FIT2086 Exam Revision
Supplementary Questions

Daniel F. Schmidt

November 7, 2017

Contents

1 Introduction

This document contains some extra examples of the types of questions you will be asked on the exam.

2 Short Answer Questions

Please provide 2-3 sentence description of the following terms:

1. The principal of maximum likelihood
2. A mixture model
3. A random forest
4. Penalized regression
5. A random variable

3 Maximum Likelihood Estimation

A random variable Y is said to follow a Gamma distribution with an integer shape parameter equal to α, and a rate parameter β, if

$$\mathbb{P}(Y = y \,|\, \alpha, \beta) = \frac{\beta^\alpha}{(\alpha - 1)!} Y^{\alpha-1} \exp\left(-\beta Y\right)$$

where $y > 0$ is a non-negative continuous number. Imagine we observe a sample of n non-negative real numbers $\mathbf{y} = (y_1, \ldots, y_n)$ and want to model them using a Gamma distribution. *(hint: remember that the data is independently and identically distributed).*

1. Write down the Gamma distribution likelihood function for the data $\mathbf{y}$ (i.e., the joint probability of the data under a Gamma distribution with shape parameter α and rate parameter β).
2. Write down the negative log-likelihood function of the data $\mathbf{y}$ under a Gamma distribution with shape parameter α and rate parameter β.
3. Assume that α is known (i.e., we do not have to estimate it but it is a given constant). Derive the maximum likelihood estimator for β.

4 Confidence Intervals and p-values

A car company runs a fuel efficiency test on a new model of car. They perform 6 tests, and in each test they drive the car until the fuel tank is empty, then calculate the liters of fuel consumed per one-hundred kilometers of distance covered. The observed efficiencies (in litres per 100 kilometers, $L/100km$) were:

$$\mathbf{y} = (7.87, 8.10, 9.07, 8.83, 7.60, 8.91).$$

From previous efficiency experiments the car company has estimated the population standard deviation in fuel efficiency recordings (i.e., the experimental error) to be 0.3 ($L/100km$). We can assume that a normal distribution is appropriate for our data, and that the population standard deviation of fuel efficiency recordings for our experiment is the same as the population fuel efficiency recordings of previous experiments.

1. Using our sample, estimate the population mean fuel effiency for this brand of car. Calculate a 95% confidence interval for the population mean fuel efficiency and summarise your results appropriately.
2. The car company runs the same set of tests, on the same set of cars, but with a different brand of fuel. The new observed fuel efficiencies (again, in $L/100km$) were

   $$\mathbf{y}_B = (7.74, 7.74, 8.22, 7.88, 7.85, 8.27).$$

   The company wants to know if this fuel has made any difference to the fuel efficiency. Again, we can assume the population standard deviation for this new set of fuel efficiency measurements is known to be 0.3 $L/100km$. Using this information, please provide a p-value for testing the null hypothesis that the mean fuel efficiency for the two fuel types is the same. Please interpret this p-value.

5 Random Variables

Suppose Y_1 and Y_2 are two random variables distributed as per $Y_1 \sim \text{Poi}(2)$ and $Y_2 \sim \text{Poi}(4)$. Remember that $\text{Poi}(\lambda)$ denotes a Poisson distribution with rate parameter λ, which means the random variable follows the probability distribution:

$$\mathbb{P}(Y = y \mid \lambda) = \frac{\lambda^Y \exp(-\lambda)}{Y!}.$$

Recall that if $Y \sim \text{Poi}(Y)$, then $\mathbb{E}[Y] = \lambda$ and $\mathbb{V}[Y] = \lambda$. Let $S = Y_1 + Y_2$ denote the sum of these two variables; then:

1. What is the value of $\mathbb{E}[S]$?
2. What is the value of $\mathbb{V}[S]$?
3. If the probability that $S = 0$?
4. What is the value of $\mathbb{E}[Y_1 Y_2]$?

6 Appendix I: Standard Normal Distribution Table

$\lvert z \rvert$	$\mathbb{P}(Z < -\lvert z \rvert)$	$\mathbb{P}(Z < \lvert z \rvert)$	$\lvert z \rvert$	$\mathbb{P}(Z < -\lvert z \rvert)$	$\mathbb{P}(Z < \lvert z \rvert)$
0.000	0.500000	0.500000	2.047	0.020353	0.979647
0.093	0.462943	0.537057	2.140	0.016196	0.983804
0.186	0.426204	0.573796	2.233	0.012789	0.987211
0.279	0.390096	0.609904	2.326	0.010020	0.989980
0.372	0.354912	0.645088	2.419	0.007790	0.992210
0.465	0.320924	0.679076	2.512	0.006009	0.993991
0.558	0.288375	0.711625	2.605	0.004598	0.995402
0.651	0.257471	0.742529	2.698	0.003491	0.996509
0.744	0.228382	0.771618	2.791	0.002630	0.997370
0.837	0.201237	0.798763	2.884	0.001965	0.998035
0.930	0.176125	0.823875	2.977	0.001457	0.998543
1.023	0.153093	0.846907	3.070	0.001071	0.998929
1.116	0.132151	0.867849	3.163	0.000781	0.999219
1.209	0.113273	0.886727	3.256	0.000565	0.999435
1.302	0.096403	0.903597	3.349	0.000406	0.999594
1.395	0.081455	0.918545	3.442	0.000289	0.999711
1.488	0.068326	0.931674	3.535	0.000204	0.999796
1.581	0.056894	0.943106	3.628	0.000143	0.999857
1.674	0.047024	0.952976	3.721	0.000099	0.999901
1.767	0.038577	0.961423	3.814	0.000068	0.999932
1.860	0.031410	0.968590	3.907	0.000047	0.999953
1.953	0.025381	0.974619	> 4.000	< 0.000032	> 0.999968

Table 1: Cumulative Distribution Function for the Standard Normal Distribution $Z \sim N(0,1)$

7 Appendix II: Formulae

Formulae are collated on this page, some of which may be useful in answering this exam.

Probability and Random Variables

Expectation of a RV: $\mathbb{E}[X] = \sum_{x \in \mathcal{X}} x \cdot \mathbb{P}(X = x)$

Marginal probability formula: $\mathbb{P}(Y = y) = \sum_{x \in \mathcal{X}} \mathbb{P}(Y = y, X = x)$

Conditional probability formula: $\mathbb{P}(Y = y \,|\, X = x) = \dfrac{\mathbb{P}(Y = y, X = x)}{\mathbb{P}(X = x)}$

Bayes' Rule: $\mathbb{P}(Y = y \,|\, X = x) = \dfrac{\mathbb{P}(X = x \,|\, Y = y)\mathbb{P}(Y = y)}{\mathbb{P}(X = x)}$

Differentiation

$\dfrac{d}{dx}\{a\, f(x)\} = a \dfrac{d}{dx}\{f(x)\}$

$\dfrac{d}{dx}\{x^k\} = k x^{k-1}$

$\dfrac{d}{dx}\{\log x\} = \dfrac{1}{x}$

Chain rule: $\dfrac{d}{dx}\{f(g(x))\} = \dfrac{d}{d\,g(x)}\{f(g(x))\} \cdot \dfrac{d}{dx}\{g(x)\}$

Confidence Interval and Hypothesis Test for Mean with Known Variance

Let $\hat{\mu}$ be the sample mean of a sample of size n with population variance σ^2. Then a $100(1-\alpha)\%$ confidence interval for μ is

$$\left(\hat{\mu} - z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}},\ \hat{\mu} + z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}\right)$$

where $z_{\alpha/2}$ is the $100(1-\alpha/2)$-percentile of the standard normal distribution $N(0,1)$. To test the null hypothesis $H_0 : \mu = \mu_0$, calculate

$$p = \begin{cases} 2\,\mathbb{P}(Z < -|z_{\hat{\mu}}|) & \text{if } H_0 : \mu = \mu_0 \text{ vs } H_A : \mu \neq \mu_0 \\ 1 - \mathbb{P}(Z < z_{\hat{\mu}}) & \text{if } H_0 : \mu \leq \mu_0 \text{ vs } H_A : \mu > \mu_0 \\ \mathbb{P}(Z < z_{\hat{\mu}}) & \text{if } H_0 : \mu \geq \mu_0 \text{ vs } H_A : \mu < \mu_0 \end{cases}.$$

where $Z \sim N(0,1)$, and

$$z_{\hat{\mu}} = \frac{\hat{\mu} - \mu_0}{\sqrt{\sigma^2/n}}.$$

Confidence Interval and Hypothesis Test for Difference of Means with Known Variances

Let $\hat{\mu}_x$, $\hat{\mu}_y$ be the sample means from two samples of size n_x, n_y, and σ_x^2, σ_y^2 be the known population variances of the two samples. The $100(1-\alpha)\%$ confidence interval for $\mu_x - \mu_y$ is

$$\left(\hat{\mu}_x - \hat{\mu}_y - z_{\alpha/2}\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}},\ \hat{\mu}_x - \hat{\mu}_y + z_{\alpha/2}\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}\right)$$

where $z_{\alpha/2}$ is the $100(1-\alpha/2)$-percentile of the standard normal distribution $N(0,1)$. To test the null hypothesis $H_0 : \mu_x = \mu_y$, calculate

$$p = \begin{cases} 2\,\mathbb{P}(Z < -|z_{(\hat{\mu}_x - \hat{\mu}_y)}|) & \text{if } H_0 : \mu_x = \mu_y \text{ vs } H_A : \mu_x \neq \mu_y \\ 1 - \mathbb{P}(Z < z_{(\hat{\mu}_x - \hat{\mu}_y)}) & \text{if } H_0 : \mu_x \leq \mu_y \text{ vs } H_A : \mu_x > \mu_y \\ \mathbb{P}(Z < z_{(\hat{\mu}_x - \hat{\mu}_y)}) & \text{if } H_0 : \mu_x \geq \mu_y \text{ vs } H_A : \mu_x < \mu_y \end{cases}.$$

where $Z \sim N(0,1)$, and

$$z_{\hat{\mu}_x - \hat{\mu}_y} = \frac{\hat{\mu}_x - \hat{\mu}_y}{\sqrt{\sigma_x^2/n_x + \sigma_y^2/n_y}}$$