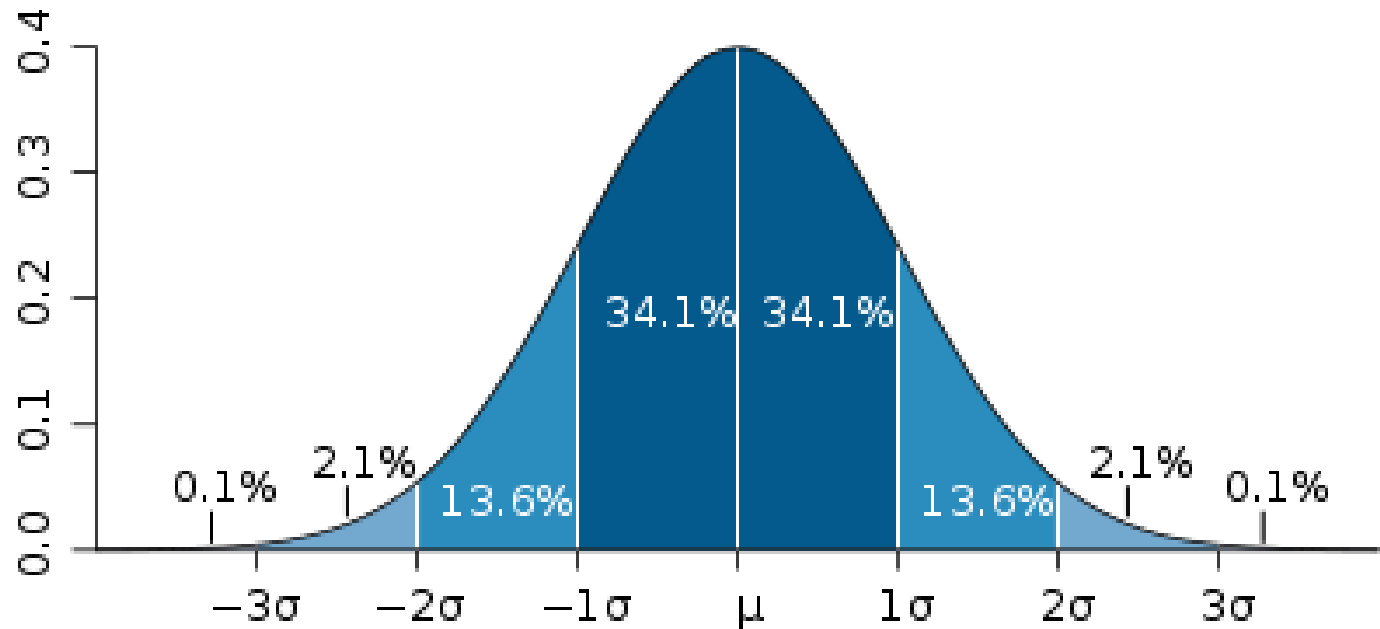
or probability of bodily
temperature being
above 38 degrees



Source: https://en.wikipedia.org/wiki/Coin_flipping

Probability distribution

Expresses the relative probabilities of different values taken by a random variable



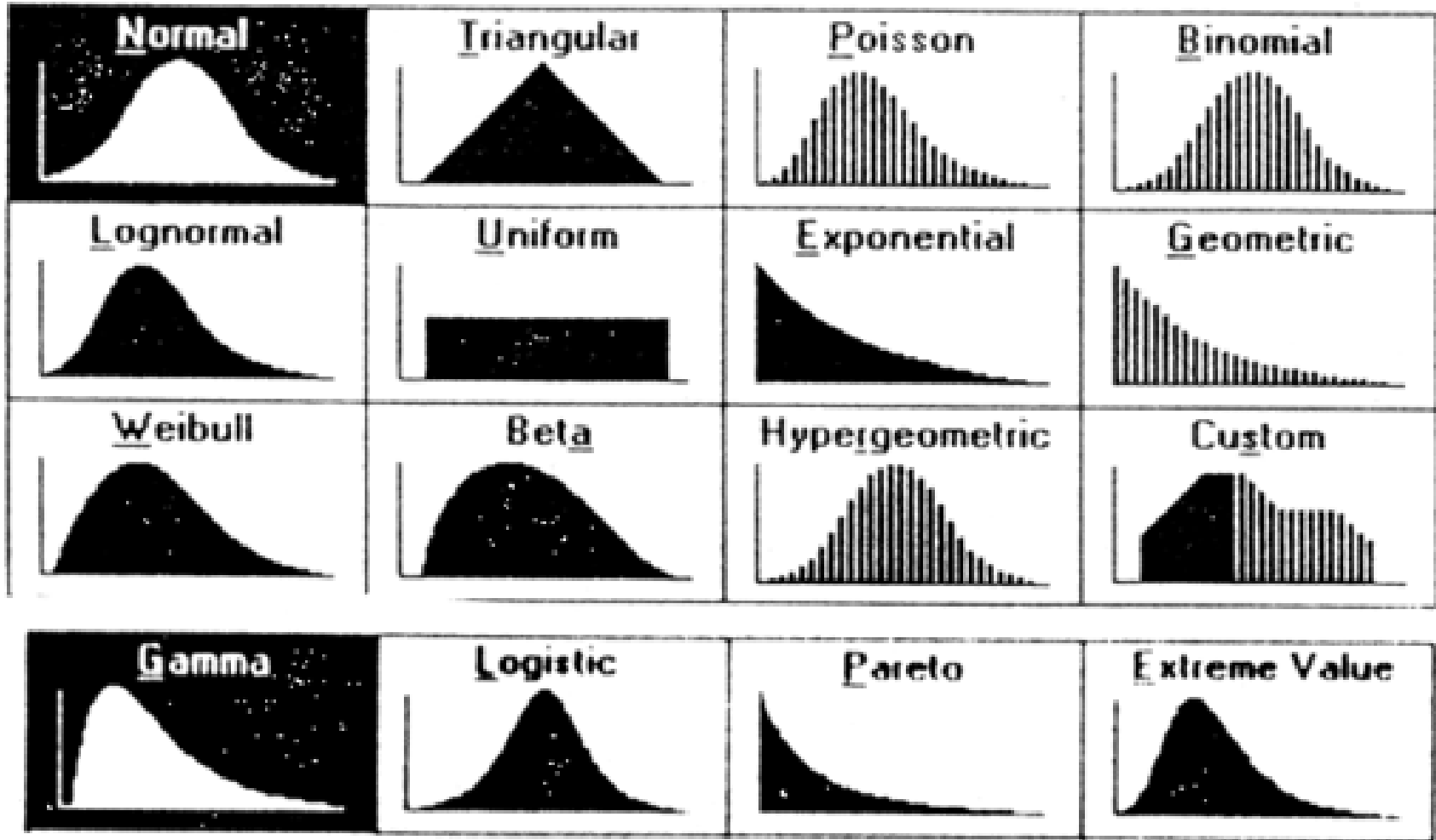
Source: http://en.wikipedia.org/wiki/Probability_distribution

e.g., grade distribution in a university course

Parametric probability distributions

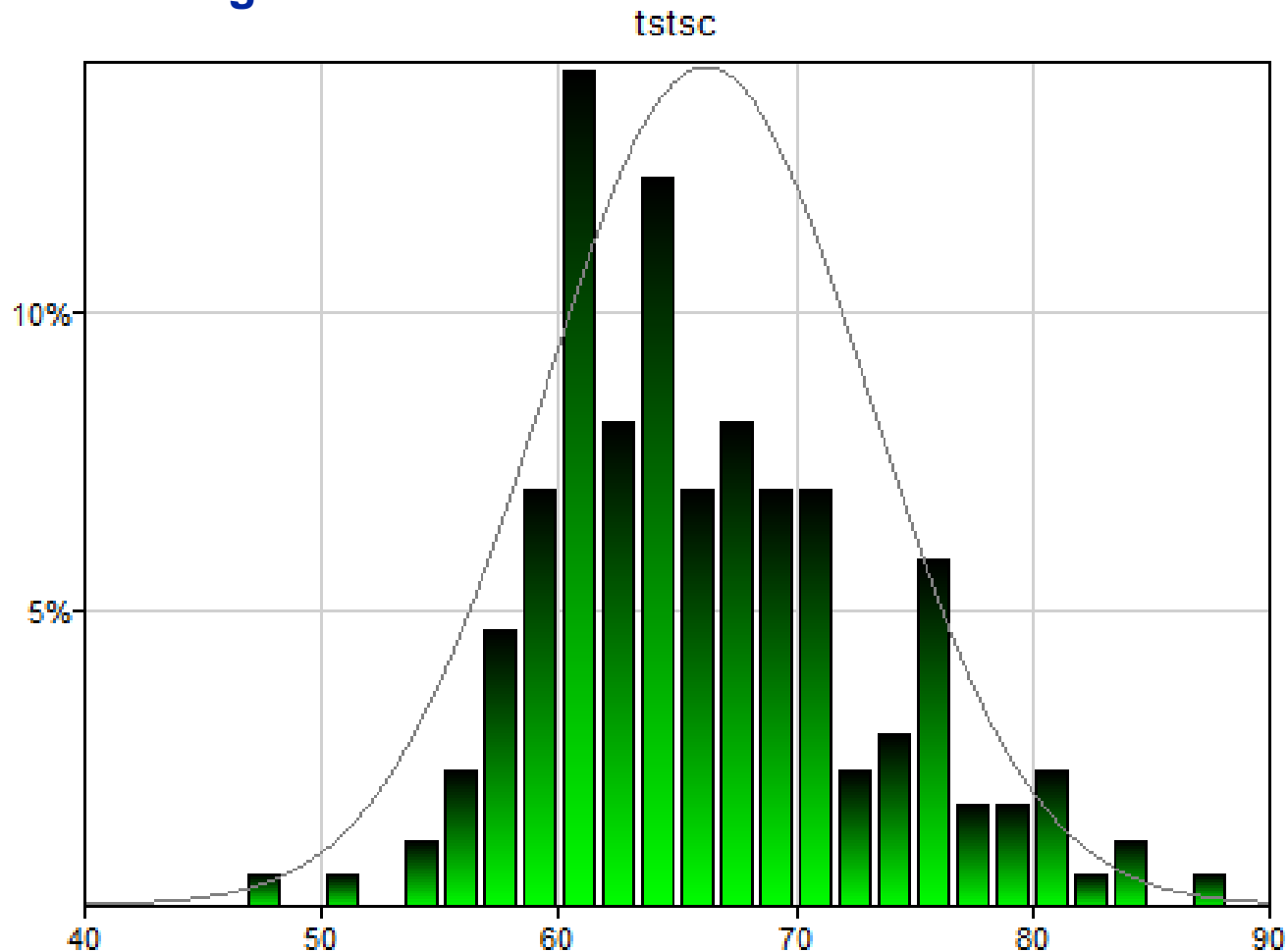
- There is a sizeable set of known/described ways that values of a variable can be distributed.
- Some of these: Normal, Log-Normal, Uniform, Beta, Exponential, Triangular, Bernoulli, Binomial, Weibull, etc.
- Some distributions are very common, e.g., Normal (a.k.a. Gaussian) distribution.
- Explained by the Central Limit Theorem (a.k.a. “order out of chaos”):
 - When you sum infinitely many random variables, the sum is going to be distributed normally.
 - You don’t really need infinitely many: as few as 12 is enough when components are uniform, typically 30 or so gives beautiful Normals.
- There are tests for goodness of fit of data to distributions.

Common parametric probability distributions

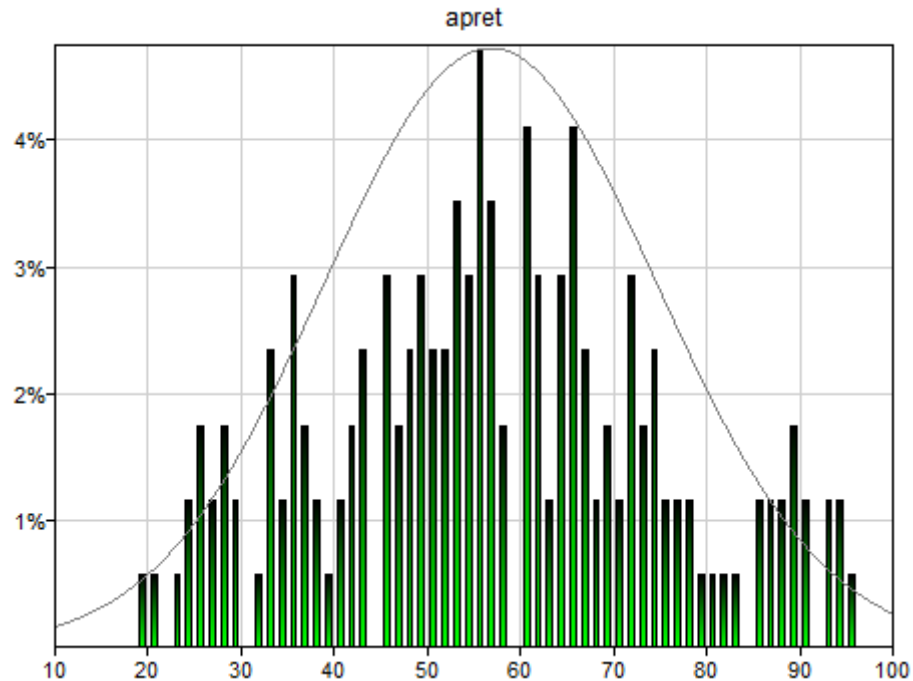
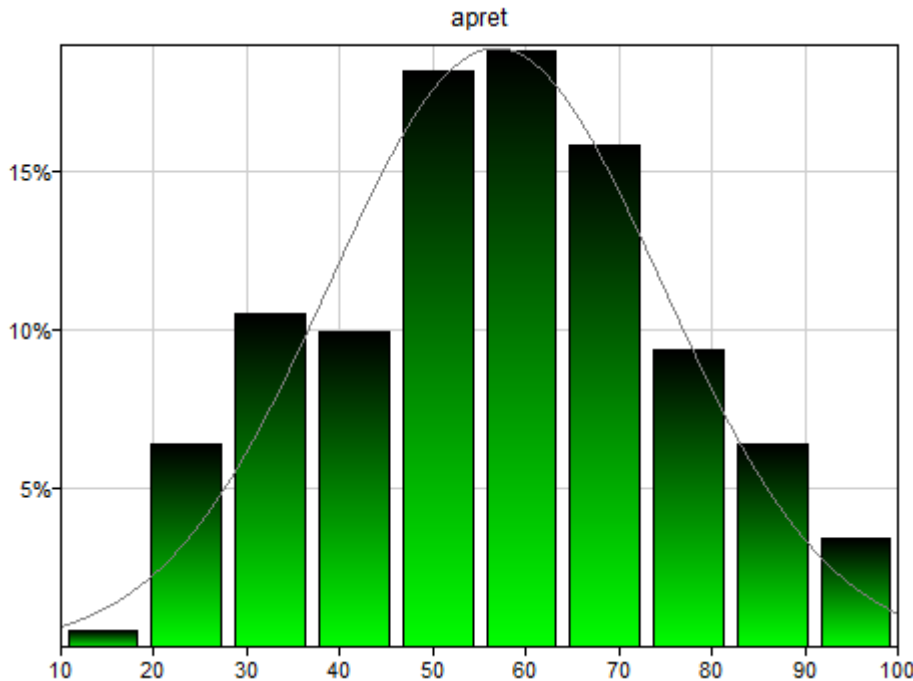


Histograms

What distributions data comes from can be seen very nicely on plots called “histograms.”

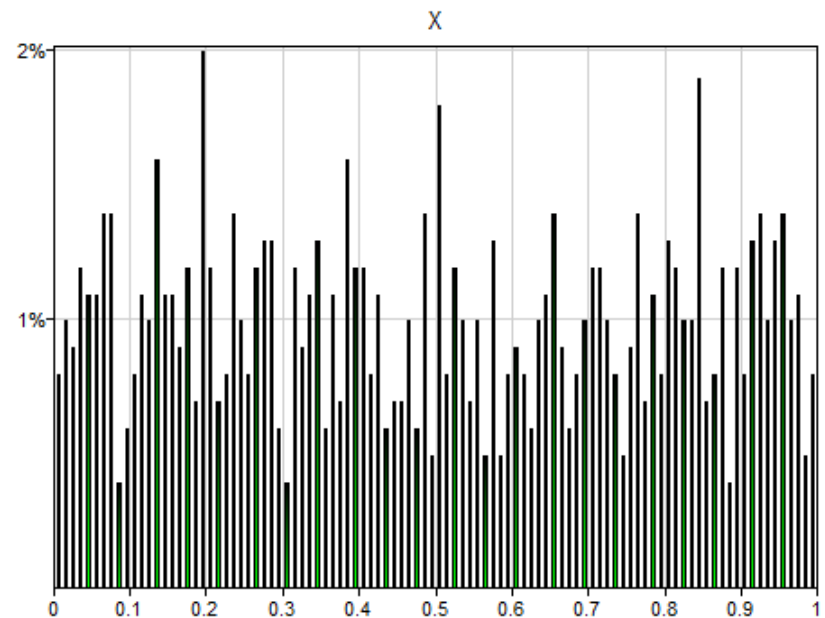
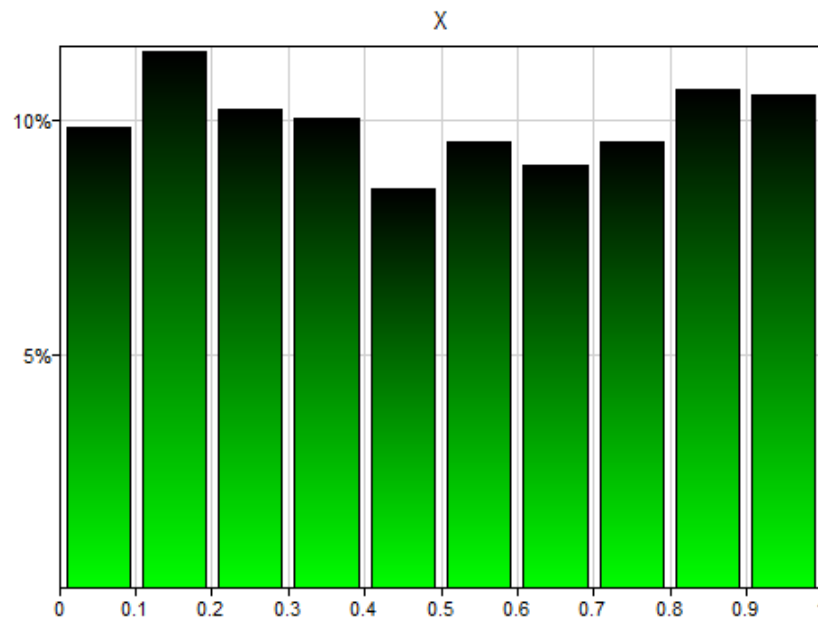


Histograms



Bin size affects the form, good bin size is essentially an art: I'm not aware of any research on automatic selection of bins. I am aware of at least one computer program that does it right (see <https://www.bayesfusion.com/> ☺).

Histograms

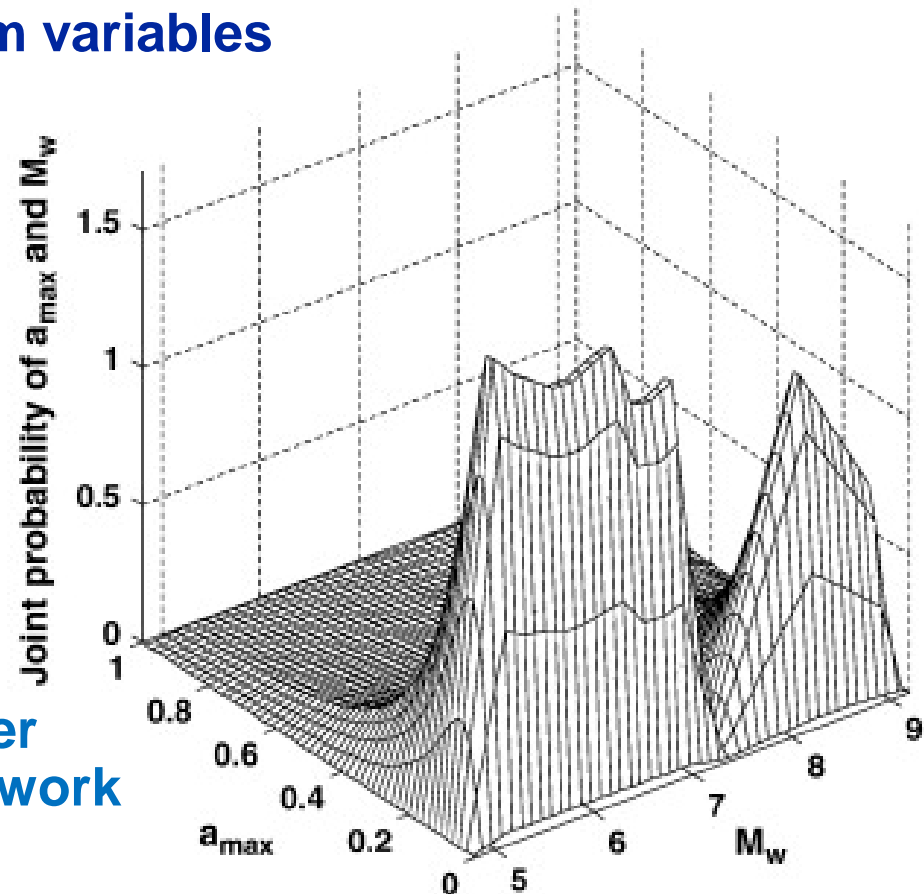


The effect of bin size is not that strong in case of some distributions (here: uniform distribution).

Joint Probability Distribution

Joint probability distribution

Expresses the probability of events defined over several random variables



e.g., probability distribution over grades and the amount of work in a university course

Source: <http://www.sciencedirect.com/science/article/pii/S0013795208002731>

Joint probability distribution

Expresses the probability of events defined over several random variables



Source: <http://postrecession.wordpress.com/tag/risk-aversion/>

Joint probability distribution

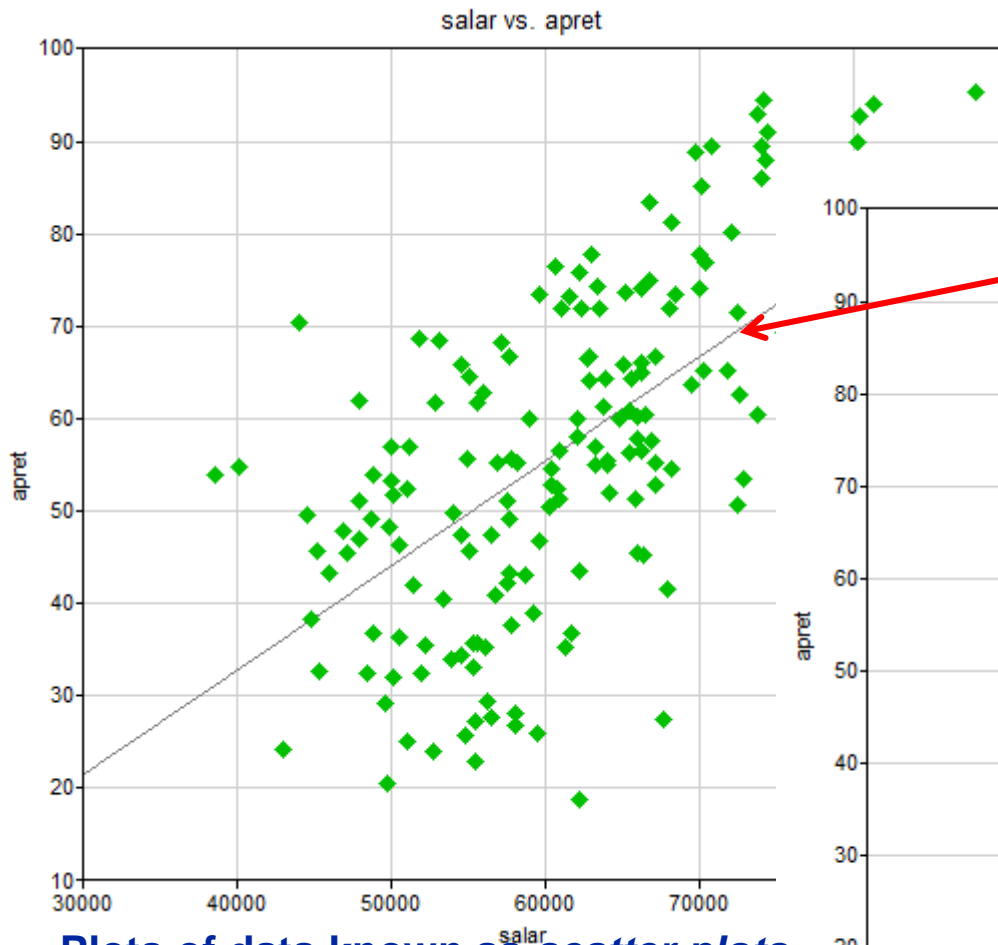
Joint probability distributions are much more interesting than probability distributions over single variables

Why?

Given the value of some of the variables in the joint probability distribution, we can estimate the probability distributions over the remaining variables.

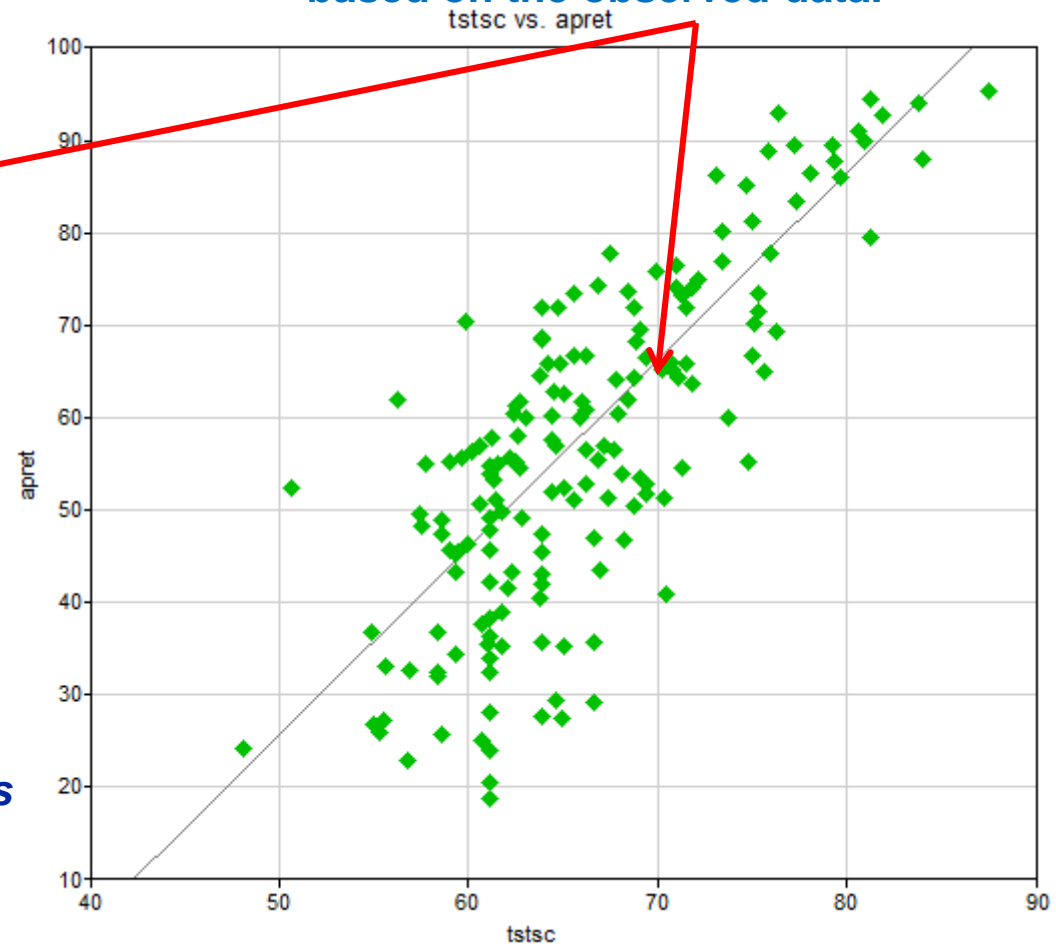
e.g., we can predict the grade distribution in a university course given the amount of work that we put into the course

Joint probability distributions



Plots of data known as *scatter plots* give an idea of the joint probability distribution between two variables.

Sometimes, we are interested in the linear relationship between variables and derive a linear regression line based on the observed data.



Correlation

- We are often looking for the information about tendency to vary together rather than independently.
- Correlation is a measure of the extent to which two random variables X & Y are linearly related (watch out: correlation may not capture non-linear dependences!).
- Originally introduced by Francis Galton to replace causation. Later, after statisticians had realized that it cannot fully represent causality, they clearly distanced from it (“Correlation does not mean causation.”).
- Can make sense (smoking and lung cancer) but can also be very tricky (examples: hospitals and dying, good surgeon and success of an operation, ice cream consumption and drowning).

Correlation matrix

	spend	apret	top10	rejr	tstsc	pacc	strat	salar
spend	-							
apret	0.601231	-						
top10	0.675656	0.642464	-					
rejr	0.633544	0.514958	0.643163	-				
tstsc	0.71491	0.782183	0.798807	0.628601	-			
pacc	-0.23673	-0.302834	-0.207505	-0.0715207	-0.164223	-		
strat	-0.561755	-0.458311	-0.247857	-0.283617	-0.465226	0.131858	-	
salar	0.711838	0.635852	0.637648	0.606777	0.715472	-0.37524	-0.347673	-

“Correlation does not mean causation”

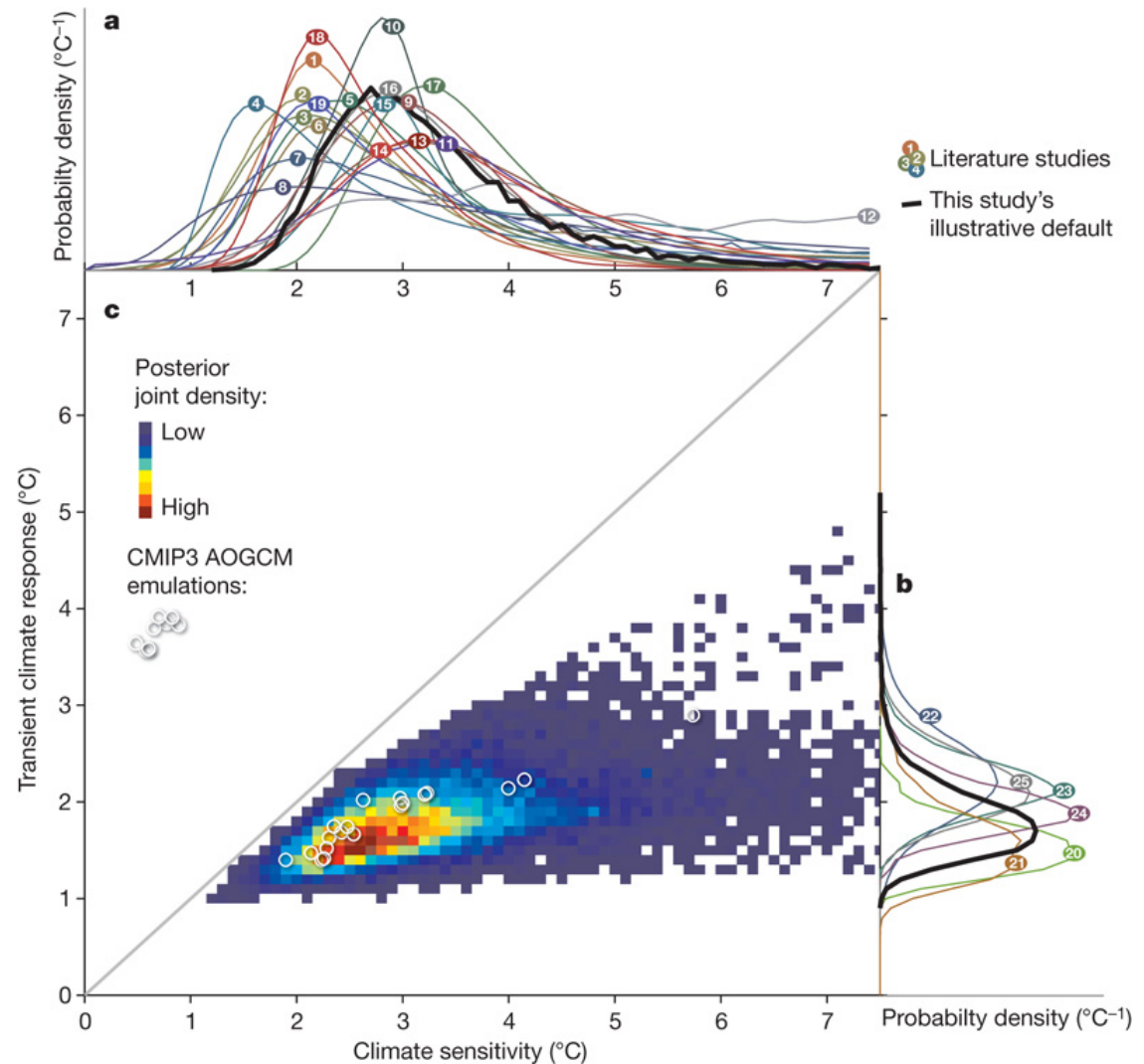
Cliché but indeed often true: Correlation between two variables by itself (i.e., in absence of other information) does not tell us much about the causal structure



Marginal probability distribution

Defined as the probability distribution over a single variable (when there are more variables ☺).

Can be derived from a joint probability distribution.

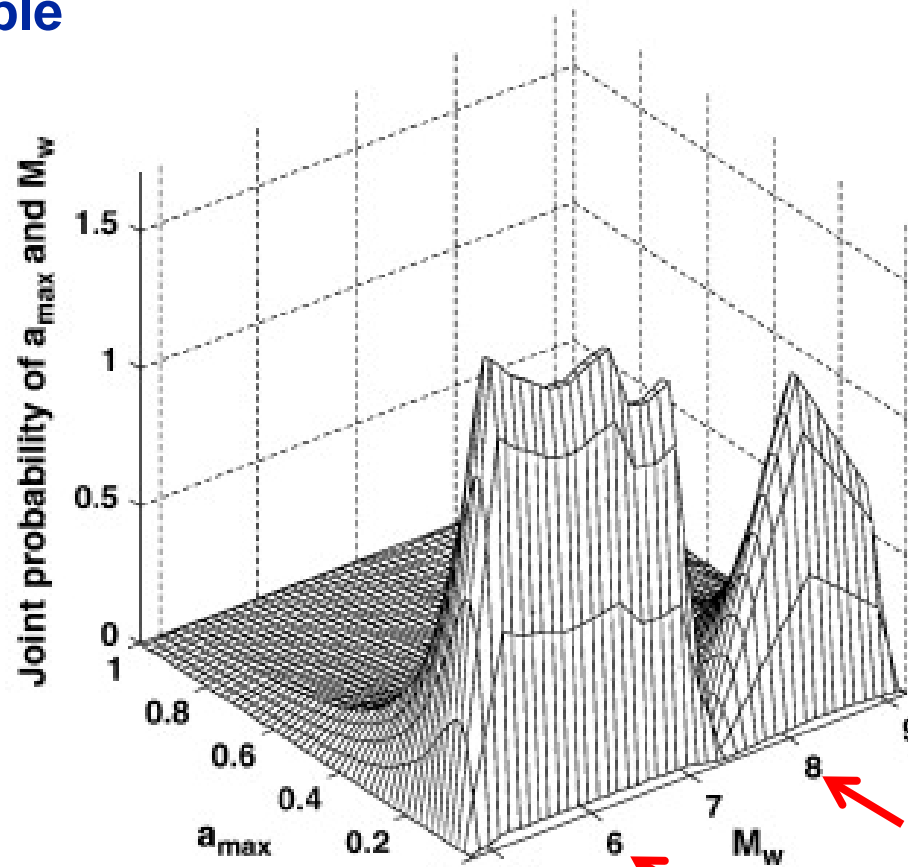


Source: http://www.nature.com/nature/journal/v458/n7242/fig_tab/nature08017_F1.html

Conditional probability distribution

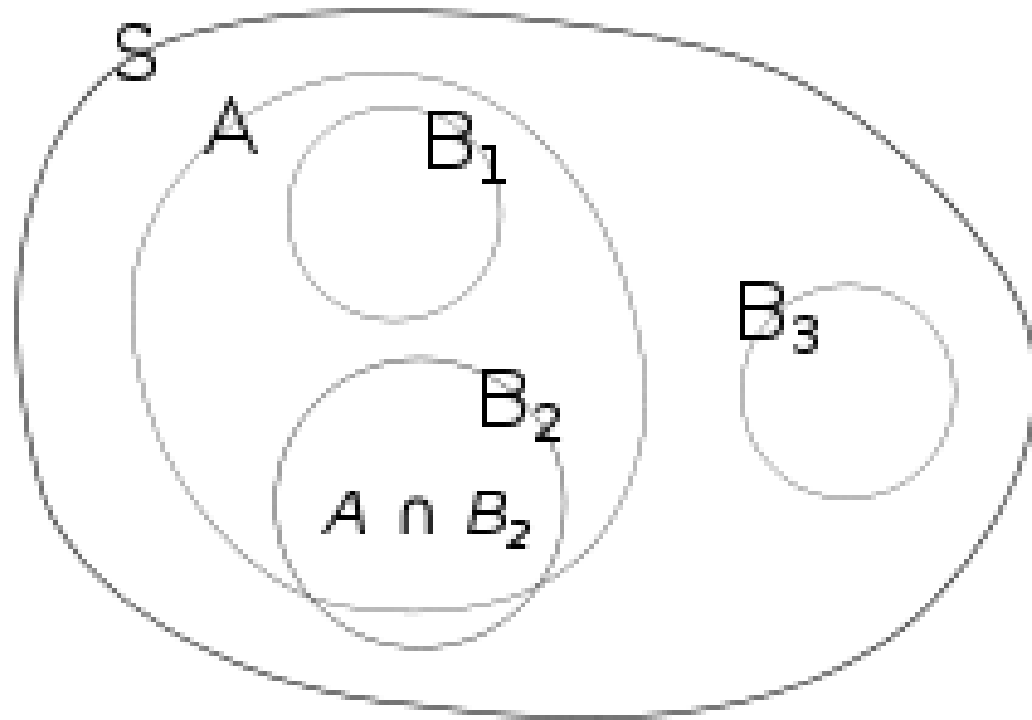
Once we know the value of one of the variables, we can make a statement about the probability distribution over the other variable

It is going to be different for different values of the first variable



Source: <http://www.sciencedirect.com/science/article/pii/S0013795208002731>

Venn diagrams



Source: http://en.wikipedia.org/wiki/Conditional_probability

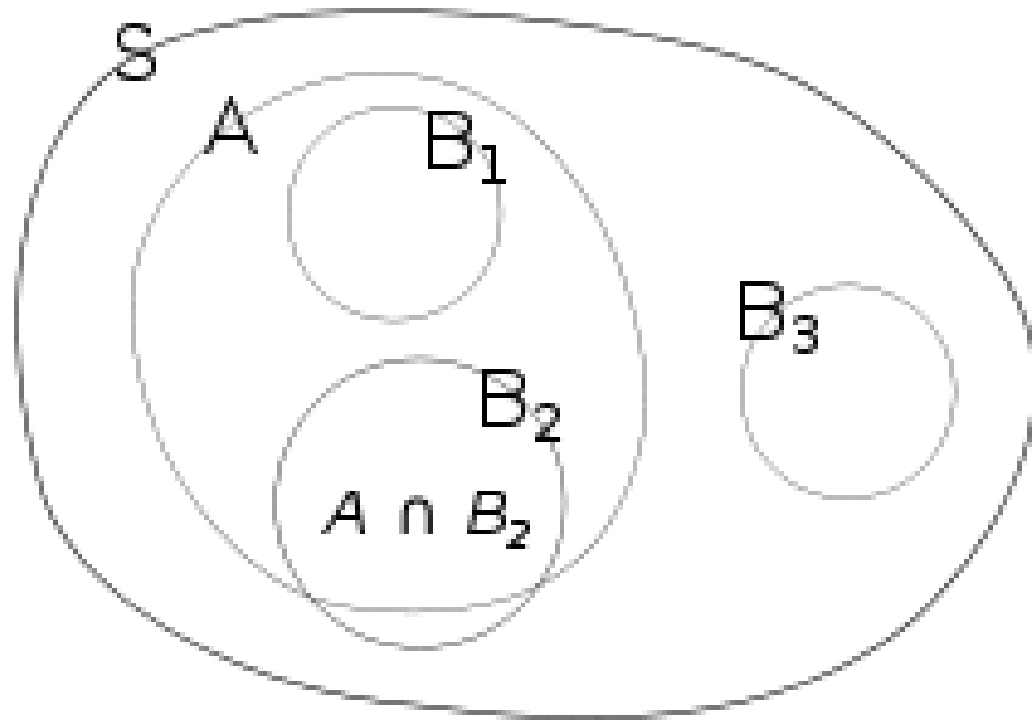
Conditional probability

Definition: $P(A|B) = P(A \cap B) / P(B)$

$$P(A|B_1) = ?$$

$$P(A|B_2) = ?$$

$$P(A|B_3) = ?$$



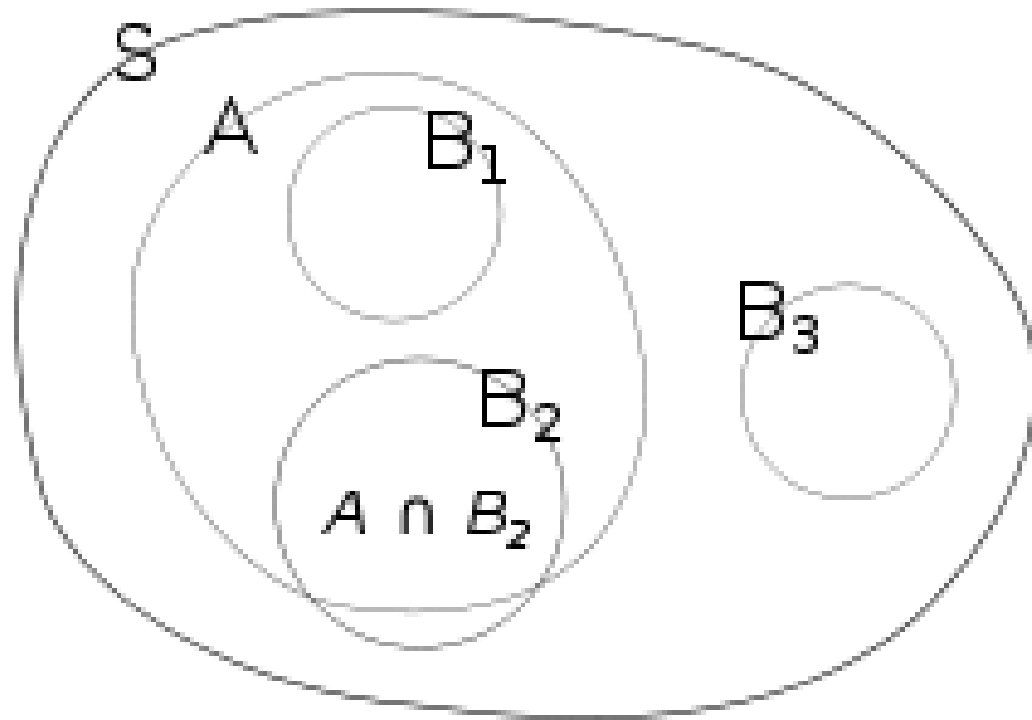
Independence

Mathematical definition: $A \perp B \Leftrightarrow P(A, B) = P(A) P(B)$

$A \perp B_1 ?$

$A \perp B_2 ?$

$A \perp B_3 ?$



Independence: Common sense

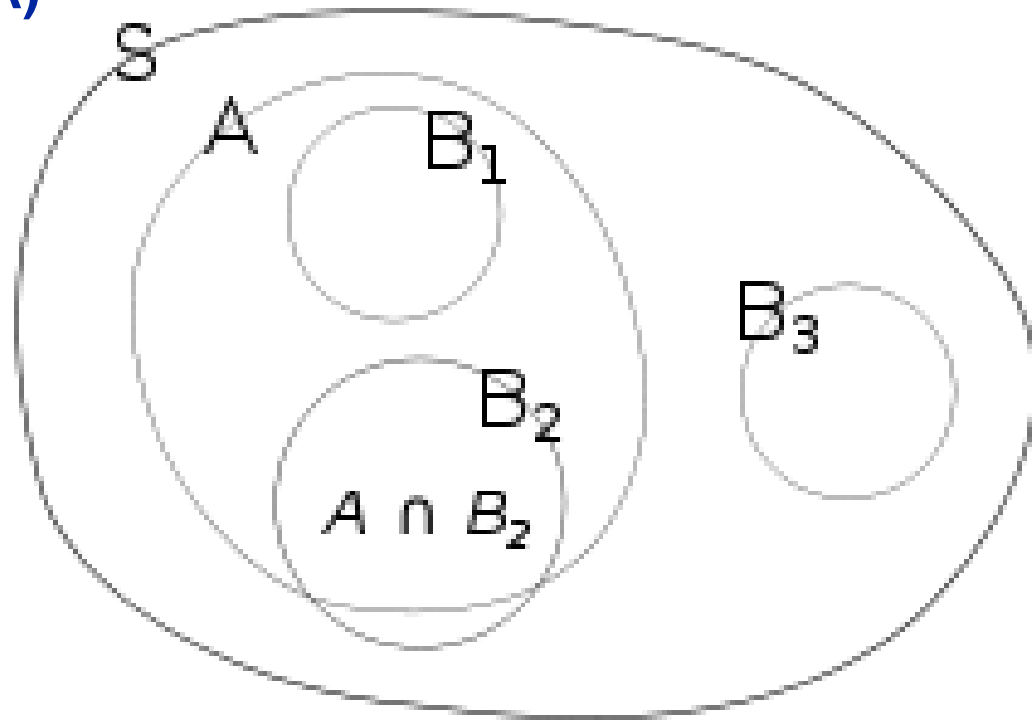
The following is straightforward to derive from the definition of independence (assuming $P(B) > 0$):

$$A \perp B \Leftrightarrow P(A|B) = P(A)$$

$$A \perp B_1 ?$$

$$A \perp B_3 ?$$

$$A \perp B_2 ?$$



Bayesian Probability Theory

Bayes theorem

An easy to prove theorem, obtained from the definition of conditional probability:

From

$$P(A|B) = P(A,B) / P(B)$$

and

$$P(B|A) = P(A,B) / P(A)$$

we have

$$P(A|B) = P(B|A) / P(B) P(A)$$

Posterior (a.k.a. a-posteriori)
probability

Prior (a.k.a. a-priori) probability

Bayes theorem gives us a mechanism for
changing our opinion in light of new evidence!

Bayes theorem example

Let the prevalence of syphilis in the population of young people planning to get married in Pennsylvania be 0.001.

Let a (mandatory) test, required for obtaining the marriage license have sensitivity of 0.98 and specificity of 0.95.

What is the probability that your fiancée, who tested positive for syphilis, has syphilis?

$$P(S|+) = P(+|S)/P(+) P(S) \quad (\text{Bayes theorem})$$

$$P(+) = P(+|S) P(S) + P(+|\sim S) P(\sim S) \quad (\text{theorem of total probability})$$

$$P(+) = 0.98 \cdot 0.001 + 0.05 \cdot 0.999 = 0.05093$$

$$P(S|+) = 0.98 / 0.05093 \cdot 0.001$$

Posterior (a.k.a. a-posteriori)
probability

Prior (a.k.a. a-priori) probability

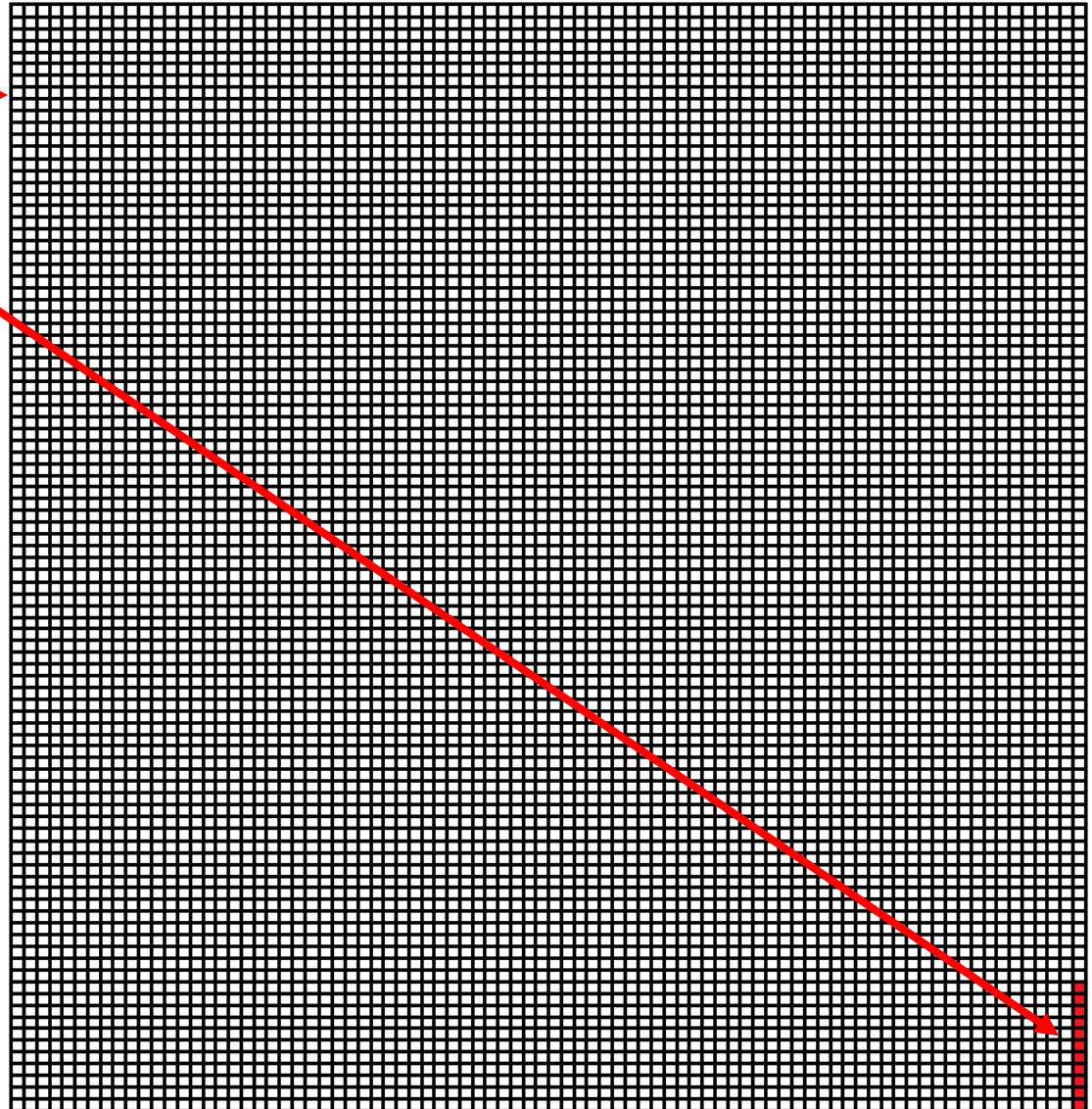
0.01924

A better human interface to the same problem

Imagine a population
of 10,000 individuals.

Prevalence of 0.001
means that 10 out of
the 10,000 will have
the disease.

Let us screen them all.



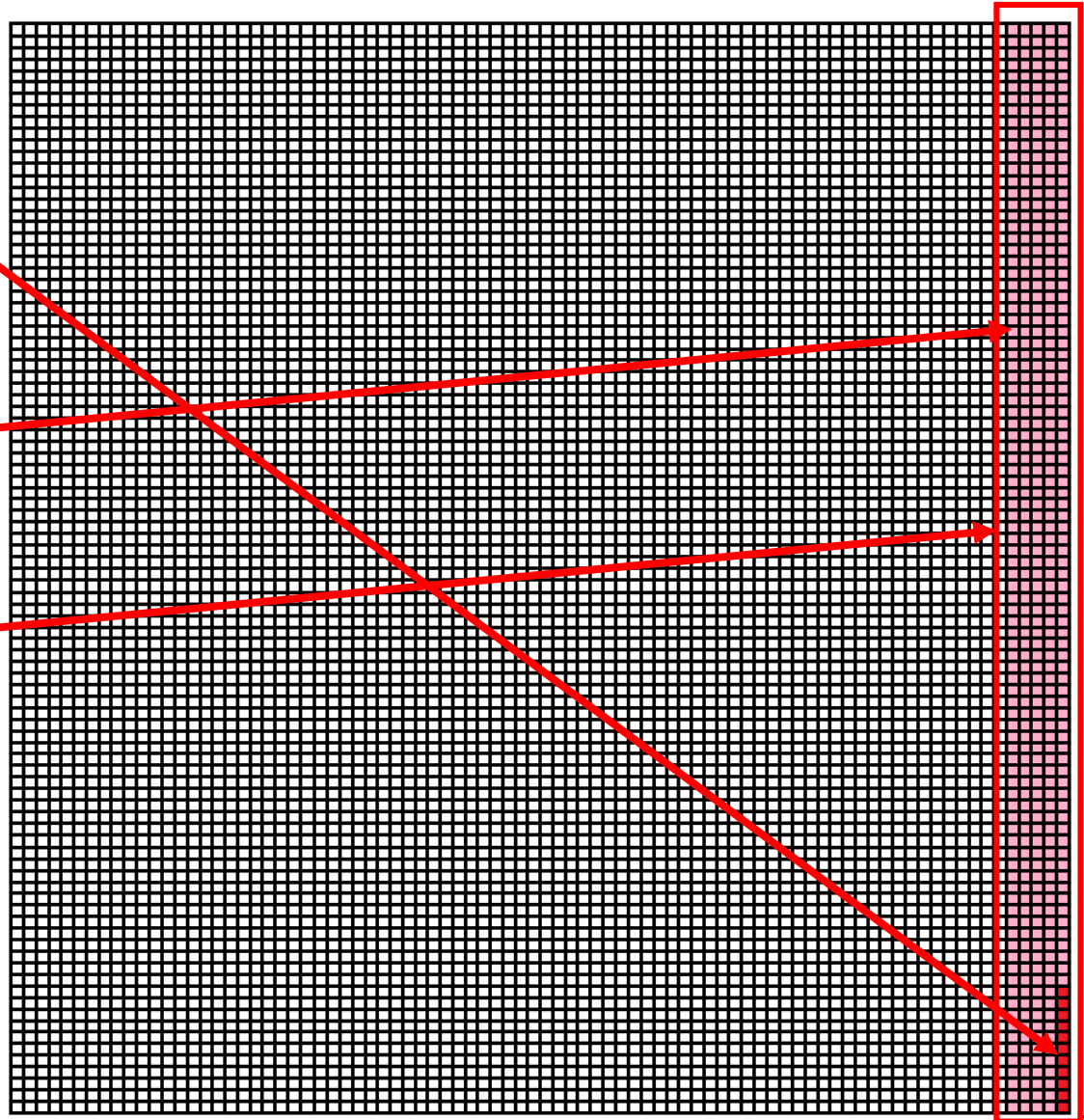
A better human interface to the same problem

With sensitivity of 98%, 9.8 of the 10 diseased will be correctly detected.

With specificity of 95%, we will have 5% (of 9,990), which is 499.5 false positives.

Now, among all those who tested positive, roughly $9.8/(9.8+499.5) \approx 2\%$ will be diseased.





Is it easier to understand 😊?



Bayes theorem and Bayesian statistics

A versatile and powerful theory that seems to solve a variety of problems, originating from an 18th century English mathematician, Rev. Thomas Bayes (http://en.wikipedia.org/wiki/Thomas_Bayes)



the theory 
that would
not die 

how bayes' rule cracked
the enigma code,
hunted down russian
submarines & emerged
triumphant from two
centuries of controversy 
sharon bertsch mcgrayne

Bayes Theory is so “hot” that a lightly written book “The Theory That Would Not Die,” published in 2011, has become a bestseller

Recommended video:

<http://www.youtube.com/watch?v=8oD6eBkjF9o>

Bayesian modeling is reliable and it solves hard problems.

It can use both, data and expert knowledge.

What is the relation of Bayesian statistics to classical statistics?



Classical statisticians: “We have no clue ☹. Probability is a limiting frequency. A nuclear war is not a repetitive process.”

Bayesians: “0.24 😊. Probability is a measure of belief”

What is the relation of Bayesian statistics to classical statistics?

- **Bayesians:** “Probability is a measure of belief” (as opposed to “limiting frequency”), so it is subjective!
- **Classical statisticians** accuse Bayesians of “hocus pocus” with the prior distributions (“How do you know them?”).
- **Bayesian statistics** comes with so called “limit theorems,” which say that no matter what distribution you choose for your prior, you will eventually converge to the true distribution if you observe enough evidence.
- “Today’s prior is yesterday’s posterior.”
- Of course, there is nothing wrong with starting with “the right distribution” in the beginning (In other words, it would be unwise to ignore available statistics).
- But even if you don’t have them, you can still do useful work, even if you have to just guess the priors.

