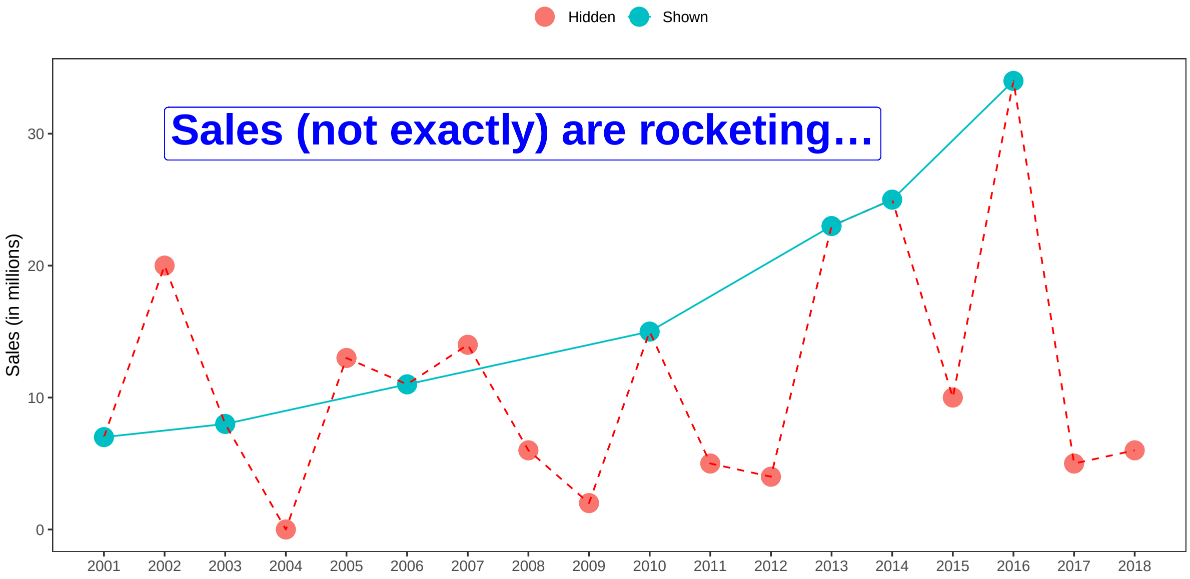
Solution to TECH3 Retake exam August 2025

Problem 1

a) The y-axis does not start at zero, which makes it seem as if the placebo supplement is giving a much lower energy score compared to the VitaBoost+365. The company is thus trying to exchaggerate the difference. Looking at the values on the y-axis, the difference is very small actually.

> **Sensor guide**
>
> 4 points for pointing out that the axis does not start at zero and 4 points for pointing out that the range of the y-axis is really small. Two points for saying something about the company's motives.

b) This plot is a result of **cherrypicking**. The company has only included "good years" such that the development over time shown in the figure seems promising for the future. A steady increase over the years. The text "Sales are ROCKETING!" amplifies the message. Had they kept all the data points, the more honest figure below would be presented to the shareholders.



> **Sensor guide**
>
> 5 points for noting that some years are missing. 3 points for suspecting cherrypicking (not nesessarily using that word). 2 points for sayings something like "convincing shareholders that the company is doing well". Note that this latter point including the figure is not possible for the students to make.

Problem 2

a) i, iii, iv, ii

The values are: i)-0.7 ii) 0.9 iii) 0.05 iv) 0.5.

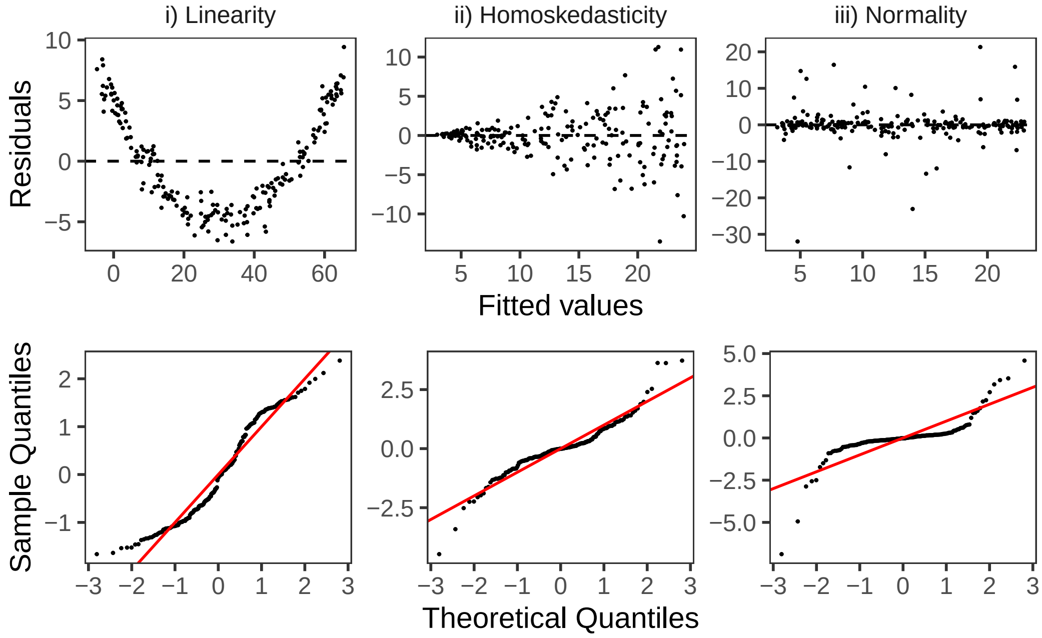
> Sensor guide
>
> 10 points for correct ordering. 2 points for each correct pairwise relative position. E.g. if answered iii, iv, ii, i gives 6 points, while iv, iii,ii, i is just 2 points.

b) i) The linearity assumption is violated. We can clearly see that the relationship between fitted values and the residuals follow a second degree polynomial.

ii) The homoskedasticity assumption is violated. We clear see from the fitted values vs residuals plot that the spread of the residuals is increasing as the fitted values become larger.

iii) The residuals are not normally distributed, based on the qq-normality plot. We see that the points follow a clear S-shape, indicating that the residual distribution has heavier tails compared to the normal distribution.

> Sensor guide
>
> 2 points for correct assumption being violated (6 points). 1 point for argument (i-ii) and 2 points for argumment in (iii).



Problem 3

Milk Jugs	0	1	2	3	4	5	6
Probability	0.05	0.15	0.30	0.25	0.15	0.07	0.03

a) Let X denote the number of milk jugs the cow produces. The probability that the cow produces milk tomorrow, is $P(X > 0) = 1 - P(X = 0) = 1 - 0.05 = 0.95$. The probability that she produces milk every day for a week is $P(\text{produces milk})^7 = 0.95^7 = 0.698$.

Sensor guide

5 points for each correct probability.

b) The expected number of milk jugs:

$$E(X)=\sum_{x=0}^{6} xP(X=x)=0\cdot 0.1+1\cdot 0.15+2\cdot 0.30+3\cdot 0.25+4\cdot 0.15+5\cdot 0.07+6\cdot 0.03=2.63$$

We find the standard deviation using the “short-cut” formula:

$$E(X^2)=\sum_{x=0}^{6} x^2P(X=x)=0^2\cdot 0.1+1^2\cdot 0.15+2^2\cdot 0.30+3^2\cdot 0.25+4^2\cdot 0.15+5^2\cdot 0.07+6^2\cdot 0.03=8.83$$

$$\mathrm{Var}(X)=E(X^2)-E(X)^2=8.83-2.63^2=1.91$$

Thus,

$$\mathrm{SD}(X)=\sqrt{\mathrm{Var}(X)}=\sqrt{1.91}=1.38.$$

Sensor guide

5 points for expectation and 5 points for standard deviation.

c) We are looking for the probability that the cow is producing more than 3500 liters in one year. Let S be the yearly production in liters and X_i denote the production for day $i=1,\ldots,365$. Then

$$S=3.75\sum_{i=1}^{365} X_i$$

, giving

$$E(S)=E(3.75\sum_{i=1}^{365} X_i)=3.75\sum_{i=1}^{365} E(X_i)=3.75\cdot 365\cdot 2.63=3600$$

and

$$\mathrm{Var}(S)=\mathrm{Var}(3.75\sum_{i=1}^{365} X_i)=3.75^2\sum_{i=1}^{365}\mathrm{Var}(X_i)=365\cdot 3.75^2\cdot 1.91=9803$$

or $\mathrm{SD}(S)=\sqrt{\mathrm{Var}(X)}=\sqrt{9803}=99.0$.

The CLT allows us to approximate the distribution of S by a normal distribution with mean 3600 and standard deviation 99. To find the probability of producing more than 3500 per year, we use the Python code in alternative (x): $P(S\geq 3500)=1-P(S\leq 3500)=1-0.1562=0.8438$

```
from scipy.stats import norm
print("x)",  norm.cdf(3500, loc = 3600, scale = 99.0))
```

Sensor guide

3 points for finding the correct expectation and 3 for standard deviation. 3 points for application of CLT and 1 point for finding the correct value.

d) Doing a Monte Carlo simulation, we do not need the central limit theorem. We can just sample daily milk production from the cow using the stated probabilities. We draw a random number, X_i, from 0-6 with the stated probabilities 365 times, i.e. $i = 1, \dots, 365$. The we find the yearly production in liters by

$$S = \frac{3.75\,\text{liter}}{\text{gallon}} \sum_{i=1}^{365} X_i,$$

which gives us the yearly production in liters. We can then calculate the value of this production, V, as $V = S\,\text{liter} \cdot 5\frac{\text{NOK}}{\text{liter}} = 5S\,\text{NOK}$. By repeating this procedure many times, say 10,000, we get the sampling distribution of the income from MacDonald's cow. This can e.g. be visualized by a histogram.

The following python code is not necessary for a complete solution:

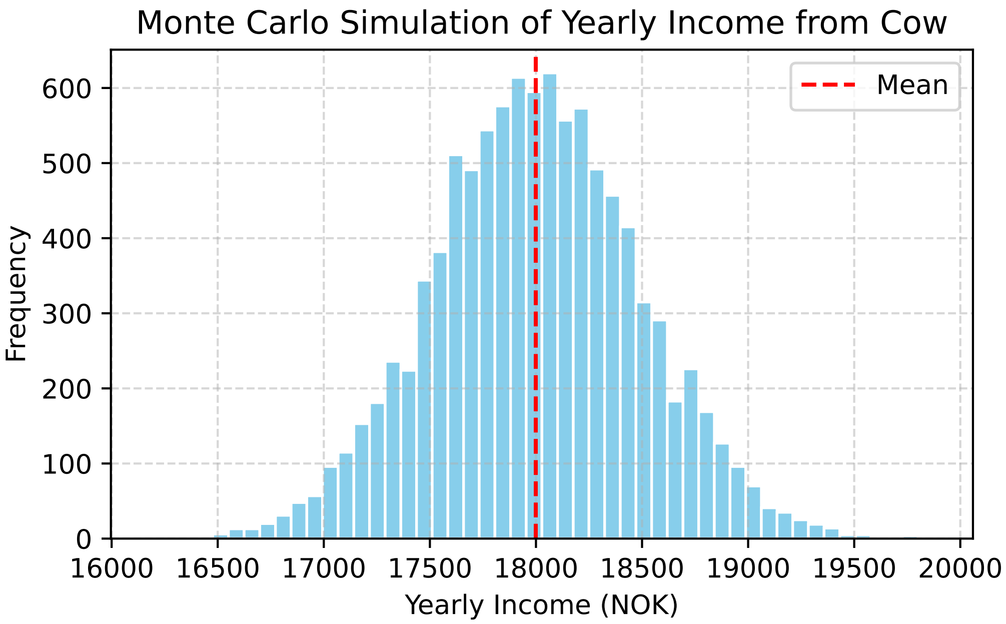
```python
import numpy as np
import matplotlib.pyplot as plt

# Parameters
n_sim = 10000            # Number of simulations
days = 365               # Days per year
jugs = np.arange(7)      # Possible daily milk jugs: 0 to 6
probs = [0.05, 0.15, 0.30, 0.25, 0.15, 0.07, 0.03]  # Given probabilities

gallon_to_liter = 3.75
price_per_liter = 5  # in NOK

# Monte Carlo simulation
np.random.seed(123)
total_income = []

for _ in range(n_sim):
    daily = np.random.choice(jugs, size=days, p=probs)
    total_jugs = np.sum(daily)
    liters = total_jugs * gallon_to_liter
    income = liters * price_per_liter
    total_income.append(income)

# Convert to numpy array
total_income = np.array(total_income)

# Summary statistics
mean_income = np.mean(total_income)
ci_low, ci_high = np.percentile(total_income, [2.5, 97.5])

# Print results
print(f"Expected yearly income: {mean_income:.2f} NOK")
```

```
Expected yearly income: 17999.06 NOK
```

```python
print(f"95% Confidence interval: [{ci_low:.2f}, {ci_high:.2f}] NOK")
```

```
95% Confidence interval: [17043.75, 18975.00] NOK
```

```
# Plot histogram
plt.hist(total_income, bins=50, color='skyblue', edgecolor='white');
plt.title("Monte Carlo Simulation of Yearly Income from Cow");
plt.xlabel("Yearly Income (NOK)");
plt.ylabel("Frequency");
plt.axvline(mean_income, color='red', linestyle='dashed', linewidth=1.5,
label='Mean');
plt.legend();
plt.grid(True, linestyle='--', alpha=0.5);
plt.tight_layout();
plt.show();
```



Sensor guide

Not necessary to write Python code, but if anyone does, focus on the understanding and not perfect syntax or runnable code. The answer should contain an explanation of simulating 365 days of milk production from the giving distribution (not using CLT). It should also mention that this is done a large number of times and how they find the probability from the simulated quantities. 7 points for description of the simulation procedure and 3 points specifically for the explanation of how to find the distribution. Deduct 2 points if the student use CLT in the Monte Carlo simulation.

Problem 4

	Low BR	Med BR	High BR	Total
Few storks	0.235	0.059	0.118	0.412
Many storks	0.294	0.118	0.176	0.588
—	—	—	—	—
Total	0.529	0.177	0.294	1

a) If A and B are independent, then the probability of B should not change if we condition on A. The probability of B (many storks) is 0.588. If we condition on A (high birth rate), we have

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{0.176}{0.294} = 0.598 \neq 0.588 = P(B).$$

Since $P(B|A) \neq P(B)$, the two events are not independent.

ℹ Sensor guide

10 point for correctly proving that events are not independent. Note that the deviation from independence here is quite small.

b)
- H_0 : The birth rate category and storks pairs category are independent events.
- H_A : The birth rate category and storks pairs category are not independent events.

The p-value is very high (0.94), which indicates that there is very weak evidence against the null hypothesis. For any reasonable significance level, we will not reject the null hypothesis that the two events are independent.

ℹ Sensor guide

5 points for hypotheses and 5 points for conclusion.

c) In (a) we selected a random country in the dataset (treating the sample as the whole population). Any deviation from the mathematical definition of independence would lead us to conclude that the two variables are not independent. In (b), we account for the data only being a sample from a unknown population, and use the p-value to evaluate the evidence for/against the null hypothesis.

ℹ Sensor guide

4 points for explaining what was done in the two exercises and 2 points for highlighting the contrast between "within sample" in (a) and "population" in (b).

d) This is a correlation test, where the null hypothesis is that the correlation $H_0 : \rho = 0$, and the alternative hypothesis is that $H_A : \rho \neq 0$. With a significance level of 5%, we would reject the null hypothesis and claim that the correlation is nonzero. The is a linear relationship between the number of stork pairs and the birth rate.

ℹ Sensor guide

5 points for hypotheses and 5 points for conclusion.

e) **Does more storks imply higher birth rates?** No, this is not a causal relationship. **Correlation does not imply causality.** There is a linear association between birth rates and number of stork pairs, but here we are doing an **observational study**, and not a controlled experiment. Causal analysis is therefore not possible here. I would suspect that this is an example of a **spurious correlation**.

ℹ Sensor guide

The point here is that the students show they understand the difference between correlation and causation. 0 points if anyone says yes to the question. 10 points for a good discussion including the highlighted keywords.

f) Here the area of the country (in 1000 km^2) is lineary associated with both birth rate and the number of stork pairs in the country. Perhaps this is a **confounder variable**, meaning that both birth rate and stork pairs increase with land area of the country. Country with large land area will often have more contryside/rural areas, where people tend to have more children (higher birth rate) and more suitable for storks to thrive.

ℹ Sensor guide

The main concept here is confounder variable. We want some kind of discussion saying that both birth rate and the number of storks may depend on a common thing. Area of the country is perhaps not the main confounder, but a proxy for more rural areas where large families and storks may flourish.

Problem 5

a) Find an expression for $\log \mu(x)$. What kind of relationship is there between logarithmic mortality rate $\log \mu$ and the age of a customer (x)? Taking the logarithm of μ we get:

$$\log \mu(x) = \log(ae^{bx}) = \log a + \log e^{bx} = \log a + b\,x.$$

The relationship between μ and x linear.

ℹ Sensor guide

5 points for deriving the correct mathematical expression and 5 points for recognizing it as linear.

b) Identifying the intercept as -8.8024 and the slope associated with age as 0.0788. That is $\widehat{\log \mu} = \widehat{\log a} + \hat{b}x = -8.8024 + 0.0788\,x$. The intercept is the log mortality associated with a person of age 0. Note that the text says the Gompertz model is mainly used for adult populations, so we do not expect this number to be accurate for the log infant mortality. The slope parameter has the interpretation of quantifying the increase in log mortality rate associated with being one year older.

ℹ Sensor guide

4 points for setting up the equation with the correct estimates. 3 points each for interpretation of intercept and slope.

c) The log-mortality rate of a 70-year old: $\widehat{\log \mu}(70) = -8.8024 + 0.0788\,70 = -3.2864$.

The Gompertz model assumes that the log-mortality rate increases linearly as a function of age. As indicated in (b), the Gompertz model will not be able to take into account the increase in mortality at birth. This is not very problematic in a pension setting since it is only those who live until retirement that get any payouts.

ℹ Sensor guide

5 points for corrrect prediction of 70-year old log-mortality. 5 points for the discussion.

Model formulation:

$$\log \mu = \underbrace{\beta_0}_{\text{intercept}} + \underbrace{\beta_1 \cdot \text{Gender}}_{\text{level shift}} + \underbrace{\beta_2 \cdot \text{Cohort}}_{\text{Cohort trend}} + \underbrace{\beta_3 \cdot x}_{\text{Age trend}} + \underbrace{\beta_4 \cdot x \cdot \text{Cohort}}_{\text{cohort-specific slope}} + \underbrace{\beta_5 \cdot x \cdot \text{Gender}}_{\text{gender-specific slope}}$$

d) As the text suggest, we know that life expectancy is increasing as a function of time. A child born in 1900 had a lower life expectancy than one born in 2000. This is captured in the model by the cohort terms. All else fixed, the difference in log mortality of an x year old female born in 1900 and one born in $1900 + h$ is $\beta_2 h + \beta_4 x h$. From the regression output, $\widehat{\beta}_2 = \texttt{Cohort} = -0.0236$ and $\widehat{\beta}_4 = \texttt{Age:Cohort} = 0.0002$, such that the expected log mortality for the same age is decreasing with 0.0236 per year while the age-related slope is increasing with 0.0002 per year.

Note that $\widehat{\beta}_3 = \texttt{Age} = -0.3441$ is negative, which may seem counter-intuitive since mortality increases with age. However, if we consider a female born in 1991 (Cohort = 1991), the age-related slope will be $\widehat{\beta}_3 + \widehat{\beta}_4 \cdot 1991 = -0.3441 + 0.0002 \cdot 1991 = 0.0549$, which is positive.

Sensor guide

2 points for noting that it has to do with the cohort terms. Additional 4 points for more detailed discussion connecting the actual estimated coefficients. 4 points for a good explanation for the change of sign in age-related slope.

e) Here we keep the age (x=70) and cohort fixed, but the Gender is opposite. The difference in log mortality is therefore $\beta_1 + \beta_5 x = \beta_1 + \beta_5 \cdot 70$. From the output, $\widehat{\beta}_1 = \texttt{Gender[T.Male]} = 0.4694$ and $\widehat{\beta}_5 = \texttt{Age:Gender[T.Male]} = -0.0029$. The difference in log mortality rates is therefore $0.4694 - 0.0029 \cdot 70 = 0.266$. Hence, the husband has a higher predicted log-mortality than his wife, according to the model. We can find the relative difference by taking $e^{0.266} = 1.30$, i.e. 30% higher mortality for the husband. He is thus expected to die earlier than his wife, and therefore get a higher payout per month.

Sensor guide

If students are able to identify that β_1 is positive, leading to a higher prediction of log-mortality for the husband, give 6 points. For full credit, the other gender-related term should be included in the discussion.