Solutions

Problem 1

a) This graph is problematic because the y-axis, which is not shown, does not start at zero. This gives an illusion that the graduation rate increased a lot during Obama's predidency, while the increase is just 7 percentage points from 2007-08 to 2013-14. Also, the change from 2009-10 to 2010-11 is clearly exaggerated in the graph, compared to similar increases in the following years. This is likely connected to Obama's presidency starting in 2009 and supporting the message the White House wanted to share with this graphic. Also, the book stacks is an example of chart junk or visual clutter, which should be avoided.

> ℹ Sensor guide
>
> Note that the students may have other points than what is suggested here, but the suggested points are maybe the most obvious. The missing y-axis, exaggerated change from 78% to 79%, and chart junk/visual clutter are the main things to look for in a good answer.

b) First: **Never use a pie chart. Ever.** Second, this particular pie chart consist of three (apparent) proportions, that do not add to 100% and are very similar. If the numbers were not displayed on the chart, identifying the largest value would be challenging, since humans are often poor at accurately perceiving volumes. The chart itself therefore does not add anything that is not presented by the displayed percentages. A barplot would have been better.

> ℹ Sensor guide
>
> Note that the students may have other points than what is suggested here, but the suggested points are maybe the most obvious. Main elements: Proportions do not add to 100%, human perception of volumes, barplot as a better alternative. Reserve 2 points for a sentence similar to: Never use a pie chart. Ever.

Problem 2

a) i) *A p-value tells us the probability that the null hypothesis is true.*
Explanation: The p-value gives the probability of observing data as extreme as (or more extreme than) what was observed, *assuming the null hypothesis is true.* It does not provide the probability that the null hypothesis itself is true.

ii) *A small p-value proves that the alternative hypothesis is true.*
Explanation: A small p-value suggests that the observed data is unlikely under the null

hypothesis, but it does not *prove* the alternative hypothesis. Statistical inference cannot provide absolute proof.

iii) *A Type I error occurs when we fail to reject a false null hypothesis.*
Explanation: A Type I error occurs when we *incorrectly reject* a true null hypothesis. Failing to reject a false null hypothesis is a **Type II error**.

> Sensor guide
>
> 3 points for each statement. An additonal point for explaining what rejecting a false null hypothesis is in iii).

b) i) *In linear regression, the R^2 value always increases when more predictors are added, so we should keep adding predictors.*
Explanation: While R^2 typically increases or stays the same when predictors are added, this does not mean adding predictors is always good. Adding irrelevant predictors can lead to overfitting and reduce model interpretability and predictive performance. Adjusted R^2 accounts for this by penalizing for more predictors.

ii) Correlation measures how strong a nonlinear relationship is between two variables.
Explanation: Correlation (e.g., Pearson's correlation) measures the strength of a *linear* relationship between variables. It does not capture nonlinear associations well.

iii) *In a simple linear regression, the residuals should be normally distributed for the model to be valid.*
Explanation: Normality of residuals is important primarily for prediction intervals, and for small samples, confidence intervals and hypothesis tests. The regression model itself (i.e., estimation of coefficients) does not require normally distributed residuals.

> Sensor guide
>
> 3 points for (i)-(ii) and 4 for (iii).

Problem 3

a) The probability that a single workshop has at least one participant is $P(X \geq 1) = 1 - P(X = 0) = 1 - 0.10 = 0.90$. Assuming the workshops are independent, the probability of at least one participant five consecutive weeks is $0.9^5 = 0.59$.

> Sensor guide
>
> 5 points for each of the probabilities.

b) The expected number of participants is

$$\mu = E(X) = \sum_{x=0}^{4} x\, P(X = x) = 0 \cdot 0.1 + 1 \cdot 0.25 + 2 \cdot 0.3 + 3 \cdot 0.2 + 4 \cdot 0.15 = 2.05.$$

To find the variance, we first find $E(X^2)$;

$$E(X^2) = \sum_{x=0}^{4} x^2\, P(X = x) = 0^2 \cdot 0.1 + 1^2 \cdot 0.25 + 2^2 \cdot 0.3 + 3^2 \cdot 0.2 + 4^2 \cdot 0.15 = 5.65,$$

and then use the *short-cut formula*

$$\sigma^2 = \mathrm{Var}(X) = E(X^2) - \mu^2 = 5.65 - 2.05^2 = 1.45.$$

The standard deviation is then

$$\sigma = \mathrm{SD}(X) = \sqrt{\sigma^2} = \sqrt{1.45} = 1.20.$$

ℹ Sensor guide

5 points for each correct calculation: expectation and standard deviation.

c) Let $W_i = -4,200+(2,200-250)X_i$ denote the profit of the workshop of week $i = 1, \dots, 52$. We then have that the expected profit is

$$\begin{aligned}\mu &= E(W_i) = E(-4,200 + (2,200 - 250)X_i) \\ &= -4,200 + 1,950\, E(X_i) = -4,200 + 1,950\, 2.05 = -202.5\end{aligned}$$

and variance of the profit is

$$\begin{aligned}\sigma^2 &= \mathrm{Var}(W_i) = \mathrm{Var}(-4,200 + 1,950\, X_i) = 1,950^2\, \mathrm{Var}(X_i) \\ &= 1,950^2 \cdot 1.20^2 = 5,475,600,\end{aligned}$$

which gives a standard deviation of $\sigma = \sqrt{5,475,600} = 2,340$. Let $S = \sum_{i=1}^{52} W_i$ denote the total profit after 52 workshops. From the CLT for sums, we have that S is approximately $N(\mu \cdot n, \sigma^2 \cdot n) = N(-202.5 \cdot 52, 2340^2 \cdot 52) = N(-10,530, 16,874^2)$.

The probability of a negative capital, i.e. the probability of S being less than 50,000, can thus be approximated by

$$P(S < -50,000) = 0.009665,$$

from the python output option (i).

```python
from scipy import stats
print("i)",   stats.norm.cdf(-50000, loc = -10530, scale = 16874))
print("ii)",  stats.norm.cdf(-50000, loc = -567, scale = 19568))
print("iii)", stats.norm.cdf(-50000, loc = -8240, scale = 17785))
```

```
i) 0.009665088561003676
ii) 0.005764923931026473
iii) 0.009436092522358941
```

Sensor guide

Note that there may be several ways to the same conclusion here. In the suggested solution we apply the CLT to sum of weekly profits. One could also apply it to the sum of weekly income, and transform the approximately normally distributed variable afterwards. However, we have not learned in the course that linear combinations of normally distributed variables is also normal.
3 points for each of the expectation and standard deviation/variance calculation. 3 points for correct application of CLT for sums and 1 point for choosing the correct probability in the end.

d) To create a Monte Carlo simulation of this, I would simulate 52 weeks of participant numbers (x_i, $i = 1, \dots, 52$) from the distribution given in the text. I would then calculate the total surplus for that year of workshops, using the formula:

$$\text{surplus} = \text{NOK}(2,200 - 250) \sum_{i=1}^{52} x_i - \text{NOK}\, 3,500 \cdot 52.$$

I would repeat the procedure above, say 10,000 times, and calculate the relative frequency of the simulations where the surplus is less than $-\text{NOK}50,000$.

```python
# Monte Carlo simulation:
import numpy as np
surplus = np.zeros(10000)
for i in range(surplus.size) :
  # Simulate number of participants in 52 weeks of workshops:
  x = np.random.choice([0, 1, 2, 3, 4],
                       size=52,
                       p=[0.10, 0.25, 0.30, 0.20, 0.15])
  # Calculate total surplus:
  surplus[i] = -4200*52 + (2200-250)*x.sum()

print("# Probability of going bankrupt: ", np.mean(surplus< -50000))
```

```
# Probability of going bankrupt:  0.0103
```

> 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 a explanation of simulating 52 weeks of parcipitation numbers and a surplus function of that. 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 the probability calculation.

Problem 4

a)

We have that

$$P(V) = \frac{\text{number of visitors making a purchase}}{\text{number of visitors}} = \frac{307}{1029} = 0.298,$$

and

$$P(V|D) = \frac{\text{number of visitors purchasing while being offered discount}}{\text{number of customers being offered discount}} = \frac{163}{505} = 0.210.$$

Since $P(V) \neq P(V|D)$, V and D are not independent.

> Sensor guide
>
> Note that the student may choose other probabilities to compare, and should of course get full credit for that. 5 points for correct calculations and 5 points for correct conclusion.

b)

$$\begin{aligned} H_0 &: \text{ Visitor purchase and visitor being offered discount are independent events.} \\ H_A &: \text{ Visitor purchase and visitor being offered discount are dependent events.} \end{aligned}$$

- The *classical approach* would be to compare the p-value with a predetermined significance level, say $\alpha = 5\%$, and then conclude that, since the p-value $= 0.0927 > 0.05$, we do not reject the null hypothesis and conclude that the two events are independent: **The discount does not have an effect on the whether a purchase is made**.
- The p-value of 0.0927 indicates **weak evidence against the null hypothesis**, suggesting that the two events are independent. We could see this p-value as a measure of belief in the discount nudge.

> Sensor guide
>
> 5 points for correct hypotheses and 5 points for correct conclusion. We do not require students to give both the classical approach and the “measure of belief” interpretation of the p-value.

c)

- In a) we selected a customer randomly from the sample, and could therefore treat the proportions in the table as the true probabilities when drawing a random customer. Any deviation from the rule of independent events would make us conclude that the events are dependent, which was the conclusion we reached.
- In b), we performed a formal statistical test to assess independence in the population. The chi-squared test indicated that the variables are independent in the population, in contrast to (a).

> Sensor guide
>
> 10 points total. 4 points for explaining what we did in the two exercises and 2 points for highlighting the contrast between “within sample” in (a) and “population” in (b).

d) Benefit of the Proportion Z-test:

- The proportion z-test provides a **directional test** (one-sided), allowing us to test a specific hypothesis — here, whether the nudge increases the proportion of reusable bag buyers.
- It also provides a confidence interval for the difference in proportions, giving an estimate of the magnitude of the effect, not just whether it exists.
- In contrast, the chi-squared test is **non-directional** (two-sided) and only tells us if there is a relationship, not the direction or size of the difference.

Practical Implications from the Output:

- The one-sided p-value is 0.0464 is relatively low, which indicates moderate evidence against the null hypothesis that discount nudging does not increase the proportion of purchases.
- The observed effect size is $10/505 - 34/524 = -0.045$, while the population effect size is between -10.4% and 0.8% with 95% probability.
- Practically, this means that implementing a discount nudge does not seem to have any significant effect on whether visitors end up purchasing.
- Whether the company should implement this kind of discount, would depend on the potential costs of the nudging held up against the effect we have observed.

ℹ Sensor guide

5 points for explaining additional benefit. Look for the keywords directional vs non-directional tests. 5 points for practical implications. Should involve some discussion of the p-value and the effect size. Should also include some discussion of the consequences for the company.

e) First, let R denote customer clicking the recommendation and V is still customer making a purchase. We then have from the text that $P(V) = 0.03$, $P(R|V) = 0.8$ and $P(R|V^c) = 0.10$. The manager says that clicking is a very strong signal of purchase intent, but he has it the wrong way. He is assessing the probability of clicking recommendations, given that the customer bought the product. He should rather consider the probability of buying the product conditioned on clicking the recommendation. We find the probability of clicking the recommendations by the *law of total probability + conditional probability*:

$$P(R) = P(R|V)P(V) + P(R|V^c)P(V^c) = 0.8 \cdot 0.03 + 0.10 \cdot 0.97 = 0.121,$$

where we have used that $P(V^c) = 1 - P(V) = 1 - 0.03 = 0.97$ (the complement rule). We then have, using *Bayes law*, $P(V|R) = P(R|V)\frac{P(V)}{P(R)} = 0.8 \cdot \frac{0.03}{0.121} = 0.20$. Even though 80% of buyers clicked, most clickers are non-buyers, and the probability that a clicker actually buys is only about 20%. This is much lower than the marketing manager intuitively thinks. The marketing manager ignores the base rate of buyers (only 3%) and focuses just on conditional probabilities (like “80% of buyers clicked”). This is an example of the **base rate fallacy** or *Bayesian trap*.

In terms of the relative increase from the unconditional probability of buying, $P(V) = 0.03$, to the conditional probability of buying given that the customer clicked, $P(V|R) = 0.20$, one could argue that clicking improve the likelihood of purchase significantly even though the conditional probability $P(V|R)$ is relatively low.

ℹ Sensor guide

5 points for calculation and 5 points for discussion. Keywords to look for is base rate fallacy or Bayesian trap. If these are not mentioned, subtract 1 points from the discussion part.

Problem 5

a) $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ for installation $i = 1, \ldots, 100$.

The fitted model is $y_i = 97,520 + 1,938 \cdot x_i$.

The estimated intercept, $\hat{\beta}_0 = 97,520$ has the interpretation of the representing the fixed costs associated with any installation, while $\hat{\beta}_1 = 1,938$ is the estimated cost of an additional m^2 of installed solar panel.

Sensor guide

4 points for setting up the equation, inserting the correct values for the estimated coefficients. 3 points for interpretation of the intercept and 3 for the slope parameter.

b) The assumptions are: Linearity, independent errors, constant variance, normality (optional), and no multicolinearity. The model only has one covariate, so multicolinearity is not an issue here. The linearity assumption seems to be fulfilled as the relationship between the size of the installation and the cost looks linear from the figure showing the data. The residuals vs fitted values scatter plot has evenly distributed points with no clear pattern, indicating that the errors are independent of the fitted values and that the variance of the errors is constant. The qqplot shows points on the line, indicating that normality can be assumed. This means all assumptions are fulfilled here.

Sensor guide

1 points for each of the correct assumptions and 1 point for the evaluation of each assumption.

c) Predicting costs of a 140 m^2 installation: $\hat{y} = 97520 + 1938 \cdot 140 = 368,840$ NOK. Note that this is an extrapolation, since the largest installation in the training data of the model is around 80 m^2. The consequence here is that this prediction is highly uncertain. Perhaps there are additional costs related to such a large installation. The company should make sure their margin is large enough to account for the uncertainty associated with this.

Sensor guide

5 points for calculating the prediction correctly and 5 points for discussion. Keyword to look for: Extrapolation.

d) The ENOVA support scheme is linear given by $s_i = 7,500 + 1250 \cdot 0.275\, x_i = 7,500 + 344\, x_i$. Unlike the costs of the installation, the support scheme is deterministic, and thus has no error term. Since the maximum subsidy is 32,500, the formula for s_i will only hold up to this maximum. That is,

$$32,500 \geq 7,500 + 344x \quad \Rightarrow \quad x \leq \frac{32,500 - 7,500}{344} = 72.7$$

Thus,

$$s_i = \max(32,500,\; 7,500 + 344\, x_i) = \begin{cases} 7,500 + 344\, x_i, & 0 \leq x_i \leq 72.7 \\ 32,500, & x_i > 72.7. \end{cases}$$

ℹ Sensor guide

5 points for setting up the linear formula for S_i and 5 points specifying the range of x values correctly.

e) An 80% prediction interval based on the Monte Carlo simulation is given by the 10% and 90% percentiles: $[-20784, -33]$. We notice that the 90% percentile is -33, which means slightly more than 90% of the simulated profits are negative, and thus slightly less than 10% probability of positive profits after 12 years.

ℹ Sensor guide

5 points for correct probability and 5 points for correct prediction interval.