Fundamental Concepts from Statistics

Chirayu Wongchokprasitti, PhD

University of Pittsburgh

Center for Causal Discovery

Department of Biomedical Informatics

chw20@pitt.edu

http://www.pitt.edu/~chw20

Overview

- **Fundamentals**
  (Uncertainty, probability, variance, sampling, randomness, elements of data analysis. Describing and displaying data, correlation.)
- **Bayes theorem and Bayesian probability theory**
- **Joint probability distribution**
  (The foundation of any analytic technique. Conditional probability distribution, Bayes theorem, prior and posterior probability distribution. Tools for representing joint probability distributions: probability trees, Bayesian networks, equation-based models)
- **Representations of the joint probability distribution**

Why statistics ☺?

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

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 ☺?

"Data don't make any sense,
we will have to resort to statistics."

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.

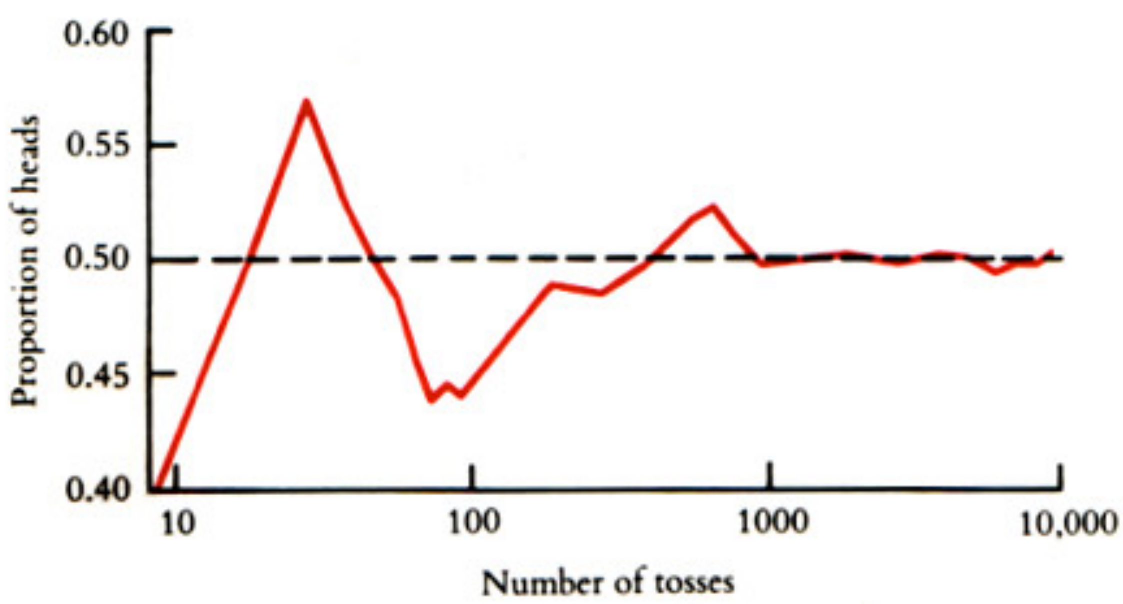
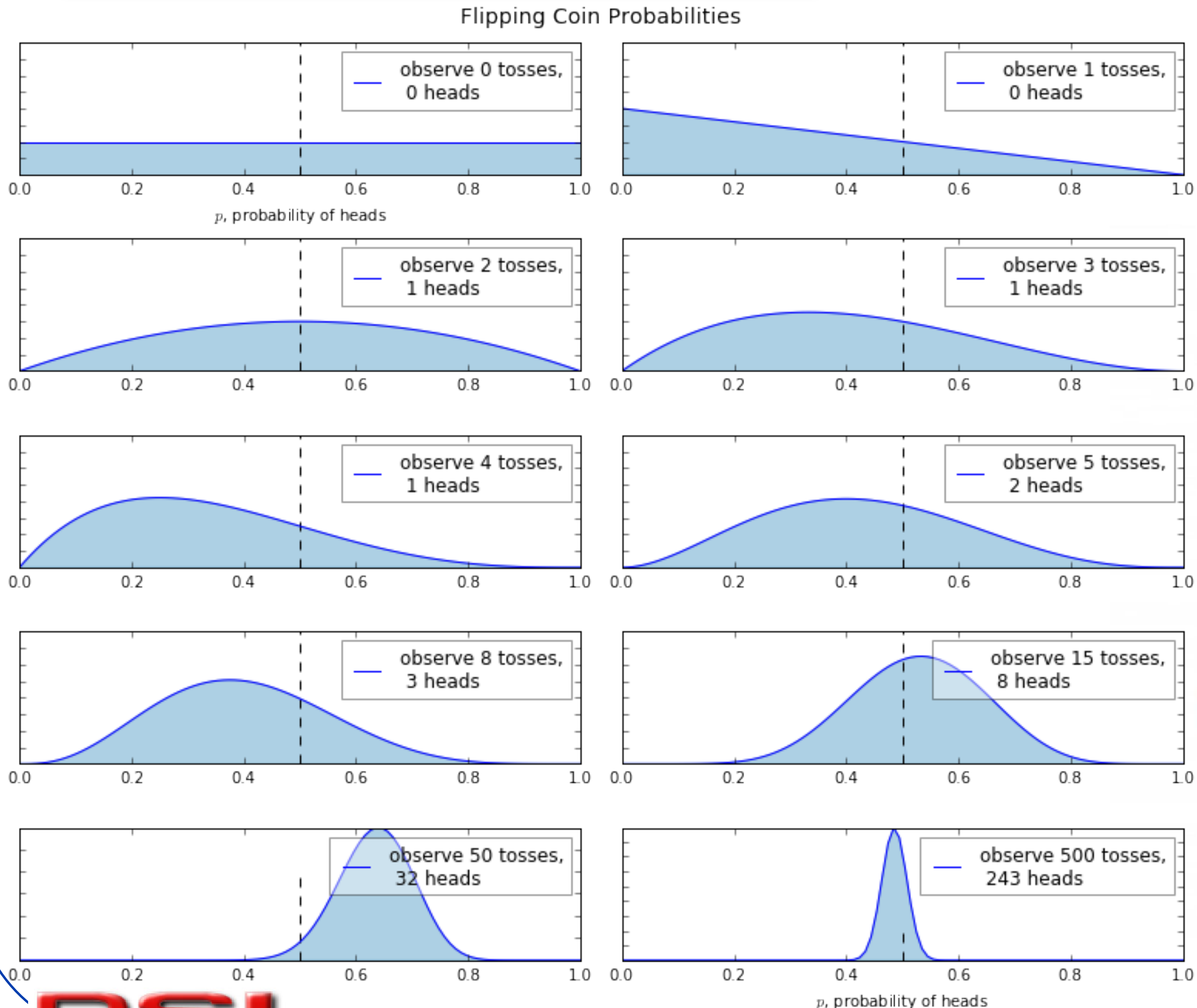


Figure 7.2 Percent of heads versus number of tosses in Kerrich's coin-tossing experiment. [David Freedman et al., *Statistics* Norton, 1978.]

Flipping Coin Probabilities



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.**
- **...**

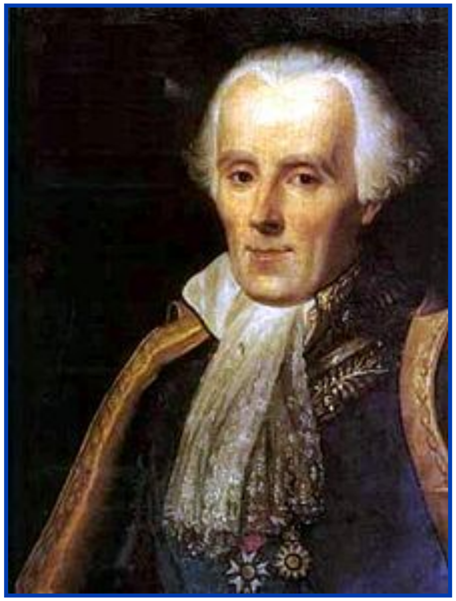
A Brief Review of Probability Theory and Statistics

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)



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.

Some features of statistical analysis

- Questions that we ask (in statistics but also in science in general) concern systems, i.e., parts of the real world that can be reasonably studied in separation.
- We want to make inference from a sample to a population (unless we can make the entire population a sample)!
- Ideal sampling should be random, giving every member of the population an equal chance of being selected
- In that case, we hope (but have a whole statistics for us) that the sample is representative, i.e., has approximately the characteristics of the population.
- If the sample is not random, then unknown/known factors may bias the sample (such as experimenter's biases, political factors, etc.).
- Even in case of random sampling (the ideal) there is no guarantee for a representative sample, but we can get arbitrarily close (in terms of probability) to the population.

Describing and displaying data

Statistics provides tools for describing and displaying data

Example:

- **What causes low student retention in U.S. colleges?**
- **Over 120 variables (only 8 in the picture on the right-hand side) measured across 204 universities (total of over 24,000 numbers).**
- **Note variables (columns) and data points (rows).**

spend	apret	top10	rejr	tstsc	pacc	strat	salar
9855	52.5	15	29.474	65.063	36.887	12	60800
10527	64.25	36	22.309	71.063	30.97	12.8	63900
7904	37.75	26	25.853	60.75	41.985	20.3	57800
6601	57	23	11.296	67.188	40.289	17	51200
7251	62	17	22.635	56.25	46.78	18.1	48000
6967	66.75	40	9.718	65.625	53.103	18	57700
8489	70.333	20	15.444	59.875	50.46	13.5	44000
9554	85.25	79	44.225	74.688	40.137	17.1	70100
15287	65.25	42	26.913	70.75	28.276	14.4	71738
7057	55.25	17	24.379	59.063	44.251	21.2	58200
16848	77.75	48	26.69	75.938	27.187	9.2	63000
18211	91	87	76.681	80.625	51.164	12.8	74400
21561	69.25	58	44.702	76.25	26.689	9.2	75400
20667	65	68	22.995	75.625	28.038	11	66200
10684	61.75	26	8.774	66	33.99	9.5	52900
11738	74.25	32	25.449	66.875	27.701	12	63400
10107	74	43	11.315	71	29.096	16.2	66200
7817	65.75	36	33.709	64.25	52.548	17.7	54600
7050	26	11	0	55.313	55.651	18.8	59500
9082	83.5	73	64.668	77.375	43.185	13.6	66700
11706	60	56	16.937	73.75	39.479	12.7	62100
7643	49.25	23	36.635	62.813	39.302	18.7	57700
25734	90	77	67.758	80.938	44.133	10	80200
20155	86	84	69.31	79.688	48.766	17.6	74000
29852	94.5	84	75.009	81.313	51.363	10.6	74100
7980	68.5	34	9.122	63.875	35.294	16.3	53100
8446	57	23	29.65	64.625	36.181	14.8	63200
24636	92.75	88	70.653	81.875	43.464	12.8	80300
7396	68.75	34	13.469	63.889	39.05	14.8	51900
24256	81.25	68	35.556	75	26.736	11.5	68200
7263	54	28	49.583	68.125	42.149	13.4	48839
7005	46.75	50	36.236	68.188	33.875	22.5	59600
10454	77.75	34	23.784	67.5	33.333	11.2	70000

Measures of central tendency and spread

Measures of central tendency:

- mode (value occurring with the greatest frequency)
- median (mid-most score in a series)
- mean (arithmetic average)
- trimmed mean

Measures of spread:

- ranges: crude range (highest, lowest), extended range (or corrected range) adds one unit to the range (to account for a possible error in measurement), trimmed ranges (drop *x*% of extreme points on both sides)
- variance $\sigma^2 = \Sigma_i (x_i - \mu)^2/n$
- standard deviation $\sigma = \mathrm{sqrt}(\sigma^2)$
- average deviation $\Sigma_i (x_i - \mu)/n$

Basic statistics

Excel

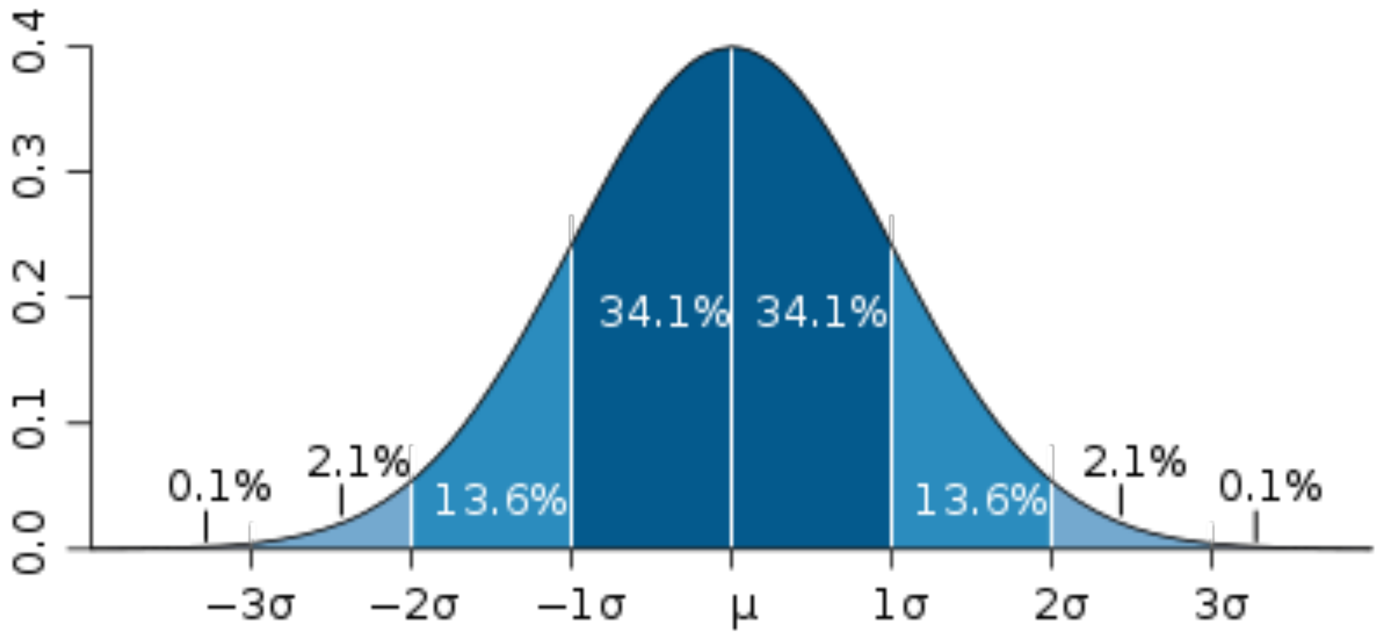
Apgra	
Mean	56.72107647
Median	55.7085
Mode	72
Standard Deviation	18.07709676
Variance	326.7814274
Kurtosis	-0.55445012 8
Skewness	0.089185832
Range	76.5
Minimum	18.75
Maximum	95.25
Sum	9642.583
Count	170

GeNIe

	Mean	Variance	StdDev	Min	Max	Count
spend	10974.5	3.02507e+007	5500.07	4125	35863	170
apret	56.7211	326.781	18.0771	18.75	95.25	170
top10	38.4588	547.859	23.4064	8	98	170
rejr	30.6542	292.345	17.0981	0	84.067	170
tstsc	66.1642	48.6549	6.97531	48.125	87.5	170
pacc	43.1731	171.746	13.1052	8.964	76.253	170
strat	16.0865	16.0521	4.0065	7.2	29.2	170
salar	61357.6	9.60946e+007	9802.79	38640	87900	170

Probability distribution

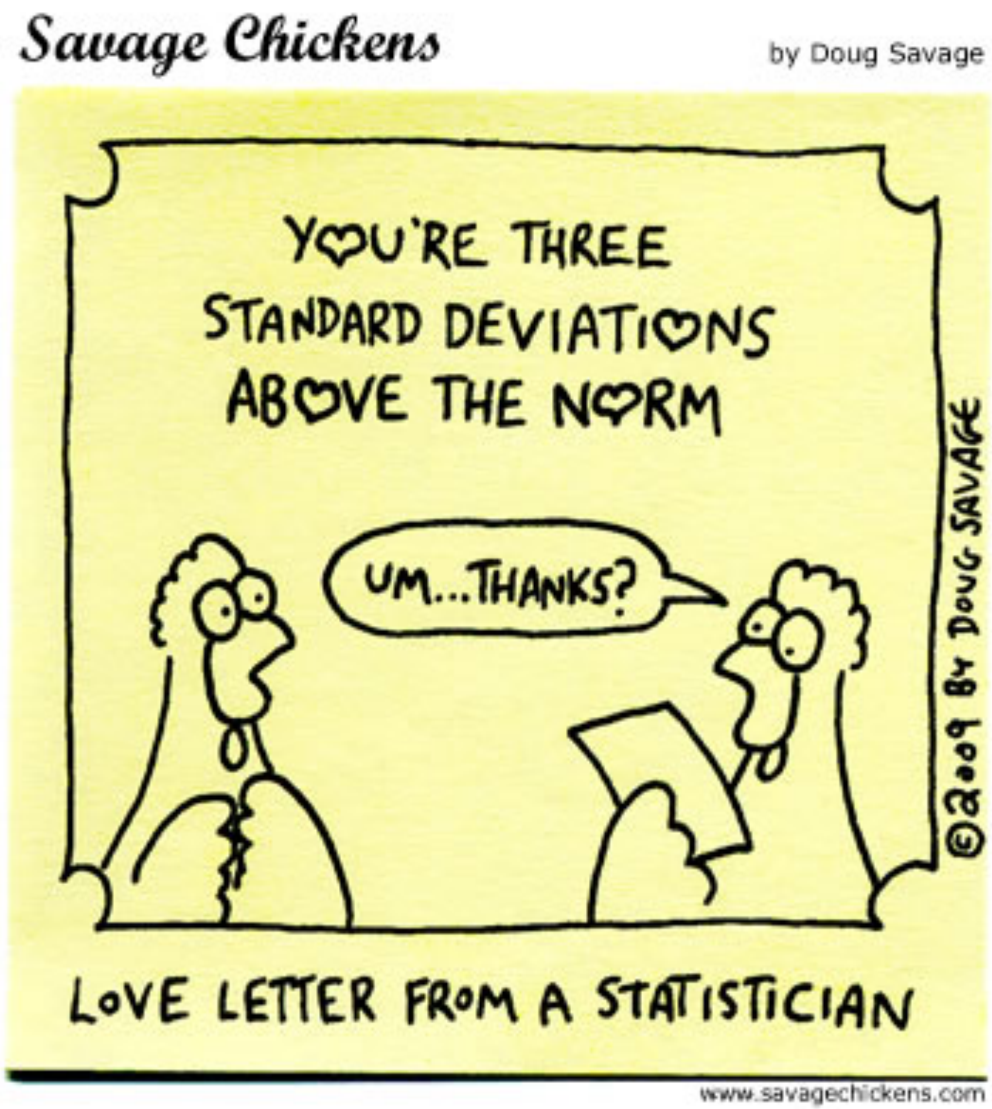
Expresses the relative probabilities of different values taken by a random variable



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

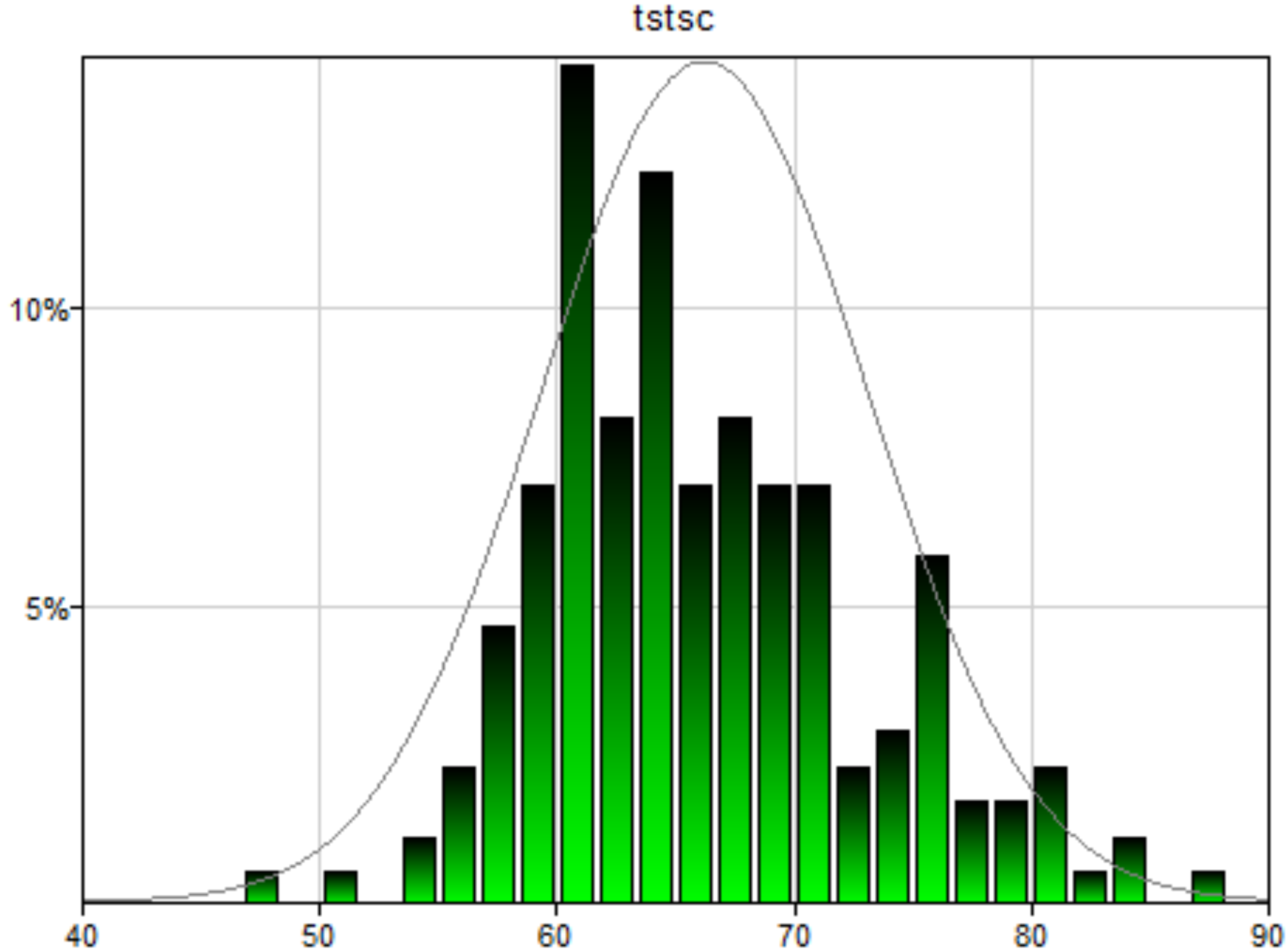
e.g., grade distribution in a university course

Standard deviation

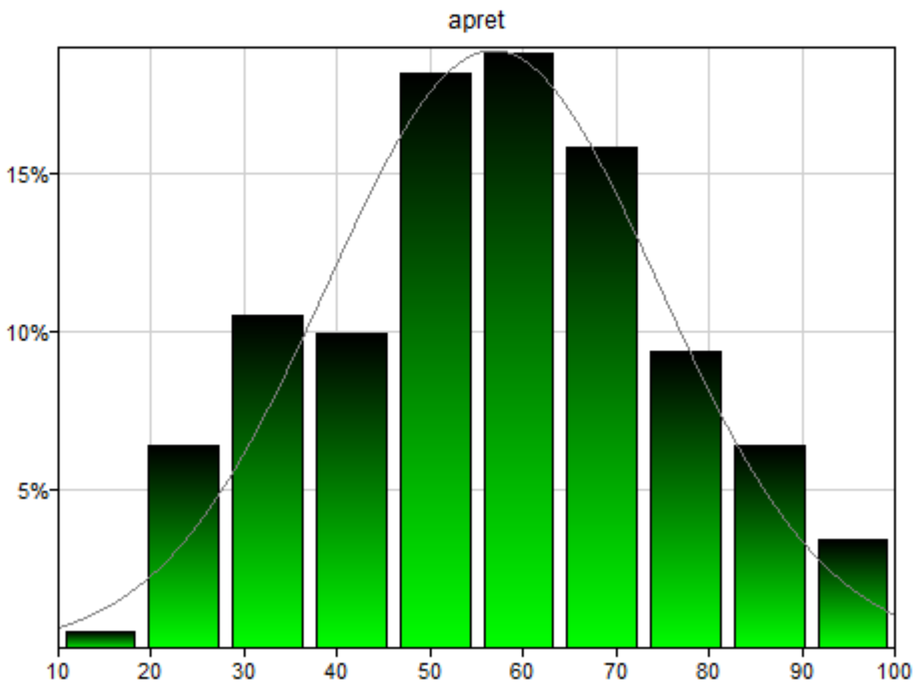


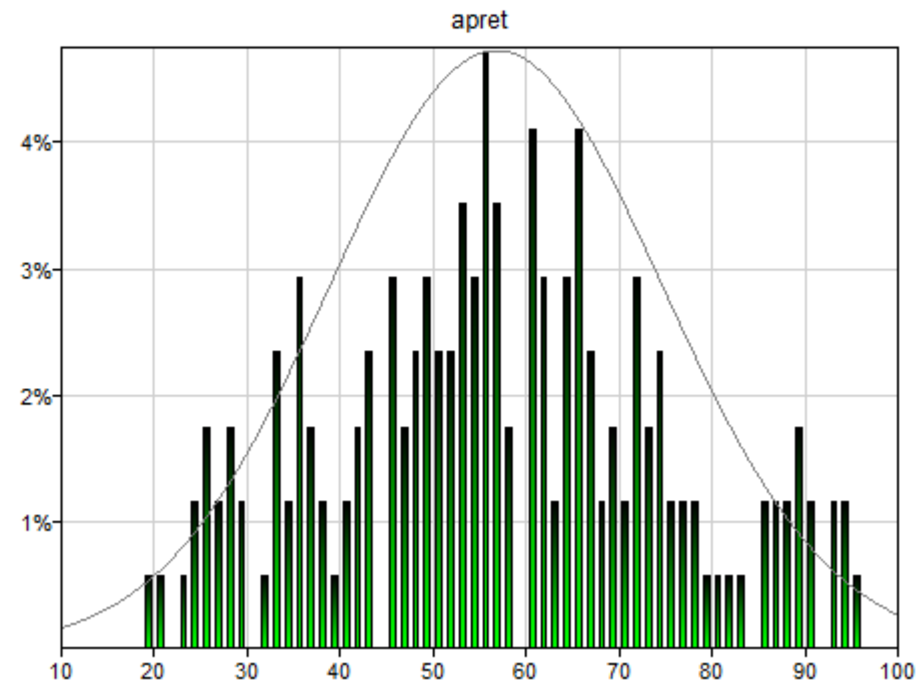
Histograms

Values that a variable takes in a data set can be seen very nicely on plots called “histograms”



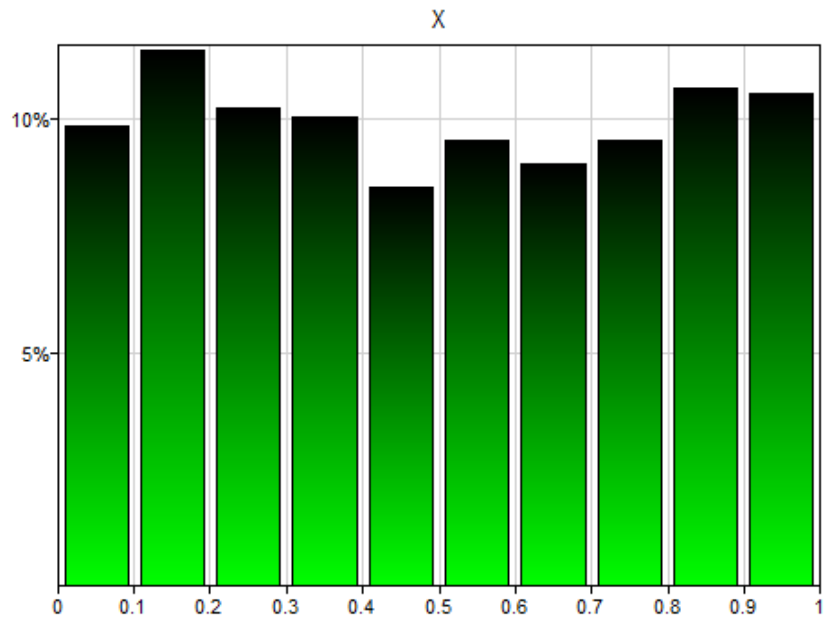
Histograms

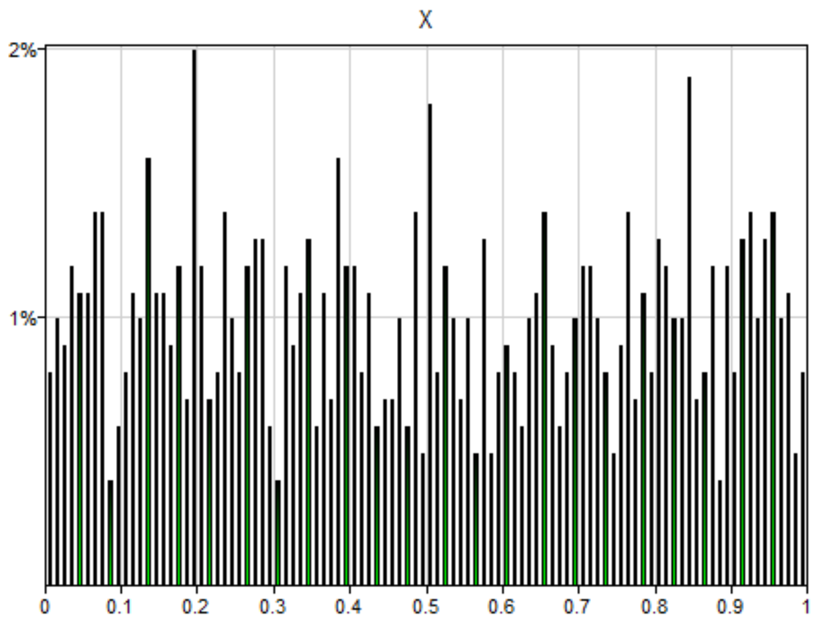




Bin size affects the form, good bin size is essentially an art: I'm not aware of any research on automatic selection of bins. I am aware of at least one computer program that does it right (see http://genie.sis.pitt.edu/).

Histograms



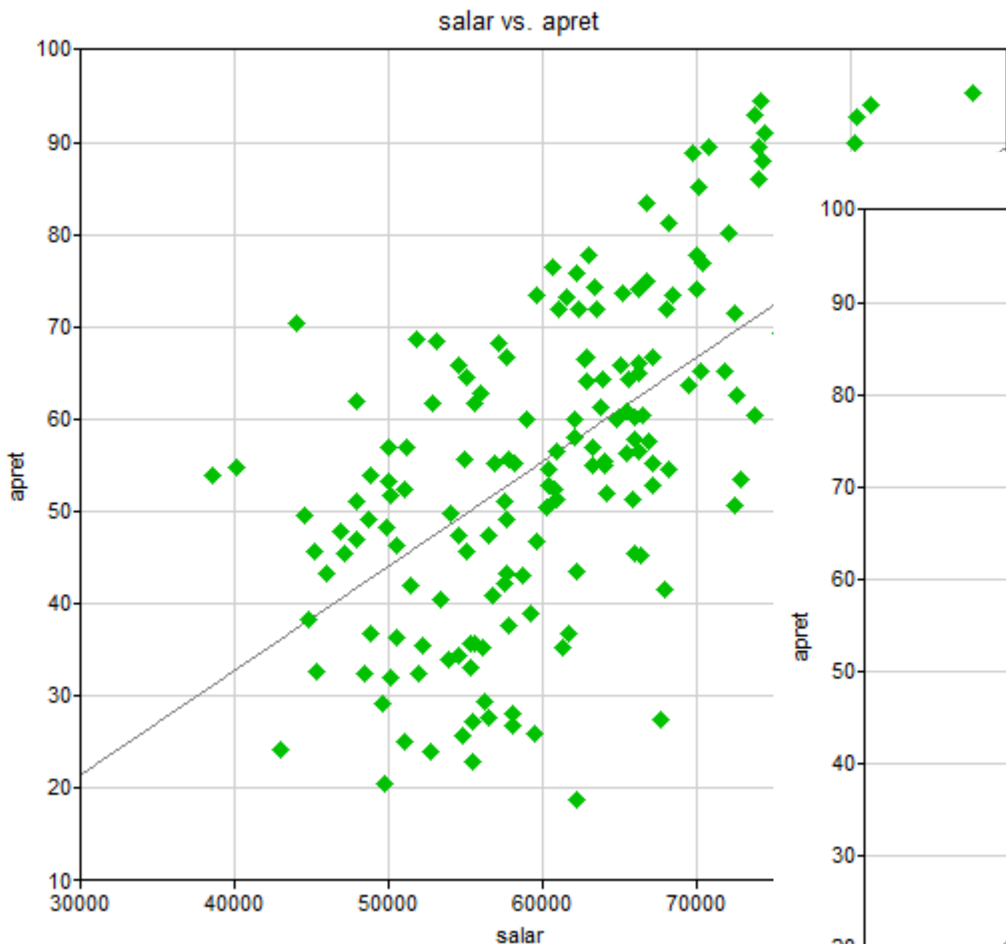


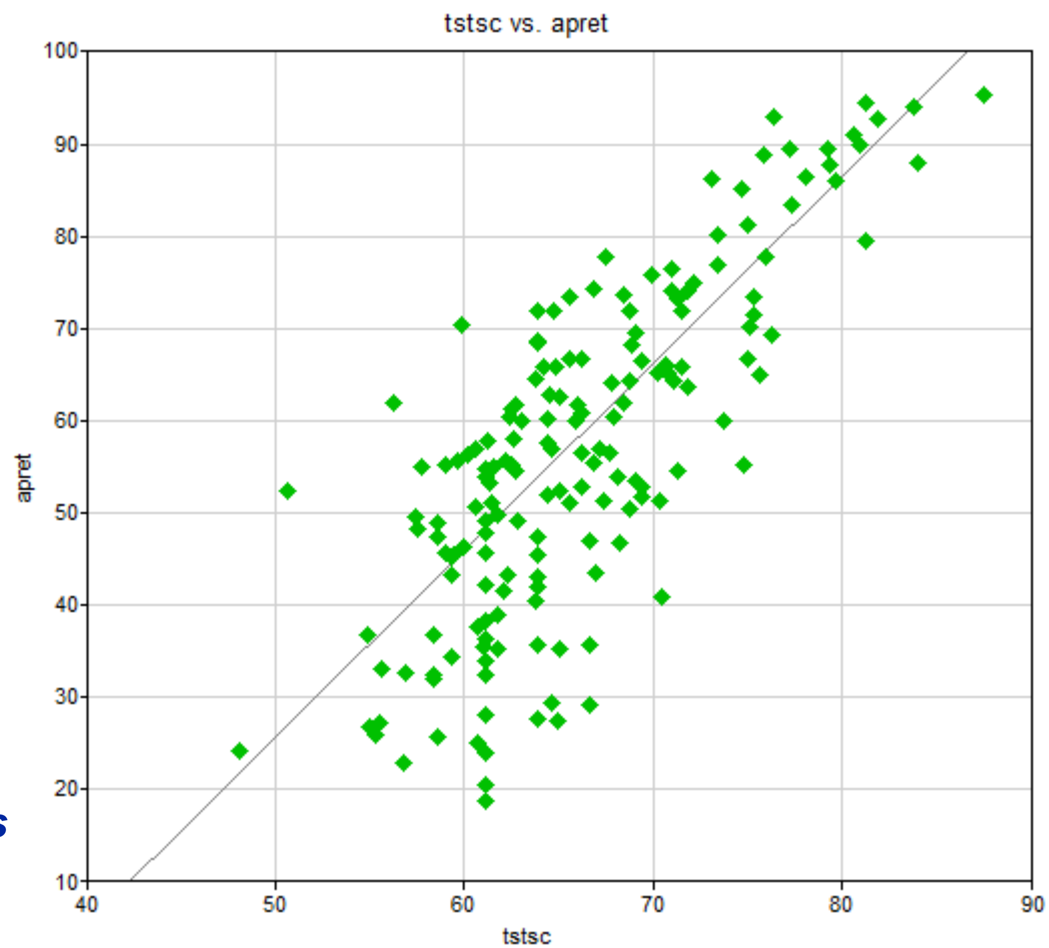
The effect of bin size is not that strong in case of some distributions (here: uniform distribution).

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.

Scatter plots



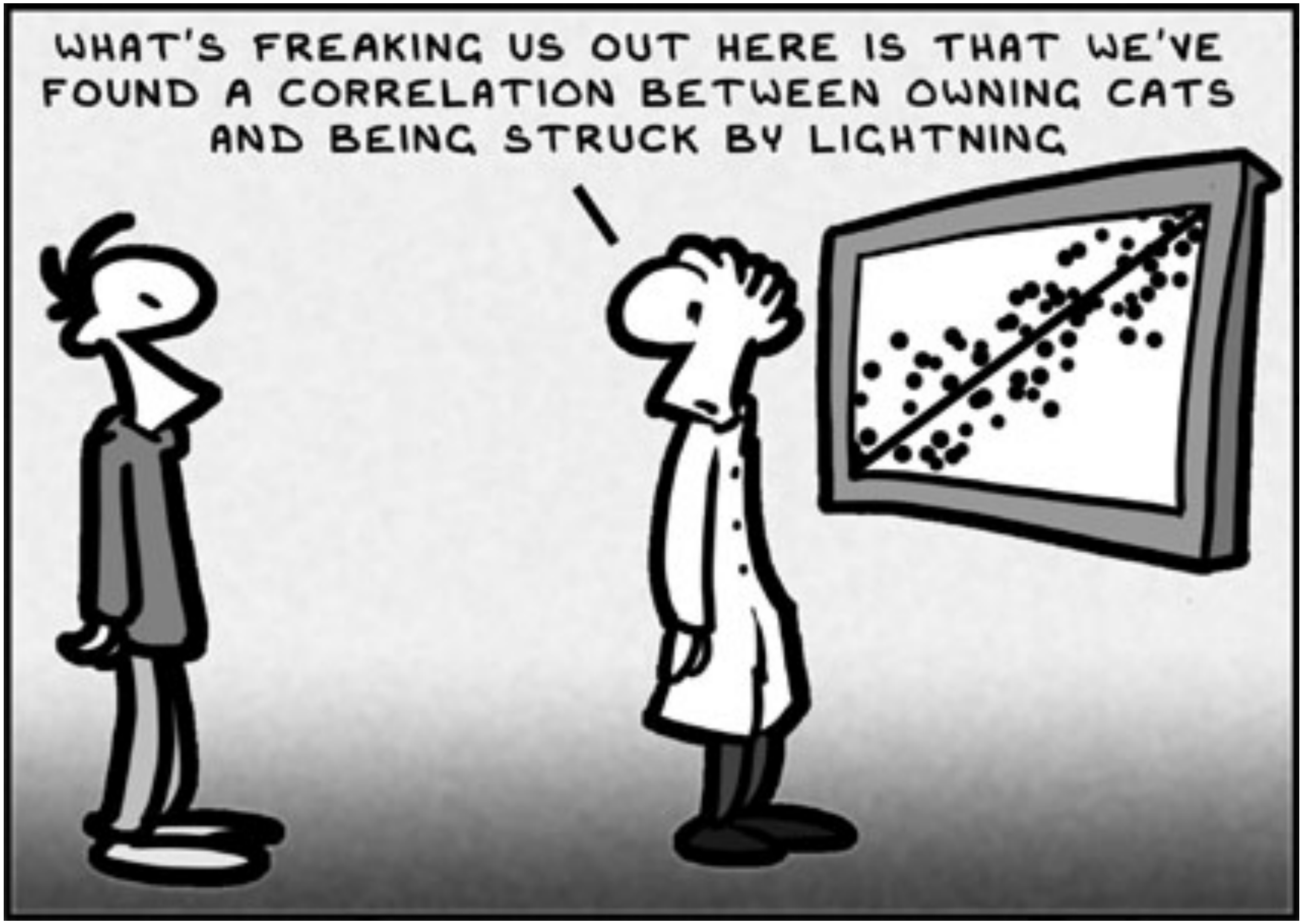


Plots of data known as *scatter plots* give an idea of the joint probability distribution between two variables.

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 dying, ice cream consumption and drowning).

Correlation



Correlation matrix

	spend	apret	top10	rejr	tstsc	pacc	strat	salar
spend	1							
apret	0.601231	1						
top10	0.675656	0.642464	1					
rejr	0.633544	0.514958	0.643163	1				
tstsc	0.71491	0.782183	0.798807	0.628601	1			
pacc	-0.23673	-0.302834	-0.207505	-0.0715207	-0.164223	1		
strat	-0.561755	-0.458311	-0.247857	-0.283617	-0.465226	0.131858	1	
salar	0.711838	0.635852	0.637648	0.606777	0.715472	-0.37524	-0.347673	1

Correlation does not mean causation

Cliché but certainly true: A single correlation by itself does not tell us much about the causal structure

Linear regression

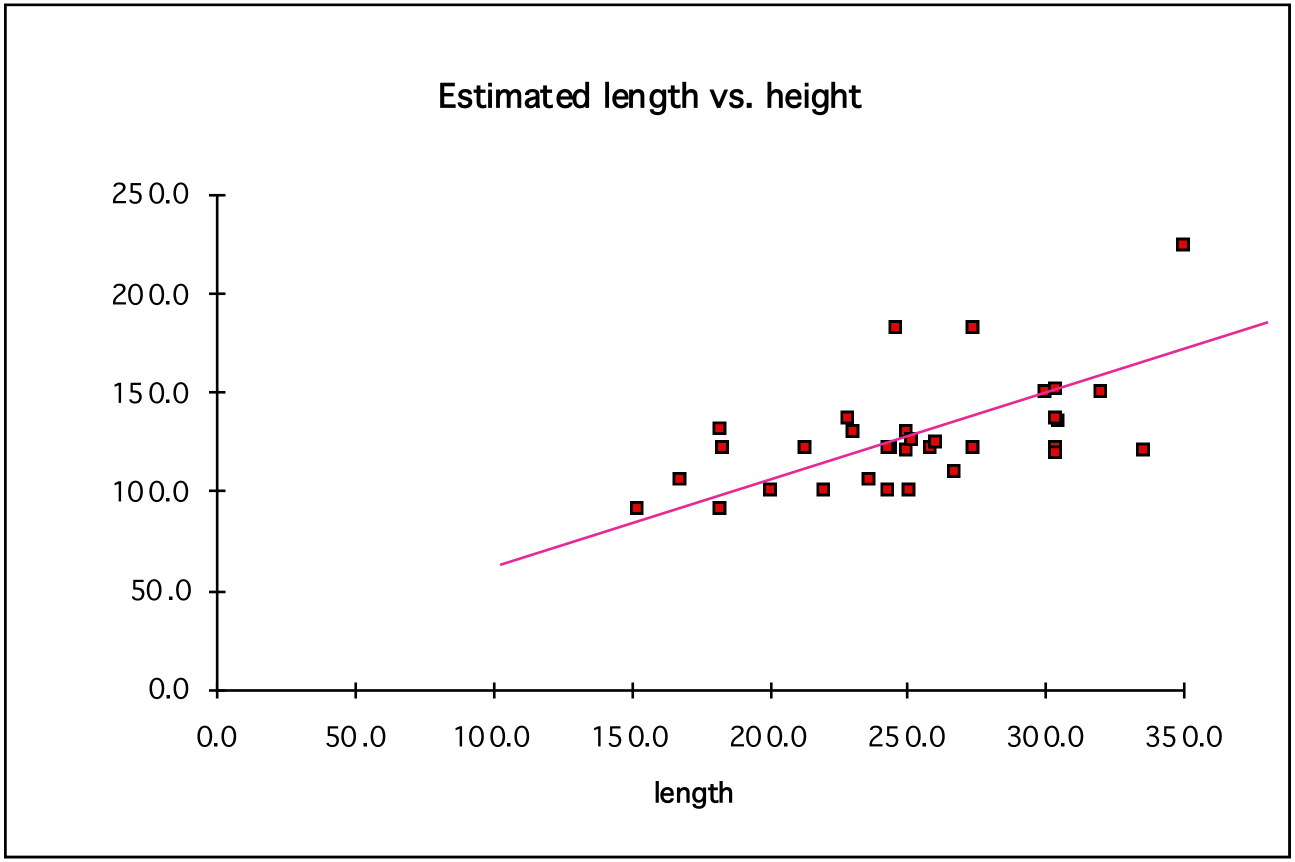
- Scatter plots portray the relationship between two quantitative variables. We would like to summarize the relationship more briefly.
- The simplest interesting relationship is linear (straight-line) dependence of a response variable y on an explanatory variable x.
- A straight line that describes the dependence of one variable on another is called a regression line.
- Regression line allows us to predict (approximately) the value of one variable if we know the value of the other variable.

Linear regression

We fit a line to the data, the line equation is $Y = b_0 + b_1X$

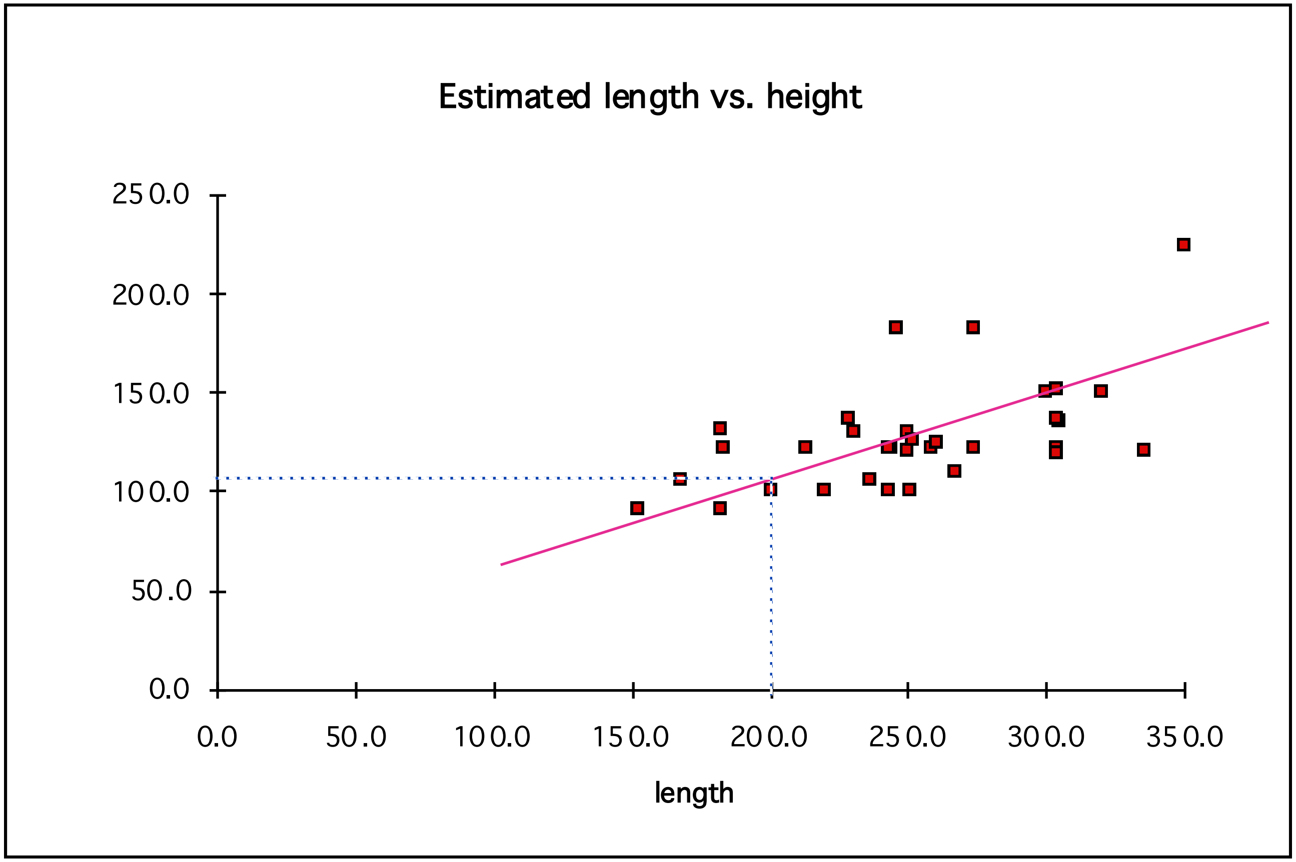
Note1: b_0, b_1 are intercept, coefficient parameters, respectively

Note2: Linear Regression ≠ Linear Model ($Y = b_0 + b_1X + b_2X^2$)



Linear regression: Prediction

Can we predict what an INFSCI 1000 student will estimate for height if she estimated the length to be 200 cm?

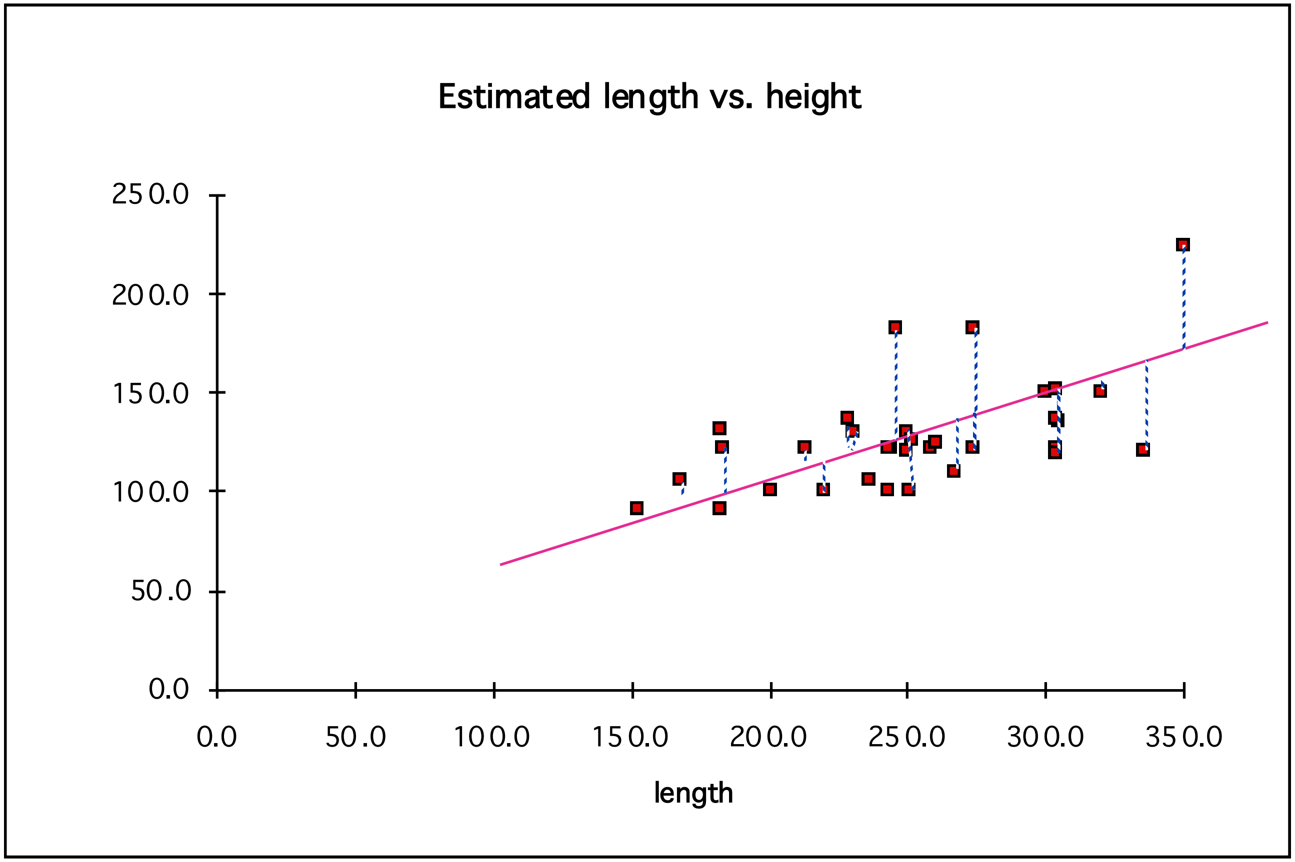


Least-squares regression

- **How do we actually fit the line to our data points?**
- **You can visually try to draw a line across the data point until you are satisfied with the fit, but we would like to have a procedure that is somewhat objective and reproducible.**
- **There are many mathematical ways of fitting a line to a set of data. The oldest and most commonly used is the method of least squares.**

Least-squares regression

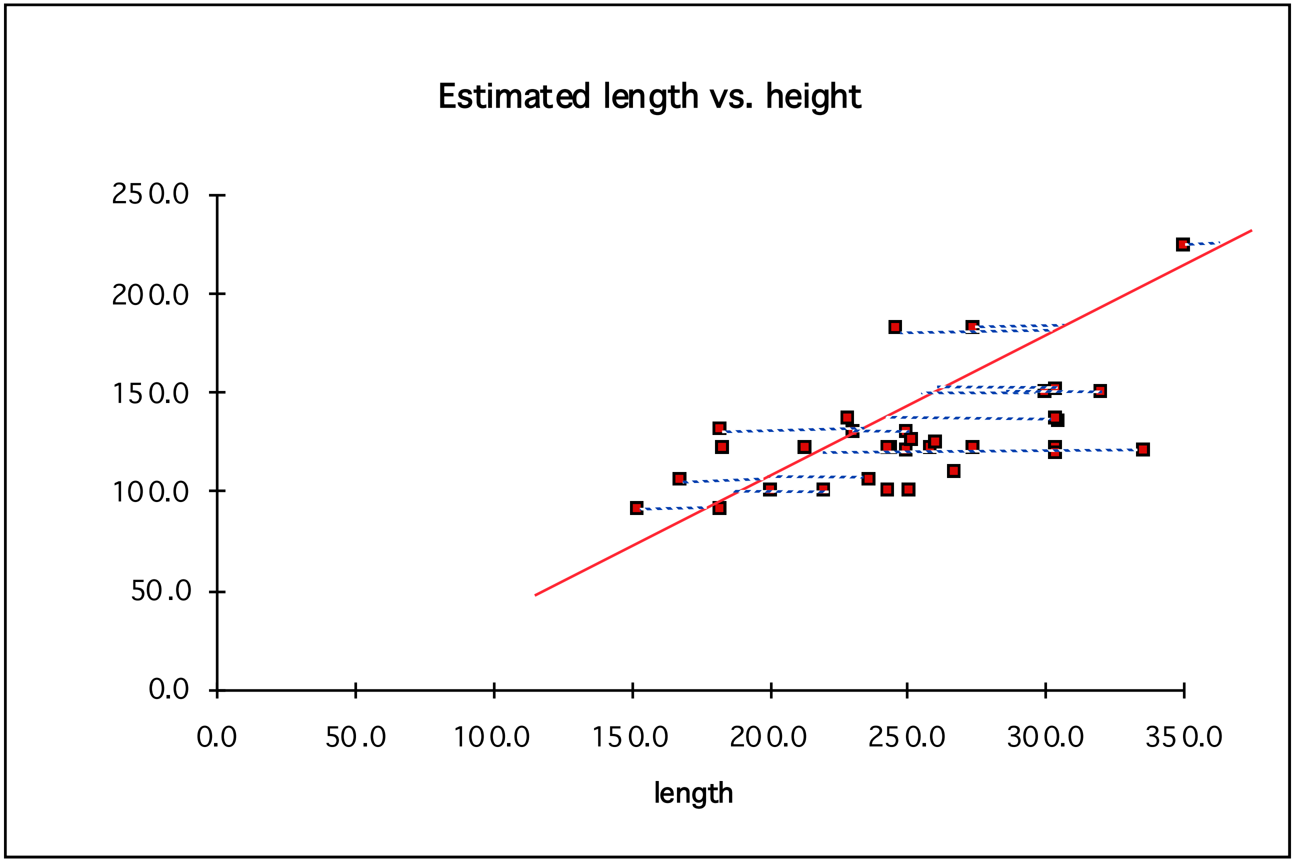
The idea: minimize the sum of squares of the deviations of the data points from the line in the vertical direction.



Most statistical packages implement least-squares regression.

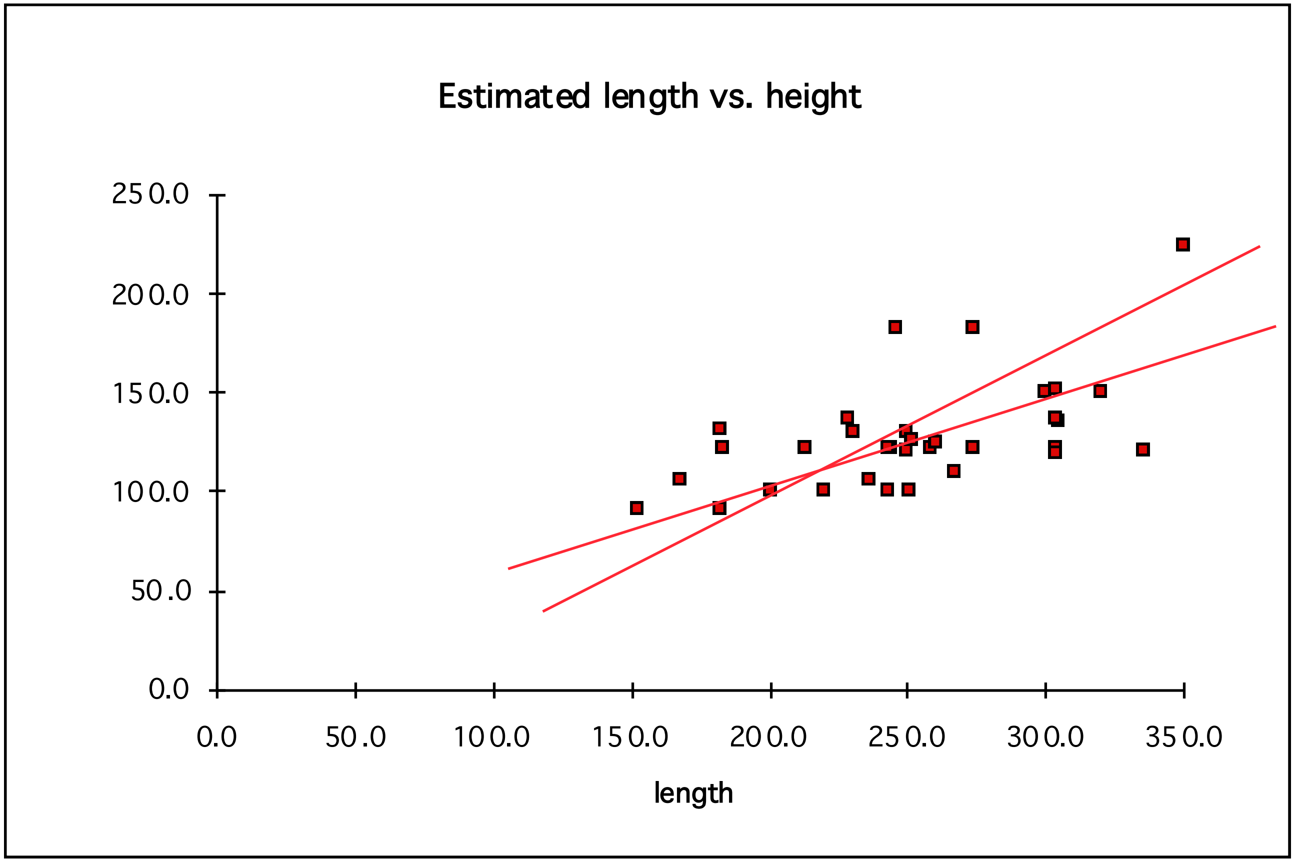
Asymmetry of regression

Choice of explanatory variable affects the parameters of the regression line



Asymmetry of regression

The two regression lines are going to be different in general



Linear regression: An example

Line fitting (or in general curve fitting to the interactions).

***e.g.*, linear regression results of the influence of *tstsc* on *apret* and *apgra* (175 universities).**

```
apret =  13.2 + 1.02 tstsc, R-sq(adj) = 50.5%
apgra = -78.7 + 2.04 tstsc, R-sq(adj) = 62.0%
```

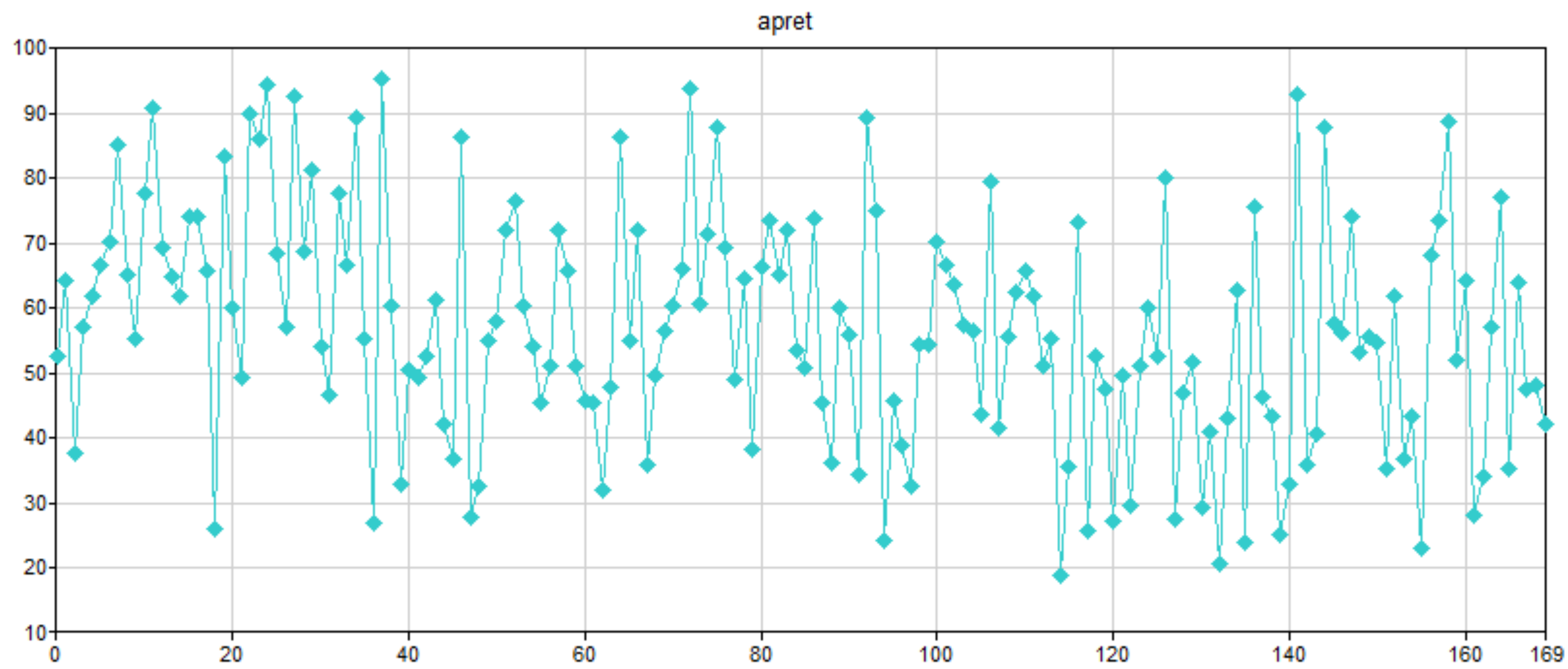
Can be also in multidimensional space.

***e.g.*, linear regression results of the influence of *tstsc* and *top10* on *apret* and *apgra* (175 universities).**

```
apret =  33.4 + 0.142 top10 + 0.634 tstsc, R-sq(adj) = 52.6%
apgra = -68.4 + 0.0283 top10 + 1.87 tstsc, R-sq(adj) = 62.5%
```

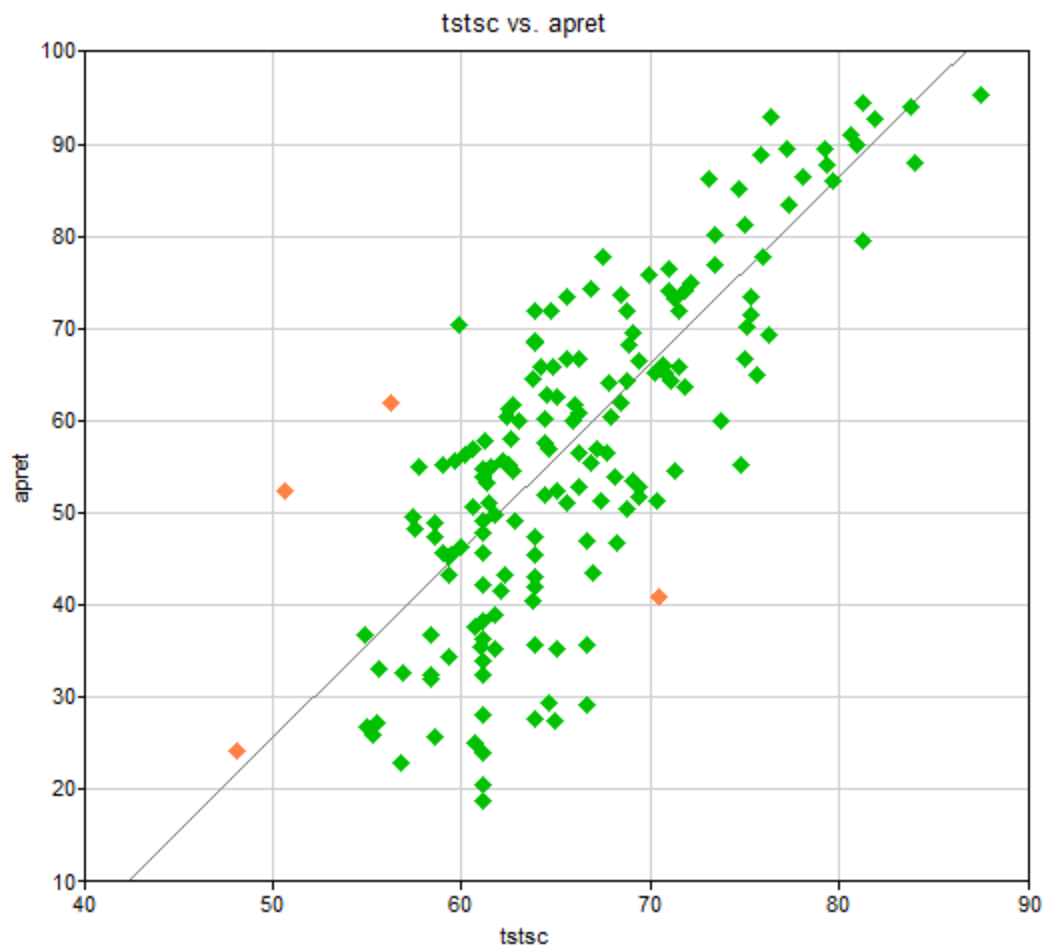
Time series

- Measurements of variables that vary over time.
- This is often a matter of assumption: regular, static data also vary over time but we assume that they do not.



Outliers

- Values that come about because of errors in measurements, transcription, etc., or because of momentary failure in our assumptions.
- We remove them because they are potentially violating our assumptions.
- How to distinguish them? Typically done "manually." Visual inspection is usually very helpful.



Bayesian Probability Theory

Basic Notations

- Random variable
  - An element / event whose status is unknown
  - A = “It will snow tomorrow.”
- Domain
  - The set of values a random variable can take:
  - “A = The coin will flip to Head side”: Binary
  - “A = Number of Steelers wins in 2015”: Discrete
  - “A = % change in Facebook stock in 2015”: Continuous

Axioms of Probability (Kolmogorov's Axioms)

1. $0 \leq P(A) \leq 1$
2. $P(true) = 1$, $P(false) = 0$
3. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

$P(H) = 0.5$

$P(T) =$

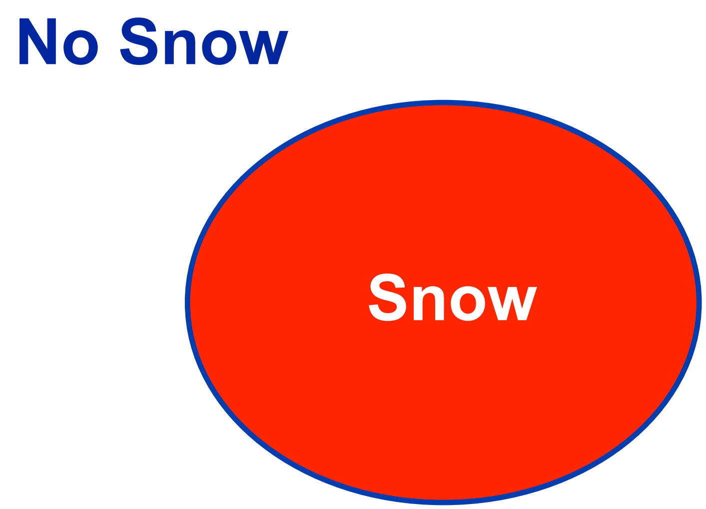
$P(H,H) =$

$P(H,H,H) =$

$P(X_1 = X_2 = X_3) =$

Prior (Belief or Knowledge)

Degree of belief in an event in the absence of any other information



P(Snow Tomorrow) = 0.9
P(No Snow Tomorrow) = 0.1

Conditioning (Conditional Probability)

- Probabilistic conditioning specifies how to revise beliefs based on new information.
- Take all background information into account. This gives the prior probability.
- For Example:

P(Slept in class) = 0.5

P(Slept in class | liked class) = 0.25

P(Didn't sleep in class | liked class) = 0.75

Slept	Liked
0	1
0	1
1	0
0	0
1	0
1	1
0	1
1	0

Product Rule

Definition of conditional probability:

$$P(X_1 \mid X_2) = \frac{P(X_1, X_2)}{P(X_2)}$$

Product rule gives an alternative, more intuitive formulation:

$$P(X_1, X_2) = P(X_1 \mid X_2)P(X_2) = P(X_2 \mid X_1)P(X_1)$$

Product rule general form:

$$P(X_1, \ldots, X_n) = P(X_1, \ldots, X_t \mid X_{t+1}, \ldots, X_n)P(X_{t+1}, \ldots, X_n)$$

Chain Rule

Chain rule is derived by successive application of product rule:

$$= P(X_1,\ldots,X_{n-1},X_n)$$

$$= P(X_1,\ldots,X_{n-1})P(X_n \mid X_1,\ldots,X_{n-1})$$

$$= P(X_1,\ldots,X_{n-2})P(X_{n-1} \mid X_1,\ldots,X_{n-2})P(X_n \mid X_1,\ldots,X_{n-1})$$

$$= \ldots$$

$$= P(X_1)P(X_2 \mid X_1)\ldots P(X_{n-1} \mid X_1,\ldots,X_{n-2})P(X_n \mid X_1,\ldots,X_{n-1})$$

$$= \prod_{i=1}^{n} P(X_i \mid X_1,\ldots,X_{i-1})$$

Bayes theorem

An easy to prove theorem, derived from the product rule:

From

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

and

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

we have

$$\boxed{P(A|B)} = P(B|A)\,/\,P(B)\,\boxed{P(A)}$$

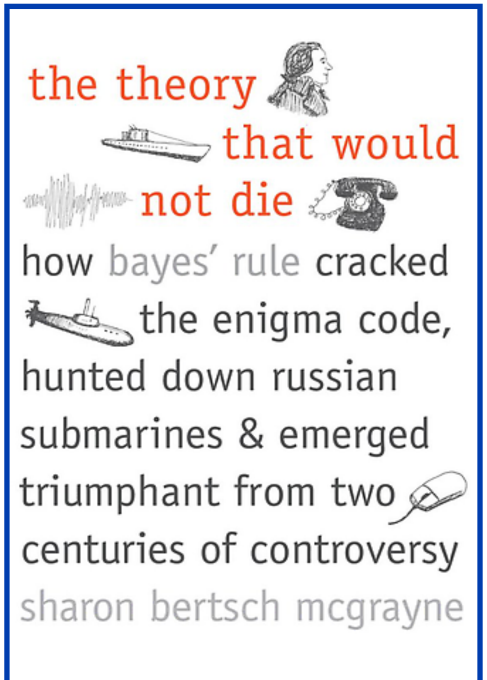
Posterior (a.k.a. a-posteriori) probability

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

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

Bayes theorem and Bayesian statistics

A versatile and powerful approach 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)



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

Bayesian modeling is reliable and it solves hard problems.

It can use both, data and expert knowledge.

Recommended video:
http://www.youtube.com/watch?v=8oD6eBkjF9o

What is the relation of Bayesian statistics to classical statistics?

What is p()?

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?

- 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.
- 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.

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(S)/P(+) (Bayes theorem)

P(+) = P(+|S) P(S) + P(+|~S) P(~S) (theorem of total probability)

P(+) = 0.98 0.001 + 0.05 0.999 = 0.05093

P(S|+) = 0.98*0.001 / 0.05093 0.001

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

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

0.01924

Joint Probability Distribution

Joint probability distribution

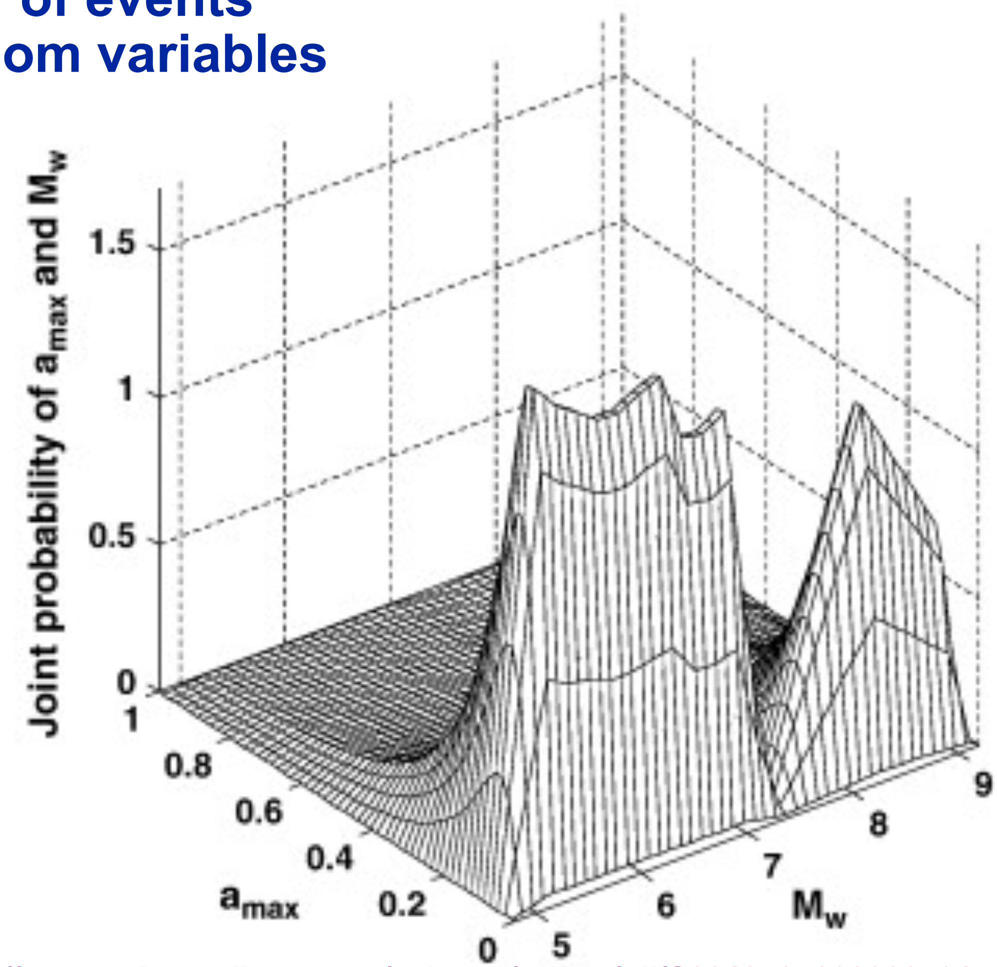
Expresses the probability of events defined over several random variables

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

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

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

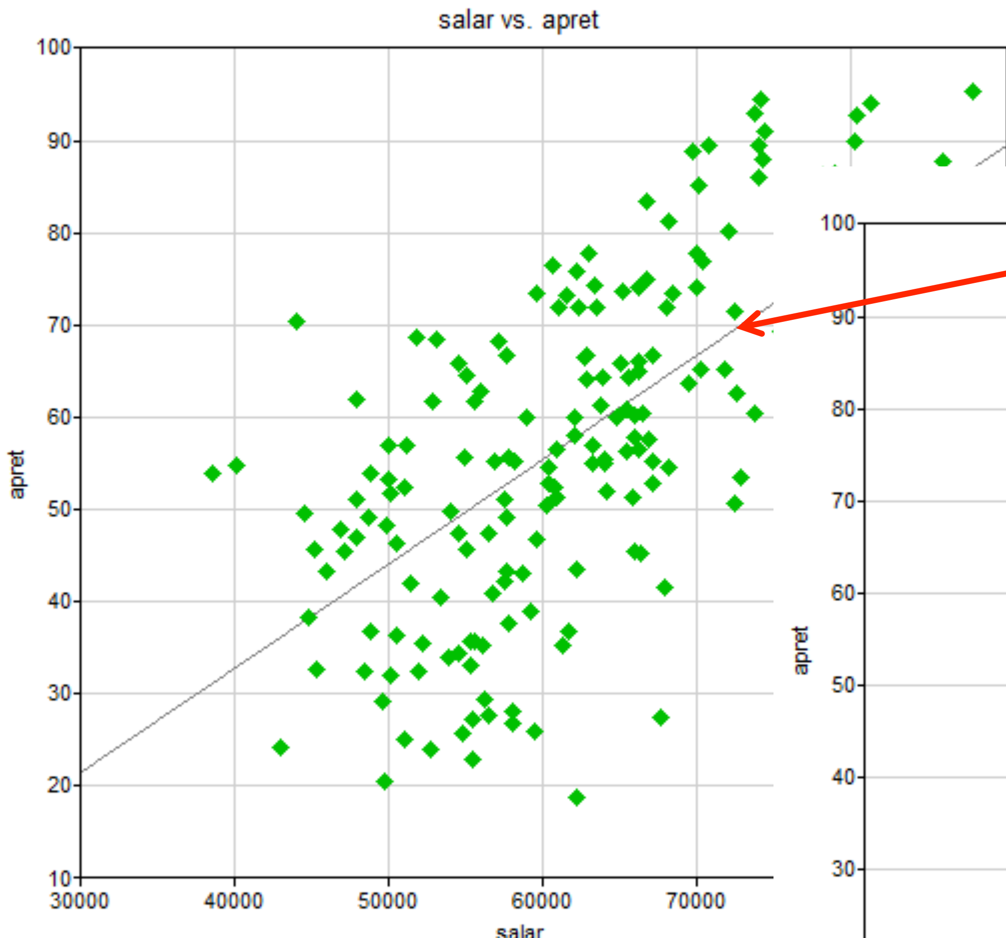
Why?

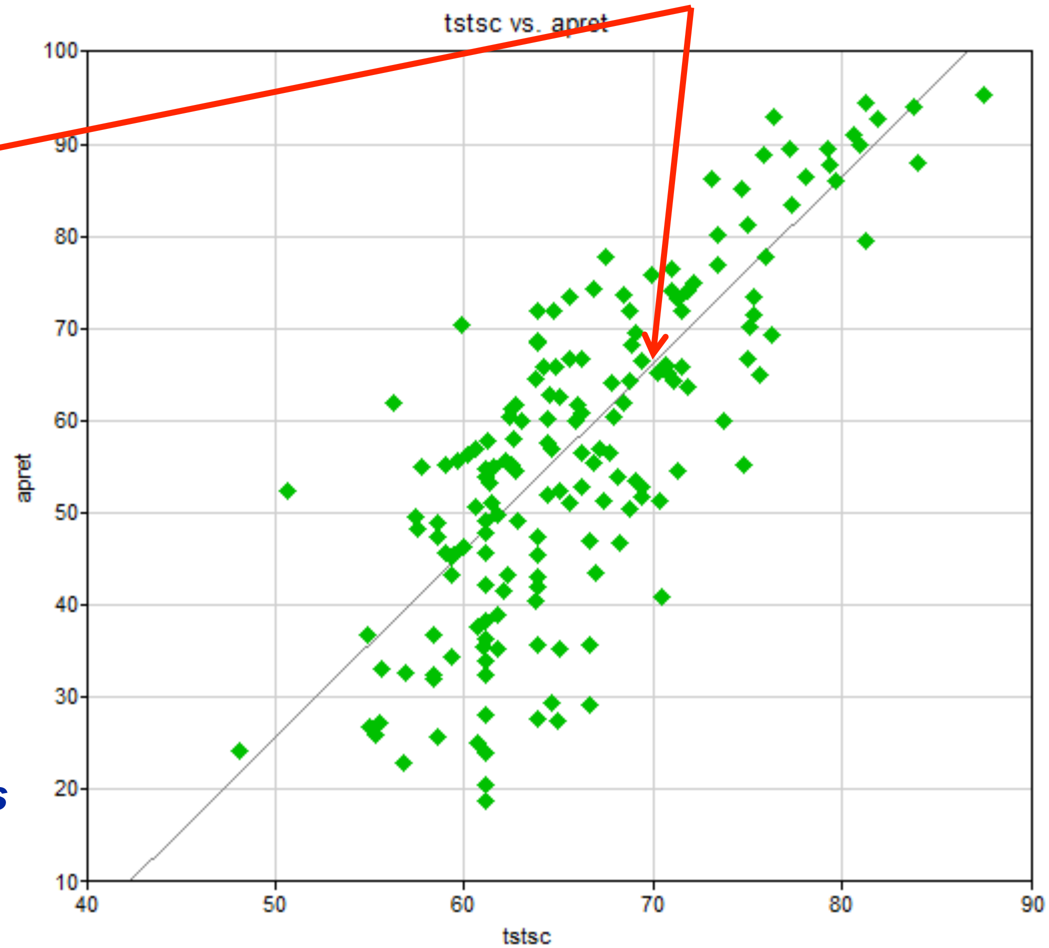
Given the value of some of the variables in the join probability distribution, we can estimate the probability distributions over the remaining variables.

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

Joint probability distributions

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



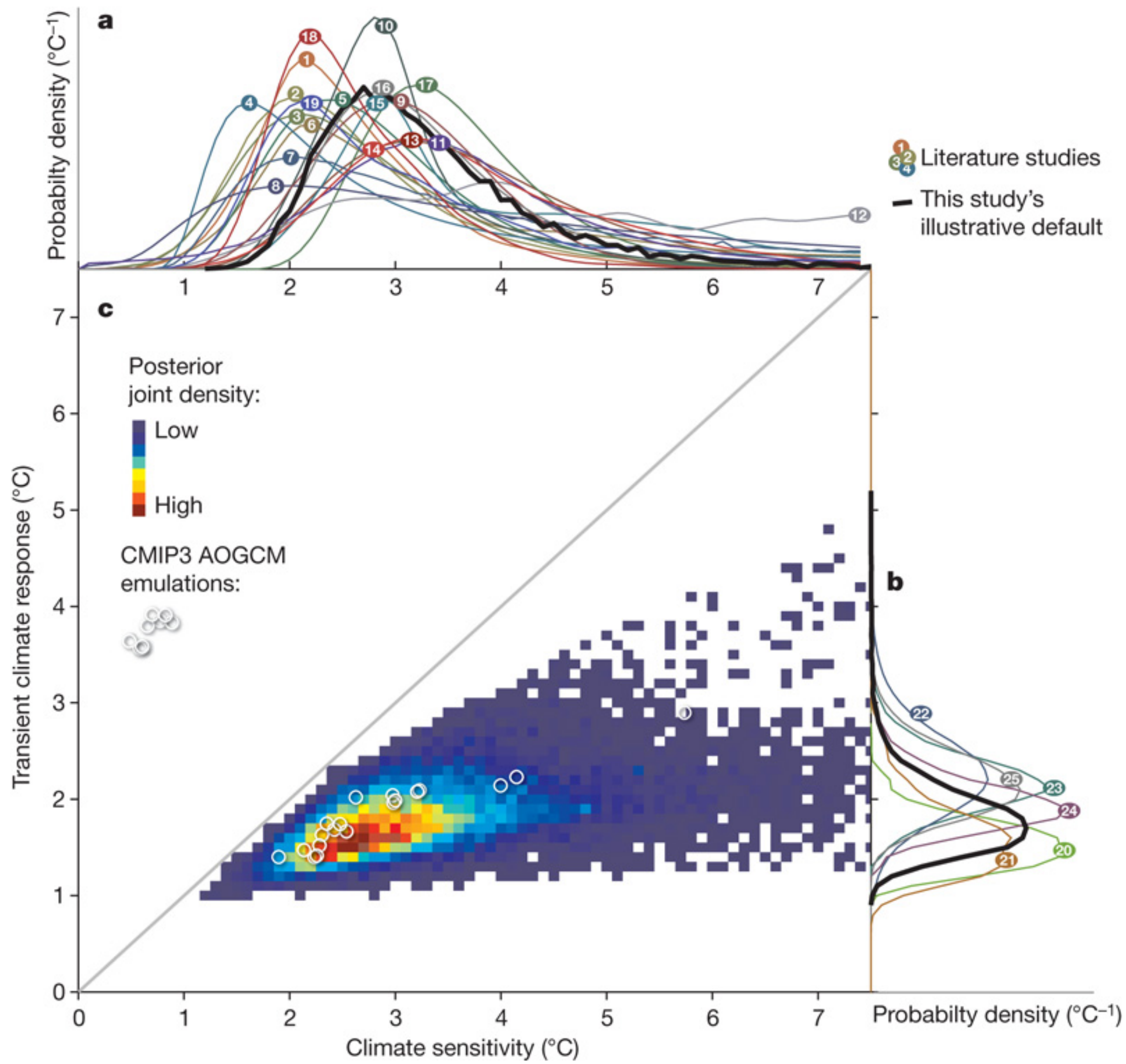


Plots of data known as *scatter plots* give an idea of the joint probability distribution between two variables.

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.

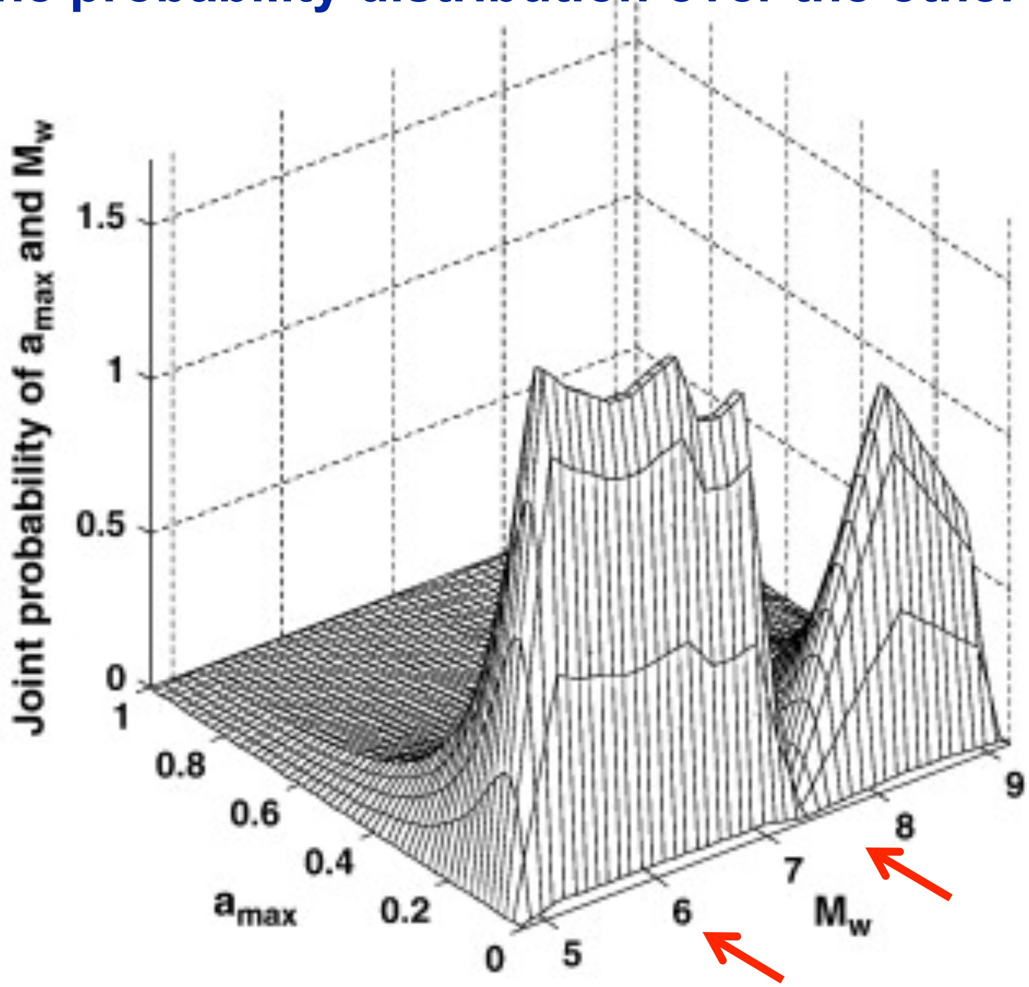


Source: http://www.nature.com/nature/journal/v458/n7242/fig_tab/nature08017_F1.html

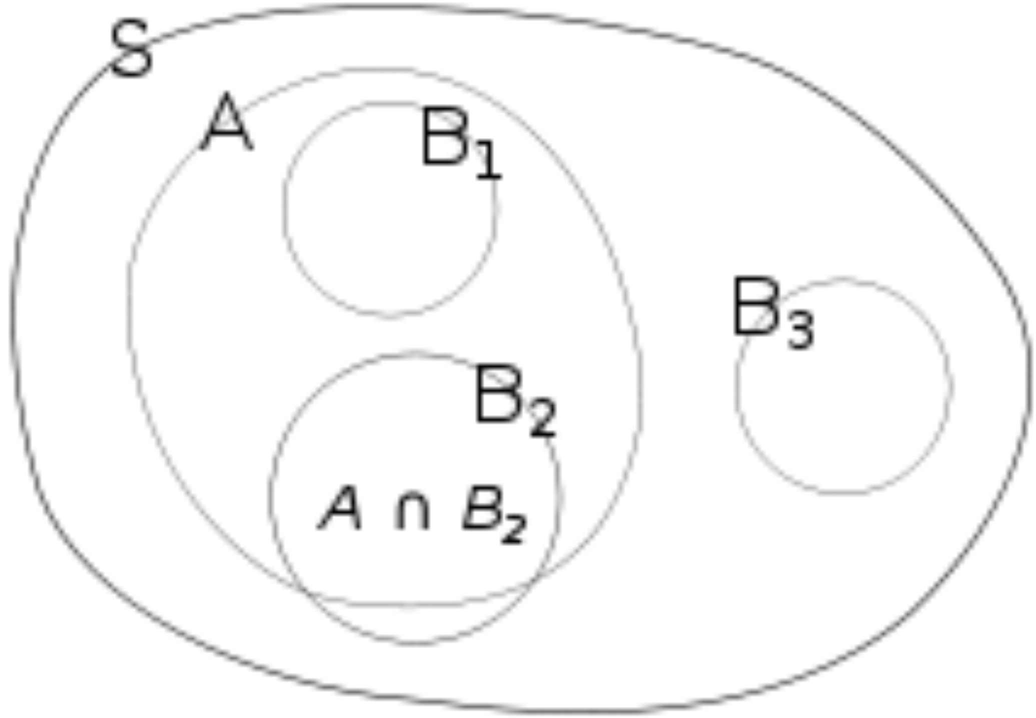
Conditional probability distribution

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



Venn diagrams

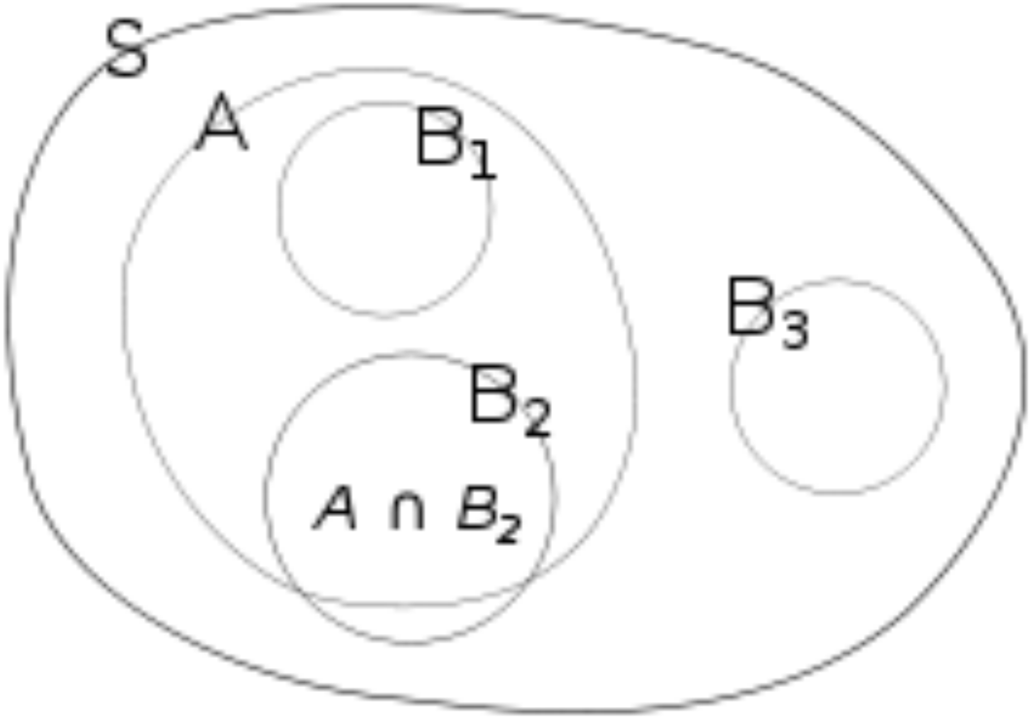


Source: http://en.wikipedia.org/wiki/Conditional_probability

Conditional probability

Definition: P(A|B) = P(A,B) / P(B)

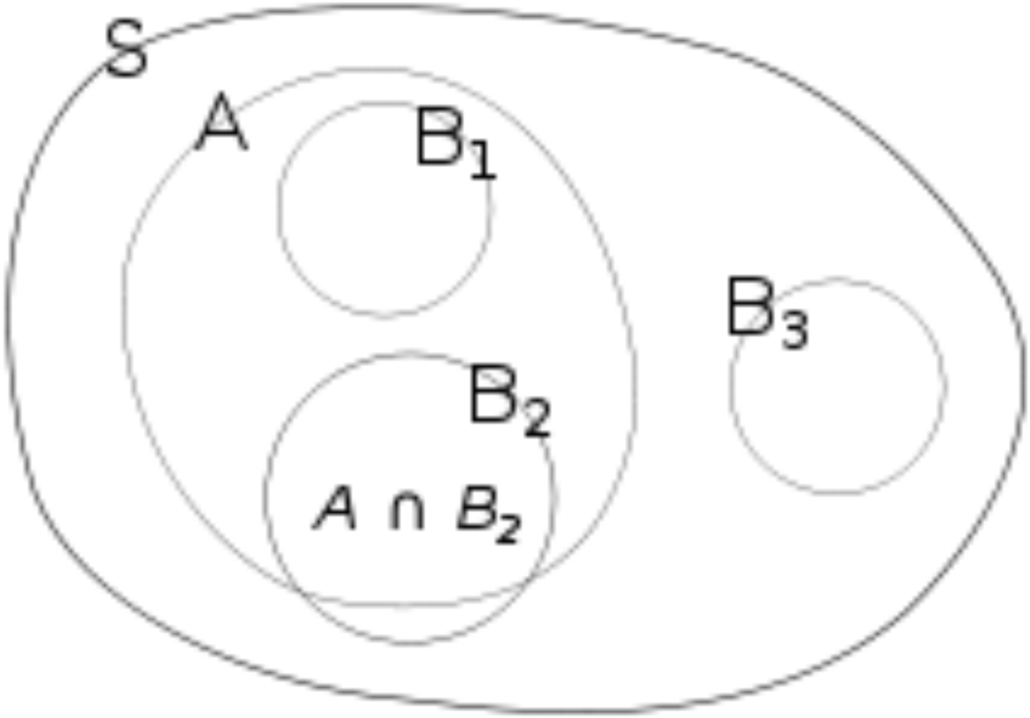
$P(A|B_1) = ?$
$P(A|B_2) = ?$
$P(A|B_3) = ?$



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

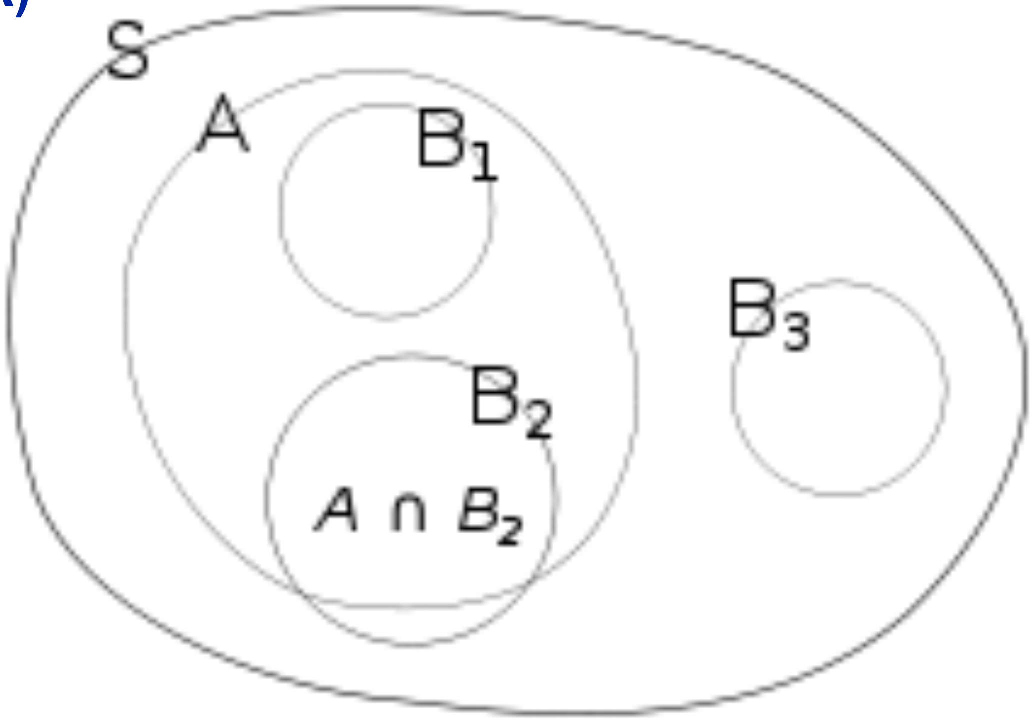
The following is straightforward to derive from the definition of independence (assuming P(B) > 0):

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

$A \perp B_1$?

$A \perp B_3$?

$A \perp B_2$?



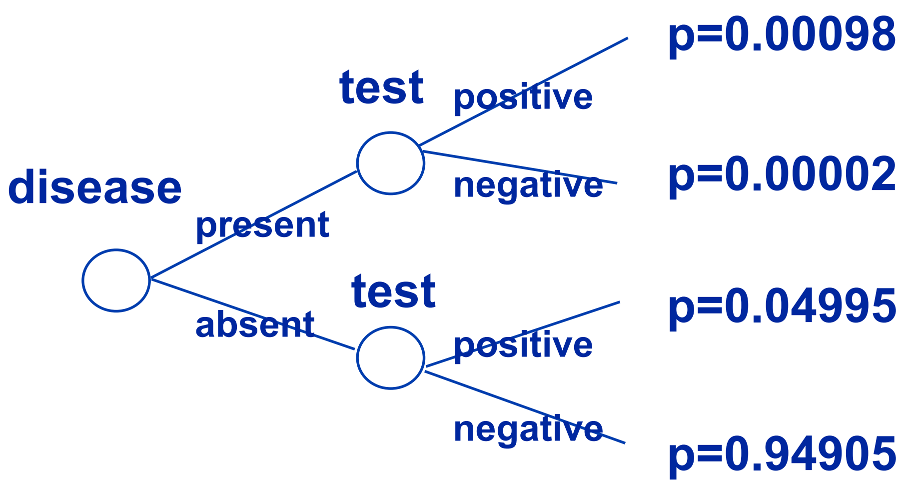
Representations of the Joint Probability Distribution

Probabilistic knowledge representations

- A probabilistic (Bayesian) model encodes the *joint probability distribution* over its variables.
- Knowledge of the joint probability distribution is sufficient to derive any marginal and conditional probability over the model's variables.

Probability trees

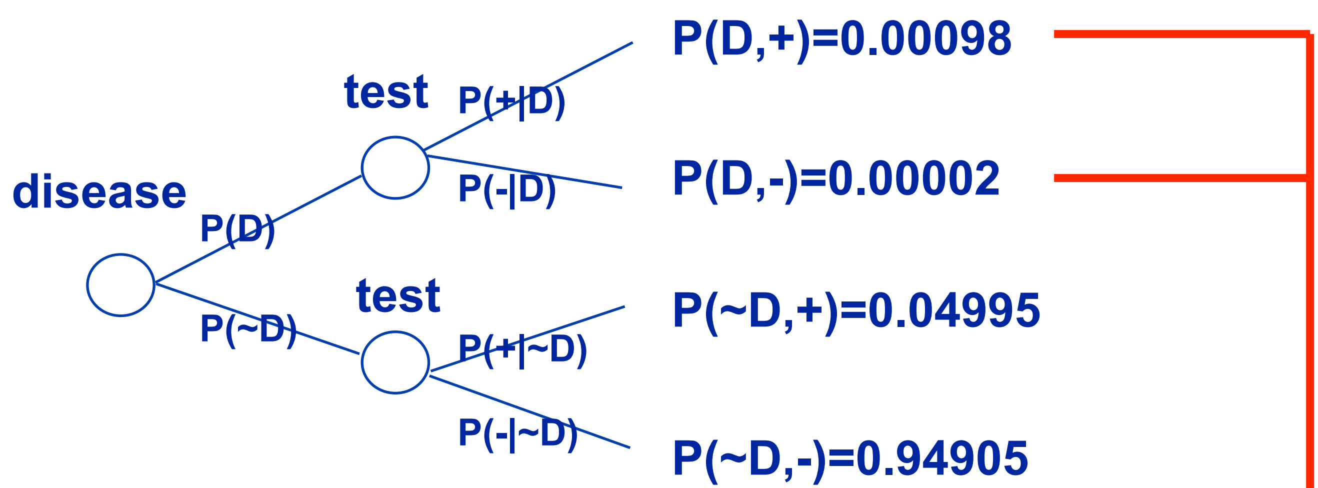
The simplest and quite natural graphical representation of a joint probability distribution over discrete variables



P(disease present ∧ test positive) = P(D ∩ +) = 0.00098
P(disease present ∧ test negative) = P(D ∩ +) = 0.00002
P(disease absent ∧ test positive) = P(~D ∩ +) = 0.04995
P(disease absent ∧ test negative) = P(~D ∩ +) = 0.94905

Computation in probability trees

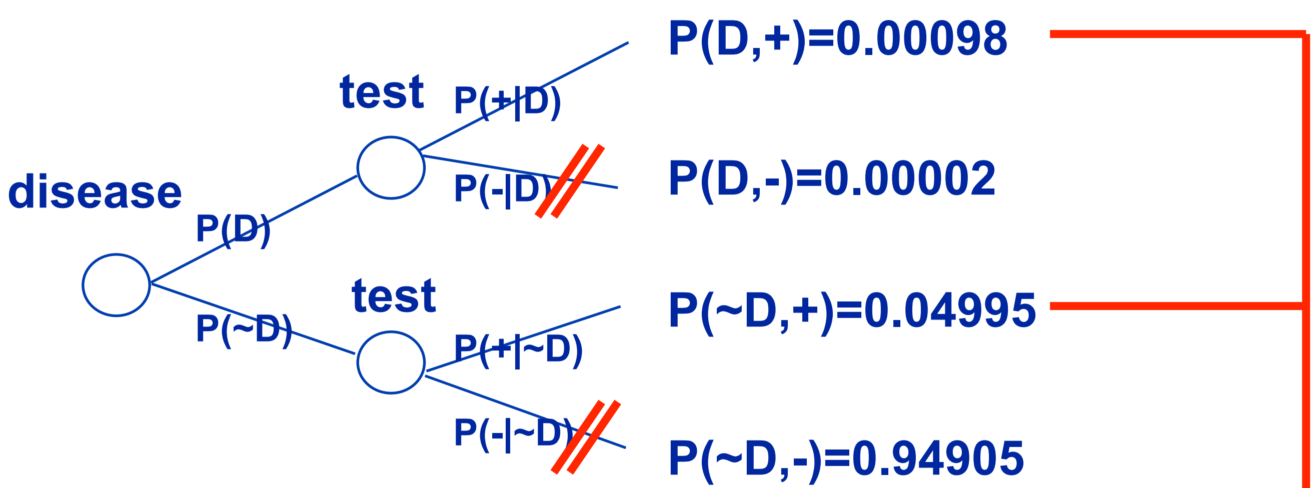
We can calculate any marginal or conditional probability distribution from the joint probability distribution encoded in the tree.



What is the probability of the disease present?
P(D) = 0.00098+0.00002 = 0.001

Computation in probability trees

We can calculate any marginal or conditional probability distribution from the joint probability distribution encoded in the tree.

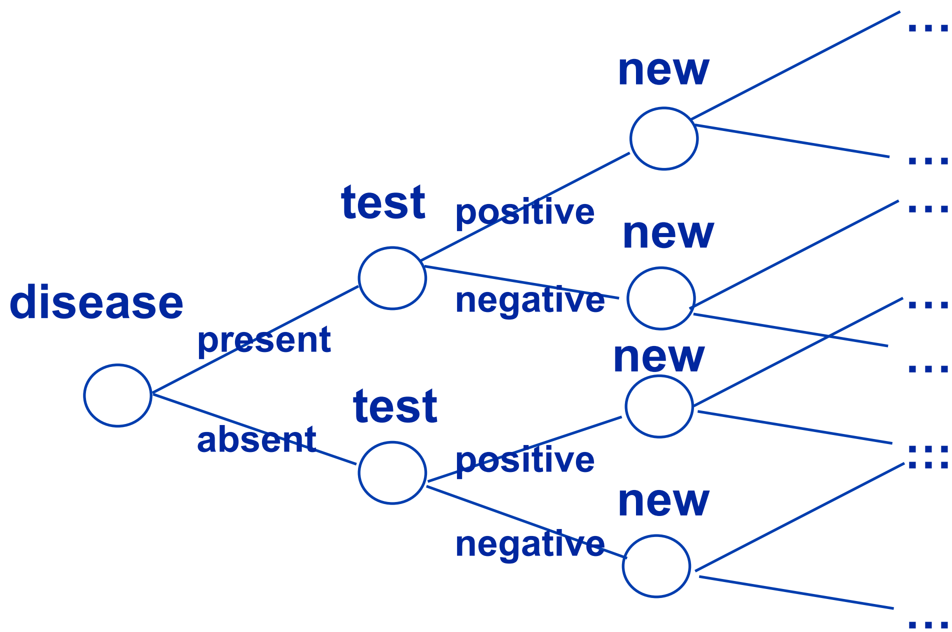


What is the probability of the disease present given a positive test result? Observation of a positive test result makes some of the branches of the tree impossible. What we need to do is just renormalize the remaining, possible (i.e., those that are compatible with the evidence) branches!

P(D|+) = 0.00098/(0.00098+0.04995) ≈ 0.01924

What is wrong with probability trees?

Trees grow exponentially with the number of variables



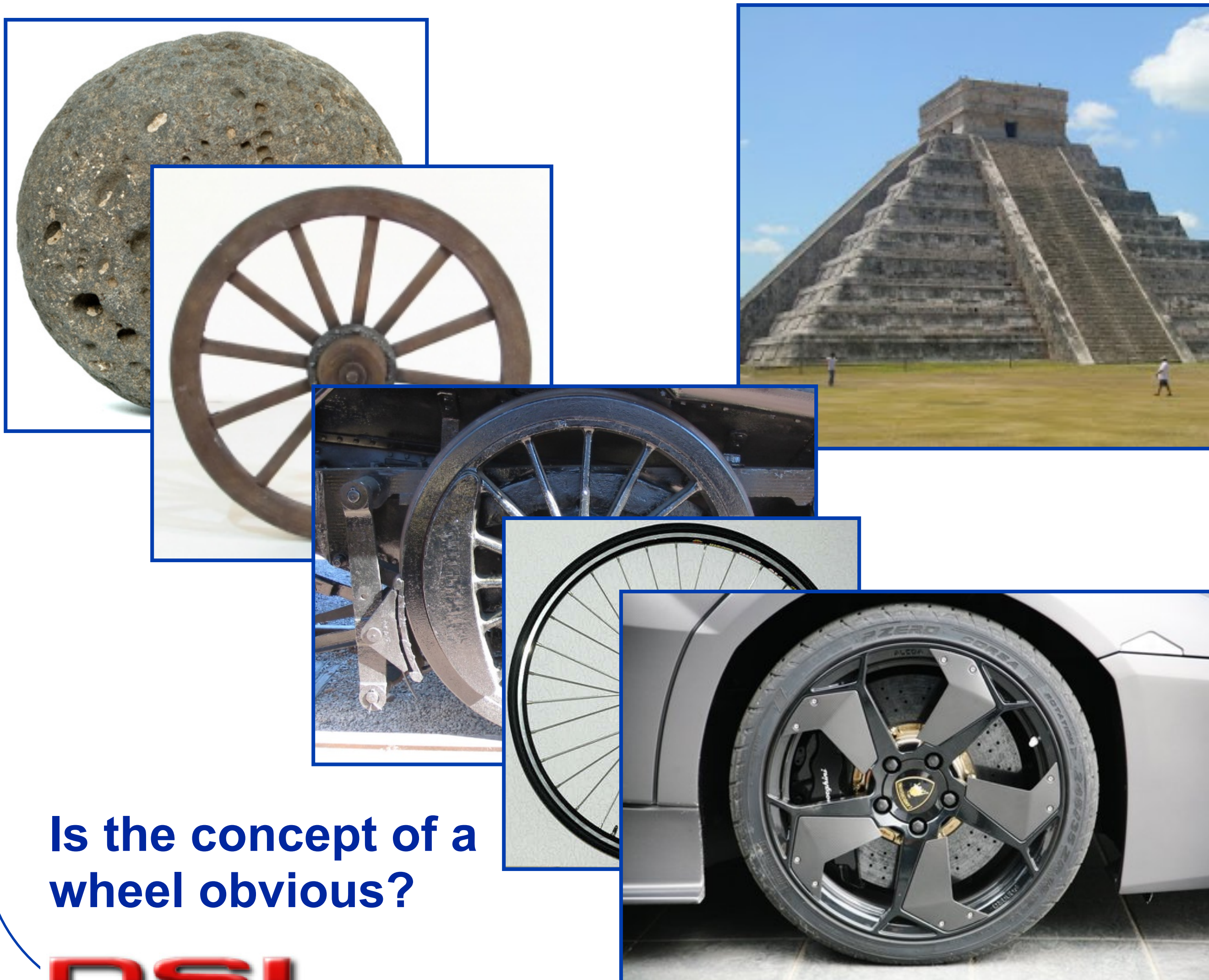
For n binary variables, we will have 2^n branches.
When n=10, the total number of branches is 1,024
When n=11, it is 2,048
…
When n=20, it is 1,048,576 (which is a lot ☺)

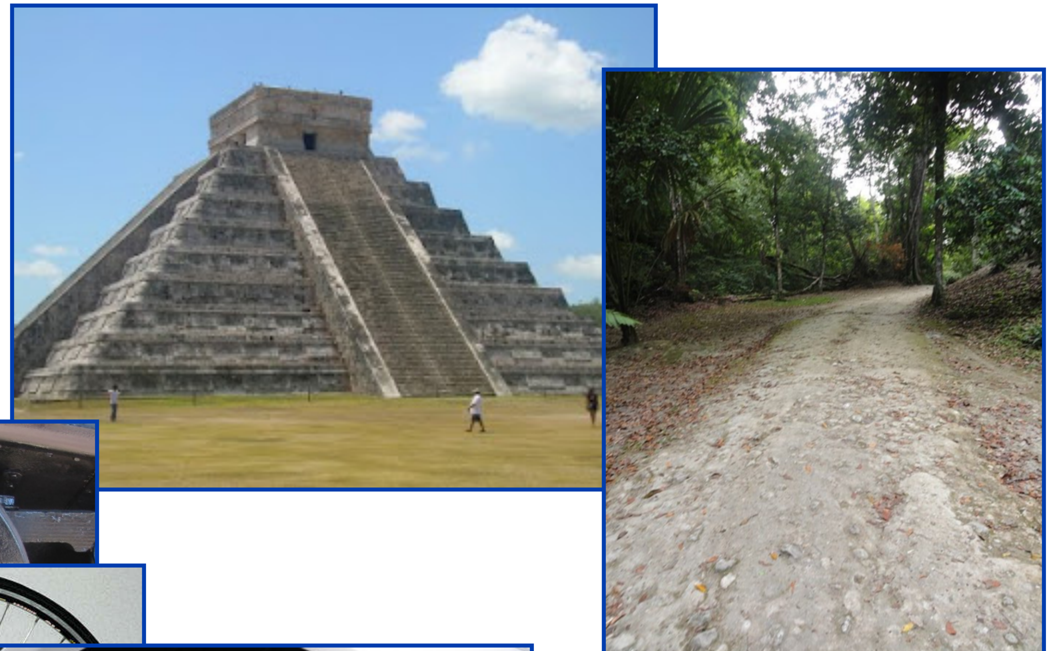
Great idea (only 30-40 years old)

Use independences among variables in the joint probability distribution to reduce the number of parameters in its representation!

Due to seminal work on probabilistic independence by A. Philip Dawid and Judea Pearl

All brilliant ideas are obvious (once we have them ☺)





Is the concept of a wheel obvious?

Then why none of the civilizations in the Americas had it?

Factorability of the joint probability distribution

Every joint probability distribution can be factorized, i.e., rewritten as a product of prior and conditional probability distributions of each of the model's variables

$$f(X_1, X_2, \ldots, X_n) = f(X_1 \mid X_2, X_3, \ldots, X_n)\, f(X_2 \mid X_3, \ldots, X_n) \ldots f(X_{n-2} \mid X_{n-1}, X_n)\ f(X_{n-1} \mid X_n)\, f(X_n)$$

e.g., four variables (a, b, c, d), we have:

$$P(A,B,C,D)=P(A|B,C,D)\ P(B|C,D)\ P(C|D)\ P(D)$$
$$P(A,B,C,D)=P(A|B,C,D)\ P(B|C,D)\ P(D|C)\ P(C)$$
...
$$P(A,B,C,D)=P(B|A,C,D)\ P(D|A,C)\ P(A|C)\ P(C)$$
...

There are n! different directed graphs corresponding to various ways of factorizing a joint probability distribution over n variables.

For n=4, we have 4!=24 different factorizations.

Factorability of the joint probability distribution

- **Any factorization can be simplified if we consider independencies among variables.**
- **Those factorizations that become the simplest are better than others in terms of efficiency of representation.**

e.g., suppose we know that $B \perp D | C$, $D \perp A | C$, and $A \perp C$

We can simplify

$$P(A,B,C,D)=P(B|A,C,D)\,P(D|A,C)\,P(A|C)\,P(C)$$

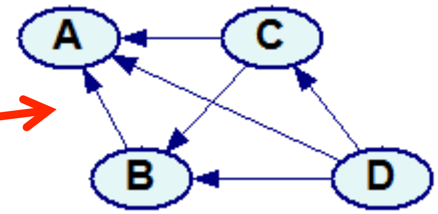
into

$$P(A,B,C,D)=P(B|A,C)\,P(D|C)\,P(A)\,P(C)$$

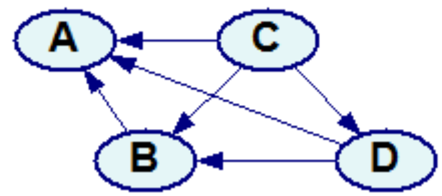
Bayesian networks

- This underlies the very idea of Bayesian networks.
- We draw a directed graph with arc from the conditioning variables to the variables in the factorization.

$P(A,B,C,D)=P(A|B,C,D)\ P(B|C,D)\ P(C|D)\ P(D)$

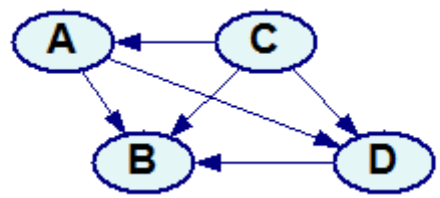


$P(A,B,C,D)=P(A|B,C,D)\ P(B|C,D)\ P(D|C)\ P(C)$



...

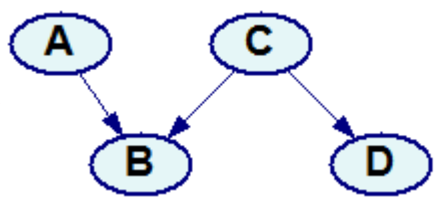
$P(A,B,C,D)=P(B|A,C,D)\ P(D|A,C)\ P(A|C)\ P(C)$



...

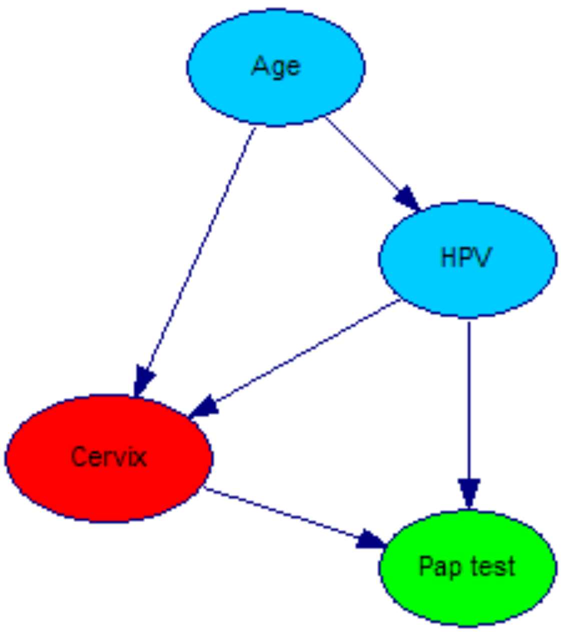
$B \perp D|C,\ D \perp A|C,\ A \perp C$

$P(A,B,C,D)=P(B|A,C)\ P(D|C)\ P(A)\ P(C)$



Bayesian networks

A Bayesian network [Pearl 1988] is a directed acyclic graph (DAG) consisting of:

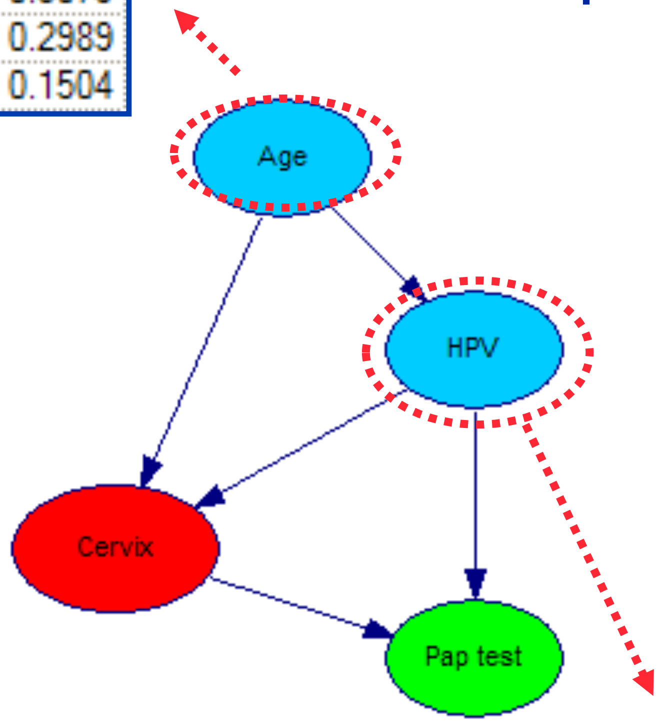


- The qualitative part, encoding a domain's variables (nodes) and the probabilistic (usually causal) influences among them (arcs).
- The quantitative part, encoding the joint probability distribution over these variables.

Bayesian networks: Numerical parameters

a1_below_20	0.0416
a2_20_29	0.2012
a3_29_45	0.3079
a4_45_60	0.2989
a5_60_up	0.1504

Prior probability distribution tables for nodes without predecessors (Age)

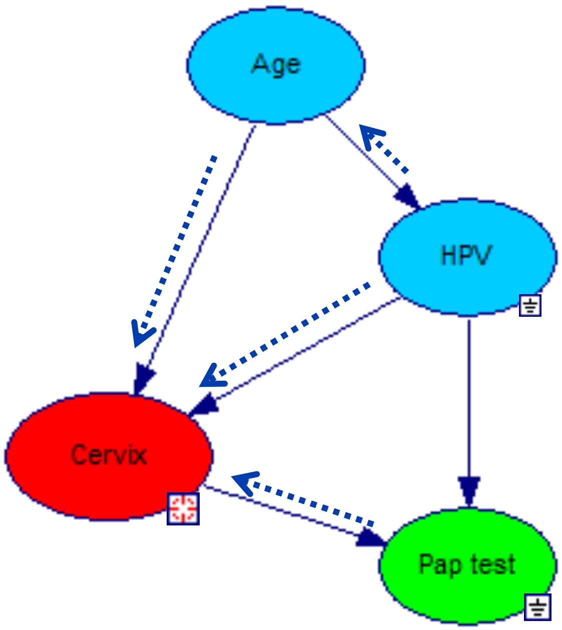


Conditional probability distributions tables for nodes with predecessors (HPV, Pap test, Cervix)

Age	a1_below_20	a2_20_29	a3_29_45	a4_45_60	a5_60_up
NA	0.8652	0.8387	0.7904	0.8055	0.8851
Negative	0.069	0.0901	0.1782	0.1765	0.1012
Positive	0.0613	0.0667	0.0282	0.0142	0.0082
Qns	0.0045	0.0045	0.0032	0.0038	0.0055

Reasoning in Bayesian networks

The most important type of reasoning in Bayesian networks is updating the probability of a hypothesis (e.g., a diagnosis) given new evidence (e.g., medical findings, test results).



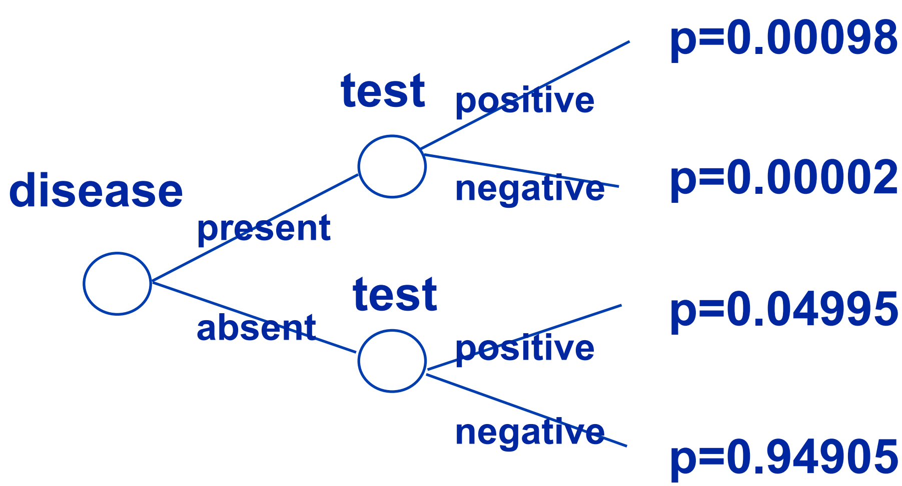
Example:

What is the probability of invasive cervical cancer in a (female) patient with high grade dysplasia with a history of HPV infection?

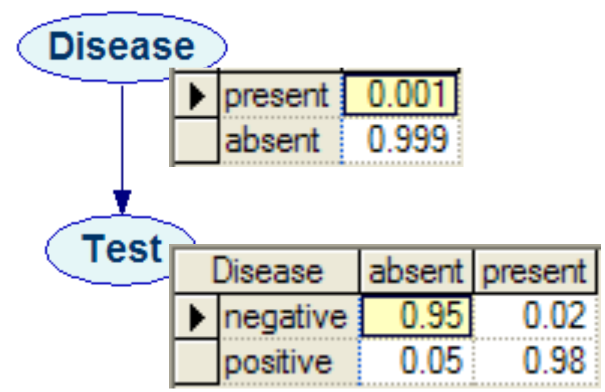
P(CxCa | HPV=positive, HSIL=yes)

Probability trees and Bayesian networks

probability tree

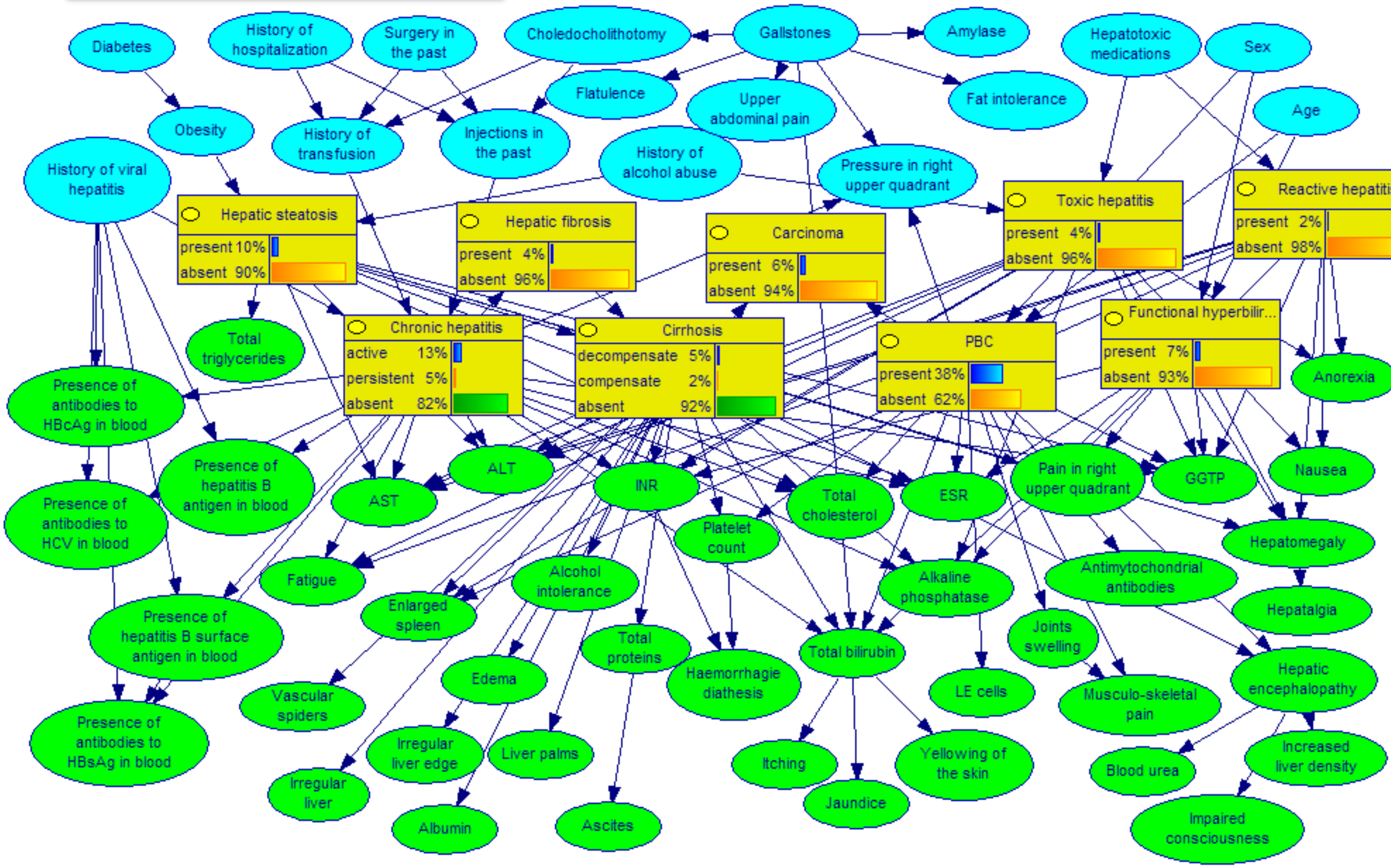


Bayesian network



The two representations are equivalent
But, when there are independences in the domain,
Bayesian networks are much, much more efficient!

HEPAR II Model



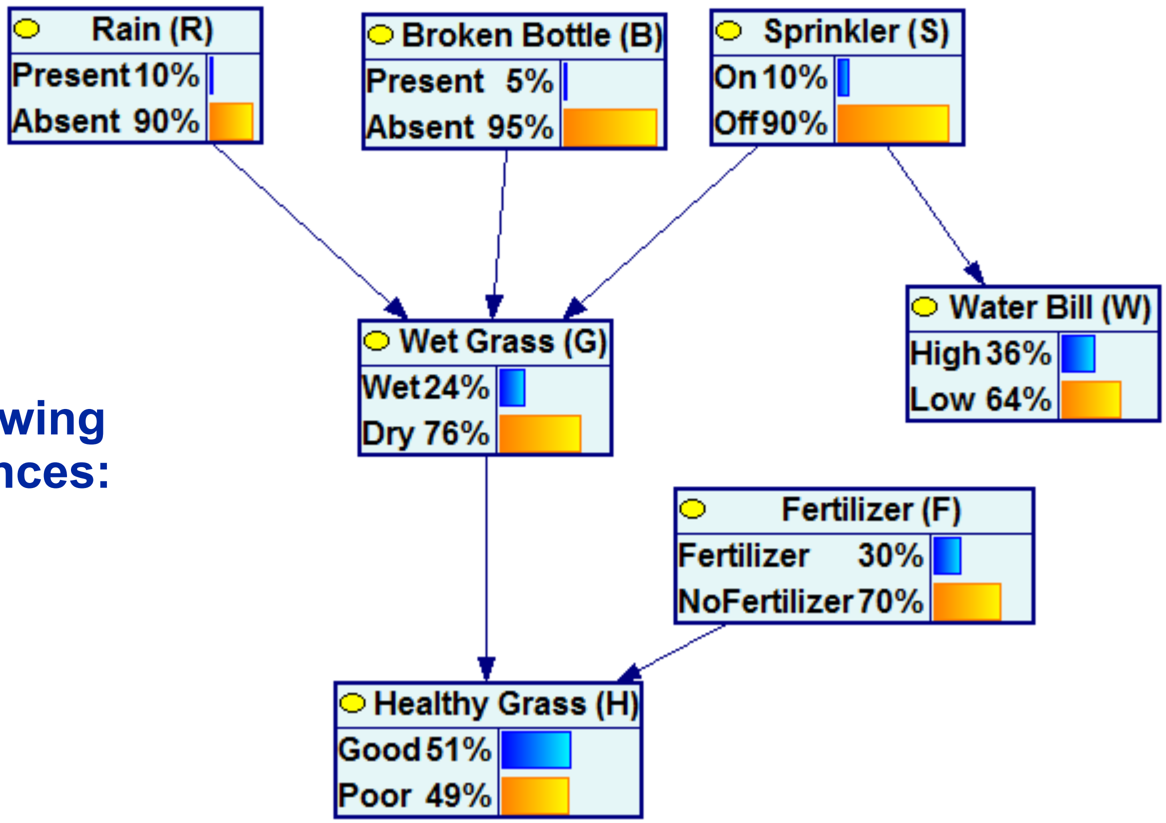
70 variables; 2,139 numerical parameters (instead of over $2^{70}-1 \approx 10^{21}$!)

Independences: Markov condition

- Allows to read back dependences and independences from the graph.
- Informally speaking, it is an assumption that ties directed probabilistic graphs with probability, specifying how a directed graphs represents independence.
- A node is independent of its non-descendants given its predecessors (D-separation).

Markov condition: Example

$P(H,G,W,R,B,S,F)=P(H|G,F)\ P(G|R,B,S)\ P(W|S)\ P(R)\ P(B)\ P(S)\ P(F)$



This graph implies the following (conditional) independences:

$R \perp B$, $R \perp S$, $B \perp S$, $R \perp F$, $B \perp F$, $S \perp F$

$R \perp W$, $B \perp W$, $W \perp F$, $G \perp F$

$R \perp H|G$, $B \perp H|G$, $S \perp H|G$, $W \perp H|G$

$W \perp *|S$

$R \perp W|G,S$, $B \perp W|G,S$

Equation-based systems and graphical models

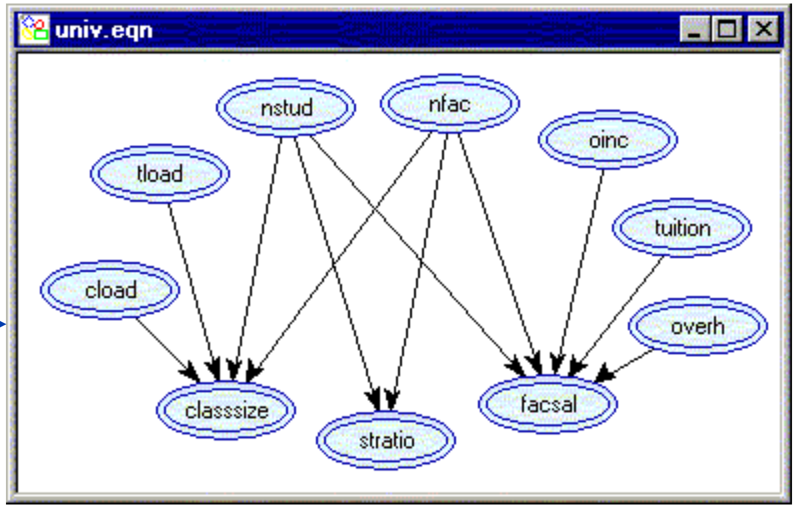
classsize = (*nstud* * *cload*) / (*nfac* * *tload*)
facsal = (*oinc* + *tuition* * *nstud*) / (*nfac* * (1 + *overh*))
stratio = *nstud* / *nfac*

Core equations

cload = 15
tload = 6
nstud = 22102
nfac = 3006
oinc = 30000000
tuition = 12000
overh = 0.48

Equations for exogenous variables

Together they determine the structure of the model



Equation-based systems: Reversibility of causal ordering

classsize **= (*nstud* * *cload*) / (*nfac* * *tload*)**

facsal **= (*oinc* + *tuition* * *nstud*) / (*nfac* * (1 + *overh*))**

stratio **= *nstud* / *nfac***

cload **= 15**

tload **= 6**

nstud **= 22102**

~~***nfac*** **= 3006**~~ **stratio = 10**

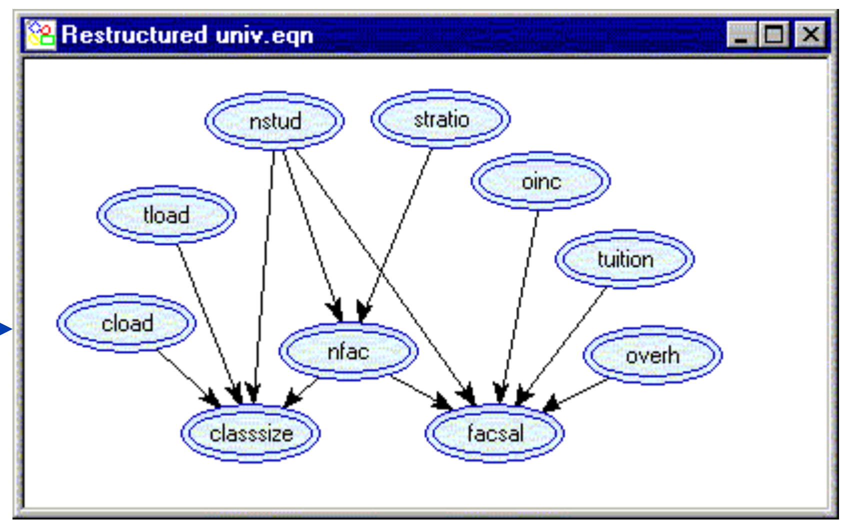
oinc **= 30000000**

tuition **= 12000**

overh **= 0.48**

Setting *stratio* to be exogenous at the expense of *nfac*

The new model structure



Explication of the asymmetries due to Herb Simon (early 1950s)

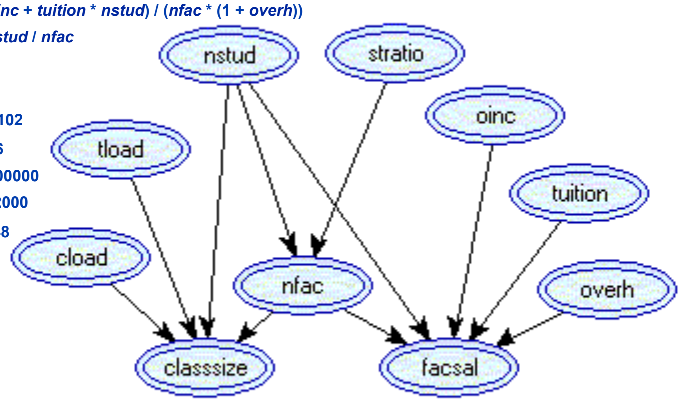
Advantages of directed graphs

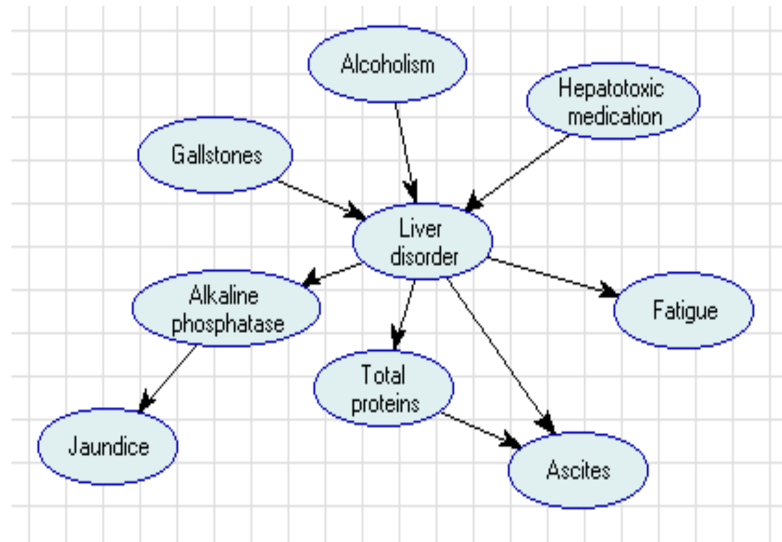
- May be built to reflect the causal structure of a model (helps with obtaining insight into the problem)
- Can accommodate representation of uncertainty
- Can be reconfigured as needed
- Have sound theoretical foundations: We are dealing here with probability theory and decision theory
- We can talk (almost) the same language with statisticians, philosophers, and scientists

Family of directed graphs (a bigger picture)

(a.k.a. “influence nets,” “causal diagrams,” etc.)

*classsize = (nstud * cload) / (nfac * tload)*
*facsal = (oinc + tuition * nstud) / (nfac * (1 + overh))*
stratio = nstud / nfac
cload = 15
tload = 6
nstud = 22102
nfac = 3006
oinc = 30000000
tuition = 12000
overh = 0.48





Both, systems of equations and joint probability distributions can be pictured by directed acyclic graphs.

Further Readings

- GeNie Software Documentation: https://dslpitt.org/genie/wiki/Main_Page
- An Introduction to Statistical Learning with Applications in R: http://www-bcf.usc.edu/~gareth/ISL/ (Chapter 1-2)
- Probabilistic Programming and Bayesian Methods for Hackers: (Chapter 1) http://nbviewer.jupyter.org/github/CamDavidsonPilon/Probabilistic-Programming-and-Bayesian-Methods-for-Hackers/blob/master/Chapter1_Introduction/Chapter1.ipynb
- Causation, Prediction, and Search: https://www.cs.cmu.edu/afs/cs.cmu.edu/project/learn-43/lib/photoz/.g/scottd/fullbook.pdf (Chapter 1)