Test Mathematical Statistics (Module 5),
October 28, 2022, 8.45-11.45 h.

A formula sheet is added. A regular scientific calculator is allowed, a programmable calculator with graphical interface is not.

Part A: Basic concepts

(a) [1 point] What is an estimator?

(b) [2 points] The c.d.f. of a random variable X is $x \mapsto F(x) = P(X \leq x)$. In the lecture we discussed a strategy to construct estimators by replacing probabilities by relative frequencies. Using this principle, propose an estimator for the c.d.f. given i.i.d. data $X_1, \ldots, X_n$.

(c) [1 point] During the lecture, we worked with two estimators for the variance of i.i.d. observations. One estimator has a normalization factor $1/n$ and the other has a normalization factor $1/(n-1)$. Provide a statistical reason why both estimators are of interest.

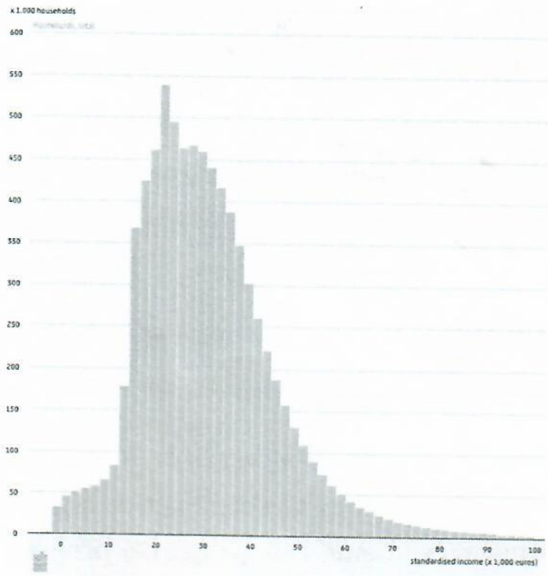
(d) [1 point] What is a dummy variable? Provide an example.

(e) [3 points] For a random variable X with mean μ and variance σ^2, the skewness is defined by $E(X-\mu)^3/\sigma^3$. What does the skewness measure? Compute the skewness for the Bernoulli distribution with success probability p. For which value(s) of p do we have positive/negative/vanishing skewness?

(f) [2 points] What defines a nonparametric test? What are the hypotheses in the Shapiro-Wilks test?

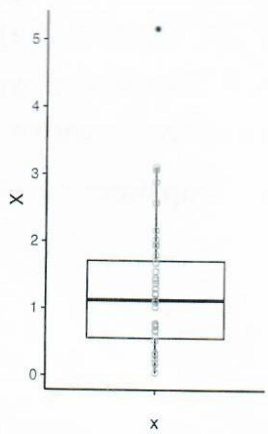
Part B: Visual interpretation of data

1. [1 point] A histogram of the Dutch income distribution is displayed below. How could the data be transformed such that the distribution resembles more a normal distribution?



Source:CBS

2. [2 points] Below a boxplot is displayed with data points plotted on top (every little circle is one data point). Make a drawing for the c.d.f. of this dataset that is as accurate as possible.



Part C: Theory

3. (a) [1 point] Show that for any $\theta > 0$

$$f_\theta(x) = \begin{cases} \theta x^{\theta-1}, & 0 < x \leq 1, \\ 0, & \text{otherwise} \end{cases}$$

is a p.d.f.

(b) [2 point] Determine the c.d.f. and the quantile function for the distribution with p.d.f. f_θ and find closed-form expressions for the first quartile and the median of this distribution.

(c) [3 point] Denote by P_θ the distribution with p.d.f. f_θ. Suppose that we observe an i.i.d. sample $X_1, \ldots, X_n \sim P_\theta$ and the parameter space is $\Theta = (0, \infty)$. Show that

$$\widehat{\theta} = -\frac{n}{\sum_{i=1}^n \log(X_i)}$$

is the maximum likelihood estimator. (*To get full points you also have to show that the estimator indeed maximizes the likelihood*)

(d) [3 points] Show that $1/\widehat{\theta}$ is an unbiased estimator for $1/\theta$. *Hint: You may use the fact that* $\lim_{x \downarrow 0} x^\alpha \log(x) = 0$ *for all* $\alpha > 0$.

(e) [3 points] Based on the first moment, compute the moment estimator for θ.

(f) [1 point] To compare the moment estimator with the MLE for estimation of θ, we would like to compute the mean-squared error (MSE) for both of them. However, this seems very hard and might even be analytically intractable. What else can we do to determine which estimator is better?

(g) [2 points] Show that the statistic $T(X_1, \ldots, X_n) = \sum_{i=1}^n \log(X_i)$ is sufficient.

(h) [3 points] Show that the α-level uniformly most powerful test for $H_0 : \theta = \theta_0$ vs $H_1 : \theta = \theta_1$ with $0 < \theta_0 < \theta_1$ is of the form $\sum_{i=1}^n \log(X_i) > c_\alpha$ for a suitable constant c_α.

4. [2 points] Consider the multiple linear regression model in matrix -vector notation $\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ and assume that $X^\top X$ is invertible and let $\widehat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top \mathbf{Y}$. Show that

$$\|\mathbf{Y} - X\boldsymbol{\beta}\|_2^2 = \|\mathbf{Y} - X\widehat{\boldsymbol{\beta}}\|_2^2 + \|X(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_2^2,$$

where $\|\cdot\|_2^2$ denotes the squared Euclidean norm. Deduce from this relation that $\widehat{\boldsymbol{\beta}}$ is the least squares estimator.

5. [3 points] Suppose we include the same explanatory variable two times in our multiple regression model. Working in the matrix-vector notation $\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ for multiple regression, show that the matrix $X^\top X$ is not invertible which means we cannot work with the usual definition of the least squares estimator $\widehat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top Y$.

6. Someone from the sales department approaches you for statistical consultation. To collect automatically tolls from cars passing a tunnel, they want to buy a system that automatically detects and reads the license plates of cars. Before they start negotiating the prices, they want to perform a quality test with all the available systems. For a given system, denote by p the (unknown) probability that it misreads a license plate. Moreover, it is assumed that for different cars, the chance to misread the license plate is independent of each other.

   (a) [1 point] Show that for a given system and n measurements, we can rewrite this as statistical model, where we are given n i.i.d. observations from the Bernoulli distribution with parameter p.

   (b) [4 points] Work in the same setup as in Part (a) and let $\alpha \leq 1/2$. Assuming that n is very large, find a value $T_{\alpha,n}$ such that if $p \leq 0.01$, we have

   $$P\big(\text{number of misread license plates} > T_{\alpha,n}\big) \leq \alpha + \text{a small remainder term.}$$

   Hint: In principle, $T_{\alpha,n} = \infty$ is a solution. Since this is statistically useless, it does not count. One should try to find the smallest possible $T_{\alpha,n}$ satisfying the inequality above.

   (c) [1 point] The quality test for the license plate detection systems is to remove all systems for which we have $(1-\alpha)$-confidence that $p > 0.01$. Formulate a procedure based on Part (b) to remove all such systems.

Formula Sheet Mathematical Statistics

Probability Theory

$E(X+Y) = E(X) + E(Y)$ $\quad$ $E(X-Y) = E(X) - E(Y)$ $\quad$ $E(aX+b) = aE(X) + b$

$var(X) = E(X^2) - (EX)^2$ $\quad$ $var(aX+b) = a^2 var(X)$

If X and Y are independent: $var(X+Y) = var(X) + var(Y)$, $var(X-Y) = var(X) + var(Y)$

$var(T) = E(var(T|V)) + var(E(T|V))$

Distribution	Probability/Density function	Range	$E(X)$	$var(X)$
Binomial (n,p)	$\binom{n}{x} p^x (1-p)^{n-x}$	$0,1,2,\dots,n$	np	$np(1-p)$
Poisson (μ)	$e^{-\mu}\mu^x/x!$	$0,1,2,\dots$	μ	μ
Uniform on (a,b)	$1/(b-a)$	$a<x<b$	$(a+b)/2$	$(b-a)^2/12$
Exponential (λ)	$\lambda\exp(-\lambda x)$	$x \geq 0$	$1/\lambda$	$1/\lambda^2$
Gamma (α,β)	$x^{\alpha-1}\exp\left(-\frac{x}{\beta}\right)/(\Gamma(\alpha)\beta^\alpha)$	$x>0$	$\alpha \times \beta$	$\alpha \times \beta^2$
Chi-square (χ_f^2)	is the Gamma distribution with $\alpha = f/2$ and $\beta = 2$			

Testing procedure in 8 steps

1. Give a probability model of the observed values (the statistical assumptions).
2. State the null hypothesis and the alternative hypothesis, using parameters in the model.
3. Give the proper test statistic.
4. State the distribution of the test statistic if H_0 is true.
5. Compute (give) the observed value of the test statistic.
6. State the test and **a.** Determine the rejection region or **b.** Compute the p-value.
7. State your statistical conclusion: reject or fail to reject H_0 at the given significance level.
8. Draw the conclusion in words.

Bounds for Confidence Intervals:

* $\hat{p} \pm c\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
* $\overline{X} \pm c\frac{S}{\sqrt{n}}$ and $\left(\frac{(n-1)S^2}{c_2}, \frac{(n-1)S^2}{c_1}\right)$
* $\overline{X} - \overline{Y} \pm c\sqrt{S^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$, with $S^2 = \frac{n_1-1}{n_1+n_2-2}S_X^2 + \frac{n_2-1}{n_1+n_2-2}S_Y^2$ or: $\overline{X} - \overline{Y} \pm c\sqrt{\frac{S_X^2}{n_1}+\frac{S_Y^2}{n_2}}$
* $\hat{p}_1 - \hat{p}_2 \pm c\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$
* (regression) $\hat{\beta}_i \pm c \times se(\hat{\beta}_i)$ and $\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm cS\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}}$, with $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{x}$.

  $S_{xy} = \sum_i (x_i - \bar{x})(y_i - \bar{y})$, $\hat{\beta}_1 = \frac{S_{xY}}{S_{xx}}$, $se(\hat{\beta}_1) = \frac{S}{\sqrt{S_{xx}}}$ and $S^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$.

Prediction intervals: $\overline{X} \pm c\sqrt{S^2\left(1+\frac{1}{n}\right)}$

(regression) $\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm cS\sqrt{1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}}$

Test statistics

* X (number of successes for a binomial situation)
* $T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$ and S^2
* $T = \frac{(\overline{X} - \overline{Y}) - \Delta_0}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$, with $S^2 = \frac{n_1 - 1}{n_1 + n_2 - 2} S_X^2 + \frac{n_2 - 1}{n_1 + n_2 - 2} S_Y^2$ or: $Z = \frac{(\overline{X} - \overline{Y}) - \Delta_0}{\sqrt{\frac{S_X^2}{n_1} + \frac{S_Y^2}{n_2}}}$
* $F = \frac{S_X^2}{S_Y^2}$
* $Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$, with $\hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$
* (regression) $T = \hat{\beta}_i / se(\hat{\beta}_i)$ and $F = \frac{SS_{Regr}/k}{SS_{Error}/(n-k-1)}$

Adjusted coefficient of determination: $R_{adj}^2 = 1 - \frac{n-1}{n-k-1} \times \frac{SS_{Error}}{SS_{Total}}$

Analysis of categorical variables

* 1 row and k columns: $\chi^2 = \sum_{i=1}^{k} \frac{(N_i - E_0 N_i)^2}{E_0 N_i}$ $(df = k - 1)$
* $r \times c$-cross table: $\chi^2 = \sum_{j=1}^{c} \sum_{i=1}^{r} \frac{\left(N_{ij} - \hat{E}_0 N_{ij}\right)^2}{\hat{E}_0 N_{ij}}$, with $\hat{E}_0 N_{ij} = \frac{\text{row total} \times \text{column total}}{n}$

  and $df = (r-1)(c-1)$.

Non-parametric tests

* Sign test: $X \sim B\left(n, \frac{1}{2}\right)$ under H_0
* Wilcoxon`s Rank sum test: $W = \sum_{i=1}^{n_1} R(X_i)$,

  under H_0 with: $E(W) = \frac{1}{2} n_1 (N+1)$ and $var(W) = \frac{1}{12} n_1 n_2 (N+1)$

Test on the normal distribution

* Shapiro – Wilk`s test statistic: $W = \frac{\left(\sum_{i=1}^{n} a_i X_{(i)}\right)^2}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}$