18.05 Exam 2 in-class review solutions
Spring 2022

Problem 1. *The following data is from a random sample:*

1, 1, 1, 2, 3, 5, 5, 8, 12, 13, 14, 14, 14, 14, 18, 100.

Find the first, second and third quartiles.

Solution: The first quartile is the value where 25% of the data is below it. We have 16 data points so this is between the 4th and 5th points, i.e. between 2 and 3. It is reasonable to take the midpoint and say 2.5.

The second quartile is between 8 and 12, we say 10.

The third quartile is 14.

Problem 2. MLE examples. *For each of the following, there is an unknown parameter and some data. Give the likelihood function and find the MLE.*

(a) *We have a coin with probability of heads θ. We toss it 10 times and get 3 heads.*

Solution: Likelihood $= P(x = 3|\theta)$. Log likelihood $= \ln(P(x = 3\ |\theta)) = \ln\binom{10}{3} + 3\ln(\theta) + 7\ln(1-\theta)$.

Take the derivative and set to 0: $\dfrac{3}{\theta} - \dfrac{7}{1-\theta} = 0 \ \Rightarrow\ \boxed{\hat{\theta} = \dfrac{3}{10}}$.

(b) *Wait time follows exp(λ). In 5 independent trials wait 3, 5, 4, 5, 2*

Solution: Exponential density is $f(x|\lambda) = \lambda e^{-\lambda x}$. So, since the sum of the data is 19,

$$\text{Likelihood } = f(\text{data}|\lambda) = \lambda^5 e^{-19\lambda}.$$

Log likelihood $= 5\ln(\lambda) - 19\lambda$.

Take the derivative and set to 0: $\dfrac{5}{\lambda} - 19 = 0 \ \Rightarrow\ \boxed{\hat{\lambda} = \dfrac{5}{19}}$.

(c) *We have 4, 6, 8, 12 and 20-sided dice. One is chosen at random and rolled twice giving resulting in a 9 and a 5.*

Solution: We give the likelihood in a table. The hypothesis is *theta*, the number of sides on the chosen die.

Hypothesis θ	Likelihood $P(\text{data} \mid \theta)$
4-sided	0
6-sided	0
8-sided	0
12-sided	1/144
20-sided	1/400

Read directly from the table: MLE = 12-sided die.

Problem 3. MLE examples

For each of the following, there is an unknown parameter and some data. Give the likelihood function and find the MLE.

(a) *In this problem there are two unknown parameters μ and σ.*

Independent samples $x_1, \dots, x_n$ are drawn from a $N(\mu, \sigma^2)$ distribution.

Solution: For the exam do not focus on the calculation here. You should understand the idea that we need to set the partial derivatives with respect to μ and σ to 0 and solve for the critical point $(\hat{\mu}, \hat{\sigma^2})$.

The density is $f(x|\mu,\sigma) = \dfrac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

The likelihood is

$$L(\mu,\sigma) = \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n e^{-\frac{\sum(x_i-\mu)^2}{2\sigma^2}}.$$

So, log likelihood is $l(\mu,\sigma) = -n\ln(\sqrt{2\pi}) - n\ln(\sigma) - \dfrac{\sum(x_i-\mu)^2}{2\sigma^2}$.

Taking partial derivatives and setting them to 0:

$$\frac{\partial l}{\partial \mu} = \frac{2\sum(x_i-\mu)}{2\sigma^2} = 0$$
$$\frac{\partial l}{\partial \sigma} = -\frac{n}{\sigma} + \frac{\sum(x_i-\mu)^2}{\sigma^3} = 0.$$

Soving these equations, we get $\hat{\mu} = \overline{x}$, $\hat{\sigma^2} = \frac{\sum(x_i-\hat{\mu})^2}{n}$.

(b) *One sample $x = 6$ drawn from a uniform$(0,\theta)$ distribution.*

Solution: The likelihood is

$$L(\theta) = \begin{cases} 0 & \text{if } \theta < 6 \\ \frac{1}{\theta} & \text{if } \theta \geq 6 \end{cases}$$

Because of the term $1/\theta$ in the likelihood, the likelihood is at a maximum when θ is as small as possible. **Solution:** : $\hat{\theta} = 6$.

(c) *One sample x drawn from a uniform$(0,\theta)$ distribution.*

Solution: This is identical to part (b), except the exact value of x is not given.

Answer: $\hat{\theta} = x$.

Problem 4. Discrete prior-discrete likelihood.

Jon has 1 four-sided, 2 six-sided, 2 eight-sided, 2 twelve sided, and 1 twenty-sided dice. He picks one at random and rolls a 7.

(a) *For each type of die, find the posterior probability Jon chose that type.*

Solution: Make a table. (The last column is included for part (d).)

Hypothesis θ	Prior $P(\theta)$	Likelihood $\phi(x_1 = 7 \mid \theta)$	Bayes numerator	posterior $f(\theta \mid x_1 = 7)$	likelihood $P(x_2 = 8 \mid \theta)$
4-sided	1/8	0	0	0	0
6-sided	1/4	0	0	0	0
8-sided	1/4	1/8	1/32	$1/32T \approx 0.536$	1/8
12-sided	1/4	1/12	1/48	$1/48T \approx 0.357$	1/12
20-sided	1/8	1/20	1/160	$1/160T \approx 0.107$	1/20
Total	1		$T = \frac{1}{32} + \frac{1}{48} + \frac{1}{160}$	1	

The posterior probabilities are given in the 5th column of the table. The total probability $T = \frac{7}{120}$ is also the answer to part (c).

(b) *What are the posterior odds Jon chose the 20-sided die?*

Solution: $\text{Odds}(\text{20-sided} \mid x_1 = 7) = \dfrac{P(\text{20-sided} \mid x_1 = 7)}{P(\text{not 20-sided} \mid x_1 = 7)} = \dfrac{1/160T}{1/32T + 1/48T} = 0.12.$

(c) *Compute the prior predictive probability of rolling a 7 on the first roll.*

Solution: $P(x_1 = 7) = T = 7/120.$

(d) *Compute the posterior predictive probability of rolling an 8 on the second roll.*

Solution: See the last two columns in the table. $P(x_2 = 8 \mid x_1 = 7) = \frac{1}{32T} \cdot \frac{1}{8} + \frac{1}{48T} \cdot \frac{1}{12} + \frac{1}{160T} \cdot \frac{1}{20} = \frac{49}{480}$.

Problem 5. *Suppose $x \sim$ binomial$(30, \theta)$, $x = 12$. If we have a prior $f(\theta) \sim$ Beta$(1, 1)$ find the posterior for θ.*

Solution: To be able to talk about this, let's call x the number of successes. So the data is 12 successes and 18 failures. We know how a Beta prior updates with a binomial likelihood: Prior $f(\theta) \sim \text{Beta}(1, 1)$ gives $\boxed{\text{posterior } f(\theta \mid x = 12) \sim \text{Beta}(13, 19).}$

Alternate method. We can also do this with a table. Notice that we don't bother to specify the normalizing constants since the posterior has the form of a Beta(13,19) distribution.

Hypothesis θ	Prior $f(\theta)\,d\theta$	Likelihood $\phi(x = 12\|\theta)$	Bayes numerator	posterior $f(\theta \mid x = 6)$
θ	$c_1\theta^0(1-\theta)^0\,d\theta$	$c_2\theta^{12}(1-\theta)^{18}$	$c_3\theta^{12}(1-\theta)^{18}\,d\theta$	$c_4\theta^{12}(1-\theta)^{18}\,d\theta$
Total	1		$T = \int_0^1 c_3\theta^{12}(1-\theta)^1\,d\theta$	1

Problem 6. *Suppose $x \sim$ geometric(θ), $x = 6$. If we have a prior $f(\theta) \sim$ Beta$(4, 2)$ find the posterior for θ.*

Solution: Our definition of the geometric distribution means that $x = 6$ represents 6 successes in a row and then 1 failure. The updating has the same rule as in the previous problem:

Prior: $f(\theta) \sim \text{Beta}(4, 2)$ gives $\boxed{\text{posterior } f(\theta \mid x = 6) \sim \text{Beta}(10, 3).}$

(We could also do this problem using an update table.)

Problem 7. *In the population IQ is normally distributed: $\theta \sim N(100, 15^2)$. An IQ test finds a person's 'true' IQ + random error $\sim N(0, 10^2)$. Someone takes the test and scores 120.*

Find the posterior pdf for this person's IQ.

Solution: Prior, $f(\theta) \sim \text{N}(100, 15^2)$, $x \sim \text{N}(\theta, 10^2)$.

So we have, $\mu_{\text{prior}} = 100$, $\sigma^2_{\text{prior}} = 15^2$, $\sigma^2 = 10^2$, $n = 1$, $\overline{x} = x = 120$.

Applying the normal-normal update formulas: $a = \frac{1}{15^2}$, $b = \frac{1}{10^2}$. This gives

$\mu_{\text{post}} = \frac{100/15^2 + 120/10^2}{1/15^2 + 1/10^2} = 113.8$, $\sigma^2_{\text{post}} = \frac{1}{1/15^2 + 1/10^2} = 69.2$

Problem 8. **z and one-sample t-test.** *For both problems use significance level $\alpha = 0.05$. Assume the data 2, 4, 4, 10 are independent draws from a $N(\mu, \sigma^2)$ distribution.*

Take H_0: $\mu = 0$; *H_A: $\mu \neq 0$.*

(a) *Assume $\sigma^2 = 16$ is known and test H_0 against H_A.*

Solution: We have $\overline{x} = 5$, $s^2 = \frac{9+1+1+25}{3} = 12$

We'll use z for the test statistic (we could also use $\overline{x}$).

$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{5}{2} = 2.5.$$

The null distribution for z is $\mathrm{N}(0, 1)$. This is a two-sided test so the rejection region is

$$(z \leq z_{0.975} \text{ or } \ z \geq z_{0.025}) \ = \ (-\infty, -1.96] \cup [1.96, \infty)$$

Since our z-statistic in the rejection region we reject H_0 in favor of H_A.

Repeating the test using a p-value:

$$p = P(|z| \geq 2.5 \,|\, H_0) = 2 * \texttt{pnorm}(-2.5, 0, 1) \approx 0.012$$

Since $p < \alpha$ we reject H_0 in favor of H_A.

(b) *Now assume σ^2 is unknown and test H_0 against H_A.*

Solution: We have $\bar{x} = 5$, $s^2 = \frac{9+1+1+25}{3} = 12$

We'll use $t = \frac{\bar{x}-\mu}{s/\sqrt{n}}$ for the test statistic. The null distribution for t is t_3. For the data we have $t = 5/\sqrt{3}$. This is a two-sided test so the p-value is

$$p = P(|t| \geq 5/\sqrt{3} | H_0) = 2 * \texttt{pt}(-5/\texttt{sqrt}(3), 3) \approx 0.06318$$

Since $p > \alpha$ we do not reject H_0.

Problem 9.

Two-sample t-test *Suppose that we have data from 1408 women admitted to a maternity hospital for (i) medical reasons or through (ii) unbooked emergency admission. The duration of pregnancy is measured in complete weeks from the beginning of the last menstrual period.*

(i) Medical: 775 observations with $\bar{x} = 39.08$ and $s^2 = 7.77$.

(ii) Emergency: 633 observations with $\bar{x} = 39.60$ and $s^2 = 4.95$.

(a) *Set up and run a two-sample t-test to investigate whether the duration differs for the two groups.*

Solution: The pooled variance for this data is

$$s_p^2 = \frac{774(7.77) + 632(4.95)}{1406}\left(\frac{1}{775} + \frac{1}{633}\right) = 0.0187$$

The t statistic for the null distribution is

$$\frac{\bar{x} - \bar{y}}{s_p} = -3.8064$$

Rather than compute the two-sided p-value using `2*pt(-3.8064,1406)` we simply note that with 1406 degrees of freedom the t distribution is essentially standard normal and 3.8064 is almost 4 standard deviations. So

$$P(|t| \geq 3.8064) = P(|z| \geq 3.8064)$$

which is very small, much smaller than $\alpha = 0.05$ or $\alpha = 0.01$. Therefore we reject the null hypothesis in favor of the alternative that there is a difference in the mean durations.

(b) *What assumptions did you make?*

Solution: We assumed the data was normal and that the two groups had equal variances. Given the big difference in the sample variances this assumption might not be warranted.

Note: there are significance tests to see if the data is normal and to see if the two groups have the same variance.

Problem 10. *Three treatments for a disease are compared in a clinical trial, yielding the following data:*

	Treatment 1	*Treatment 2*	*Treatment 3*
Cured	*50*	*30*	*12*
Not cured	*100*	*80*	*18*

Use a chi-square test to compare the cure rates for the three treatments

Solution: The null hypothesis H_0 is: all three treatments have the same cure rate.

Under H_0 the MLE for the cure rate is: (total cured)/(total treated) $= 92/290 = 0.317$.

Given H_0 we get the following table of observed and expected counts. We include the fixed values in the margins

	Treatment 1	Treatment 2	Treatment 3	
Cured	50, 47.6	30, 34.9	12, 9.5	92
Not cured	100, 102.4	80, 75.1	18, 20.5	198
	150	110	30	

Pearson's chi-square statistic: $X^2 = \sum \dfrac{(O_i - E_i)^2}{E_i} = 2.13.$

Likelihood ratio statistic: $G = 2\sum O_i \ln(O_i/E_i) = 2.12.$

To compute the expected counts, we need all the marginal counts. If we made up data that had these same marginal counts we could put values in 2 of the cells freely and then all the others are determined. Thus, degrees of freedom $= 2$. Using R we compute the p-value using the χ^2 distribution with 2 degrees of freedom.

```
p = 1 - pchisq(2.12, 2) = 0.346
```

(We used the G statistic, but we would get essentially the same answer using X^2.)

For the exam you would have to use the χ^2 table to estimate the p-value. In the $df = 2$ row of the table 2.12 is between the critical values for $p = 0.3$ and $p = 0.5$.

The problem did not specifiy a significance level, but a p-value of 0.35 does not support rejecting H_0 at any common level. We do not conclude that the treatments have differing efficacy.

Problem 11. ANOVA. *The table shows recovery time in days for three medical treatments.*

T_1	T_2	T_3
6	8	13
8	12	9
4	9	11
5	11	8
3	6	7
4	8	12

Note: *For $\alpha = 0.05$, the critical value of $F_{2,15}$ is 3.68.*

(a) *Set up and run an F-test for H_0 vs. H_A.*

Solution: It's not stated but we have to assume independence and normality.

$n = 3$ groups, $m = 6$ data points in each group.

F-stat: $f \sim F_{n-1,n(m-1)} = F_{2,15}$.

Group means: (Treatments 1-3): $\overline{y}_1 = 5$, $\overline{y}_2 = 9$, $\overline{y}_3 = 10$.

Grand mean: $\overline{y} = 8$.

Group variances: $s_1^2 = 16/5$, $s_2^2 = 24/5$, $s_3^2 = 28/5$.

$MS_B = \frac{6}{2}(14) = 42$, $MS_W = \frac{68}{15}$, $f = \frac{MS_B}{MS_W} = \frac{42}{68/15} = 9.264$.

(b) *Based on the test, what might you conclude about the treatments?*

Solution: Since $9.264 > 3.68$, at a significance level of 0.05 we reject the null hypothesis that all the means are equal. That is, we conclude that the recovery time is not the same for all 3 treatments.

18.05 Introduction to Probability and Statistics
Spring 2022