NHH

SCHOOL EXAMINATION TECH3

Spring, 2025

Date: June 4. 2025

Time: 09:00-13:00

Number of hours: 4

An invigilator can contact course responsible by phone: +47 995 72 636

SUPPORT MATERIALS PERMITTED DURING THE EXAMINATION:

Calculator Yes ☑ No ☐

Dictionary: one bilingual dictionary permitted.

No other support materials permitted.

Number of pages, including front page: 17

Note that in **appendix D** (at the end), you will find a collection of formulas and definitions that **may** be useful for some of the exercises. Each task (a,b,c, etc.) is worth the **same amount of points** for the grading.

Problem 1

You are presented two graphs. Figure 1a) is an illustration that was published on Twitter by the official White House account for the Obama administration showing an increase in high school graduation rate under Obama's presidency (2009-2017). The second graph was presented on Fox News during the 2012 Republican nomination. For each of the graphs, (a) and (b), discuss in **a few sentences** whether you think these are good graphics? Justify your answer.

a) b)

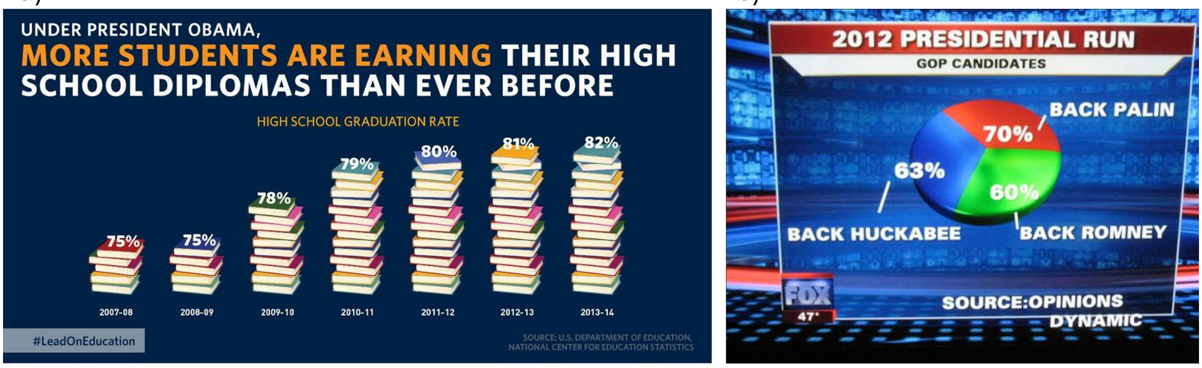


Figure 1: a) Graduation rate under President Obama, posted on Twitter (now X), by the White House. b) Fox News presenting support for various Republican nominees for the 2012 US election.

Problem 2

Below are two lists of statements. Give a **short** explanation why each are wrong.

a) i) A p-value tells us the probability that the null hypothesis is true.

ii) A small p-value proves that the alternative hypothesis is true.

iii) A Type I error occurs when we fail to reject a false null hypothesis.

b) i) In linear regression, the R^2 value always increases when more predictors are added, so we should keep adding predictors.

ii) Correlation measures how strong a nonlinear relationship is between two variables.

iii) In a simple linear regression, the residuals should be normally distributed for the model to be valid.

Problem 3

A startup offering hands-on craft beer brewing workshops is piloting a weekend course format. Each session runs on Saturdays only and is limited to 4 participants due to space and brewing setup constraints. Based on early demand, the number of participants **X** per Saturday follows this discrete distribution:

Participants (X)	0	1	2	3	4
Probability P(X=x)	0.10	0.25	0.30	0.20	0.15

Each participant pays **NOK 2,200**, and costs the company **NOK 250** in materials. There is also a fixed cost of **NOK 4,200** per workshop. The company starts with **NOK 50,000 in capital** and plans to run the workshop every Saturday for one year (**52 weeks**).

a) What is the probability that a single workshop has at least one participant? What is the probability that the workshop has at least one participant every Saturday for five consecutive weeks?

b) Show that the expected number of participants per week is 2.05 and that the standard deviation is 1.20.

c) Assuming the number of participants each week is independent, use the **Central Limit Theorem** to estimate the probability that after 52 weeks, the company's ending capital is negative(i.e. their initial capital of **NOK 50,000** has been spent).

> The Central Limit Theorem for a sum
>
> Let $X_1, X_2, ..., X_n$ be a sequence of independent and identically distributed random variables having a distribution with expectation $E(X) = \mu_X$ and finite variance $Var(X) = \sigma^2$. Then the sum $\sum_{i=1}^{n} X_i$ is approximately normally distributed with expectation $n \cdot \mu$ and variance $n \cdot \sigma^2$, i.e. $N(n \cdot \mu, n \cdot \sigma^2)$.

Use the following pre-calculated values if needed:

```
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
```

d) Sketch a Monte Carlo simulation routine (in pseudocode or Python-style code) to estimate the probability in part (c).

Problem 4

An online retailer is experimenting with a personalized discount strategy. For every new visitor, a fair virtual coin is flipped, and if it shows heads, the visitor is shown a **limited-time 10% off!** banner for a specific product. The goal is to see whether offering a visible discount increases the likelihood of purchase.

From a controlled sample of 1029 site visitors, the following purchasing data is recorded:

	Discount Banner: No	Discount Banner: Yes	Total
Did Not Purchase	380	342	722
Purchased	144	163	307
Total	524	505	1029

a) Suppose one of these customers is randomly selected. Are the following two events independent? V: The visitor made a purchase, D: The visitor was shown the discount banner.

Doing a chi-squared test of independence based on the above contingency table in Python we get the following output:

```
Chi-squared statistic: 2.83

Degrees of freedom: 1

P-value: 0.0927
```

b) Formulate the null- and alternative hypothesis for this test. What is your conclusion based on the output?

c) What is the difference between the conclusion you draw in a) and the conclusion you draw in b)?

Below follows the Python output of a one-sided proportion z-test, where the alternative hypothesis is that the proportion of visitors that purchased in the discount banner population is higher than in the no discount banner population.

```
Z-statistic: -1.6811

One-sided P-value: 0.0464

95% Confidence interval for difference in proportions: (-0.104, 0.008)
```

d) What is the additional benefit of the proportion test compared to the chi-squared test? What practical implications can you draw from the output above?

In a separate experiment for a different product, the retailer studied whether clicking a product recommendation is a good predictor of purchasing. They found:

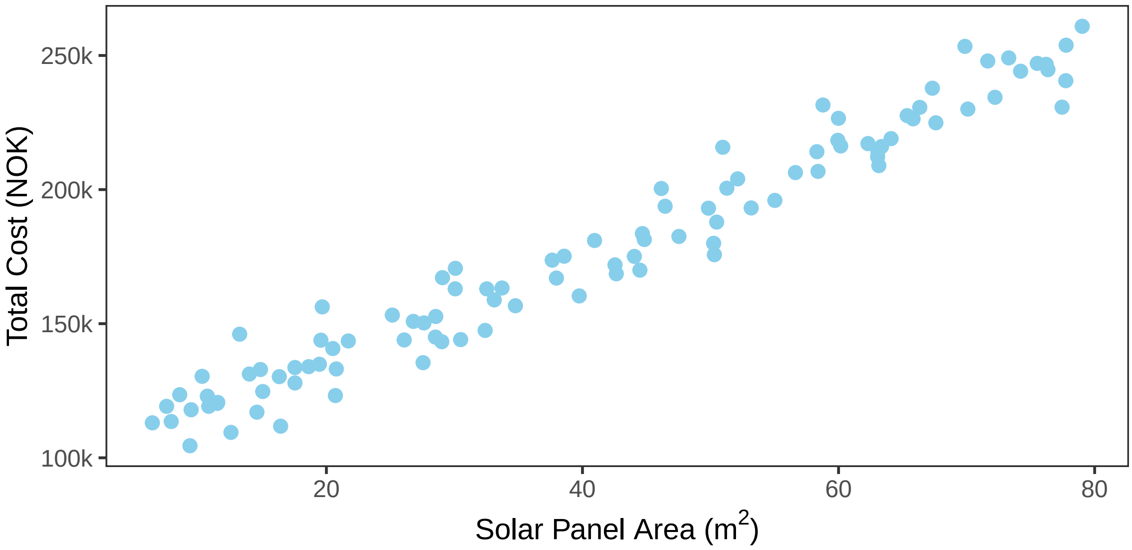
- Only **3%** of site visitors actually bought the product.
- **80%** of those who bought the product **clicked the recommendation**.
- Just **10%** of those who **did not buy the product** also clicked the recommendation.

The marketing manager sees that most buyers clicked and very few non-buyers did — and concludes that **clicking is a very strong signal** of purchase intent.

e) Do you agree with the marketing manager? Support your answer with calculations.

Problem 5

A company called *Watt's Up Solar* installs solar panels on roofs of private houses, and the company wants to simplify their pricing procedure. They distinguish between costs associated with all installations and costs that increase with the size of the installation, measured by the area (in m^2) of solar panels being installed. The model is only intended for internal use, and the company adds a margin on top of the output from the model. For installation i, let x_i denote the roof area covered by the installation (in m^2), and y_i denote the total costs (in NOK) of the installation. The data is presented in the figure below.



The company uses a linear regression model and the summary of the fitted model is presented in **Appendix A**.

a) Set up the equation for the model inserting the estimated coefficients. Give an interpretation of the coefficients for this particular context.

Diagnostic plots for the fitted model are presented in **Appendix B**.

b) What are the assumptions of a linear regression and are they fulfilled here?

c) A new customer has a gigantic roof. The customer wants to install 140 m^2 of solar panels. What will the internal cost of this project be? Comment on the prediction.

ENOVA is a Norwegian state-owned enterprise that offers subsidies to private households willing to invest in improving energy efficiency of the house - including investing in solar panels. The support scheme for solar panel installations[1] starts at 7500 NOK and increases with 1250 NOK per kilowatt installed effect. The installed effect of the solar panels the company uses is 0.275kilowatt-hour/m^2. The maximum total subsidy is 32,500 NOK per household.

d) Set up an equation describing the relationship between the subsidy of an installation i, s_i, and the size of the installation, x_i. For what values of x_i does the equation hold?

The government proposes introducing a “Norwegian price” for electricity, offering everyone a fixed price of **0.40 NOK per kilowatt-hour**[2]. Using this information and a few assumptions, we have implemented a Monte Carlo simulation to quantify the probability that the installation is profitable after 12 years. The output from the Monte Carlo simulation is presented in **Appendix C**.

e) Use the output in **Appendix C** to find a 80% prediction interval of the profits after 12 years. Approximate the probability of the installation being profitable after 12 years.

[1] https://www.enova.no/privat/alle-energitiltak/solenergi/solcelleanlegg/.

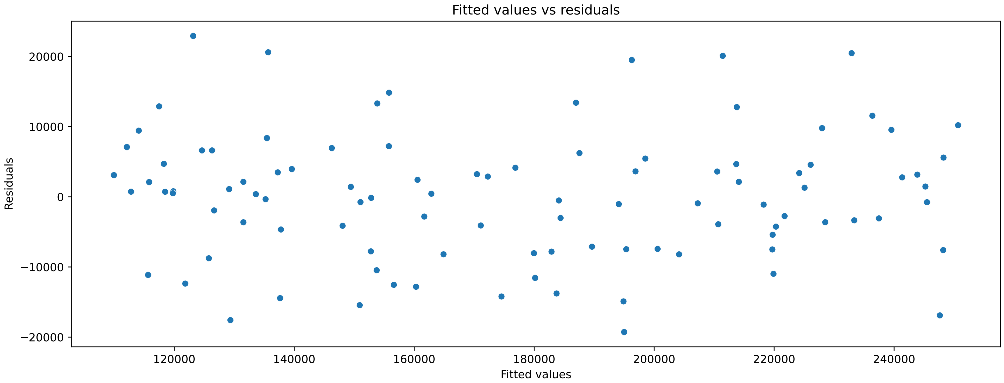
[2] https://www.regjeringen.no/no/aktuelt/norgespris-skal-sikre-forutsigbare-og-stabile-strompriser-for-folk/id3090849/.

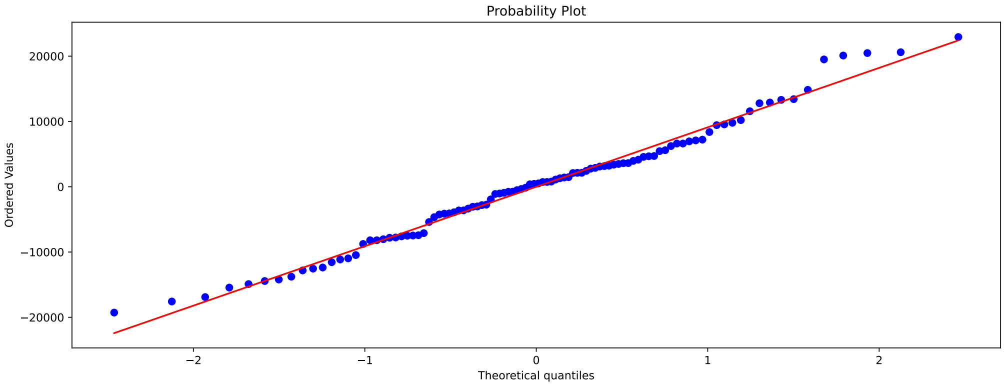
Appendix A

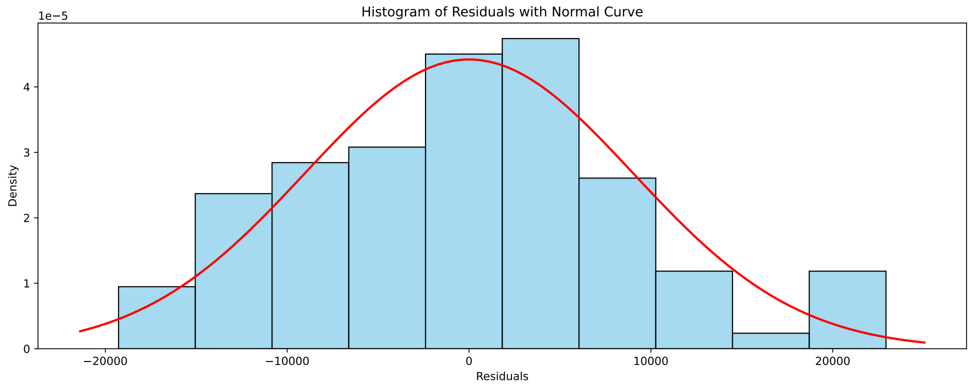
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.957
Model:                            OLS   Adj. R-squared:                  0.957
Method:                 Least Squares   F-statistic:                     2189.
Date:                Wed, 4 June 2025   Prob (F-statistic):           7.58e-69
Time:                        09:00:00   Log-Likelihood:                -1052.2
No. Observations:                 100   AIC:                             2108.
Df Residuals:                      98   BIC:                             2114.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   9.752e+04   1917.747     50.853      0.000    9.37e+04    1.01e+05
x           1937.8685     41.418     46.788      0.000    1855.675    2020.062
==============================================================================
Omnibus:                        0.900   Durbin-Watson:                   2.285
Prob(Omnibus):                  0.638   Jarque-Bera (JB):                0.808
Skew:                           0.217   Prob(JB):                        0.668
Kurtosis:                       2.929   Cond. No.                         97.9
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is
    correctly specified.
```

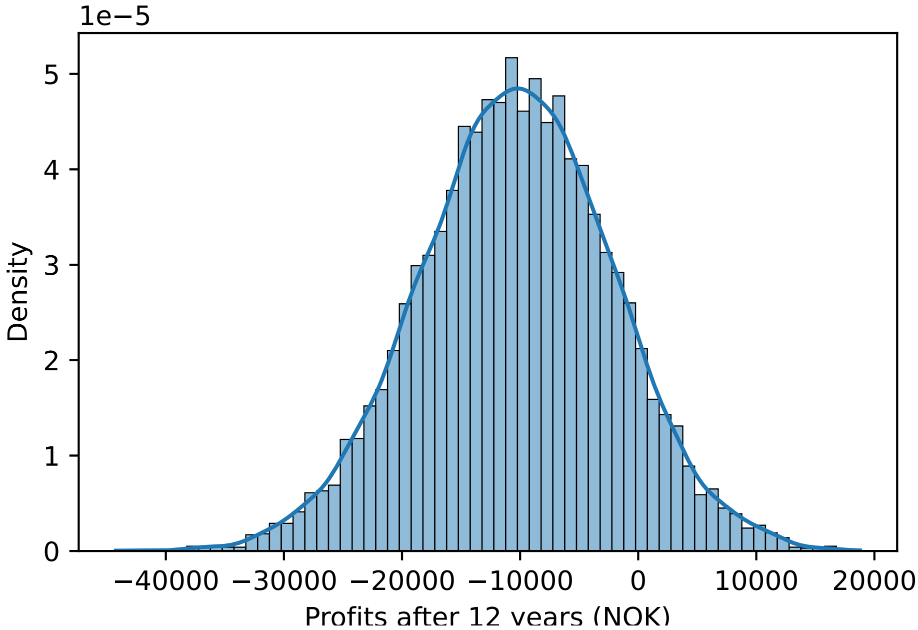
Appendix B


Fitted values vs residuals
Residuals
Fitted values
20000
10000
0
−10000
−20000
120000
140000
160000
180000
200000
220000
240000


Probability Plot
Ordered Values
Theoretical quantiles
20000
10000
0
−10000
−20000
−2
−1
0
1
2


Histogram of Residuals with Normal Curve
1e−5
Density
Residuals
4
3
2
1
0
−20000
−10000
0
10000
20000

Appendix C




```
Percentiles of profits after 12 years

0% percentile: -44227
10% percentile: -20784
20% percentile: -17177
30% percentile: -14486
40% percentile: -12368
50% percentile: -10319
60% percentile: -8223
70% percentile: -6087
80% percentile: -3498
90% percentile: -33
100% percentile: 18776
```

Appendix D: Formulas

Module 1: Visualizing and summarizing data

Summary statistics

- Mean
$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$$
- Standard deviation
$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}$$
- Variance
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$$
- Covariance
$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)(y_i - \bar{y}_n)$$
- Median: Middle observation.
- Mode: Most frequent observation.

Linear combination of variables

If $Z = a \cdot X + b \cdot Y$, then:
- $\overline{Z} = a \cdot \overline{X} + b \cdot \overline{Y}$.
- $S_Z^2 = a^2 \cdot S_X^2 + b^2 \cdot S_Y^2 + 2ab \cdot S_{XY}$.

Module 2: Probability, random variables, probability distributions and simulations.

Sample space

The sample space S is the set of all possible outcomes for an experiment. For example, when throwing a dice, $S = \{1, 2, 3, 4, 5, 6\}$.

Kolmogorov's axioms and probabilty rules

1. Range: The probabiltity $P(A)$ of any event A satisfies $0 \leq P(A) \leq 1$.
2. Something will happen: If S is the sample space, then $P(S) = 1$.
3. Union of disjoint events: If $A_1, A_2, A_3, \ldots,$ are pairwise disjoint events, then

$$P(A_1 \cup A_2 \cup A_3 \cup \ldots) = P(A_1) + P(A_2) + P(A_3) + \cdots .$$

From these we can derive

4. Complement rule: $P(A) + P(A^c) = 1$
5. General rule of unions: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

6. Law of total probability:

$$P(A) = P(A \cap B_1) + P(A \cap B_2) + \cdots + P(A \cap B_k),$$

where $B_1, \ldots, B_k$ are disjoint events.

The law of conditional probability.

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Independence

Two events, A and B, are independent if $P(A \cap B) = P(A)\,P(B)$.

Bayes law

$$P(B|A) = P(A|B) \cdot \frac{P(B)}{P(A)}.$$

Law of total probability + conditonal probability

If we split the sample space into n disjoint $B_1, \ldots, B_n$, then

$$P(A) = P(A|B_1) \cdot P(B_1) + P(A|B_2) \cdot P(B_2) + \cdots + P(A|B_n) \cdot P(B_n).$$

Discrete expectations

Let X be a discrete random variable with probability distribution function $p(x)$. The expectation of X, $E(X)$ is calculated by

$$\mu_x = E(X) = \sum_{\text{all values of } x} x \cdot p(x).$$

For a general function, g, we have that

$$E(g(x)) = \sum_{\text{all values of } x} g(x) \cdot p(x).$$

The variance shortcut

The definition of the variance is

$$\text{Var}(X) = E([X - \mu_x]^2) = \sum_x (x - \mu_x)^2 p(x).$$

Often it is quicker to use the "shortcut" formula:

$$\text{Var}(X) = E(X^2) - \mu_x^2.$$

The Bernoulli distribution

Let X be Bernoulli(p) distributed. Then X is a binary variable, with possible values of X being 0 or 1, and the probability distribution function of X is

$$P(X = 1) = p, \quad P(X = 0) = 1 - p, \quad p \in (0, 1).$$

We refer to p as the probability of success.

Probability density function (pdf)

For continuous random variables, the probability density function (pdf), $f(x)$, describes the distribution. A pdf must fulfill that for all values of x, $f(x) \geq 0$ and

$$\int_{-\infty}^{\infty} f(x)dx = 1.$$

Continuous cumulative probabilities

Let X be a continuous random variable with pdf $f(x)$, for $x \in \mathcal{R}$. The cumulative distribution function, $F(x) = P(X \leq x)$, can be found by

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\, dt.$$

Continuous expectations

Let X be a continuous random variable with density function $f(x)$, for $x \in \mathcal{R}$. The expectation of X, $E(X)$ is calculated by

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x)\, dx.$$

For a general function, g, we have that

$$E(g(x)) = \int_{-\infty}^{\infty} g(x) \cdot f(x)\, dx.$$

Joint vs marginal probability distributions

A marginal distribution function is for one random variable, as we have seen so far. The joint distribution function gives the probability of any outcome of two or more random variables. Let X and Y be discrete random variables. Their joint distribution function is $p_{xy}(x, y) = P(X = x \cap Y = y)$. If X and Y are independent, $p_{xy}(x, y) = p_x(x)p_y(y)$.
Given the joint distribution function, we can find the marginal distribution functions using the law of total probability:

$$p_x(x) = \sum_{\text{all values of } y} p_{xy}(x, y), \quad p_y(y) = \sum_{\text{all values of } x} p_{xy}(x, y).$$

Similarly, if X and Y are continuous, their marginal density functions can be found by integrating their joint density function

$$f_x(x) = \int_{\text{all values of } y} f_{xy}(x,y)dy, \quad f_y(y) = \int_{\text{all values of } x} f_{xy}(x,y)dx.$$

Again, if X and Y are independent, $f_{xy}(x,y) = f_x(x)f_y(y)$.
Depending on X and Y being discrete or continuous, we can find

$$E(XY) = \sum_x \sum_y xy\, p_{xy}(x,y) \quad \text{or} \quad E(XY) = \int_x \int_y xy\, f_{xy}(x,y)\, dx\, dy.$$

Covariance and correlation

Let X and Y be two random variables. The covariance between X and Y is a measure of the linear dependence between X and Y, and is defined by

$$\text{Cov}(X,Y) = E[(X - E(X))(Y - E(Y))] = E(X\,Y) - E(X)\,E(Y).$$

If X and Y are independent, $\text{Cov}(X,Y) = 0$. If X and Y are normally distributed, $\text{Cov}(X,Y) = 0$ also implies that X and Y are independent.
Since the covariance can be a bit difficult to interpret, we often use correlation for measuring the linear relationship between X and Y. The correlation, ρ, is a standardization of covariance and is a number in $[-1, 1]$, defined by

$$\rho_{xy} = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}.$$

Some rules relating to covariances:

$$\begin{aligned}
\text{Cov}(X,a) &= 0, \\
\text{Cov}(X,X) &= \text{Var}(X), \\
\text{Cov}(aX+b, cY+d) &= ac\text{Cov}(X,Y), \\
\text{Cov}(X, Y+Z) &= \text{Cov}(X,Y) + \text{Cov}(X,Z), \\
\text{Var}(aX+bY) &= a^2\text{Var}(X) + b^2\text{Var}(Y) + 2ab\text{Cov}(X,Y),
\end{aligned}$$

where X,Y,Z are random variables and a,b,c,d are constants.

Module 3: Estimation, sampling distributions and resampling

Population and sample

A **population** is a collection of all items of interest to our study. The numbers we obtain from the **population** is called parameters. A **sample** is a subset of the population. The numbers obtained from the sample are called **statistics**.

Estimator

An **estimator** is a function of the sample that provides an estimate of the unknown parameter. An estimator is a statistic since we use the sample to compute it.

Unbiased and consistent estimators

Let $\hat{\theta}_n$ be an estimator of the parameter θ_0. An estimator is **unbiased** if $E(\hat{\theta}_n) = \theta_0$. An estimator is **consistent** if

$$lim_{n\to\infty} E(\hat{\theta}_n) = \theta_0 \quad \text{and} \quad lim_{n\to\infty} Var(\hat{\theta}_n) = 0$$

Sampling distribution

The **sampling distribution** is the distribution of our statistic/estimator across samples.

The Central Limit Theorem

Let $X_1, X_2, ..., X_n$ be a sequence of independent and identically distributed random variables having a distribution with expectation $E(X) = \mu_X$ and finite variance $Var(X) = \sigma^2$. Then

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ is approximately } N(\mu, \frac{\sigma^2}{n}).$$

Thus, for a sufficiently large sample size n, $\bar{X}$ is approximately normally distributed with mean μ and variance $\frac{\sigma^2}{n}$.

The Central Limit Theorem for a sum

Let $X_1, X_2, ..., X_n$ be a sequence of independent and identically distributed random variables having a distribution with expectation $E(X) = \mu_X$ and finite variance $Var(X) = \sigma^2$. Then the sum $\sum_{i=1}^{n} X_i$ is approximately normally distributed with expectation $n \cdot \mu$ and variance $n \cdot \sigma^2$, i.e. $N(n \cdot \mu, n \cdot \sigma^2)$.

Discrete expectations

Let X be a discrete random variable with probability distribution function $p(x)$. The expectation of X, $E(X)$ is calculated by

$$\mu_x = E(X) = \sum_{\text{all values of } x} x \cdot p(x).$$

For a general function, g, we have that

$$E(g(x)) = \sum_{\text{all values of } x} g(x) \cdot p(x).$$

Expectation and variance rules

Let $X_1, X_2, \ldots, X_n$ be random variables. For constants, a and b,

$$E(a + b\,X_1) = a + b\,E(X_1), \quad \text{and} \quad \text{Var}(a + b\,X_1) = b^2\,\text{Var}(X_1).$$

We also have that

$$E(X_1 + X_2 + \cdots + X_n) = E(X_1) + E(X_2) + \cdots + E(X_n).$$

If $X_1, X_2, \ldots, X_n$ are independent,

$$\text{Var}(X_1 + X_2 + \cdots + X_n) = \text{Var}(X_1) + \text{Var}(X_2) + \cdots + \text{Var}(X_n).$$

The variance shortcut

The definition of the variance is

$$\text{Var}(X) = E([X - \mu_x]^2) = \sum_x (x - \mu_x)^2 p(x).$$

Often it is quicker to use the "shortcut" formula:

$$\text{Var}(X) = E(X^2) - \mu_x^2.$$

Module 4: Designing studies, hypothesis testing, and quantifying effects

Hypothesis testing's six steps

1. Formulate a hypothesis of interest
2. Specify the null and alternative hypotheses
3. Collect some data
4. Fit a model to the data and compute a test statistic
5. Determine the probability of the observed result under the null hypothesis
6. Assess the "statistical significance" of the result

Type I/II errors and the significance level

Type I error is rejecting the null hypothesis when it is true. Type II error is failing to reject the null, when it is false. The significance level of a hypothesis test is the probability of making a Type I error.

Confidence interval: The theoretical approach

$$\text{CI} = \text{point estimate} \pm \text{critical value} \times \text{standard error}$$

If the data is normally distributed, and we want a $(1 - \alpha)100\%$ confidence interval, the interval can be found using the formula

$$\text{CI} : \overline{x} \pm t_{1-\alpha/2,n-1} \times \frac{s}{\sqrt{n}},$$

where $t_{1-\alpha/2,n-1}$ is a critical value in a t-distribution with $n-1$ degrees of freedom, fulfilling $P(T \leq t_{1-\alpha/2,n-1}) = 1-\alpha$. For $n = 30$ and $\alpha = 0.05$ (95% CI) the critical value can be found in Python by:

```
from scipy import stats
alpha = 0.05
n=30
print(stats.t.ppf(q=1-alpha/2, df=n-1))
```

```
2.045229642132703
```

Cohen's d

Cohen's d is an effect size and can be calculated as

$$\text{Cohen's d} = \frac{\text{mean difference}}{\text{standard deviation}}.$$

Module 5: Measuring relationships and fitting models

Pearson chi-squared test for discrete distributions

We have a null hypothesis formulated as a *discrete* probability distribution $p_1, \ldots, p_k$ of observing possible outcomes $u_1, \ldots, u_k$. When we observe n outcomes from this distribution, we would expect $e_i = p_i \cdot n$ observations of outcome u_i. If we *have* observed n outcomes from the distribution, and outcome u_i has occurred f_i times, we then wonder whether the observed frequencies (f_i) differ so much from the *expected* frequencies (e_i) that we no longer believe that $p_1, \ldots, p_k$ is the true probability distribution.
The test statistic is given by:

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i},$$

which follows a χ^2-distribution with $k-1$ degrees of freedom if the null hypothesis is true.

Independent events (revisited)

For two events A and B with positive probability of occurring, the following three statements are **equivalent**:

$$P(A \cap B) = P(A) \cdot P(B)$$

$$P(A \mid B) = P(A)$$

$$P(B \mid A) = P(B)$$

Chi-squared test for independence

Assume we have n observations that can be characterized by two categorical variables, A and B. Assume that variable A can be classified into r categories $a_1, a_2, ..., a_r$, and variable B can be classified into s categories $b_1, b_2, ..., b_s$. We organize the observations in a **contingency table**, where each cell is the observed frequencies of observation having $a = a_i$ and $b = b_j$, denoted by f_{ij}:

	b_1	b_2	$\cdots$	b_s	**Sum**
a_1	f_{11}	f_{12}	$\cdots$	f_{1s}	$f_{1\cdot}$
a_2	f_{21}	f_{22}	$\cdots$	f_{2s}	$f_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
a_r	f_{r1}	f_{r2}	$\cdots$	f_{rs}	$f_{r\cdot}$
Sum	$f_{\cdot 1}$	$f_{\cdot 2}$	$\cdots$	$f_{\cdot s}$	n

Under the null hypothesis of A and B being independent, the **expected frequency** in each cell is given by:

$$e_{ij} = \frac{f_{i\cdot} \cdot f_{\cdot j}}{n}$$

We then wonder whether the observed frequencies (f_{ij}) differ so much from the *expected* frequencies (e_{ij}) that we no longer believe that A and B are independent. The test statistic is:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(f_{ij} - e_{ij})^2}{e_{ij}},$$

which approximately follows a χ^2-distribution with $(r-1)(s-1)$ degrees of freedom, provided the expected frequencies are large enough (typically at least 5 in each cell). A large value of χ^2 indicates a greater discrepancy between observed and expected frequencies and thus provides evidence **against the null hypothesis of independence**.

Two-Proportion Z-Test

We are testing whether there is a difference in the proportion of successes between two independent groups. Let p_1 be population proportion in **Group 1** and p_2 be population proportion in **Group 2**. In a one-sided test or directional test we are testing:

$$H_0 : p_1 = p_2 \qquad H_A : p_1 > p_2$$

while for a two sided test we are testing

$$H_0 : p_1 = p_2 \quad \text{vs.} \quad H_a : p_1 \neq p_2$$

Let $\hat{p}_1$ and $\hat{p}_2$ be the **sample proportions** of success in group 1 and 2, respectively, and let p_{pool} be the total proportion. The test statistic is given as:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{SE} \quad \text{where} \quad SE = \sqrt{p_{\text{pool}}(1 - p_{\text{pool}}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

Under the null hypothesis H_0, the test statistic Z approximately follows a **standard normal distribution**: $Z \sim \mathcal{N}(0, 1)$

Covariance and correlation

Let X and Y be random variables with $\mu_x = E(X)$ and $\mu_Y = E(Y)$, then

$$\operatorname{Cov}(X, Y) = E\Big[(X - \mu_x)(Y - \mu_y)\Big] = E(X \cdot Y) - \mu_x \mu_y.$$

The correlation is

$$\rho = \frac{\operatorname{Cov}(X, Y)}{\operatorname{SD}(X)\operatorname{SD}(Y)}$$