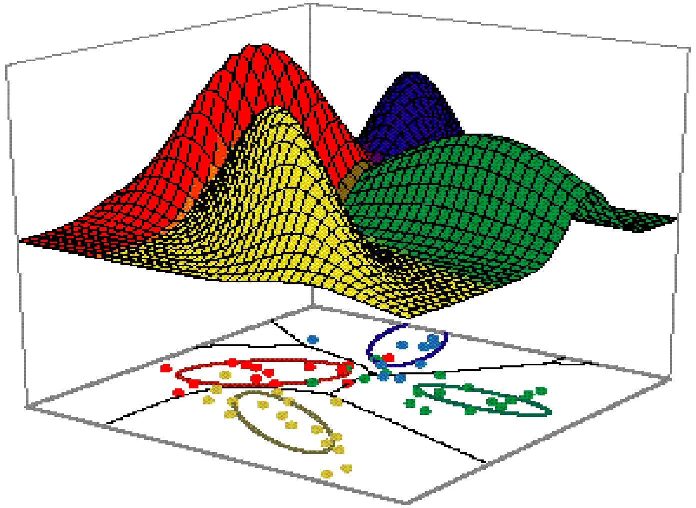
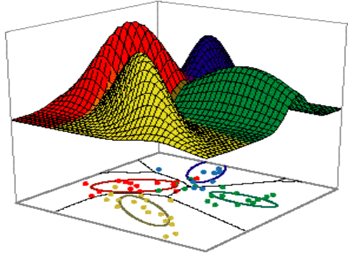
Part 3: Experiment Design



Types of experiments
Types of data
Data pre-processing
Feature selection / Dimensionality
Selecting classifier structure
Data set partitioning

Some materials in these slides were taken from Pattern Classification (2nd ed) by R. O. Duda, P. E. Hart and D. G. Stork, John Wiley & Sons, 2000 and Empirical Methods for Artificial Intelligence by Paul R. Cohen, MIT Press, 1995

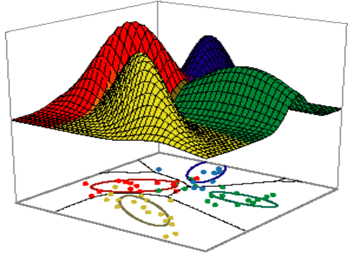
Pattern Classification as objects of Empirical Study



Paul R. Cohen:

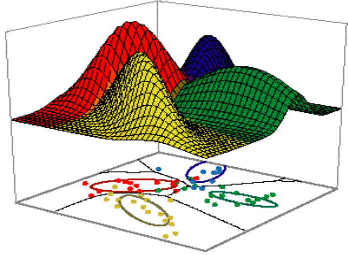
"Studying AI systems is not very different from studying moderately intelligent animals such as rats. One obliges the agent (rat or program) to perform a task according to an experimental protocol, observing and analyzing the macro-and micro-structure of its behavior. Afterward, if the subject is a rat, its head is opened up or chopped off; and if it is a program, its innards are fiddled with."

Empirical Methods for Artificial Intelligence



- Note about Cohen’s text…
  - Builds statistical models by testing for correlation/causation between factors and results/outcomes
  - In pattern classification, we typically employ methods to find relations between factors and outcomes (classes) automatically
    - e.g. train neural network connection weights to implicitly capture (nonlinear) relationships between multiple factors and outcomes.
  - Cohen’s methods may be useful for:
    - Elucidating black-box classifier (e.g., which input features are responsible for classification)
    - Determining the impact of a design variable (e.g. number of hidden nodes in an ANN classifier) on the classifier accuracy
  - Cohen text provides useful methods for experiment design and statistically rigorous reporting of results/accuracies.

Types of Empirical Studies



- Exploratory studies
  - Yield causal hypotheses to be tested in observation/manipulation experiments.
  - Collect lots of data; analyse in many ways looking for suggestive regularities/patterns.
- Assessment studies
  - Establish baselines & ranges
- Manipulation experiments
  - Test hypotheses about causal influences of factors by manipulating them and noting effects on outcomes/measured variables
- Observation experiments
  - Natural/quasi-experimental experiment
  - Observe associations between factors and measured variables to disclose effects of factors on outcomes.

“The Data Analytic Question”

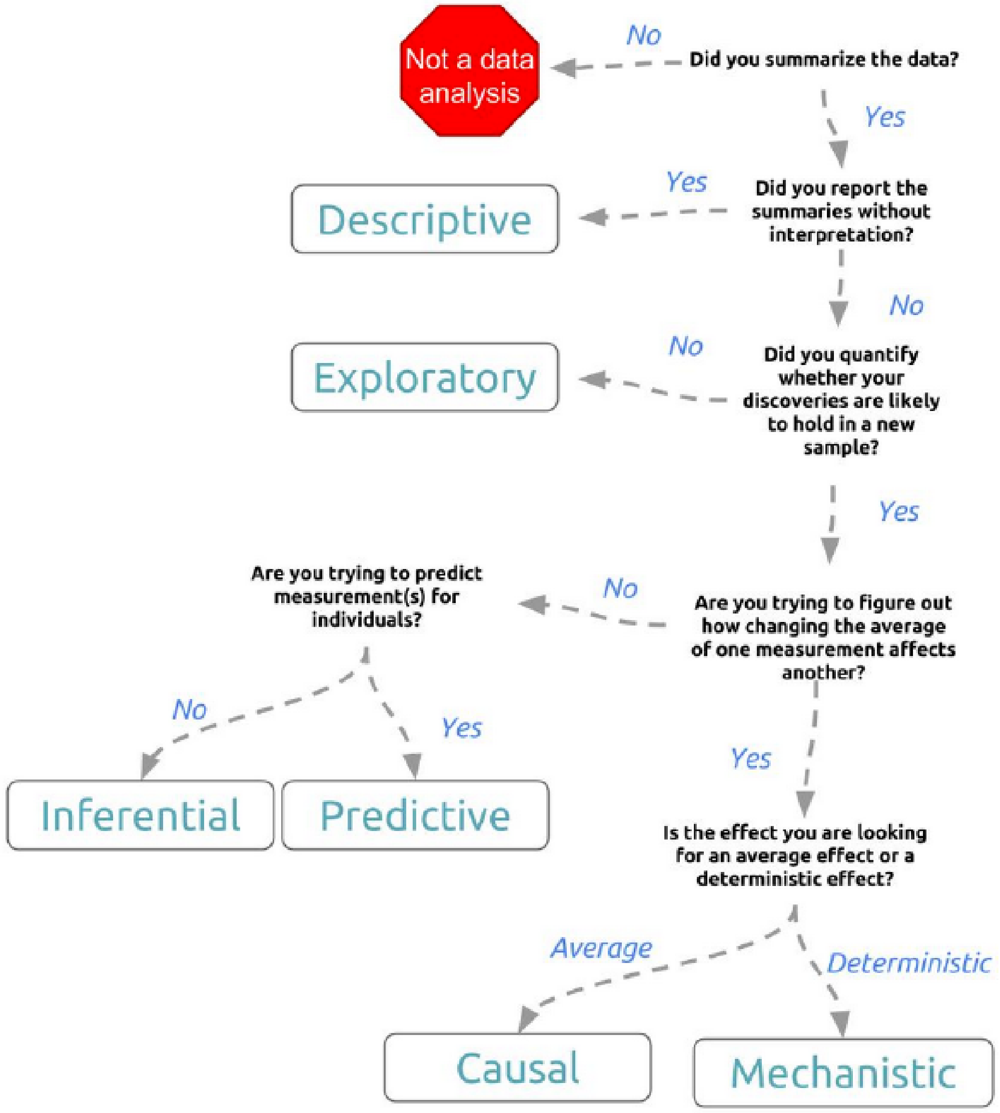
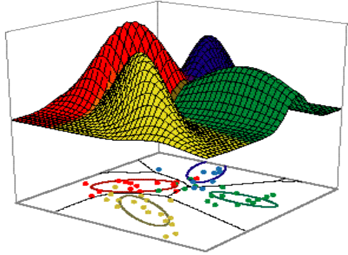
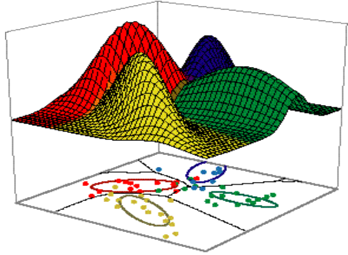


Figure 2.1 The data analysis question type flow chart



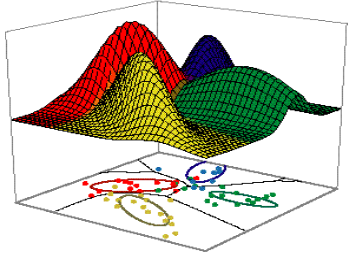
From Jeffrey Leek, <u>The Elements of Data Analytic Style</u>, 2015

Steps in Pattern Classification



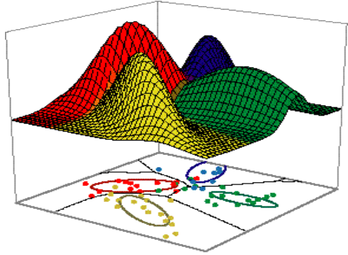
- Data pre-processing
- Selecting a learning algorithm *(throughout course)*
- Feature Selection / Representation
- Data set partitioning
- Training *(throughout course)*
- Testing & reporting results *(next week)*
- Meta-learning / CME *(later)*

Data pre-processing

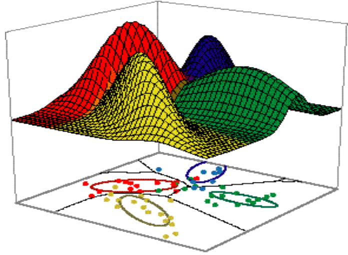


- Data may have to be normalized
  - Some classifiers are not scale invariant. Features with larger ranges may have undue influence on the classifier.
- Outliers may have to be identified and possibly removed.
- Missing data may have to be identified and possibly replaced.
- Will use visualization techniques and statistical tests to examine the data
- Always plot data to look for patterns, correlations, outliers, evidence of asymptotes / measurement saturation, etc.

Preprocessing Data - Outline



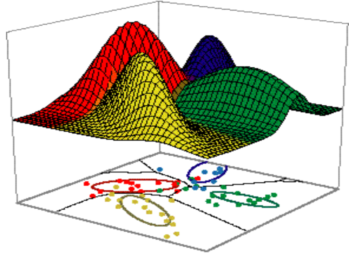
- Scales of data / transformations
- Analysis of a single variable
  - Histograms, central tendency, spread
- Analysis of pairs of variables
  - Categorical/nominal (contingency tables, χ^2 test)
  - Continuous (scatter plots, line fitting)
  - Pearson correlation test, Spearman Rank index
- Time series
- Outlier detection



Scales of Data

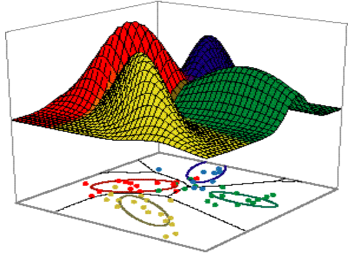
- Categorical Scale
  - Measurement assigns a category label to an individual.
  - AKA nominal when value=name
  - e.g. {cancerous, benign}
- Ordinal Scale
  - Can be ranked, but arithmetic transformations and distances between values are meaningless.
  - e.g. {low, medium, high}
- Interval Scales
  - Distances between values meaningful
  - e.g. temperature in Celsius
- Ratio Scale
  - Distances and ratios between values meaningful
  - e.g. height, temperature in Kelvin

Scales of Data



- Statistical analysis methods developed for all scales. For example:
  - Histograms: categorical
  - Spearman rank correlation: ordinal
- Can transform between types
  - Ratio→ordinal: sort data, assign increasing rank
    - [0.4, 99, 5] → [1, 3, 2]
    - Loss of information
  - Ordinal → categorical
    - e.g. transform measured age into age group label

Transforming Data

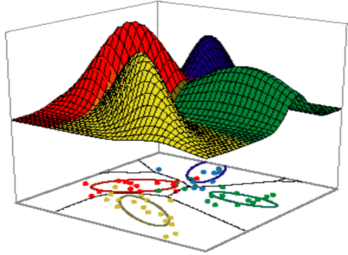


- Transformations may be useful for pre-processing
  - E.g., smooth data using moving average, apply logarithm to ‘spread’ low-valued data, etc.
  - May cause patterns to emerge.
    - But is it the transformation, or the data?
  - Turn to measurement theory…
    - E.g. Can apply any monotonic transformation to ordinal data and rank preserved:

      If $x>y \rightarrow \log(x) > \log(y)$
    - Not so for interval data:

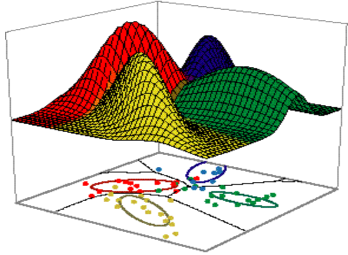
      $(x-y) \neq \log(x) - \log(y)$
  - Be careful to apply only <u>valid</u> transformations so that data continues to reflect reality.
    - Would not compute average of ordinal values… or certainly not treat like average of ratio values.

Analysis of single variable

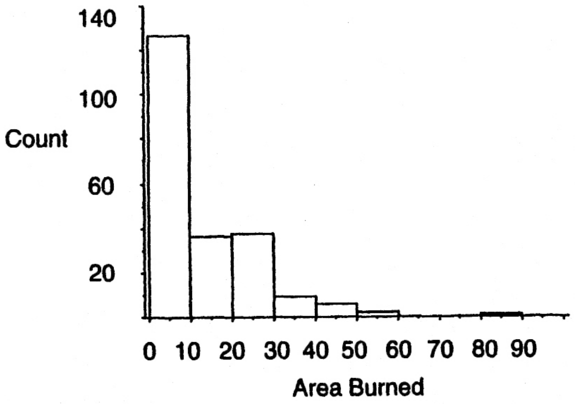


1. Histograms
2. Measures of central tendency
   - Mean vs. median vs. mode
   - Sensitivity to outliers & skew
   - Trimmed mean
3. Measures of spread
   - SD/VAR, inter-quartile range, min/max/range
   - Sensitivity to outliers

Analysis of single variable



1. Histograms
   - Must first bin values for ratio/interval/ordinal data
   - Plot frequency of each bin
   - Gaps suggest 2 things:
     - 1) Something is suppressing particular values (e.g. no floor 13 in buildings)
     - 2) There is another factor which unequally influences values left/right of gap. Try partitioning on other factor.



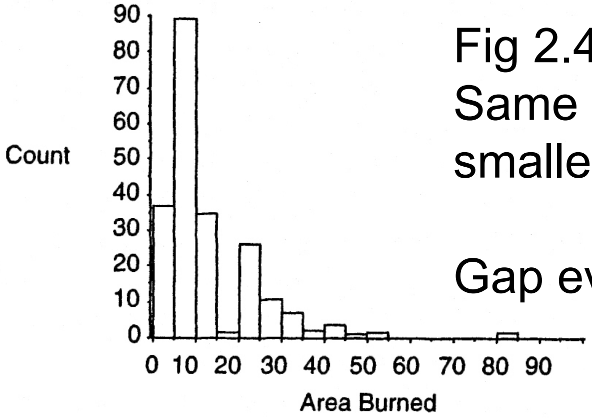
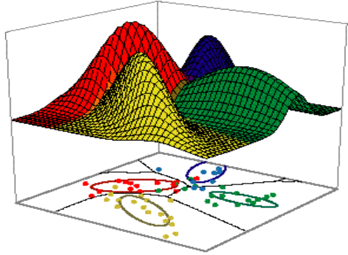


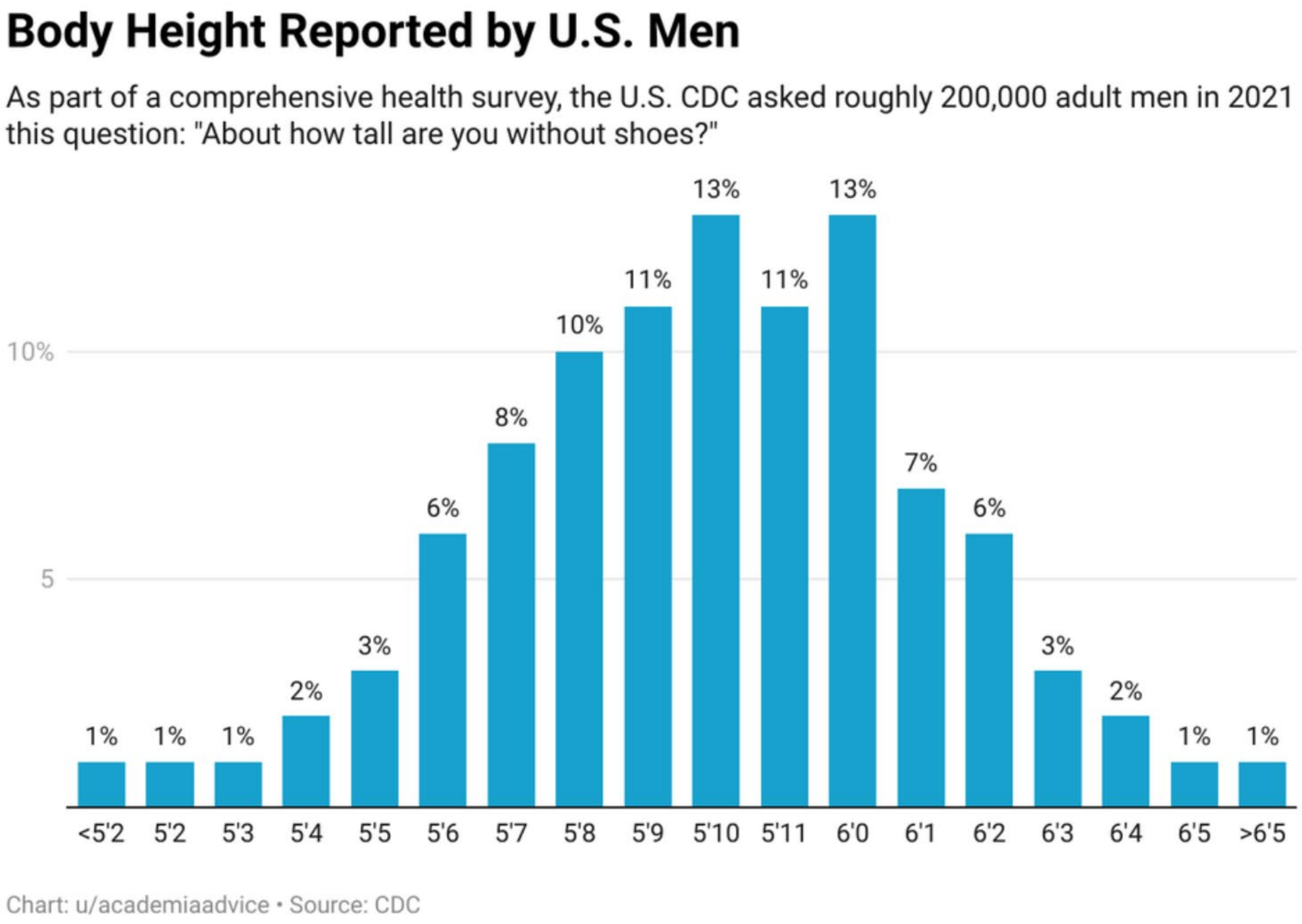
Fig 2.4
Same data,
smaller bins.

Gap evident.

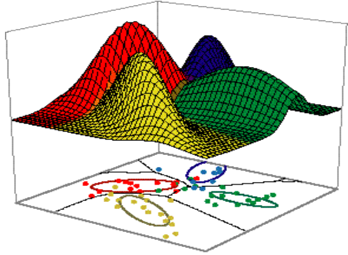


Analysis of single variable

1. Histograms



Analysis of single variable



1. Histograms
   - Note effect of bin size! Try a few before drawing conclusions…
   - Look for ‘stable shape’

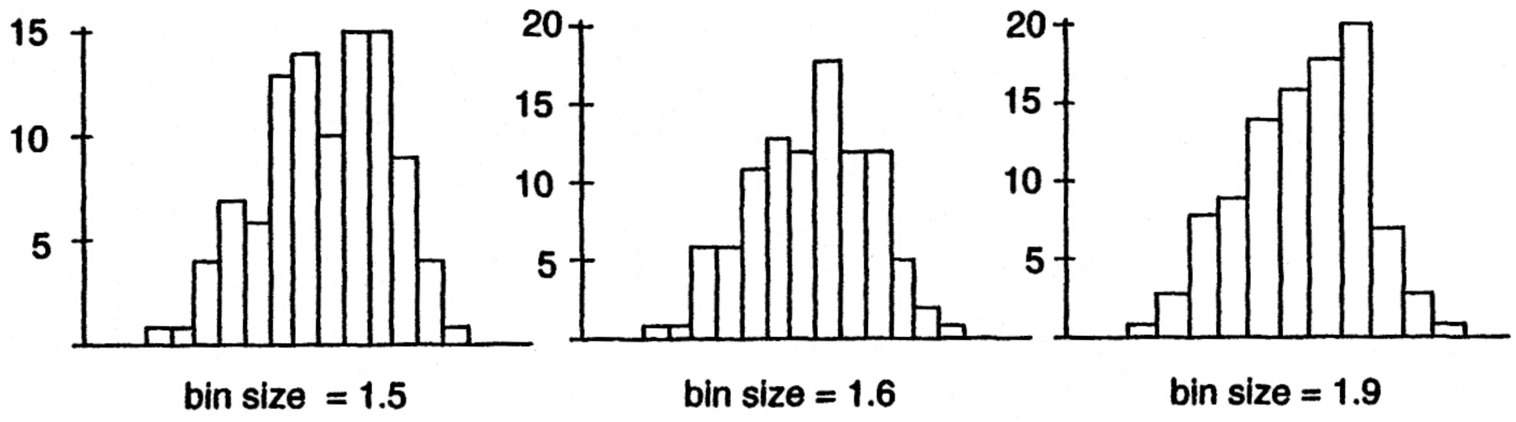
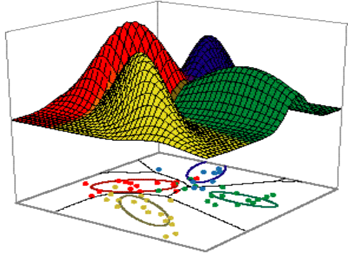


Figure 2.6 Changing bin size affects the apparent shape of a distribution.

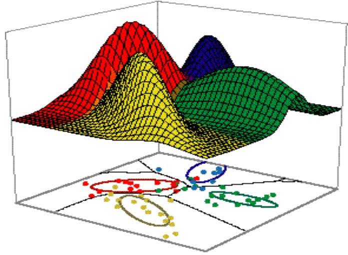
Analysis of single variable



2. Measures of central tendency

- **Mean**: arithmetic mean, average, denoted as $\overline{x}$
  - Sum of values divided by number of values
- **Median**: sort values in non-decreasing order. Median is the value that splits the distribution in half.
  - Interpolate between 2 values for even sample sizes.
  - Half of the values are larger than the median, half smaller.
- **Mode**: The most common value in a distribution.
  - Round or bin continuous values.
  - Multi-modal distributions:
    - Displays 2 separate central tendencies. Subjective decision.
    - In fig 2.4, may have 2 modes: ‘5 to 10’ and ’20 to 25’
      - (even though ’20-25’ is not the 2^{nd} largest peak…)

Analysis of single variable



2. Measures of central tendency
   - Mean = Median = Mode only for perfect unimodal symmetric distributions.
   - Real data often skewed and bumpy
     - **Skew**: Bulk of data is at one end. Bulk on one side, long tail on the other.
       - ‘right skewed’/’positive skew’ if mean>median & tail on right

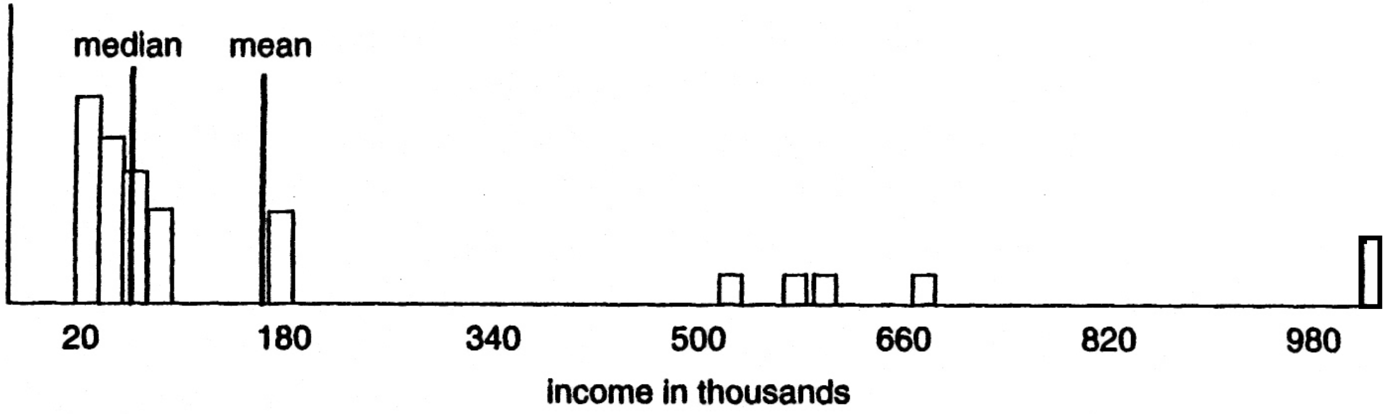
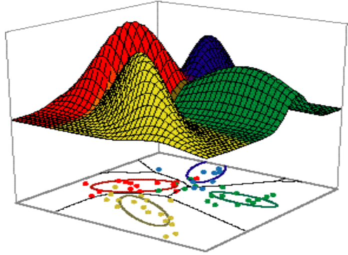


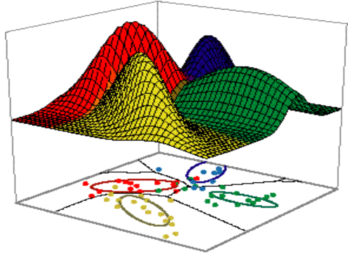
Figure 2.7 A highly skewed distribution.

Analysis of single variable



2. Measures of central tendency
   - For skewed distributions, median preferred to mean.
     - Mean easily shifted by extreme values.
     - E.g. mean income = \$178K while median = \$44K
   - Sensitivity to outliers
     - *Outlier* = value that is very large/small & uncommon
     - Median is more *robust to outliers*
       - e.g. remove 1 millionaire, median drops \$500, mean by \$32K!
   - Trimmed mean
     - Sort data, remove a fraction of upper/lower ends then take mean.

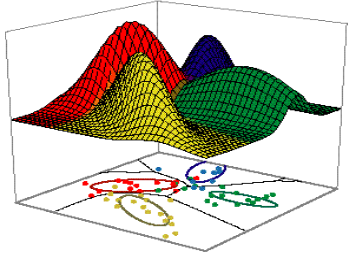
Analysis of single variable



3. Measures of spread
   - Min/max/range
     - Simplest measures of spread.
     - Range is (max – min).
   - Inter-quartile range:
     - Divide sorted distribution into 4 contiguous parts (quartiles) of same size.
     - Measure difference between highest value in 3rd quartile to lowest value in 2nd quartile.
     - More robust to outliers than simple range.
       - e.g. Compute range vs. inter-quartile range for:

         1, 1, 2, 3, 3, 5, 5, 5, 6, 6, 40, 100
   - Variance: Sum of squared distances between each datum and mean divided by the number of samples (or N-1).
     - Standard Deviation: square root of variance.
     - Variance is a mean. Therefore highly sensitive to outliers.

Analysis of pairs of categorical/ordinal variables



- Why study pairs of variables?
  - Distributions of 1 var show simple effects
  - Joint distributions show interaction effects
- Use contingency table:

Table 2.3 The joint distribution of *Outcome* and *WindSpeed*.

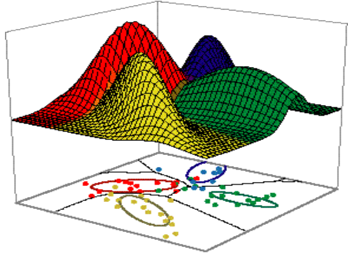
Wind Speed	Outcome=success	Outcome=failure	Totals
Low	85	35	120
Medium	67	50	117
High	63	43	106
Totals	215	128	343

- More informative using row marginal counts:

Table 2.4 The distribution in table 2.3 expressed as percentages.

Wind Speed	Outcome=success	Outcome=failure	Totals
Low	71 percent	29 percent	120
Medium	57 percent	43 percent	117
High	59 percent	41 percent	106
Totals	215	128	343

Analysis of pairs of categorical/ordinal variables



- Look for relationship between variables
  - In contingency table

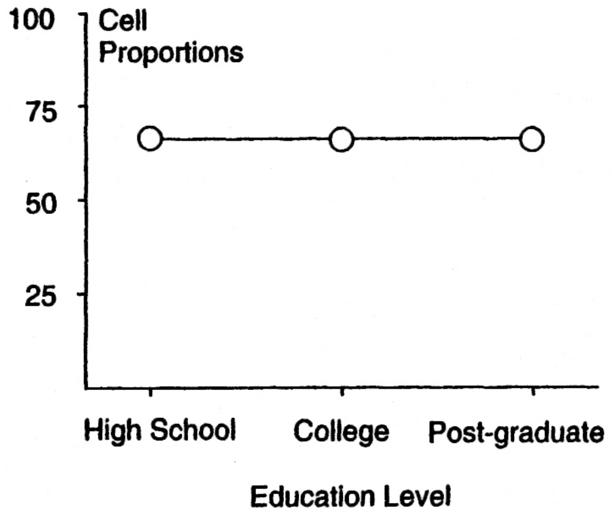
Table 2.7 Fewer individuals are found at higher educational levels, and more are married, but these are independent effects.

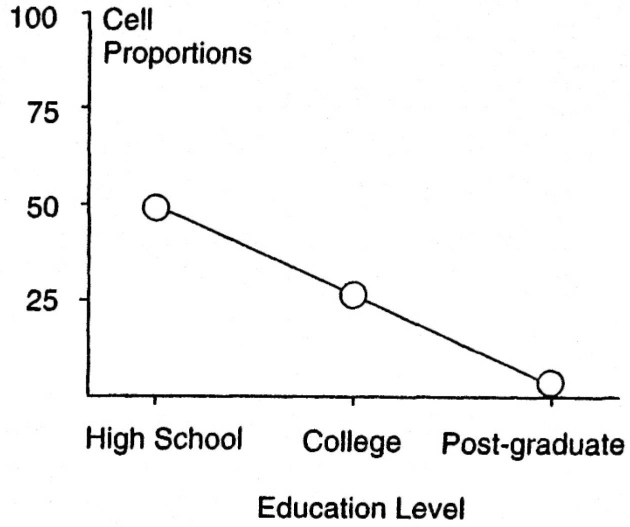
Educational level	Unmarried	Married	Totals
Postgraduate	16 (32 percent)	34 (68 percent)	50
College	35 (35 percent)	65 (65 percent)	100
High school	50 (33 percent)	100 (67 percent)	150
Totals	101	199	300

Table 2.8 A dependency between educational level and marital status.

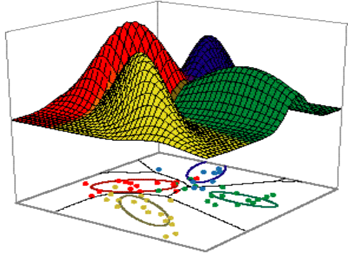
Educational level	Unmarried	Married	Totals
Postgraduate	50 (100 percent)	0 (0 percent)	50
College	75 (75 percent)	25 (25 percent)	100
High school	75 (50 percent)	75 (50 percent)	150
Totals	200	100	300

  - By plotting cell proportions





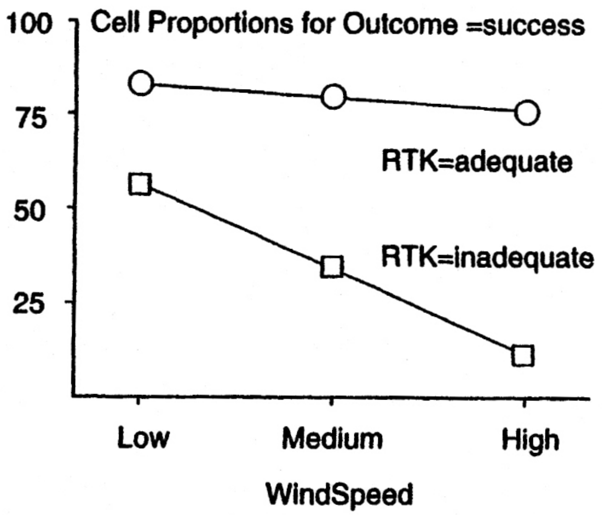
Analysis of multiple categorical/ordinal variables



- Extend to multiple variables:

Table 2.10 A three-dimensional contingency table.

		Outcome =success	Outcome =failure	Total
RTK=Adequate	WindSpeed=low	30 (86 percent)	5 (14 percent)	35
	WindSpeed=medium	32 (80 percent)	8 (20 percent)	40
	WindSpeed=high	53 (77 percent)	16 (23 percent)	69
RTK=Inadequate	WindSpeed=low	55 (65 percent)	30 (35 percent)	85
	WindSpeed=medium	35 (45 percent)	42 (55 percent)	77
	WindSpeed=high	10 (27 percent)	27 (73 percent)	37

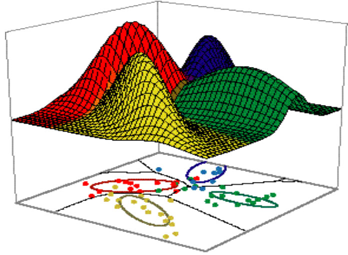


Use multiple plots for each value of extra variable.

Here, RTK=adequate is always above inadequate…
→ RTK has an effect on outcome

Also, lines are not parallel
→ windspeed affects outcome differently depending on RTK

Quantitative Analysis of contingency tables, the $\mathcal{X}^2$ test



- (Pearson's) $\mathcal{X}^2$ test: are 2 variables independent?
  1. Complete contingency table with observed cell counts
  2. Compute expected cell counts under the assumption that they are independent
     - Indep → can use row/column margins (total probability)
  3. Compare expected and observed cell counts

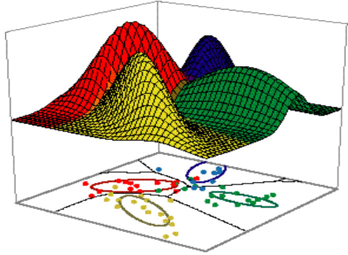
$$\chi^2 = \sum \frac{(f_0 - f_e)^2}{f_e}$$

f_e = expected frequency

f_o = observed frequency

  4. Compute/look up probability of total difference.
     - Large values of $\mathcal{X}^2$ indicate that variables are NOT independent.
     - Must correct for number of cells (degrees of freedom)
       - df = (NumRows-1)(NumCols-1)

Quantitative Analysis of contingency tables, the $\mathcal{X}^2$ test



	Outcome = success	Outcome = failure	Total
WindSpeed=low	30	5	35
WindSpeed=medium	32	8	40
WindSpeed=high	53	16	69
Total	115	29	144

Table 2.13 The expected frequencies f_e for table 2.11.

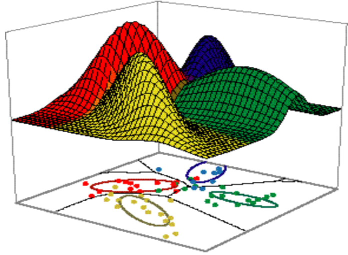
	Outcome =success	Outcome =failure	Total
WindSpeed=low	27.95	7.05	35
WindSpeed=medium	31.94	8.06	40
WindSpeed=high	55.1	13.9	69
Total	115	29	144

Pr(wind=low) = 35/144=.243
Pr(success) = 115/144=.799
If indep, Pr(low & success)=.243*.799=.194
f_e = 144 * Pr(low&success) = 144 * .194 = 27.95
OR f_e=35*115/144=27.95

$$\chi^2 = \frac{(30-27.95)^2}{27.95} + \frac{(5-7.05)^2}{7.05} + \ldots$$
$$= 1.145$$

Here, P($\mathcal{X}^2$>=1.145) = 0.56 (for d.f.=2) → do not reject null hypothesis

Quantitative Analysis of contingency tables, the χ^2 test



- Special case of 2 binary variables:

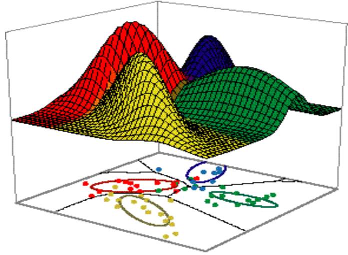
	Y=0	Y=1
X=0	a	b
X=1	c	d

$$\chi^2 = \frac{(ad - bc)^2 N}{(a+b)(c+d)(a+c)(b+d)}$$

$$N = a + b + c + d$$

df =1 (degree of freedom)

Analysis of pairs of Continuous Variables



- Use scatter plot
  - Look for relationship (linear or otherwise)
  - Random scatter, horizontal & vertical lines sign of independence

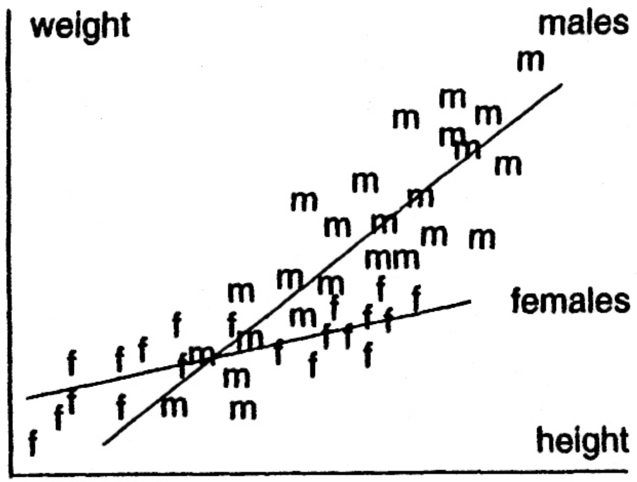
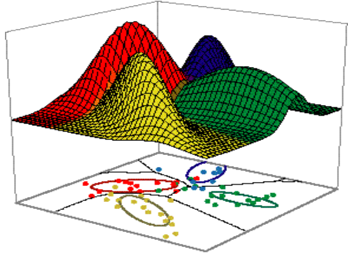


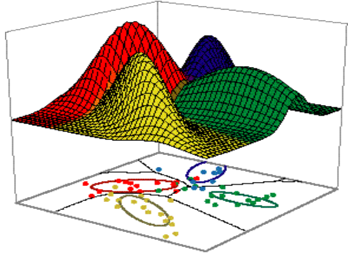
Figure 2.19 Point coloring shows that the relationship between two variables depends on a third.

Quantitative Analysis of pairs of continuous variables



- Fitting lines to scatter plots
  - Can use linear regression
    - ‘Residual’ is measure of how much of the variation in the data is NOT accounted by the best-fit linear relationship.
    - https://www.reddit.com/r/dataisbeautiful/comments/xeo0lk/visualizing_the_sum_of_squares_and_r%C2%B2_calculation/
  - By eye is often just as good or better
  - Three-group resistant line:
    - Sort points by x, divide into low, med, high
    - Find x and y medians of each group
    - Draw line between lowest median (x,y) and highest
    - Use middle group median to shift line up/down.
  - Can use nonlinear or piecewise linear function
  - Use prior knowledge
    - Special/key points, range of variables, etc.

Quantitative Analysis of pairs of continuous variables



A = linear

B = Linear with log transformation

C = quadratic
(note that line goes
above 100%!!)

D = quad with ln

E = piece-wise linear
(black dot known
to be special

F = piecewise linear with ln

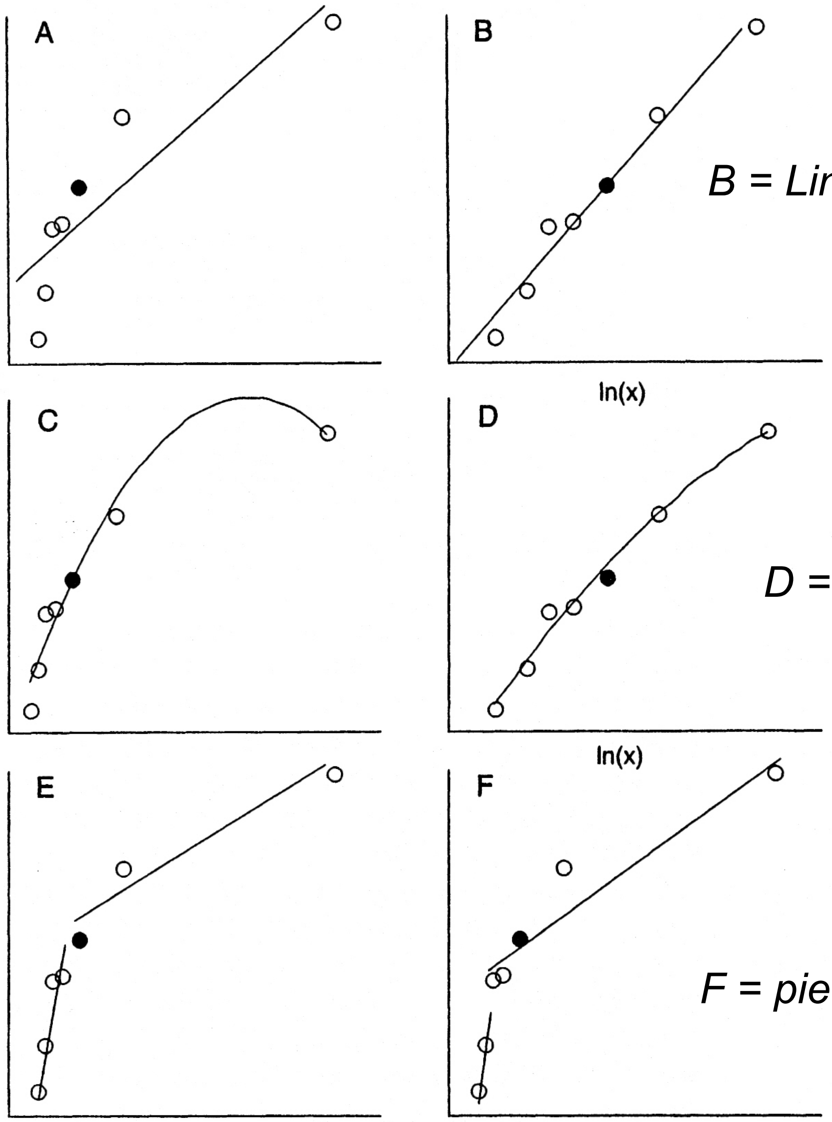
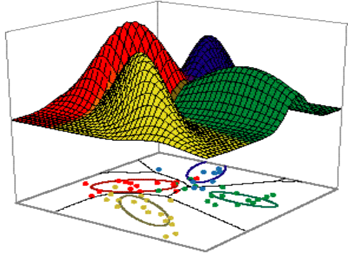


Figure 2.22 Six different fits to the same data.

Quantitative Analysis of pairs of continuous variables



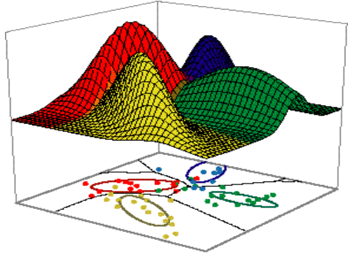
- Sample covariance:

$$\mathrm{cov}(x,y)=\frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{n-1}$$

- Standardize to [-1,1] using Pearson’s Correlation Coefficient:

$$r_{xy}=\frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{(n-1)\sigma_x\sigma_y}$$

Quantitative Analysis of pairs of continuous variables



	A	B	C
COV	+26.25	-24.25	-2.75
r_{xy}	+0.995	-0.919	-0.104

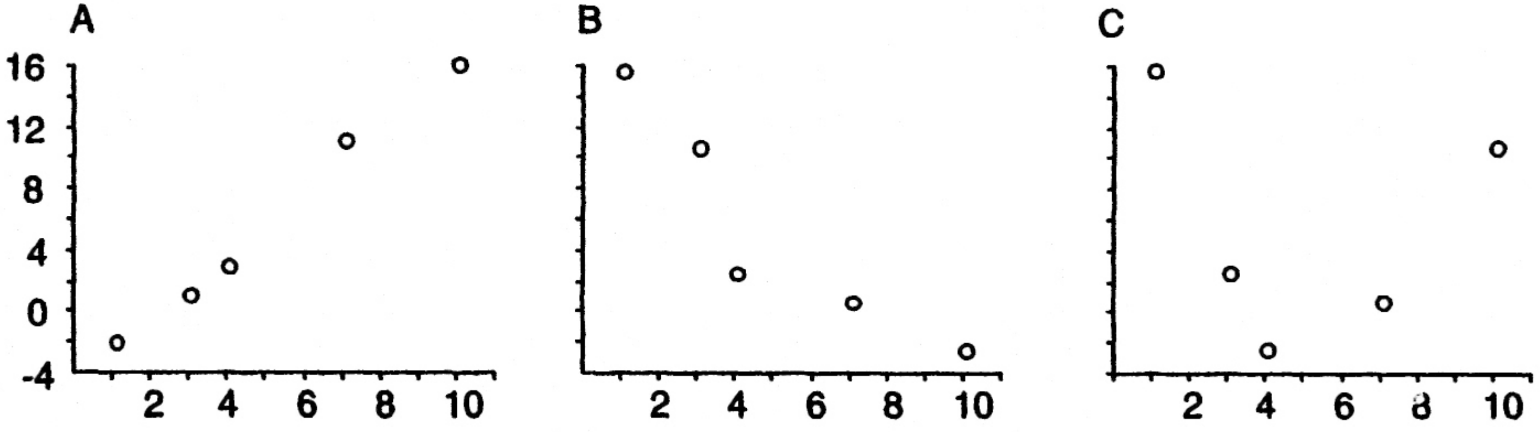
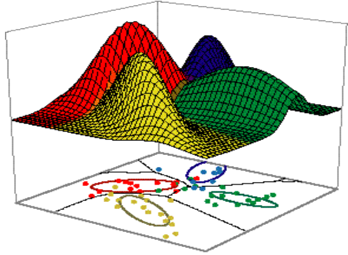


Figure 2.24 Positive, negative and nonlinear relationships between *x* and *y*.

Quantitative Analysis of pairs of continuous variables



- Always visually inspect scatter plot before drawing conclusions from r_{xy}.
  - Here all plots have same r_{xy}.
  - Effect of outliers visible in C and D.

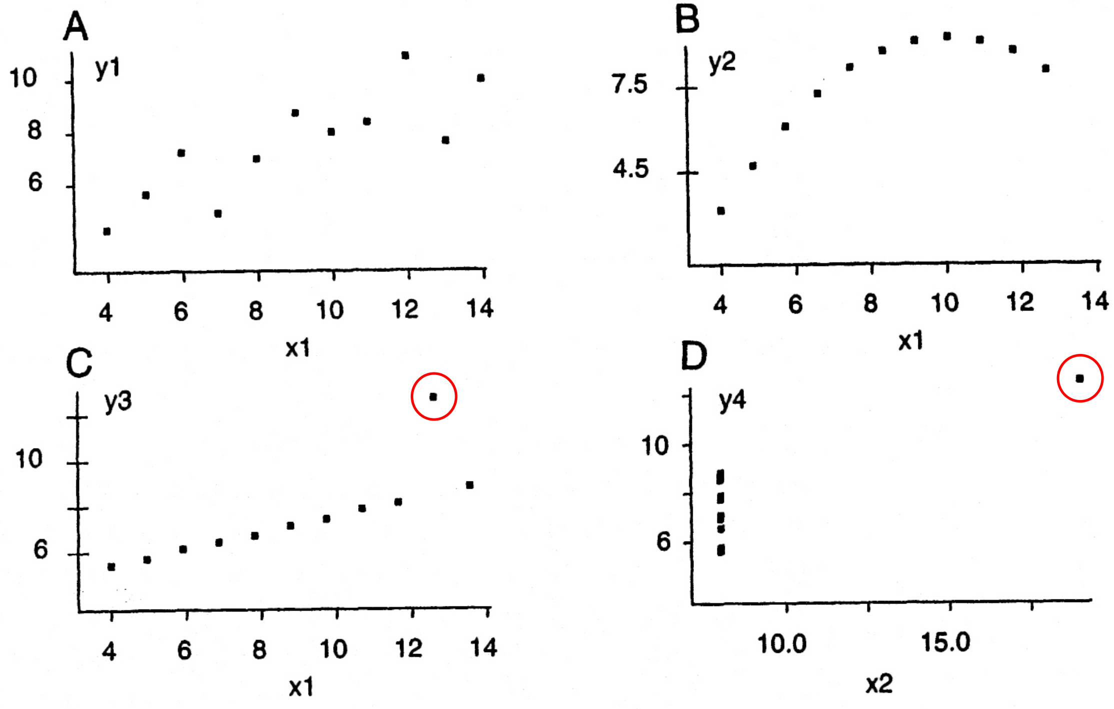
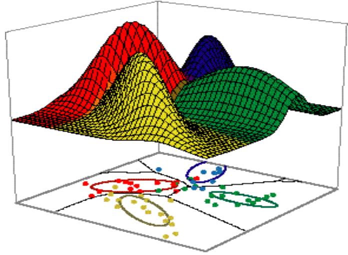


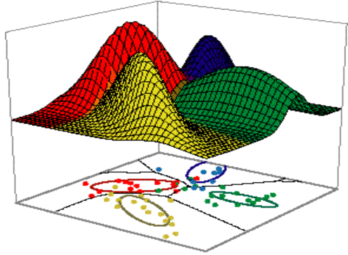
Figure 2.25 Why the correlation coefficient can be misleading (from Anscombe, 1973).

Other measures of association



- Pearson correlation not the only measure
- Spearman's rank correlation
  - Similar to Pearson, but works on rank
  - Does not require linear relationship
  - Assumes inter-rank 'distances' are roughly equal
  - Less sensitive to outliers
- Kendall's tau
  - Count number of times rank of y is what it should be (for perfect correlation), given rank of x
  - Form score from number of correct rank-pairs vs. incorrect rank pairs.
  - Does not require that inter-rank 'distances' are equal

Spearman's rank correlation



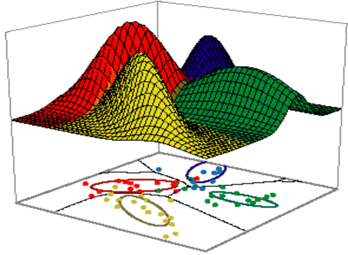
- Mathematically equivalent to Pearson correlation over ranks
- -1 ≤ *p* ≤ +1
- When no tied rank exist, simplifies to:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

d_i = the difference between each rank of corresponding values of x and y,
n = the number of pairs of values.

From Wikipedia

Spearman's rank correlation



- Example:

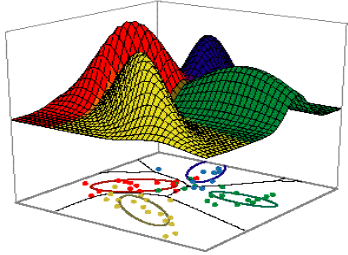
IQ (i)	Hours of TV per week (t)	rank (i)	rank (t)	d	d^2
86	0	1	1	0	0
97	20	2	6	4	16
99	28	3	8	5	25
100	27	4	7	3	9
101	50	5	10	5	25
103	29	6	9	3	9
106	7	7	3	4	16
110	17	8	5	3	9
112	6	9	2	7	49
113	12	10	4	6	36

$$\sum d_i^2 = 194 \qquad \rho = 1 - \frac{6 \times 194}{10(10^2 - 1)}$$

which evaluates to ρ = − 0.175758

From Wikipedia

Spearman's rank correlation

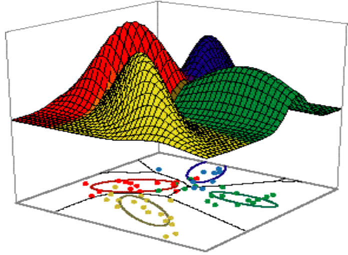


- Determining significance
  - Want to test if it is significantly different from 0
    - Null hypothesis is that there is no significant difference
  - Can use Student's T-test distribution
    - Transform $\rho \rightarrow$ t (Kendal & Stuart, 1948)
  - Better approach is to perform permutation trials:
    - (also called a randomization test, re-randomization test, or an exact test; accounts for ties in values)

- *Note that we will spend more time on hypothesis testing in Part 4…*

From Wikipedia

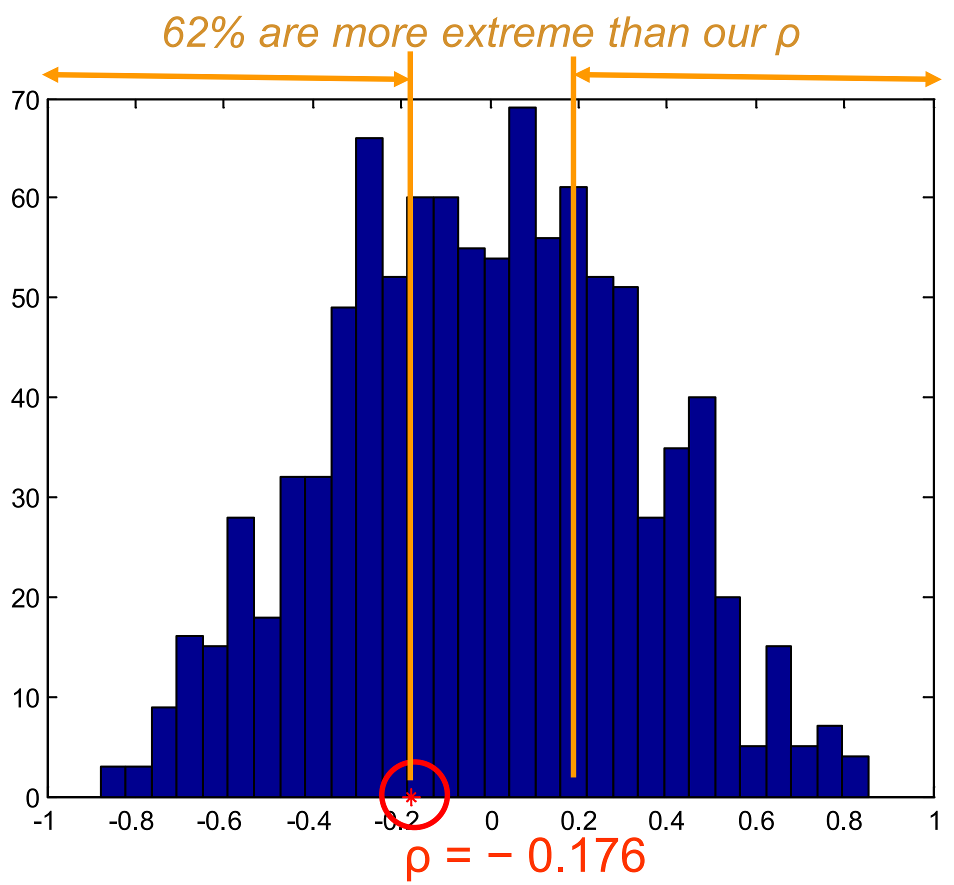
Permutation test



- Procedure
  - Randomly shuffle ranks of y, re-compute ρ.
  - Repeat several times (1000’s)
  - Count how many times a value as extreme as our *p* was observed.
    - E.g. If 95% of the data is less than |*p*|, then we can reject the null hypothesis at 5% confidence level.
      - ‘p-value’ is measure of residual uncertainty. Here p=0.05.
        - ‘probability of being wrong when rejecting H_0’
- Advantages
  - Permutation tests exist for any test statistic, regardless of whether distribution is known.
  - Can combine dependent tests on mixtures of categorical, ordinal, and ratio data

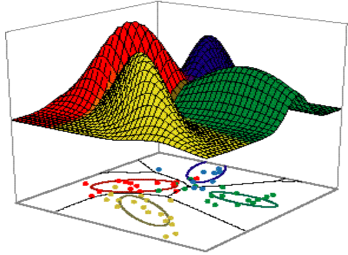
Permutation test

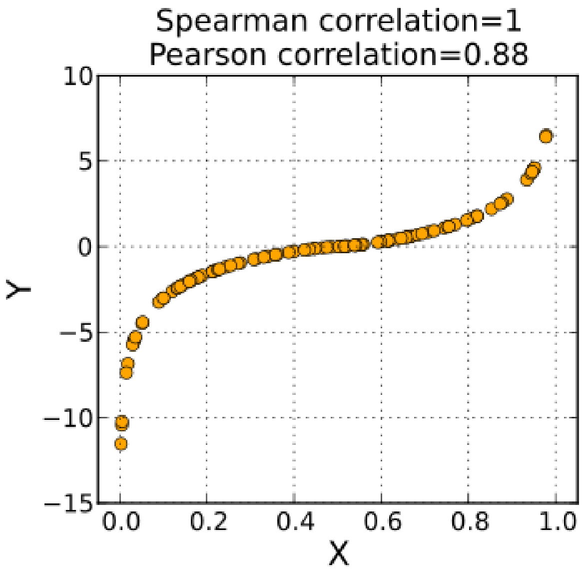
- From our example, ran 1000 permutations:



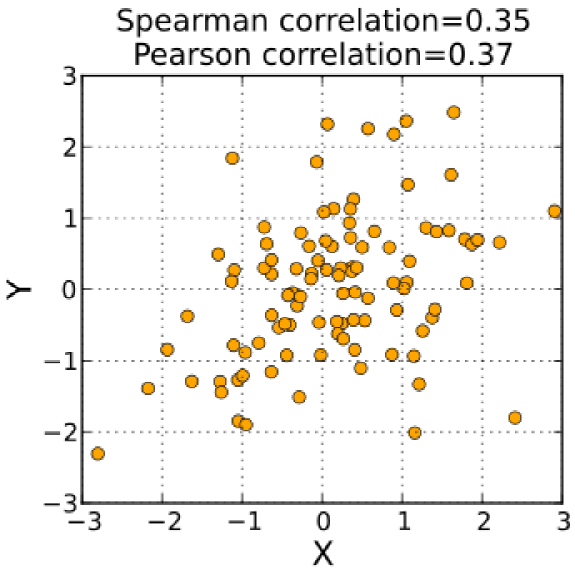
- 62% of permuted (i.e., H_0-enforced) rho values are more extreme than our observed ρ
  - Can't reject null hypothesis with sufficient confidence

Spearman vs. Pearson





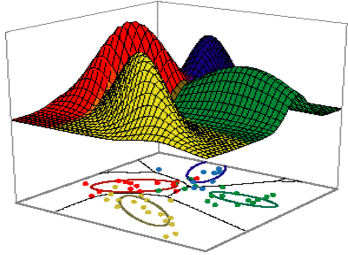
Monotonic, but nonlinear relationship



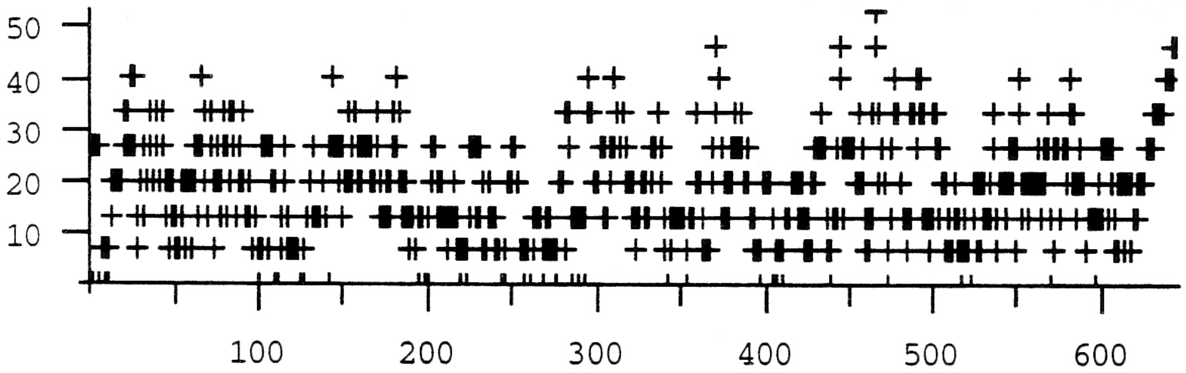
Weak relationship

https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient

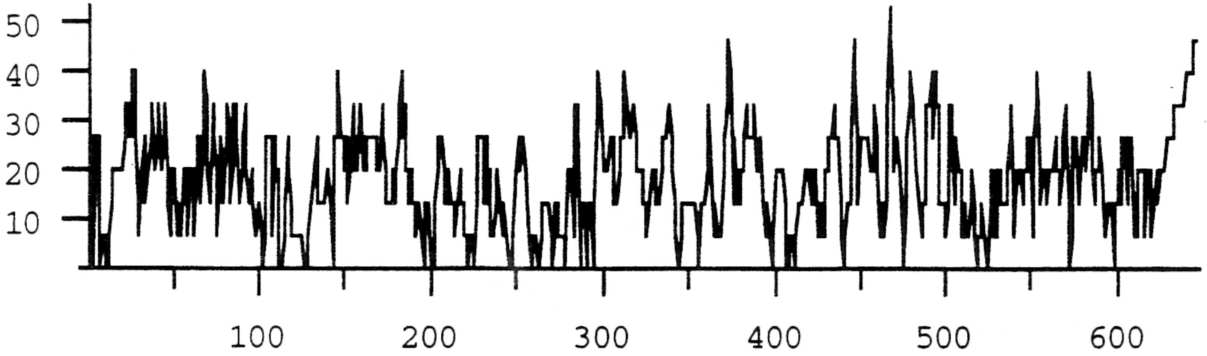
Time series analysis

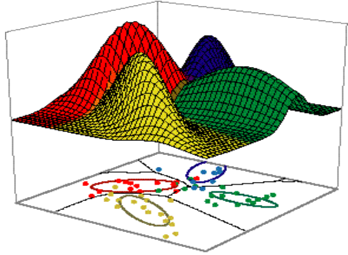


- Measurements may be time series
  - E.g. protein sequence data, behaviour of system over time
- Typically visualize using a line graph
  (instead of scatter plot)



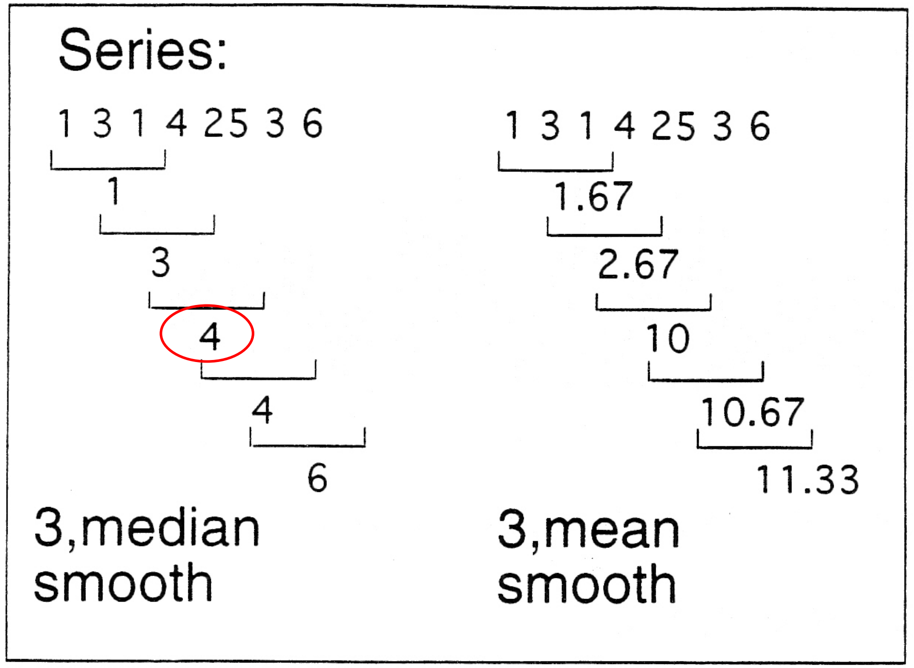
Ugly
Hard to read





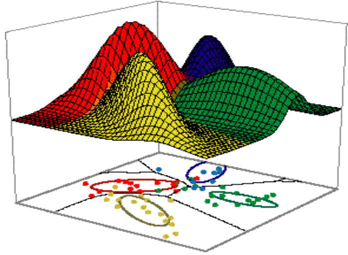
Time series analysis

- Smoothing:
  - Mean smoothing: apply sliding window, replace middle value with local mean (i.e. moving average filter)
  - Median smoothing: apply sliding window, replace center value with median



- Can have very complex smoothing filters such as:
  1) Repeatedly smooth with a 3,median filter until no change is observed
  2) Smooth with a 2-mean filter once
  3) Apply a Hanning operation (convolve with [0.25 0.5 0.25])

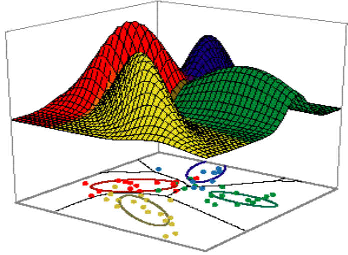
Time series analysis



- Correlation between signals
  - Compute cross-correlation between two signals:

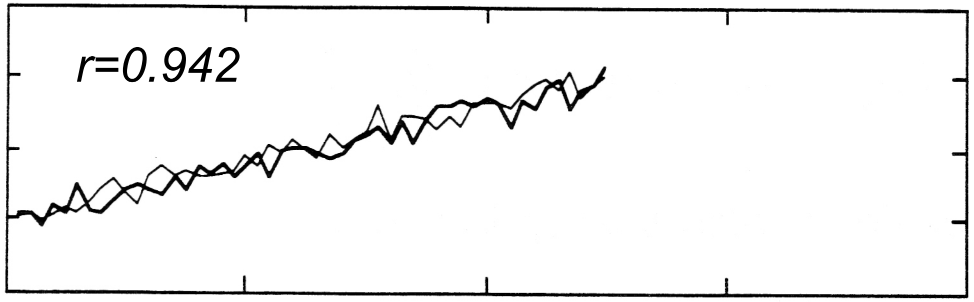
$$\Phi_{xy}(j) = \sum_n x(n)y(n-j)$$

  - Measures cross-correlation between x and y shifted by j
- Auto-correlation: measures relationships between a signal and itself at given lag.
  - E.g. weather tomorrow strongly correlated with today, less so with last week…

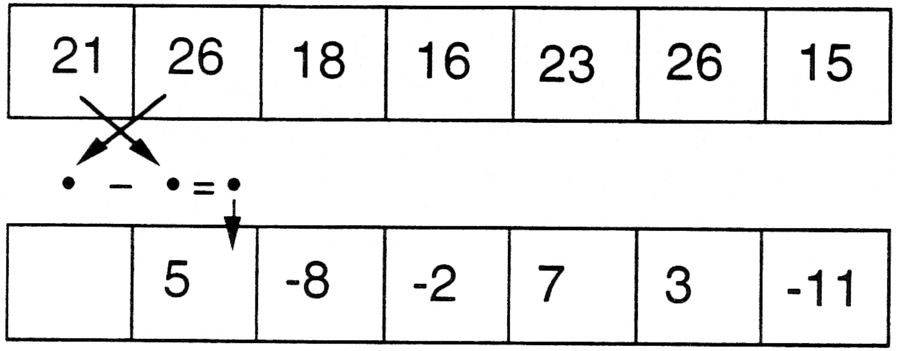


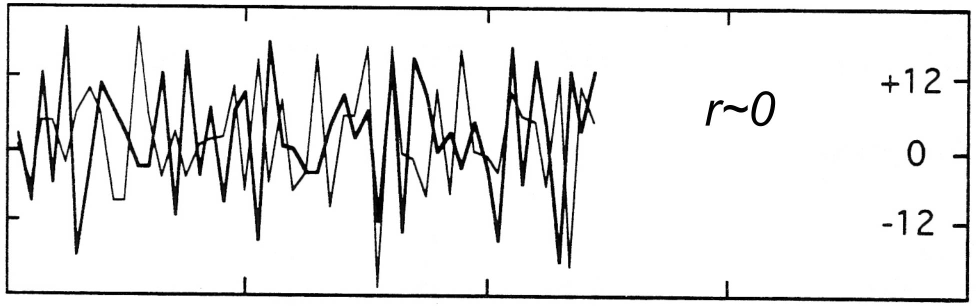
Time series analysis

- Dereferencing:
  - May want to replace signal with first difference to remove unwanted trend before computing cross-correlation

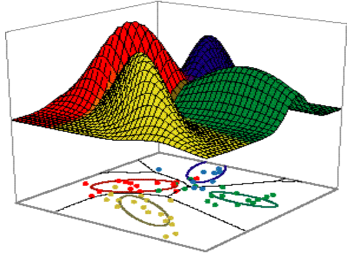


Remove trend by differencing:



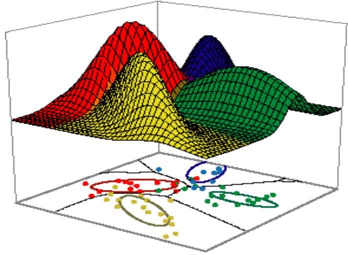


Data pre-processing - Outliers



- Outlier detection:
  - Many techniques / rules of thumb
  - Qualitative rule: “Outliers are farther from main mass of distribution than those points are to each other.”
  - Quantitative rule: Define outlier as any point which is greater than 3 standard deviations from mean
    - Recall 99.7% of data within 3σ of μ for normal…
    - Many other rules are available…

Data pre-processing - Outliers

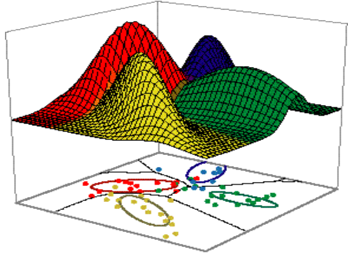


- Dealing with outliers:
  - Given sequence[1]:

    4 7 9 3 4 11 12 1304 10 15 12 13 17
  - Can see smooth curve from ~5 to ~15.
    - 1304 should not unduly affect curve.
    - Best approach is generate smooth curve with a large residual.
  - However, 1304 is in the data for <u>some reason</u>!
  - Final representation of reality:
    - We have a smooth curve from 5-15
    - Make a memo to ourselves to figure out what caused the extreme value.
  - *"You don't have to look at all the data all the time."*

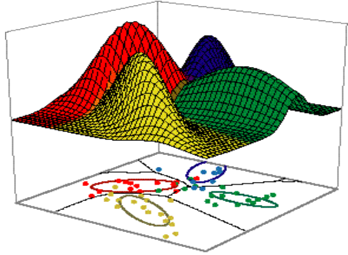
[1] Adapted from Tukey, John. 1977. <u>Exploratory Data Analysis</u>. Addison-Wesley.

Data pre-processing - Outliers



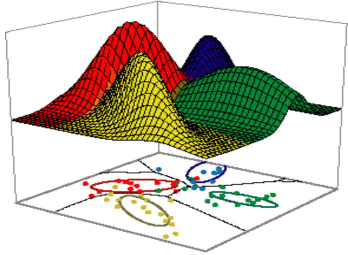
- Victoria J Hodge and Jim Austin. A Survey of Outlier Detection Methodologies. *Artificial Intelligence Review*, 22(2):85-126, 2004.
  - Categorized outlier detection techniques into three major groups:
    - Unsupervised outlier detection techniques:
      - There is no prior knowledge about the data under analysis.
    - Supervised outlier detection techniques:
      - There are samples of both normal and abnormal data.
    - Semi-supervised outlier detection techniques:
      - There are only samples of the normal data and no samples of abnormal ones.

Missing Features During Training



1. Exclude training sample with missing data
2. Impute with ‘typical’ value for that feature
   - Ignores what we do know about this sample… (other features that do have values; including class??)
3. Impute value from “similar” record (“closest fit”)
   - Uses known features to identify similar record (globally or only within same class); use value from that record
   - But many of the features may be irrelevant to the missing feature… Can focus on “similar/related” features (attribute clustering)
4. Create a “reduced-feature” model
   - (see next-next slide)
5. Use alternate model or “view”
   - See ‘surrogate node’ in Decision Tree slides
   - Use Attribute clustering to identify features that behave similarly
6. Pay to have the missing data collected

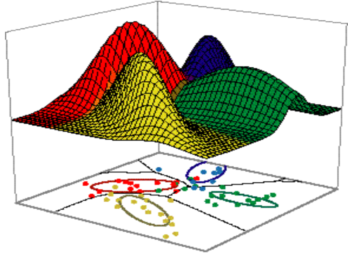
Missing Features During Operation



1. Exclude test sample with missing data
   - Refuse to classify this sample
2. Impute with ‘typical’ value for that feature
   - See counter-example on next slide
3. Impute value from “similar” record (“closest fit”)
4. Use a “reduced-feature” model
   - Pre-train and store? Compute online? Combinatorial…
5. Use alternate model
   - Use a model that does not rely on missing feature
6. Pay to have the missing data collected

More details: Saar-Tsechansky, M. and Provost, F. Handling missing values when applying classification models. *Journal of machine learning research*, 2007.

Missing Features



- Assume missing feature x_1, but measure x_2
- If replace with mean of x_1 → predict ω_3
- Instead, look at marginal (integrate out x_1)

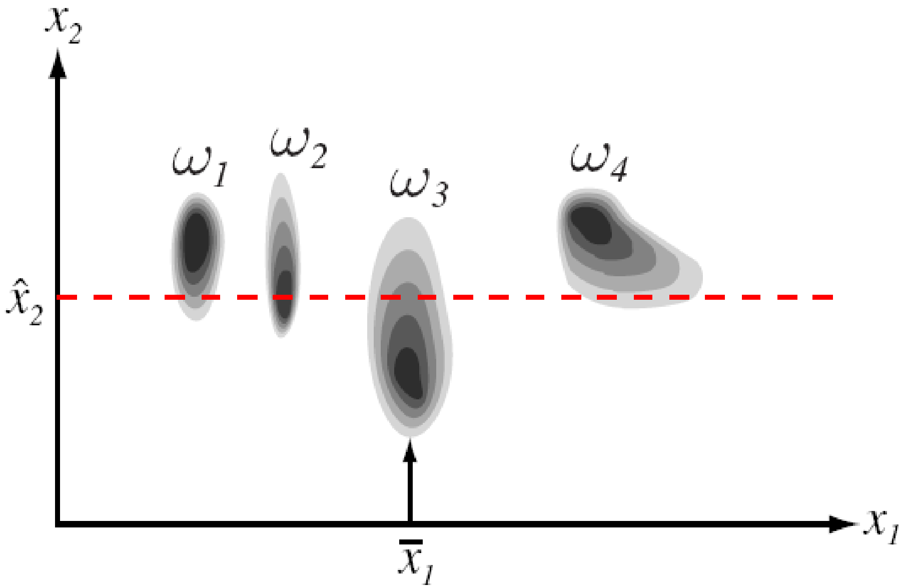
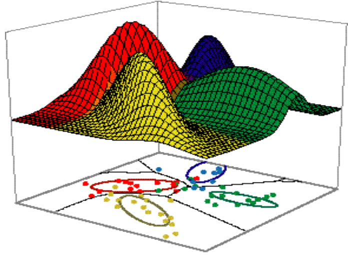


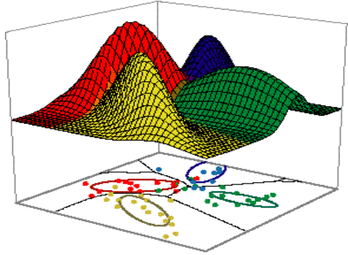
FIGURE 2.22. Four categories have equal priors and the class-conditional distributions shown. If a test point is presented in which one feature is missing (here, x_1) and the other is measured to have value $\hat{x}_2$ (red dashed line), we want our classifier to classify the pattern as category ω_2, because $p(\hat{x}_2|\omega_2)$ is the largest of the four likelihoods. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*.

Steps in Pattern Classification



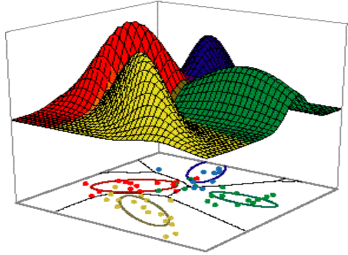
- Data pre-processing
- Selecting a learning algorithm *(throughout course)*
- Feature Selection / Representation
- Data set partitioning *(Part 4)*
- Training *(throughout course)*
- Testing & reporting results *(Part 4)*
- Meta-learning / CME *(later)*

Selecting a learning algorithm *(throughout course)*



- No Free Lunch Theorem.
  - Averaged over all possible problems, no classifier type provides universal advantage
  - E.g. select loaded dice to give advantage for a game.
    - Guaranteed to perform worse on another game.
  - "…no pattern classification method is inherently superior to any other, or even to random guessing; it is the type of problem, prior distribution, and other information which determines which classifier will give the best performance." Duda, Hart, Stork p454
  - "Even popular and theoretically grounded algorithms will perform poorly on some problems, ones in which the learning algorithm and the posterior happen not to be matched" Duda et al, p458
  - If a classifier outperforms another on a particular problem, it is <u>due to the fit between the method & the problem</u>, not the superiority of the algorithm.
  - Focus on prior information, data distribution, amount of training data, and cost function.

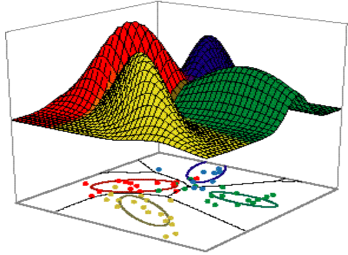
Classifier Complexity



- Minimum Description Length Principle
  - Represent classifier algorithm as a string of bits for execution in a general (Turing) computer
  - What is minimum number of bits required to describe the classifier?
  - Tied closely to 'complexity' of classifier
    - e.g. number of nodes in an ANN, number of branches in a decision tree, etc.
- VC Dimension* (Vapnik-Chervonenkis)
  - Measures the maximum number of points that can be completely 'shattered' by a decision boundary resulting from a 2-class classifier
    - Shattered in the sense of points from class A & B are isolated from each other by decision boundaries for every possible class assignment (to the points) and arbitrary point placement
      - (general position: ignore pathological cases such as 3 co-linear points etc)
    - Single linear discriminant vs. highly complex K-NN decision boundary
  - VC Dimension has been calculated for several standard pattern classification methods (DTs, SVN, ANN, etc)

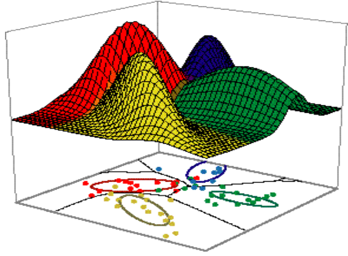
* *Andrew Moore, "VC Dimension Tutorial",* https://autonlab.org/assets/tutorials/vcdim08.pdf

Selecting a learning algorithm *(throughout course)*

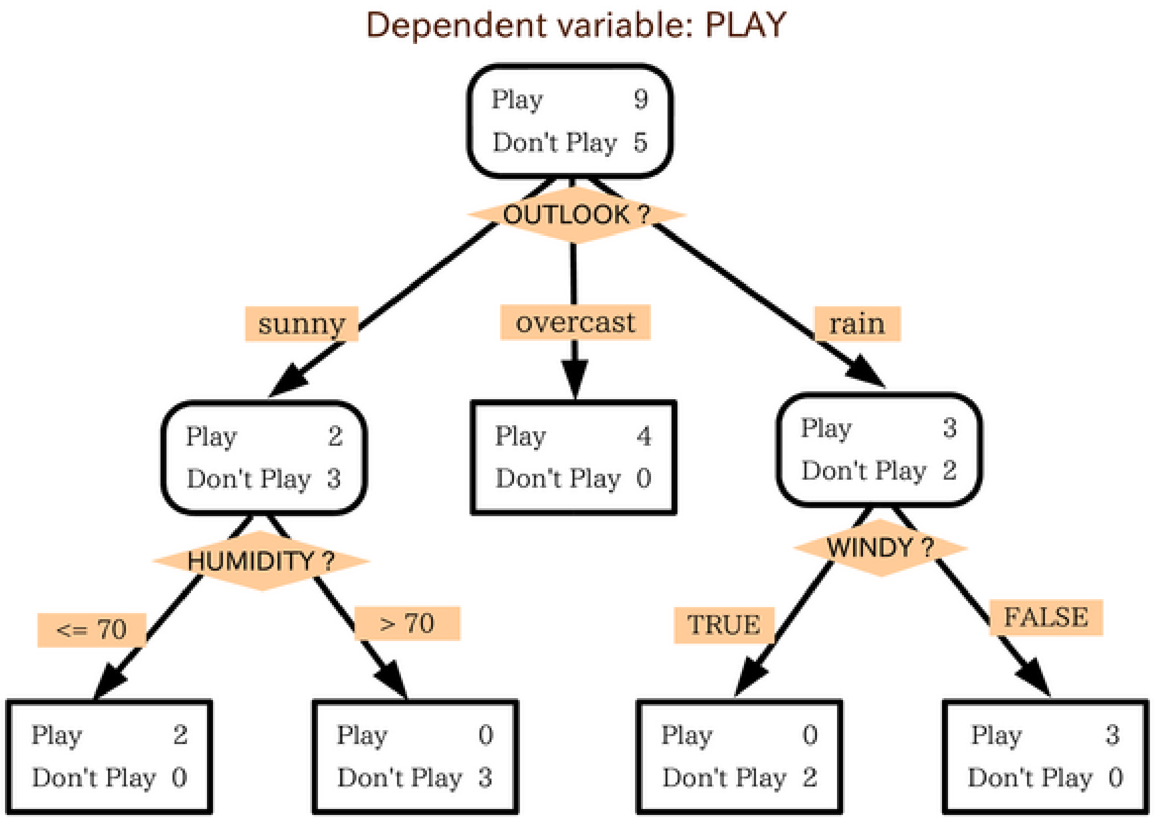


- Other considerations
  - Data types of features (representation) may suggest classifier
    - Categorical/ordinal may be better suited to a decision tree than a neural network
  - HW implementation issues
    - Trained ANN may actually result in a simple decision boundary that is easier to implement.
  - ‘White box’ vs. ‘Black box’
    - White box approaches permit easy interpretation of inner workings
    - Black box approaches may appear to work ‘by magic’
    - Will your client be satisfied with a black box approach?
      - Confidence in predictions when you cannot explain decision process directly in terms of the raw data?
      - E.g. medical informatics

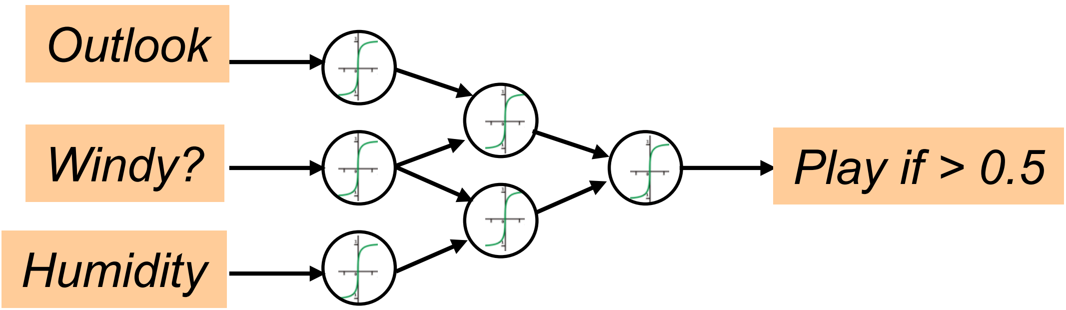
White box vs. black box

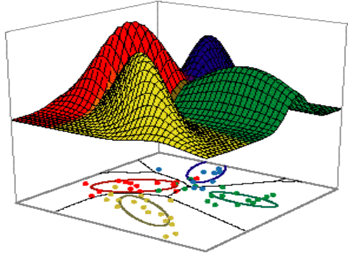


White box



Black box

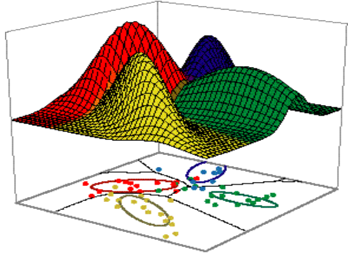




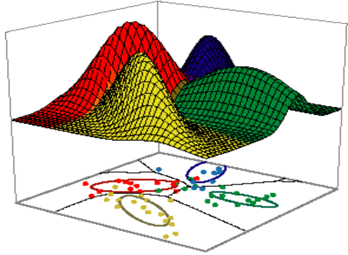
Feature Selection / Representation

- Features are actually functions that map individuals to data scales. Create representations from individuals.
  - Height(Jim)=75”
- There is no ‘natural’ representation of data
  - E.g. Paint colours could be represented using:
    - Categorical: {‘red’, ‘blue’, ‘yellow’, …}
    - Interval: luminocity measurements
    - Ratio: count photons
  - Seek a representation of reality while avoiding deluding ourselves

Feature Selection / Representation



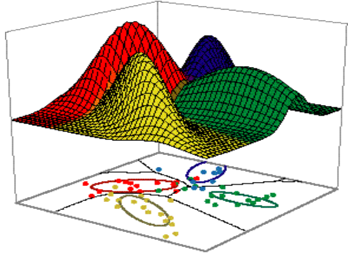
- "Ugly Duckling Theory": analogous to "No Free Lunch"
  - No problem-independent "best" set of features or feature attributes.
  - In the absence of assumptions, there is no reason to prefer any representation.
  - Example:
    - Let f_1='blind_in_right_eye' and f_2='blind_in_left_eye'
    - Define similarity as number of shared features
    - $x_1=[1,0]$ is maximally different from $x_2=[0,1]$
    - Does that make sense? For some problems yes, for some no.
    - Is f_1='blind_in_right_eye' and f_2='same_in_both_eyes' better?
- Depends on prior information about the problem.



Feature Selection / Representation

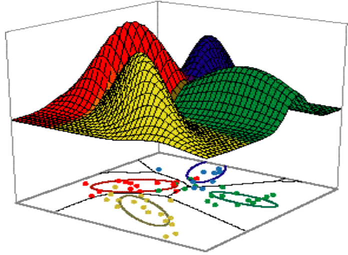
- Adding features increases the complexity of the classifier
  - Increases chance of overfitting training data and failing to generalize to new data
- Adding features which are independent of each other can only improve accuracy
  - *But, see warning above…*

Problems of dimensionality



- How does classifier accuracy depend on the dimensionality of feature set?
  - Secondary: how does this impact computational complexity of the classifier?
- The good news:
  - More features may mean increased accuracy
- The bad news:
  - The curse of dimensionality

Accuracy, dimension and training sample size



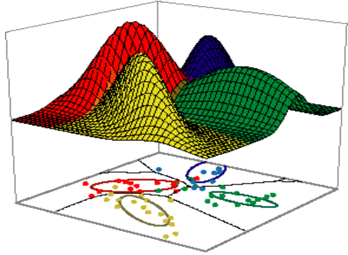
- If features are independent, can show that more features lead to decreased error
  - e.g. assume 2-class multivariate normal with same covariance: $p(\mathbf{x}|\omega_i) \sim N(\mu_i, \Sigma)$, i=1,2
  - For equal prior probabilities, Bayes error is:

$$P(e) = \frac{1}{\sqrt{2\pi}} \int_{r/2}^{\infty} e^{\frac{-u^2}{2}} du$$

We will show this next week…

- Here, *r* is Mahalanobis distance: $r^2 = (\mu_1 - \mu_2)^t \Sigma^{-1} (\mu_1 - \mu_2)$
  - Therefore, error decreases as *r* increases
  - For conditionally independent features: $\Sigma = diag(\sigma_1^2, \ldots, \sigma_d^2)$

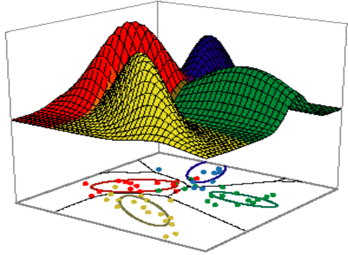
$$r^2 = \sum_{i=1}^{d} \left(\frac{\mu_{i1} - \mu_{i2}}{\sigma_i} \right)^2$$



Accuracy, dimension and training sample size

- We have: $$r^2 = \sum_{i=1}^{d} \left(\frac{\mu_{i1} - \mu_{i2}}{\sigma_i} \right)^2$$
- Therefore, as *d* increases, *r* increases, and the Bayes error decreases.
  - Benefit of each feature depends on:
    1. difference between class means in that dimension, and
    2. variance in that dimension.
- So more features should decrease error!
  - (assuming conditional independence)
  - (assuming probabilistic structure of problem known)
  - Limitless decrease in Bayes error… *(impossible)*

Accuracy, dimension and training sample size



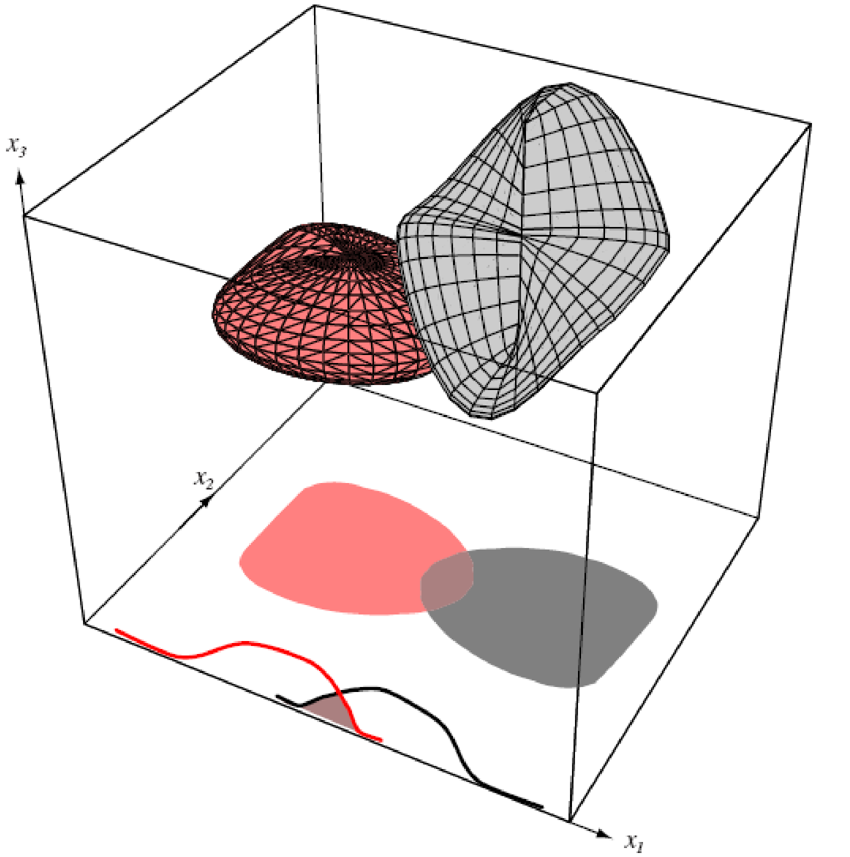
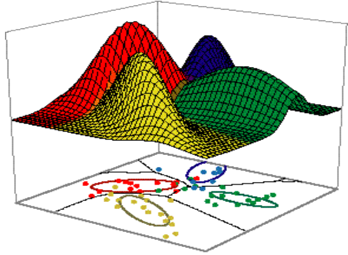


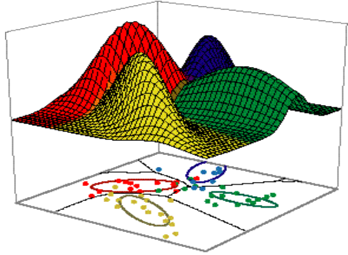
FIGURE 3.3. Two three-dimensional distributions have nonoverlapping densities, and thus in three dimensions the Bayes error vanishes. When projected to a subspace—here, the two-dimensional $x_1 - x_2$ subspace or a one-dimensional x_1 subspace—there can be greater overlap of the projected distributions, and hence greater Bayes error. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Accuracy, dimension and training sample size



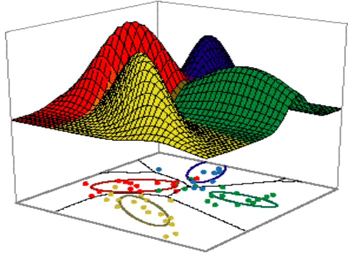
- Have shown that adding features should decrease error…
  - (assuming conditional independence)
  - (assuming probabilistic structure of problem known)
  - Limitless decrease in Bayes error… *(impossible)*
- Why don’t we actually see a continuous decrease in error?
  - Recall other sources of error:
    - estimation error due to limited training data (gets worse with higher *d*)
    - model error (from violating 2 assumptions above)

Curse of dimensionality



- Demand for large number of training samples often grows exponentially with dimensionality of feature space
  - High dimension (discriminant) functions have the potential to be much more complex than low dimensional functions.
  - Therefore require more data to fix more parameters.
  - Can reduce this problem through application of valid prior knowledge/assumptions about the problem
    - e.g. assume form of distribution

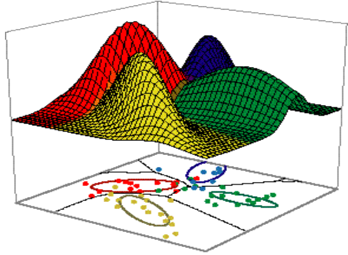
Feature selection



- "Good feature subsets contain features highly correlated with the class, yet uncorrelated with each other" Hall (1999)
- Some methods will actually do worse with more features
  - May be overly sensitive to noisy features
  - May overweight redundant features
- Use *feature selection* to mitigate these effects
  - Choose a subset of features based on merit
- Some classifiers that use a *complexity fit* implicitly incorporate feature selection into training
  - e.g. limiting the depth of a decision tree limits the number of features used in the final decision.
  - e.g. weight elimination in ANN limits the number of inputs permitted to impact the final result
  - e.g. regularize parameter estimation problem by pseudo-Bayesian estimation (weighted between prior and observed)

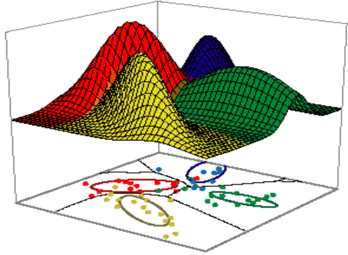
*Sholom Weiss and Casmir Kulikowski, Computer Systems That Learn, Morgan Kaufmann, 1991.

Reducing dimensionality



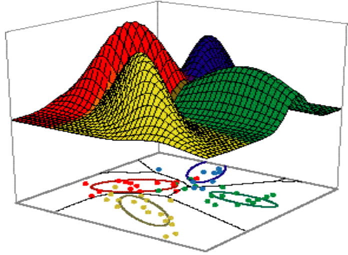
- Several options for reducing dimensionality
  - Manually select subset of features based on merit
    - Can pre-screen individual features for ability to discriminate between classes (recall $r^2 = \sum_{i=1}^{d} \left(\frac{\mu_{i1} - \mu_{i2}}{\sigma_i} \right)^2$, or information gain)
    - Can pre-screen subsets of features (pairs, etc.)
    - Cluster similar/redundant features based on covariance
  - Automated dimension reduction
    - Compute linear combination of features, then choose subset
      - Principal Component Analysis *(unsupervised)*
      - Fisher's Linear Discriminant *(supervised)*
      - Multiple Discriminant Analysis
    - Attribute clustering (e.g. using mutual information)

Reducing dimensionality



- Principal Component Analysis
  - Choose a line that best <u>represents</u> data in least-square sense
  - Choose $d' < d$ eigenvectors with largest eigenvalues
- Fisher's Linear Discriminant
  - Project onto hyperplane (line for 2D) which best <u>discriminates</u> between classes
- Multiple Discriminant Analysis
  - Multi-class extension of Fisher's linear discriminant
  - Project from d-dimensional space to $c-1$ dim. space

Reducing dimensionality



- Fisher's linear discriminant example

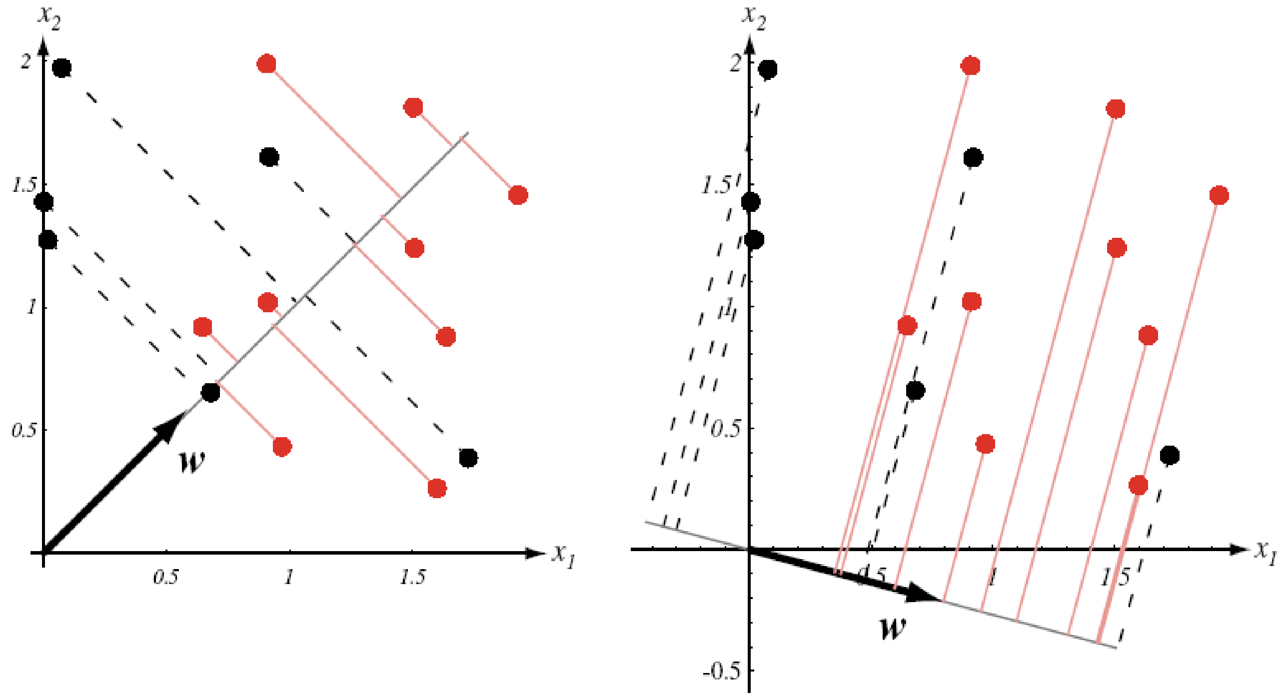
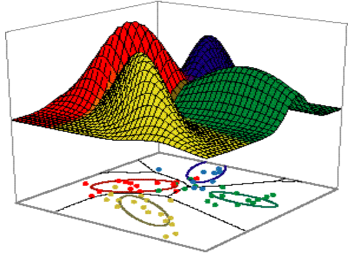


FIGURE 3.5. Projection of the same set of samples onto two different lines in the directions marked **w**. The figure on the right shows greater separation between the red and black projected points. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Reducing dimensionality



- Multiple discriminant analysis example

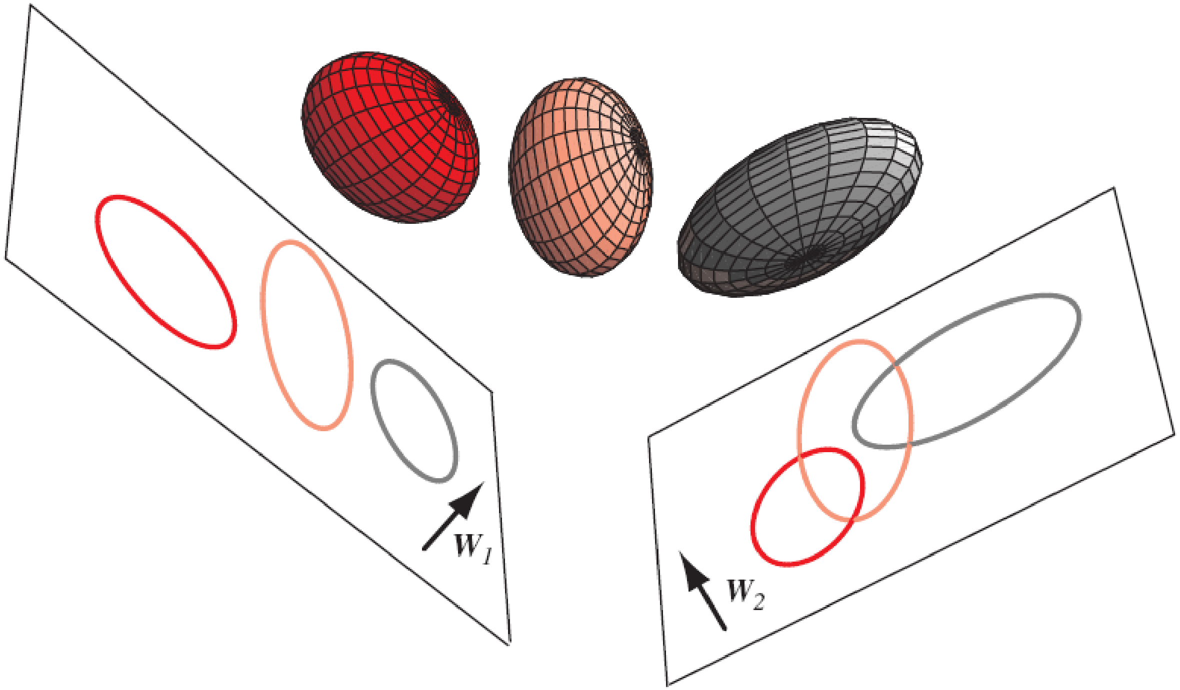
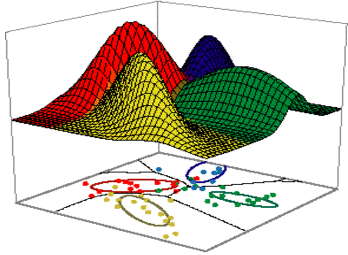


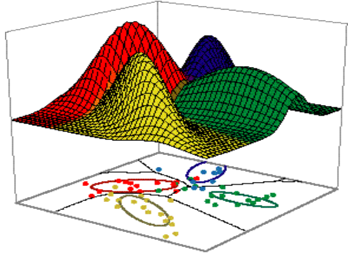
FIGURE 3.6. Three three-dimensional distributions are projected onto two-dimensional subspaces, described by a normal vectors $\mathbf{W}_1$ and $\mathbf{W}_2$. Informally, multiple discriminant methods seek the optimum such subspace, that is, the one with the greatest separation of the projected distributions for a given total within-scatter matrix, here as associated with $\mathbf{W}_1$. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Two methods of feature selection



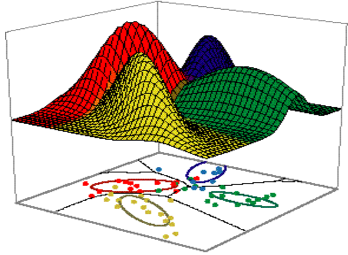
- 1) Filter method
  - Select features based on their distribution between the two classes
  - A) supervised (know class of each point)
    - Seek features which are tightly distributed within each class and show a difference between classes
  - B) unsupervised (don’t know class of each point)
    - Looking for uncorrelated sets of features
    - Look for features with high variance
      - (*A feature whose values are identical for all samples is not useful/informative*)

Two methods of feature selection



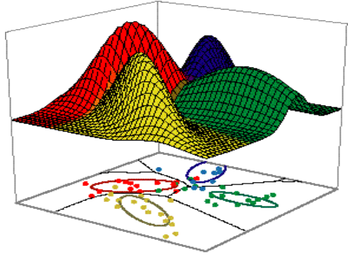
- 1) Filter method
- 2) Wrapper method
  - Select features based on the resulting accuracy of a classifier trained using those features
  - Iteratively select features, train classifier, test classifier, go back and adjust features
    - Can use forward selection, genetic algorithms, etc.
  - *Potential for over-fitting??*

Data set partitioning



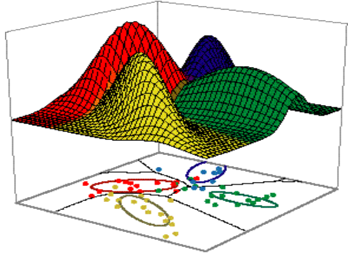
- Goal of pattern classification is to learn from training data in order to perform accurately over new future data (generalization)
- We also need to estimate true error rate
  - Therefore, need independent training and test data
  - Validation data needed when iteratively optimizing hyperparameters.
- Most problems have limited samples
  - Must decide how many to use for training, validation, and testing.
  - Must balance 2 goals:
    - Need sufficient training data to learn from
    - Need sufficient test data to accurately predict performance over future data
- Underlying assumption that all training points were drawn i.i.d. from some distribution.
  - Assume that future test points will be drawn from the same distribution

Data set partitioning



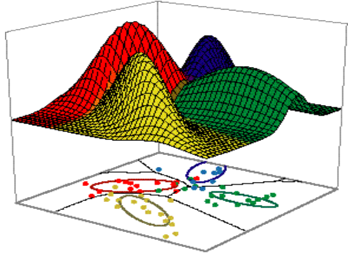
- Several strategies to maximize use of data
  - Hold-out
  - N-fold cross-validation
  - Leave-one-out / jackknife
- Each has ramifications for calculating expected error over new data.
- Will discuss advantages of each in Part 4.

Data set partitioning



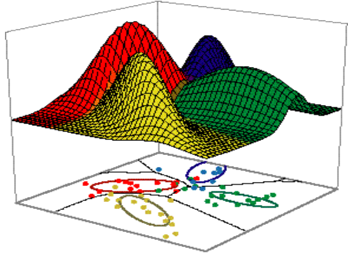
- **Class imbalance**
  - Occurs when one class is far more prevalent than the other class(es)
- Problem:
  - Classifiers tend to always predict overrepresented class and ignore rare class
    - E.g., if 90% of data is from class 1, then always choosing class 1 leads to 90% accuracy!
- Solution:
  - Random undersampling (of overrepresented class)
  - Random oversampling (of underrepresented class)
    - Can also create new synthetic data, e.g. by adding noise to existing data
  - Add synthetic samples to the minority class (e.g., SMOTE)
  - Adjusting cost/loss function (more later)
    - Make errors on rare class more costly / increase penalty for these errors

Training *(throughout course)*



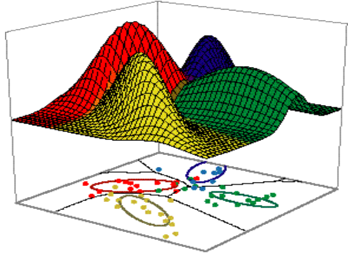
- Each form of classifier has a different approach to training.
  - Parametric classifiers assume structure of distribution of class data, and try to estimate parameters of the distribution. The decision boundary is a product of the estimated distributions.
  - Non-parametric classifiers attempt to define the decision boundary directly.
  - Some classifier structures have multiple training algorithms available (e.g. neural networks)
- What themes are common to all classifiers?
  - Many training algorithms have <u>architectural parameters</u> (hyperparameters) to set before training of model can begin
    - Parameter sweeping
    - Require separate dataset
  - Many training algorithms have a <u>stopping criteria</u> or a way to go back and <u>prune</u> the classifier in order to limit/reduce its complexity
    - Promote generalization
  - Do not aim for perfect classification on the training data if <u>Bayes error</u> (theoretical best possible error rate) is non-zero

Testing & reporting results *(Part 4)*



- How do we accurately measure and report the accuracy of a pattern classifier?
- How do we objectively compare two classifiers over a given problem?
- How can we predict how well a classifier will generalize, given its performance over our training data / testing data?

Meta-learning / CME *(later)*



- Can we resample the training data to squeeze more performance out of our classifier?
- Can we combine multiple copies of the same classifier trained slightly differently to achieve more accuracy?
- Can we combine multiple heterogeneous classifier to improve accuracy?