1 Part A: Statistical Inference

Problem 1

(a) *Express $P(X_i = 1)$ as a function of a and π.*

We have

$$\begin{aligned} P(X_i = 1) &= P(X_i = 1, \text{Group} = A) + P(X_i = 1, \text{Group} = B) \\ &= P(X_i = 1 \mid \text{Group} = A) \cdot P(\text{Group} = A) + P(X_i = 1 \mid \text{Group} = B) \cdot P(\text{Group} = B) \\ &= a \cdot P(\text{Group} = A) + (1 - a) \cdot P(\text{Group} = B) \\ &= a \cdot \pi + (1 - a) \cdot (1 - \pi) \, . \end{aligned}$$

Similarly, we have

$$\begin{aligned} P(X_i = 0) &= P(X_i = 0, \text{Group} = A) + P(X_i = 0, \text{Group} = B) \\ &= P(X_i = 0 \mid \text{Group} = A) \cdot P(\text{Group} = A) + P(X_i = 0 \mid \text{Group} = B) \cdot P(\text{Group} = B) \\ &= (1 - a) \cdot P(\text{Group} = A) + a \cdot P(\text{Group} = B) \\ &= (1 - a) \cdot \pi + a \cdot (1 - \pi) \, . \end{aligned}$$

(Optional) Because the random variable X_i can only have 2 potential values, we can now easily check whether our expressions are correct as $P(X_i = 1) + P(X_i = 0) = 1$. We have

$$\begin{aligned} P(X_i = 1) + P(X_i = 0) &= a \cdot \pi + (1 - a) \cdot (1 - \pi) + (1 - a) \cdot \pi + a \cdot (1 - \pi) \\ &= a \cdot \pi + (1 - \pi) - a + \pi \cdot a + \pi - \pi \cdot a + a - \pi \cdot a \\ &= a \cdot \pi + 1 - \pi + \pi - \pi \cdot a \\ &= 1 \, . \end{aligned}$$

(b) *Give expressions for the joint probability of the observed data $X_1, X_2, \ldots, X_n$, and the corresponding likelihood function and loglikelihood function.*

We aim to express the computation for

$$P(X_1, X_2, \ldots, X_n) \, .$$

Because we know that the n samples are drawn with the i.i.d. assumption, it holds that

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i) \, .$$

Hence,

$$\begin{aligned} P(X_1, X_2, \ldots, X_n) &= \prod_{i=1}^{n} P(X_i) \\ &= \prod_{i=1}^{n} \begin{cases} P(X_i = 1) & \text{if } X_i = 1 \\ P(X_i = 0) & \text{if } X_i = 0 \end{cases} \\ &= P(X_i = 1)^m \cdot P(X_i = 0)^{n-m} \\ &= (a \cdot \pi + (1-a) \cdot (1-\pi))^m \cdot ((1-a) \cdot \pi + a \cdot (1-\pi))^{n-m} \end{aligned} \tag{1}$$

is the likelihood function of the observed samples.

Let us now derive the experession for the log likelihood of the data:

$$\begin{aligned} \log P(X_1, X_2, \ldots, X_n) &= \log \left[(a \cdot \pi + (1-a) \cdot (1-\pi))^m \cdot ((1-a) \cdot \pi + a \cdot (1-\pi))^{n-m} \right] \\ &= m \cdot \log \left[a \cdot \pi + (1-a) \cdot (1-\pi) \right] + (n-m) \cdot \log \left[(1-a) \cdot \pi + a \cdot (1-\pi) \right] \end{aligned} \tag{2}$$

(c) *Derive an expression for the maximum likelihood estimator π_{ML} of π.*

We now aim to estimate π. To this end, we denoted the estimated value for π as π_{ML} in Equation 2 and derive the log likelihood w.r.t. π_{ML}. We have

$$\begin{aligned} &\frac{\partial}{\partial \pi_{ML}} \log P(X_1, X_2, \ldots, X_n) \\ &= \frac{\partial}{\partial \pi_{ML}} m \cdot \log \left[a \cdot \pi_{ML} + (1-a) \cdot (1-\pi_{ML}) \right] + (n-m) \cdot \log \left[(1-a) \cdot \pi_{ML} + a \cdot (1-\pi_{ML}) \right] \\ &= m \cdot \frac{1}{a \cdot \pi_{ML} + (1-a) \cdot (1-\pi_{ML})} \cdot (a - (1-a)) + (n-m) \cdot \frac{1}{(1-a) \cdot \pi_{ML} + a \cdot (1-\pi_{ML})} \cdot ((1-a) - a) \, . \end{aligned}$$

We now find the critical points by setting the gradient to zero:

$$\begin{aligned} \frac{\partial}{\partial \pi_{ML}} \log P(X_1, X_2, \ldots, X_n) &\stackrel{!}{=} 0 \\ m \cdot \frac{1}{a \cdot \pi_{ML} + (1-a) \cdot (1-\pi_{ML})} \cdot (a - (1-a)) + (n-m) \cdot \frac{1}{(1-a) \cdot \pi_{ML} + a \cdot (1-\pi_{ML})} \cdot ((1-a) - a) &= 0 \end{aligned} \tag{3}$$

where Equation 3 holds whenever $(1-a) = a$ as both additive terms are zero. It is easy to see that $a = 0.5$ is the only real solution for this equation. However, as we are interested in the maximum likelihood estimator of π_{ML}, setting $a = 0.5$ prohibits us from finding the maximum value. This is because if $a = 0.5$, we have $P(X_i = 1) = 0.5 = P(X_i = 0)$, and therefore the likelihood function is constant, regardless of the value of π. Therefore, we need to consider the case where $a \neq 0.5$.

We now solve Equation 3 for π_{ML} while assuming that $q \neq 0.5$. We have:

$$
\begin{aligned}
m \cdot \frac{1}{a \cdot \pi_{ML} + (1-a) \cdot (1-\pi_{ML})} \cdot (a - (1-a)) &= -(n-m) \cdot \frac{1}{(1-a) \cdot \pi_{ML} + a \cdot (1-\pi_{ML})} \cdot ((1-a) - a) \\
m \cdot \frac{1}{a \cdot \pi_{ML} + (1-a) \cdot (1-\pi_{ML})} \cdot (a - (1-a)) &= (n-m) \cdot \frac{1}{(1-a) \cdot \pi_{ML} + a \cdot (1-\pi_{ML})} \cdot (a - (1-a)) \\
m \cdot \frac{1}{a \cdot \pi_{ML} + (1-a) \cdot (1-\pi_{ML})} &= (n-m) \cdot \frac{1}{(1-a) \cdot \pi_{ML} + a \cdot (1-\pi_{ML})} \\
m \cdot \frac{1}{(2a-1) \cdot \pi_{ML} + 1 - a} &= (n-m) \cdot \frac{1}{(1-2a) \cdot \pi_{ML} + a} \\
m \cdot \frac{1}{(2a-1) \cdot \pi_{ML} + 1 - a} &= (n-m) \cdot \frac{1}{a - (2a-1) \cdot \pi_{ML}} \\
m \cdot (a - (2a-1) \cdot \pi_{ML}) &= (n-m) \cdot ((2a-1) \cdot \pi_{ML} + 1 - a) \\
am - (2a-1) \cdot m \cdot \pi_{ML} &= (n-m) \cdot (2a-1) \cdot \pi_{ML} + (n-m) \cdot (1-a) \\
am - (n-m) \cdot (1-a) &= (n-m) \cdot (2a-1) \cdot \pi_{ML} + (2a-1) \cdot m \cdot \pi_{ML} \\
am - (n-m) \cdot (1-a) &= \pi_{ML} \cdot (2a-1)((n-m) + m) \\
am - (n-m) + a \cdot n - am &= \pi_{ML} \cdot (2a-1)(n) \\
n \cdot (a-1) + m &= \pi_{ML} \cdot (2a-1)(n) \\
\frac{n \cdot (a-1) + m}{(2a-1)(n)} &= \pi_{ML} \\
\frac{n \cdot (a-1) + m}{n} \cdot \frac{1}{2a-1} &= \pi_{ML} \\
\left(a - 1 + \frac{m}{n}\right) \cdot \frac{1}{2a-1} &= \pi_{ML}
\end{aligned} \tag{4}
$$

We verify that this is indeed a maximum by plugging the result from Equation 4 into the second derivative. For ease of notation, we will use $b = (1-a)$. We have

$$
\begin{aligned}
&\frac{\partial^2}{\partial {\pi_{ML}}^2} \log P(X_1, X_2, \ldots, X_n) \\
= &\frac{\partial}{\partial \pi_{ML}} m \cdot (a - (1-a)) \cdot \frac{1}{a \cdot \pi_{ML} + (1-a) \cdot (1-\pi_{ML})} + (n-m) \cdot ((1-a) - a) \cdot \frac{1}{(1-a) \cdot \pi_{ML} + a \cdot (1-\pi_{ML})} \\
= &-m \cdot (a - (1-a)) \cdot \frac{1}{(a \cdot \pi_{ML} + (1-a) \cdot (1-\pi_{ML}))^2} \cdot (a - (1-a)) \\
&-(n-m) \cdot ((1-a) - a) \cdot \frac{1}{((1-a) \cdot \pi_{ML} + a \cdot (1-\pi_{ML}))^2} \cdot ((1-a) - a) \\
= &-m \cdot (a - (1-a)) \cdot \frac{1}{(a \cdot \pi_{ML} + (1-a) \cdot (1-\pi_{ML}))^2} \cdot (a - (1-a)) \\
&-(n-m) \cdot (a - (1-a)) \cdot \frac{1}{((1-a) \cdot \pi_{ML} + a \cdot (1-\pi_{ML}))^2} \cdot (a - (1-a)) \\
= &-m \cdot (a - (1-a))^2 \cdot \frac{1}{(a \cdot \pi_{ML} + (1-a) \cdot (1-\pi_{ML}))^2} - (n-m) \cdot (a - (1-a))^2 \cdot \frac{1}{((1-a) \cdot \pi_{ML} + a \cdot (1-\pi_{ML}))^2} \\
= &-m \cdot (a-b)^2 \cdot \frac{1}{(a \cdot \pi_{ML} + b \cdot (1-\pi_{ML}))^2} - (n-m) \cdot (a-b)^2 \cdot \frac{1}{(b \cdot \pi_{ML} + a \cdot (1-\pi_{ML}))^2} \\
\leq &-m \cdot \frac{1}{(a \cdot \pi_{ML} + b \cdot (1-\pi_{ML}))^2} - (n-m) \cdot \frac{1}{(b \cdot \pi_{ML} + a \cdot (1-\pi_{ML}))^2}
\end{aligned}
$$

$$
\begin{aligned}
&= -m \cdot \frac{1}{\left(a \cdot \pi_{ML} + (1-a) \cdot (1-\pi_{ML})\right)^2} - (n-m) \cdot \frac{1}{\left((1-a) \cdot \pi_{ML} + a \cdot (1-\pi_{ML})\right)^2} \\
&= -m \cdot \alpha - (n-m) \cdot \beta \\
&\leq -m - (n-m) \\
&\leq 0
\end{aligned}
$$

with $\alpha > 0$ and $\beta > 0$ from the squaring of the denominator. Hence, we know that the second derivative is negative whenever $n > 0$. From this, it follows that our estimator π_{ML} is indeed a maximum.

(d) *Can you think of an easier way to arrive at the same estimator as the maximum likelihood estimator?*

We observe that X_i follows a Bernoulli distribution. We now first show that this holds.

Recall that the definition of a Bernoulli distribution is given by

$$
\begin{aligned}
P(X = 1) &= p \\
P(X = 0) &= 1 - p \, .
\end{aligned}
$$

Therefore, we now aim to show that $P(X_i)$ follows a Bernoulli distribution. From (a), we have that We have

$$
P(X_i = 1) = a \cdot \pi + (1-a) \cdot (1-\pi)
$$

and

$$
P(X_i = 0) = (1-a) \cdot \pi + a \cdot (1-\pi) \, .
$$

We multiply out the expression to get

$$
\begin{aligned}
P(X_i = 1) &= a \cdot \pi + 1 - \pi - a \cdot (1-\pi) \\
&= a \cdot \pi + 1 - \pi - a + a \cdot \pi \\
&= 2a \cdot \pi + 1 - \pi - a \\
&= 1 + 2a \cdot \pi - \pi - a \qquad (5) \\
P(X_i = 0) &= (1-a) \cdot \pi + a \cdot (1-\pi) \\
&= \pi - a \cdot \pi + a - a \cdot \pi \\
&= -2a \cdot \pi + \pi + a \, . \qquad (6)
\end{aligned}
$$

Therefore, with $p = 1 + 2a \cdot \pi - \pi - a$ (from Equation 5), we have

$$\begin{aligned} P(X = 0) &= 1 - p \\ &= 1 - P(X = 1) \\ &= 1 - (1 + 2a \cdot \pi - \pi - a) \\ &= 1 - 1 - 2a \cdot \pi + \pi + a \\ &= -2a \cdot \pi + \pi + a \end{aligned} \tag{7}$$

where Equation 7 matches the expression in Equation 6. Therefore, we have shown that X_i follows a Bernoulli distribution with parameter $p = 1 + 2a \cdot \pi - \pi - a$.

Following from this result, because the random variable m is the sum of n i.i.d. Bernoulli random variables, we know that m follows a Binomial distribution with parameters n and p. Hence, the likelihood of observing m ones in n samples is given by the probability density function of the Binomial distribution:

$$\Pr(X = m) = \binom{n}{m} p^m (1-p)^{n-m} \, . \tag{8}$$

While Equation 8 differs from the likelihood function we derived in Equation 1, we can see that the $\binom{n}{m}$ term cancels out when deriving the log likelihood; First, the log likelihood transforms the multiplication into a sum with a log term, and then when deriving w.r.t. π_{ML}, the $\log \binom{n}{m}$ becomes zero and therefore disappears. Hence, we end up with the same expression for the log likelihood.

As the Binomial distribution is well-known, we know that the maximum likelihood estimator of p of the Binomial distribution is given by $p_{ML} = \frac{m}{n}$, which is the proportion of successes, i.e., the number of $X_i = 1$ in n samples.[1]

Finally, to derive the maximum likelihood estimator of π_{ML}, we equate p_{ML} and $P(X_i = 1)$ (as $P(X_i = 1) = p$), and then solve for π_{ML}:

$$\begin{aligned} P(X_i = 1) &= p_{ML} \\ 1 + 2a \cdot \pi_{ML} - \pi_{ML} - a &= \frac{m}{n} \\ 2a \cdot \pi_{ML} - \pi_{ML} &= \frac{m}{n} + a - 1 \\ \pi_{ML} \cdot (2a - 1) &= \frac{m}{n} + a - 1 \\ \pi_{ML} &= \frac{\frac{m}{n} + a - 1}{(2a - 1)} \, . \end{aligned} \tag{9}$$

We clearly observe that Equation 9 matches Equation 4, hence, we found the same maximum likelihood estimator of π_{ML} as in (c).

(e) *Show that π_{ML} is an unbiased estimator of π.*

We aim to show that $\mathbb{E}[\pi_{ML}] = \pi$. To this end, we will make use of the knowledge introduced in (d), namely

[1] See `https://en.wikipedia.org/wiki/Binomial_distribution#Estimation_of_parameters`.

that X_i follows a Bernoulli distribution.

We have

$$\begin{aligned}\mathbb{E}\left[\pi_{ML}\right] &= \mathbb{E}\left[\frac{\frac{m}{n}+a-1}{(2a-1)}\right] \\ &= \mathbb{E}\left[\frac{\frac{\sum_{i=1}^{n} X_i}{n}+a-1}{(2a-1)}\right] \\ &= \frac{\sum_{i=1}^{n}\mathbb{E}[X_i]+a\cdot n-1\cdot n}{(2a-1)\cdot n}\ . \end{aligned} \tag{10}$$

We now compute the expectation of X_i:

$$\begin{aligned}\mathbb{E}[X_i] &= P(X_i=1)\cdot 1 + P(X_i=0)\cdot 0 \\ &= P(X_i=1) \\ &= a\cdot\pi + (1-a)\cdot(1-\pi) \\ &= (1-a)+(2a-1)\cdot\pi\ . \end{aligned} \tag{11}$$

Finally, we can substitute Equation 11 into Equation 10 to get

$$\begin{aligned}\mathbb{E}\left[\pi_{ML}\right] &= \frac{\sum_{i=1}^{n}\mathbb{E}[X_i]+a\cdot n-1\cdot n}{(2a-1)\cdot n} \\ &= \frac{\sum_{i=1}^{n}\left((1-a)+(2a-1)\cdot\pi\right)+a\cdot n-1\cdot n}{(2a-1)\cdot n} \\ &= \frac{n\cdot\left((1-a)+(2a-1)\cdot\pi\right)+a\cdot n-1\cdot n}{(2a-1)\cdot n} \\ &= \frac{\left((1-a)+(2a-1)\cdot\pi\right)+a-1}{(2a-1)} \\ &= \frac{-(a-1)+(2a-1)\cdot\pi+(a-1)}{(2a-1)} \\ &= \frac{(2a-1)\cdot\pi}{(2a-1)} \\ &= \pi\ . \end{aligned} \tag{12}$$

We have shown that the expected value of π_{ML} is equal to the true value of π (Equation 12). Therefore, we can conclude that our estimator for π_{ML} is unbiased.

(f) *Derive an expression for the variance of π_{ML}. Analyse its dependence on a.*

We have

$$\begin{aligned}
\mathbb{V}[\pi_{ML}] &= \mathbb{V}\left[\frac{\frac{m}{n}+a-1}{(2a-1)}\right] \\
&= \mathbb{V}\left[\frac{m}{n}+a-1\right]\cdot\frac{1}{(2a-1)^2} \\
&= \mathbb{V}\left[\frac{1}{n}\left(m+a\cdot n-1\cdot n\right)\right]\cdot\frac{1}{(2a-1)^2} \\
&= \mathbb{V}\left[m+a\cdot n-1\cdot n\right]\cdot\frac{1}{(2a-1)^2\cdot n^2} \\
&= \mathbb{V}\left[\sum_{i=1}^{n}X_i+a\cdot n-1\cdot n\right]\cdot\frac{1}{(2a-1)^2\cdot n^2} \\
&= \mathbb{V}\left[\sum_{i=1}^{n}X_i\right]\cdot\frac{1}{(2a-1)^2\cdot n^2} \\
&= \sum_{i=1}^{n}\mathbb{V}\left[X_i\right]\cdot\frac{1}{(2a-1)^2\cdot n^2}\ . \qquad (13)
\end{aligned}$$

We now compute the variance of X_i. We have:

$$\begin{aligned}
\mathbb{E}[X_i^2] &= P(X_i=1)\cdot 1^2+P(X_i=0)\cdot 0^2 \\
&= P(X_i=1) \\
&= (1-a)+(2a-1)\cdot\pi\ , \qquad (14)
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}[X_i]^2 &= \left((1-a)+(2a-1)\cdot\pi\right)^2 \\
&= (1-a)^2+(2a-1)^2\cdot\pi^2+2\cdot(1-a)\cdot(2a-1)\cdot\pi \\
&= 1^2+a^2-2a+4a^2\pi^2+\pi^2-4a\cdot\pi^2+(2-2a)\cdot(2a-1)\cdot\pi \\
&= 1+a^2-2a+4a^2\pi^2+\pi^2-4a\cdot\pi^2+(4a-2-4a^2+2a)\cdot\pi \\
&= 1+\pi\left(6a-4a^2-2\right)+\pi^2(4a^2-4a+1)-2a+a^2\ . \qquad (15)
\end{aligned}$$

Therefore, the variance of X_i is given by

$$\begin{aligned}
V(X_i) &= \mathbb{E}\left[X_i^2\right] - E[X_i]^2 \\
&= \pi(2a-1) + 1 - a - \left(1 + \pi\left(6a - 4a^2 - 2\right) + \pi^2(4a^2 - 4a + 1) - 2a + a^2\right) \\
&= \pi(2a-1) + 1 - a - 1 - \pi\left(6a - 4a^2 - 2\right) - \pi^2(4a^2 - 4a + 1) + 2a - a^2 \\
&= \pi(2a-1) - \pi\left(6a - 4a^2 - 2\right) - \pi^2(4a^2 - 4a + 1) + a - a^2 \\
&= \pi\left(2a - 1 - (6a - 4a^2 - 2)\right) - \pi^2(4a^2 - 4a + 1) + a - a^2 \\
&= \pi\left(2a - 1 - 6a + 4a^2 + 2\right) - \pi^2(4a^2 - 4a + 1) + a - a^2 \\
&= \pi\left(1 - 4a + 4a^2\right) - \pi^2(1 - 4a + 4a^2) + a - a^2 \\
&= \left(1 - 4a + 4a^2\right) \cdot (\pi - \pi^2) + a - a^2 \,. \qquad (16)
\end{aligned}$$

Finally, to compute the variance of the estimator of π_{ML}, we plug Equation 16 into Equation 13 and get

$$\begin{aligned}
\mathbb{V}[\pi_{ML}] &= \sum_{i=1}^{n} \mathbb{V}\left[X_i\right] \cdot \frac{1}{(2a-1)^2 \cdot n^2} \\
&= \sum_{i=1}^{n} \left(\left(1 - 4a + 4a^2\right) \cdot (\pi - \pi^2) + a - a^2\right) \cdot \frac{1}{(2a-1)^2 \cdot n^2} \\
&= n \cdot \left(\left(1 - 4a + 4a^2\right) \cdot (\pi - \pi^2) + a - a^2\right) \cdot \frac{1}{(2a-1)^2 \cdot n^2} \\
&= \left(\left(1 - 4a + 4a^2\right) \cdot (\pi - \pi^2) + a - a^2\right) \cdot \frac{1}{(2a-1)^2 \cdot n} \\
&= \left((2a-1)^2 \cdot (\pi - \pi^2) + a - a^2\right) \cdot \frac{1}{(2a-1)^2 \cdot n} \\
&= \frac{(2a-1)^2 \cdot (\pi - \pi^2)}{(2a-1)^2 \cdot n} + \frac{a - a^2}{(2a-1)^2 \cdot n} \\
&= \frac{\pi - \pi^2}{n} + \frac{a - a^2}{(2a-1)^2 \cdot n} \,. \qquad (17)
\end{aligned}$$

We correctly observe that the variance is undefined for $a = 0.5$. Furthermore, we observe that the variance decreases as n increases, which is to be expected as we get more evidence with more data. Lastly, we observe that whenever a approaches either 0 or 1, the variance decreases. The same phenomenon is observed for the true value of π.

Figure 1 shows the log-scaled variance of π_{ML} as a function of a and π, which clearly displays the behavior previously described. In short, our monte-carlo simulation confirms that the variance of the estimator of π_{ML} is maximal whenever a approaches 0.5, and furthermore increased as π approaches 0.5. This makes sense, as whenever $a = 0.5$ (or close), there is very little information gained from the answer of the users, a yes could be from both groups. Hence, the variance of the estimator will be very large in this case. Furthermore, whenever one group is much larger than the other (i.e., $\pi \to 0$ or $\pi \to 1$), we exect the answers to be more consistent, and hence the estimation of π will be more accurate. Thus, the variance decreases as π approaches the extremes of 0 and 1.

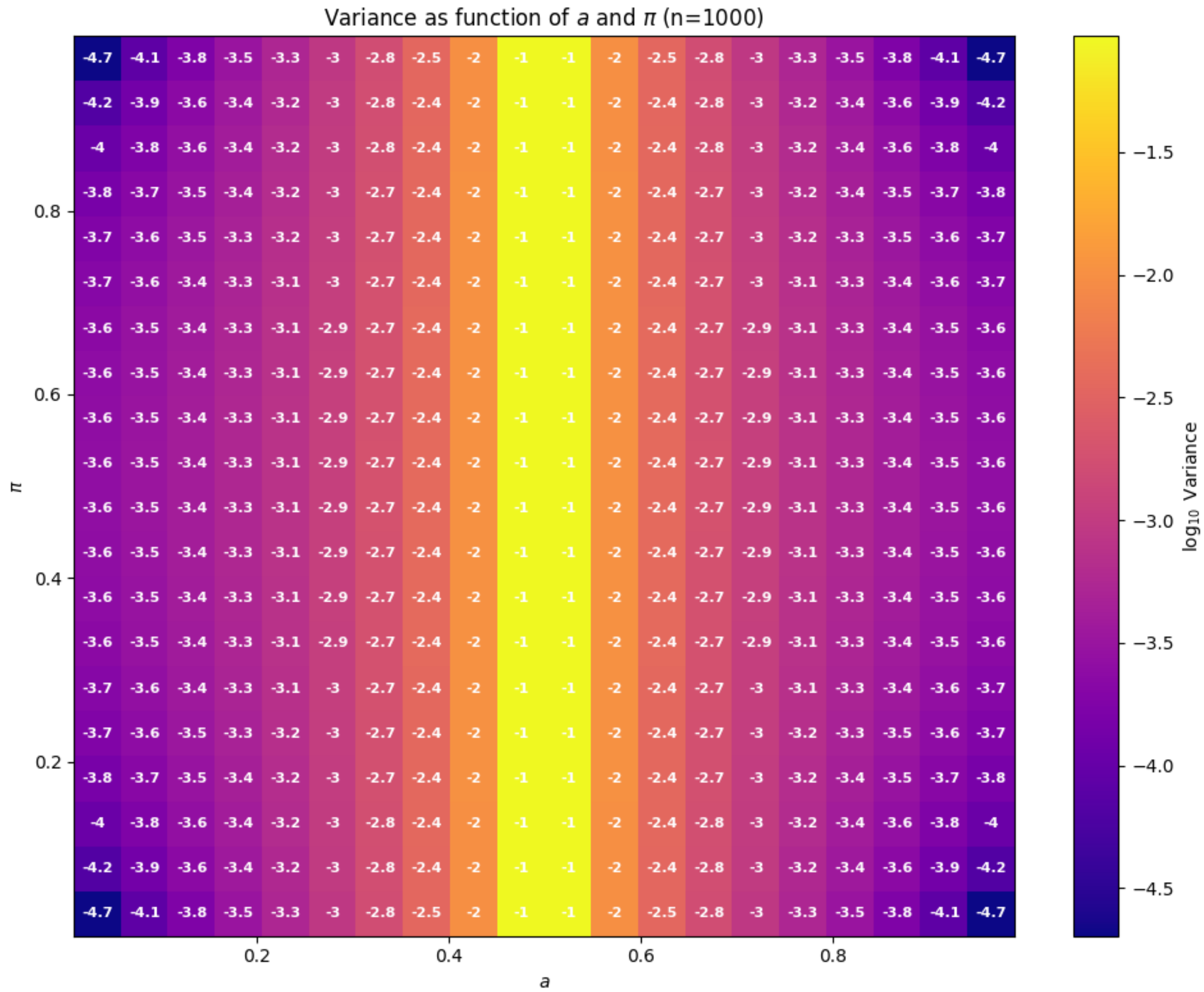


Figure 1: Variance of π_{ML} as function of a and π for different values of n. Here, the values are log-scaled for better display. Note that we restrict $a \in [0.01, 0.99]$ and $\pi \in [0.01, 0.99]$ for the plot. Hence, very low values are expected whenever a or π exceeds these bounds.

(g) *Give an expression for the bias of this naive estimator. Is it asymptotically unbiased?*

We aim to recover the value of π from direct questioning. However, persons belonging to the sensitive group will lie with probability ℓ, whereas persons belonging to the non-sensitive group will answer truthfully and therefore lie with probability 0.

Naivily, we could estimate the value of π as the fraction of persons answering yes to the question:

$$\pi_{\text{naive}} = \frac{\sum_{i=1}^{n} X_i}{n} \ .$$

We can model the whether a person belongs to the sensitive group as a hidden variable $S_i \in \{0, 1\}$ with $S_i = 1$ describing that the person belongs to the sensitive group, and $S_i = 0$ describes that a person belongs to the non-sensitive group.

Naturally, we have that

$$P(S_i = 1) = \pi$$
$$P(S_i = 0) = 1 - \pi$$

which is distributed according to a Bernoulli distribution with parameter $p = \pi$.

We can now express the probability of a person answering yes or no to the question by marginalizing over the hidden variable S_i:

$$\begin{aligned} P(X_i = 1) &= P(X_i = 1, S_i = 1) + P(X_i = 1, S_i = 0) \\ &= \pi \cdot (1 - \ell) + 0 \\ P(X_i = 0) &= P(X_i = 0, S_i = 1) + P(X_i = 0, S_i = 0) \\ &= \pi \cdot \ell + (1 - \pi) \, . \end{aligned}$$

We verify that the probabilities are correct by adding both $P(X_i = 1)$ and $P(X_i = 0)$ and getting 1:

$$\begin{aligned} P(X_i = 1) + P(X_i = 0) &= \pi \cdot (1 - \ell) + \pi \cdot \ell + (1 - \pi) \\ &= \pi(1 - \ell + \ell) + 1 - \pi \\ &= \pi(1) + 1 - \pi \\ &= 1 \, . \end{aligned}$$

We now compute the expeced value of the naive estimator of π to check whether the estimator is unbiased. We have

$$\begin{aligned} \mathbb{E}[\pi_{\text{naive}}] &= \mathbb{E}\left[\frac{\sum_{i=1}^{n} X_i}{n}\right] \\ &= \frac{1}{n} \cdot \sum_{i=1}^{n} \mathbb{E}[X_i] \\ &= \frac{1}{n} \cdot \sum_{i=1}^{n} \left(P(X_i = 1) \cdot 1 + P(X_i = 0) \cdot 0\right) \\ &= \frac{1}{n} \cdot \sum_{i=1}^{n} \left(P(X_i = 1)\right) \\ &= \frac{1}{n} \cdot \sum_{i=1}^{n} \left(\pi \cdot (1 - \ell)\right) \\ &= \frac{1}{n} \cdot n \left(\pi \cdot (1 - \ell)\right) \\ &= \pi \cdot (1 - \ell) \, . \end{aligned} \tag{18}$$

We can now use the expected value of π_{naive} to compute the bias of the estimator:

$$\begin{aligned}\text{Bias}(\pi_{\text{naive}}, \pi) &= \left[\mathbb{E}[\pi_{\text{naive}}] - \pi\right]^2 \\ &= \left[\pi \cdot (1-\ell) - \pi\right]^2 \\ &= \left[\pi - \pi\ell - \pi\right]^2 \\ &= \left[-\pi\ell\right]^2 \\ &= \pi^2\ell^2 \ . \end{aligned} \tag{19}$$

Therefore, only whenever $\ell = 0$, i.e., no one lies, is the naive estimator of π is unbiased. Otherwise, the estimator is biased towards and goes towards zero bias whenever either ℓ or π goes to zero.

(h) *Derive expressions for the mean squared error (MSE) of the randomized response estimator and the naive estimator. Analyse the expressions to draw conclusions about when, depending on the relevant parameters (π, a, ℓ, n), one estimator should be favored over the other.*

We aim to compute the mean squared error of our two estimators of π.

We use the Bias Variance decomposition to decompose the MSE into the bias and variance terms.[2] Therefore, we simplify the MSE into

$$\begin{aligned}\text{MSE}(\hat{\pi}, \pi) &= \mathbb{E}\left[\left(\hat{\pi} - \mathbb{E}[\hat{\pi}]\right)^2\right] + \left[\mathbb{E}[\hat{\pi}] - \pi\right]^2 \\ &= \mathbb{V}\left[\hat{\pi}\right] + \left[\mathbb{E}[\hat{\pi}] - \pi\right]^2 \\ &= \mathbb{V}\left[\hat{\pi}\right] + \text{Bias}(\hat{\pi}, \pi)\end{aligned}$$

Therefore, we have

[2] See `https://en.wikipedia.org/wiki/Mean_squared_error#Proof_of_variance_and_bias_relationship`.

$$
\begin{aligned}
\mathbb{V}[\pi_{\text{naive}}] &= \mathbb{V}\left[\frac{\sum_{i=1}^{n} X_i}{n}\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}[X_i] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}[X_i^2] - \mathbb{E}[X_i]^2 \\
&= \frac{1}{n^2}\sum_{i=1}^{n}[P(X_i=1)] - (P(X_i=1))^2 \\
&= \frac{1}{n^2}\sum_{i=1}^{n}(\pi\cdot(1-\ell)) - (\pi\cdot(1-\ell))^2 \\
&= \frac{n}{n^2}\left((\pi\cdot(1-\ell)) - (\pi\cdot(1-\ell))^2\right) \\
&= \frac{\pi\cdot(1-\ell) - \pi^2\cdot(1-\ell)^2}{n}
\end{aligned} \tag{20}
$$

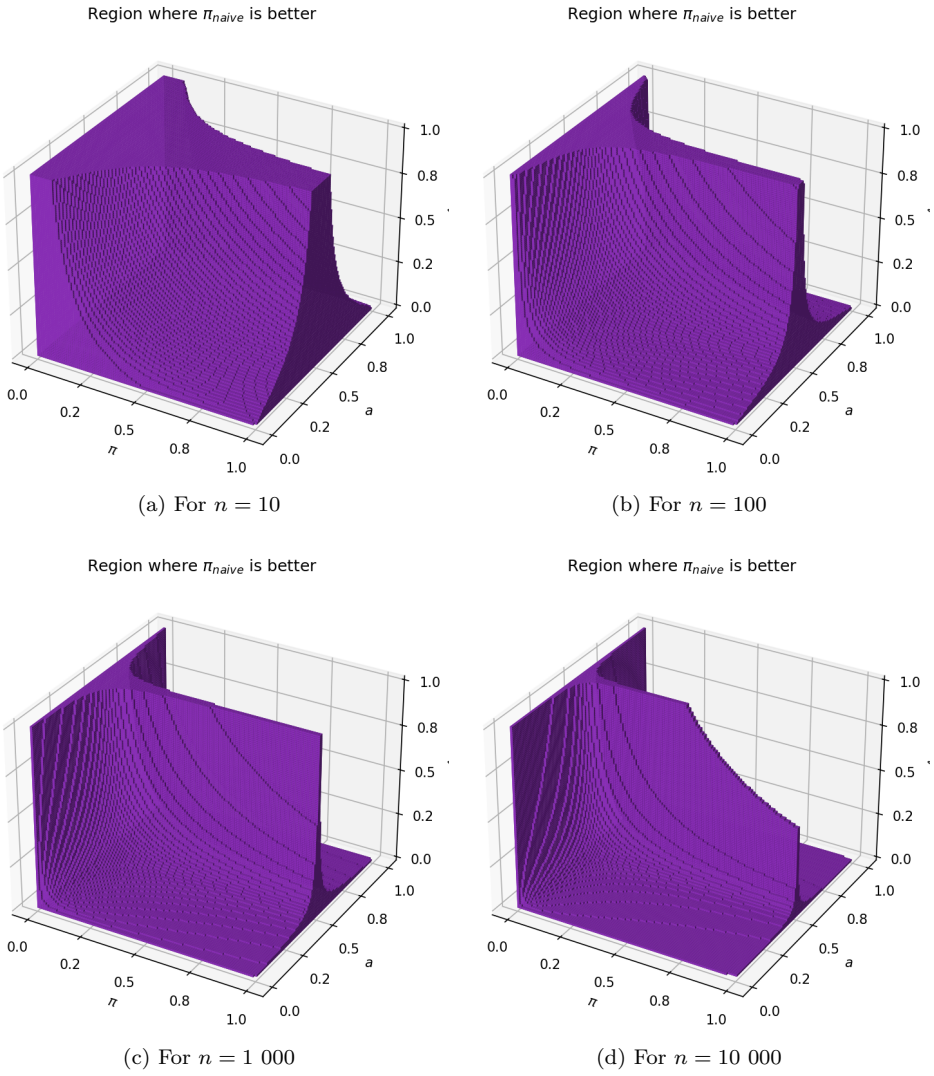
$$
\begin{aligned}
\text{MSE}(\pi_{\text{naive}}, \pi) &= \mathbb{V}[\pi_{\text{naive}}] + \text{Bias}(\hat{\pi}, \pi) \\
&\overset{(20)}{=} \frac{\pi\cdot(1-\ell) - \pi^2\cdot(1-\ell)^2}{n} + \text{Bias}(\hat{\pi}, \pi) \\
&\overset{(19)}{=} \frac{\pi\cdot(1-\ell) - \pi^2\cdot(1-\ell)^2}{n} + \pi^2\ell^2 \,.
\end{aligned}
$$

Because the π_{ML} is an unbiased estimator of π, we have that $\text{Bias}(\pi_{ML}, \pi) = 0$. Therefore, we have

$$
\begin{aligned}
\text{MSE}(\pi_{\text{ML}}, \pi) &= \mathbb{V}[\pi_{\text{ML}}] + \text{Bias}(\pi_{\text{ML}}, \pi) \\
&= \mathbb{V}[\pi_{\text{ML}}] \\
&\overset{(17)}{=} \frac{\pi - \pi^2}{n} + \frac{a - a^2}{(2a-1)^2\cdot n} \,.
\end{aligned}
$$

First, we observe that whenever no-one lies ($\ell = 0$), the left side of the MSE of π_{naive} is equal to the left side of the MSE of π_{ML}. However, as the right side of the MSE of π_{naive} is multiplied with ℓ^2, the MSE is zero for $\ell = 0$, whereas the MSE of π_{ML} is non-zero, given $a \notin \{0, 1\}$. If $a \in \{0, 1\}$, then both estimators are equal w.r.t. the MSE.

Due to the complex expressions, finding the cut-off points for when both MSE expressions are equal — and therefore finding whenever one estimator is better than the other analytically — is very complicated. While we have explained the trivial edge cases, whenever we have $a \notin \{0, 1\} \land \pi \notin \{0, 1\}$, we have no easy way to determine when which estimator is better than the other. To overcome this issue, we have computed the expected MSE for both estimators for a range of values for a, π, ℓ, and n and show the results in Figure 2. Figure 2 clearly shows the complex, non-linear relationship between both MSE expression.



(a) For $n = 10$

(b) For $n = 100$

(c) For $n = 1\ 000$

(d) For $n = 10\ 000$

Figure 2: Plots of the parameter space of (π, a, ℓ) for various sample sizes n. The purple voxels indicate for which parameters the MSE of π_{naive} is lower than the MSE of π_{ML}.

First, it is clear that no single best estimator exists for all parameters. The shape of the region where π_{naive} has a lower MSE than π_{ML} is quite clear: whenever $\ell = 0$, the naive estimator is better than the ML estimator. Furthermore, whenever $\pi = 0$, the naive estimator is also better. Both cases make sense, as whenever no one lies, the naive estimator is the better estimator of the two.

Figure 2 also shows that whenever a is close or equal to 0.5, the naive estimator is the better solution. Note that while our plot shows that in the subplot (d), for some parameters which include $a = 0.5$, the ML estimator is to be favored. This is not true, as it is simply an artefact of the plotting; We use only 100 points for each axis, and therefore the voxels are quite large. With increased resolution, the same plot would show that the naive estimator is the better solution for these parameters.

Lastly, for all other cases that we did not mention, the ML estimator is the better solution for large n. Refer to Figure 2 for the exact shape. This is logical, as the MSE of the naive estimator has a part independent

of n, whereas the MSE of the ML estimator is entirely divided by n. Thus, for large n, the ML estimator will go towards zero, whereas the naive estimator will remain constant.

Problem 2

		$N = 500$	$N = 1000$	$N = 5000$
$\hat{\mathbf{N}}$		590.45±311	1096.62±286	5152.14±583
Error (%)		18.09	9.66	3.04
$\hat{\sigma}_{\hat{N}}$		239.58±258	284.17±116	566.36±98
C.I.	$\alpha = 0.05$	[299.68±85; 1322.77±1298]	[687.34±137; 1842.05±609]	[4179.35±423; 6414.78±804]
	$\alpha = 0.01$	[250.36±59; 1740.30±2079]	[601.62±110; 2185.30±773]	[3922.46±383; 6884.52±890]
Coverage (%)	$\alpha = 0.05$	94.50±22.80	95.70±20.29	94.70±22.40
	$\alpha = 0.01$	99.40±07.72	99.20±08.91	99.20±08.91
Bias (≈)		8 181	9 335	23 146
Variance (≈)		96 721	81 796	339 889
MSE (≈)		104 902	91 131	363 035

Table 1: Result from 1000 MC simulations for estimating $\hat{N}$. For each value, we report on the mean (black) and standard deviation (gray). We omit the digits after the decimal point for the standard deviation for brevity. The Error (%) is the absolute difference between the estimated and the true value of N, divided by the true value of N. Hence, it measures the relative error of the estimated value to the true value.

Observation. We implement the formulas and run the Monte-Carlo simulations for 1000 runs and report the results in Table 1. We first observe that, although the absolute error between the estimated and the true value of N increases with N, the relative error decreases when the true N is larger.

Interestingly, the given formula proposed for estimating the variance of $\hat{N}$ is very accurate, as its average value is very close to the computed variance of the Monte-Carlo simulation — in our case, we report on the standard deviation, which is simply the square root of the variance.

Lastly, we computed the averaged confidence intervals across all runs for $\alpha = 0.05$ and $\alpha = 0.01$. Naturally, the confidence intervals when using $\alpha = 0.01$ are larger than when using $\alpha = 0.05$, we notice that the standard deviation of the lower confidence decreases when increasing the size of the Confidence Interval (C.I.), whereas the standard deviation of the upper-confidence increases when decreasing α.

Discussion.

the Biase, Variance, and MSE are computed on the aggregated values from the table manually. Computing them from the true values does slightly change the results (slight increase in MSE), however, the relationship between the three cases ($N = 500$, $N = 1000$, $N = 5000$) remain the same.

2 Part B: Data Analysis

2.1 Data preprocessing and experimental setup We combined two Equinox datasets containing change metrics and CK/OO metrics and used the class identifier `classname` as the join key. Defect-related metadata columns were removed from the feature set, while `bugs` was retained as the target variable.

We merged two datasets using a one-to-one inner join on `classname`. After merging, all feature columns and the target were converted to numeric values. The final dataset contains **324 classes**, **32 features**, and one target variable.

The dataset was split into a training set (67%) and a test set (33%). To preserve the distribution of bug counts, we used stratified random sampling based on the predefined bug-count bins (0, 1–2, 3–5, 6+). The stratification variable was used only for splitting and removed afterwards.

The resulting training and test sets were saved and used consistently in all analyses. Models were selected on the training set and the final selected model was evaluated on the test set.

2.2 Poisson regression

2.2.1 Stepwise model selection We applied Poisson regression to model the number of bugs. Since the response variable is count data, a generalized linear model with Poisson family and log link was used. Model selection was performed on the training set using **forward stepwise selection**, starting from an intercept-only model and using **AIC** as the selection criterion.

The final model selected by stepwise regression includes the following five predictors:

$$\text{bugs} \sim \textit{cbo} + \textit{numberOfMethodsInherited} + \textit{weightedAgeWithRespectTo} + \textit{avgLinesAddedUntil} + \textit{noc}$$

The table below shows the estimated coefficients of the selected Poisson regression model, together with standard errors, z-statistics, p-values, and 95% confidence intervals.

Variable	Coefficient	Std. Error	z	p-value	95% CI
Intercept	-1.9665	0.189	-10.388	< 0.001	[-2.337, -1.595]
cbo	0.0454	0.004	12.728	< 0.001	[0.038, 0.052]
numberOfMethodsInherited	0.0167	0.003	5.249	< 0.001	[0.010, 0.023]
weightedAgeWithRespectTo	0.0117	0.002	5.241	< 0.001	[0.007, 0.016]
avgLinesAddedUntil	0.0084	0.003	3.097	0.002	[0.003, 0.014]
noc	0.1789	0.092	1.945	0.052	[-0.001, 0.359]

All selected predictors have positive coefficients, indicating that higher structural complexity or more extensive code evolution is associated with a higher expected number of bugs. Most variables are statistically significant at the 5% level, while `noc` shows a borderline effect.

The predictive performance of the selected model was evaluated on the test set using Poisson deviance:

- **Deviance:** 0.81

2.2.2 LASSO regularization To model the number of bugs, we applied Poisson regression with L1 (LASSO) regularization. All predictors were standardized prior to model fitting, which is necessary for LASSO to penalize coefficients in a comparable way. The regularization parameter λ was selected on the training set using 5-fold cross-validation, with **Poisson deviance** as the evaluation criterion.

Cross-validation selected the following value:

- **Selected λ:** 0.166

Under the selected regularization strength, the LASSO Poisson model retained only a single non-zero coefficient:

- **Selected feature:** `cbo`

The final LASSO-selected model was evaluated on the test set:

- **Deviance:** 1.279

This deviance value is higher than that obtained by the stepwise Poisson model, indicating a trade-off between model simplicity and predictive accuracy.

2.2.3 Comparison of stepwise and LASSO The stepwise-selected model includes five predictors and achieves a lower test Poisson deviance (0.81), indicating better predictive performance. In contrast, Poisson LASSO yields a much sparser model with only one predictor but results in a higher deviance (1.279). For this dataset, stepwise selection appears more suitable when predictive accuracy is the primary goal.

2.3 Logistic regression

2.3.1 Stepwise model selection We constructed a binary target variable `bugbin`, where `bugbin` $= 1$ if `bugs` ≥ 1 and 0 otherwise. A logistic regression model was fitted and forward stepwise selection using AIC was performed on the training set.

The final model includes the following predictors:

- `cbo`, `numberOfMethodsInherited`, `weightedAgeWithRespectTo`, `numberOfRefactoringsUntil`, `avgLinesRemovedUnti`
  `noc`, `dit`

The final model was evaluated on the test set:

- **Accuracy:** 0.720

2.3.2 LASSO regularization In the `sklearn` implementation, the regularization strength is controlled by parameter C (with $\lambda = 1/C$). Cross-validation selected a large C, resulting in weak regularization and a model retaining many predictors.

Variable	Coef.	Std. Err.	z	p-value	95% CI
Intercept	-3.8173	0.629	-6.069	< 0.001	[-5.050, -2.585]
cbo	0.1191	0.025	4.845	< 0.001	[0.071, 0.167]
numberOfMethodsInherited	0.0296	0.021	1.440	0.150	[-0.011, 0.070]
weightedAgeWithRespectTo	0.0168	0.006	2.894	0.004	[0.005, 0.028]
numberOfRefactoringsUntil	-1.3044	0.525	-2.486	0.013	[-2.333, -0.276]
avgLinesRemovedUntil	0.0643	0.024	2.675	0.007	[0.017, 0.111]
noc	0.7458	0.345	2.161	0.031	[0.069, 1.422]
dit	0.8081	0.416	1.944	0.052	[-0.007, 1.623]

The test-set performance was:

- **Accuracy:** 0.729

2.3.3 Comparison of stepwise and LASSO Both approaches achieved similar classification performance. Stepwise selection yields a more interpretable and compact model, while logistic LASSO slightly improves accuracy at the cost of reduced sparsity.

Statement on the Useage of Gen. AI

- **Felix Céard-Falkenberg**
  Unless stated otherwise, all code, text, and math derivations have been entirely thought and written by me with no external help — both for gen. AI and wolframalpha anc co.

  We used Claude 4.5 Opus for generating the plots for Figure 2 as the plot is not a core of the assignment, and only aids in understanding our own results. Note that we (heavily) modified the code generated by Claude.

- **Yiquan Hu**
  For Part B, we used ChatGPT to generate required code for stepwise selection with the permissions stated in the assignment instructions and also consulted ChatGPT for general suggestions on parameter adjustment. All analysis, results and interpretations were produced by us. We only used ChatGPT lightly refine the academic word of some sentences without adding new content.