Tim Gianitsos, Alex Tsun
CS109



# Section #7 Concept Check Solutions

## 1 Lecture 19, 2-21-20: Sampling/Bootstrapping

1. Computing the sample mean is similar to the population mean: sum all available points and divide by the number of points. However, sample variance is slightly different from population variance.

   (a) Consider the equation for population variance, and an analogous equation for sample variance.

   $$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 \qquad (1) \qquad\qquad S^2_{biased} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \qquad (2)$$

   $S^2_{biased}$ is a random variable which estimates the constant $\sigma^2$. Is $E[S^2_{biased}]$ greater or less than $\sigma^2$?

   (b) Write the equation for $S^2_{unbiased}$ (known simply as $S^2$ in the slides). This is known as *Bessel's correction*.

---

1. (a) $E[S^2_{biased}] < \sigma^2$. The intuition is that the spread of a sample of points is generally smaller than the spread of all the points considered together.

   (b) $S^2_{unbiased} = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

---

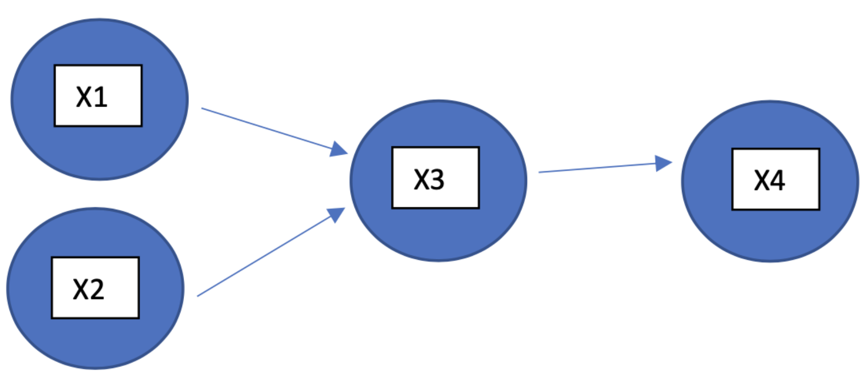
## 2 Lecture 20, 2-24-20: General Inference

Suppose $X_1, \ldots, X_4$ are discrete random variables. We will abuse notation and write $p(x_1, x_2, x_3, x_4)$ to represent $P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4)$. In your answers, feel free to do the same. For example, $p(x_1, x_3) = P(X_1 = x_1, X_3 = x_3)$. Decompose into four terms, each as simple as possible.

1. If there is no assumption of independence, then $p(x_1, x_2, x_3, x_4) =$

2. If all variables are assumed independent, then $p(x_1, x_2, x_3, x_4) =$

3. Assuming the variables follow the Bayesian network structure below, $p(x_1, x_2, x_3, x_4) =$



1. $p(x_1)p(x_2|x_1)p(x_3|x_1, x_2)p(x_4|x_1, x_2, x_3)$ (for example)
2. $p(x_1)p(x_2)p(x_3)p(x_4)$
3. $p(x_1)p(x_2)p(x_3|x_1, x_2)p(x_4|x_3)$

## 3 Lecture 21, 2-26-20: Parameters and MLE

Suppose $x_1, \ldots, x_n$ are iid samples from some distribution with density function $f_X(x; \theta)$, where $\theta$ is unknown. Recall that the likelihood of the data is

$$L(\theta) = \prod_{i=1}^{n} f_X(x_i; \theta)$$

Recall we solve an optimization problem to find $\hat{\theta}$ which maximizes $L$.

1. Write an expression for the log-likelihood, $LL(\theta) = \log L(\theta)$.
2. Why can we optimize $LL(\theta)$ rather than $L(\theta)$?
3. Why do we optimize $LL(\theta)$ rather than $L(\theta)$?

1. $LL(\theta) = \sum_{i=1}^{n} \log f_X(x_i; \theta)$
2. Logarithms are monotonic. This means that if $f(a) > f(b)$, then $\log(f(a)) > \log(f(b))$, so correctness of arg max is preserved.
3. Logs turn products into sums, which makes taking the derivative much simpler.