Uncertainty and Statistical Decision Making

Marek J. Drużdżel

Politechnika Białostocka
Wydział Informatyki

m.druzdzel@pb.edu.pl
http://www.wi.pb.edu.pl/~druzdzel/

Overview

- Scientific inquiry and uncertainty
- Describing and displaying data
- Scientific Inference
- Elements of decision theory
- Comparing the means
- Concluding remarks

Scientific inquiry and uncertainty

***"... 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.**
- **No exception as far as empirical research is concerned.**
- **We need to deal with uncertainty in empirical work.**
- **How?**
- **Of course, using probability theory!**
- **Accidents happen to prepared minds (this is a general principle!), so 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."

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.
- ...

Scientific inquiry and uncertainty

A slightly modified example from Hillel J. Einhorn & Robin M. Hogarth, Judging Probable Cause, *Psychological Bulletin* 99(1)3-19, 1986

> "… imagine that we are ignorant of the cause of birth. However, it has been suggested that sexual intercourse is related to pregnancy, and the following experiment was designed to test this hypothesis:
>
> 100 couples were allocated at random to an intercourse condition, and 100 to an non-intercourse condition."

Sample and sampling

- An important idea: We want to make inference from a sample to a population (unless we can make the entire population a sample)!
- Sampling should be random, giving every member of the population an equal chance of being selected
- 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.).
- No guarantee for a representative sample, but we can get arbitrarily close (in terms of probability) to the population.

Uncertainty: Typical experimental data

We obtain a data set that might look as follows:

	Wife	Husband	Pregnancy
1	Mary Smith	William Smith	Yes
2	Elisabeth Brown	John Brown	No
3	Nancy Green	Philip Green	?
100	Catherine Waters	Benjamin Waters	Yes

Uncertainty: Typical experimental data

What might typical results of such an experiment look like?

	pregnancy	no pregnancy	
Treatment (intercourse)	20	76	96
Control (celibacy)	5	87	92

(BTW: this is usually called a *contingency table*)

It looks quite messy. What should we conclude?
Does intercourse cause pregnancy or not?

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.

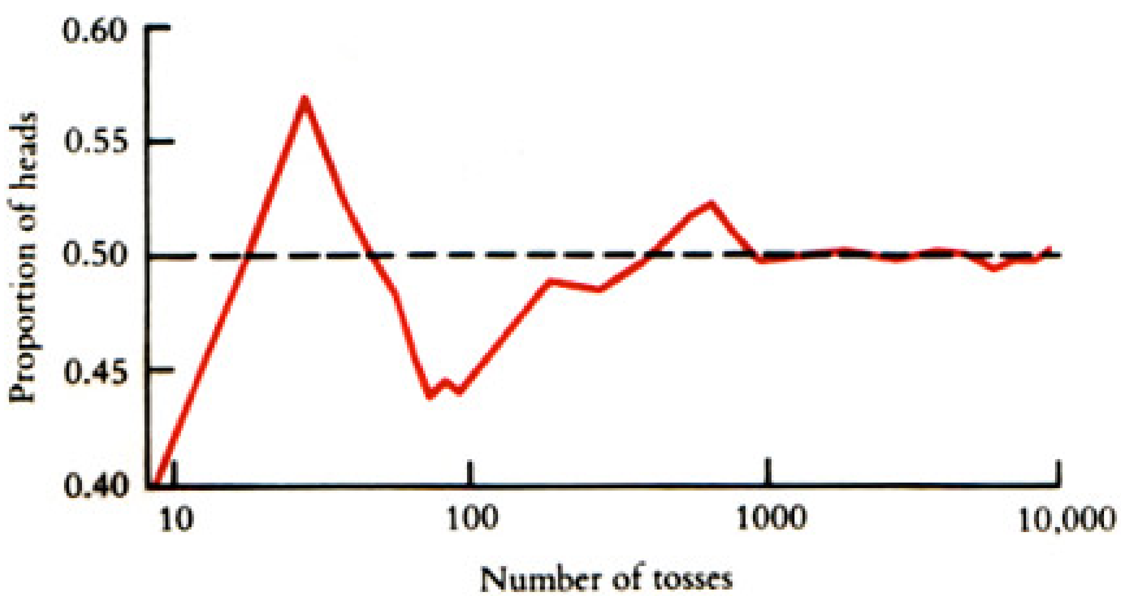


Figure 7.2 Percent of heads versus number of tosses in Kerrich's coin-tossing experiment. [David Freedman et al., *Statistics* Norton, 1978.]

Scientific inquiry as decision making

An outcome like this is not atypical for empirical studies.

Some sources of uncertainty:

- **Errors in measurement (the pregnancy test results might have been false positive or false negative).**
- **No control over subjects (who knows whether they have understood and followed the experimenter's instructions).**
- **Latent variables that we did not control for (other causes or inhibitors of pregnancy?).**
- **Subject *mortality* (some couple might have dropped out of the study, some did not show up for the test).**
- **Biased sample of subjects.**
- **Bad luck.**
- **...**

So, what should we do?

Scientific hypothesis testing is an instance of decision making under uncertainty!

Some features of scientific inquiry

Some features of scientific inquiry (based on our example):

- Scientific questions concern systems, i.e., parts of the real world that can be reasonably studied in separation.
- We are usually trying to make inferences from a sample to a population. This is to make the whole enterprise practical.
- Choosing a sample is important: If it is biased, our inferences will be invalid!
- We usually choose items randomly and give every member of the population an equal chance of being selected.
- We are working under many constraints.
- We will always have to deal with uncertainty, so we'd better get used to it and learn to deal with it.
- Scientific questions often concern causal relations.
- Basic logic and common sense are essential for structuring the problem and hypotheses. Subsequently we collect information about the system and test.

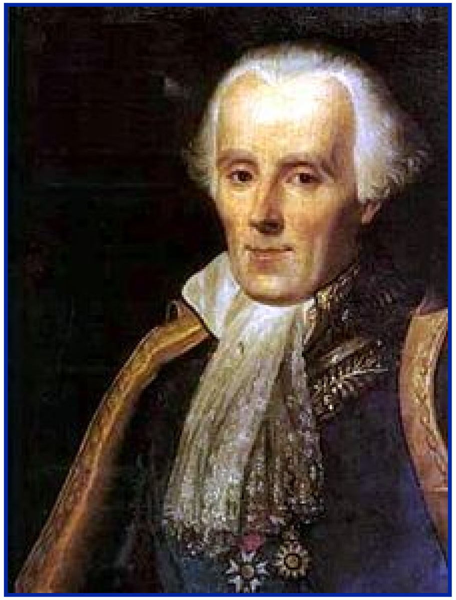
A Brief Review of Probability Theory and Statistics

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.**

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

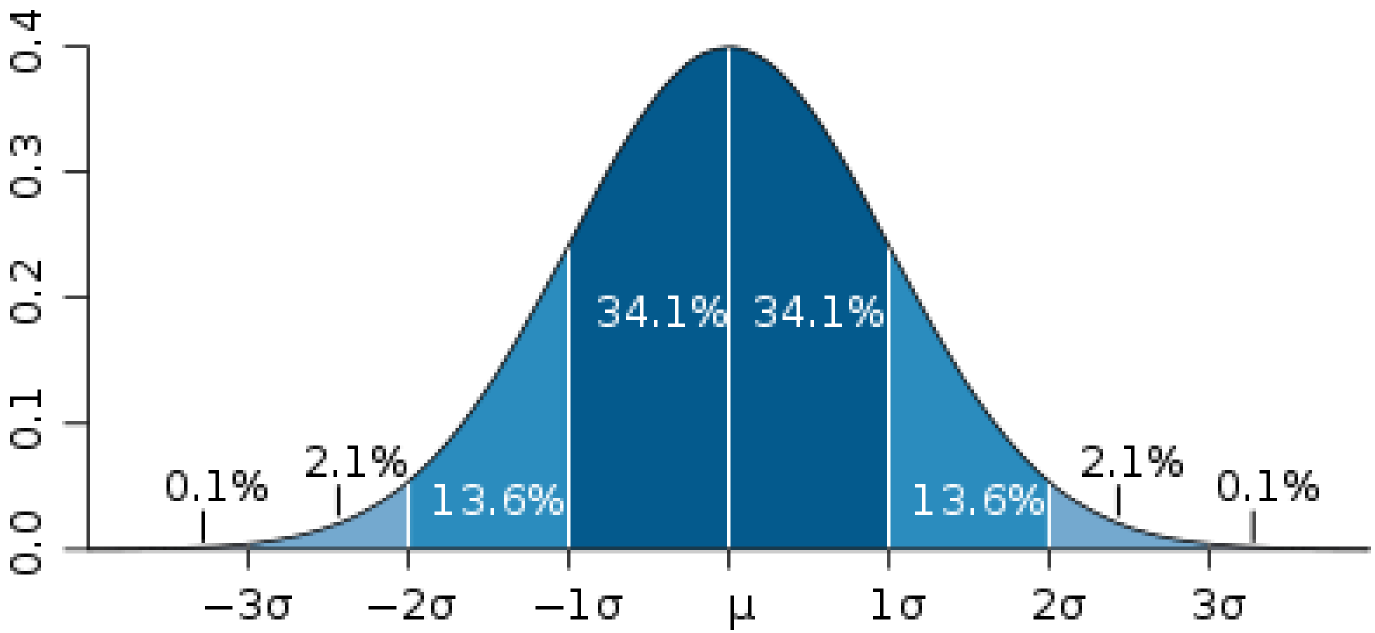
Apgra	
Mean	56.72107647
Median	55.7085
Mode	72
Standard Deviation	18.07709676
Variance	326.7814274
Kurtosis	-0.554450128
Skewness	0.089185832
Range	76.5
Minimum	18.75
Maximum	95.25
Sum	9642.583
Count	170

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

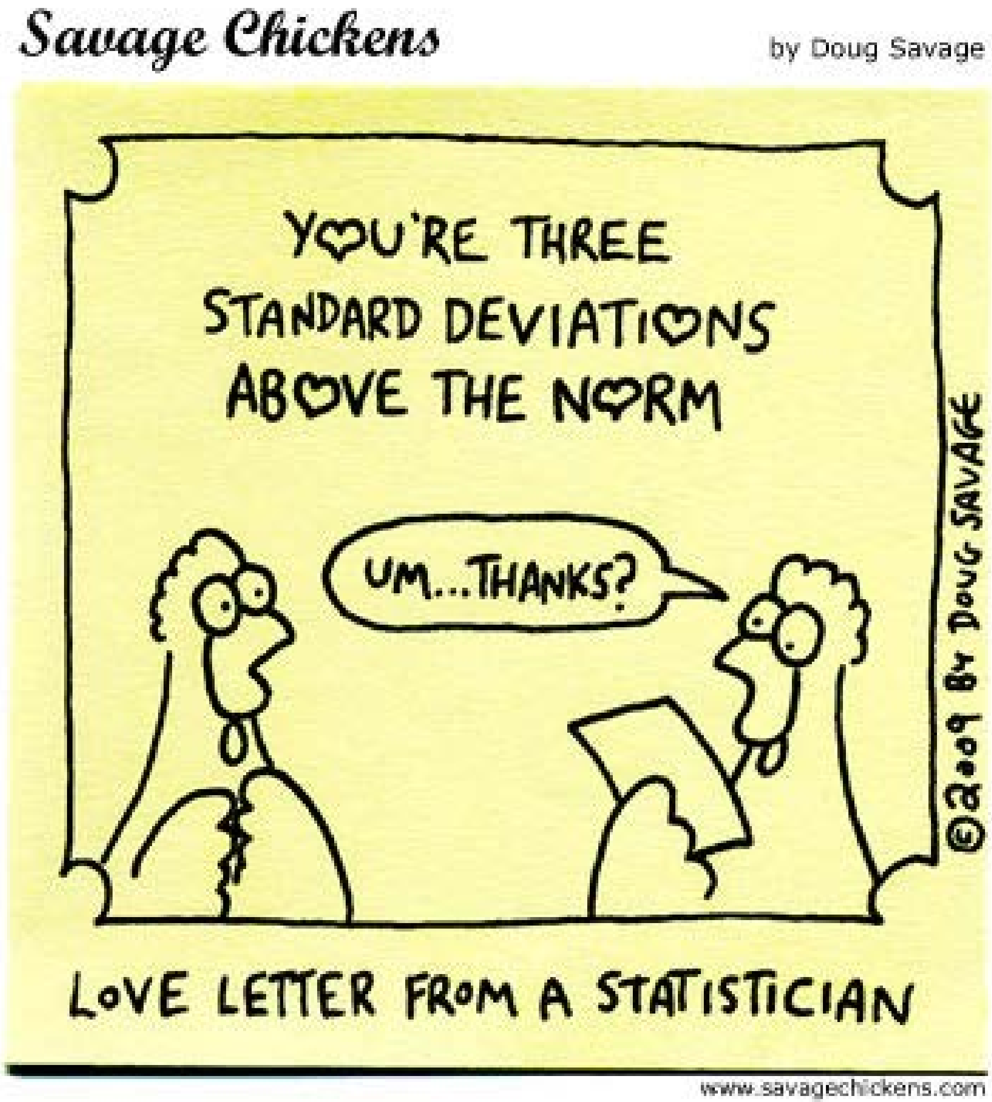
Expresses the relative probabilities of different values taken by a random variable



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

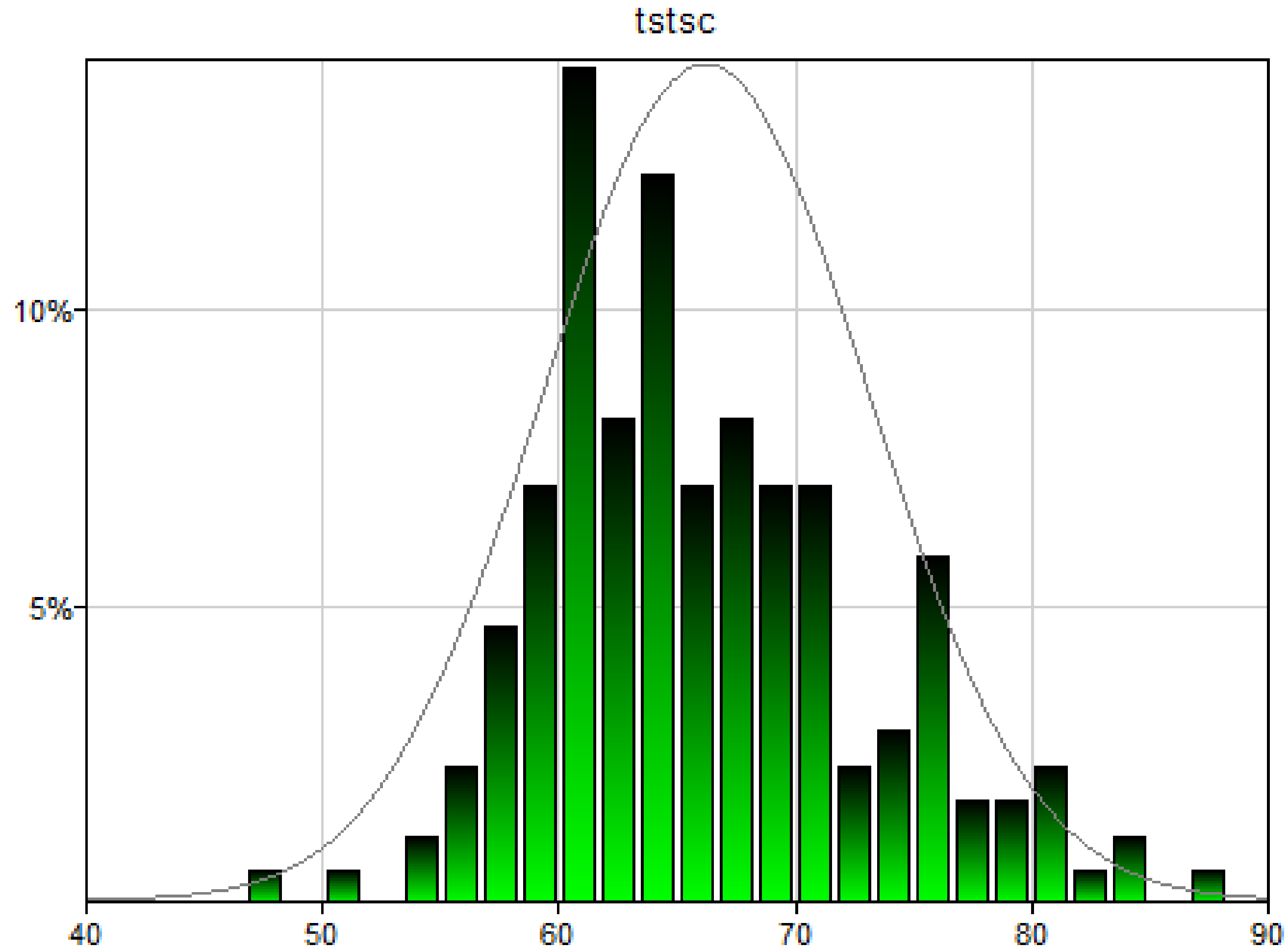
e.g., grade distribution in a university course

Standard deviation

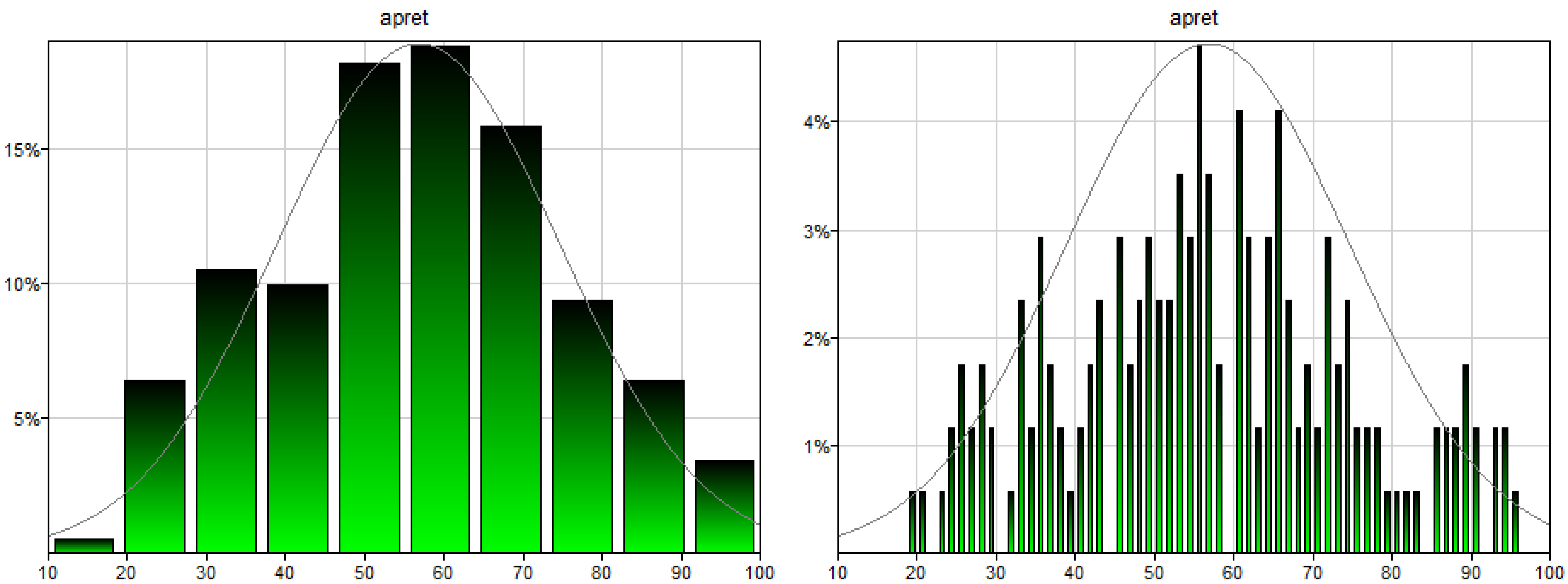


Histograms

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

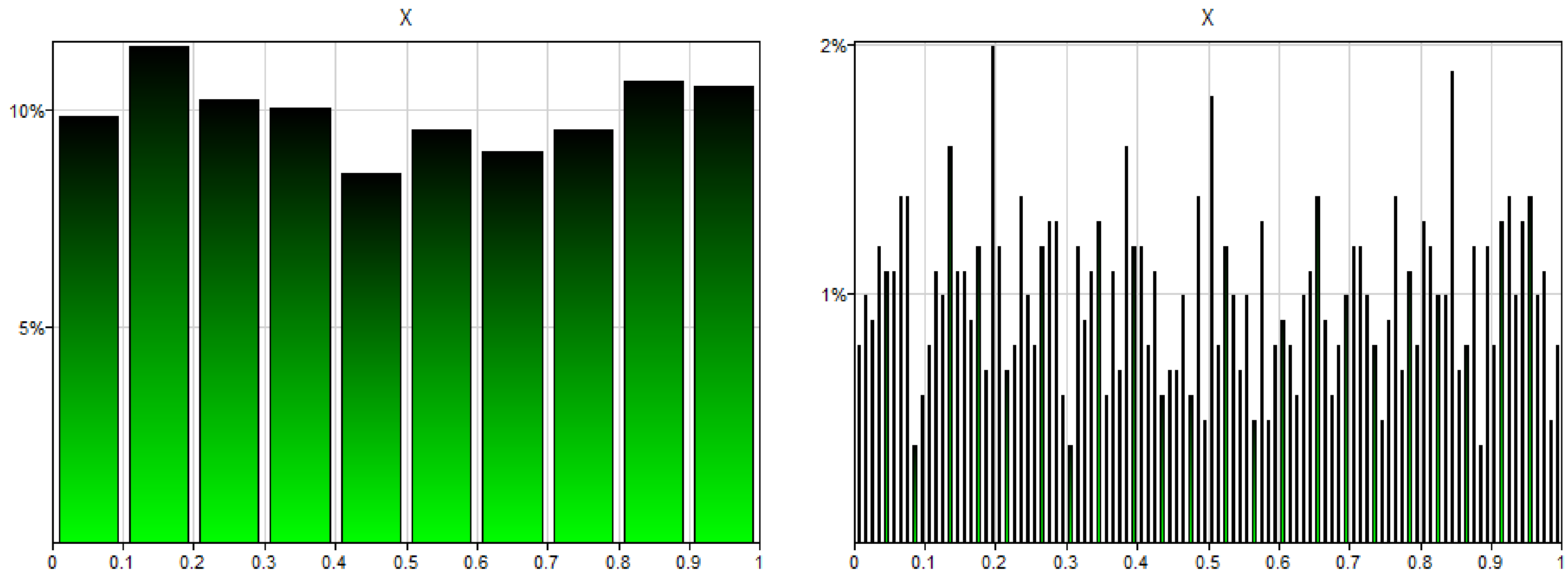


Histograms



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

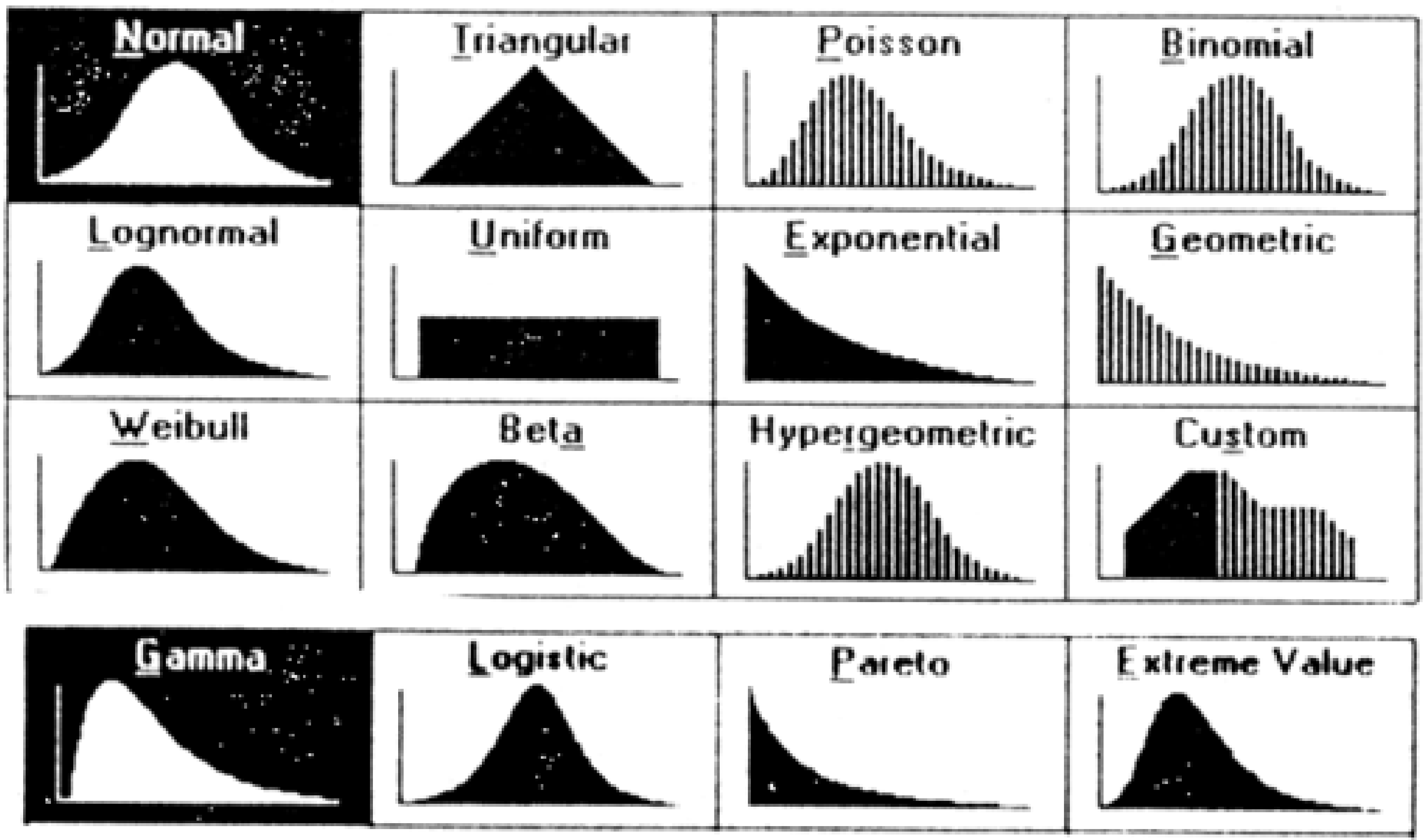


The effect of bin size is not that strong in case of some distributions (here: uniform distribution).

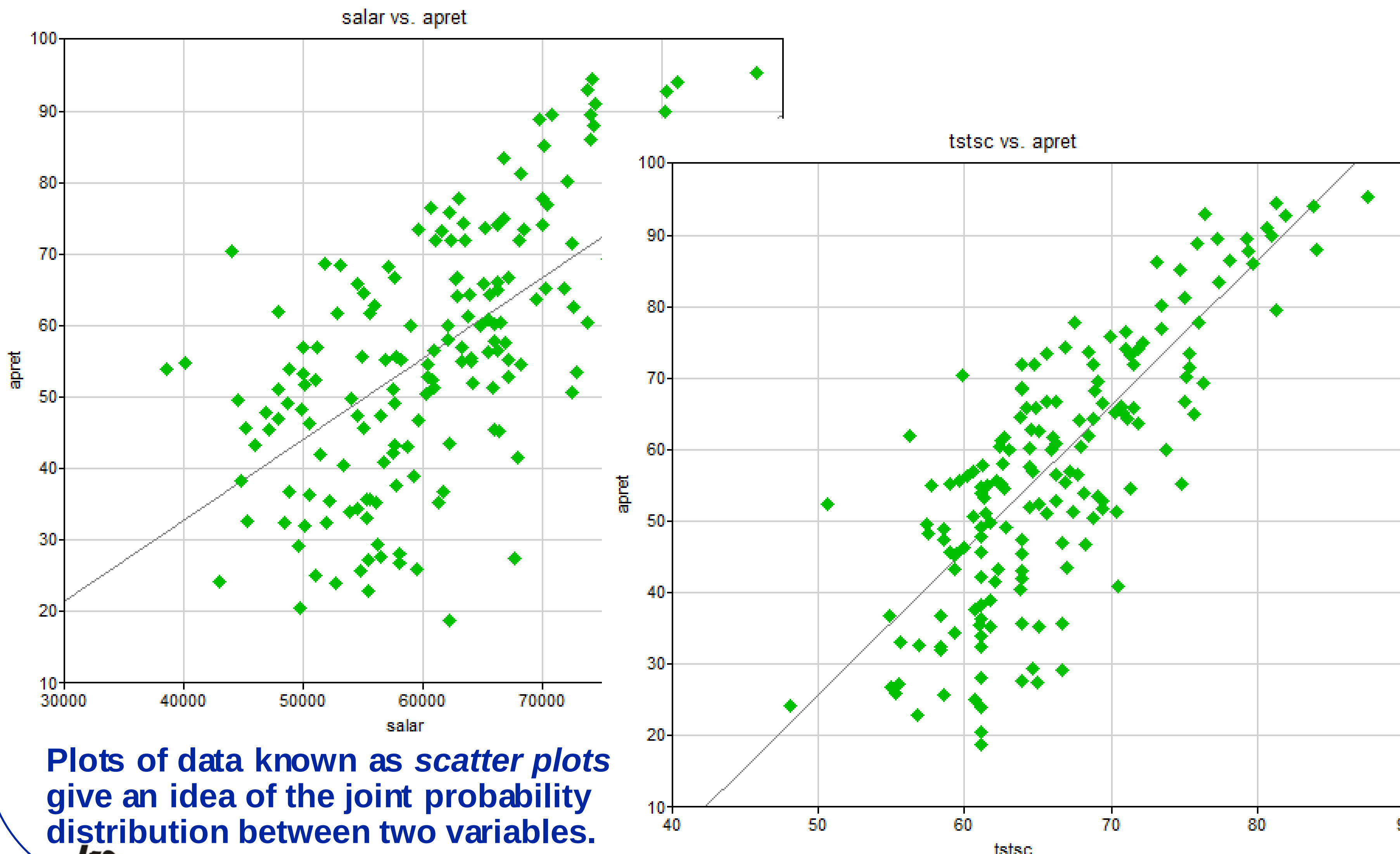
Probability distributions

- 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.

Common parametric probability distributions



Joint distribution functions

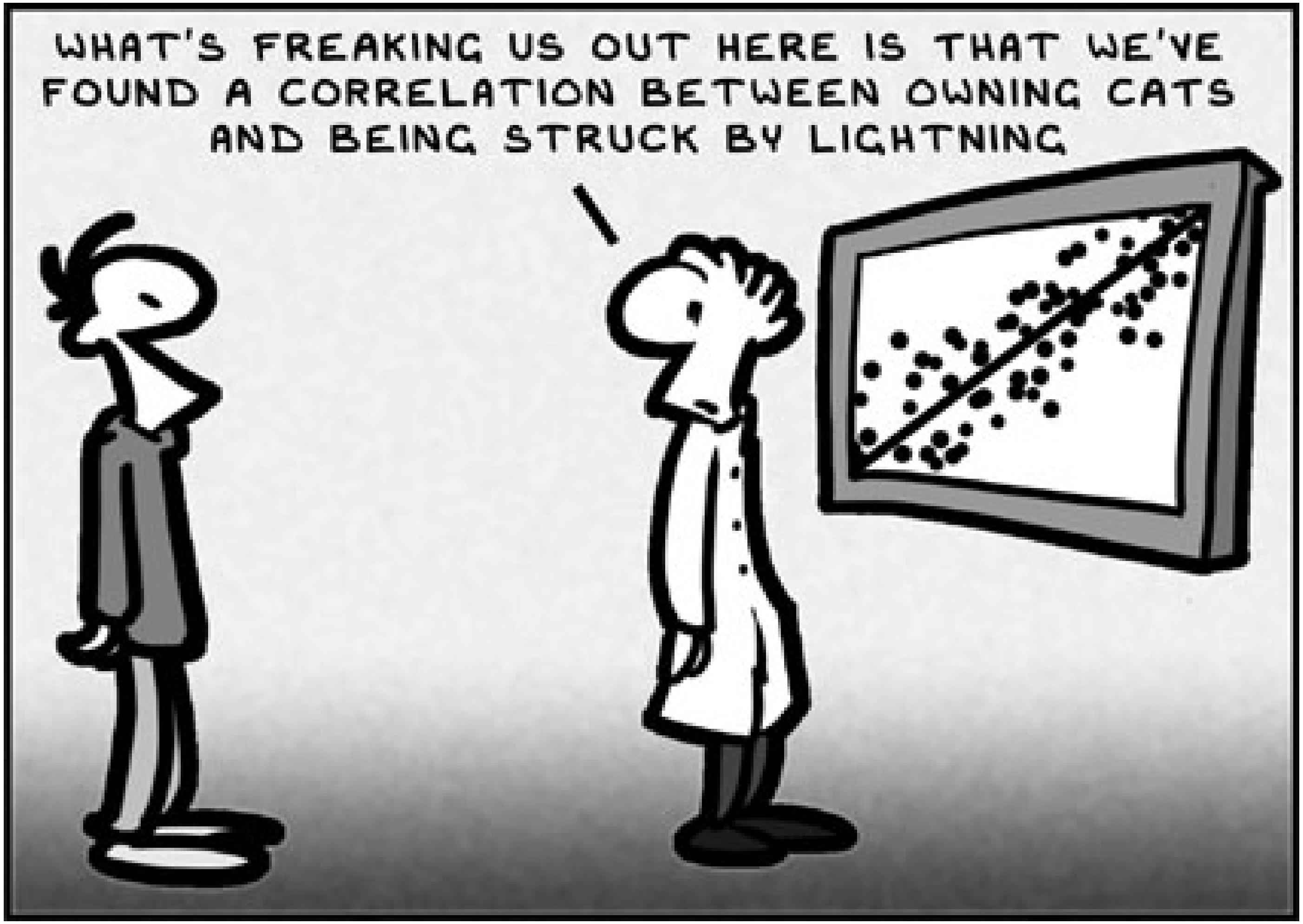


Plots of data known as *scatter plots* give an idea of the joint probability distribution between two variables.

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	-							
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”

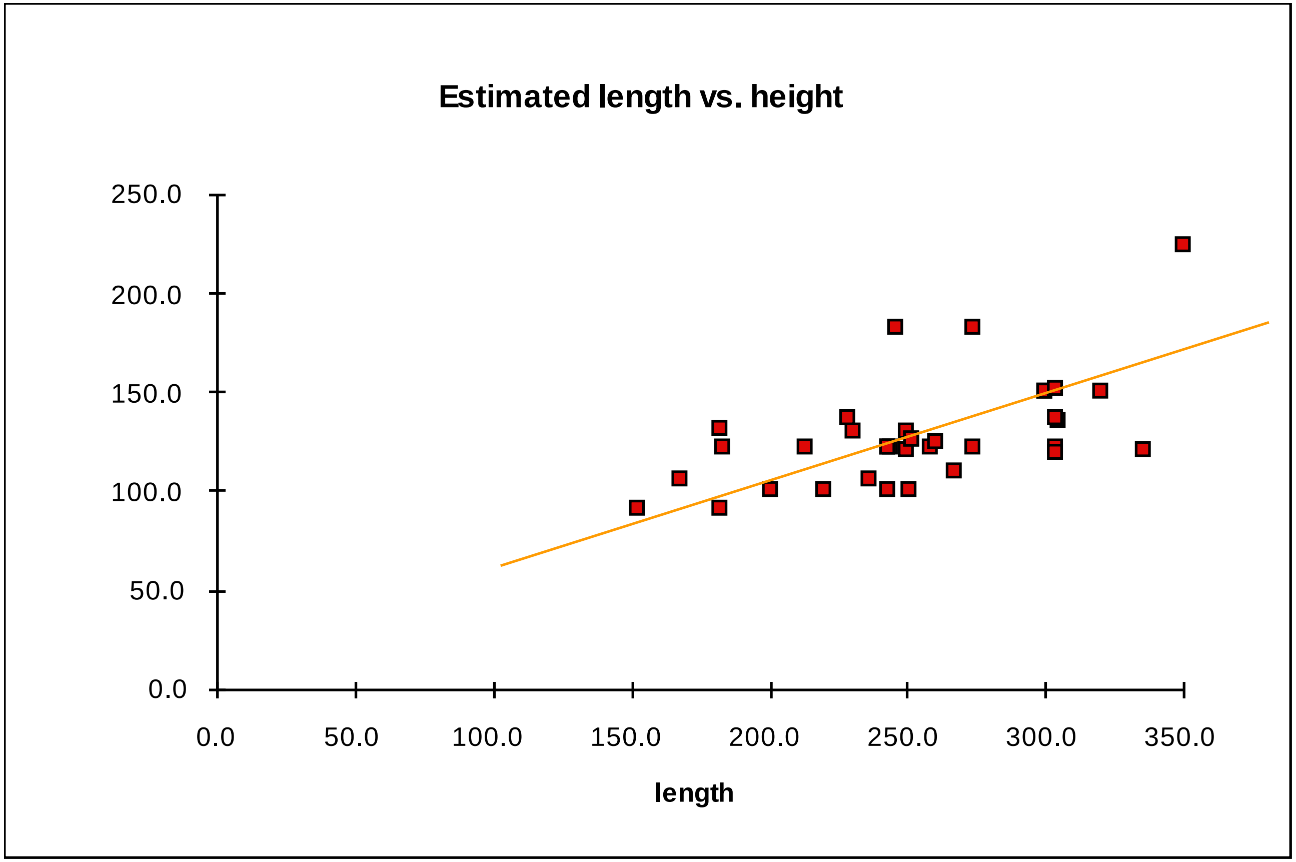
Cliché but indeed often true: Correlation between two variables by itself (i.e., in absence of other information) does not tell us much about the causal structure

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.

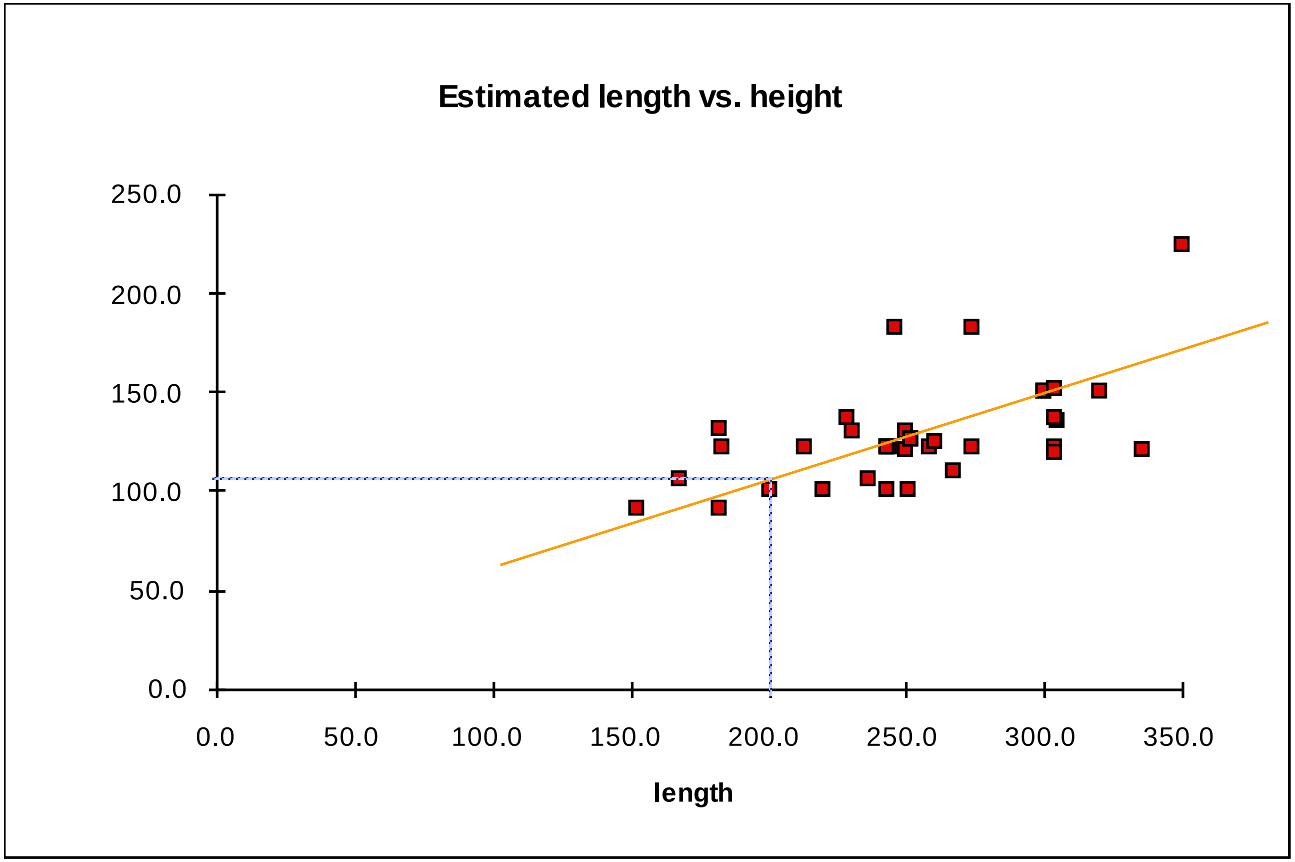
Linear regression

We fit a line to the data, the line equation is *y=a+bx*



Linear regression: Prediction

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

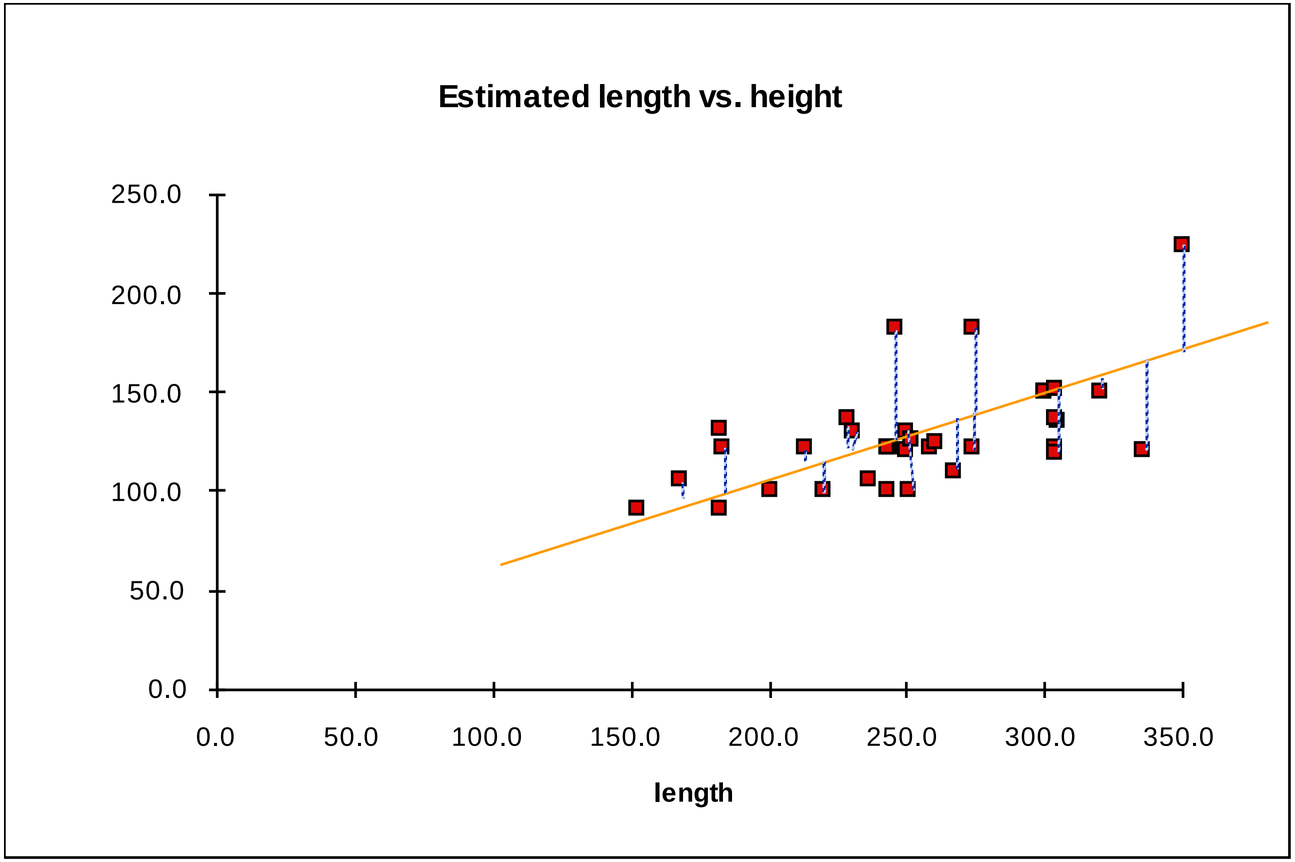


Least-squares regression

- 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

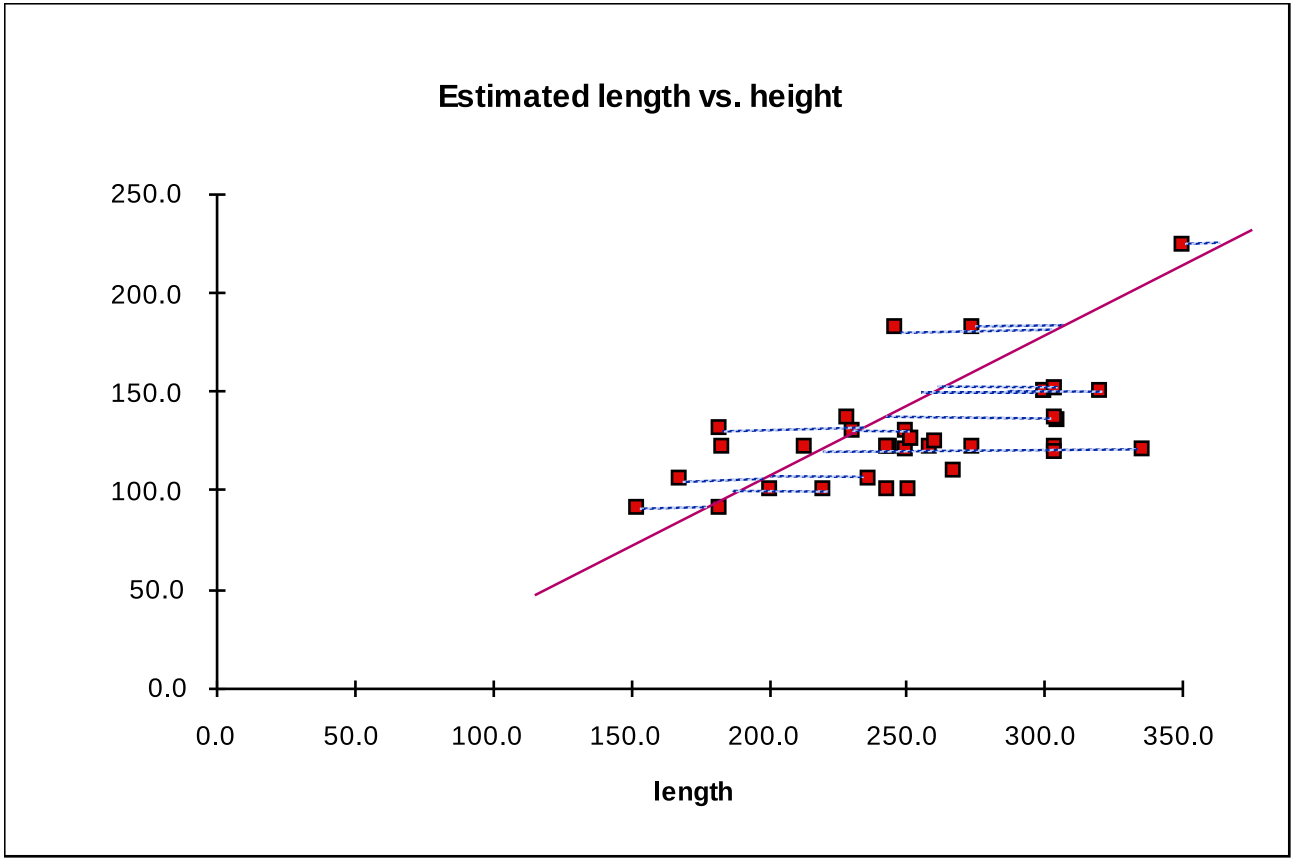
The idea: minimize the sum of squares of the deviations of the data points from the line in the vertical direction.



Most statistical packages implement least-squares regression.

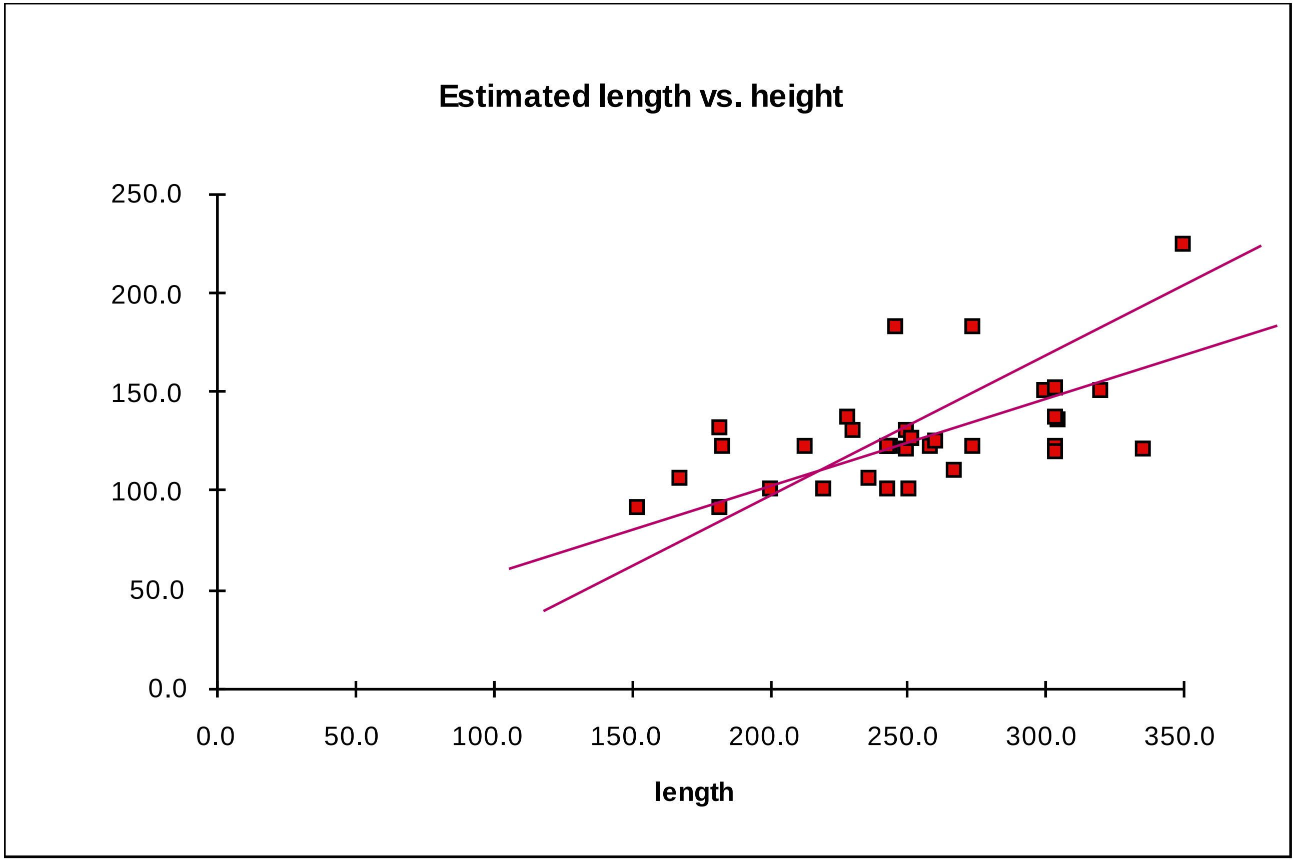
Asymmetry of regression

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%
```

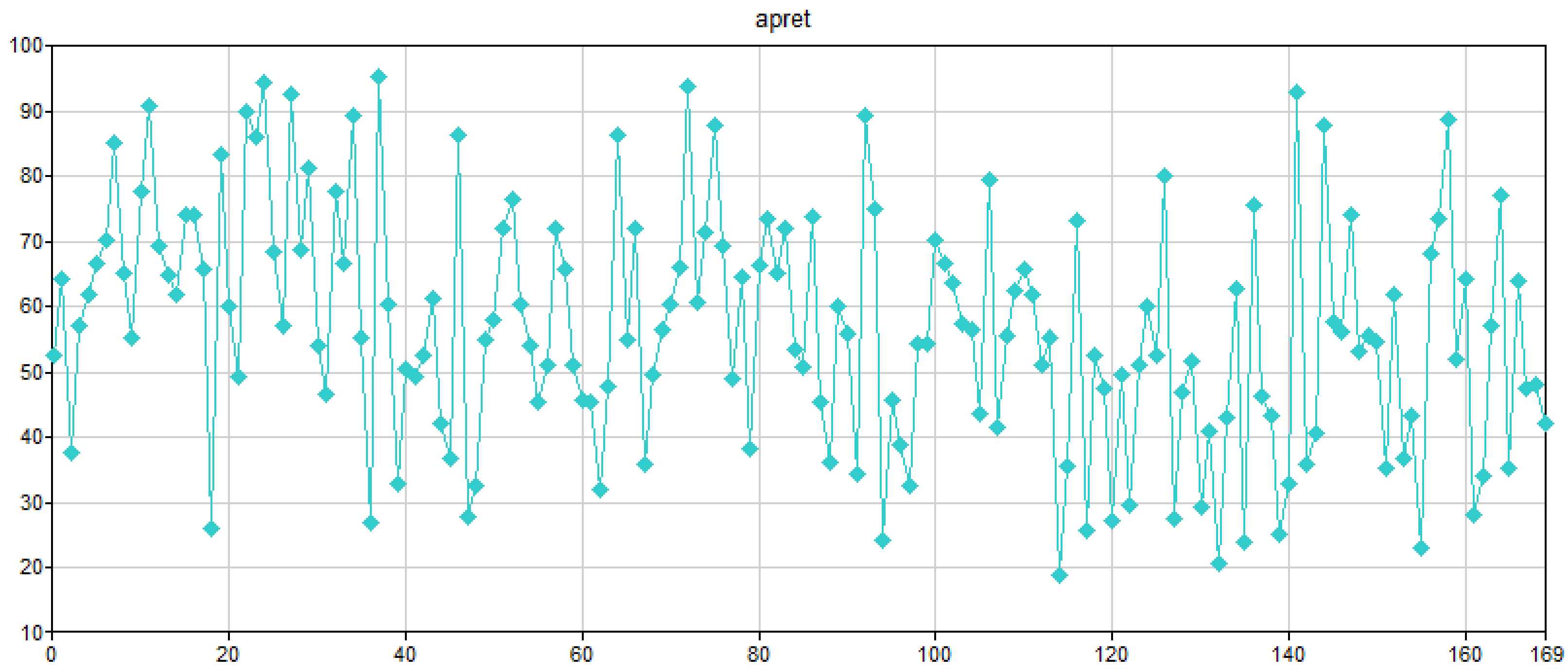
Can be also in multidimensional space.

***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

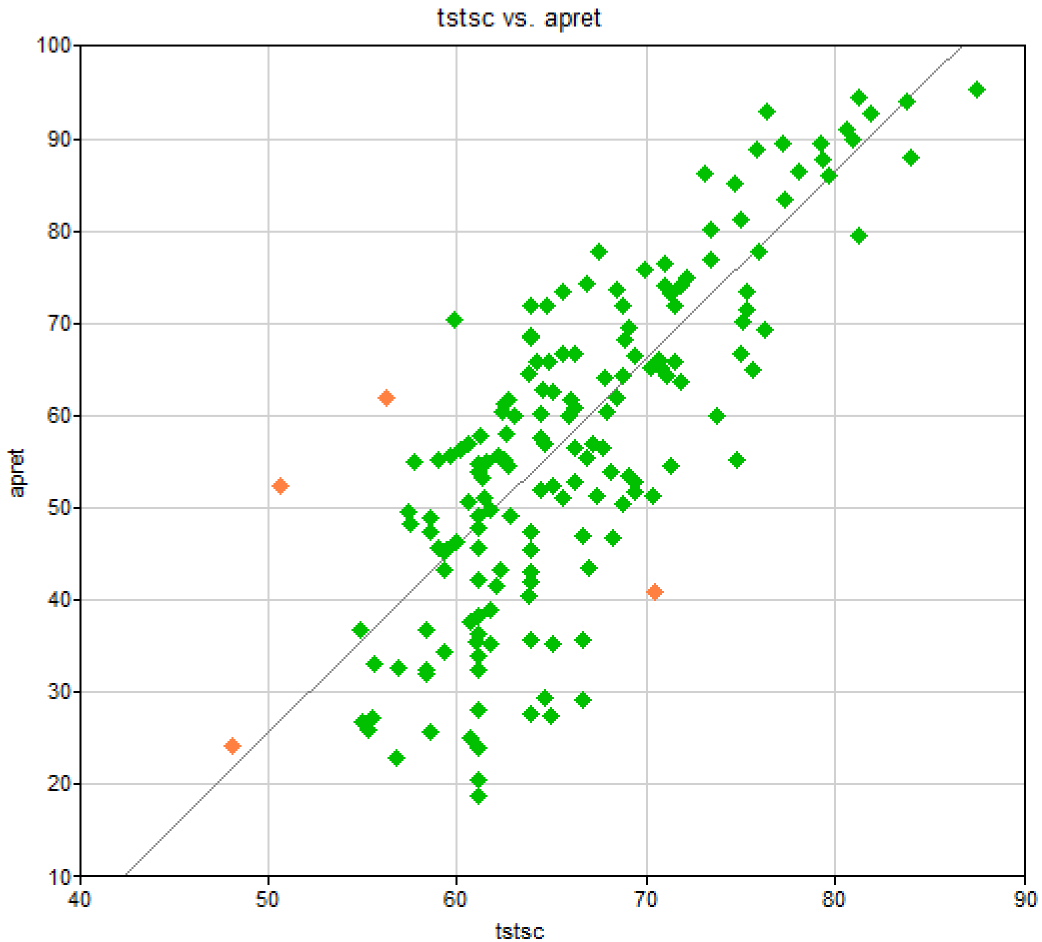
- **Measurements of variables that vary over time.**
- **This is often a matter of assumption: regular, static data also vary over time but we assume that they do not.**



We won't talk much about it in this class: There is only so much you can cover.

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.



Statistical Significance Testing

Scientific inference:
Classical hypothesis testing

How is this decision usually made?

- **The classical statistics: Significance testing.**
- **Bayesian approach.**

Classical significance testing:

- **There is no magic associated with classical hypothesis testing, it is just a tool for decision making under uncertainty, nothing more.**
- **Why do we do it this and no other way? Historical reasons.**

- **You cannot ever be sure about truth or falsity of a hypothesis (it is the classical philosophical problem of induction: how do we know after seeing 9,999 black ravens that the 10,000*th* one will be also black? How can we be sure that the sun will rise tomorrow?).**
- **You can get to the truth only with some probability.**

Elements of classical hypothesis testing

Elements of classical significance testing:

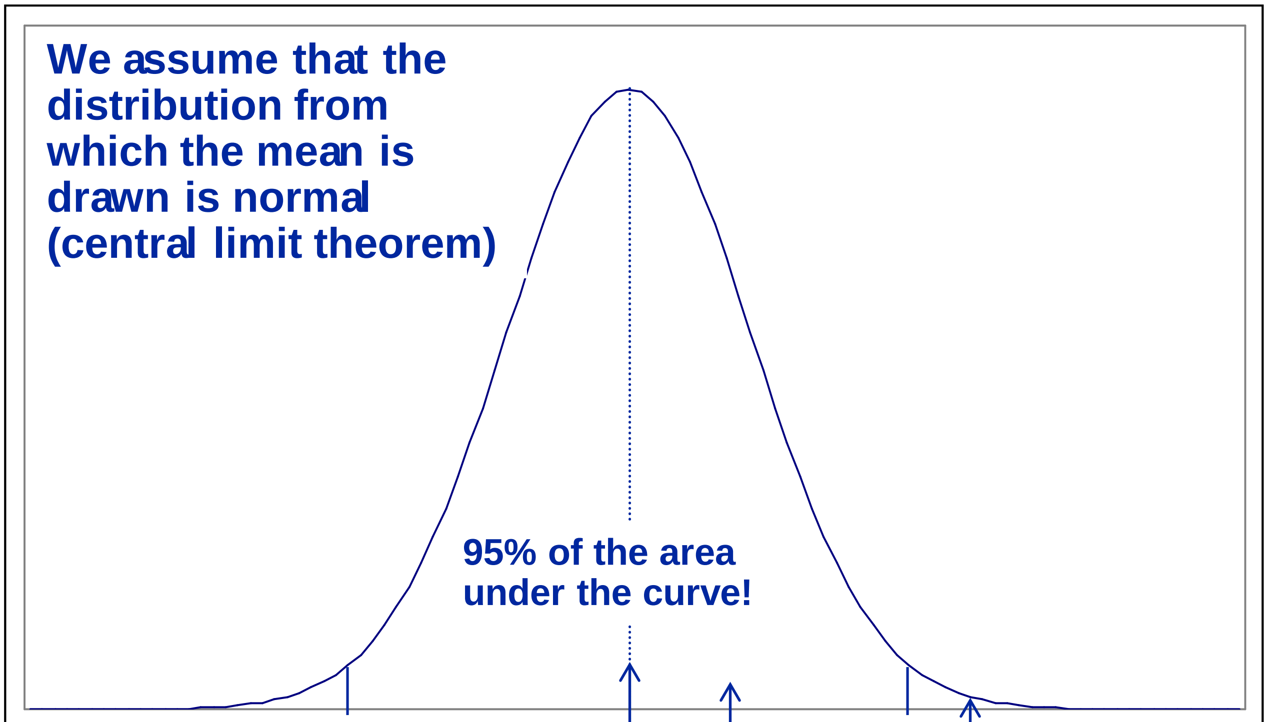
- Null hypothesis (H_0) and its complement (H_1).
- Significance level (α, *p* value).
- Statistical power (1- β), probability of rejecting H_0 given that it should be rejected.
- Sample size *n*.
- Effect size.

H_0 usually says something like "no effect" and is the more conservative one.

There is a possible confusion in terms: Effect size may be large, even if statistically not significant, may be also very small even if significant statistically.

Even if small, may be of considerable practical importance!

Comparing the means



The probability of observing a given value of the mean is proportional to the area under the curve.

μ_0

True mean

μ_1

If the observed mean falls inside the 95% range, we keep H_0

μ_1

If the observed mean falls outside the 95% range, we reject H_0

Risks of classical hypothesis testing

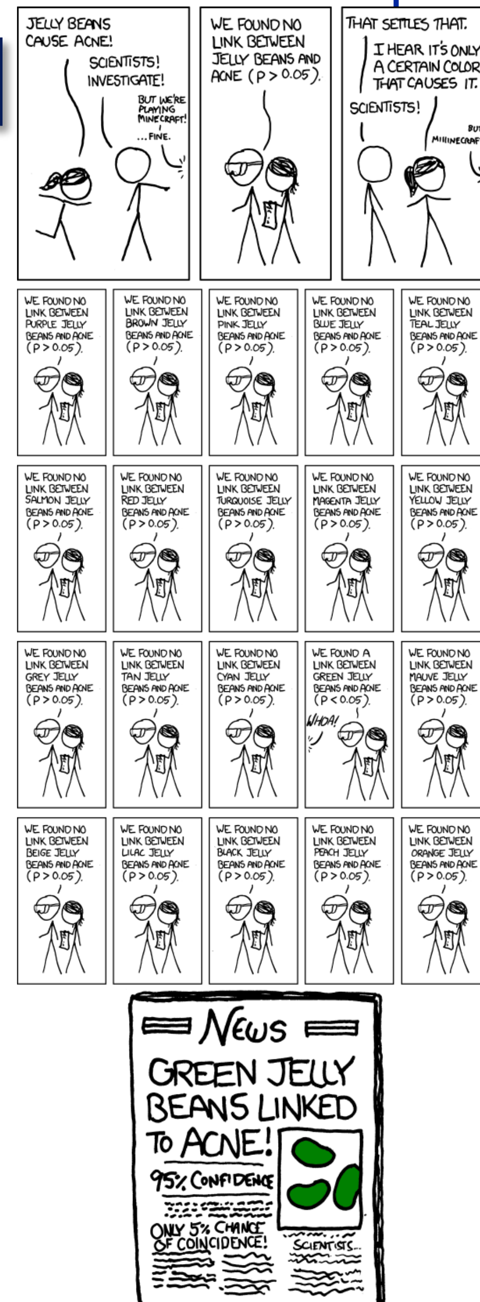
Risks related to classical significance testing (and to any decision making under uncertainty):

- Type I errors (reject H_0 when true).
- Type II errors (accept H_0 when false).

How do we deal with this risk?

- Need to consider consequences of these errors.
- What is the “correct” significance level? $\alpha = 0.05$ is a customary value, said to be proposed by Ronald Fisher at a party (let's hope it was a tea party ☺).
- Traditionally scientists believe that it is worse to risk being gullible than it is to be blind to a relationship – philosophers of science characterize it as “healthy skepticism” of the scientific outlook.
- What about cure for AIDS? Is this conservatism still OK?
- Statistical power $(1 - \beta)$, probability of rejecting H_0 when we should, i.e., $1 - \beta$.
- Also, note that we do not usually specify β!
- Power curve is the plot of the power of a test as we vary one of its parameters (α, β, variance, sample size, effect size).

Testing multiple hypotheses



Caution!

If you test many hypothesis using the classical significance testing, you run an increasing risk of accidental errors.

Elements of Decision Theory

Elements of decision theory

The theoretically sound way of making decisions under uncertainty

- Decision making: we need to consider uncertainty and preferences. These are measured in terms of probability and utility respectively.
- Some special cases are easy:
  - it is better to be rich and healthy than poor and sick;
  - it's better to start a project that has a high chance of succeeding and a high payoff than to start a project that has a low chance of succeeding and a low payoff.

  We can reason qualitatively but it is possible to do it within this framework.
- Probability is a measure of uncertainty.
- Utility is a measure of preference that combines with probability as mathematical expectation.

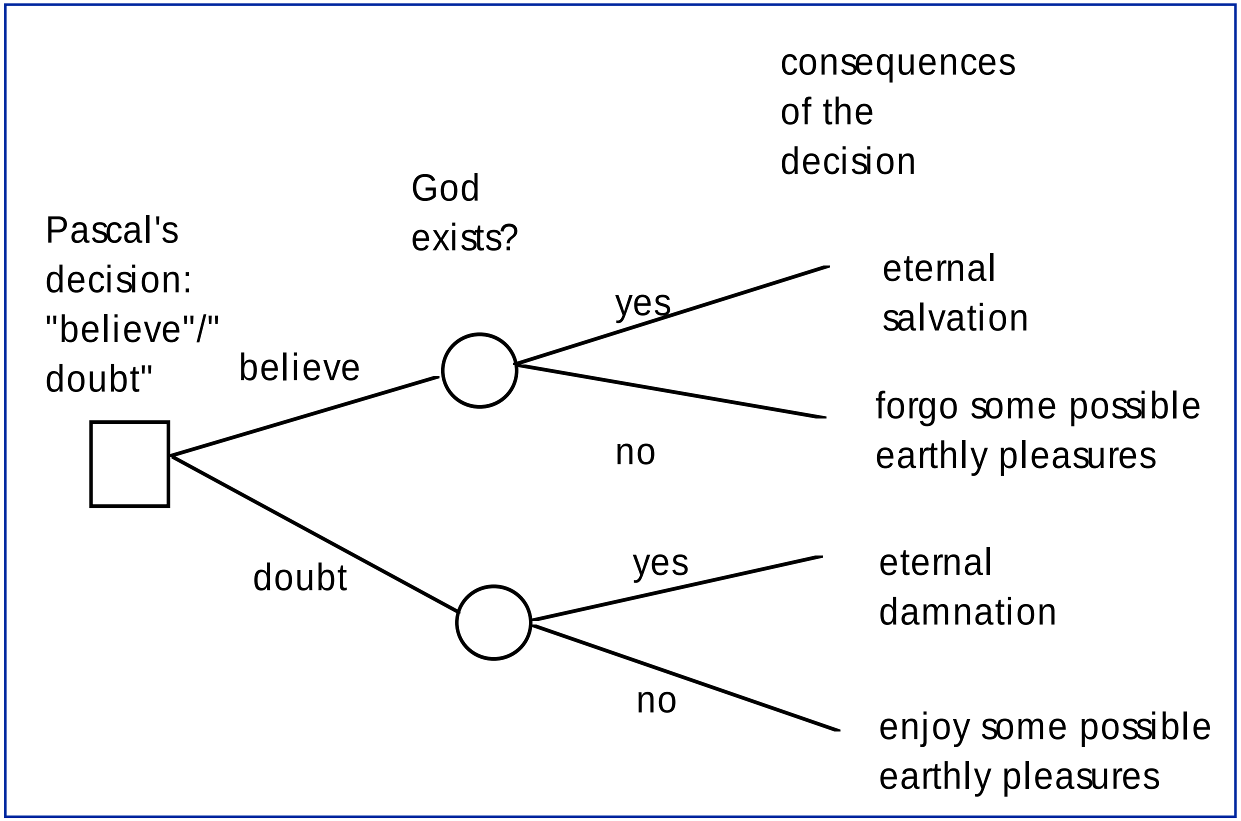
Pascal’s wager

Pascal’s wager: Should we believe in God or not?

	God exists	God does not exist
believe	eternal salvation	forgo some earthly pleasures in your life
doubt	eternal damnation	enjoy some earthly pleasures in your life

Pascal’s wager

Pascal’s wager: Decision tree



EU(believe) = p ∞ + (1-p)(-ε) = ∞
EU(doubt) = p (- ∞) + (1-p) ε = - ∞

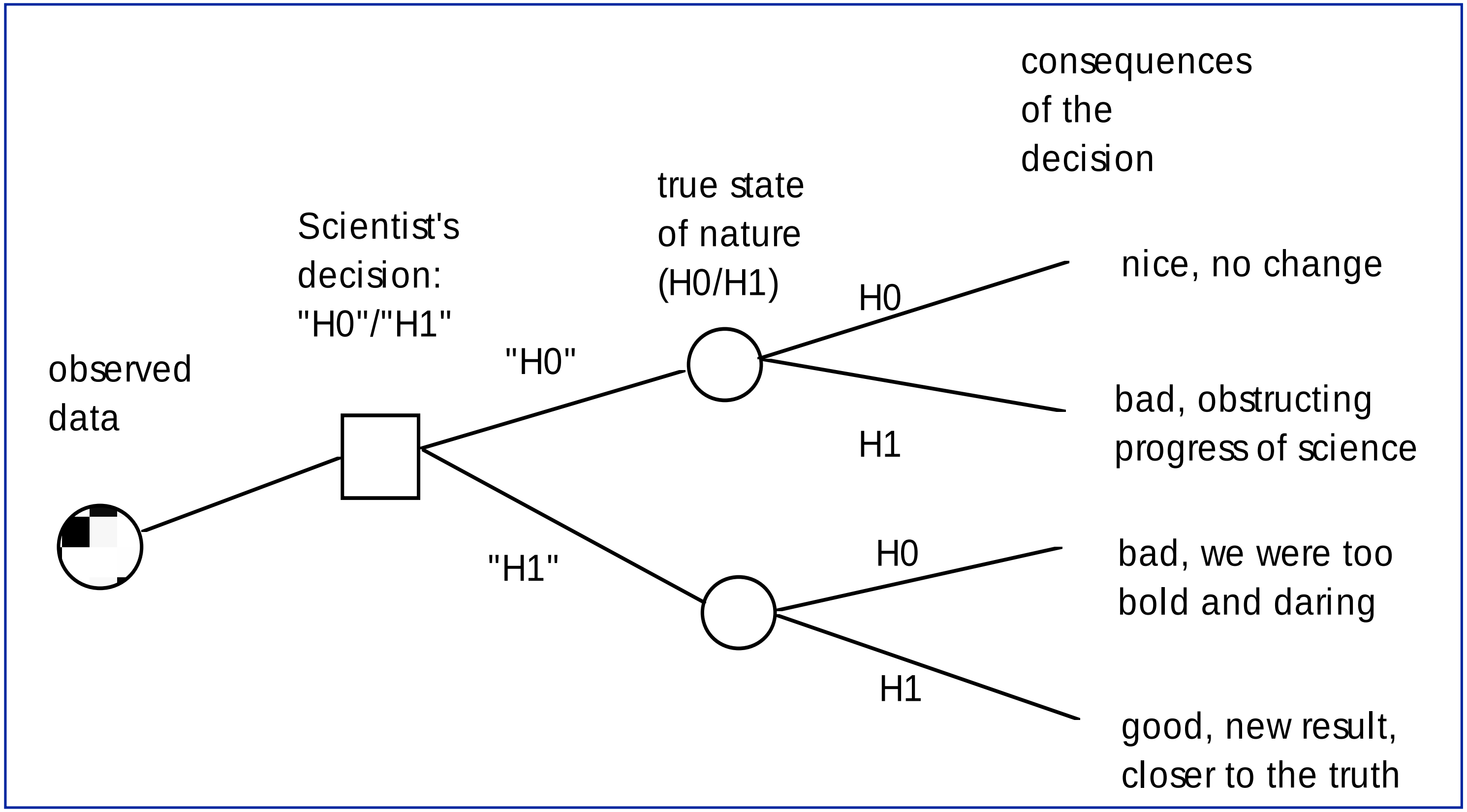
The only rational thing is to believe ☺!

Scientific decision making

Is there any difference between this and scientific decision making?

- **Research may involve issues as difficult and important as smoking, cholesterol, etc.**
- **Then the congressmen or senators will pick up our results, stand up in the congress and enact laws that will either save lives or make our lives unnecessarily uncomfortable.**
- **Another example of decision making under uncertainty, quite close to hypothesis testing: Dilemma of a referee. Should I recommend this paper for publication or not? Advancement of science *vs.* the risk of going into a wrong path.**
- **Statistical inference in science is a decision process and most books will make you aware of the importance to consider consequences.**
- **The elements of decision theory that you get here will help you in understanding what this is about.**

Classical hypothesis testing: A decision-theoretic view



Classical hypothesis testing: A decision-theoretic view

- **The main problem (of course, after determining what the value of the outcomes are) is to determine the prior probability of the hypothesis. To see that, start with $Pr(H_0|D)$ and the derive everything using Bayes theorem in terms of $Pr(H_0)$, $Pr(D|H_0)$, and $Pr(D|H_1)$.**
- **Recall the possible errors are: (type I and type II) - α and β are probabilities of these errors.**
- **Decision-theoretic (Bayesian) view allows to explore the exact relation between the significance level and the decision.**

$$Pr(H_0|D)=Pr(D|H_0)Pr(H_0) / (Pr(D|H_0)Pr(H_0)+Pr(D|H_1)Pr(H_1))$$
$$= \alpha P(H_0)/(\alpha Pr(H_0) + (1-\beta)(1-Pr(H_0)))$$
$$Pr(H_1|D)=Pr(D|H_1)Pr(H_1) / (Pr(D|H_0)Pr(H_0)+Pr(D|H_1)Pr(H_1))$$
$$= (1-\beta)(1-P(H_0))/(\alpha Pr(H_0) + (1-\beta)(1-Pr(H_0)))$$

But we don't have $p(H_0)$!

This is the reason why classical statistics rejects this approach.

Concluding remarks

- Which test/statistical decision making procedure to use? Find one that is the strongest given your conditions and assumptions (*e.g.*, use T-test or Z-test instead of ANOVA for comparison of the means).
- Scientific method is more of a philosophical outlook than a single fixed procedure.
- Science relies mainly on the method of empirical inquiry as a means of opening theoretical work to scrutiny.
- Discipline-dependent language (definitions, quantitative analysis, hypotheses, theories). "Think Yiddish, write British."
- There are often hidden assumptions, not strictly logical, but intuitive presuppositions about the nature of reality (*e.g.*, strict determinism, indeterminism).
- Explanatory act, to be influential in science, must "make sense" (this is a historical observation).