Review

- Counting & permutations
- Random events
  - Axioms, conditional probability, marginalisation, Bayes, independence
- Random variables
  - Definition, Bernoulli & Binomial RVs, indicator RVs
  - Mean & variance, correlation & conditional expectation
- Inequalities
  - Markov, Chebyshev, Chernoff
- Sample mean, weak law of large numbers
- Continuous random variables, Normal distribution, CLT
- Statistical modelling: logistic regression & linear regression

Inequalities

- Markov's Inequality. For X a non-negative random variable:

$$P(X \geq a) \leq \frac{E(X)}{a} \text{ for all } a > 0$$

- Chebyshev's Inequality. For X a random variable with mean $E(X) = \mu$ and variance $var(X) = \sigma^2$:

$$P(|X - \mu| \geq k) \leq \frac{\sigma^2}{k^2} \text{ for all } k > 0$$

- Chernoff's Inequality. For X a random variable:

$$P(X \geq a) \leq \min_{t>0} e^{-ta+\log E(e^{tX})}$$

(this is the basis for large deviations theory[1])

[1] https://en.wikipedia.org/wiki/Large_deviations_theory

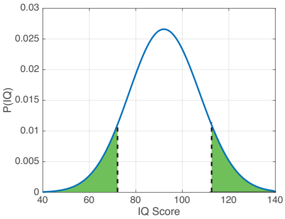
Example: Markov

An elevator can carry a load of at most 1000Kg. The average weight of a person is 80Kg. Suppose 10 people are in the elevator, use Markov's inequality to upper bound the probability that it is overloads.

- Load $S = \sum_{i=1}^{10} X_i$ where X_i is the weight of the i'th person.
- $E[S] = E[\sum_{i=1}^{10} X_i] = \sum_{i=1}^{10} E[X_i] = 10 \times 80 = 800$
- By Markov inequality, $P(S \geq 1000) \leq E[S]/1000 = 800/1000 = 0.8$
- Note that we need only information about the mean – no need for any knowledge of the distribution of people's weights

Confidence Intervals

- Recall that when a random variable lies in an interval $a \leq X \leq b$ with a specified probability we call this a confidence interval e.g. $p - 0.05 \leq Y \leq p + 0.05$ with probability at least 0.95.



- Chebyshev inequality allows us to calulate confidence intervals given the mean and variance of a random variable.
- For sample mean $\bar{X} = \frac{1}{N}\sum_{k=1}^{N} X_k$, Chebyshev inequality tells us $P(|\bar{X} - \mu| \geq \epsilon) \leq \frac{\sigma^2}{N\epsilon^2}$ where μ is mean of X_k and σ^2 is its variance.
- E.g. When $\epsilon = \frac{\sigma}{\sqrt{0.05N}}$ then $\frac{\sigma^2}{N\epsilon^2} = 0.05$ and Chebyshev tells us that $\mu - \frac{\sigma}{\sqrt{0.05N}} \leq \bar{X} \leq \mu + \frac{\sigma}{\sqrt{0.05N}}$ with probability at least 0.95.

Laws of Large Numbers

Consider N independent and identically distributed (i.i.d) random variables $X_1, \cdots X_N$ each with mean μ and variance σ^2. Let $\bar{X} = \frac{1}{N}\sum_{k=1}^{N} X_k$.

- Weak Law of Large Numbers. For any $\epsilon > 0$:

  $$P(|\bar{X} - \mu| \geq \epsilon) \to 0 \text{ as } N \to \infty$$

  That is, $\bar{X}$ **concentrates** around the mean μ as N increases. Follows from Chebyshev inequality.

- Central Limit Theorem.

  $$\bar{X} \sim N(\mu, \frac{\sigma^2}{N}) \text{ as } N \to \infty$$

  That is, as N increases the distribution of $\bar{X}$ converges to a Normal (or Gaussian) distribution. Variance $\sigma^2/N \to 0$ as $N \to \infty$. So distribution concentrates around the mean μ as N

- CLT gives us another way to estimate a confidence interval i.e. using the properties of the Normal distribution

Continuous Random Variables

Continuous random variables:

- Take on real-values
- e.g. travel time to work, temperature of this room, fraction of Irish population supporting Scotland in the rugby
- Cumulative distribution function (CDF) $F_Y(y) = P(Y \leq y)$ can be used with both discrete and continuous RVs

For continuous random variable Y we have probability density function (PDF) $f_Y(y)$:

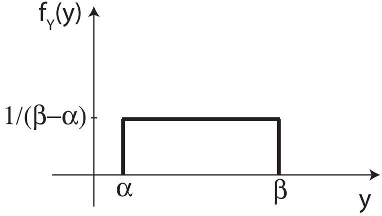
- $f_Y(y) \geq 0$
- $\int_{-\infty}^{\infty} f_Y(y)dy = 1$ (total area under PDF is 1)
- $P(a \leq Y \leq b) = \int_a^b f_Y(y)dy$ (area under PDF between a and b is the probability that $a \leq Y \leq b$)
- $F_Y(b) = \int_{-\infty}^{b} f_Y(y)dy$

Example: Uniform Random Variables

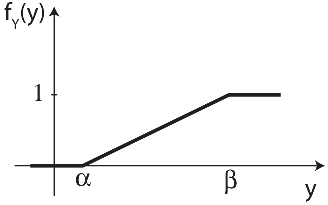
Y is a **uniform random variable** when it has PDF:

$$f_Y(y) = \begin{cases} \frac{1}{\beta-\alpha} & \textit{when } \alpha \leq y \leq \beta \\ 0 & \textit{otherwise} \end{cases}$$

PDF



CDF



- For $\alpha \leq a \leq b \leq \beta$: $P(a \leq Y \leq b) = \frac{b-a}{\beta-\alpha}$
- rand() function in Matlab.
- A bus arrives at a stop every 10 minutes. You turn up at the stop at a time selected uniformly at random during the day and wait for 5 minutes. What is the probability that the bus turns up ?

Expectation and Variance

Just replace sums with integrals when using continuous RVs

For discrete RV X	For continuous RV X
$E[X] = \sum_x xP(X = x)$	$E[X] = \int_{-\infty}^{\infty} xf(x)dx$
$E[X^n] = \sum_x x^n P(X = x)$	$E[X^n] = \int_{-\infty}^{\infty} x^n f(x)dx$

For both discrete and continuous random variables:

$$E[aX + b] = aE[X] + b$$

$$Var(X) = E[(X - \mu)^2] = E[X^2] - (E[X])^2$$

$$Var(aX + b) = a^2 Var(X)$$

Example

A detector looks for edges in an image. Conditioned on an edge being present, the detector response X is Gaussian with mean 0 and variance σ^2. When no edge is present, the detector response is Gaussian with mean 0 and variance 1. An image has an edge with probability p. What is the mean and variance of the detector response.

- Let F be the event that an edge is present.
- $E[X] = E[X|F]P(edge) + E[X|F^c]P(F^c) = 0 \times p + 0 \times (1 - p) = 0$
- $Var(X) = E[X^2|F]P(edge) + E[X^2|F^c]P(F^c) = \sigma^2 \times p + 1 \times (1 - p)$

Joint and Conditional Probability Density Functions

Joint and conditional PDFs behave much the same as probabilities:

- Joint PDF of X and Y is: $f_{XY}(x, y)$
- Conditional PDF is defined as: $f_{X|Y}(x|y) = \frac{f_{XY}(x,y)}{f_Y(y)}$

Chain rule holds for PDFs:

$$f_{XY}(x, y) = f_{X|Y}(x|y)f_Y(y) = f_{Y|X}(y|x)f_X(x)$$

We can marginalise PDFs:

$$\int_{-\infty}^{\infty} f_{XY}(x, y)dy = f_X(x)$$

Bayes Rule holds:

$$f_{Y|X}(y|x) = \frac{f_{X|Y}(x|y)f_Y(y)}{f_X(x)}$$

X are Y are independent when: $f_{XY}(x, y) = f_X(x)f_Y(y)$

Example

Suppose random variable $Y = X + M$, where $M \sim N(0, 1)$. Conditioned on $X = x$, what is the PDF of Y ?

- $f_{Y|X}(y|x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{(y-x)^2}{2})$

Suppose that $X \sim N(0, \sigma)$. What is $f_{X|Y}(x|Y)$?

- Use Bayes Rule:

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x) f_X(x)}{f_Y(y)}$$
$$= \frac{\frac{1}{\sqrt{2\pi}} \exp(-\frac{(y-x)^2}{2}) \times \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{x^2}{2\sigma^2})}{f_Y(y)}$$

- $f_Y(y)$ is just a normalising constant (so that the area under $f_{X|Y}(x|y)$ is 1).

Classification: Logistic Regression

- Label Y only takes values 0 or 1. Real-valued vector $\vec{X}$ of m observed features $X^{(1)}, X^{(2)}, \cdots, X^{(m)}$
- In **Logistic regression** our statistical model is that:

$$P(Y = 1|\Theta = \vec{\theta}, \vec{X} = \vec{x}) = \frac{1}{1 + \exp(-z)} \text{ with } z = \sum_{i=1}^{m} \theta^{(i)} x^{(i)}$$

$$P(Y = 0|\Theta = \vec{\theta}, \vec{X} = \vec{x}) = 1 - P(Y = 1|\Theta = \vec{\theta}, \vec{X} = \vec{x}) = \frac{\exp(-z)}{1 + \exp(-z)}$$

- Model has m parameters $\theta^{(1)}$, $\theta^{(2)} \cdots, \theta^{(m)}$. We gather these together into a vector $\vec{\theta}$
- Training data is RV D. Consists of n observations $d = \{(\vec{x}_1, y_1), \cdots, (\vec{x}_n, y_n)\}$
- **Maximum Likelihood** estimate: select the value of $\vec{\theta}$ which maximises $P(D|\vec{\theta})$.

Linear Regression

- Assume a linear relationship between x and Y

$$Y = \sum_{i=1}^{m} \Theta^{(i)} x^{(i)} + M$$

- $\vec{\Theta}$ is a vector of unknown (perhaps random) parameters and M is random "noise" e.g. $M \sim N(0,1)$, $\Theta^{(i)} \sim N(0,\lambda)$ with the value of λ known.
- Training data D is a set of n observed pairs $d = \{(x_1, y_1), \cdots, (x_n, y_n)\}$
- **Maximum Likelihood** estimate: select the value of $\vec{\theta}$ which maximises $P(D|\vec{\theta})$.
- **Maximum a posteriori** (MAP) estimate: select the value of $\vec{\theta}$ which maximises $P(\vec{\theta}|D)$.