Problem Sheet 13 [1]

Exercise 1. A researcher wishes to estimate the proportion of all adults who own a cell phone. He takes a random sample of $1,572$ adults; $1,298$ of them own a cell phone, hence $1298/1572 \approx 0.83$ or about 83% own a cell phone.

1. What is the population of interest?
2. What is the parameter of interest?
3. What is the statistic involved?
4. Based on this sample, do we know the proportion of all adults who own a cell phone?

Solution 1.
1. All adults, we would like to *infer* something about the total population of adults.
2. The proportion of adults who own a cell phone.
3. The proportion computed from the sample, a function of our data, approximately 83%.
4. No we do not exactly know, we only approximate it from the sample.

Exercise 2. In a clinical trial, data collection usually starts at "baseline," when the subjects are recruited into the trial but before they are assigned to treatment or control. Data collection continues until the end of follow-up. Two clinical trials on prevention of heart attacks report baseline data on smoking, shown below. In one of these trials, the randomization did not work. Which one, and why?

	Number of persons	**Percent who smoked**
(i) Treatment	1,012	49.3%
(i) Control	997	69.0%
(ii) Treatment	995	59.3%
(ii) Control	1,017	59.0%

Solution 2. It appears that randomization did not work in the first trial. If the treatment had been assigned randomly, we would expect approximately the same proportion of any characteristic in both the treatment and control arms. While this expectation holds true in the second trial, in the first trial, smokers seem to have been more likely assigned to the control arm than to the treatment arm.

[1] Exercises are based on book *Statistics* by David Freeman, Robert Pisani, and Roger Purves and the course notes of Prof. Anthony Davison

Exercise 3. Breast cancer is one of the most common malignancies among women in the U.S. If it is detected early enough—before the cancer spreads—chances of successful treatment are much better. Do screening programs speed up detection by enough to matter?

The first large-scale trial was run by the Health Insurance Plan of Greater New York, starting in 1963. The subjects (all members of the plan) were 62,000 women aged 40 to 64. These women were divided at random into two equal groups. In the treatment group, women were encouraged to come in for annual screening, including examination by a doctor and X-rays. About 20,200 women in the treatment group did come in for the screening; but 10,800 refused. The control group was offered usual health care. All the women were followed for many years.

Results for the first 5 years are shown in the table below.

		Cause of Death			
		Breast cancer		**All other**	
		Number	**Rate**	**Number**	**Rate**
Treatment group					
Examined	20,200	23	1.1	428	21
Refused	10,800	16	1.5	409	38
Total	31,000	39	1.3	837	27
Control group	31,000	63	2.0	879	28

Table 1: Deaths in the first five years of the Health Insurance Plan screening trial, by cause. Rates per 1,000 women

Epidemiologists who worked on the study found that (i) screening had little impact on diseases other than breast cancer; (ii) poorer women were less likely to accept screening than richer ones; and (iii) most diseases fall more heavily on the poor than the rich.

1. Does screening save lives? Which numbers in the table prove your point?
2. Why is the death rate from all other causes in the whole treatment group ("examined" and "refused" combined) about the same as the rate in the control group?
3. Breast cancer (like polio, but unlike most other diseases) affects the rich more than the poor. Which numbers in the table confirm this association between breast cancer and income?
4. The death rate (from all causes) among women who accepted screening is about half the death rate among women who refused. Did screening cut the death rate in half? If not, what explains the difference in death rates?

Solution 3. 1. What can be concluded from the trial is that offering screening has an effect on breast cancer death rates (1.3 compared to 2.0).

The data seems to suggest that screening is effective, as the breast cancer death rate among those examined (1.1) is about half that of the control group (2.0). However, since individuals self-selected into the screening group, those who opted for screening may have had a natural resistance to breast cancer that might not apply to other diseases unrelated to screening itself. This natural resistance would mean that the observed reduction in breast cancer death rates

might result from this underlying factor, not from the screening. This phenomenon is known as a "confounder," as it is a common cause influencing both the treatment and the outcome.

That said, there is no plausible reason to assume that there exists a variable "breast cancer resistance" that would influence a participant's decision to attend the screening, their breast cancer death rates, however, not their "all other causes" death rates. Therefore, it is likely that the observed difference in death rates is primarily due to screening.

A better trial design would involve obtaining consent for screening and then randomizing treatment within the consenting group to eliminate potential confounders.

2. The offering of breast cancer screening should have no effect on deaths from other causes. Therefore, the mortality rates for non-breast cancer causes in the treatment and control groups should be approximately the same.

3. A comparison of the breast cancer death rates in the control group and the refusal group reveals that income may be associated with breast cancer mortality. According to point (ii) below the table, poorer women were more likely to refuse screening, meaning the refusal group has a higher proportion of poorer women compared to the control group. Note that individuals from either of these two groups were not screened. Since the breast cancer rate is lower (1.5 compared to 2.0) in the group with more poorer women, we can conclude that there is an association between income and breast cancer fatality.

4. No, because *association is not causation*. It is likely that individuals who accepted the screening are generally more *health-conscious* and, as noted in point (ii), are also wealthier on average. These factors contribute to their lower mortality rates from other causes, independent of the screening itself. This demonstrates the presence of a common cause influencing both the decision to undergo screening (the treatment) and the outcome (death from other causes). Therefore, comparing these two populations directly cannot accurately determine the effect of screening on deaths from other causes.

Exercise 4. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} Unif(a, b)$, with mean $E(X_j) = \frac{a+b}{2} = \theta$.
Find the mean square error (MSE) of the average $\overline{X}$ as an estimator of θ.

Solution 4. Using the bias-variance decomposition $\text{MSE}(\overline{X}; \theta) = b(\overline{X}; \theta)^2 + \text{Var}(\overline{X})$.
Since

$$\overline{X} = E(\frac{1}{n}\sum_{j=1}^{n} X_j) = n\frac{1}{n}E(X_j) = \theta,$$

the estimator is unbiased, i.e. the bias is zero. This means that the MSE is just equal to the variance.

The variance of a $Unif(a, b)$ distributed random variable is $\frac{(b-a)^2}{12}$. Similarly to Exercise 7 from Week 11, we can show that

$$\text{Var}(\overline{X}) = \text{Var}\left(\frac{1}{n}\sum_{j=1}^{n} X_j\right) = \frac{1}{n^2}\sum_{j=1}^{n}\text{Var}(X_j) = \frac{nVar(X_j)}{n^2} = \frac{Var(X_j)}{n} = \frac{(b-a)^2}{12n}.$$

Therefore $MSE(\overline{X}; \theta) = \frac{(b-a)^2}{12n}$

Exercise 5. A firm wants to hire an IT engineer but hasn't yet decided what annual salary to offer. Not wanting to offer a salary too far removed from the average salaries offered by other firms, the recruiter decides to ask 1000 engineers their annual salaries. The survey reveals that the average salary of the 1000 engineers is 48000 CHF. Suppose that you know that the standard deviation of IT engineer salaries is 12000 CHF.

1. Give a 95% confidence interval for the average salary.
2. Is it reasonable to say that the average salary is roughly 50000 CHF?

Solution 5. 1. Denote the average salary from the survey with $\overline{X}$. Since the standard deviation is known, the 95% confidence interval for the true average salary is

$$\text{CI} = \overline{x} \pm \Phi^{-1}\big(1 - \frac{1-0.95}{2}\big) SD(\overline{X}) = 48000 \pm 1.96 \cdot \frac{12000}{\sqrt{1000}} = (47256.23, 48743.77),$$

where we used that $Var(\overline{X}) = \frac{Var(X)}{n}$, where X is the random variable denoting an IT engineer's salary.

2. It depends on how you interpret roughly, but because the 95% confidence interval does not include 50000 CHF, I would not say that the true average IT engineer salary in Switzerland is roughly 50000 CHF.

Exercise 6. Suppose $x_1, \ldots, x_6$ are a random sample from the $\mathcal{N}(\mu_1, \sigma_1^2)$ distribution and $y_1, \ldots, y_5$ are a random sample from the $\mathcal{N}(\mu_2, \sigma_2^2)$ distribution, independent of the x-s. All the means are unknown, but suppose that the variances σ_1^2 and σ_2^2 equal 2.5 and 3. We find that

$$\overline{x} = \frac{1}{6}\sum_{k=1}^{6} x_k = 49.2, \quad \overline{y} = \frac{1}{5}\sum_{k=1}^{5} y_k = 48.4.$$

1. Construct 95% confidence intervals for μ_1 and μ_2.
2. Give the distribution of the difference $\overline{X} - \overline{Y}$ of the averages, and construct a 95% confidence interval for $\mu_1 - \mu_2$.

Solution 6. 1. The confidence intervals can be constructed as follows:

$$\text{CI}_{\mu_1} = \overline{x} \pm \Phi^{-1}\big(1 - \frac{1-0.95}{2}\big) SD(\overline{X}) = 49.2 \pm 1.96 \cdot \frac{\sqrt{2.5}}{\sqrt{6}} = (47.9, 50.5),$$

and

$$\text{CI}_{\mu_2} = \overline{y} \pm \Phi^{-1}\big(1 - \frac{1-0.95}{2}\big) SD(\overline{Y}) = 48.4 \pm 1.96 \cdot \frac{\sqrt{3}}{\sqrt{5}} = (46.9, 49.9),$$

where we used that $Var(\overline{X}) = \frac{Var(X)}{n}$ and $Var(\overline{Y}) = \frac{Var(Y)}{n}$.

2. Since both $\overline{X}$ and $\overline{Y}$ are the sum of normally distributed random variables divided by a constant, they are normally distributed. By the same argument, $\overline{X} - \overline{Y}$ has a normal distribution as well.

By the linearity of expectation

$$E(\overline{X}-\overline{Y}) = E\left(\frac{\sum_{i=1}^{6} X_i}{6}\right) - E\left(\frac{\sum_{i=1}^{5} Y_i}{5}\right) = \frac{6E(X_i)}{6} - \frac{5E(Y_i)}{5} = \mu_1 - \mu_2.$$

Due to the independence of X-s and the Y-s

$$Var(\overline{X}-\overline{Y}) = Var(\overline{X}) + Var(\overline{Y}) = \frac{Var(X)}{6} + \frac{Var(Y)}{5} = \frac{2.5}{6} + \frac{3}{5} = \frac{61}{60},$$

where we used the already stated relations $Var(\overline{X}) = \frac{Var(X)}{n}$ and $Var(\overline{Y}) = \frac{Var(Y)}{n}$.
To put everything together $\big(\overline{X}-\overline{Y}\big) \sim \mathcal{N}(\mu_1 - \mu_2, \frac{61}{60})$.
Then the confidence interval for the difference of the true means is

$$\mathrm{CI}_{\mu_1-\mu_2} = (\overline{x}-\overline{y}) \pm \Phi^{-1}\big(1-\frac{1-0.95}{2}\big) SD(\overline{X}-\overline{Y}) = (49.2-48.4) \pm 1.96 \cdot \sqrt{\frac{61}{60}} = (-1.2, 2.8)$$

Note: The confidence interval for the difference of the means is not the difference of the two confidence intervals for the means!