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 \( \theta \). We toss it 10 times and get 3 heads.

**Solution:** Likelihood = \( P(x = 3|\theta) \). Log likelihood = \( \ln(P(x = 3|\theta)) = \ln\left(\frac{10}{3}\right) + 3\ln(\theta) + 7\ln(1 - \theta) \).

Take the derivative and set to 0: \( \frac{3}{\theta} - \frac{7}{1 - \theta} = 0 \Rightarrow \hat{\theta} = \frac{3}{10} \).

(b) Wait time follows \( \exp(\lambda) \). 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: \( \frac{5}{\lambda} - 19 = 0 \Rightarrow \hat{\lambda} = \frac{5}{19} \).

(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 \( \theta \) | Likelihood \( P(\text{data} | \theta) \) |
|-------------------------|------------------------------------------|
| 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 \( \mu \) and \( \sigma \).

Independent samples \( x_1, \ldots, x_n \) 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 \( \mu \) and \( \sigma \) to 0 and solve for the critical point \( (\hat{\mu}, \hat{\sigma}^2) \).

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

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 \( l(\mu, \sigma) = -n \ln(\sqrt{2\pi}) - n \ln(\sigma) - \frac{\sum (x_i - \mu)^2}{2\sigma^2} \).

Taking partial derivatives and setting them to 0:

\[
\frac{\partial l}{\partial \mu} = \frac{2 \sum (x_i - \mu)}{2\sigma^2} = 0,
\]

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

Solving these equations, we get \( \hat{\mu} = \bar{x}, \quad \hat{\sigma}^2 = \frac{\sum (x_i - \bar{x})^2}{n} \).

(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 \( \theta \) 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: \( \hat{\theta} = x \).


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).)

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Prior ( P(\theta) )</th>
<th>Likelihood ( \phi(x_1 = 7 \mid \theta) )</th>
<th>Bayes numerator</th>
<th>posterior ( f(\theta \mid x_1 = 7) )</th>
<th>likelihood ( P(x_2 = 8 \mid \theta) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>4-sided</td>
<td>1/8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6-sided</td>
<td>1/4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8-sided</td>
<td>1/4</td>
<td>1/8</td>
<td>1/32</td>
<td>1/32T \approx 0.536</td>
<td>1/8</td>
</tr>
<tr>
<td>12-sided</td>
<td>1/4</td>
<td>1/12</td>
<td>1/48</td>
<td>1/48T \approx 0.357</td>
<td>1/12</td>
</tr>
<tr>
<td>20-sided</td>
<td>1/8</td>
<td>1/20</td>
<td>1/160</td>
<td>1/160T \approx 0.107</td>
<td>1/20</td>
</tr>
<tr>
<td>Total</td>
<td>1</td>
<td>( T = \frac{1}{32} + \frac{1}{12} + \frac{1}{160} )</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>
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: Odds(20-sided $| x_1 = 7$) = \frac{P(20\text{-sided} | x_1 = 7)}{P(\text{not 20\text{-sided}} | x_1 = 7)} = \frac{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 of the table. $P(x_2 = 8 | x_1 = 7) = \frac{1/32T \cdot 1/8 + 1/32T \cdot 1/12 + 1/160T \cdot 1/20 = 49/480}$.

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

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 $\text{posterior } f(\theta | x = 12) \sim \text{Beta}(13, 19)$.

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

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

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 $\text{posterior } f(\theta | x = 6) \sim \text{Beta}(10, 3)$.

(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, $f(\theta) \sim N(100, 15^2)$, $x \sim N(\theta, 10^2)$.

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

Applying the normal-normal update formulas: $a = \frac{1}{15^2}$, $b = \frac{1}{10^2}$. This gives $\mu_{\text{post}} = \frac{100/15^2 + 120/10^2}{1/15^2 + 1/10^2} = 113.8$, $\sigma_{\text{post}}^2 = \frac{1}{1/15^2 + 1/10^2} = 69.2$.
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 $\bar{x} = 5$, $s^2 = \frac{9+1+1+25}{3} = 12$

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

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

The null distribution for $z$ is $N(0,1)$. 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:

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

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

(b) Now assume $\sigma^2$ 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 $t = \frac{5}{\sqrt{3}}$. This is a two-sided test so the $p$-value is

$$p = P(|t| \geq \frac{5}{\sqrt{3}} \mid H_0) = 2 \times \text{pt}(-5/\sqrt{3}, 3) \approx 0.06318$$

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 $2 \times \text{pt}(-3.8064,1406)$ 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:

<table>
<thead>
<tr>
<th></th>
<th>Treatment 1</th>
<th>Treatment 2</th>
<th>Treatment 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cured</td>
<td>50</td>
<td>30</td>
<td>12</td>
</tr>
<tr>
<td>Not cured</td>
<td>100</td>
<td>80</td>
<td>18</td>
</tr>
</tbody>
</table>

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: $(\text{total cured})/(\text{total treated}) = 92/290 = 0.317$.

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

<table>
<thead>
<tr>
<th></th>
<th>Treatment 1</th>
<th>Treatment 2</th>
<th>Treatment 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cured</td>
<td>50, 47.6</td>
<td>30, 34.9</td>
<td>12, 9.5</td>
</tr>
<tr>
<td>Not cured</td>
<td>100, 102.4</td>
<td>80, 75.1</td>
<td>18, 20.5</td>
</tr>
<tr>
<td></td>
<td>150</td>
<td>110</td>
<td>30</td>
</tr>
</tbody>
</table>

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

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 $\chi^2$ distribution with 2 degrees of freedom.

$$p = 1 - \text{pchisq}(2.12, 2) = 0.346$$

(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 $\chi^2$ 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 specify 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.

<table>
<thead>
<tr>
<th></th>
<th>$T_1$</th>
<th>$T_2$</th>
<th>$T_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>8</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>12</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>9</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>11</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>12</td>
<td></td>
</tr>
</tbody>
</table>

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): $\bar{y}_1 = 5, \bar{y}_2 = 9, \bar{y}_3 = 10$.

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

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

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

(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.