SCHOOL EXAMINATION TECH3

Spring Retake, 2025

Date: August 8. 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.

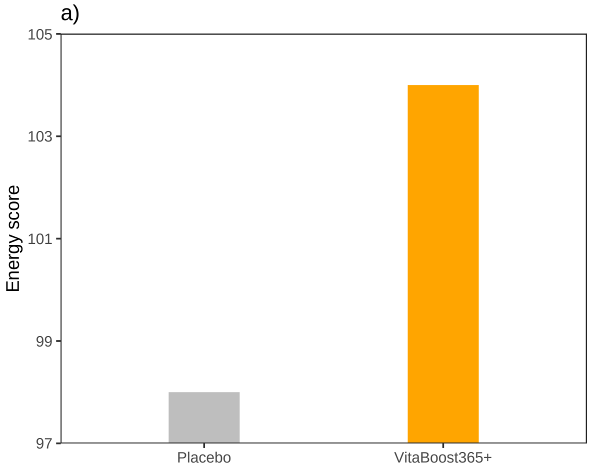
Number of pages, including front page: 15

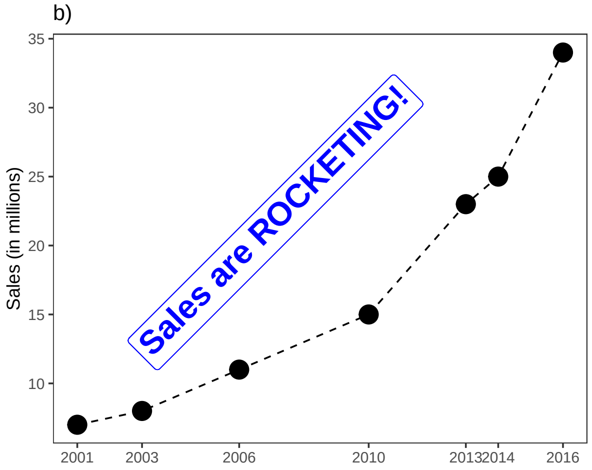
Note that in the **Appendix** (at the end), you will find a collection of formulas and definitions that **may** be useful for some of the exercises.

Problem 1

A company, called *MetaZen Pharma*, is selling a dietary supplement called **VitaBoost365+**. The first graphic (a) below was used in a nationwide campaign, where they claimed that their product lead to a large improvement in energy score compared to a placebo supplement.

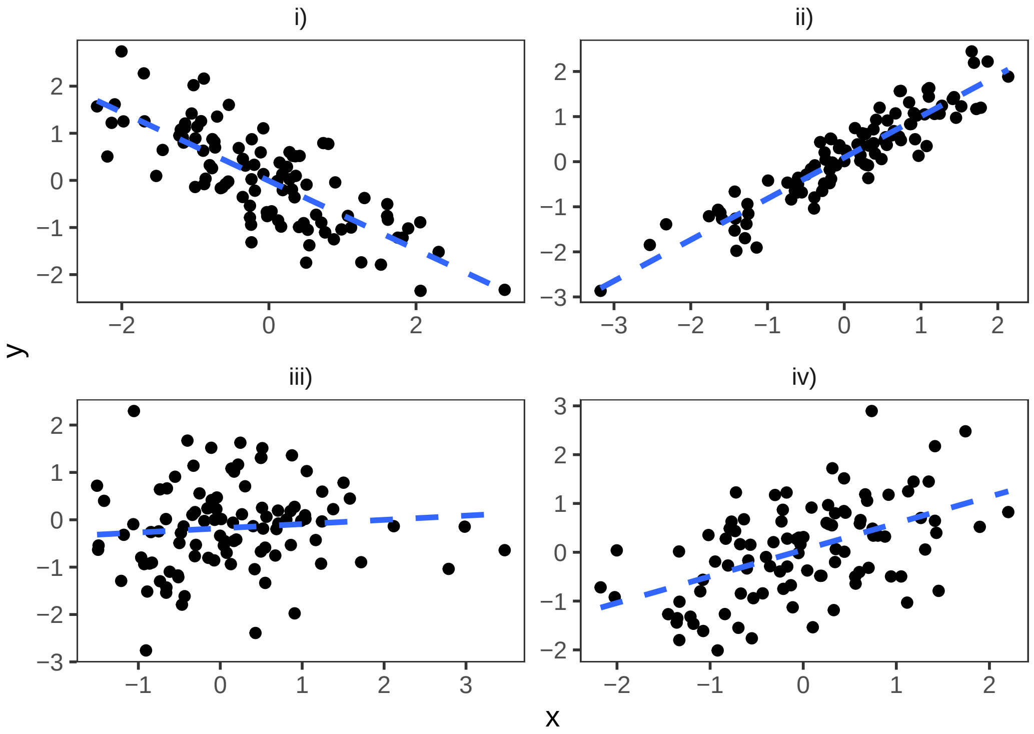
The second graphic (b) was shown at the general assembly of share holders. Discuss how these graphics have been manipulated for pushing the agenda of the company. What is the company trying to achieve here?



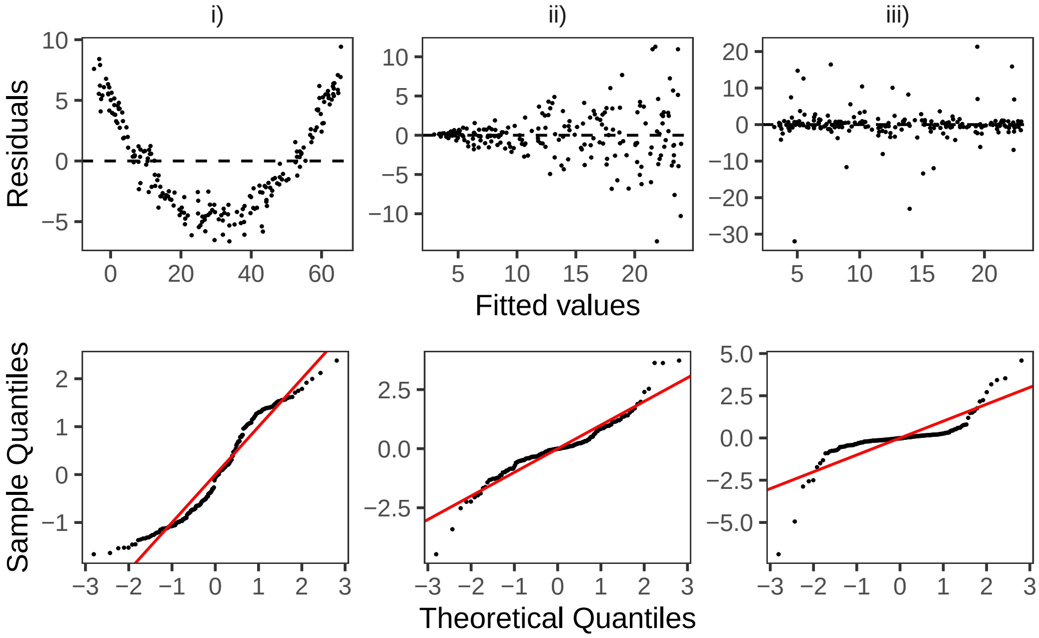


Problem 2

a) Let r be the empirical correlation between data vectors x and y. For the 4 plots below, arrange the empirical correlation between X and Y from low to high.



b) There are five assumptions to linear regression. Below, we have plotted the residual diagnostic plots *fitted values vs residuals* and *qqplot* for three models. For each fitted model in the figure, one assumption is violated. Which? **Justify your answer with a short explanation**.



Problem 3

Old MacDonald had an organic farm and on that farm he had a cow. He was worried that the cow was producing too little milk, and started to count how many milk jugs the cow filled each day. After a few months, he has estimated the following probabilities:

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) What is the probability that the cow fills any milk jugs tomorrow? What is the probability that she fills any milk jugs every day for a week?

b) Show that the expectation and standard deviation of the number of milk jugs produced by the cow are respectively, **2.63** and **1.38**.

The farmer sells his milk through a dairy company, *Trine Dairy*. To be allowed to sell the milk through the company, he must produce a minimum requirement of **3500 liters** per year. Each milk jug is **1 gallon = 3.75 liters**.

c) Using the Central limit theorem, what is the probability that the cow will produce more than the minimum requirement? You may find the following Python code and output useful:

```
from scipy.stats import norm
print("i)",   norm.cdf(3500, loc = 960, scale = 99.0))
print("ii)",  norm.cdf(3500, loc = 960, scale = 26.4))
print("iii)", norm.cdf(3500, loc = 960, scale = 136.8))
print("iv)",  norm.cdf(3500, loc = 26.4, scale = 3600))
print("v)",   norm.cdf(3500, loc = 26.4, scale = 960))
print("vi)",  norm.cdf(3500, loc = 99.0, scale = 3600))
print("vii)", norm.cdf(3500, loc = 99.0, scale = 960))
print("viii)",norm.cdf(3500, loc = 976.8, scale = 3600))
print("ix)",  norm.cdf(3500, loc = 976.8, scale = 960))
print("x)",   norm.cdf(3500, loc = 3600, scale = 99.0))
print("xi)",  norm.cdf(3500, loc = 3600, scale = 26.4))
print("xii)", norm.cdf(3500, loc = 3600, scale = 976.8))
```

Alternative	Output	Alternative	Output	Alternative	Output
`i)`	1.000000	`ii)`	1.000000	`iii)`	1.000000
`iv)`	0.832700	`v)`	0.999852	`vi)`	0.827600
`vii)`	0.999802	`viii)`	0.758314	`ix)`	0.995710
`x)`	0.156223	`xi)`	0.000076	`xii)`	0.459229

MacDonald receives **5 NOK** per liter of milk he delivers.

d) Sketch how you would implement a Monte Carlo simulation for the income distribution from Old MacDonald's cow. Python code is not necessary.

Problem 4

According to legend, storks (the bird) are responsible for delivering babies to new parents. This folktale, popularized in European traditions and fairy tales, imagines storks flying through the sky with newborns wrapped in cloth bundles, gently placing them on doorsteps.
Scientist Robert Matthews wanted to do an observational study on this legend and gathered information about the number of stork pairs and the birth rate in several countries ($n = 17$).

As a starting point, we split number of storks pairs into two categories: few storks and many storks, and the birth rate into three categories, low, medium and high birth rate. We then count the number of countries in each category. This gives the following table:

	Low BR	Med BR	High BR	Total
Few storks	4	1	2	7
Many storks	5	2	3	10
—	—	—	—	—
Total	9	3	5	17

a) We randomly select a country from the dataset. Let event A be that the country has a high birth rate, and let event B be that the country has many storks. Are events A and B independent?

b) We do a chi-square contingency test on the table above. Formulate the hypotheses and give your conclusion based on the python output below:

```
Chi-squared statistic: 0.1187301587301587

Degrees of freedom: 2

P-value: 0.9423626691840816
```

c) Explain the difference between the conclusion in (a) and (b).

Robert Matthews did a **correlation test** to see if he could observe a relation between number of storks pairs and birth rate.

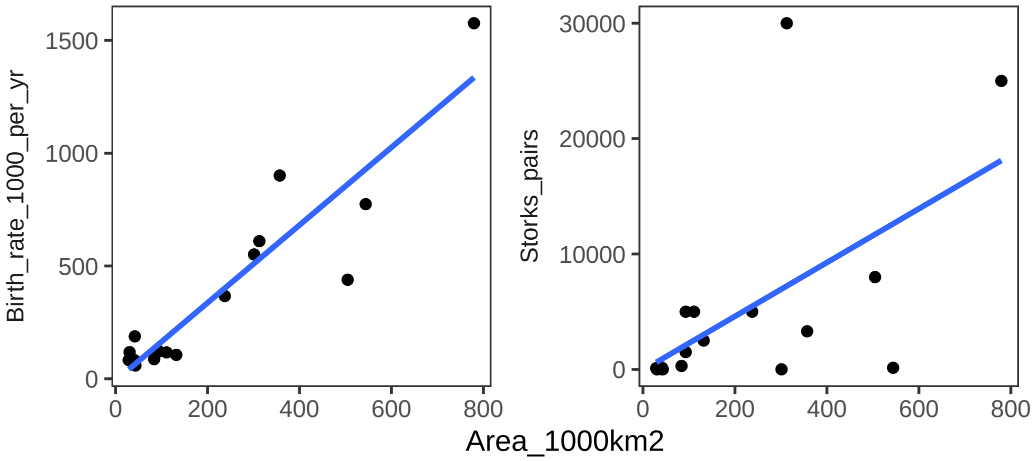
```
Correlation: 0.620, p-value: 0.008
```

d) What are the hypotheses for this test and what is your conclusion?

e) Does more storks imply higher birth rates?

Below is a plot with two panels. On the x-axis in both plots is the area (in 1000km^2) of the different countries, and on the y-axes we have plotted Birth rate per 1000 people per year (left) and number of stork pairs (right).

f) Use the figures and write a **short** discussion about the relationship between number of stork pairs in a country and said country's birth rate.



Problem 5 A pension provider, *The Final Paycheck Co.*, is responsible for estimating the future liabilities of its retirement plans. To do so accurately, it must model the life expectancy of individuals in its portfolio. This requires an understanding of how mortality rates vary with age. Mortality rate is the usually calculated by age and year, often also split by gender, and it is defined as the number of people of a certain age that died in a certain year divided by the number of alive people of the same age at the beginning of that year (people at risk of dying).

One widely used model in demography and actuarial science is the Gompertz model, which describes the mortality rate $\mu(x)$ at age x as increasing exponentially:

$$\mu(x) = a\, e^{bx},$$

where a and b are parameters to be estimated. Traditionally, the Gompertz model has mainly been used for adult populations.

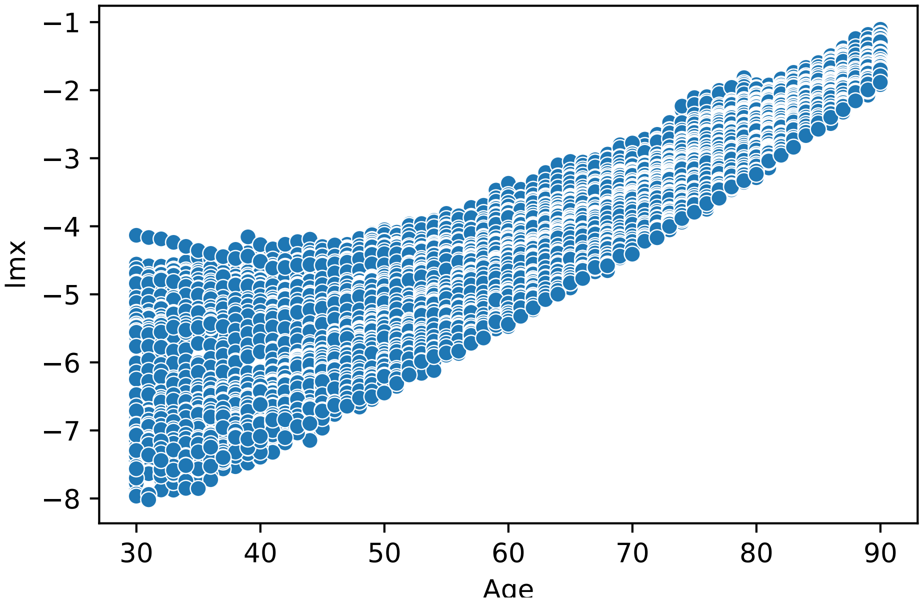
a) Find an expression for $\log \mu(x)$. What kind of relationship is there between logarithmic mortality rate $\log \mu$ and the age of a customer (x)?

The Final Paycheck Co. use official mortality data for Norway (publicly available from mortality.org). The dataset consists of the following columns

- `Age`: Integer age, between 30 and 90.
- `Year`: Integer year
- `Cohort`: Defined by Year-Age (integer). Birth-year of cohort.

- `mx`: Mortality rate for year and age group.
- `lmx`: Log mortality rate for year and age group.

We have plotted the log-mortality rate (`lmx`) against age (`Age`):



b) Below is Python output from fitting a Gompertz model. Specify the model with the estimated coefficients. Interpret the coefficients.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    lmx   R-squared:                       0.823
Model:                            OLS   Adj. R-squared:                  0.823
Method:                 Least Squares   F-statistic:                 5.055e+04
Date:                Fri,  8 Aug 2025   Prob (F-statistic):               0.00
Time:                        09:00:00   Log-Likelihood:                -10615.
No. Observations:               10858   AIC:                         2.123e+04
Df Residuals:                   10856   BIC:                         2.125e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -8.8024      0.022   -401.513      0.000      -8.845      -8.759
Age            0.0788      0.000    224.829      0.000       0.078       0.080
==============================================================================
Omnibus:                       10.131   Durbin-Watson:                   0.057
Prob(Omnibus):                  0.006   Jarque-Bera (JB):               10.109
Skew:                          -0.073   Prob(JB):                      0.00638
Kurtosis:                       3.031   Cond. No.                         222.
==============================================================================
```

c) According to the model, what is the log-mortality rate of a 70-year old? Can you think of reasons why the Gompertz model is mainly used for adult populations?

People are living longer now than they did in previous generations and females live longer than males. However, the current model treats age and mortality independently of when a person was born and gender. We add `Gender` to the dataset (Female/Male) and fit the following model:

The model can be formulated as:

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

where Gender is 0 if person is female and 1 if male and Cohort is the year they were born. The model is estimated using Python and the summary is printed below.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    lmx   R-squared:                       0.949
Model:                            OLS   Adj. R-squared:                  0.949
Method:                 Least Squares   F-statistic:                 8.022e+04
Date:                Fri,  8 Aug 2025   Prob (F-statistic):               0.00
Time:                        09:00:00   Log-Likelihood:                -8293.2
No. Observations:               21716   AIC:                         1.660e+04
Df Residuals:                   21710   BIC:                         1.665e+04
Df Model:                           5
Covariance Type:            nonrobust
======================================================================================
                         coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------
Intercept             35.9080      0.301    119.304      0.000      35.318      36.498
Gender[T.Male]         0.4694      0.017     27.467      0.000       0.436       0.503
Age                   -0.3441      0.005    -72.121      0.000      -0.353      -0.335
Age:Gender[T.Male]    -0.0029      0.000    -10.559      0.000      -0.003      -0.002
Cohort                -0.0236      0.000   -148.158      0.000      -0.024      -0.023
Age:Cohort             0.0002   2.54e-06     87.045      0.000       0.000       0.000
==============================================================================
Omnibus:                     2342.138   Durbin-Watson:                   0.560
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             3500.207
Skew:                           0.811   Prob(JB):                         0.00
Kurtosis:                       4.114   Cond. No.                     1.46e+07
==============================================================================
```

d) Discuss how this model captures improvements in longevity over time using the provided model equation and the printout summary. Why does the parameter associated with **Age** change signs?

The pension the company provides is a **life annuity**, which gives the customer a monthly payout as long as they live. The size of the monthly payout depends on the life expectancy of the customer.

A female customer (aged 70) calls the customer hotline - she has discovered that her husband (70) is receiving a higher monthly payout from the company pension compared to her and feels discriminated. They have checked, and all characteristics of their pension is equal except their gender and the monthly payout.

e) Use the printed summary from the fitted model and the provided equation to explain and quantify this difference.

Appendix: 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,

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

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,

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

The variance shortcut

The definition of the variance is

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

Often it is quicker to use the “shortcut” formula:

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

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:

```python
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

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

The correlation is

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