Part One

Task 1

The company Fränzi and Friends developed a 2nd-generation quick test at home for SARS-Cov-2, which is pending regulatory agency's review. The test has been shown to have a sensitivity of 99.5% and a specificity of 99.5%. Suppose that Fred uses the test by Fränzi and Friends and the test was positive. Assume that 5% of the population is in fact infected. Was is your guess about the probability that Fred is indeed infected?

To determine the probability that Fred is indeed infected given that he tested positive, we can use Bayes' theorem.

$$\begin{aligned}
&\text{Given:} \\
S_n &= 99.5\% = 0.995 \text{ Sensitivity (true positive rate)} \\
S_p &= 99.5\% = 0.995 \text{ Specificity(true negative rate)} \\
P(\text{Infected}) &= 5\% = 0.05 \text{ Prevalence (prior probability)}
\end{aligned}$$

$$\begin{aligned}
&\text{We want to find:} \\
P(\text{Infected} \mid \text{Positive}) &= \frac{P(\text{Positive} \mid \text{Infected}) \cdot P(\text{Infected})}{P(\text{Positive})}
\end{aligned}$$

$$\begin{aligned}
&\text{Calculate } P(\text{Positive}) : \\
P(\text{Positive}) &= P(\text{Positive} \mid \text{Infected}) \cdot P(\text{Infected}) + P(\text{Positive} \mid \text{Not Infected}) \cdot P(\text{Not Infected}) \\
P(\text{Positive} \mid \text{Infected}) &= S_n = 0.995 \\
P(\text{Positive} \mid \text{Not Infected}) &= 1 - S_p = 1 - 0.995 = 0.005 \\
P(\text{Not Infected}) &= 1 - P(\text{Infected}) = 1 - 0.05 = 0.95 \\
P(\text{Positive}) &= 0.995 \cdot 0.05 + 0.005 \cdot 0.95 \\
P(\text{Positive}) &= 0.04975 + 0.00475 = 0.0545
\end{aligned}$$

$$\begin{aligned}
&\text{Apply Bayes' theorem:} \\
P(\text{Infected} \mid \text{Positive}) &= \frac{0.995 \cdot 0.05}{0.0545} \\
P(\text{Infected} \mid \text{Positive}) &= \frac{0.04975}{0.0545} \\
P(\text{Infected} \mid \text{Positive}) &\approx 0.9128
\end{aligned}$$

Therefore, the probability that Fred is indeed infected given that he tested positive is approximately 91.28%.

Task 2

```
import numpy as np
import matplotlib.pyplot as plt

# Function to calculate the probability using Bayes' theorem
def calculate_posterior_prob(sensitivity, specificity, prevalence):
    prevalence = prevalence / 100
    false_positive_rate = 1 - specificity
    P_positive = (sensitivity * prevalence) + (false_positive_rate * (1 - prevalence))
    P_infected_given_positive = (sensitivity * prevalence) / P_positive
    return P_infected_given_positive * 100

# Parameters
sens = 99 / 100
prevalence_range = np.linspace(0.001, 50, 500)  # from 0.001% to 50%
specificities = [99 / 100, 99.9 / 100, 99.99 / 100, 99.999 / 100]

# Plotting
plt.figure(figsize=(12, 8))

for spec in specificities:
    probabilities = [calculate_posterior_prob(sens, spec, prev) for prev in
    prevalence_range] plt.plot(prevalence_range,
    probabilities, label=f'Specificity = {spec * 100}%')

plt.xlabel('Infection Prevalence (%)')
plt.ylabel('Probability of Infection Given Positive Test (%)')
plt.title('Probability of Infection Given Positive Test as a Function of
Prevalence and Specificity')
plt.legend()
plt.grid(True)
plt.show()
```

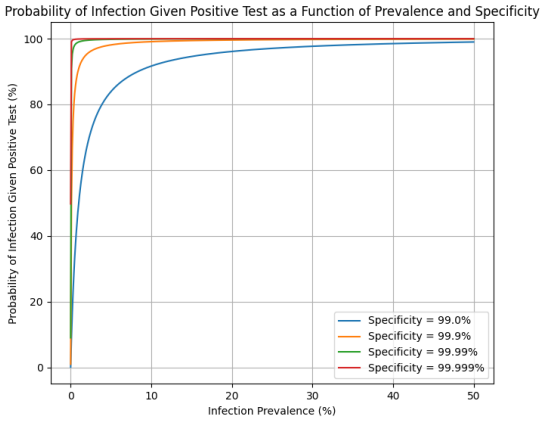

Figure 1: Caption

Task 3

What are you interpretations of the results?

The graph shows the probability that Fred is indeed infected given a positive test result, plotted against the prevalence of the infection in the population, for different specificities of the test.

1. High Specificity Results in High Posterior Probability:
   (a) As specificity increases, the probability that Fred is indeed infected given a positive test result approaches 100% very quickly, even for low prevalence rates. This is because a higher specificity means a lower false positive rate, making a positive test result more reliable.
   (b) For example, with a specificity of 99.999%, the probability of being infected given a positive test result is almost 100% immediately, even if the prevalence is very low. This indicates that the test is extremely reliable at ruling out false positives.
2. Effect of Prevalence on Posterior Probability:
   (a) For lower specificities (like 99%), the probability of being infected given a positive test result reaches 100% only when the prevalence is relatively higher. This is because, with lower specificity, there are more false positives, making the test result less reliable at low prevalence rates.
   (b) As prevalence increases, the posterior probability increases for all specificities, but the rate at which it reaches 100% is faster for higher specificities.

Higher specificity significantly enhances the reliability of a positive test result, especially in low prevalence settings. For public health strategies and individual decision-making, using tests with higher specificities can greatly reduce the likelihood of false positives, providing more accurate assessments of infection status.

Part Two

Cao and Moult (BMC Genomics, 2014) reported studied overlap between drug targets and GWAS hits

Task 1

Assuming we know nothing about a gene (let's call it gene WKN1), what is the probability that the gene is a target for a disease listed here?

Table 1 indicates that there are 24.0 drug targets on average for each of the 88 diseases surveyed. However, we also need the total number of unique drug target genes to calculate the overall probability.

To simplify the calculation, if we assume that the 856 drug target genes mentioned in the text (without considering overlap between diseases) represent unique genes, the calculation for the probability p that a random gene (like WKN1) is a drug target is given by:

$$p = \frac{\text{Number of unique drug target genes}}{\text{Total number of human genes}}$$

Using an estimated total number of human genes as approximately 20,000, we calculate:

$$p = \frac{856}{20,000} = 0.0428$$

Task 2

Assuming that we know nothing about another gene WKN2but that it is a GWAS hit for a disease, what is the probability that WKN2 is a target for that disease?

To find the probability that a gene WKN2, known to be a GWAS hit for a disease, is a target for that disease: We'll specifically look at the relationship between the number of GWAS reported genes and the number of those that are also drug targets for the same disease.

From Table 1, we observe the following summary: - There are a total of 23 instances where GWAS reported genes are also drug targets for the same disease. - There are 88 diseases listed with GWAS reported genes totaling 2,568 (the sum of GWAS reported genes across all diseases, $\sum_{i=1}^{88} \text{GWAS reported genes}_i$).

To calculate the probability p that a GWAS hit (like WKN2) is also a drug target for that same disease, we would use the proportion of GWAS reported genes that are drug targets:

$$p = \frac{\text{Total instances where GWAS genes are drug targets for the same disease}}{\text{Total GWAS reported genes}}$$

Using the given numbers:

$$p = \frac{23}{2568} \approx 0.00895638629283489$$

The probability that a gene known to be a GWAS hit (like WKN2) for a disease is also a target for that disease is approximately 0.90%.