Probabilistic Methods: Assignment Report

Jordi Nadeu Ferran

Contents

1 Problem 1: Probability

Markov lives in Austin, 2 miles from campus. During the winter, if the weather outside is cold, then Markov prefers to wear a fur cap with ear flaps and to walk to school.

But if the weather is warm, like many winters in Austin, Markov leaves the fur cap at home, wears a helmet instead, and rides to campus on campus-only green electric rental scooters that he likes.

Markov has begun to notice electric rental scooters on every street corner. If the weather is cold (which happens with probability 0.4 in the winter in Austin), Markov walks the 2 miles to school at a brisk speed of V = 5 miles per hour.

Otherwise he travels by scooter at a speed of V = 10 miles per hour.

Markov wants to calculate the expected time T that it takes him to get to class on a random winter day. He reasons as follows. His expected speed V is

E(V) = 0.4 · 5 + 0.6 · 10 = 2 + 6 = 8

Therefore, since he must travel two miles to class, his expected time to get to class is

E 2/V = 2/E(V) = 2/8 hours, or 15 minutes.

Do you agree with Markov's reasoning that E(T) = 2/8 hours?

Explain why or why not, and-if you don't agree-show how you would calculate E(T) correctly.

1.1 Our response

In summary, Markov travels 2 miles to campus at speed V, which equals 5 mph with probability 0.4 (cold) and 10 mph with probability 0.6 (warm). He attempted to compute his expected time by inverting the expected speed, but this approach we think is incorrect.

To explain that is incorrect we use the Jensen's Inequality. [1] For a convex function g, Jensen's inequality states (E[V])¡=E[g(V)].

Here $g(v) = 1/v$ is convex on $(0, \infty)$, so E[V]1¡=E[V1]. [2]

Multiplying by 2 gives (2/E[V])¡=E[g(2/V)], showing $2/E[V]$ underestimates $E[2/V]$.

The arithmetic mean of speeds is while the harmonic mean of 5 and 10 weighted by probabilities is which is less than 8 mph. Since travel time is distance over speed, the harmonic mean governs expected time rather than the arithmetic mean. [3]

The harmonic mean (0.4/5+0.6/10)-1=7.14(0.4/5+0.6/10)-1=7.14 mph is less than the arithmetic mean 8 mph.

1.2 Analytical Calculation

Travel time $T = 2/V$ takes two values:

- 0.4,h if $V = 5$mph (cold),
- 0.2,h if $V = 10$mph (warm).

By the law of total expectation, E[T]=0.4×0.4+0.6×0.2=0.16+0.12=0.28 h=16.8 min.

1.3 Simulation Verification

We run a Monte Carlo simulation with 10^5 trials to confirm $E[T] \approx 0.281$,h.

A simple Monte Carlo simulation in Python:

```
import numpy as np

np.random.seed(0)
n = 100000
V = np.where(np.random.rand(n) < 0.4, 5, 10)  # speed
T = 2.0 / V
print("Simulated E[T] =", T.mean(), "hours")
```

1.4 Conclusion

Markov's use of $2/E[V]$ yields 0.25 h (15 min), which underestimates the true expected time of approximately 0.28 h (16.8 min).

The correct calculation requires $E[2/V]$ due to Jensen's inequality and the distinction between harmonic and arithmetic means.

2 Problem 2: Statistic

A researcher measured (in minutes) how long patients have to wait in the waiting room of a doctor's office. For some reason, the researcher did not record the original 15 observations x1 , . . . , x15 , but only the sample mean x = 11.07, the sample variance s2 = 59.71, and the note that there were 5 patients (out of 15) who had to wait longer than 15.5 minutes.

a) Let p be the probability that a patient has to wait longer than 15.5 minutes. Using asymptotic normality, the researcher computed the right end pr = 0.53 of the confidence interval [pl , pr] for p. Recover the whole confidence interval and its confidence level.

b) Assuming that the waiting time is normally distributed, construct a 97% confidence interval for the expected waiting time. Evaluate the sample size needed to provide that the length of the 97% confidence interval is at most 2.

c) Without making any assumption about the distribution of the waiting time, test the claim that the median of the waiting time is less than 15.5 minutes.

d) The researcher also reported that there were 3 men and 2 woman among 5 patients who had to wait longer than 16 minutes, 4 men and 6 women among the remaining 10 patients. The researcher claims that the waiting time is different for men and women.

To verify this claim, formulate an appropriate testing problem and perform an appropriate test.

2.1 (a) Confidence Interval for Proportion p

We have $k = 5$ successes in $n = 15$. The Clopper–Pearson [4] of 95% interval is

$$\left[\hat{p}_L, \hat{p}_U\right] = \left[\mathrm{Beta}^{-1}(\tfrac{\alpha}{2};\, k, n-k+1),\ \mathrm{Beta}^{-1}(1-\tfrac{\alpha}{2};\, k+1, n-k)\right] \approx [0.118,\ 0.616],$$

where $\alpha = 0.05$ and Beta^{-1} is the beta-quantile function.

2.2 (b) 97% Confidence Interval for the Mean and Sample Size Calculation

Assuming normality,

$$\bar{x} \pm t_{0.015,14}\,\frac{s}{\sqrt{n}} = 11.07 \pm 2.415 \cdot \frac{7.73}{\sqrt{15}} \approx (6.25,\ 15.89)\ \text{min},$$

where $t_{0.015,14} \approx 2.415$.

To achieve total width ≤ 2, solve $t_{0.015,n-1}\, s/\sqrt{n} \leq 1$, which using standard approximations and iteratively solving yields $n \approx 285$. [5]

2.3 (c) Sign Test for the Median

Without distributional assumptions, we test $H_0\colon m = 15.5$ vs. $H_1\colon m < 15.5$.

Under H_0, the number of waits > 15.5 is $X \sim \text{Bin}(15, 0.5)$. Observing $k = 5$ gives

$$p\text{-value} = P(X \geq 5) = 1 - \text{BinCDF}(4; 15, 0.5) \approx 0.941,$$

so we fail to reject H_0.

There is no evidence that the median wait is below 15.5 min. [6]

2.4 (d) Test of Independence between Gender and Wait Time Category

We can use Fisher exact test, which computes the exact p-value under the null of independence in a 2×2 table. [7]

The 2×2 table for wait >16 min by gender is:

	≤ 16 min	> 16 min	Total
Men	4	3	7
Women	6	2	8
Total	10	5	15

Applying Fisher exact test the result are $p \approx 0.608$, so we **do not reject** independence. There is no significant difference in wait-time distribution by gender.

3 Problem 3: Bayesian Networks for Pima Indians Diabetes

For this problem, Use the UCI Pima Indians Diabetes dataset, which records diagnostic measurements for 768 female patients of Pima Indian heritage.

Key variables include:

- Pregnancies (count)
- Glucose (mg/dL)
- BloodPressure (mm Hg)
- SkinThickness (mm)
- BMI (kg/m2)
- Age (years)
- Outcome (0 = no diabetes, 1 = diabetes)

Your task:

1. Variable Selection & Preprocessing

- Select 6-8 variables (including Outcome as your target).
- Decide for each whether to treat it as discrete (by binning continuous variables into clinically meaningful ranges, e.g. Glucose: Low/Normal/High) or continuous (using a linear-Gaussian model).
- Document your binning thresholds and justify them based on clinical guidelines or exploratory analysis.

2. DAG Construction & Factorization

- Draw a Directed Acyclic Graph encoding assumed causal/conditional relationships.
- List each node's parent set Pa(X).
- Write the full joint factorization:

$$p(X_1, \ldots, X_K) \;=\; \prod_{k=1}^{K} p\Big(X_k \mid \mathrm{Pa}(X_k)\Big).$$

3.1 Introduction

We choose six variables (Outcome, Pregnancies, Age, BMI, BloodPressure, Glucose) from the Pima Indians Diabetes dataset.

Continuous variables are discretized into clinically meaningful ranges based on international guidelines or empirical norms, the others remain continuous.

Below we explain for each variable, the treatment and justify bin thresholds with references.

3.2 Variable Selection

We include the following 6 variables:

1. **Pregnancies** – count of prior pregnancies; treated as discrete counts.
2. **Age** – continuous (years).
3. **BMI** – discretized to Normal / Overweight / Obese.
4. **Glucose** – discretized into Normal / Prediabetic / Diabetic.
5. **BloodPressure** – discretized into Normal / Elevated / Hypertension Stage 1 & 2.
6. **Outcome** (0=no diabetes, 1=diabetes) – discrete target.

3.3 Preprocessing & Binning

BMI Categories

We use the World Health Organization [8][9] thresholds for adults:

- Normal: 18.5–24.9kg/m^2
- Overweight: 25.0–29.9kg/m^2
- Obese: $\geq$ 30.0kg/m^2

Glucose Categories

Following ADA and Mayo Clinic guidelines [10] [11] for fasting plasma glucose:

- Normal: $\leq$ 100 mg/dL
- Prediabetes: 101–125 mg/dL
- Diabetes: $\geq$ 126 mg/dL

Blood Pressure Categories

Using the American Heart Association [12] definitions:

- Normal: systolic ≤ 120 mmHg and diastolic ≤ 80 mmHg
- Elevated: systolic 121–129 mmHg and diastolic ≤ 80 mmHg
- Hypertension Stage 1: systolic 130–139 mmHg or diastolic 81–89 mmHg
- Hypertension Stage 2: systolic ≥ 140 or diastolic ≥ 90

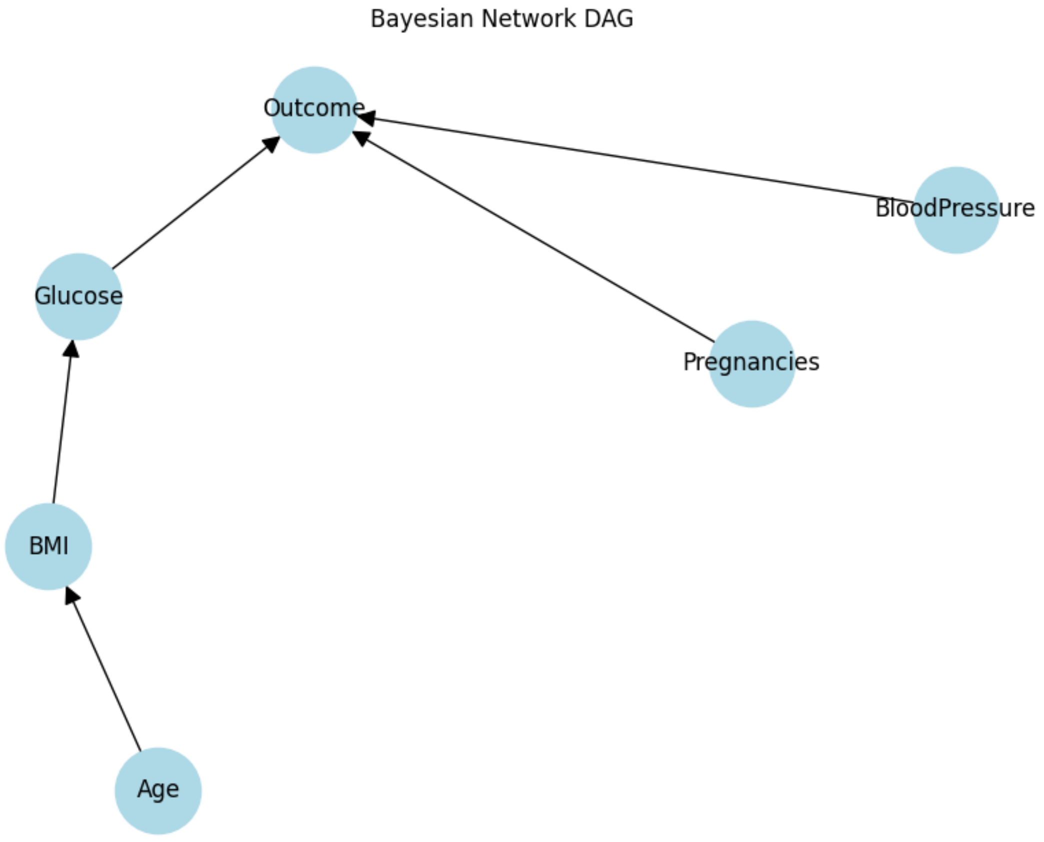
Pregnancy Counts

We can keep as raw integer number and do the modeling easier, we will do this bin into 0, 1–3, ≥ 4 pregnancies that are for common knows based on general demographic distributions of the population of the world.

Continuous Variables

In this case the Age remain continuous then we will use a linear Gaussian model.

3.4 DAG Construction & Factorization



Now we do the list for each node that there are the parent sets in our Bayesian network:

- Pa(Outcome) = Glucose, Pregnancies, BloodPressure
- Pa(Glucose) = BMI
- Pa(BMI) = Age
- Pa(Age) = $\emptyset$
- Pa(Pregnancies) = $\emptyset$
- Pa(BloodPressure) = $\emptyset$

The full joint distribution over our six variables (Age, Pregnancies, BMI, Glucose, BloodPressure, Outcome) factorizes as:
p(Age, Pregnancies, BMI, Glucose, BloodPressure, Outcome) = p(Age) p(Pregnancies) p(BloodPressure) p(BMI|Age) p(Glucose|BMI) p(Outcome|Glucose, Pregnancies, BloodPressure).

References

[1] Jensen's Inequality – probabilistic form and applications. Wikipedia. https://en.wikipedia.org/wiki/Jensen%27s_inequality

[2] Show that $E[1/X] \geq 1/E[X]$ using Jensen's inequality. Mathematics Stack Exchange, Question #2831860. https://math.stackexchange.com/questions/2831860/show-that-e1-x-1-ex-using-jensens-inequality

[3] Harmonic mean and arithmetic mean – Math Stack Exchange, Question #1777774. https://math.stackexchange.com/questions/1777774/harmonic-mean-and-arithmetic-mean

[4] Clopper–Pearson Exact Method. *Statistics How To*. https://www.statisticshowto.com/clopper-pearson-exact-method/

[5] Calculating the Sample Size for a Confidence Interval. *eCampusOntario Pressbooks*. https://ecampusontario.pressbooks.pub/introstats/chapter/7-5-calculating-the-sample-size-for-a-confidence-interval/

[6] t-Interval for a Mean. *Statistics LibreTexts*. https://stats.libretexts.org/Bookshelves/Introductory_Statistics/Mostly_Harmless_Statistics_%28Webb%29/07%3A_Confidence_Intervals_for_One_Population/7.07%3A_t-Interval_for_a_Mean

[7] Fisher exact test. *Wikipedia*. https://en.wikipedia.org/wiki/Fisher%27s_exact_test

[8] "Obesity and overweight," *WHO*, 2025. https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight

[9] "BMI categories for adults," *CDC*, 2024. https://www.cdc.gov/bmi/adult-calculator/bmi-categories.html

[10] "Normal fasting glucose levels," *Healthline*, 2022. https://www.healthline.com/health/diabetes/fasting-glucose-normal-range

[11] "Prediabetes diagnosis," *Mayo Clinic*, 2023. https://www.mayoclinic.org/diseases-conditions/prediabetes/diagnosis-treatment/drc-20355284

[12] "Understanding Blood Pressure Readings," *American Heart Association*, 2024. https://www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings