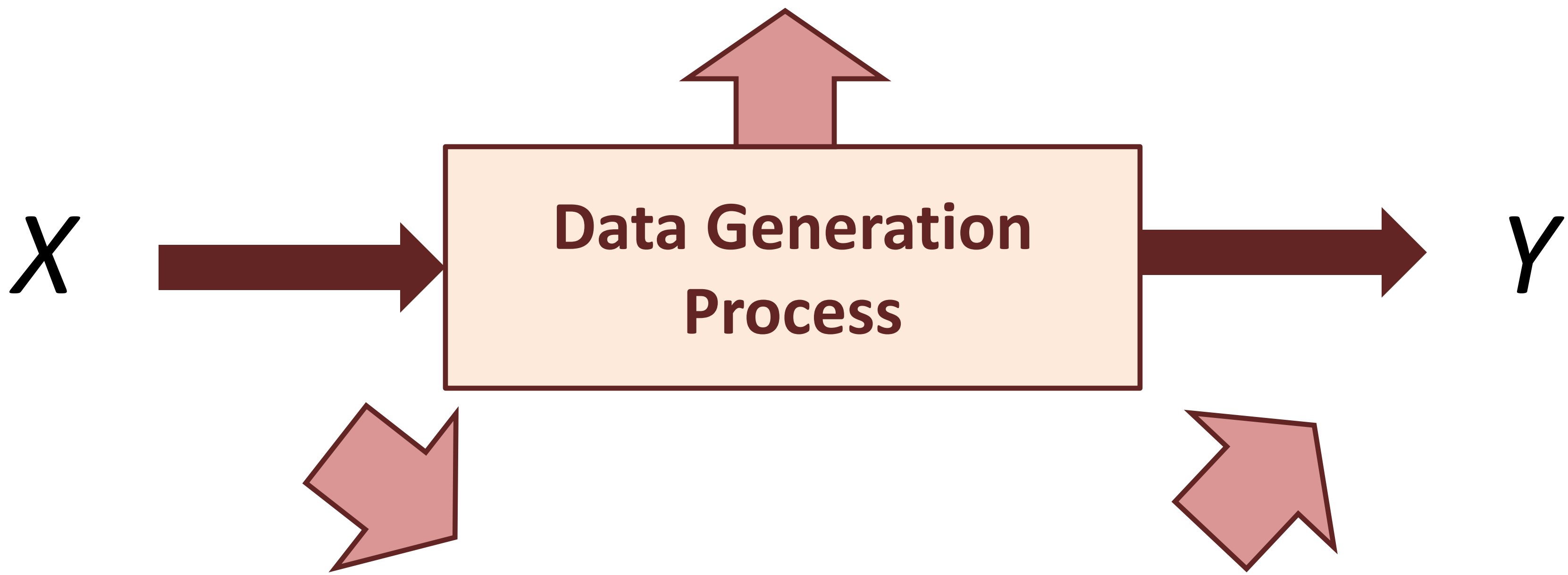
Hypothesis Testing and the boundaries between Statistics and Machine Learning

The Data Science and Decisions Lab, UCLA

April 2016

Statistical Inference vs Statistical Learning

Statistical inference tries to draw conclusions on the data generation model



Statistical learning just tries to predict Y

Statistics vs Machine Learning: Two Cultures

Leo Breiman,
“Statistical Modeling: The Two Cultures,”
Statistical Science, 2001

Descriptive vs. Inferential Statistics

- **Descriptive statistics: describing a data sample (sample size, demographics, mean and median tendencies, etc) without drawing conclusions on the population.**

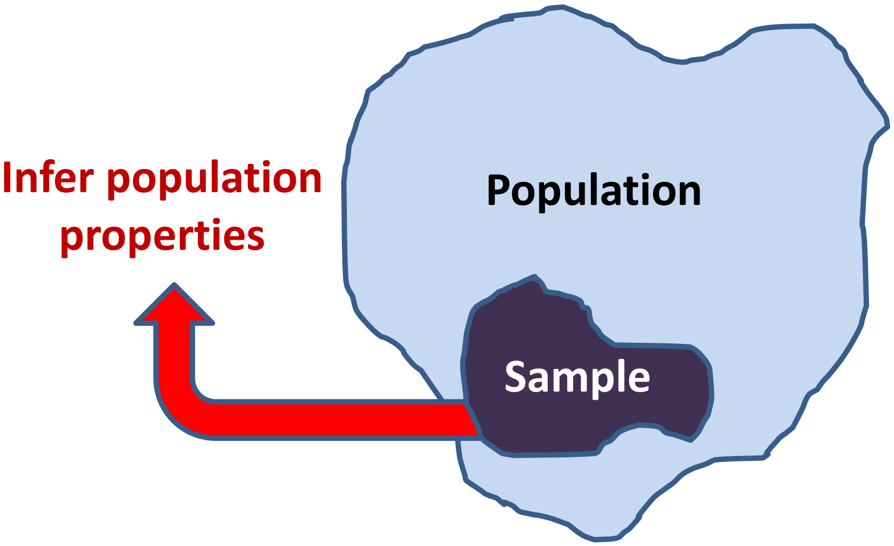
- **Inferential (inductive) statistics: using the data sample to draw conclusions about the population (conclusion still entail uncertainty = need measures of significance)**

- **Statistical Hypothesis testing is an inferential statistics approach but involves descriptive statistics as well!**

Statistical Inference problems

- **Inferential Statistics problems:**

  - **Point estimation**
  - **Interval estimation**
  - **Classification and clustering**
  - **Rejecting hypotheses**
  - **Selecting models**

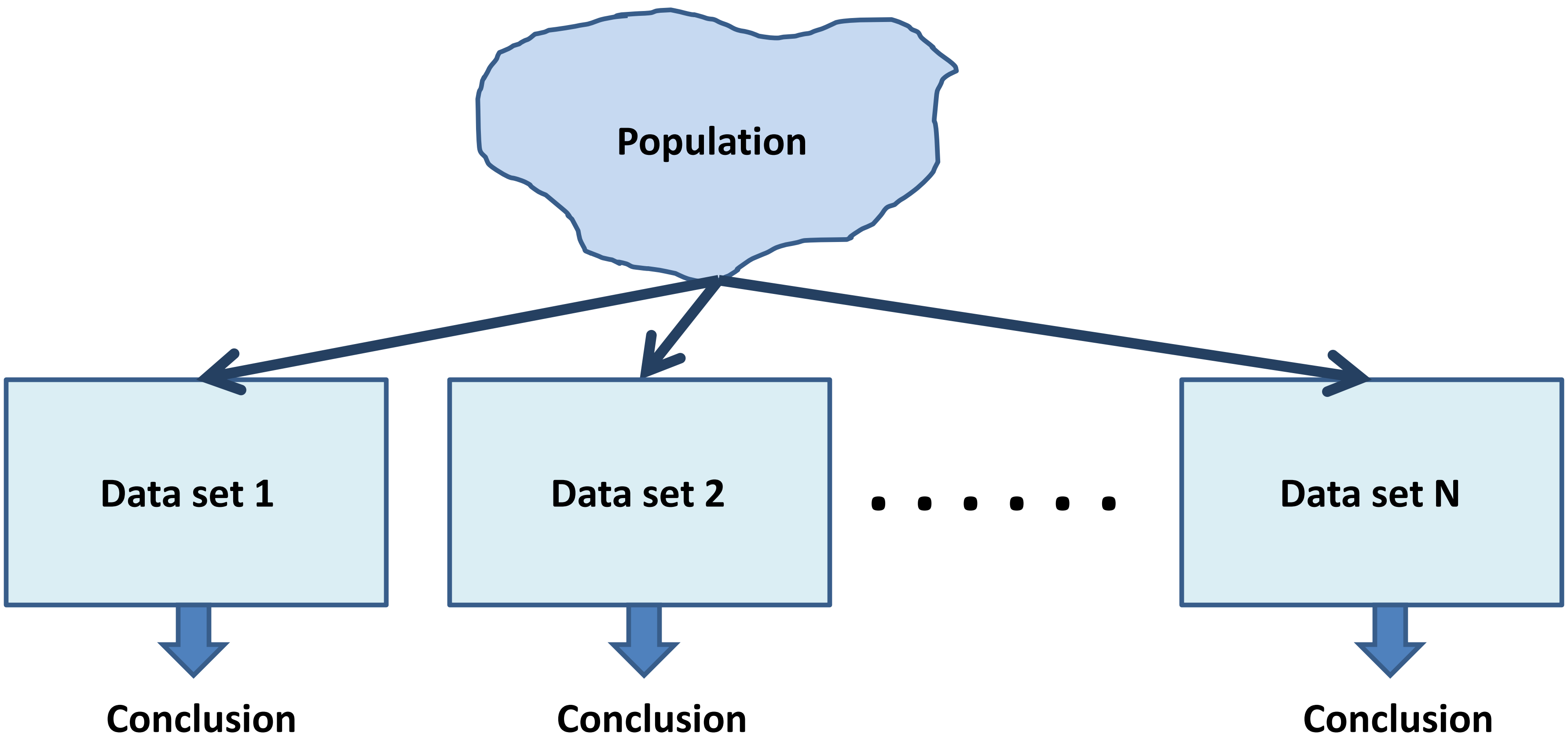


Frequentist vs. Bayesian Inference

- **Frequentist inference:**
- **Key idea:** ***Objective interpretation* of probability -** any given experiment can be considered as one of an infinite sequence of possible repetitions of the same experiment, each capable of producing statistically independent results.
- Require that the correct conclusion should be drawn with a given (high) probability among this set of experiments.

Frequentist vs. Bayesian Inference

- **Frequentist inference:**



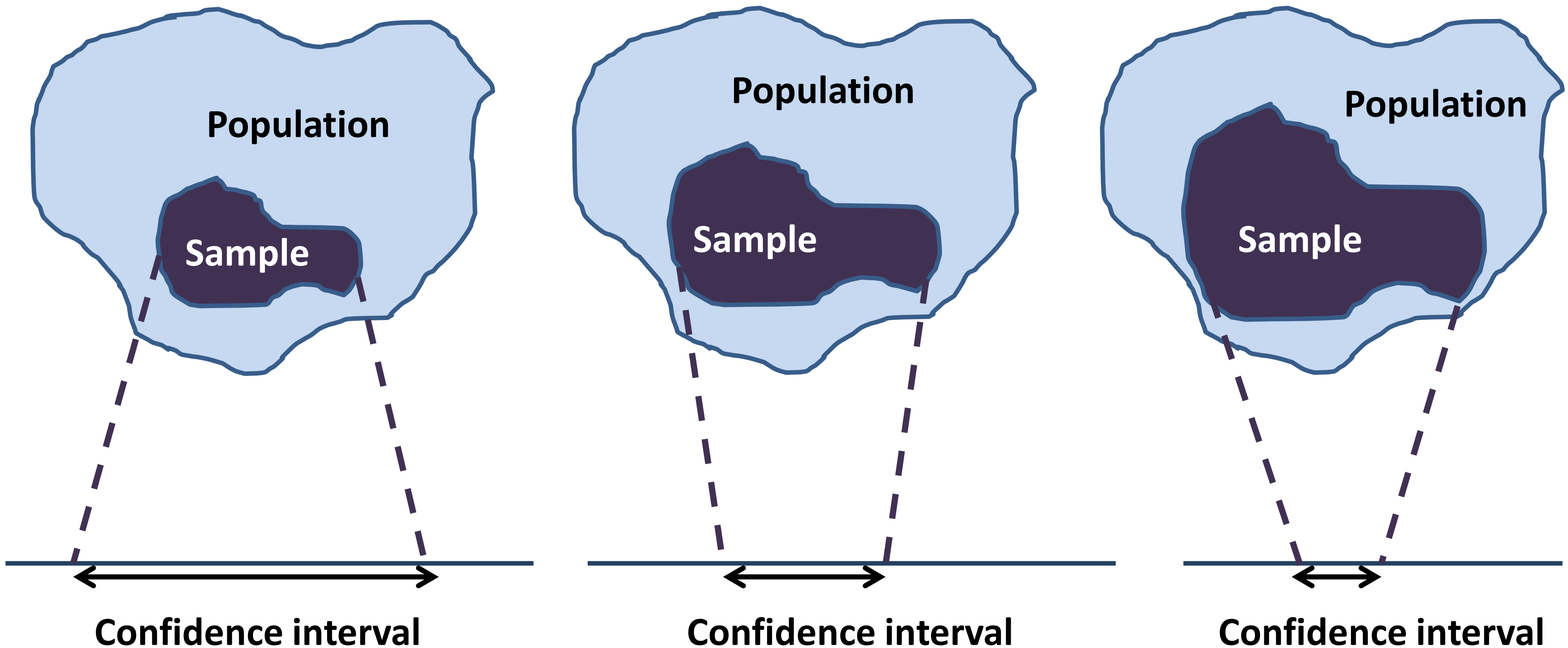
The same conclusion is reached with high probability by resampling the population and repeating the experiment.

Frequentist vs. Bayesian Inference

- **Frequentist inference:**
- **Measures of significance: p-values and confidence intervals**
- Frequentist methods are objective: you can do significance testing or confidence interval estimation without defining any explicit (subjective) utility function.
- Do not need to assume a prior distribution over model parameters, but assume they take fixed, unknown values.

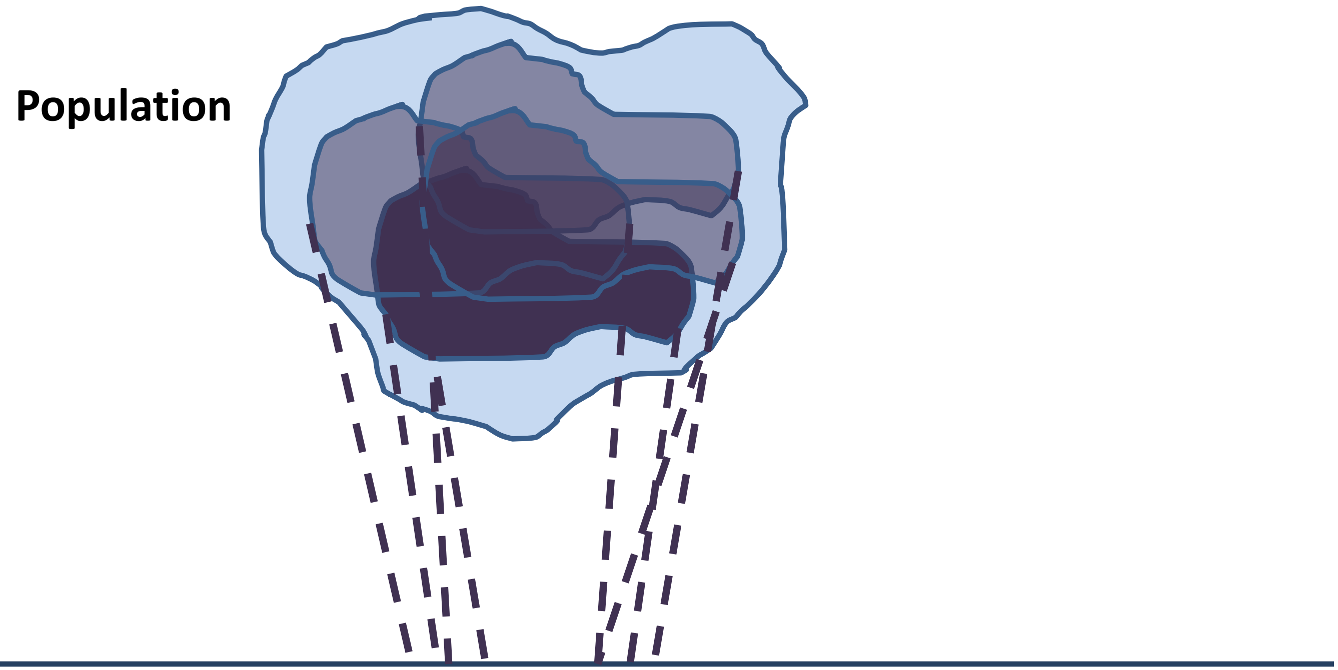
Frequentist vs. Bayesian Inference

- **Frequentist inference:**
- **Measures of significance:** p-values and confidence intervals



Frequentist vs. Bayesian Inference

- **Frequentist inference:**



Confidence intervals are random!

The fraction of such intervals that contain the true parameter = confidence level 1-δ

Frequentist vs. Bayesian Inference

- <u>**Bayesian inference**</u>**:**

- **Key idea:** ***Subjective interpretation* of probability –** statistical propositions that depend on a posterior belief that is formed having observed data samples. Subjective because it depends on prior beliefs (and utility functions).

- **Measures of significance: credible intervals and Bayes factors**.

Frequentist vs. Bayesian Inference

- **Bayesian inference:**

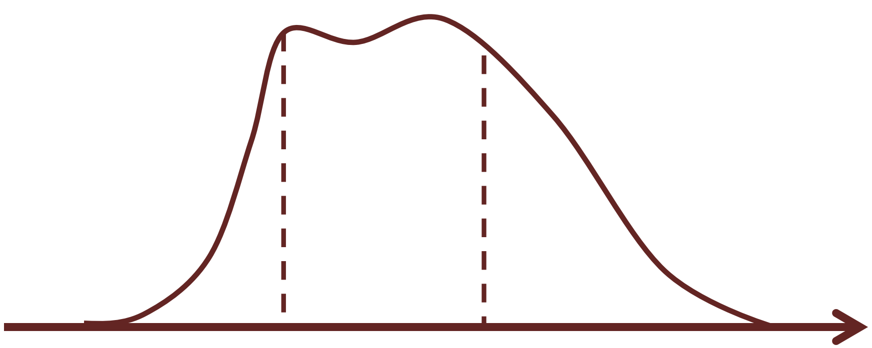
- **Credible intervals for parameter estimation:** an interval in the domain of a posterior probability distribution or predictive distribution used for interval estimation

- **Credible intervals vs. Confidence intervals:** Bayesian intervals treat their bounds as fixed and the estimated parameter as a random variable, whereas frequentist confidence intervals treat their bounds as random variables and the parameter as a fixed value.

Frequentist vs. Bayesian Inference

- **Bayesian inference: Estimation problems**
- **Credible intervals vs. Confidence intervals**

Posterior distribution (belief)



95% credible interval

$$P(\theta_1 < \theta < \theta_2 \mid x) = 95\%$$

Random 95% confidence interval

$$P\left(\hat{\theta}_1(x) < \theta < \hat{\theta}_2(x)\right) = 95\%$$

$$P\left(\hat{\theta}_1(x) < \theta < \hat{\theta}_2(x) \mid x\right) = 95\%$$

Since parameter is not random, this probability is either 0 or 1

Frequentist vs. Bayesian Inference

- **Bayesian inference: Comparison problems**

- **Bayes factors vs. p-values**

- **Bayesian factors are natural alternative to classical hypothesis testing that measure the strength of evidence through "risk rations"**

$$K = \frac{P(X|H_o)}{P(X|H_1)} = \frac{\int P(\theta_o|H_o)P(X|\theta_o, H_o)d\theta_o}{\int P(\theta_1|H_1)P(X|\theta_1, H_1)d\theta_1}$$

- **Guidelines on the value of K: K<1 negative, K>100 decisive, etc.**

Statistical Hypothesis Testing: Problems

- **Many statistical inference problems involve hypothesis testing:**
- **Examples:**

- **Which model best fits the data?**
- **Is treatment X more effective for males than females?**
- **Is smoking a risk factor for coronary heart diseases?**
- **Is the chance of a certain intervention being successful depends on a specific feature of the patient?**
- **Does this subpopulation of patients belong to the same category?**

- **Usually a Yes-No question. Inference = answer this question from a data sample. <u>Understanding the data independent of any specific ML algorithm</u>**

Statistical Hypothesis Testing: the setting

- **Usually we want to test:**

1) Whether two samples can be considered to be from the same population.
2) Whether one sample has systematically larger values than another.
3) Whether samples can be considered to be correlated.

Significance of conclusions: predict the likelihood of an event associated with a given statement (i.e. the hypothesis) occurring by chance, given the observed data and available information.

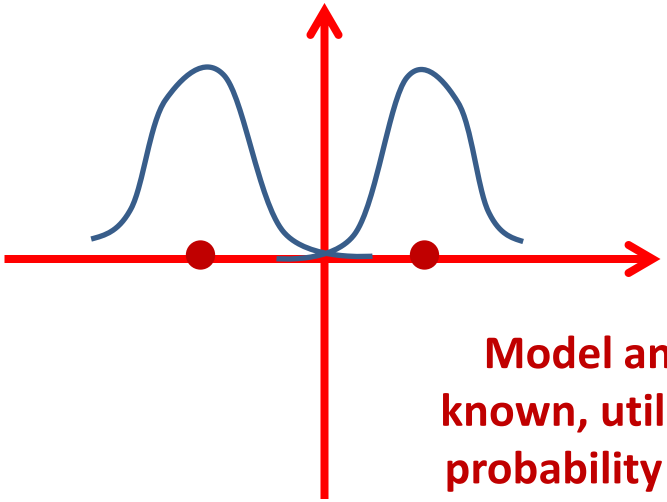
Testing is usually objective: frequentist significance measures!

Statistical Hypothesis Testing: the setting

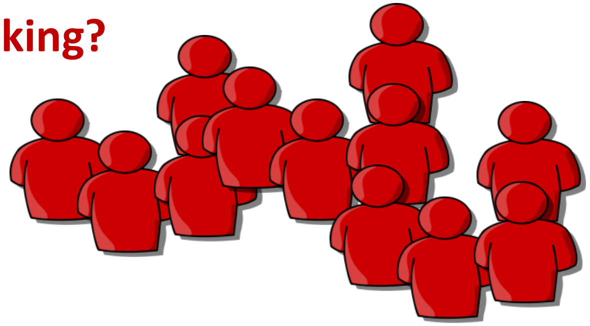
Testing is usually objective: frequentist significance measures!

- **Complex phenomena (no solid model), inference not necessary associated with specific utility (need objective conclusions)**

- **e.g. signal detection vs. medical study**



Effect of smoking?



Model and priors known, utility is error probability (Bayesian risk) = Bayesian framework

Complex population, priors may not be easy to construct, no utility = frequentist framework

Statistical Hypothesis Testing: Main steps

Identify the null and alternative hypotheses

Model and statistical assumptions

Select an appropriate test and relevant **test statistic**

Find the distribution of the statistic

Find a threshold on the static given the significance level α

Reject the null hypothesis if statistic is less than threshold

Compute the p-value

Reject the null hypothesis if p-value is less than the significance level α

Parametric and Nonparametric Hypothesis Testing

Parametric Tests	Nonparametric Tests
Assumes a certain parametric form of the underlying distribution	Assumes no specific functional form on the underlying distribution
Less applicability, more statistical power	More applicability, less statistical power

Null Hypothesis test

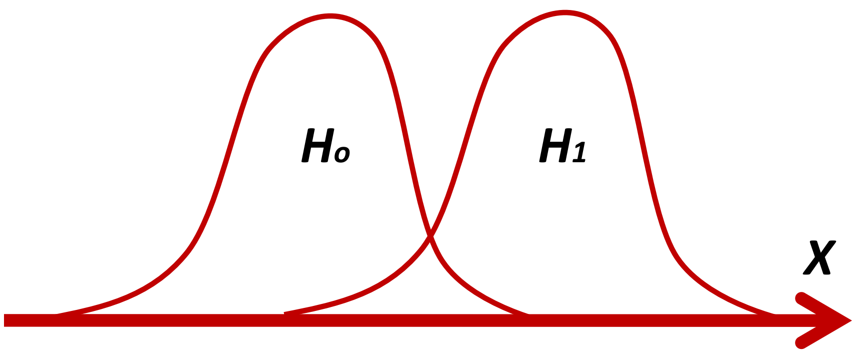
H_0: **Statement is true**
H_1: **Statement is not true**

We want to accumulate enough evidence to reject the null hypothesis.

Parametric and Nonparametric Hypothesis Testing

Parametric Tests

Nonparametric Tests



Distribution-free, but need other assumptions!

Bayesian framework =
Neyman-Pearson Lemma: **Likelihood is the test statistic, and can be always used to find a UMP**
Frequentist framework =
Can compute closed form p-vaues

One-sample vs. two-sample tests

One-sample tests	Two-sample tests
Testing whether a coin is fair	**Testing whether two coins have the same probability of heads**
$H_0 : p = 0.5$ $H_1 : p \neq 0.5$	$H_0 : p_1 = p_2$ $H_1 : p_1 \neq p_2$

Measures of significance, type-I and type-II errors, and criteria for rejection

Significance level α: the rate of false positive (type-I) errors, called the size of the test.

Significance power 1-β: the rate of false negative (type-II) errors, 1- β is called the power of the test.

$$\alpha = P(\text{reject } H_0 | H_0 \text{ is correct})$$

$$\beta = P(\text{do not reject } H_0 | H_0 \text{ is incorrect})$$

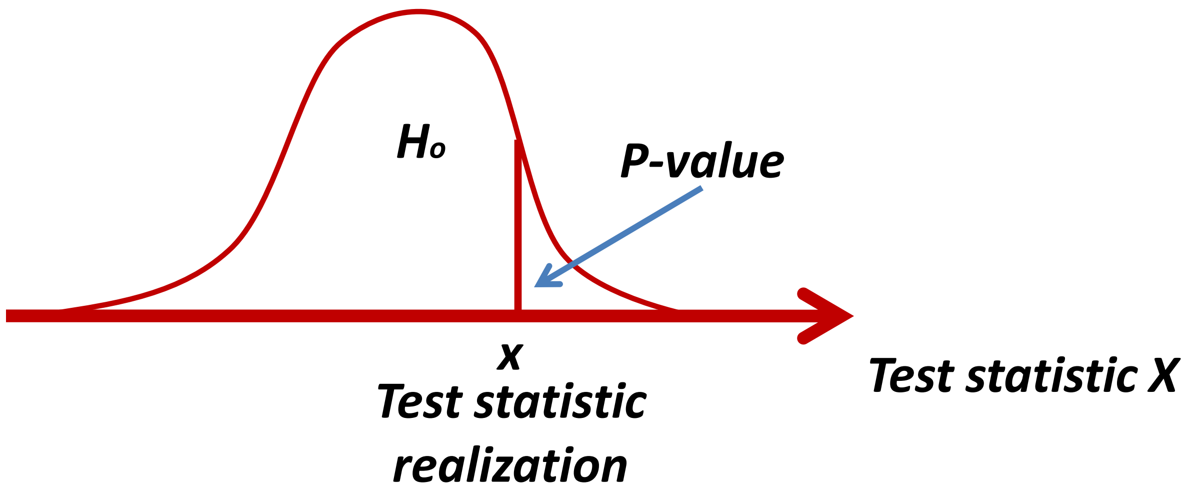
	H_0 is correct	H_0 is incorrect
Reject null hypothesis	false positive type I error (α)	true positive
Fail to reject null hypothesis	true negative	false negative type II error (β)

Measures of significance, type-I and type-II errors, and criteria for rejection

The null hypothesis is rejected whenever the p-value is less than the significance level α

P-value computation

$$p = P(X < x|Ho)$$



Should we pick a parametric or non-parametric test?

	1-sample	2-sample independent	2-sample dependent (paired)
Parametric	t-test	t-test Welch's t-test	paired t-test
Non-parametric	sign test Wilcoxon signed-rank test	median test Mann-Whitney U-test	sign test Wilcoxon signed-rank test

Should we pick a parametric or non-parametric test?

Decide the hypothesis and whether the test is one sample or two-sample

Pick an appropriate parametric test

Test the validity of Assumptions of the parametric test

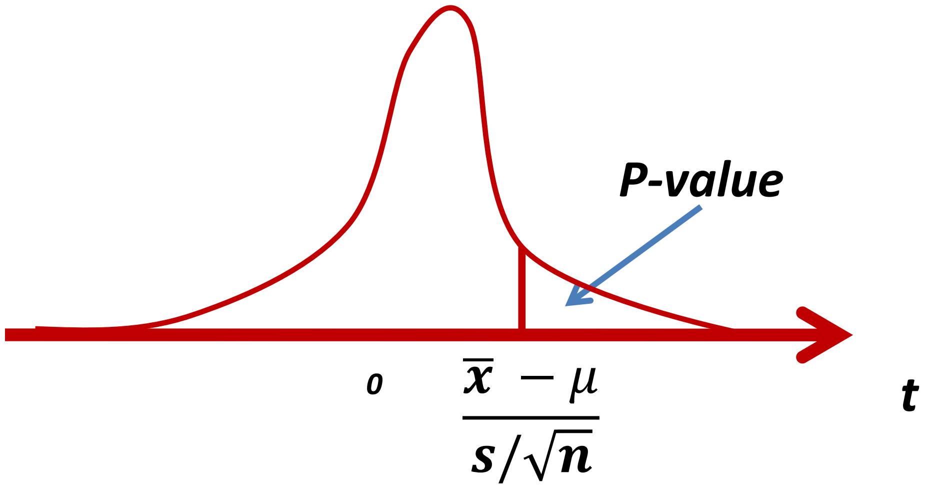
The t-test

- Assumes that the data is normally distributed: the Shapiro-Wilk test is used to check the validity of that assumption
- The test statistic follows a Student-t distribution
- One sample t-test: test whether the data sample has a mean that is close to a hypothetical mean
- Two sample t-test: test whether two data samples have significantly different means

One-sample t-test

- **Null hypothesis:** **the population mean is equal to some value μ_0**

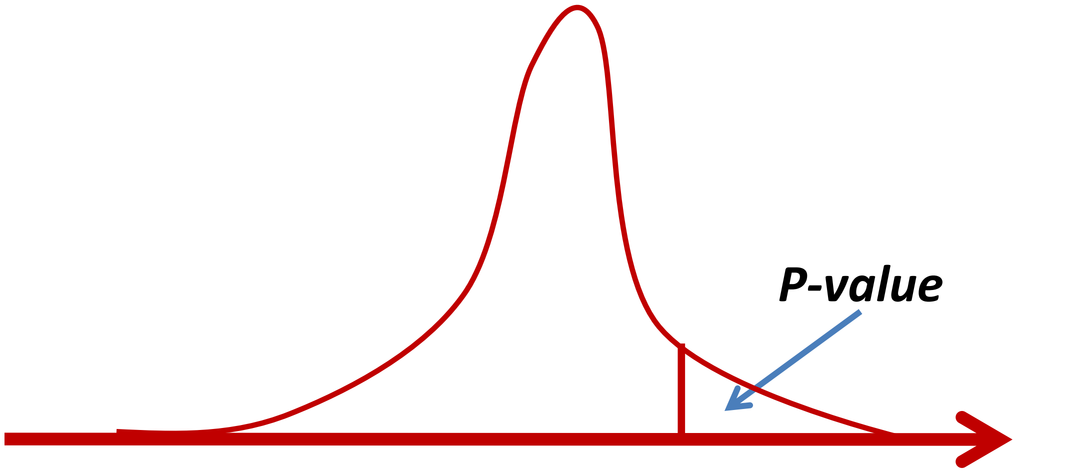
$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \qquad t \sim \mathcal{T}_{n-1}$$



Independent Two-sample t-test

- **Null hypothesis:** the population mean of two groups are equal

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{12}\sqrt{n_1^{-1} + n_2^{-1}}} \qquad t \sim \mathcal{T}_{n_1+n_2-2}$$

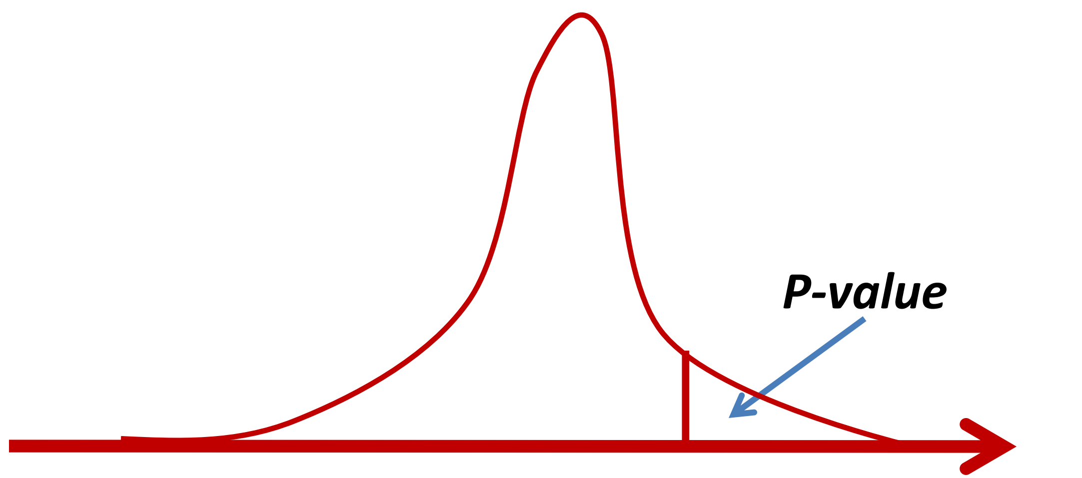


$$\frac{\bar{x}_1 - \bar{x}_2}{s_{12}\sqrt{n_1^{-1} + n_2^{-1}}}$$

Welch's t-test

- **Null hypothesis:** **the population mean of two groups are equal, but does not assume both groups have the same variance**

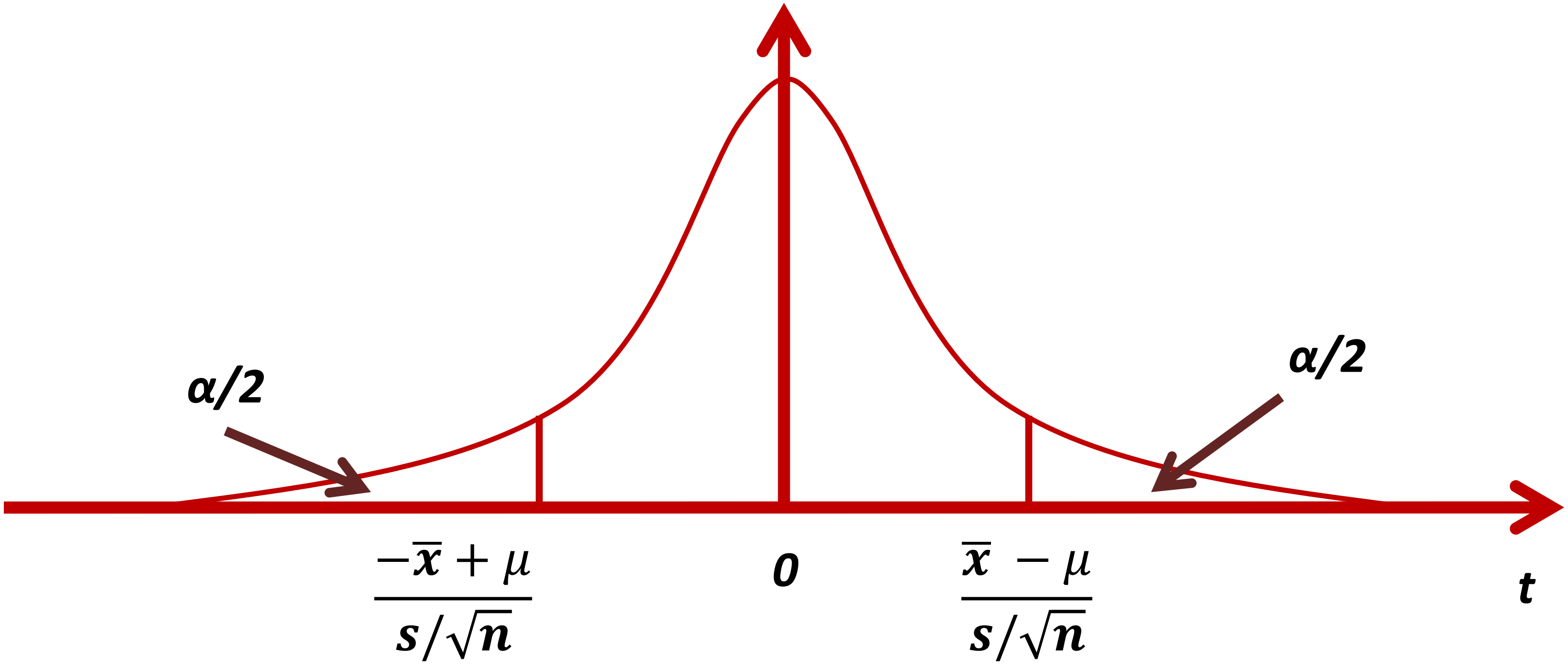
$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$



$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Typical cutoff on the t-statistic

- **Two tailed tests:** the p-values are computed with regard to the two sides of the Student-t distribution, e.g. if significance level is 0.05, then area under each side is 0.025



Typical cutoff on the t-statistic

- **Typical significance level is α = 0.05, the CDF of Student-t distribution is tabulated**

Magic number t = 2. (t-statistic cutoff)

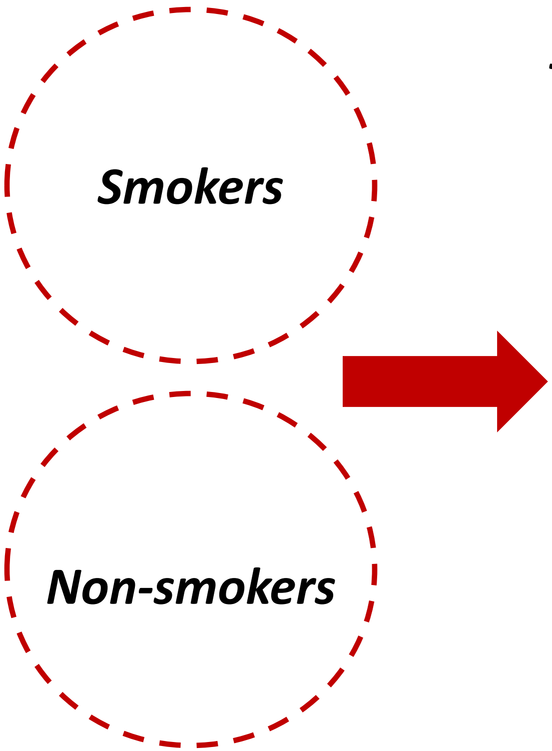
	PROPORTION IN ONE TAIL					
	0.25	0.10	0.05	0.025	0.01	0.005
	PROPORTION IN TWO TAILS					
df	0.50	0.20	0.10	0.05	0.02	0.01
1	1.000	3.078	6.314	12.706	31.821	63.657
2	0.816	1.886	2.920	4.303	6.965	9.925
3	0.765	1.638	2.353	3.182	4.541	5.841
4	0.741	1.533	2.132	2.776	3.747	4.604
5	0.727	1.476	2.015	2.571	3.365	4.032
6	0.718	1.440	1.943	2.447	3.143	3.707
7	0.711	1.415	1.895	2.365	2.998	3.499
8	0.706	1.397	1.860	2.306	2.896	3.355
9	0.703	1.383	1.833	2.262	2.821	3.250
			1.812	2.228	2.764	3.169
			1.796	2.201	2.718	3.106
			1.782	2.179	2.681	3.055
			1.771	2.160	2.650	3.012
			1.761	2.145	2.624	2.977
			1.753	2.131	2.602	2.947
			1.746	2.120	2.583	2.921
			1.740	2.110	2.567	2.898
			1.734	2.101	2.552	2.878
			1.729	2.093	2.539	2.861
			1.725	2.086	2.528	2.845
			1.721	2.080	2.518	2.831
			1.717	2.074	2.508	2.819
			1.714	2.069	2.500	2.807
			1.711	2.064	2.492	2.797
25	0.684	1.316	1.708	2.060	2.485	2.787
26	0.684	1.315	1.706	2.056	2.479	2.779
27	0.684	1.314	1.703	2.052	2.473	2.771
28	0.683	1.313	1.701	2.048	2.467	2.763
29	0.683	1.311	1.699	2.045	2.462	2.756
30	0.683	1.310	1.697	2.042	2.457	2.750
40	0.681	1.303	1.684	2.021	2.423	2.704
60	0.679	1.296	1.671	2.000	2.390	2.660
120	0.677	1.289	1.658	1.980	2.358	2.617
∞	0.674	1.282	1.645	1.960	2.326	2.576

	PROPORTION IN ONE TAIL					
	0.25	0.10	0.05	0.025	0.01	0.005
	PROPORTION IN TWO TAILS					
df	0.50	0.20	0.10	0.05	0.02	0.01
30	0.683	1.310	1.697	2.042	2.457	2.750
40	0.681	1.303	1.684	2.021	2.423	2.704
60	0.679	1.296	1.671	2.000	2.390	2.660
120	0.677	1.289	1.658	1.980	2.358	2.617
∞	0.674	1.282	1.645	1.960	2.326	2.576

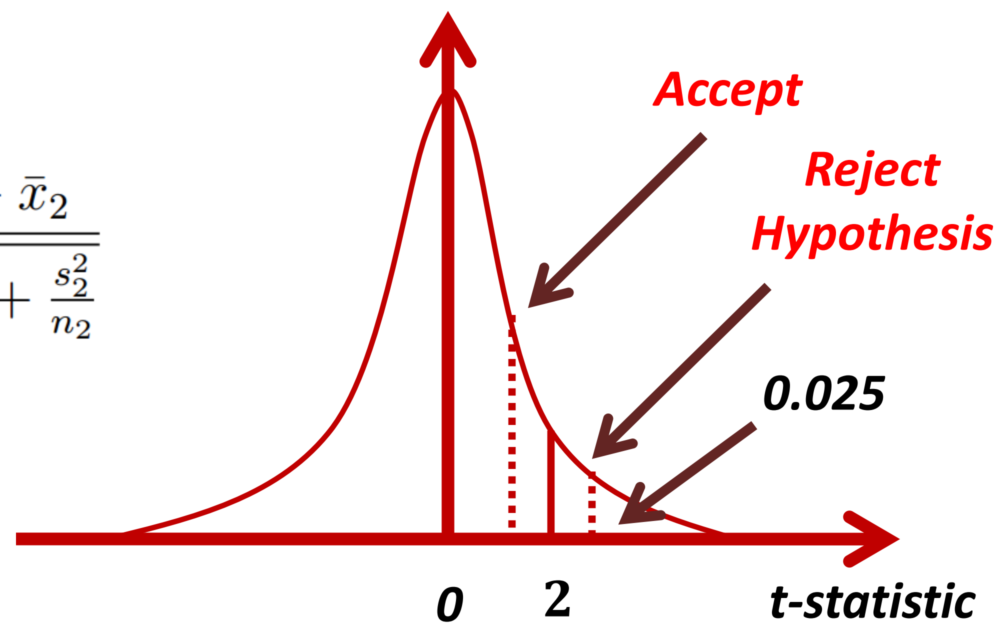
Typical cutoff on the t-statistic

- **Typical Statistical Inference in a research study:**
- **Research Question:** Is smoking a risk factor for high blood pressure?

Control demographic features (ages, ethnicities, etc)



$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$



Multiple testing

- **Testing multiple hypotheses simultaneously**
- **Should we take decision on every hypothesis separately using its marginal p-value? NO!**
- **Multiple testing matters! We may care about the whole set of tests, need a method to control false discoveries**
- **Example:**
- If $\alpha = 0.05$, and we are doing 100 tests, then the probability of making at least one true null hypothesis is rejected is given by

$$1 - (1 - 0.05)^{100} = 0.994$$

Multiple testing: p-value adjustments and type-I errors control

- **For testing M hypotheses, we have a vector of t-statistics and p-values as follows**

$$[t_1, t_2, \dots, t_M], [p_1, p_2, \dots, p_M]$$

When people say "adjusting p-values for the number of hypothesis tests performed" what they mean is controlling the Type I error rate.

Type-I error notions for multiple testing

$$\text{FWER} = \mathbb{P}\left(\sum_{i=1}^{M} \mathbf{1}_{\{t_i > t_i^*\}} \geq 1\right) \qquad \text{FDR} = \mathbb{E}\left[\frac{\sum_{i=1}^{M} \mathbf{1}_{\{t_i > t_i^*\}}}{M}\right]$$

P-value adjustment methods

Single-step methods

Individual test statistics are compared to their critical values simultaneously

Bonferroni method

Sequential methods

Stepdown methods therefore improve upon single-step methods by possibly rejecting `less Signiant' hypotheses in subsequent steps.

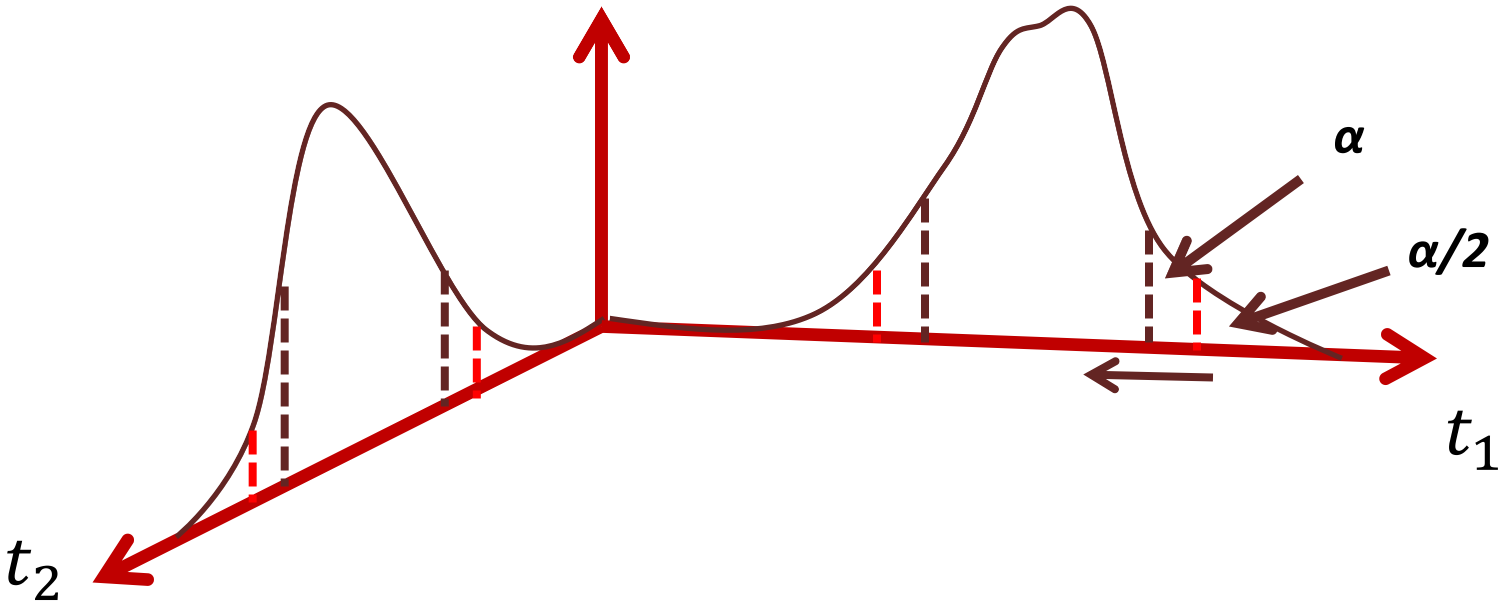
Holm’s method

Bonferroni method

- **Reject any hypothesis with a p-value less than $\frac{\alpha}{M}$**

- $\tilde{p} = \min(M\,p, 1)$

- No assumption on dependency structure, all p-values are treated similarly



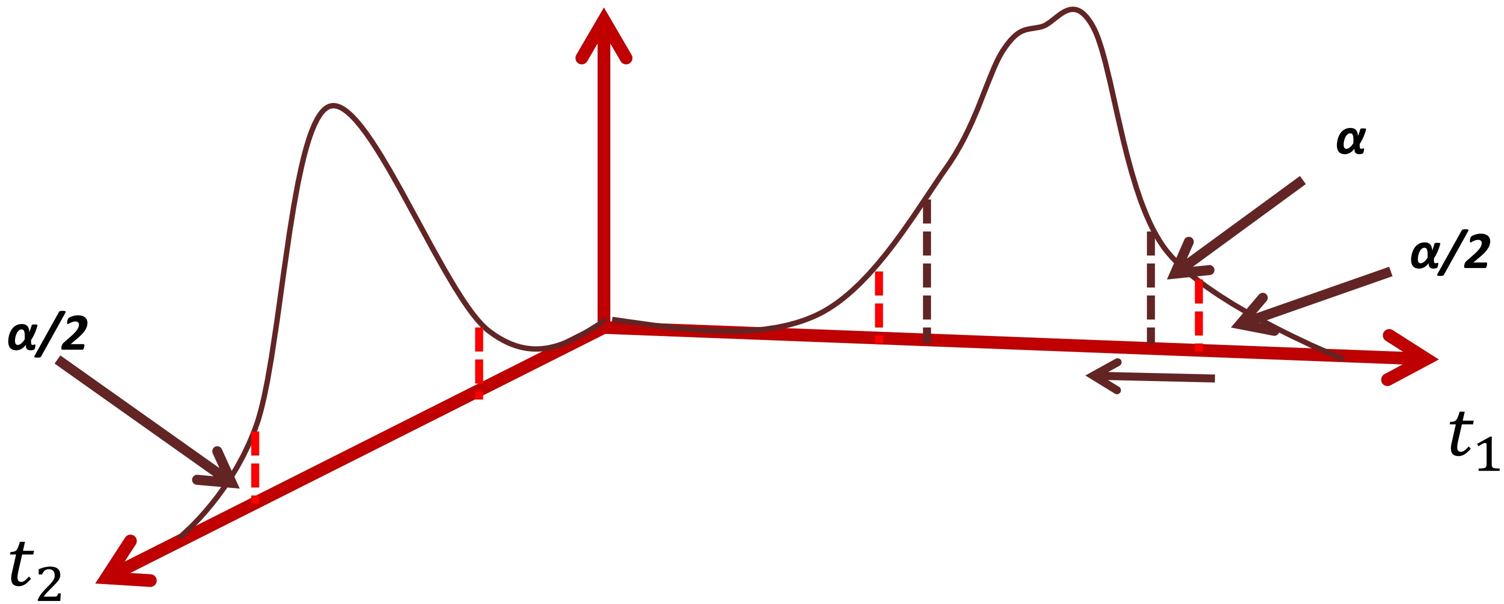
Bonferroni method: criticism

- **Counter-intuitive:** interpretation of finding depends on the number of other tests performed
- **High probability of type-II errors**: not rejecting the general null hypothesis when important effects exist.

> **"Bonferroni adjustments are, at best, unnecessary and, at worst, deleterious to sound statistical inference"**
> **Perneger (1998)**

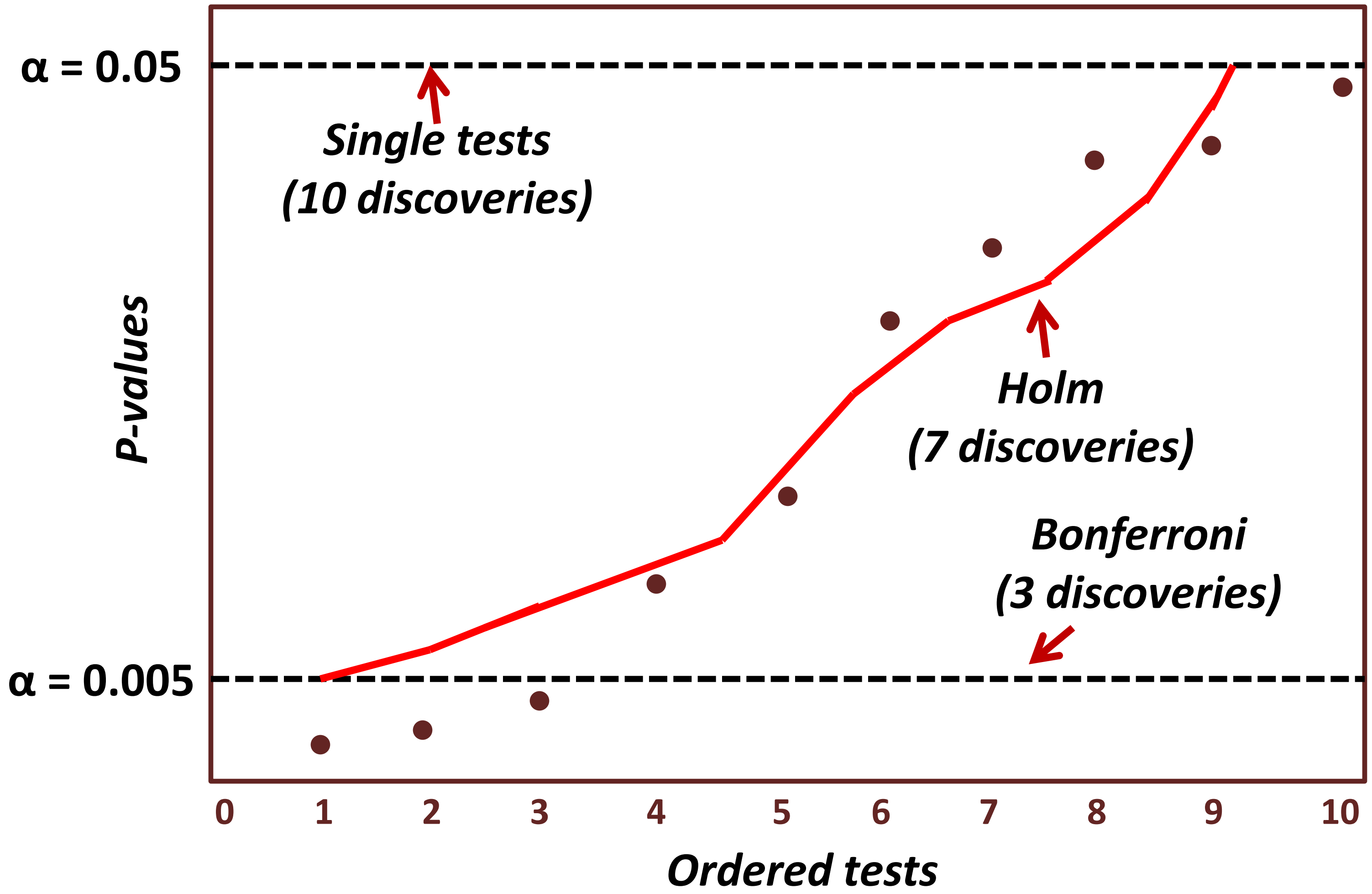
Holm's sequential method

- **Scales different p-values differently based on their significance**
- **Order the p-values by their magnitudes** $p_{(1)} < p_{(2)} < \cdots < p_{(M)}$
- $\tilde{p}_{(i)} = \min((M - i + 1)\, p_{(i)}, 1)$



Holm's vs. Bonferroni Discoveries

- For M = 10, α = 0.05



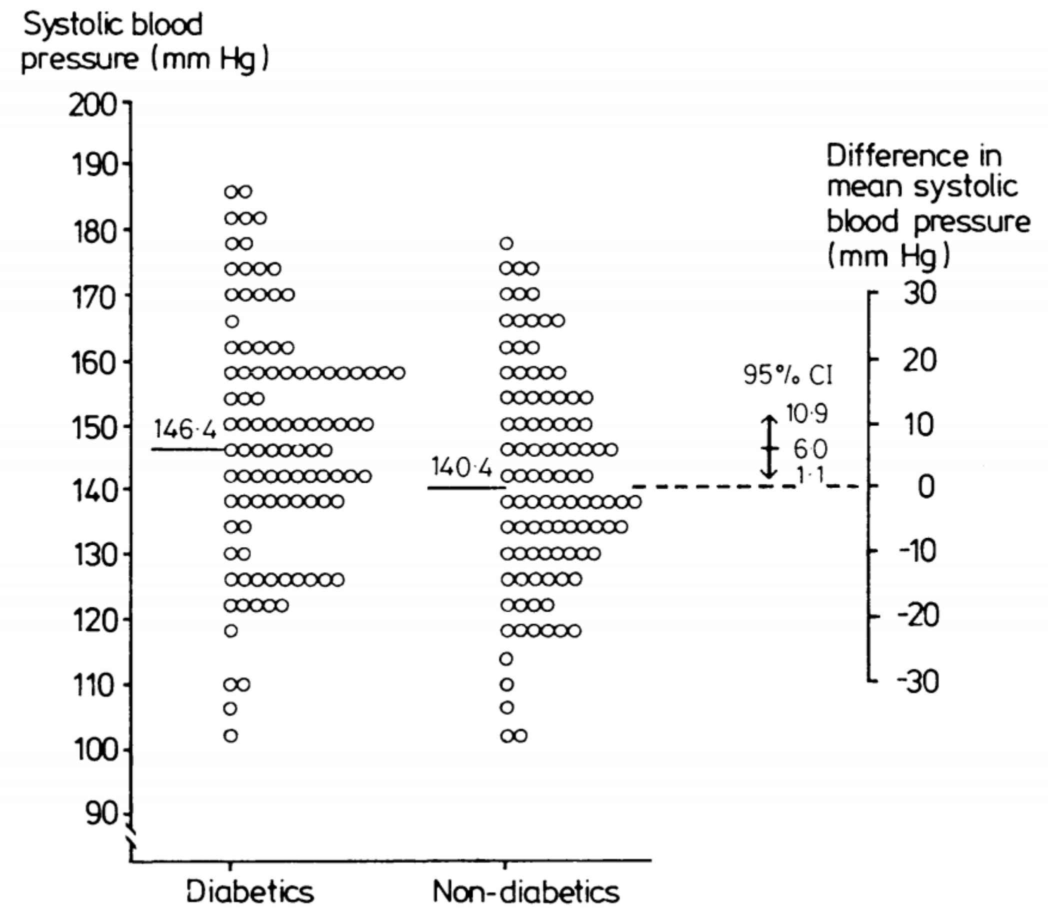
Multiplicity of testing in medical literature

P. Ioannidis, ``Why most published research findings are false?", PLoS medicine, 2005

Y. Hochberg, and Y. Benjamini, ``More powerful procedures for multiple significance testing", Stat. in Medicine, 1990.

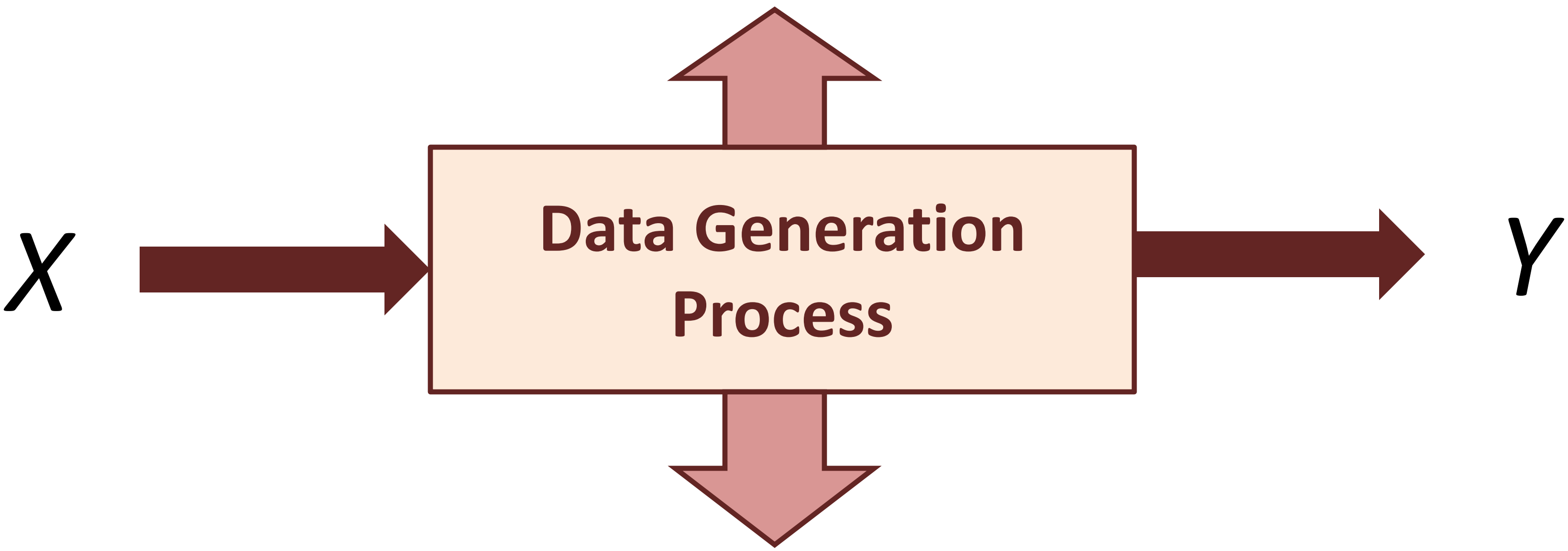
Hypothesis testing or parameter estimation?

- **Confidence intervals on t-static instead of p-values if the numeric values are themselves of interest**



The marriage of Statistical Learning and Statistical Inference

Statistical inference: draw conclusions about the population in order to select better predictors



Statistical learning: use inference to build a better predictor for Y

Personalization as a multiplicity of nested tests

- **Key idea: clustering is based on inference of subpopulation properties independent of the classifier and its complexity**

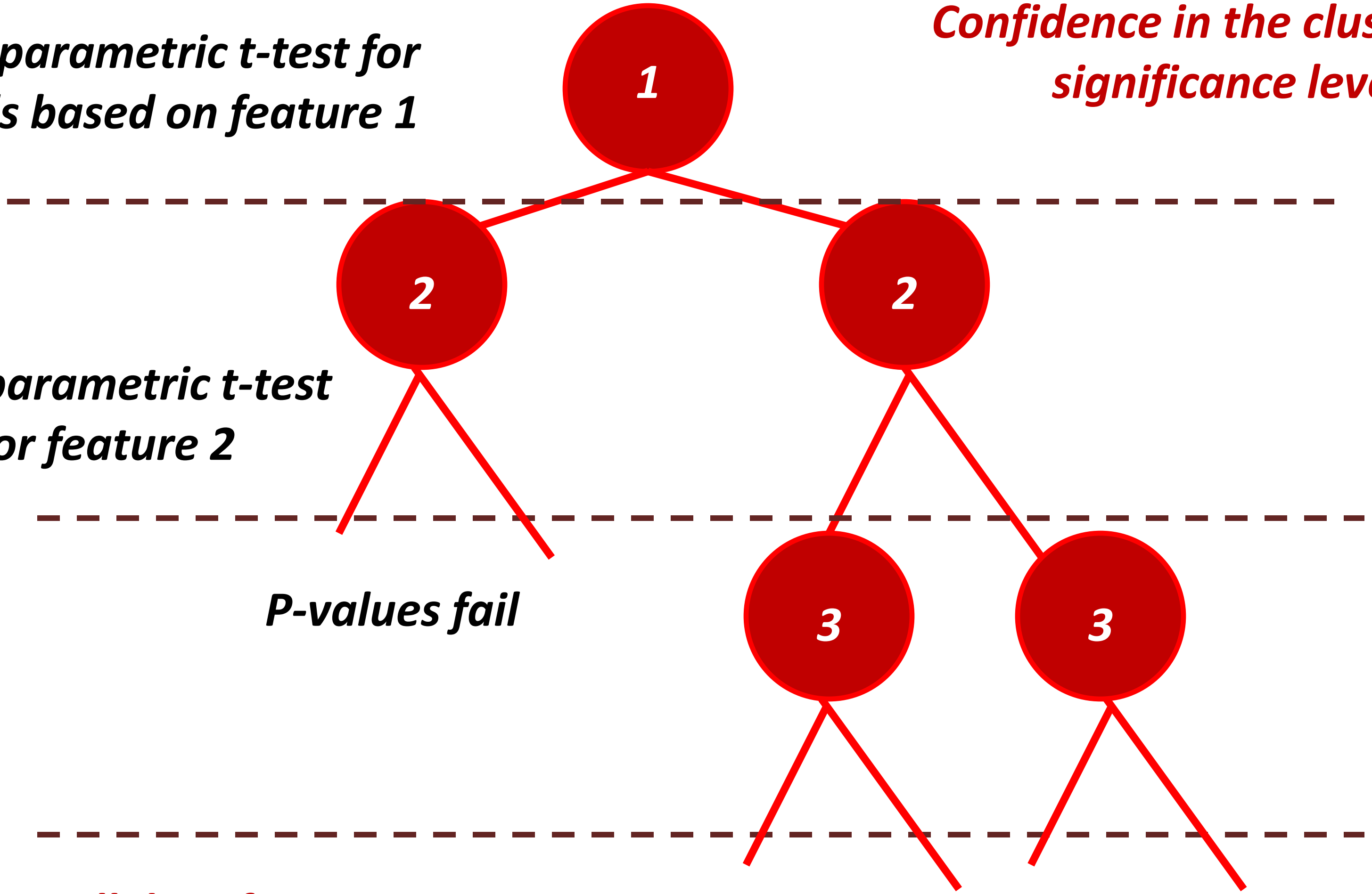
- **Group homogeneous subpopulation together**

- **FWER is now an analog of a PAC confidence bound on the homogeneity of subpopulations!!**

Classification algorithms that conduct research

Non-parametric t-test for labels based on feature 1

Confidence in the clustering = significance level!



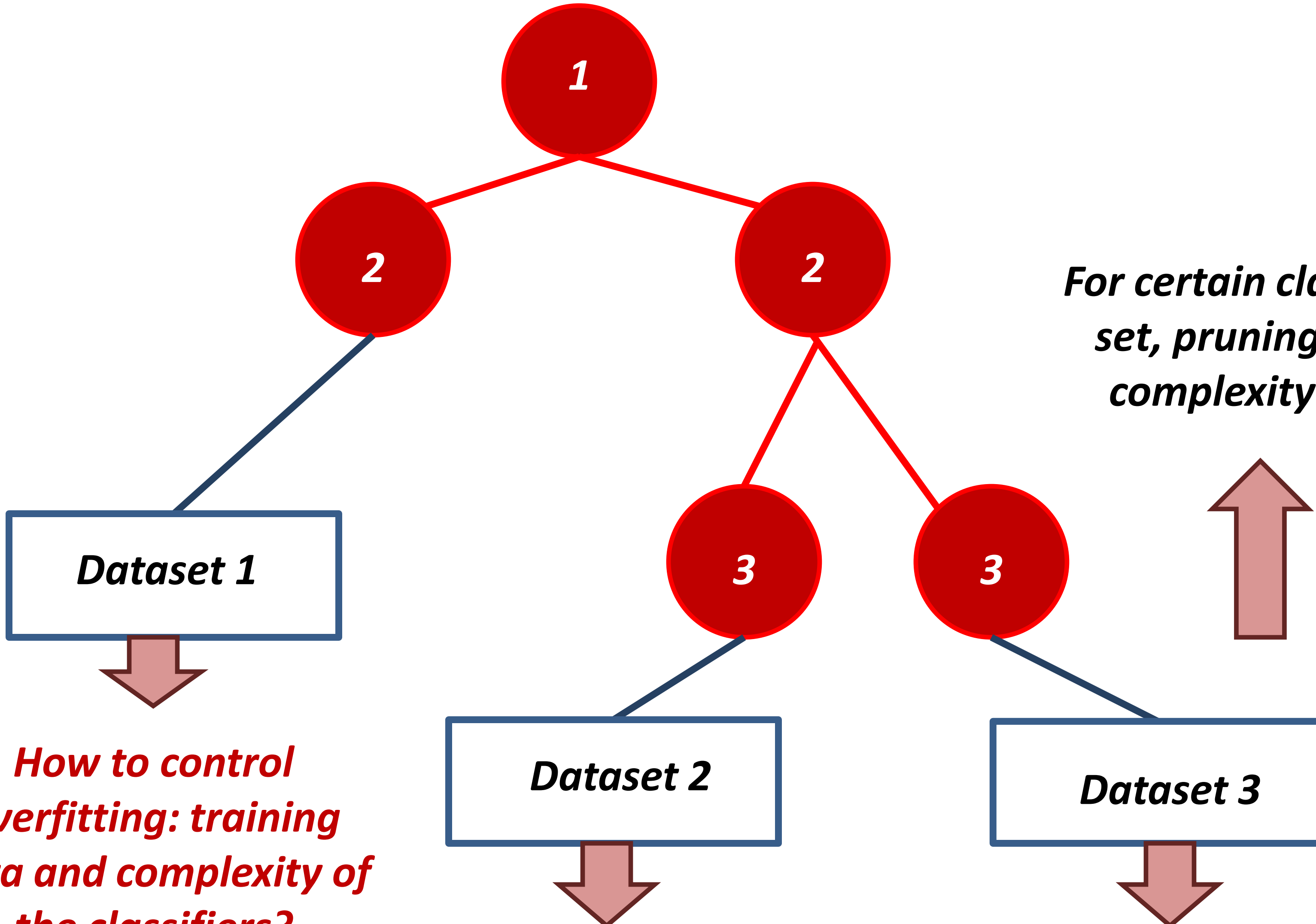
Non-parametric t-test for feature 2

P-values fail

Use all data for statistical inference!

P-values fail

Classification algorithms that conduct research



For certain classifier set, pruning by a complexity test

How to control overfitting: training data and complexity of the classifiers?