Probability and Statistics MATH-235
Autumn 2024 Mats J. Stensrud

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?

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%

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.

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

		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.

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?

Exercise 4. Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} Unif(a, b)$, 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.

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?

Exercise 6. Suppose $x_1, \dots, x_6$ are a random sample from the $\mathcal{N}(\mu_1, \sigma_1^2)$ distribution and $y_1, \dots, y_5$ 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 $\mu_1 - \mu_2$.