# Unit Schedule: Modules

| Module | Week | Content | Ross |
|---|---|---|---|
| **1.** | 1 | Introduction to modelling for data science and to R | 1,2 |
| **2.** | 2 | Probabilities and bias | 3 |
| | 3 | **Expectations** | 4 |
| | 4 | Distributions | 5 |
| **3.** | 5 | Statistical inference | 6&7 |
| | 6 | Hypothesis testing | 7&8 |
| **4.** | 7 | Dependence and linear regression | 9 |
| | 8 | classification and clustering | |
| **5.** | 9 | Comparing means | 10 |
| | 10 | Random number generation and simulation | |
| **6.** | 11 | Validation and complexity | 15 |
| | 12 | Modelling | |

FIT5197 Statistical Data Modelling

# Module 2

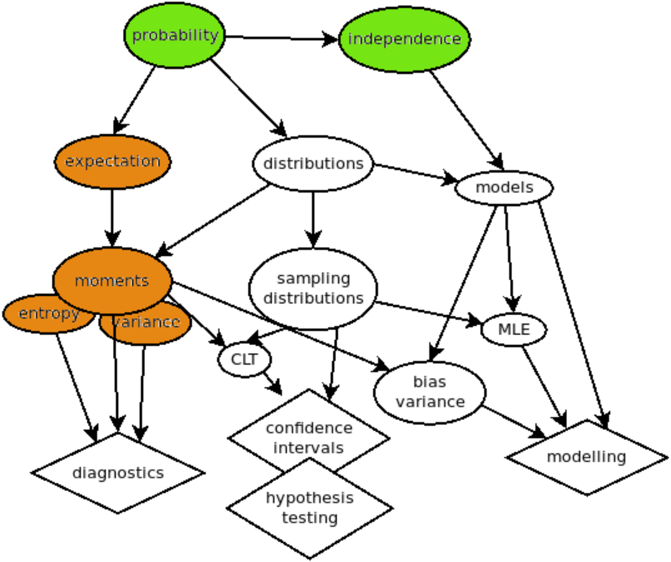
# Expectations and Other Measures

2020 Lecture 3

Monash University

Revision at *https://flux.qa/43FMK4*

# Concept Map for This Unit



# Expected Values (ePub sections 2.2, 2.5, Ross 4.4-4.7, 4.9)

# Outline


Measuring Things in Average

Expected Values

Entropy and Coding

Dependence

Chebyshev's Inequality

Weak Law of Large Numbers

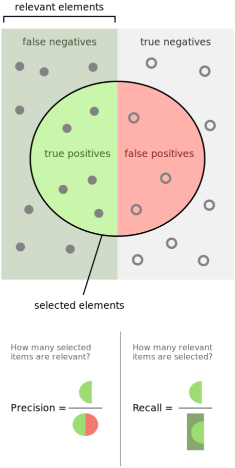
# Measuring Things in Average

# Cancer Decisions

Suppose your GP runs a cheap test on you which returns positive for bowel cancer. She recommends you visit a specialist for a 2nd diagnosis and possible surgery to remove the section of bowel. The second diagnosis consists of the specialist doing a biopsy followed by running a test on the sample. The biopsy has a considerable cost and is not fully reliable. The surgery has a greater cost (both expense and the loss of body part), and a chance of failure (cancer still exists).

- what properties would you like of the GP's test to improve your situation?
- what properties would you like of the specialist's test to improve your situation?
- suppose the specialist's test is positive, and you agree to surgery, what properties would you like of surgery outcomes?

# Evaluating Binary Predictions




**sensitivity**, **recall**, **hit rate**, or **true positive rate** (TPR)

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{P}} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} = 1 - \mathrm{FNR}$$

**specificity**, **selectivity** or **true negative rate** (TNR)

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{N}} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} = 1 - \mathrm{FPR}$$

**precision** or **positive predictive value** (PPV)

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} = 1 - \mathrm{FDR}$$

**accuracy** (ACC)

$$\mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{P}+\mathrm{N}} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}$$

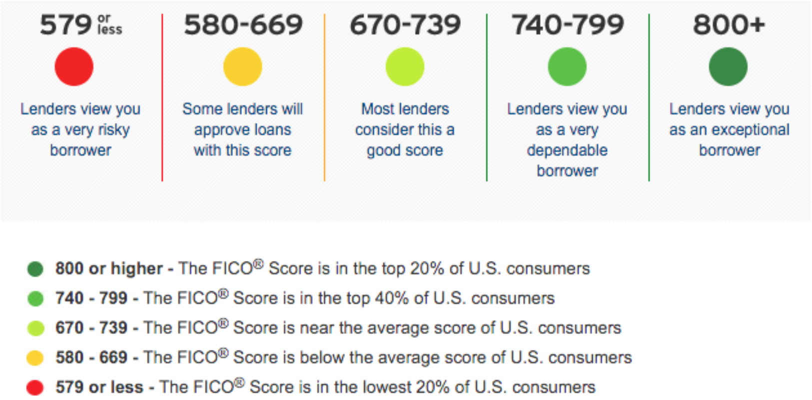
see *precision and recall* on Wikipedia

# Cancer Decisions, cont.

- what properties would you like of the GP's test to improve your situation?
  - high recall (mostly)
  - good precision (probably not possible)
- what properties would you like of the specialist's test to improve your situation?
  - high accuracy
  - not too expensive
- suppose the specialist's test is positive, and you agree to surgery, what properties would you like of surgery outcomes?
  - low chance of failure
  - lower expense and life cost

# FICO Scores

Fair Isaac Corporation produces *credit scores*.



# Computing FICO Scores

**Sample FICO® Scoring Model**
Example: Partial Model

| Category | Characteristic | Attributes | Points |
|---|---|---|---|
| Payment History | Number of months since the most recent derogatory public record | No public record | 75 |
| | | 0 – 5 | 10 |
| | | 6 – 11 | 15 |
| | | 12 – 23 | 25 |
| | | 24+ | 55 |
| Outstanding Debt | Average balance on revolving trades | No revolving trades | 30 |
| | | 0 | 55 |
| | | 1 – 99 | 65 |
| | | 100 – 499 | 50 |
| | | 500 – 749 | 40 |
| | | 750 – 999 | 25 |
| | | 1000 or more | 15 |
| Credit History Length | Number of months in file | Below 12 | 12 |
| | | 12 – 23 | 35 |
| | | 24 – 47 | 60 |
| | | 48 or more | 75 |
| Pursuit of New Credit | Number of inquiries in last 6 mos. | 0 | 70 |
| | | 1 | 60 |
| | | 2 | 45 |
| | | 3 | 25 |
| | | 4+ | 20 |
| Credit Mix | Number of bankcard trade lines | 0 | 15 |
| | | 1 | 25 |
| | | 2 | 50 |
| | | 3 | 60 |
| | | 4+ | 50 |

14 

# FICO Scores and Probabilities

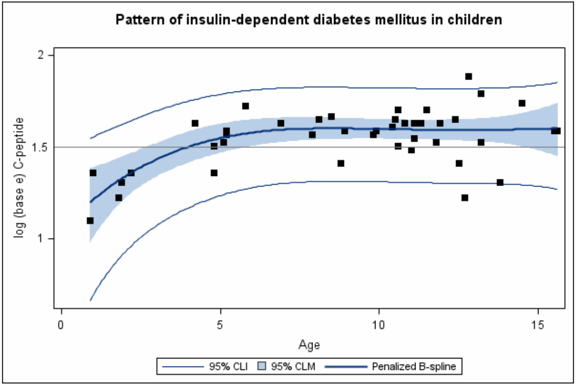
- we expect the FICO score is *calibrated* with probability of not defaulting
  - the lower the score for a consumer, the higher the probability of defaulting on a loan
  - this is probability as frequency: for a given score at a given time, there is a true but unknown probability of default
  - it is also affected by the kind of loan

# FICO Scores and Probabilities

- we expect the FICO score is *calibrated* with probability of not defaulting
  - the lower the score for a consumer, the higher the probability of defaulting on a loan
  - this is probability as frequency: for a given score at a given time, there is a true but unknown probability of default
  - it is also affected by the kind of loan
- bank managers adjust the “acceptance” FICO score to suit their financial targets
  - more loans? decrease acceptance score!

# Predicting Real Values

- SAS *prediction with error bars*
- don't just want prediction, may also want confidence bands or upper/lower limits



# Making Decisions

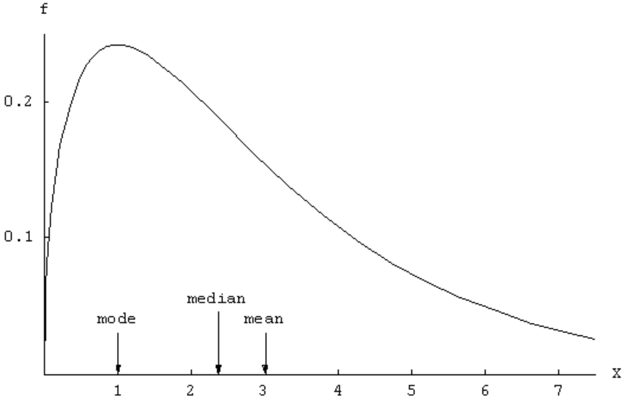
- **should** understand costs of various outcomes
- **need** to estimate recall, precision, or other measures of quality for categorical/binary decisions
- **or** may use a calibrated score for cruder control
- **need** to estimate means and ranges for real valued predictions

# Making Decisions

- **should** understand costs of various outcomes
- **need** to estimate recall, precision, or other measures of quality for categorical/binary decisions
- **or** may use a calibrated score for cruder control
- **need** to estimate means and ranges for real valued predictions

theoretical tool for estimation is the **expected value**

# Example: Central Tendency




- in a skewed distribution, the mode, median and mean do not line up!
- long tail on right and hump on left means skewed to the right

# Characterising Distributions

**central tendency:** where abouts is the distribution mainly located? what is its centre?

**deviation:** how much does it vary? what is the rough spread of the distribution?

**skew:** is it anti-symmetric? does it have a long tail in some direction?

**NB.** suppose you don't have modern graphics devices, just some tables of numbers and want to measure the above!

# Characterising Distributions

central tendency: where abouts is the distribution mainly located? what is its centre?

deviation: how much does it vary? what is the rough spread of the distribution?

skew: is it anti-symmetric? does it have a long tail in some direction?

**NB.** suppose you don't have modern graphics devices, just some tables of numbers and want to measure the above!

theoretical tool for characterisation is the **expected value**

# The Most Important Formula

The fundamental result for understanding modern algorithms.

$$MSE(\mathcal{H}) = \overline{bias}(\mathcal{H})^2 + \frac{1}{|\mathcal{H}|}\overline{variance}(\mathcal{H}) + \left(1 - \frac{1}{|\mathcal{H}|}\right)\overline{covariance}(\mathcal{H})$$

This is developed by Uedo and Nakano 1996. **NB.** bias, variance and covariance are all defined as expected values

You need to understand this to understand modern machine learning algorithms.

# Outline


Measuring Things in Average

Expected Values

Entropy and Coding

Dependence

Chebyshev's Inequality

Weak Law of Large Numbers

# Expected Values

# Expected Values

- Given a distribution, we can define the expected value of the RV:

$$\mathbb{E}\left[X\right] = \sum_{x \in \mathcal{X}} x\, p(x)$$

recalling that $p(x) \equiv p(X = x)$.

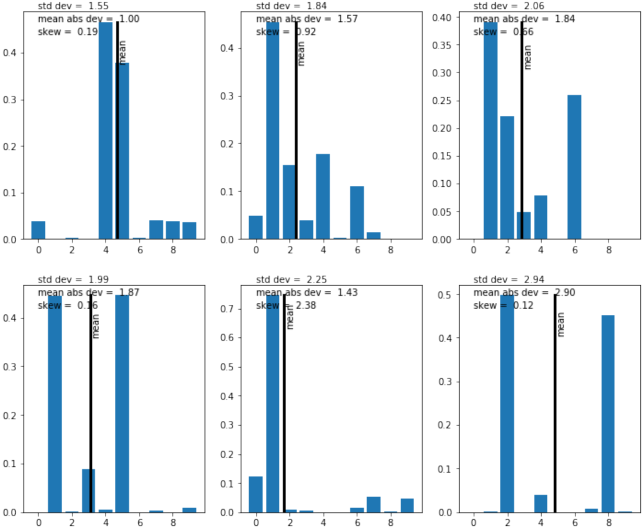
- The expected value is the average value over $\mathcal{X}$, weighted by the probability of each particular $x \in \mathcal{X}$ appearing.
- For continuous RVs, replace the sum with an integral:

$$\mathbb{E}\left[X\right] = \int x\, p(x) dx$$

# Expected Values

- Given a distribution, we can define the expected value of the RV:
$$\mathbb{E}\left[X\right] = \sum_{x \in \mathcal{X}} x\, p(x)$$
recalling that $p(x) \equiv p(X = x)$.
- The expected value is the average value over $\mathcal{X}$, weighted by the probability of each particular $x \in \mathcal{X}$ appearing.
- For continuous RVs, replace the sum with an integral:
$$\mathbb{E}\left[X\right] = \int x\, p(x) dx$$
- Example:
$p(X = 1) = 0.5,\ p(X = 2) = 0.4,\ p(X = 3) = 0.1$:
$$\mathbb{E}\left[X\right] = 1 \cdot 0.5 + 2 \cdot 0.4 + 3 \cdot 0.1 = 1.6$$

# Distributional Properties




mean = $\mathbb{E}[X]$

std-dev = $\sqrt{\mathbb{E}[(X - mean)^2]}$

mean-abs-dev = $\mathbb{E}[|X - mean|]$

skew = $\dfrac{\mathbb{E}\left[(X - mean)^3\right]}{(std\text{-}dev)^3}$

# Expected Values, cont.

- More generally:

$$\mathbb{E}\left[f(X)\right] = \sum_{x \in \mathcal{X}} f(x)p(x)$$

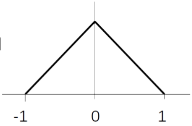
  where $f(x)$ is any function of $x$.
- Also for $n$ a non-negative integer,
  - $\mathbb{E}\left[X^n\right]$ is called the $n$-th moment
  - $\mathbb{E}\left[(X - \mathbb{E}\left[X\right])^n\right]$ is called the $n$-th central moment
  - see also *skewness* and *kurtosis*
- Example:
  $p(X = 1) = 0.5$, $p(X = 2) = 0.4$, $p(X = 3) = 0.1$:

$$\begin{aligned}
\mathbb{E}\left[\ln X\right] &= 0.5 \cdot \ln 1 + 0.4 \cdot \ln 2 + 0.1 \cdot \ln 3 = 0.3871 \\
\mathbb{E}\left[\ln(1/p(X))\right] &= 0.5 \cdot \ln 1/0.5 + 0.4 \cdot \ln 1/0.4 + 0.1 \cdot \ln 1/0.1 \\
&= 0.5 \cdot 0.693 + 0.4 \cdot 0.916 + 0.1 \cdot 2.303 = 0.943
\end{aligned}$$

  where $\ln x$ is the natural logarithm.

# Expected Values, e.g.

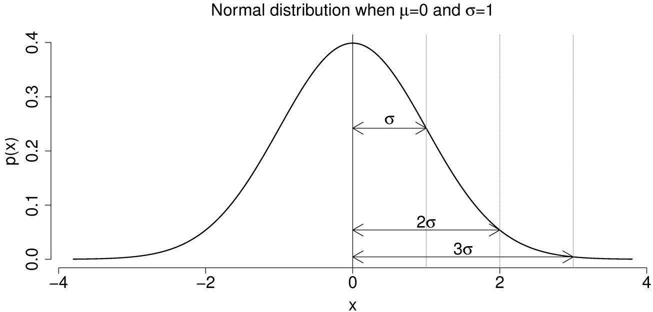
Example: consider the simple triangular distribution $p(x) = 1 - |x|$ for $x \in [-1, 1]$:




$$
\begin{aligned}
\mathbb{E}[x] &= \int_{-1}^{0} x(1+x)\mathrm{d}\,x + \int_{0}^{1} x(1-x)\mathrm{d}\,x \\
&= \int_{1}^{0} x'(1-x')\mathrm{d}\,x' + \int_{0}^{1} x(1-x)\mathrm{d}\,x = 0 \\
\mathbb{E}\left[x^2\right] &= \int_{-1}^{0} x^2(1+x)\mathrm{d}\,x + \int_{0}^{1} x^2(1-x)\mathrm{d}\,x \\
&= -\int_{1}^{0} x'^2(1-x')\mathrm{d}\,x' + \int_{0}^{1} x^2(1-x)\mathrm{d}\,x \\
&= 2\int_{0}^{1} x^2(1-x)\mathrm{d}\,x = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}
\end{aligned}
$$

Similarly, $\mathbb{E}\left[x^{2n+1}\right] = 0$ and $\mathbb{E}\left[x^{2n}\right] = \frac{2}{(2n+1)(2n+2)}$.

# Expected Values, Gaussian




- normal (Gaussian) distribution has PDF

$$p(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{-(x-\mu)^2/2\sigma^2}$$

- expected values not simple to do!

# Statistical Dispersion

- Expected values let us define important properties such as the variance:

$$\begin{aligned}\mathbb{V}[X] &= \mathbb{E}\left[(X-\mathbb{E}[X])^2\right] \\ &= \sum_{x\in\mathcal{X}}(x-\mathbb{E}[X])^2 p(x)\end{aligned}$$

$\Longrightarrow$ the expected squared deviation around the mean

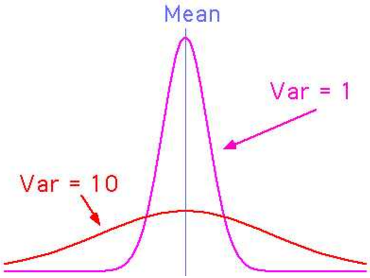
- The larger $\mathbb{V}[X]$ the more variation around the mean
- The standard deviation is equal to $\sqrt{\mathbb{V}[X]}$.
- Alternatively the mean absolute deviation, $\mathbb{E}[|X-\mathbb{E}[X]|]$, is less often used because it is harder to determine analytically.
  $\Longrightarrow$ the expected absolute deviation around the mean

# Variance Examples

- Example:
  $p(X = 1) = 0.5$, $p(X = 2) = 0.4$, $p(X = 3) = 0.1$; recall that in this case, $\mathbb{E}[X] = 1.6$, so:

  $$\mathbb{V}[X] = (1-1.6)^2 \cdot 0.5 + (2-1.6)^2 \cdot 0.4 + (3-1.6)^2 \cdot 0.1 = 0.44$$

- Example: Gaussian



# Variance, cont.

- A useful alternative expression for variance is:

$$
\begin{aligned}
\mathbb{V}[X] &= \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \\
&= \mathbb{E}\left[X^2 - 2X\mathbb{E}[X] + \mathbb{E}[X]^2\right] \\
&= \mathbb{E}\left[X^2\right] - 2\mathbb{E}[X]\,\mathbb{E}[X] + \mathbb{E}[X]^2 \\
&= \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2
\end{aligned}
$$

  where the third step follows from properties of sums/integrals
- Variance is sum of expected squared value of $X$, minus square of expected value of $X$
  $\Longrightarrow$ Use this to find variance for our example on previous slide

# Expectations and Independent RVs

- In general, expectation of a function of two RVs is

$$\mathbb{E}\left[f(X, Y)\right] = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} f(x, y) p(x, y)$$

- Fact 1: Due to linearity of expectation, we have

$$\mathbb{E}\left[f(X) + g(Y)\right] = \mathbb{E}\left[f(X)\right] + \mathbb{E}\left[g(Y)\right]$$

for all RVs $X$ and $Y$, and

# Expectations and Independent RVs

- In general, expectation of a function of two RVs is

$$\mathbb{E}\left[f(X, Y)\right] = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} f(x, y) p(x, y)$$

- Fact 1: Due to linearity of expectation, we have

$$\mathbb{E}\left[f(X) + g(Y)\right] = \mathbb{E}\left[f(X)\right] + \mathbb{E}\left[g(Y)\right]$$

  for all RVs $X$ and $Y$, and
- Fact 2: For independent RVs, we have

$$\mathbb{E}\left[f(X)g(Y)\right] = \mathbb{E}\left[f(X)\right] \mathbb{E}\left[g(Y)\right]$$

  implying that

$$\mathbb{V}\left[X + Y\right] = \mathbb{V}\left[X\right] + \mathbb{V}\left[Y\right]$$

  for $X$ and $Y$ independent.

# Existence of Expected Values

- Expected values do not always exist
- If $\mathcal{X}$ is *finite*, then $\mathbb{E}[X]$ always exists
- However, in general, $\mathcal{X}$ will not be finite
- $\mathcal{X}$ is usually the set of integers $\mathbb{Z}$ or real numbers $\mathbb{R}$
  $\Longrightarrow$ for these, expectations are not guaranteed to exist
- In contrast, the quantiles (such as median) *always* exist

# Outline

Measuring Things in Average

Expected Values

Entropy and Coding

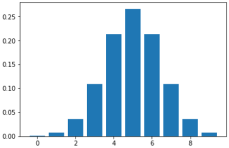
Dependence

Chebyshev's Inequality

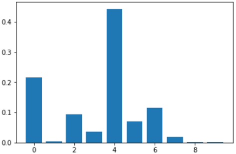
Weak Law of Large Numbers

# Entropy and Coding

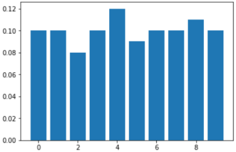
# Variance for Discretes




(A)




(B)

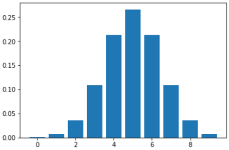



(C)

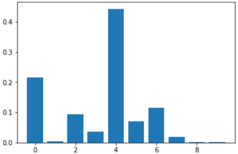
- what would be a measure of variance for discretes?
- there is no value ordering, so variance measures are meaningless

  i,e,, rather then figure (A), we have (B)
- would like (C) to have low "variance"
- would like (B) to have higher "variance"

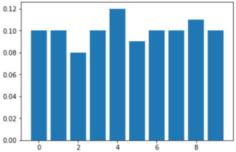
# Variance for Discretes, cont.




(A) size=6.15




(B) size=4.89




(C) size=9.95

- want a notion of "effective number of values"
- call it size for now
- illustrated for the plots above

  e.g. (C) almost uniform, so just less than 10
- am using size=$2^{H(\vec{p})}$
  - $H()$ is the entropy function computed to base 2
  - $\vec{p}$ is the probability vector

# Entropy for Probability Vectors

**Definition:** *Entropy* is a function $H(\vec{p})$ on prob. vectors, $\vec{p}$

$$H(\vec{p}) = \sum_{i=1}^{K} p_i \log_2 \left( \frac{1}{p_i} \right)$$

where $K$ is the dimension of $\vec{p}$.

- using log to base 2 so that entropy of Bernoulli(0.5) distribution is 1
- $\lim_{p \to 0} p \log_2 \left( \frac{1}{p} \right) = 0$, so well defined when $p_i = 0$
- measured in units of **bits**
- uniform distribution, for $\vec{p} = (1/K, ..., 1/K)$, $H(\vec{p}) = \log_2 K$
- if $H(\vec{p}) = 0$ then $p_i = 1$ for some $i$

# Entropy for Discretes

Alternatively, if $X$ is a discrete variable without loss of generality having outcomes $\mathcal{X} = \{1, 2, ..., K\}$, and $p(X{=}i) = p_i$ for $i = 1, ..., K$, then $H(X)$ is defined as $H(\vec{p})$.

- entropy defined as an expected value

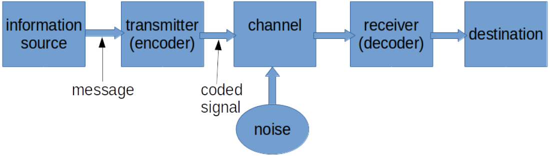
$$H(X) = \mathbb{E}\left[\log_2 1/p(X)\right]$$

- suppose further we use $p(X|Y{=}y)$, then the entropy is denoted $H(X|Y{=}y)$
- suppose $Y$ has outcomes in $\mathcal{Y}$, then define conditional entropy $H(X|Y)$ as

$$H(X|Y) = \sum_{y \in \mathcal{Y}} p(Y{=}y) H(X|Y{=}y)$$

- $X$ and $Y$ are independent if and only if $H(X|Y) = H(X)$

# Simple Communication Model

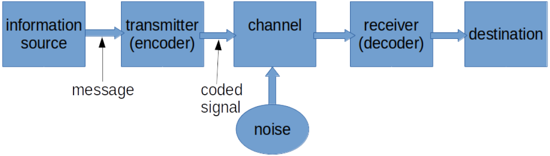



- a message is to be encoded into a binary signal (string of 0/1's) and sent across a channel to be decoded and so received
- e.g. noise-free communication

| | |
|---|---|
| message: | "hello" |
| encoding: | "001011101110101000010011111101010 ..." |
| transmitter: | written on a block of disk |
| receiving: | read Boolean string off the disk, without noise |
| decoding: | convert Boolean string back to "hello" |

- e.g. noisy communication, on reading the Boolean string might be corrupted!

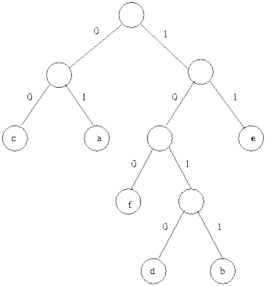
# Encoding and Decoding




- assuming a noise-free channel, how do you convert your message into a Boolean string so it can be read back OK?
- the encoding and decoding must be prearranged so they match up
- what properties would you like of your codes:
  - short messages?
  - unambiguous decoding?

**NB.** the encoder-decoder framework is a dominant paradigm in unsupervised neural networks and natural language translation

# Encoding Binary Codes



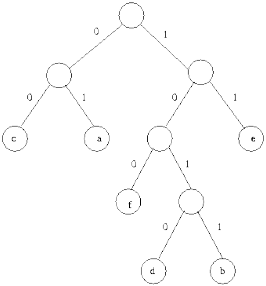

the tree defines a binary code for letters

| | | | | | |
|---|---|---|---|---|---|
| 'a' '01' | 'b' '1011' | 'c' '00' | 'd' '1010' | 'e' '11' | 'f' '100' |

encode "fab" $\longrightarrow$

- reading path through tree to leaves gives "100", "01", "1011";
- but there are no spaces in our binary strings, so must mash together "100011011"

# Decoding Binary Codes



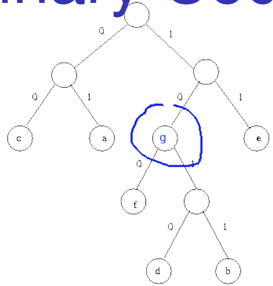

the tree defines a binary code for letters

| ‘a’ | ‘b’ | ‘c’ | ‘d’ | ‘e’ | ‘f’ |
|---|---|---|---|---|---|
| ‘01’ | ‘1011’ | ‘00’ | ‘1010’ | ‘11’ | ‘100’ |

decode “011011001010” $\longrightarrow$

- trace through the tree, match prefix “01”: “a1011001010”
- next, match prefix “1011”: “ab001010”
- next, match prefix “00”: “abc1010”
- get “abcd”

# Binary Codes



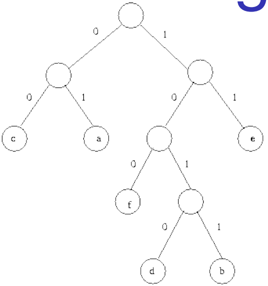

**suppose a new symbol “g” is given code ’10’ in this tree**

- when the receiver/decoder sees ’10’ in their message they have to decide
  - is it a “g” **or** is it the prefix of “f”, “d” or “b”
- this causes ambiguity about the symbol being received
- symbols can only be at the leaves of a tree to avoid ambiguity
- so no symbol’s code can be the prefix of any other symbol’s code

# Binary Prefix Codes

- a binary *prefix code* for the symbols assigns a binary string to every symbol such that no code is the prefix of another
- a prefix code guarantees we can recognise the end of the code when receiving a symbol
- every binary prefix code has a corresponding binary tree form with symbols at the leaves
- if some leaves are empty the code is inefficient and the code could be rearranged to eliminate the unused leaves

# Code Lengths

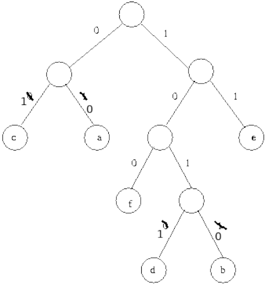



tree also gives code lengths for letters

| ‘a’ | ‘b’ | ‘c’ | ‘d’ | ‘e’ | ‘f’ |
|---|---|---|---|---|---|
| ‘01’ | ‘1011’ | ‘00’ | ‘1010’ | ‘11’ | ‘100’ |
| 2 | 4 | 2 | 4 | 2 | 3 |

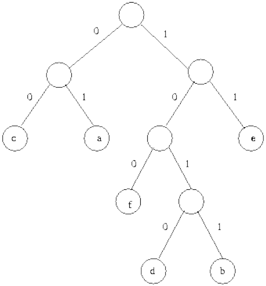
reordering tree has same code lengths

| ‘a’ | ‘b’ | ‘c’ | ‘d’ | ‘e’ | ‘f’ |
|---|---|---|---|---|---|
| ‘00’ | ‘1010’ | ‘01’ | ‘1011’ | ‘11’ | ‘100’ |
| 2 | 4 | 2 | 4 | 2 | 3 |




i.e., main properties of the tree are defined by the code lengths

# Code Lengths




tree also gives code lengths for letters

| ‘a’ ‘01’ | ‘b’ ‘1011’ | ‘c’ ‘00’ | ‘d’ ‘1010’ | ‘e’ ‘11’ | ‘f’ ‘100’ |
|---|---|---|---|---|---|
| 2 | 4 | 2 | 4 | 2 | 3 |

- can ask the question, “what is the average code length for a 100 letter message?”
  - but need a distribution over letters
- can ask the alternative question, “what code yields minimum average code length for a given distribution?”
  - we want to save on transmission costs!

# Kraft Inequality

- each symbol has a code length $l_1, ..., l_K$ in a binary tree

**Theorem: *Kraft Inequality* for Prefix Codes:**
Given code lengths $l_1, ..., l_K$, then $\sum_{k=1}^{K} 2^{-l_k} \leq 1$ if and only if there is a corresponding binary prefix code.

- so we can check if a code is a prefix code merely by evaluating $\sum_{k=1}^{K} 2^{-l_k}$ on the code lengths!

# Expected Code Length

- have $K$ symbols occurring with probability $p_1, ..., p_K$

**Definition:** the **expected code length** is defined as

$$\mathbb{E}\left[l_k\right] = \sum_{k=1}^{K} p_k l_k$$

- one interpretation of a "good" code is that it minimises expected code length for the probabilities of the symbols
- we want to do this to reduce transmission costs

e.g., find an encoding-decoding algorithm for typical blobs in a SQL database which is 1% more efficient, then Oracle saves $100M

# Codelengths and Probabilities

- given code lengths $\vec{l}$, the probabilities that minimise the expected code length is $p_k = 2^{-l_k}$
  - wrong direction! we want to find the codes given the probabilities

**Lemma:** Given probabilities, there is a binary prefix code with expected code length $\leq 1 + \sum_{k=1}^{K} p_k \log_2 1/p_k$.

Proof: plug $l_k = \lceil \log_2 1/p_k \rceil$ into Kraft inequality

- this gives us a heuristic way to build a code:
  1. make the code length for symbol $k$ be $l_k = \lceil \log_2 1/p_k \rceil$
  2. build a tree using this (assigning shortest symbols first)
  3. try to reduce inefficiencies
- lower average codelengths can be obtained by chunking symbols into groups before trying to build a code

# Coding and Entropy: Summary

- we wish to encode an item from a dictionary chosen with probability vector $\vec{p}$
- binary prefix codes are a way to encode without redundancy or confusion
- entropy $H(\vec{p})$ in bits is a lower bound on the average binary code length for any such codes
- a prefix code with average code length less than $1 + H(\vec{p})$ can always be built matching a probability $\vec{p}$
- the value $2^{H(\vec{p})}$ is a good measure for discrete (unordered) variables comparable to variance

# Outline


Measuring Things in Average

Expected Values

Entropy and Coding

Dependence

Chebyshev's Inequality

Weak Law of Large Numbers

# Dependence

# Covariance/Correlation

For two variables $X$ and $Y$ we can define the covariance:

$$\begin{aligned}\operatorname{cov}(X, Y) &= \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]\end{aligned}$$

and from this, we can define the correlation:

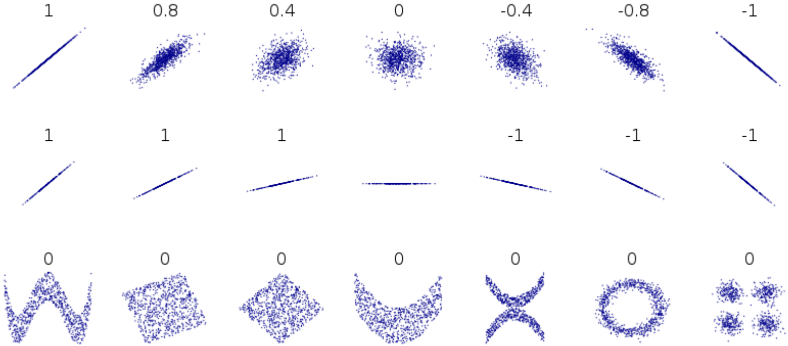
$$\operatorname{corr}(X, Y) = \frac{\operatorname{cov}(X, Y)}{\sqrt{\mathbb{V}[X]\,\mathbb{V}[Y]}}$$

$\Longrightarrow$ Compare to the sample correlation formula in Lecture 1.

Also, let $Z_X = \frac{X - \mathbb{E}[X]}{\sqrt{\mathbb{V}[X]}}$ and similarly for $Z_Y$, then

$$\operatorname{corr}(X, Y) = \operatorname{cov}(Z_X, Z_Y)$$

# Correlation, Examples




- strength of linearity
- positive or negative
- affected by outliers
- slope not relevant (due to standardising)

# Covariance/Correlation, cont.

- Positive covariance/correlation:
  $\Longrightarrow$ if $X > \mathbb{E}[X]$ then likely $Y$ is *greater* than $\mathbb{E}[Y]$
- Negative covariance/correlation:
  $\Longrightarrow$ if $X > \mathbb{E}[X]$ then likely $Y$ is *less* than $\mathbb{E}[Y]$

# Covariance/Correlation, cont.

- Positive covariance/correlation:
  $\Longrightarrow$ if $X > \mathbb{E}[X]$ then likely $Y$ is *greater* than $\mathbb{E}[Y]$
- Negative covariance/correlation:
  $\Longrightarrow$ if $X > \mathbb{E}[X]$ then likely $Y$ is *less* than $\mathbb{E}[Y]$
- Covariance between $(-\infty, \infty)$,
  - Depends on scale (unit of measurement) of variables $X$ and $Y$
- Correlation between $[-1, 1]$,
  - Independent of scale of variables
- If $X$, $Y$ independent, $\operatorname{cov}(X, Y) = \operatorname{corr}(X, Y) = 0$
  $\Longrightarrow$ Converse is **not** true!

# Why is $|\mathrm{corr}(X, Y)| < 1$? (optional)

- consider the centered vectors of data
  $\vec{u} = (x_1 - \overline{x}, ..., x_N - \overline{x})$ and $\vec{v} = (y_1 - \overline{y}, ..., y_N - \overline{y})$
- geometric reasoning says

$$|\vec{u}^T \vec{v}| = |\vec{u}||\vec{v}||\cos\theta| \leq |\vec{u}||\vec{v}|$$

  where $\theta$ is the angle between $\vec{u}$ and $\vec{v}$
- rearranging, we get

$$1 \geq \left(\frac{\vec{u}^T \vec{v}}{N}\right)^2 \frac{N}{\vec{u}^T \vec{u}} \frac{N}{\vec{v}^T \vec{v}}$$

- now let $N \to \infty$

$$1 \geq \mathrm{cov}(X, Y)^2 \frac{1}{\mathbb{V}[X]} \frac{1}{\mathbb{V}[Y]} = \mathrm{corr}(X, Y)^2$$

# Covar./Correl. Example

- Example: Probability distribution of $X$, $Y$:

| | $X=1$ | $X=2$ | $X=3$ |
|---|---|---|---|
| $Y=1$ | 0.05 | 0.15 | 0.1 |
| $Y=2$ | 0.25 | 0.15 | 0.3 |

- To find covariance, we need expected values (using sum rule):

$$\begin{aligned}
\mathbb{E}[X] &= p(X=1)\cdot 1 + p(X=2)\cdot 2 + p(X=3)\cdot 3 \\
&= (0.05+0.25)\cdot 1 + (0.15+0.15)\cdot 2 + (0.1+0.3)\cdot 3 \\
&= 2.1 \\
\mathbb{E}[Y] &= p(Y=1)\cdot 1 + p(Y=2)\cdot 2 \\
&= (0.05+0.15+0.1)\cdot 1 + (0.25+0.15+0.3)\cdot 2 \\
&= 1.7
\end{aligned}$$

# Covar./Correl. Example cont.

- Example: Probability distribution of $X$, $Y$:

| | $X = 1$ | $X = 2$ | $X = 3$ |
|---|---|---|---|
| $Y = 1$ | 0.05 | 0.15 | 0.1 |
| $Y = 2$ | 0.25 | 0.15 | 0.3 |

- Then $\operatorname{cov}(X, Y)$ is

$$\begin{aligned}&(1-1.7)(0.05(1-2.1)+0.15(2-2.1)+0.1(3-2.1))\\+&(2-1.7)(0.25(1-2.1)+0.15(2-2.1)+0.3(3-2.1))\\=&-0.0862\end{aligned}$$

- Challenge: see if you can calculate $\operatorname{corr}(X, Y)$.

# Outline

Measuring Things in Average

Expected Values

Entropy and Coding

Dependence

Chebyshev's Inequality

Weak Law of Large Numbers

# Chebyshev's Inequality

# Chebyshev's Inequality

**Theorem: Chebyshev's Inequality:**
If $X$ is a RV with mean $\mu$ and variance $\sigma^2$, then for any $k > 0$

$$p\left(\frac{|X-\mu|}{\sigma} \geq k\right) \leq \frac{1}{k^2}$$

- At least $(1 - \frac{1}{k^2})$ $x100\%$ of the data lies within $k$ standard devations of the mean.
- Named after P. Chebyshev (1821-1894)
- This inequality allows us to compute (bounds on) probabilities even when only the mean and variance are known

# Chebyshev's Inequality, cont.

- Chebyshev's bound if only $\mathbb{E}[X] = 0$, $\mathbb{V}[X] = 1$ is known:
  - $p(|X| \geq 1) \leq 1$;
  - $p(|X| \geq 2) \leq 0.25$;
  - $p(|X| \geq 3) \leq 0.1112$;

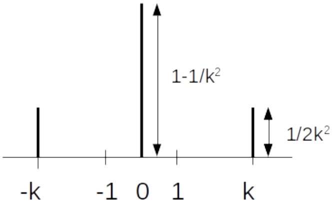
- Compare to the situation for a standard normal distribution, that we know $X \sim N(0, 1)$:
  - $p(|X| \geq 1) = 0.3173$;
  - $p(|X| \geq 2) = 0.0455$;
  - $p(|X| \geq 3) = 0.0027$.

$\Longrightarrow$ Chebyshev's bounds very general but not always accurate.

# Chebyshev's Inequality, Worst Case

What distribution is worst case for $k$: $p\left(\frac{|X-\mu|}{\sigma} \geq k\right) = \frac{1}{k^2}$?

- all the probability is at 3 discrete points, *i.e.* $x_i = \mu$ or $x_i = \mu \pm k\sigma$
- shown below worst case for $\mu = 0$, $\sigma = 1$



# Chebyshev's Inequality Proof
(optional)

If $X$ is a RV with mean $\mu$ and variance $\sigma^2$. Then

$$\begin{aligned}
\sigma^2 &= \int (x-\mu)^2 p(x) \mathrm{d}\, x \\
&\geq \int_{|x-\mu|\geq k\sigma} (x-\mu)^2 p(x) \mathrm{d}\, x \\
&\geq \int_{|x-\mu|\geq k\sigma} (k\sigma)^2 p(x) \mathrm{d}\, x \\
&= (k\sigma)^2 p\left(\frac{|X-\mu|}{\sigma} \geq k\right)
\end{aligned}$$

Note, from the steps we can also argue that the worst case distribution given previously is unique.

# Chebyshev's for Samples

Replace our distribution by the induced sample distribution (i.e., each point in the sample is equally likely).

**Theorem:**
For a sample $S = \{x_1, ..., x_N\}$ of variable $X$ with mean $\overline{x}$ and sample standard deviation $s_x$, then for any $k > 0$

$$\left|\left\{x_i \,:\, \frac{|x_i - \overline{x}|}{s_x} \geq k\right\}\right| \leq \frac{N}{k^2}$$

That is, the number of data points at least $k\, s_x$ from the mean is no more than $\frac{N}{k^2}$.

- Allows us to compute (bounds on) properties of the sample with only knowledge of the sample mean and standard deviation.

# Outline

Measuring Things in Average

Expected Values

Entropy and Coding

Dependence

Chebyshev's Inequality

Weak Law of Large Numbers

# Weak Law of Large Numbers

# Weak Law of Large Numbers

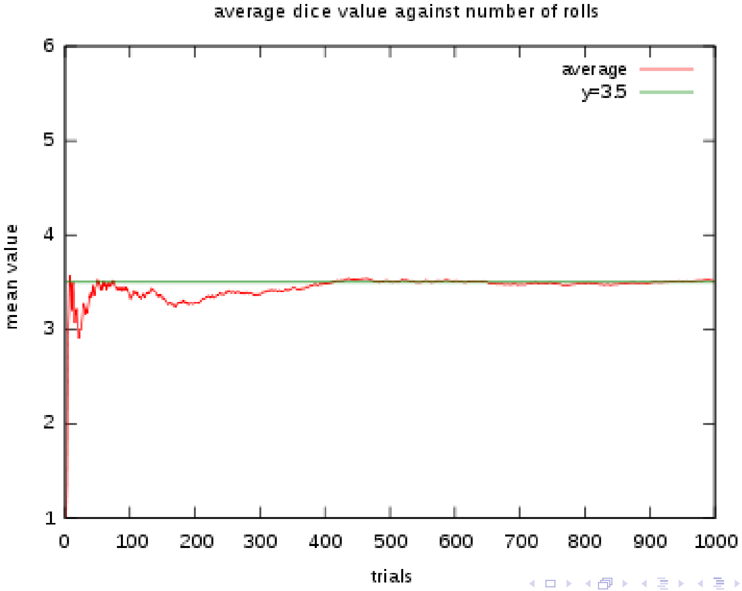
An important application of Chebyshev's inequality is to prove the weak law of large numbers.

**Theorem: Weak law of large numbers:**
Let $X_1, \ldots, X_n$ be RVs with $\mathbb{E}[X_i] = \mu$; then for any $\varepsilon > 0$

$$p\left(\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| > \varepsilon\right) \to 0 \text{ as } n \to \infty.$$

- Informally, you can think of this result as saying that the mean of a sample of random variables converges to the expected value as the sample size grows larger.

# Law of Large Numbers, e.g.



# Summary

# Revision of Probability

- $p(X = x, Y = y)$ is joint probability of $X = x$ and $Y = y$.
  - Sum-rule (marginal probability):

$$p(X = x) = \sum_{y} p(X = x, Y = y)$$

  - Conditional probability

$$p(X = x \mid Y = y) = \frac{p(X = x, Y = y)}{p(Y = y)}$$

- Cumulative distribution function (for ordered $x$):

$$p(X \leq x) = \sum_{x \leq x} p(X = x)$$

- Also: $p(X > x) = 1 - p(X \leq x)$.

# Revision of Expected Value

- Let $p(X = x) \equiv p(x)$; expectation and variance of $f(X)$:

$$\begin{aligned}\mathbb{E}[f(X)] &= \sum_x p(x) f(x) \\ \mathbb{V}[f(X)] &= \mathbb{E}\left[(X - \mathbb{E}[f(X)])^2\right]\end{aligned}$$

  with integral replacing sum for continuous RVs.

- Some useful rules:
  - $\mathbb{E}[f(X) + g(Y)] = \mathbb{E}[f(X)] + \mathbb{E}[g(Y)]$
  - $\mathbb{E}[cf(X)] = c\mathbb{E}[f(X)]$
  - $\mathbb{V}[cf(X)] = c^2\mathbb{V}[f(X)]$

- If $X$, $Y$ are independent RVs
  - $\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)]\,\mathbb{E}[g(Y)]$
  - $\mathbb{V}[f(X) + g(Y)] = \mathbb{V}[f(X)] + \mathbb{V}[g(Y)]$

# Revision of Entropy

- define
$$H(X) = \mathbb{E}\left[\log_2 1/p(X)\right]$$
- if X has domain $\mathcal{X}$ of dimension $K$, then $0 \leq H(X) \leq K$
- if $H(X) = 0$ then $p(X{=}x) = 1$ for some $x \in \mathcal{X}$
- entropy can be justified as the lower bound in bits of a binary prefix code to encode a realisation of $X$
- for $X$, $Y$
$$H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$$
- if $X$, $Y$ are independent RVs
$$H(X, Y) = H(X) + H(Y)$$

End of Week 3