Resit Mathematical Statistics
November 12, 2021, 8.45-11.45 h.
Instructors: Annika Betken and Johannes Schmidt-Hieber

A formula sheet is added. A regular scientific calculator is allowed, a programmable calculator ("GR") is not.

Part A: Basic concepts

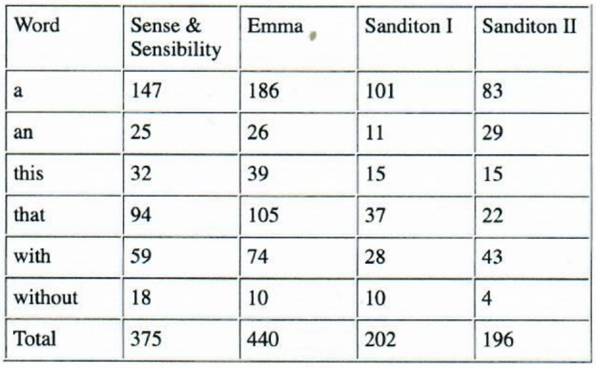
(a) [1 point] Provide an example of a categorical variable.

(b) [1 point] In a regression analysis with many explanatory variables, the coefficient of determination R^2 is 0.81 and the adjusted R^2 is 0.39. Out of these two statistics, which one should we report? Motivate your choice.

(c) [1 point] For which type of testing problem can Fisher's exact test be applied?

(d) [1 point] Describe the difference between uncorrelated and independent random variables.

(e) [1 point] There is a debate whether some of Jane Austen's novels might have been written by a ghost writer. The table below gives some word frequencies.



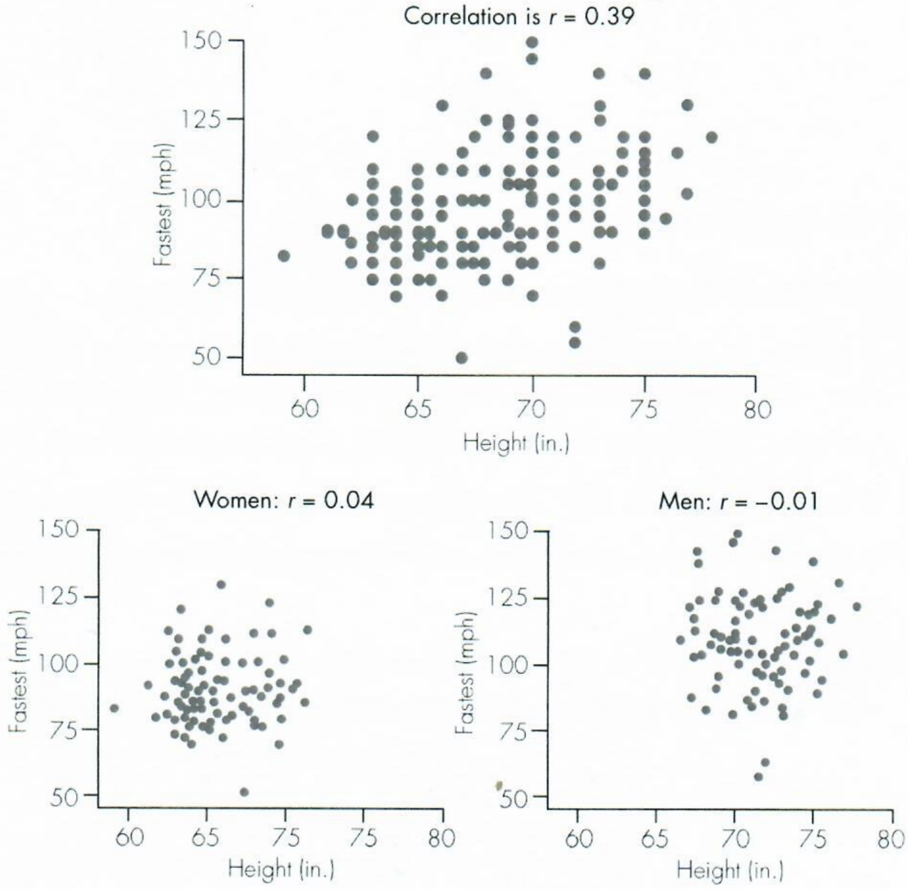
Word	Sense & Sensibility	Emma	Sanditon I	Sanditon II
a	147	186	101	83
an	25	26	11	29
this	32	39	15	15
that	94	105	37	22
with	59	74	28	43
without	18	10	10	4
Total	375	440	202	196

Which statistical test can be used for this dataset in order to check whether a ghostwriter has been involved.

5

Part B: Visual interpretation of data

1. [2 points] The height (in inches) and the fastest driven speed with a car (in mph) is measured for several students. The data are given in the scatterplot below next to two additional scatterplots that only look at the female and male participants in the study. For each of the three scatterplots the sample correlation is denoted by r. What do we learn from these plots about the relationship between height, gender and the fastest speed driven with a car?



Correlation is $r = 0.39$

Women: $r = 0.04$

Men: $r = -0.01$

2. [2 points] The random variable X is generated by the following procedure. Toss a fair coin, that is, heads an tails occur with probability 1/2. If heads appears, draw X from a $\mathcal{N}(2, 1)$ distribution and if tails appears draw X from a $\mathcal{N}(-2, 1)$ distribution. Make a plot of the probability density function of X (for the solution it is enough to get the shape right).

4

Part C: Theory

3. Suppose we observe n independent random variables $X_i \sim \text{Poisson}(\mu)$, $i = 1, \dots, n$, where $\text{Poisson}(\mu)$ denotes the Poisson distribution with intensity μ (see the formula sheet for the p.m.f., the expectation and the variance).

   (a) [2 points] Compute the maximum likelihood estimator $\widehat{\mu}$ for μ.

   (b) [2 points] Consider the class of estimators $a\widehat{\mu}$ with a a real number. For which value of a is the mean squared error (MSE) minimized? Let a^* be the value minimizing the MSE. Is $a^*\widehat{\mu}$ an estimator?

4. (a) [2 points] The p.d.f. of the exponential distribution is $f(x) = e^{-x}$ for $x \geq 0$ and $f(x) = 0$ for $x < 0$. Show that the quantile function is $Q(y) = -\log(1 - y)$.

   (b) [1 point] What is the first quartile, the median and the third quartile for the exponential distribution?

5. Consider the Gaussian multivariate regression model $Y = X\beta + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, I)$, assuming that $X^\top X$ is invertible. Recall that $\widehat{\beta}^{\text{MLE}} = (X^\top X)^{-1}X^\top Y$ is the MLE. The residuals are given by $Y - X\widehat{\beta}^{\text{MLE}}$.

   (a) [2 points] Show that the residuals are given by the formula

   $$\big(I - X(X^\top X)^{-1}X^\top\big)\varepsilon.$$

   (b) [2 points] Show that the distribution of the residuals is given by

   $$\mathcal{N}\big(0, I - X(X^\top X)^{-1}X^\top\big).$$

   (c) [1 point] Under the assumptions on linear regression, the residuals have consequently a normal distribution. In practice, the assumptions on linear regression are often not met and the residuals have a different distribution. To test whether the assumptions on the regression model are met, one could be tempted to apply Shapiro-Wilk's test on normality. Which assumption of Shapiro-Wilk's test on normality is violated?

   (d) [+2 Bonus] What could be done to make Shapiro-Wilk's test on normality applicable?

12(14)

6. [2 points] Fix $\alpha \in (1/2, 1)$. Let I_1 and I_2 be $1-\alpha$ confidence intervals for the parameters θ_1 and θ_2, respectively. Define the set $A = \{u + v : u \in I_1, v \in I_2\}$. Show that for the parameter $\rho = \theta_1 + \theta_2$,

$$P(\rho \in A) \geq 1 - 2\alpha.$$

This means that A is a $(1-2\alpha)$-confidence interval for ρ.

2