Brief Review
Chapter 1-3

Language of Probabilities

- Define random variables
- Express our beliefs about behaviour of these RVs, and relationships to other RVs
- Examples:
  - $p(x)$ Gaussian means we believe X is Gaussian distributed
  - $p(y \mid X = x)$—or written $p(y \mid x)$— is Gaussian says that conditioned on x, then y is Gaussian; but $p(y)$ might not be Gaussian
  - $p(w)$ and $p(w \mid \text{Data})$

PMFs and PDFs

- Discrete RVs have PMFs
  - outcome space: e.g, $\Omega = \{1, 2, 3, 4, 5, 6\}$
  - event space: powerset (e.g., event {1,2})
  - examples: probability table, Poisson
- Continuous RVs have PDFs
  - outcome space: e.g., $\Omega = [0, 1]$
  - event space: Borel field (e.g., event [0.01, 0.02])
  - example: Gaussian, Gamma

Probability Mass Functions

Ω = discrete sample space
$\mathcal{E} = \mathcal{P}(\Omega)$

Probability mass function:

1. $p : \Omega \to [0, 1]$
2. $\sum_{\omega \in \Omega} p(\omega) = 1$

The probability of any event $A \in \mathcal{E}$ is defined as

$$P(A) = \sum_{\omega \in A} p(\omega)$$

Probability Density Functions

Ω = continuous sample space
$\mathcal{E} = \mathcal{B}(\Omega)$

Probability density function:

1. $p : \Omega \to [0, \infty)$
2. $\int_\Omega p(\omega) d\omega = 1$

Who has never seen an integral?

The probability of any event $A \in \mathcal{E}$ is defined as

$$P(A) = \int_A p(\omega) d\omega.$$

CONDITIONAL DISTRIBUTIONS

Conditional probability distribution:

$$p(y|x) = \frac{p(x,y)}{p(x)}$$

If p(x,y) is small, does this imply that p(y|x) is small?

An example for conditional distributions

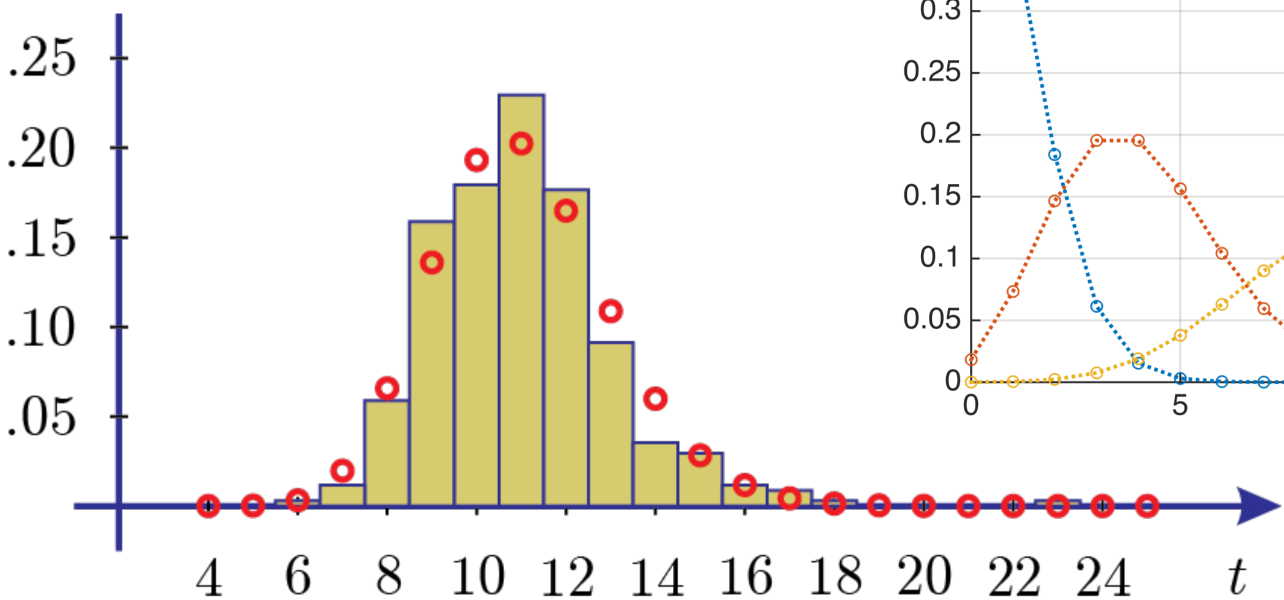
- Two **types of books**: fiction (X=0) and non-fiction (X=1)
- Let Y correspond to **number of pages**
- What is the difference between p(Y = 10 | X = 0) and p(Y = 10, X = 0)?
  - p(Y = 10, X = 0) = probability that a book is fiction and has 10 pages (imagine randomly sampling a book with eyes closed in the library)
  - p(Y = 10 | X = 0) = probability that a fiction book has 10 pages (imagine randomly sampling a book **in the fiction section** of the library with eyes closed)

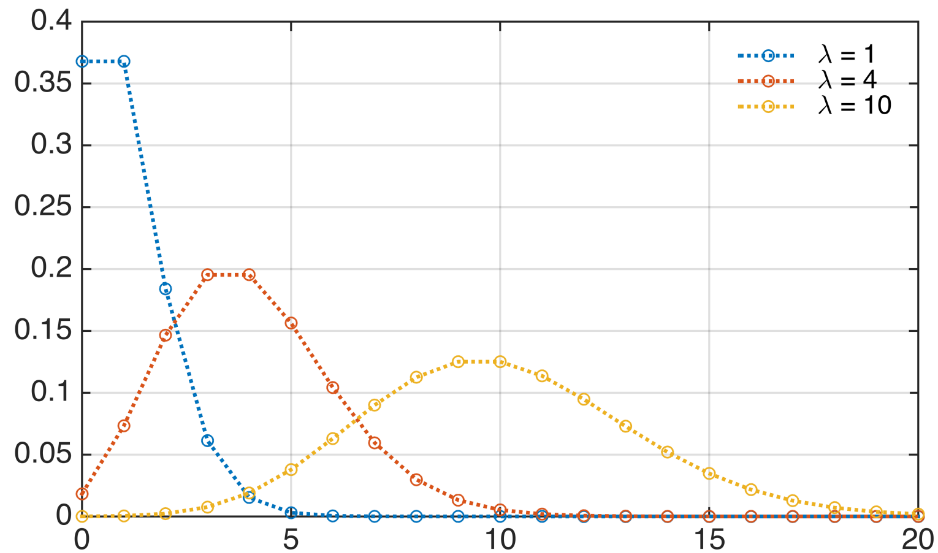
An example for conditional distributions

- Two **types of books**: fiction (X=0) and non-fiction (X=1)
- Let Y correspond to **number of pages**
- What distribution might we have for p(y | X = 0) and p(y | X = 1)?
- How about p(y)?

Recall this Think-Pair-Share

- How might you use a given Poisson distribution, that models commute times?
- How might you pick lambda for a Poisson distribution, to model commute times?





$$p(k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Chain rule and Bayes rule

Recall chain rule: $p(x, y) = p(x|y)p(y) = p(y|x)p(x)$

Bayes rule: $$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

Independence of Random Variables

X and Y are **independent** if:

$$p(x, y) = p(x)p(y)$$

X and Y are **conditionally independent** given Z if:

$$p(x, y|z) = p(x|z)p(y|z)$$

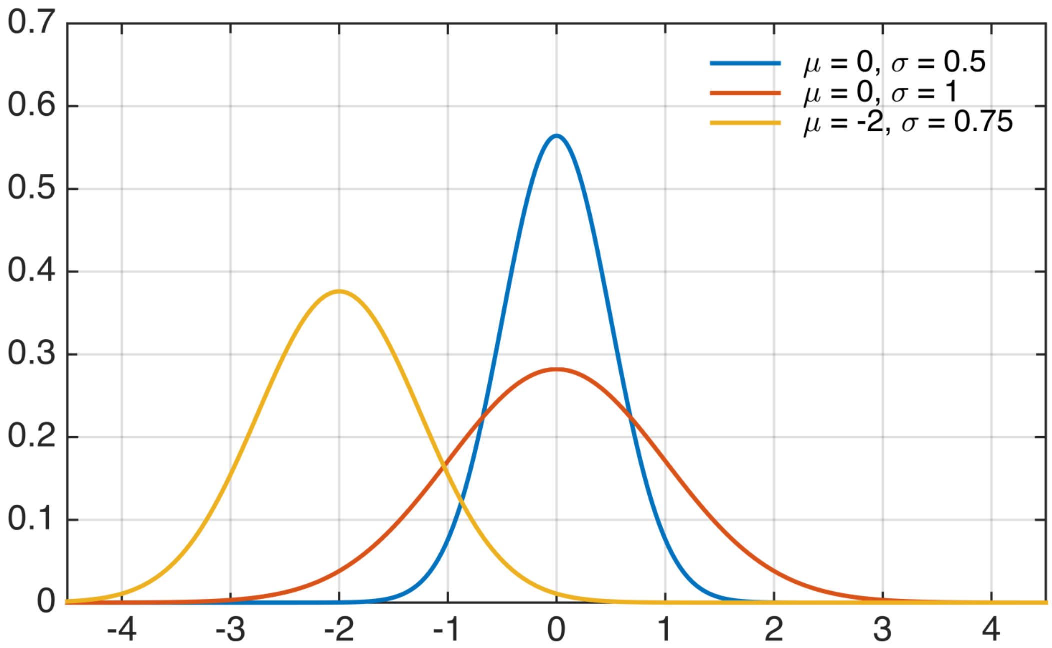
Conditional Independence Examples
Example 7 in the notes

- Imagine you have a biased coin (does not flip 50% heads and 50% tails, but skewed towards one)
- Let Z = bias of a coin (say outcomes are 0.3, 0.5, 0.8 with associated probabilities 0.7, 0.2, 0.1)
  - what other outcome space could we consider?
  - what kinds of distributions?
- Let X and Y be consecutive flips of the coin
- Are X and Y independent?
- Are X and Y conditionally independent, given Z?

**(Basic example about an important issue in ML: hidden variables)

Expected value (Mean, Average)

$$\mathbb{E}[X] = \begin{cases} \sum_{x \in \mathcal{X}} x p(x) & X : \text{discrete} \\ \int_{\mathcal{X}} x p(x) dx & X : \text{continuous} \end{cases}$$



Conditional Expectations

$$\mathbb{E}\left[Y|X=x\right] = \begin{cases} \sum_{y\in\mathcal{Y}} yp(y|x) & Y: \text{discrete} \\ \\ \int_{\mathcal{Y}} yp(y|x)dy & Y: \text{continuous} \end{cases}$$

Different expected value, depending on which x is observed

Properties of Expectations

- $E[cX] = c\,E[X]$, for a constant c
- $E[X + Y] = E[X] + E[Y]$ (linearity of expectation)
- If X and Y independent, then $E[XY] = E[X]\,E[Y]$
- $E[Y] = E[E[Y \mid X]]$, where outer expectation over X
  - called Law of Total Expectation

Properties of Variances

- $V[c] = 0$ for a constant c
- $V[c X] = c^2 V[X]$
- $V[X + Y] = V[X] + V[Y] + 2 \, Cov[X,Y]$
- If X and Y are independent, $V[X + Y] = V[X] + V[Y]$
  - i.e., $Cov[X,Y] = 0$

Sample average is an unbiased estimator

Obtain instances $x_1, \ldots, x_n$

What can we say about the sample average?

This sample is random, so we consider i.i.d. random variables $X_1, \ldots, X_n$

Reflects that we could have seen a different set of instances x_i

$$\begin{aligned} \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] \\ &= \frac{1}{n}\sum_{i=1}^{n}\mu \\ &= \mu \end{aligned}$$

For any one sample $x_1, \ldots, x_n$, unlikely that $\displaystyle\frac{1}{n}\sum_{i=1}^{n} x_i = \mu$

Bias and variance

- Bias of the sample average estimator
  - Bias(Xbar) = E[Xbar] - mu = 0
- Variance of of the sample average estimator
  - Var(Xbar) = sigma^2 / n
- Reflects that variability over possible sample averages you could’ve seen

Concentration Inequality

Confidence Interval:

$$\Pr\left(\left|\bar{X} - \mathbb{E}[\bar{X}]\right| \geq \epsilon\right) \leq \delta.$$

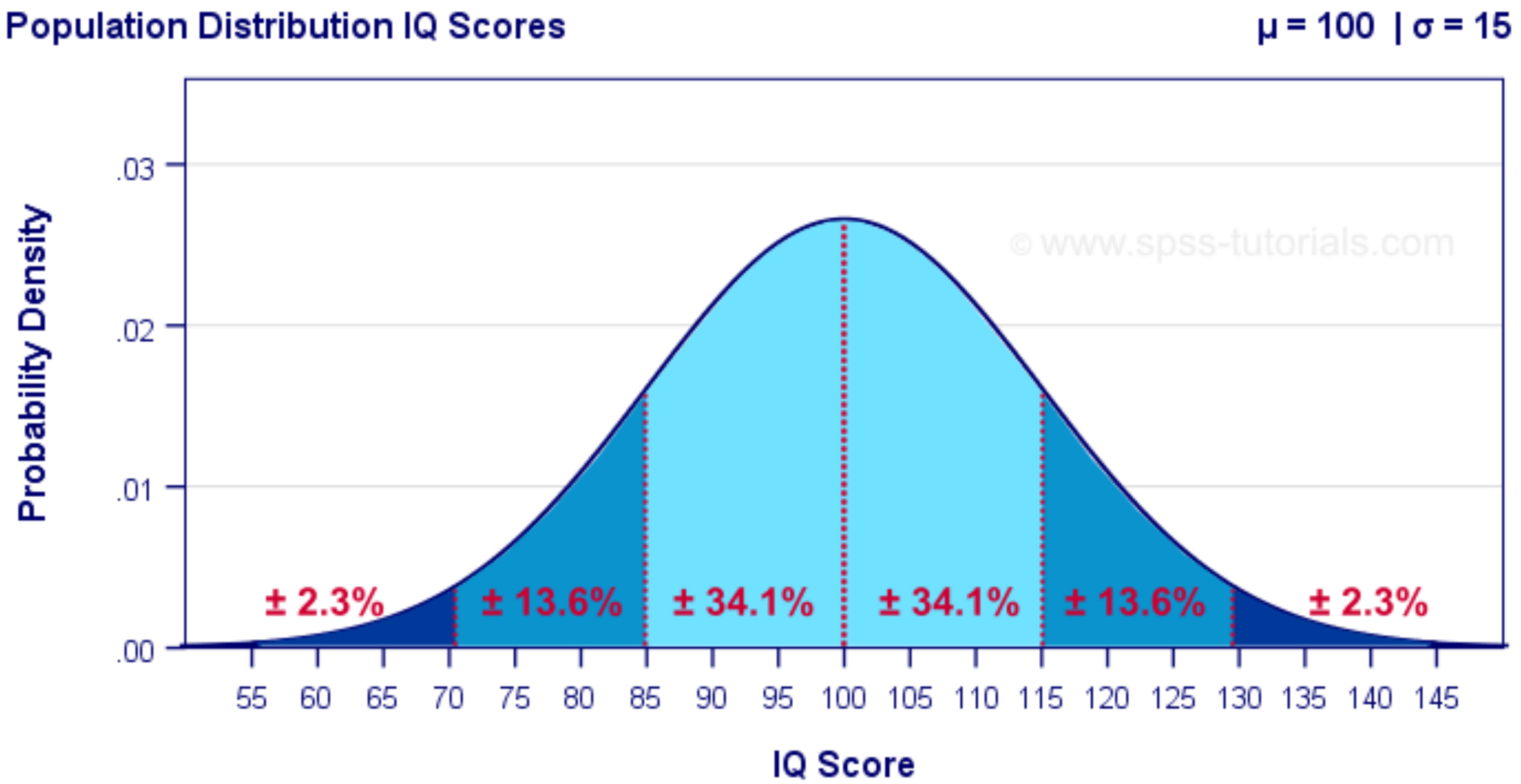
Chebyshev's:

$$\Pr(|\bar{X} - \mathbb{E}[\bar{X}]| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}$$ = delta

Interval under Gaussian Assumption

Gaussian Xi $\Pr(|\bar{X} - \mu| \geq 1.96\sigma/\sqrt{n}) = 0.95$

Unknown dist. Xi $\Pr(|\bar{X} - \mu| \geq 4.47\sigma/\sqrt{n}) = 0.95$



Consistency, Convergence Rate and Sample Complexity

- Consistency: Estimator -> True Value in the limit of infinite data
- Convergence Rate: the speed at which the estimator converges to its limit point
  - rate was typically O(1/sqrt(n)) for us
  - what is rate of estimator that returns 0?
- Sample Complexity: # of samples needed to reach a level of accuracy epsilon
  - upper bounded by 1.96 sigma/sqrt(n)

Question 1. [40 MARKS]

Recall that the expected value of a random variable X is $\mathbb{E}[X] = \sum_{x \in \mathcal{X}} p(X = x)x$, where $\mathcal{X}$ is the set of possible values of X, and the variance is given by $\text{Var}[X] = \mathbb{E}[(X^2 - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$. Suppose you have a coin that has probability p of coming up heads and $1 - p$ of coming up tails. You flip the coin n times. Let the random variable X denote the number of heads you see.

Part (a) [5 MARKS]

What is the outcome space $\mathcal{X}$ for this X?

Part (b) [5 MARKS]

Recall that the probability of seeing k successes in n independent Bernoulli trials is $\binom{n}{k} p^k (1-p)^{n-k}$. Write an expression for $P(X = x)$, in terms of x.

Part (c) [5 MARKS]

Let $X_1, X_2, \ldots, X_n$ correspond to the coin flip outcomes for the n flips. Express X in terms of these X_i.

Part (d) [10 MARKS]

Show that $\mathbb{E}[X] = np$.

Part (e) [15 MARKS]

Derive an expression for the variance, $\text{Var}[X]$.

Question 2. [20 MARKS]

Imagine you are given an estimator, Y, with $\text{Bias}(Y) = 1/\sqrt{n}$. (Recall that bias is $\text{Bias}(Y) \doteq \mathbb{E}[Y] - \mu$ where μ is the unknown parameter for which Y is an estimate.) Is Y a consistent estimator? Explain why or why not.

Question 3. [40 MARKS]

Imagine you have n iid random variables $X_1, X_2, \ldots, X_n$, with $\mathbb{E}[X_i] = \mu$ and $\text{Var}(X_i) = \sigma^2$ for all i. Let $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ be the sample average estimator. To get confidence intervals we used concentration inequalities. Using Chebyshev's inequality, we can say that

$$P(|\bar{X} - \mathbb{E}[\bar{X}]| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2} \tag{1}$$

Part (a) [10 MARKS]

What is $\mathbb{E}[\bar{X}]$?

Part (b) [30 MARKS]

Derive a 95% confidence interval for $\mathbb{E}[\bar{X}]$, using the above inequality. Show your steps.